Universidad Autónoma de Madrid | Escuela de Doctorado

Tesis Doctoral

Deep Learning with Controlled Flows: Expressivity, Generalization and Generation

Antonio Álvarez López

Programa de Doctorado en Matemáticas

Dirección:
Enrique Zuazua Iriondo
Rafael Orive Illera

Madrid, 2026

Contact

Name	Antonio Álvarez López
Email	antonio.alvarezl@uam.es
Office	213, Module 17
Department	Departamento de Matemáticas
Faculty	Ciencias, Universidad Autónoma de Madrid
Address	C/ Francisco Tomás y Valiente, 7
Postal code	28049
City	Madrid, Spain

Ph. D. Advisors

Enrique Zuazua Iriondo

Universidad Autónoma de Madrid

Rafael Orive Illera

Universidad Autónoma de Madrid

Instituto de Ciencias Matemáticas, CSIC-UAM-UCM-UC3M

Funding

This thesis has been partially supported by the Spanish Ministry of Science, Innovation and Universities through the pre-doctoral fellowship FPU21/05673 and project PID2023-146872OB-I00; by the Madrid Government under the multiannual Agreement with UAM (Excellence of the University Research Staff, V PRICIT); and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 777822.

Agradecimientos

Embarcarme en un doctorado siempre fue—con naturales altibajos—una meta personal. Hoy, tres años y medio más tarde, el viaje llega a su destino final, y resulta inevitable echar la vista atrás para comprobar que quien lo comenzó no es exactamente el mismo que lo termina. Y aunque escribir una tesis exige un esfuerzo de introspección y largas horas de trabajo en solitario, la realidad es que difícilmente habría llegado hasta aquí sin haberme rodeado tan bien, y me siento afortunado por ello.

Asumo de antemano la injusticia de no poder nombrar a todo el mundo, por lo que me centraré, por fuerza, en aquellas personas más estrechamente vinculadas a mi experiencia doctoral. Espero que nadie sienta su ausencia como un olvido; tengo muy presente a cada persona que ha contribuido a moldearme. Sirvan estas líneas como agradecimiento colectivo antes de poner nombres y apellidos, comenzando, cómo no, por mis dos directores de tesis.

Enrique: realizar una tesis bajo la tutela de un matemático de tu talla ha sido, a partes iguales, una gran motivación y un gran desafío. Me he esforzado por aprovechar al máximo cada día. A tu lado he aprendido mucho de matemáticas, pero creo que tu influencia va más allá de lo puramente académico: me has aportado visión, claridad y determinación, y estoy convencido de que ese ha sido tu mayor regalo. Seguramente sea el tiempo quien descubra la huella que dejas en mi formación, pero recordar nuestros primeros días como director y alumno es una prueba irrefutable de que el *método Zumosol* ha surtido efecto.

Rafa: no solo has sido mi guía; también te has convertido en mi compañero. Nadie ha compartido conmigo tantas horas de trabajo, tantas reflexiones y tantas dudas—demasiadas por mi parte, lo confieso—desde septiembre de 2022. Soy consciente del esfuerzo que has hecho por abrirme puertas y brindarme las mejores oportunidades. Gracias por confiar en mí, por tu inagotable paciencia y por tener el consejo adecuado cada vez que me tocó remar a contracorriente. Al margen de dónde siga mi camino, ten por seguro que seguiré dándote la lata el tiempo que me permitas.

Agradezco también a los miembros de mi tribunal de tesis y, en particular, a los lectores, por haber accedido a revisarla. En especial, a Davide, por el tiempo que me ha dedicado en nuestras reuniones.

Sigo con mis colaboradores, esas personas que me han visto crecer más de cerca en los últimos años. Domènec, recuerdo con cariño cómo me acogiste en Londres durante mi primer curso. Allí lograste transmitirme tu pasión por este trabajo, y me formaste a base de muchas horas de trabajo conjunto que fueron mis primeros pasos en este mundillo. Borjan, gracias a tu acogida en Boston y París pude exponerme al más alto nivel. Desde tu estilo directo y sincero, me empujaste a ir más allá y a forjar mi propia cultura; tus consejos han sido oro. Martín, gracias por ser mi compadre dondequiera que nos viésemos; me llevo grandes historias contigo y la inmensa suerte de habernos entendido siempre tan bien, una sintonía que nos llevó a escribir juntos un *paper* aun siendo dos novatos. Y, ya fuera de mi *familia académica*, debo agradecer a Arselane, colaborador en mi primer trabajo, a quien le deseo una exitosa carrera. Y a Marcos por haber sido un apoyo constante, por haber creído en mí desde el primer momento y por su contagiosa tenacidad.

Tampoco olvido a quienes colaboran en los proyectos en curso que nutren este trabajo. A Lorenzo, por su nobleza y generosidad. Al profesor Qianxiao y a sus estudiantes, Jingpu y Liyan, por ser grandes anfitriones de un periodo investigador muy estimulante.

Sigo con mis demás *hermanos* de tesis. A Nico, por ser mi primer contacto real con esta comunidad durante un ya lejano viaje hacia Benasque en 2022. A Albert, con quien he tenido la suerte de compartir la mayor parte de este viaje y de ser testigos mutuos de nuestra evolución. A Ziqi y Ziqian, por vuestra energía positiva y hospitalidad. Finalmente, a los dos Danis: ha sido bonito veros crecer desde el día en que empezasteis (e incluso durante vuestros TFMs). Espero haberos ayudado, al menos, una parte de lo que otros hicieron por mí en su momento.

A mis compañeros, presentes y pasados, de la cátedra en Erlangen: Alberto, Christian, Christopher, Dragos, Enrico, Giovanni, Giulia, Giulio, Haitian, Hannes, Helon, Hugo, Ilias, Kang, Maureen, Michael, Nicolas, Pedro, Sebastián, Sol, Thys, Timothée, Yubiao, Yue, Zhengping, y un largo etcétera. Gracias por hacer la rutina más fácil, y muy especialmente por las cenas en el *Lê&Vi*.

Una de las cosas más bonitas que ofrece el doctorado es la posibilidad de viajar, conocer y colaborar con gente de todo el mundo. He podido disfrutar de ese privilegio, y así quiero dedicar unas líneas a mis amigos del Colegio de España: Gustavo, Jorge, Martín, Ramonda, entre otros; así como a mis compañeros de aventuras en Singapore: Abraham, Ainara, Dani, Diego, Javi, Louis, Manu, Martín, Mikel, Rodrigo, etc. Y también quiero mencionar aquí a Carlos Esteve y a los profesores Juan-Pablo y Florian, quienes han sido para mí un apoyo importante en diversas etapas de mi tesis.

Viajamos ahora a Madrid, y en concreto a Cantoblanco. Hacer un doctorado en la UAM tiene un encanto especial gracias a la calidad humana de su gente. Para empezar, agradezco a los grandísimos matemáticos del despacho 613: Diego, Jorge, Pablo y Eduardo Tablate—mentor, pero sobre todo, amigo leal. También a los ya doctores: Celia, Irene, Ismael, Jorge, Luis, el dúo vasco de Mikel y Peio, Rodrigo. A mis coetáneos: Alicia, Carlos, Claudia, David, Esther, Henrique, Itahisa, Javi Casado, Javi Minguillón. También a quienes, sin ser exactamente de la UAM, en mi mente parecen serlo: Andrés, Chema, Daniel, Enrique, Joaquín, José Antonio, Mafer, Pablo. Y, por último, a mi paisano Carlos. Llevamos ya unos diez años aguantándonos y compartiendo el día a día.

Dejando a un lado la academia, quiero acordarme de mis amigos. Primero, agradezco a aquellos que han estado presentes en el día a día madrileño y que han sido mi rincón de desconexión durante estos cuatro años: a Jorge, con quien llegué a Madrid hace ya cinco años, atravesando esta etapa hombro con hombro; a Iago, con quien comparto una amistad particular que nos ha llevado a ser compañeros de piso; a Adri, pura autenticidad y sintonía; al *capitán* Juan Dolz, Juanda, Marcos, Miguel; y también a Carlos y Edu, amigos dentro y fuera de la cancha. Y, por supuesto, a los imprescindibles que ya llevaban a mi lado aún más años: Pablo Morales, Pablo Romero, Víctor Carneiro.

A mis amigos de la carrera: Alfredo, Andrés, Antón, Antonio, Carlos, César, Iago, Joserra, Luis, Mar. Aunque ya no coincidamos tanto, es bonito ver cómo cada uno va forjando su propio camino.

A los amigos que me llevé de un memorable año en Granada, hoy repartidos por España: Adri, Alberto, Andrea, Clara, Diego, Javier, Liher, Manu, Matías, Nerea, Rubén, Sandra, y a los demás que formaron parte de aquella etapa. Sin olvidar a mis canarios: Aarón, Ale, Darío, Diego, Jorge, Prada.

A los que llevan ya muchos años a mi lado: Alba, Alicia, Andrés, Brais, Carlín, Carlos, Carol, Edu, Fer, Goris, Íñigo, Jaco, Loren, Lucas, Luís do Río, Luis García, Manu, María, Marta, Miguel, Nico, Pablo Pardo, Pablo Sánchez, Sifre y un injusto etcétera. Y, con especial cariño, a Cris.

A mi familia, el pilar sobre el que se asienta lo que soy. A mi madre, por su amor incondicional y por estar siempre pendiente de aquello que yo no alcanzo a ver. A mi padre, por su entrega y por ser mi mayor inspiración, mi primer guía y, seguramente, el porqué de esta tesis. A Alberto, mi hermano (el de verdad), por ser la persona con la que más recuerdos comparto. A todos mis tíos y mis primos. Y, por último, con un recuerdo especial, a mis abuelos Carmen y José, María y Jesús.

Esta tesis es, en buena medida, fruto de todos los aquí mencionados.

Antonio Álvarez López
Madrid, marzo de 2026.

Abstract

This thesis develops a mathematical framework for deep learning models based on controlled flows, with particular emphasis on neural ODEs. Its central idea is to view learning as the selection of time-dependent controls in a dynamical system so that the induced flow satisfies a data-driven objective. Under this viewpoint, expressivity becomes a controllability problem, while statistical performance is tied to the regularity and complexity of the underlying dynamics.

The first part of the thesis studies supervised learning through controlled flows acting on finite point clouds. For classification, we formulate the task as a relaxed simultaneous controllability problem and develop constructive methods with improved minimax complexity bounds, especially in high dimensions. For regression, we study exact interpolation for several neural ODE architectures and establish explicit trade-offs between width and depth, where depth is measured either by the time horizon or by the number of control switches. We also analyze autonomous regimes and derive complementary approximate and probabilistic controllability results. Finally, we address generalization through a cell controllability framework, which leads to structured nonparametric estimators and explicit population risk bounds.

The second part is devoted to generative modeling. In this setting, learning is formulated as the transport of a base distribution toward a target distribution through controlled flows. We investigate universality with respect to the Wasserstein, total variation, and Kullback–Leibler distances, showing how the class of reachable distributions depends on both the discrepancy and the geometry of the control family, and relating these properties to the point interpolation capacity of the system. We also obtain constructive approximation results for continuous normalizing flows under suitable tail assumptions.

The third part illustrates the scope of this dynamical framework in two directions: modeling distributional dynamics from snapshot data, motivated by digital health applications, and the mean-field analysis of Transformer-type attention dynamics on the sphere, where we characterize the discrete and sparse structure of stationary measures.

Altogether, this thesis provides a unified controllability-based perspective on expressivity, approximation, generalization, and generative modeling in continuous-depth learning systems.

Keywords: Learning and adaptive systems; Artificial neural networks and deep learning; Controllability; Attainable sets; Transport equations

Resumen

Esta tesis desarrolla un marco matemático para los modelos de aprendizaje profundo basados en flujos controlados, prestando especial atención a las EDOs neuronales. Su idea central es concebir el aprendizaje como la selección de controles dependientes del tiempo en un sistema dinámico, de forma que el flujo inducido cumpla con un objetivo dictado por los datos. Bajo esta perspectiva, la expresividad se convierte en un problema de controlabilidad, mientras que el rendimiento estadístico queda ligado a la regularidad y complejidad de la dinámica subyacente.

La primera parte de la tesis estudia el aprendizaje supervisado mediante flujos controlados que actúan sobre nubes finitas de puntos. Para la clasificación, formulamos la tarea como un problema de controlabilidad simultánea relajada y desarrollamos métodos constructivos con cotas de complejidad minimax mejoradas, especialmente en altas dimensiones. Para la regresión, estudiamos la interpolación exacta en diversas arquitecturas de EDOs neuronales y establecemos balances explícitos entre anchura y profundidad, donde la profundidad se mide bien por el horizonte temporal, bien por el número de conmutaciones del control. También analizamos regímenes autónomos y derivamos resultados complementarios de controlabilidad aproximada y probabilística. Finalmente, abordamos la generalización a través de un marco de controlabilidad por celdas, lo que conduce a estimadores no paramétricos estructurados y cotas explícitas para el riesgo poblacional.

La segunda parte está dedicada al modelado generativo. En este contexto, el aprendizaje se formula como el transporte de una distribución base hacia una distribución objetivo a través de flujos controlados. Investigamos propiedades de universalidad con respecto a las distancias de Wasserstein, variación total y Kullback-Leibler, mostrando cómo la clase de distribuciones alcanzables depende tanto de la métrica de discrepancia como de la geometría de la familia de controles, y relacionando estas propiedades con la capacidad de interpolación de puntos del sistema. También obtenemos resultados constructivos de aproximación para flujos normalizadores continuos bajo hipótesis adecuadas sobre el decaimiento de las colas.

La tercera parte ilustra el alcance de este marco dinámico en dos direcciones: el modelado de dinámicas distribucionales a partir de datos transversales, motivado por aplicaciones en salud digital; y el análisis de campo medio para dinámicas de atención tipo Transformer sobre la esfera, donde caracterizamos la estructura discreta y dispersa de las medidas estacionarias.

En conjunto, esta tesis proporciona una perspectiva unificada, basada en la controlabilidad, sobre la expresividad, aproximación, generalización y modelado generativo en sistemas de aprendizaje de profundidad continua.

Keywords: Sistemas de aprendizaje y adaptativos; Redes neuronales artificiales y aprendizaje profundo; Controlabilidad; Conjuntos alcanzables; Ecuaciones de transporte

Contents

Abstract | Resumen v

Notation xi

The Big Picture | La Perspectiva General 1

Contributions | Contribuciones 17

1 Preliminaries 29
- 1.1 Controlled flows . . . 29
- 1.2 Measure transport . . . 30
- 1.3 Neural ODE toolbox . . . 31

I Supervised Learning 37

Overview 39

2 Data Classification 41
- 2.1 Introduction . . . 41
- 2.2 Setup and main results . . . 42
- 2.3 Classification via canonical separability . . . 45
- 2.4 Classification by separability in general position . . . 51
- 2.5 Simulations . . . 55
- 2.6 Discussion . . . 58
- 2.7 Perspectives . . . 59
- 2.8 Proofs . . . 60

3 Data Interpolation 65
- 3.1 Introduction . . . 65
- 3.2 Setup . . . 66
- 3.3 Results . . . 69
- 3.4 Perspectives . . . 74
- 3.5 Proofs . . . 76

4 Generalization 95
- 4.1 Introduction . . . 95
- 4.2 Setup and main results . . . 97
- 4.3 Nonparametric estimation . . . 100
- 4.4 Flow realization and complexity . . . 105
- 4.5 Perspectives . . . 109
- 4.6 Proofs . . . 110

II Generative Modeling 119

Overview 121

5 Controllability for Measure Transport 123
5.1 Introduction . . . 123
5.2 Setup and main results . . . 124
5.3 Wasserstein distance . . . 127
5.4 Total variation distance . . . 133
5.5 Perspectives . . . 140
5.6 Proofs . . . 140

6 Continuous Normalizing Flows 167
6.1 Introduction . . . 167
6.2 Setup and main results . . . 168
6.3 Remarks and extensions . . . 170
6.4 Perspectives . . . 173
6.5 Proofs . . . 174

III Applications 185

Overview 187

7 Distributional Dynamics Modeling for Digital Health 189
7.1 Introduction . . . 189
7.2 Setup . . . 190
7.3 Model formulation . . . 190
7.4 Simulations . . . 196
7.5 Medical case study . . . 199
7.6 Perspectives . . . 203
7.7 Proofs . . . 205

8 Mean-field Landscape of Transformers 209
8.1 Introduction . . . 209
8.2 Setup . . . 211
8.3 Results . . . 215
8.4 Simulations . . . 221
8.5 Outlook: Mixture-of-Experts . . . 230
8.6 Perspectives . . . 233
8.7 Proofs . . . 233

Conclusions | Conclusiones 251

Bibliography 263
Software and code . . . 277

Index 279

Notation

Sets.

$[n]$	The set $[n] \coloneqq \{1, \dots, n\}$ for $n \in \mathbb{N}$.
x_i, $x^{(i)}$	Subscripts index elements of a set or sequence; superscripts index coordinates of a vector.
A^c	The complement of a set $A \subset \mathbb{R}^d$: $A^c \coloneqq \mathbb{R}^d \setminus A$.
$\lvert A \rvert$	Cardinality of a set A.
$\mathbf{1}_A$	Indicator function of a set $A \subset \mathbb{R}^d$: $\mathbf{1}_A(x) = 1$ if $x \in A$, and 0 otherwise.
$\mathbb{R}_*$	Subsets of $\mathbb{R}$ restricted by sign: $\mathbb{R}_{>0} \coloneqq (0, +\infty)$, $\mathbb{R}_{\geq 0} \coloneqq [0, +\infty)$, etc.
$[x, y]$, (x, y)	Closed and open line segments in $\mathbb{R}^d$, respectively, connecting two points $x \neq y$.
Δ^{d-1}	The probability simplex in $\mathbb{R}^d$: $\Delta^{d-1} \coloneqq \{x \in \mathbb{R}^d_{\geq 0} \mid \sum_{i=1}^d x^{(i)} = 1\}$.
S_N	The group of all permutations of $[N]$.

Vectors and matrices.

$u \cdot v$, $\langle u, v \rangle$	Euclidean scalar product for $u, v \in \mathbb{R}^d$.
$\lVert \cdot \rVert$, $\lVert \cdot \rVert_\infty$	Euclidean norm on $\mathbb{R}^d$; sup-norm on $\mathbb{R}^d$: $\lVert u \rVert_\infty \coloneqq \max_{i \in [d]} \lvert u^{(i)} \rvert$.
$\{e_1, \dots, e_d\}$	Canonical basis of $\mathbb{R}^d$.
$\mathbf{1}$	All-ones vector in $\mathbb{R}^d$.
$\mathscr{M}_{d \times d}(\mathbb{R})$	Real $d \times d$ matrices.
$\mathscr{S}_d^+(\mathbb{R})$	Symmetric positive definite matrices.
$\lVert \cdot \rVert_{\mathrm{op}}$	Operator norm on $\mathscr{M}_{d \times d}(\mathbb{R})$.
I_d, Id	Identity matrix of size $d \times d$; identity map in the relevant space.

Relations.

$x \wedge y$, $x \vee y$	$x \wedge y \coloneqq \min\{x, y\}$ and $x \vee y \coloneqq \max\{x, y\}$.
$\lceil z \rceil$, $\lfloor z \rfloor$	Smallest integer $\geq z$ (resp. largest integer $\leq z$).
a_+, a_-	$a_+ \coloneqq a \vee 0$ and $a_- \coloneqq (-a) \vee 0$.
$P \prec Q$	Partial order on $\mathscr{S}_d^+(\mathbb{R})$: $P \prec Q$ iff $Q - P$ is positive definite.

$f \approx g$	Informal approximate equality; when a precise meaning is needed, it will be specified explicitly (e.g. $f = g + o(1)$ or $\lvert f - g\rvert \le \varepsilon$).
$f \lesssim g$	There exists $C > 0$ (independent of the relevant parameters) such that $f \le Cg$.
$f \propto g$	$f = Cg$ for some constant $C > 0$.
$f \asymp g$	$f \lesssim g$ and $g \lesssim f$ (there exist $0 < c \le C$ such that $cg \le f \le Cg$).
$f = O(g)$	Asymptotic upper bound (in a regime clear from context, e.g. $n \to \infty$): there exist $C > 0$ and $N \in \mathbb{N}$ such that $\lvert f(n)\rvert \le C\lvert g(n)\rvert$ for all $n \ge N$.
$f = o(g)$	$f(n)/g(n) \to 0$ in the regime under consideration.
$f \sim g$	$f(n)/g(n) \to 1$ in the regime under consideration.

Function spaces.

$\mathscr{C}^k$, L^p, $W^{m,p}$	Standard (differentiable, Lebesgue, and Sobolev) function spaces on $\mathbb{R}^d$.
$\mathscr{C}_b(\mathbb{R}^d;\mathbb{R}^m)$	Space of bounded continuous maps $\mathbb{R}^d \to \mathbb{R}^m$.
$\mathscr{C}^{0,\alpha}(\mathbb{R}^d;\mathbb{R}^m)$	Space of α-Hölder maps for $\alpha \in (0,1)$ and globally Lipschitz maps for $\alpha = 1$.
$\mathscr{C}^{1,1}(\mathbb{R}^d;\mathbb{R}^m)$	Space of differentiable maps with globally Lipschitz derivative.
$\mathrm{Lip}(f)$, $\mathrm{Lip}_x(f)$	Global Lipschitz constant of a map f; Lipschitz constant of f with respect to the variable x (uniformly in other variables).
$\mathcal{X}_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{R}^m)$	For any function space $\mathcal{X}$ as above: those $f : \mathbb{R}^d \to \mathbb{R}^m$ such that $f\lvert_K \in \mathcal{X}(K)$ for every compact set $K \subset \mathbb{R}^d$.
$\mathcal{X}_{\mathrm{nd}}(\mathbb{R})$	For any function space $\mathcal{X}$ as above being $d = 1$: the subset of functions $f \in \mathcal{X}(\mathbb{R})$ that are nondecreasing.

Measure and probability.

$\mathcal{B}(\mathbb{R}^d)$	Borel σ-algebra on $\mathbb{R}^d$.
$\mathcal{P}(\mathbb{R}^d)$	Space of Borel probability measures on $\mathbb{R}^d$. We use μ or ν for its elements.
$\mathcal{L}^d$	Lebesgue measure on $\mathbb{R}^d$.
$\mathrm{Unif}(\Omega)$	Uniform probability measure on some $\Omega \in \mathcal{B}(\mathbb{R}^d)$ with $0 < \mathcal{L}^d(\Omega) < \infty$, that is, $\mathrm{Unif}(\Omega) \coloneqq \frac{\mathbf{1}_\Omega}{\mathcal{L}^d(\Omega)}\mathcal{L}^d$.
$\mathcal{N}(m,\Sigma)$	Gaussian measure on $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance $\Sigma \in \mathscr{S}_d^+(\mathbb{R})$.
$\mathcal{N}(\cdot \mid m,\Sigma)$	Gaussian density on $\mathbb{R}^d$ with mean $m \in \mathbb{R}^d$ and covariance $\Sigma \in \mathscr{S}_d^+(\mathbb{R})$: $\mathcal{N}(x \mid m,\Sigma) \coloneqq (2\pi)^{-d/2}\det(\Sigma)^{-1/2}e^{-\frac{1}{2}(x-m)^\top\Sigma^{-1}(x-m)}, \qquad x \in \mathbb{R}^d.$
$\mu \ll \nu$	Absolute continuity: $\mu(A) = 0$ for every $A \in \mathcal{B}(\mathbb{R}^d)$ such that $\nu(A) = 0$.
$\mu_n \rightharpoonup \mu$	Weak convergence: $\int f \,\mathrm{d}\mu_n \to \int f \,\mathrm{d}\mu$ for all $f \in \mathscr{C}_b(\mathbb{R}^d)$.

$\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$	Lebesgue-absolutely continuous measures. We typically use ρ to denote their densities (i.e., $\mu = \rho\mathcal{L}^d$).
$\mathcal{P}_p(\mathbb{R}^d)$	Measures with finite p-th moment: $\{\mu \in \mathcal{P}(\mathbb{R}^d) \mid \int \|x\|^p \,\mathrm{d}\mu(x) < \infty\}$.
$\mathsf{T}_{\#}\mu$	Pushforward of μ by a measurable map $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$: $(\mathsf{T}_{\#}\mu)(A) \coloneqq \mu(\mathsf{T}^{-1}(A))$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.

Calculus on the sphere.

$\mathbb{S}^{d-1}$	Unit sphere in $\mathbb{R}^d$.
σ_d	Uniform probability measure on $\mathbb{S}^{d-1}$.
$\mathsf{T}_x\mathbb{S}^{d-1}$	Tangent space to the sphere at $x \in \mathbb{S}^{d-1}$: $\mathsf{T}_x\mathbb{S}^{d-1} \coloneqq \{v \in \mathbb{R}^d \mid v \cdot x = 0\}$.
$\mathsf{T}\mathbb{S}^{d-1}$	Tangent bundle of the sphere: $\mathsf{T}\mathbb{S}^{d-1} \coloneqq \{(x, v) \in \mathbb{S}^{d-1} \times \mathbb{R}^d \mid v \in \mathsf{T}_x\mathbb{S}^{d-1}\}$.
$\mathsf{P}_x^{\perp}$	$\mathsf{P}_x^{\perp} \coloneqq \mathsf{Id} - xx^{\top}$ is the orthogonal projection onto $\mathsf{T}_x\mathbb{S}^{d-1}$ for each $x \in \mathbb{S}^{d-1}$.
$\nabla_{\mathbb{S}} f$	Riemannian gradient: $\nabla_{\mathbb{S}} f(x) \coloneqq \mathsf{P}_x^{\perp} \nabla \tilde{f}(x)$ for differentiable $f : \mathbb{S}^{d-1} \to \mathbb{R}$, where $\tilde{f}$ is any differentiable extension of f to a neighborhood of x.
$\mathrm{div}_{\mathbb{S}}\, v$	Spherical divergence of a tangent field v, also defined via differentiable extensions.

Deep learning and flows.

$\mathcal{X}$, $\mathcal{Y}$	Input space (typically $\mathbb{R}^d$); label space (e.g., $\mathbb{R}^m$ or a finite set).
$\mathcal{D}$, $\mathcal{D}_N$	Dataset. The subscript in $\mathcal{D}_N$ indicates size N (e.g., $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$).
σ	Activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Acts component-wise when applied to vectors.
p, q	Width (number of neurons) of each hidden layer: p for one-layer models, (p, q) for two-layer models.
L	Depth parameter: number of layers (discrete) or time horizon (continuous).
θ, $\theta(\cdot)$	Network parameters (weights/biases). Written as $\theta(t)$ when time-dependent.
v_θ, Φ_t^θ	Vector field parameterized by θ; Φ_t^θ is the associated flow map at time t.
$\mathcal{U}$	Set of admissible input controls (i.e., the space to which $\theta(\cdot)$ belongs).
$\mathcal{V}$	Family of admissible vector fields (control family).
$\mathcal{V}^{\bullet}$	Superscript indicates the specific activation function: use + if $\sigma(s) = s_+$, use tr if $\sigma_{\mathrm{tr}}(s) = 1 \wedge s_+$, use $\triangle$ if $\sigma_{\triangle}(s) = (1 - \|s - 1\|)_+$. If omitted, σ is generic.
$\mathcal{V}_{\bullet}$	Subscript indicates architectural constraints: p (width per layer), 1 (single neuron per layer), a (autonomous), sa (semi-autonomous), ℓ (linear-in-control).
μ_{B}, ρ_{B}	Base (or reference) probability measure and its density.
$\mu_\star$, $\rho_\star$	Target probability measure and its density.
μ_t^θ, ρ_t^θ	Pushforward measure and density at time t evolved by the flow: $\mu_t^\theta \coloneqq (\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}$.

The Big Picture

> *"Michael Jordan,* [...] *one of the leading figures in modern machine learning, advocates* [...] *prioritizing the conjunction of this 'intelligence augmentation' approach with the development of an 'intelligent infrastructure', derived from the constant and ubiquitous digitization of all human activity, rather than focusing exclusively on the development of strong AI* [...]. *Or, if you will, that we should worry less about 'science fiction dreams or superhuman nightmares' and more about better understanding these technologies and guiding their development as they become increasingly present, influential, and disruptive."*
>
> — José R. Dorronsoro
> Inaugural lecture at Universidad Autónoma de Madrid
> Academic year 2006–2007 [Dor22]

The learning paradigm

Over the last few decades, we have witnessed a silent revolution that began to accelerate in the 2010s and has, in recent years, fundamentally transformed society, driven by the ubiquitous integration of technology into every facet of modern life, from science and education to healthcare. Under the broad public label of *artificial intelligence* (AI), much of this change is powered by *machine learning* (ML): the discipline concerned with algorithms that infer patterns from data to make predictions or decisions, rather than relying on hard-coded rules [Mit97].

First coined by Arthur Samuel in 1959 ([Sam59]), the concept of *machine learning* has evolved into a highly formalized discipline that, from a modern mathematical perspective, is best viewed as a framework for high-dimensional approximation and inference. Its core premise is to use data generated by a process that lacks an explicit or tractable mathematical formulation—typically modeled as an underlying function or probability law—to build a parametric model that reproduces observed behaviors and generalizes to unseen scenarios.

This approach marks a conceptual departure from numerical analysis. Whereas classical methods typically seek to approximate known continuous operators and derive error bounds from first principles, ML must infer structure solely from finite (and potentially noisy) samples. This creates a paradox: we possess *recipes* that work remarkably well, yet the momentum is largely empirical. Therefore, the major open challenge today lies in building a rigorous mathematical foundation capable of explaining *why* these methods work, *when* they fail, and *how* they should be improved.

Despite the theoretical gap, its success is undeniable: ML has redefined the state of the art across a vast spectrum of applications. In computer vision and natural language processing, deep models now drive everything from medical diagnostics to large-scale language models. Generative models have moved from research prototypes to widely deployed tools able to synthesize text, images, audio, and code. Beyond consumer technology, ML has become integral to scientific and industrial workflows, from simulation-based inference to surrogate modeling and decision support.

For the mathematician, this proliferation of applications poses a unique challenge. The task extends beyond solving a well-posed problem; it requires abstracting empirical phenomena into rigorous structures. In a landscape where heuristics and scaling laws often dictate practice, analysis aims to distinguish the mechanisms that are truly essential from mere implementation details.

Providing a comprehensive historical review of machine learning is beyond the scope of this thesis. The field's trajectory is complex, marked by an immense multitude of contributors, foundational breakthroughs, and periods of stagnation known as *AI winters*. Rather than a historical account, our objective is to present a mathematically motivated introduction to several of the most prominent models. For an exhaustive historical overview, we refer the reader to [Fra20].

To structure this exploration, it is convenient to distinguish a few canonical learning paradigms (for a comprehensive introduction, we refer to standard references such as [Bis06; HTF09; Bac24]). In *supervised learning*, one observes pairs of input–label data (x_i, y_i)—where the input x_i represents the observed features (e.g., an image or a physical state) and the label y_i represents the desired output or category—and seeks a predictor that generalizes beyond the data; this is the setting of classification and regression. In *unsupervised learning*, one observes only the inputs (x_i) and aims to extract structure: low-dimensional representations, clusters, or a probabilistic model for the data distribution. A particularly important unsupervised task is *generative modeling*, where one seeks to learn (an approximation of) the data distribution itself in order to generate realistic synthetic samples [GBC16]. Finally, in *reinforcement learning*, data are collected adaptively through interactions with an environment, and learning is driven by reward feedback [SB18].

From neurons to networks

While these paradigms differ in their objectives and available data, they currently converge on a common modeling framework. At the heart of their empirical success lies *deep learning*—a term that appears in the literature at least as early as [Dec86]—which encompasses a class of machine learning methods relying on high-capacity models trained via gradient-based optimization [LBH15; BGC17].

These models are built upon artificial *neural networks*. While modern architectures are best understood as flexible, hierarchical mathematical functions composed of simple nonlinear operations rather than literal biological models, their early inspiration stems from the interplay of neurons in the human nervous system—as their name evokes.

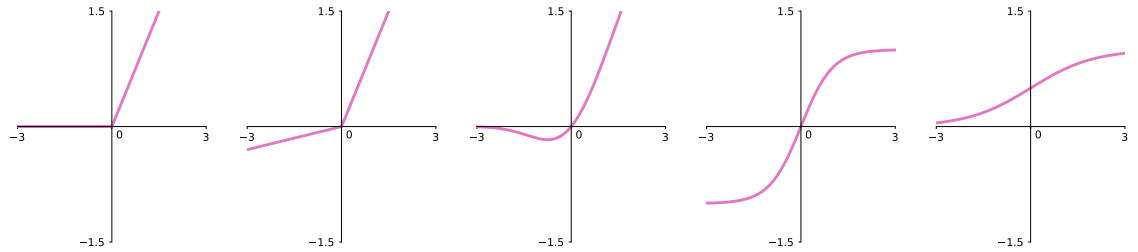
Historically, the conceptual foundation for these units was laid by McCulloch and Pitts [MP43]. They abstracted the local action of biological neurons—propagating or canceling impulses—into simple *threshold* logic gates, demonstrating that networks of such units could compute boolean functions. While this established the template for neural computation, their model relied on fixed, hand-crafted connections; it possessed no mechanism for adaptation or learning.

The critical shift from static logic circuits to adaptive models was achieved by Rosenblatt in the late 1950s [Ros58]. Building upon this conceptual template, he introduced the *perceptron*. This represented the true *germ* of machine learning: a parametric model paired with an explicit mathematical rule to automatically adjust its weights, embodying the core idea that a decision mechanism could be *fit from data* rather than explicitly programmed.

In its simplest form, a perceptron is any parameterized function of the form

$$f_\theta(x) = w\,\sigma(a \cdot x + b), \tag{1}$$

where $x \in \mathbb{R}^d$ is the state, $\theta \equiv (w, a, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ is the parameter, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a fixed nonlinearity called the *activation function*. Historically, smooth and bounded functions were



(a) ReLU $s \mapsto s_+$ (b) Leaky ReLU $s \mapsto s \vee (0.1 \cdot s)$ (c) GeLU $s \mapsto s\,\Phi(s)$ (d) Tanh $s \mapsto \frac{e^s - e^{-s}}{e^s + e^{-s}}$ (e) Sigmoid $s \mapsto \frac{1}{1+e^{-s}}$

Figure 1: Representation of some activation functions $\sigma : \mathbb{R} \to \mathbb{R}$. For GeLU, $\Phi(s) = (1 + \operatorname{erf}(s/\sqrt{2}))/2$ is the cumulative distribution function of the standard Gaussian.

commonly employed (e.g., tanh or the logistic sigmoid). However, in modern deep learning, the predominant standard is the *rectified linear unit* (ReLU). The *Gaussian error linear unit* (GeLU) is also common, particularly in Transformer architectures [HG16]. The analytic expressions of these functions, along with other popular alternatives, are detailed in Figure 1.

In approximation theory, functions of the form (1) are also known as *ridge functions* [Pin15]. The standard ML terminology calls them *neurons*, and interprets $a \cdot x + b$ as an affine pre-activation, σ as a nonlinear firing mechanism, and w as an output modulation.

A single neuron only yields translations and rescalings of the activation functions depicted in Figure 1, and therefore possesses limited expressive power. The natural next step is to superimpose p such units, which leads to the family of *shallow* (*single-hidden-layer*) networks of prescribed *width* $p \in \mathbb{N}$:

$$\mathcal{F}_p \coloneqq \left\{ x \mapsto f_\theta(x) = \sum_{i=1}^{p} w_i\, \sigma(a_i \cdot x + b_i), \qquad \theta \equiv (w_i, a_i, b_i)_{i \in [p]} \in (\mathbb{R} \times \mathbb{R}^d \times \mathbb{R})^p \right\}. \tag{2}$$

Mathematically, $\mathcal{F}_p$ constitutes a class of *nonlinear* models. This contrasts sharply with classical *linear* approximation schemes—such as Fourier series, splines, or polynomial interpolation—where the dictionary of basis functions is fixed a priori, and only the coefficients are learned. In a shallow network, the dictionary is *adaptive*: the inner parameters a_i and b_i allow the basis functions $x \mapsto \sigma(a_i \cdot x + b_i)$ to shift and reshape, effectively tuning the basis itself to the target function.

Despite its flexibility, the architecture remains structurally elementary. This rendered shallow networks the primary candidate for early rigorous analysis, particularly regarding density results (known in ML parlance as *universal approximation theorems*). The foundational results established that, as the width p increases, these networks can approximate target functions with arbitrary precision.

From the perspective of approximation theory, the fundamental question of whether the span of translations and scalings of a single base function is dense dates back at least to 1932 with Norbert Wiener's Tauberian theorems in [Wie32], which settled the problem for Lebesgue spaces (L^1 and L^2). While Wiener established the mathematical roadmap for this type of universality, his Fourier-based approach inherently requires the base function to be integrable. This condition is necessarily violated by the standard activation functions employed in modern neural architectures, such as those depicted in Figure 1, which do not decay at infinity.

Addressing these specific gaps for neural networks—namely, the use of non-integrable activations and the requirement of uniform convergence on compact sets—Cybenko in [Cyb89] was the first to explicitly prove the density of $\mathcal{F} \coloneqq \bigcup_{p\geq 1} \mathcal{F}_p$ for *sigmoidal* activations. Hornik subsequently extended this to broader classes of activations and stronger topologies in [Hor93]. A definitive characterization was later provided by Leshno et al. in [Les+93], where the authors proved a necessary and sufficient condition: the density property holds if and only if the activation σ is not a polynomial. For a comprehensive treatment of these ridge-function approximations, the monograph [Pin99] by Pinkus remains the standard reference.

The power of depth

By composing multiple layers, one obtains *deep* feedforward networks. Formally, a network of *depth* L is defined as the function:

$$f_\theta(x) = T_L \circ \sigma \circ T_{L-1} \circ \sigma \circ \cdots \circ \sigma \circ T_1(x), \tag{3}$$

where $x \in \mathbb{R}^d$, each $T_\ell(y) = A_\ell y + b_\ell$ is an affine map between Euclidean spaces of prescribed dimension, and σ acts component-wise. When the T_ℓ in (3) are implemented by fully-connected linear layers (i.e. A_ℓ is an unconstrained matrix), such feedforward networks are commonly referred to as *multilayer perceptrons* (MLPs). A graph representation of this architecture is provided in Figure 2a.

The transition from summation (width) to function composition (depth) unlocks a dramatic increase in the expressivity of the function class. A compelling analogy can be found in the algebra of polynomials. Summing monomials of a fixed degree d, we are inevitably restricted to the subspace of polynomials of that specific degree. In contrast, composing two such polynomials has a multiplicative effect on the degree. Consequently, the size of the class grows exponentially with depth, whereas it only grows linearly with width.

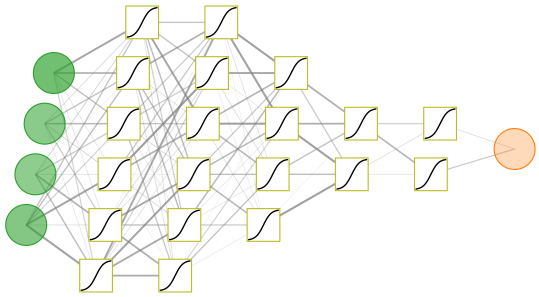
This analogy suggests that deep architectures are more efficient at representing structurally complex functions. In the realm of neural networks, the compositional nature translates into a hierarchical reuse of intermediate features, which is central to the success of modern deep learning.

Despite this theoretical advantage, the ability to *represent* complex functions is only half the story; one must also be able to *train* the model efficiently. For decades, the potential of deep architectures was hindered by the difficulty of solving the underlying high-dimensional nonconvex optimization problem. To address this, *training* is typically formulated as the minimization of a loss function via variants of *gradient descent*.

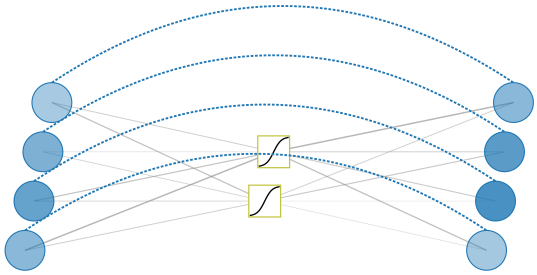
The practical feasibility of this approach relies on the efficient computation of gradients. While the algorithmic breakthrough that enabled multi-layer optimization was the popularization of the *backpropagation* algorithm [RHW86]—essentially a recursive application of the chain rule—its full potential remained latent until large-scale and parallel computing resources to optimize these parameters became widely available. This synergy between algorithmic efficiency and raw computational power is what ultimately propelled neural networks to their current dominance.

Residual networks and the continuous-depth viewpoint

Even with backpropagation, severe instabilities emerge as depth increases. The most notorious bottlenecks are the *vanishing* and *exploding* gradient phenomena [BSF94]. Because gradients are computed via the iterated chain rule, they involve a sequential product of layer-wise Jacobian matrices. Unless the spectrum of these matrices remains close to unity, the gradient norm usually



(a) Graph representation of an MLP with an input layer in $\mathbb{R}^4$ (each circle is a coordinate), followed by $L = 5$ hidden layers of variable width using $\sigma = \tanh$, and a scalar output. Weights are edges and opacity is proportional to their magnitude.



(b) A single ResNet block. The state neurons evolve through one hidden layer of width $p = 2$. Identity skip connections are depicted as curved arcs.

Figure 2: Comparison of neural network architectures.

undergoes exponential decay or growth with depth. This instability renders the optimization of deep architectures ill-conditioned.

A major turning point was the introduction of *residual networks* (ResNets) in the breakthrough paper [He+16]. ResNets modify the layer-to-layer update by adding *skip connections* (see Figure 2b). Writing x_ℓ and θ_ℓ for the feature representation and the parameters at layer ℓ, a residual network with L layers and input $x_0 \in \mathbb{R}^d$ takes the form

$$x_{\ell+1} = x_\ell + f_{\theta_\ell}(x_\ell), \qquad (\ell = 0, \dots, L-1), \tag{4}$$

where f_{θ_ℓ} is a nonlinear map (typically a small feedforward network with only a few layers). The identity term provides a direct path for gradient propagation and empirically stabilizes training, enabling networks with hundreds of layers.

From a mathematical perspective, (4) is the forward *Euler discretization* of an *ordinary differential equation* (ODE), up to a choice of time scaling (e.g., conventionally setting the step size to $\Delta t = 1$). This observation motivates *continuous-depth* models: interpreting depth as an artificial time variable $t > 0$ and letting the parameters vary continuously in time leads to neural ODEs [E17; HR18; Che+18]. A *neural ODE* evolves an initial input $x_0 \in \mathbb{R}^d$ governed by the system

$$\begin{aligned} \dot{x}(t) &= v_{\theta(t)}\big(x(t)\big), \qquad (t \geqslant 0), \\ x(0) &= x_0, \end{aligned} \tag{5}$$

where $\theta(\cdot)$ is a time-dependent parameter and v_θ is a *vector field* parameterized by a neural network[-1]. We deliberately switch notation from f_θ in (4) to v_θ in (5) to emphasize the dynamical perspective: the model specifies a *vector field* whose solution defines a *flow* $\Phi_t^\theta(x_0) \coloneqq x(t)$. Operationally, this flow acts as a continuum of *infinitesimal* neural network layers.

The continuous framework of (5) is often more tractable analytically, and places deep learning in the orbit of *control theory* [FZ03]. Classically, control studies *dynamical systems* driven by an external *input*,

$$\dot{x}(t) = v(x(t), u(t)), \qquad (t \geqslant 0), \tag{6}$$

[-1]The field v_θ can also incorporate known physical terms (yielding hybrid models) or impose specific mathematical structures. In this thesis, however, we will focus on the canonical case of a standard neural network.

and asks whether one can steer the state to a desired target (*reachability/controllability*, see [Son98]) and how to do so while optimizing a performance criterion (*optimal control*, see [Tré05]). Standard references include [Pon+62; Kha02; AS04; Ber17].

In this thesis, we will focus particularly on the first concept, which is fundamentally tied to the feasibility of the learning problem in machine learning. Broadly speaking, we say that the system (6) is *controllable* if, for any initial state $x_0 \in \mathbb{R}^d$ and any desired target state $x_1 \in \mathbb{R}^d$, there exists an admissible control signal $u(\cdot)$ and a finite time $T > 0$ such that the corresponding trajectory satisfies $x(0) = x_0$ and $x(T) = x_1$. A seminal reference establishing the algebraic foundations of this property for linear systems is [Kal60]. In the nonlinear setting, the geometric viewpoint is classically rooted in works such as [Cho40; JS72]; a modern reference is [Cor07].

In the neural ODE setting, the time-dependent parameter $\theta(\cdot)$ plays the role of the control input $u(\cdot)$. From this viewpoint, training becomes an optimal control problem: the goal is to select a control trajectory $\theta(\cdot)$ such that the induced flow map minimizes a learning objective that depends on the terminal map [E17; Zua22].

This viewpoint has proved useful both conceptually and technically: it connects expressivity to the notion of controllability [AS22; RZ23; Che+25c], and it supports optimality conditions and training algorithms based on Pontryagin-type maximum principles (including mean-field formulations that reflect the probabilistic nature of learning) [EHL18; Li+18; Ben+19; Bon+23].

Computationally, this perspective frames backpropagation as the integration of an *adjoint backward equation*. Popularized in [Che+18], this method allows for training with constant memory cost—independent of the number of (discretized) layers—by reconstructing trajectories reversibly, though often at the cost of numerical stability. A theoretical discussion of these issues can be found in [Mas+20].

Finally, when the input x is random, (5) transports *its law* through time. This observation leads naturally to *continuous normalizing flows*, introduced alongside neural ODEs in [Che+18] and developed further in scalable variants such as [Gra+19; Onk+20], where one learns dynamics that push forward a simple *base distribution* towards the *data distribution*.

Beyond perceptrons

There is no single *neural network* architecture. Many successful families arise by replacing the affine maps in (3) with structured operators adapted to a specific data modality.

Convolutional neural networks (CNNs), for instance, replace fully-connected linear layers by convolutions and weight sharing, resulting in *convolutional layers* of the form

$$x_{\ell+1} = \sigma\big(A_\ell * x_\ell + b_\ell\big), \qquad (\ell = 0, \dots, L-1), \tag{7}$$

where $*$ denotes a discrete convolution (applied over space or time indices). Thanks to encoding translation equivariance and locality, CNNs have been central in vision since early architectures [Lec+98] and became a cornerstone of the deep learning surge after large-scale image recognition breakthroughs [KSH12].

Sequential data (text, speech, sensor streams) prompted recurrent models, where the same transformation is iterated over time and the architecture is explicitly dynamical. A basic *recurrent neural network* (RNN) can be written as

$$\begin{aligned} h_{\ell+1} &= \sigma\big(A_h h_\ell + A_x x_\ell + b\big), \\ y_\ell &= A_y h_\ell \end{aligned} \qquad (\ell = 0, \dots, L-1), \tag{8}$$

where h_ℓ is a *hidden state*, x_ℓ is the input at step ℓ, and the weight matrices are shared across time. Classical RNNs can be difficult to train over long horizons due to vanishing or exploding gradients; gated architectures such as *long short-term memory* (LSTM) networks were introduced to mitigate these issues by controlling information flow through multiplicative gates [HS97; BSF94]. It is worth noting that (8) also admits continuous-time formulations that are naturally studied as input-output systems. A classical reference analyzing the controllability of such systems is [SS97].

More recently, *transformers* [Vas+17] have become a dominant paradigm for *sequence modeling*. A transformer layer combines three ingredients: (i) an interaction operator (*self-attention*) that mixes tokens with data-dependent weights, (ii) a pointwise perceptron (MLP) applied independently to each token, and (iii) residual connections and *normalization layers* [BKH16]. At a conceptual level, the attention mechanism at layer ℓ updates the i-th token within a sequence of n tokens by

$$x_{\ell+1}^i = \sum_{j=1}^n \alpha_{ij}\, V(x_\ell^j), \qquad \text{where} \quad \alpha_{ij} = \frac{\exp\big(s_\theta(x_\ell^i, x_\ell^j)\big)}{\sum_{k=1}^n \exp\big(s_\theta(x_\ell^i, x_\ell^k)\big)}. \tag{9}$$

where s_θ is a learned *similarity score* (the simplest being a dot product after learned linear projections, known as *dot-product attention*) and V is a learned value map (often linear). *Multi-head attention* repeats this mechanism in several learned subspaces and recombines the outputs.

While the attention mechanism captures token interactions, MLPs (or alternatively, convolutional layers) applied independently to each token are fundamental to the model's expressive power. Indeed, the theoretical guarantees establishing the approximation capabilities of Transformers rely heavily on these components. Given the myriad of architectural variants—differing in aspects such as masked or unmasked attention, encoder or decoder structure, pre- or post-layer normalization, and the precise form of the feed-forward block [PH22]—we do not attempt to model all these features simultaneously in this thesis. For now, we abstract away layer normalization and focus on the core residual attention–MLP structure found, for instance, in the GPT-2 architecture [Ope24].

Adopting the continuous-depth perspective introduced earlier in (5), it is natural to interpret this simplified Transformer block dynamically. If a residual layer of step size Δt first applies self-attention and then a perceptron, the sequential update reads:

$$y_\ell^i = x_\ell^i + \Delta t\, \mathrm{Att}_{\theta_\ell}^i(x_\ell^1, \dots, x_\ell^n), \qquad x_{\ell+1}^i = y_\ell^i + \Delta t\, f_{\theta_\ell}(y_\ell^i),$$

where $\mathrm{Att}_{\theta_\ell}^i$ is given by (9) and f_{θ_ℓ} denotes the MLP. Since $y_\ell^i = x_\ell^i + O(\Delta t)$, a first-order expansion yields

$$x_{\ell+1}^i = x_\ell^i + \Delta t\Big(f_{\theta_\ell}(x_\ell^i) + \mathrm{Att}_{\theta_\ell}^i(x_\ell^1, \dots, x_\ell^n)\Big) + O\big((\Delta t)^2\big).$$

Thus, although self-attention and the MLP are composed sequentially at the discrete level, they contribute additively in the continuous-depth limit. This formally leads to an interacting-particle system of the form

$$\dot{x}_i(t) = f_{\theta(t)}\big(x_i(t)\big) + \sum_{j=1}^n \frac{\exp\big(s_{\theta(t)}(x_i(t), x_j(t))\big)}{\sum_{k=1}^n \exp\big(s_{\theta(t)}(x_i(t), x_k(t))\big)}\, V(t) x_j(t), \qquad i \in [n]. \tag{10}$$

This viewpoint is consistent with the dynamical interpretation of Transformer layers as numerical discretizations of multi-particle systems [Lu+20; PH22].

With these perspectives in place, we now turn to a summary of the thesis contributions and an outline of the manuscript.

Perspectiva General

> *«Michael Jordan, [...] una de las grandes figuras del aprendizaje automático moderno, defiende [...] dar prioridad a la conjunción de este enfoque de "aumento de la inteligencia" con el desarrollo de una "infraestructura inteligente", derivada de la constante y omnipresente digitalización de cualquier actividad humana, en detrimento de un énfasis exclusivo en el desarrollo de la IA fuerte [...]. O, si se quiere, que debemos preocuparnos menos por "sueños de ciencia ficción o pesadillas super humanas" y más en comprender mejor estas tecnologías y guiar su desarrollo según se hacen cada vez más presentes, influyentes y disruptivas».*
>
> — José R. Dorronsoro
> Lección inaugural de la Universidad Autónoma de Madrid
> Curso académico 2006–2007 [Dor22]

El paradigma del aprendizaje

En las últimas décadas hemos presenciado una revolución silenciosa, acelerada notablemente en la década de 2010, que ha transformado la sociedad por completo, impulsada por la integración omnipresente de la tecnología en todas las facetas de la vida moderna, desde la ciencia y la educación hasta la salud. Bajo la etiqueta popular de *inteligencia artificial* (IA), una parte sustancial de este cambio está impulsada por el *aprendizaje automático* (*machine learning*, ML): la disciplina que estudia algoritmos capaces de inferir patrones a partir de datos para realizar predicciones o tomar decisiones, sin haber sido programados explícitamente para una tarea concreta [Mit97].

Acuñado originalmente por Arthur Samuel en 1959 ([Sam59]), el concepto de *machine learning* ha evolucionado hasta convertirse en una disciplina altamente formalizada que, desde una perspectiva matemática moderna, se concibe mejor como un marco para la aproximación e inferencia en alta dimensión. Su premisa central es utilizar datos generados por un proceso que carece de una formulación matemática explícita o tratable —típicamente modelado como una función o una ley de probabilidad subyacente— para construir un modelo parametrizado capaz de reproducir los comportamientos observados y generalizarlos a escenarios no vistos.

Este enfoque marca una divergencia conceptual frente al análisis numérico. Mientras que los métodos clásicos típicamente buscan aproximar operadores continuos conocidos y derivar cotas de error a partir de primeros principios, el aprendizaje automático debe inferir la estructura basándose exclusivamente en muestras finitas (y potencialmente ruidosas). Esto genera una paradoja: poseemos *recetas* que funcionan extraordinariamente bien, y sin embargo su impulso es en gran medida empírico. Por ello, el gran desafío abierto en la actualidad consiste en construir una base matemática rigurosa capaz de explicar por qué funcionan estos métodos, cuándo fallan y cómo deberían mejorarse.

A pesar de esta brecha teórica, su éxito es innegable. El ML ha redefinido el estado del arte en un espectro muy amplio de aplicaciones. En visión por computador y procesamiento del lenguaje natural, actualmente los modelos profundos impulsan desde el diagnóstico médico hasta los modelos de lenguaje a gran escala. Los modelos generativos han pasado de ser prototipos de investigación

a herramientas ampliamente desplegadas, capaces de sintetizar texto, imágenes, audio y código. Más allá de la tecnología de consumo, el ML forma parte esencial de flujos de trabajo científicos e industriales, abarcando desde la inferencia basada en simulación hasta el modelado sustituto y el soporte a la toma de decisiones.

Para el matemático, esta proliferación de aplicaciones plantea un desafío particular: la tarea no consiste solo en resolver un problema bien planteado, sino en abstraer fenómenos empíricos en estructuras rigurosas. En un panorama donde la práctica viene dictada a menudo por heurísticas y leyes de escalado, el análisis busca distinguir los mecanismos realmente esenciales de los meros detalles de implementación.

Ofrecer una revisión histórica exhaustiva del aprendizaje automático queda fuera del alcance de esta tesis. La trayectoria del campo es compleja, marcada por una inmensa multitud de contribuyentes, avances fundamentales y periodos de estancamiento conocidos como «inviernos de la IA». En lugar de un recuento histórico, nuestro objetivo es presentar una introducción con motivación matemática a varios de los modelos más prominentes. Para una visión histórica más detallada, remitimos al lector a [Fra20].

Resulta conveniente distinguir algunos paradigmas canónicos (para una introducción exhaustiva, remitimos a obras de referencia como [Bis06; HTF09; Bac24]). En *aprendizaje supervisado* se observan pares de entrada y etiqueta (x_i, y_i)—donde la entrada x_i representa las características observadas (p. ej., una imagen o un estado físico) y la etiqueta y_i la salida o categoría deseada—y se busca un predictor que generalice más allá de los datos; este es el marco de la clasificación y la regresión. En *aprendizaje no supervisado* se observan únicamente las entradas (x_i) y se pretende extraer estructura: representaciones de baja dimensión, clústeres o un modelo probabilístico subyacente. Una tarea no supervisada particularmente importante es el *modelado generativo*, donde se aprende (una aproximación de) la distribución de los datos con el fin de generar muestras sintéticas realistas [GBC16]. Por último, en *aprendizaje por refuerzo*, los datos se recogen de forma adaptativa mediante interacciones con un entorno, y el aprendizaje está guiado por señales de recompensa [SB18].

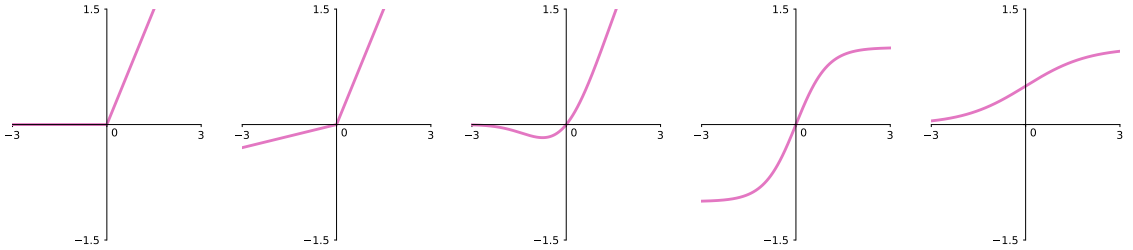
De las neuronas a las redes

Aunque estos paradigmas difieren en sus objetivos y en los datos disponibles, hoy convergen en un marco de modelado común. En el núcleo de su éxito empírico se encuentra el *aprendizaje profundo* (*deep learning*) —un término cuya aparición en la literatura se remonta, al menos, a [Dec86]—, una clase de métodos que se apoya en modelos de gran capacidad entrenados mediante optimización basada en gradientes [LBH15; BGC17].

Estos modelos se construyen sobre *redes neuronales* artificiales. Si bien las arquitecturas modernas se entienden mejor como funciones matemáticas flexibles y jerárquicas compuestas por operaciones no lineales simples, en lugar de modelos biológicos literales, su inspiración inicial proviene de la interacción de las neuronas en el sistema nervioso humano, tal y como evoca su nombre.

Históricamente, los cimientos conceptuales de estas unidades fueron establecidos por McCulloch y Pitts [MP43]. Ellos abstrajeron la acción local de las neuronas biológicas —propagar o inhibir impulsos— en simples puertas lógicas de *umbral*, demostrando que las redes formadas por estas unidades podían computar funciones booleanas. Aunque este trabajo estableció la plantilla para la computación neuronal, su modelo dependía de conexiones fijas diseñadas manualmente; carecía de cualquier mecanismo de adaptación o aprendizaje.

El salto crítico de los circuitos lógicos estáticos a los modelos adaptativos fue logrado por Rosenblatt a finales de la década de 1950 [Ros58]. Partiendo de esta base conceptual, introdujo el *perceptrón*.



(a) ReLU $s \mapsto s_+$ (b) Leaky ReLU $s \mapsto s \vee (0.1 \cdot s)$ (c) GeLU $s \mapsto s\,\Phi(s)$ (d) Tanh $s \mapsto \frac{e^s - e^{-s}}{e^s + e^{-s}}$ (e) Sigmoide $s \mapsto \frac{1}{1+e^{-s}}$

Figure 3: Representación de algunas funciones de activación $\sigma : \mathbb{R} \to \mathbb{R}$. En GeLU, $\Phi(s) = (1 + \operatorname{erf}(s/\sqrt{2}))/2$ es la función de distribución acumulada de la Gaussiana estándar.

Este representó el verdadero *germen* del aprendizaje automático: un modelo paramétrico emparejado con una regla matemática explícita para ajustar automáticamente sus pesos, materializando la idea central de que un mecanismo de decisión podía *ajustarse a partir de datos* en lugar de ser programado explícitamente.

En su forma más simple, un perceptrón es una función parametrizada dada por

$$f_\theta(x) = w\,\sigma(a \cdot x + b), \tag{1'}$$

donde $x \in \mathbb{R}^d$ es el estado, $\theta \equiv (w, a, b) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ es el parámetro, y $\sigma : \mathbb{R} \to \mathbb{R}$ es una no linealidad fija denominada *función de activación*. Históricamente, se solían emplear funciones suaves y acotadas (p. ej., la tanh o la sigmoide). Sin embargo, en el aprendizaje profundo moderno, el estándar predominante es la *unidad lineal rectificada* (ReLU). También es habitual la *unidad lineal de error gaussiano* (GeLU), en particular en arquitecturas *transformer* [HG16]. Las expresiones analíticas de estas funciones, junto con otras alternativas populares, se detallan en la Figura 3.

En teoría de la aproximación, las funciones de la forma (1′) también se conocen como *funciones de cresta* [Pin15]. La terminología estándar en ML las interpreta como *neuronas*: $a \cdot x + b$ es una preactivación afín, σ introduce la no linealidad y w modula la contribución de salida.

Una sola neurona se limita a producir traslaciones y reescalados de las funciones de activación mostradas en la Figura 3, por lo que posee un poder expresivo muy limitado. El siguiente paso natural es superponer p de estas unidades, lo que conduce a la familia de redes *poco profundas* (de una sola capa oculta) de *anchura* $p \in \mathbb{N}$ prefijada:

$$\mathcal{F}_p := \left\{ x \mapsto f_\theta(x) = \sum_{i=1}^{p} w_i\,\sigma(a_i \cdot x + b_i), \qquad \theta \equiv (w_i, a_i, b_i)_{i \in [p]} \in (\mathbb{R} \times \mathbb{R}^d \times \mathbb{R})^p \right\}. \tag{2'}$$

Desde el punto de vista matemático, $\mathcal{F}_p$ es una clase de modelos *no lineales*. Esto contrasta con los esquemas clásicos de aproximación *lineal*—como las series de Fourier, los splines o la interpolación polinómica—donde el diccionario de funciones base se fija *a priori* y solo se aprenden los coeficientes. En una red poco profunda, el diccionario es *adaptativo*: los parámetros internos a_i y b_i permiten desplazar y remodelar las funciones base $x \mapsto \sigma(a_i \cdot x + b_i)$, adaptando de hecho la base a la función objetivo.

A pesar de su flexibilidad, la arquitectura sigue siendo estructuralmente elemental. Esto convirtió a las redes poco profundas en el candidato principal para los primeros análisis rigurosos, particularmente en lo relativo a resultados de densidad (referidos en la literatura de ML como *teoremas de*

aproximación universal) Los resultados fundacionales establecieron que, a medida que aumenta la anchura p, estas redes pueden aproximar funciones objetivo con precisión arbitraria.

Desde la perspectiva de la teoría de la aproximación, la cuestión fundamental de cuándo las combinaciones lineales de traslaciones y reescalados de una única función base forman un conjunto denso se remonta, al menos, hasta 1932 con los teoremas tauberianos de Wiener en [Wie32], centrados en los espacios de Lebesgue (L^1 y L^2). Si bien Wiener trazó la hoja de ruta matemática para este tipo de universalidad, su enfoque basado en el análisis de Fourier exige inherentemente que la función base sea integrable. Esta condición es violada de forma natural por las funciones de activación estándar empleadas en las arquitecturas modernas, como las ilustradas en la Figura 3, que no decaen en el infinito.

Abordando estas lagunas específicas para las redes neuronales —a saber, el uso de activaciones no integrables y el requisito de convergencia uniforme en conjuntos compactos— Cybenko en [Cyb89] fue el primero en demostrar explícitamente la densidad de $\mathcal{F} \coloneqq \bigcup_{p\geq 1} \mathcal{F}_p$ para activaciones sigmoides. Hornik extendió posteriormente el resultado a clases más amplias de activaciones y a topologías más fuertes en [Hor93]. Una caracterización definitiva fue proporcionada más tarde por Leshno et al. en [Les+93], quienes establecieron una condición necesaria y suficiente: la densidad se cumple si y solo si la activación σ no es un polinomio. Para un tratamiento exhaustivo de estas aproximaciones mediante funciones de cresta, la monografía de Pinkus ([Pin99]) sigue siendo la referencia estándar.

El poder de la profundidad

La composición de múltiples capas da lugar a las redes *feedforward profundas*. Formalmente, una red de profundidad L puede escribirse como:

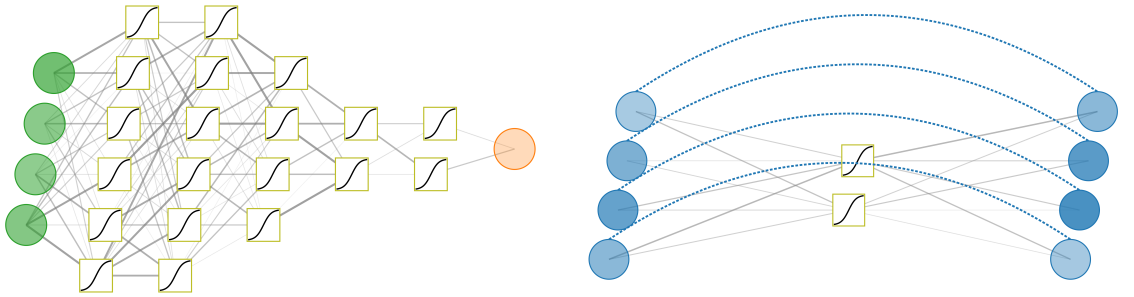
$$f_\theta(x) = T_L \circ \sigma \circ T_{L-1} \circ \sigma \circ \cdots \circ \sigma \circ T_1(x), \tag{3'}$$

donde $x \in \mathbb{R}^d$, cada $T_\ell(y) = A_\ell y + b_\ell$ es una aplicación afín entre espacios euclídeos de dimensión arbitraria, y σ actúa componente a componente. Cuando las transformaciones T_ℓ se implementan mediante capas lineales totalmente conectadas (es decir, A_ℓ es una matriz sin restricciones estructurales), estas redes se denominan comúnmente *perceptrones multicapa* (MLP). En la Figura 4a se muestra una representación gráfica de esta arquitectura.

El paso de la suma (anchura) a la composición (profundidad) permite un aumento drástico en la expresividad de la clase de funciones. Una analogía esclarecedora proviene del álgebra de polinomios. Al sumar monomios de un grado fijo d, estamos restringidos al subespacio de polinomios de ese grado. En cambio, la composición de dos polinomios tiene un efecto multiplicativo sobre el grado. En consecuencia, el tamaño de la clase crece de forma exponencial con la profundidad, mientras que con la anchura el crecimiento es puramente lineal.

Esta analogía sugiere que las arquitecturas profundas son más eficientes para representar funciones con estructura composicional. En el contexto de las redes neuronales, esto se traduce en una reutilización jerárquica de las representaciones intermedias, un mecanismo central en el éxito del aprendizaje profundo moderno.

No obstante, la capacidad de *representar* funciones complejas es solo una parte de la ecuación; también es necesario poder *entrenar* el modelo eficientemente. Durante décadas, el potencial de la profundidad estuvo limitado por la dificultad inherente a la optimización no convexa en alta dimensión. Típicamente, el entrenamiento se formula como la minimización de una función de pérdida mediante variantes del *descenso por gradiente*.



(a) Representación gráfica de un MLP con una capa de entrada en $\mathbb{R}^4$ (cada círculo es una coordenada), seguida de $L = 5$ capas ocultas de anchura variable con $\sigma = \tanh$, y una salida escalar. Los pesos se muestran como aristas cuya opacidad es proporcional a la magnitud.

(b) Un bloque de ResNet. Las neuronas de estado evolucionan a través de una capa oculta de anchura $p = 2$. Las conexiones de identidad se representan como arcos curvos.

Figure 4: Comparación de arquitecturas de redes neuronales.

La viabilidad práctica de este enfoque depende del cálculo eficiente de los gradientes. El avance que permitió escalar la optimización a redes multicapa fue la popularización del algoritmo de *retropropagación* [RHW86], basado en la aplicación recursiva de la regla de la cadena. Sin embargo, este éxito no habría sido posible sin la llegada de recursos de computación paralela y a gran escala, que finalmente permitieron optimizar estos vastos conjuntos de parámetros de manera masiva. Es esta convergencia entre el refinamiento algorítmico y la potencia de cálculo la que define el paradigma actual.

Redes residuales y el punto de vista de profundidad continua

Incluso con el uso de retropropagación, surgen inestabilidades severas al aumentar la profundidad de la red. Los obstáculos más conocidos son los problemas de *desvanecimiento* y *explosión* del gradiente [BSF94]. Dado que los gradientes se calculan mediante la aplicación recursiva de la regla de la cadena, el proceso implica un producto secuencial de matrices jacobianas. A menos que el espectro de estos operadores se mantenga próximo a la unidad, la norma del gradiente tiende a decaer o crecer exponencialmente con la profundidad, resultando en un problema de optimización mal condicionado.

Un punto de inflexión fue la introducción de las *redes residuales* (ResNets) en el artículo seminal [He+16]. Las ResNets modifican la actualización entre capas añadiendo conexiones de salto (véase la Figura 4b). Si denotamos por x_ℓ y θ_ℓ el estado y los parámetros en la capa ℓ, una red residual con L capas y entrada $x_0 \in \mathbb{R}^d$ toma la forma

$$x_{\ell+1} = x_\ell + f_{\theta_\ell}(x_\ell), \qquad (\ell = 0, \dots, L-1), \tag{4'}$$

donde f_{θ_ℓ} es una aplicación no lineal (típicamente, una red pequeña de pocas capas). El término identidad proporciona un camino directo para la propagación del gradiente y estabiliza empíricamente el entrenamiento, lo que permite utilizar redes con cientos de capas.

Desde una perspectiva matemática, la ecuación (4′) se corresponde con una discretización de Euler explícita de una ecuación diferencial ordinaria (EDO), salvo por una elección de escala temporal (p. ej., fijando convencionalmente el tamaño de paso en $\Delta t = 1$). Esta observación motiva los modelos

de *profundidad continua*: si interpretamos la profundidad como una variable temporal $t > 0$ y permitimos que los parámetros varíen de forma continua, obtenemos las *EDOs neuronales* [E17; HR18; Che+18]. Una EDO neuronal evoluciona una entrada inicial $x_0 \in \mathbb{R}^d$ por el sistema

$$\begin{aligned} \dot{x}(t) &= v_{\theta(t)}\big(x(t)\big), \qquad (t \geqslant 0), \\ x(0) &= x_0, \end{aligned} \tag{5'}$$

donde $\theta(\cdot)$ es un parámetro dependiente del tiempo y v_θ es un campo vectorial, típicamente parametrizado mediante una red neuronal[0]. Hemos cambiado deliberadamente la notación de f_θ en (4′) a v_θ en (5′) para enfatizar la perspectiva dinámica: el modelo especifica un *campo vectorial* cuya solución define un mapa de evolución o *flujo* $\Phi_t^\theta(x_0) \coloneqq x(t)$. Operativamente, este flujo actúa como un continuo de capas neuronales *infinitesimales*.

El marco continuo de (5′) suele resultar más tratable analíticamente, y sitúa al aprendizaje profundo en la órbita de la *teoría de control* [FZ03]. Clásicamente, el control estudia *sistemas dinámicos* gobernados por una *entrada* externa,

$$\dot{x}(t) = v(x(t), u(t)), \qquad (t \geqslant 0), \tag{6'}$$

y aborda cuestiones como si es posible guiar el estado hacia un objetivo deseado (*alcanzabilidad*/*controlabilidad*, véase [Son98]) y cómo hacerlo optimizando un criterio de rendimiento (*control óptimo*, véase [Tré05]). Referencias estándar incluyen [Pon+62; Kha02; AS04; Ber17].

En esta tesis nos centraremos particularmente en el primer concepto, que está fundamentalmente ligado a la viabilidad del problema de aprendizaje en *machine learning*. En términos generales, decimos que el sistema (6′) es *controlable* si, para cualquier estado inicial $x_0 \in \mathbb{R}^d$ y cualquier estado objetivo $x_1 \in \mathbb{R}^d$, existe una señal de control admisible $u(\cdot)$ y un tiempo finito $T > 0$ tales que la trayectoria correspondiente satisface $x(0) = x_0$ y $x(T) = x_1$. Una referencia seminal que establece los fundamentos algebraicos de esta propiedad para sistemas lineales es [Kal60]. En el marco no lineal, el enfoque geométrico tiene sus raíces clásicas en trabajos como [Cho40; JS72]; una referencia moderna es [Cor07].

En el contexto de las EDOs neuronales, el parámetro dependiente del tiempo $\theta(\cdot)$ actúa como la entrada de control $u(\cdot)$. Desde esta óptica, el entrenamiento se convierte en un problema de control óptimo: el objetivo es seleccionar una trayectoria de control $\theta(\cdot)$ tal que el flujo inducido minimice una función de coste dependiente del mapa final [E17; Zua22].

Esta perspectiva ha demostrado ser útil tanto conceptual como técnicamente: conecta cuestiones de expresividad con la noción de controlabilidad [AS22; RZ23; Che+25c], y fundamenta condiciones de optimalidad y algoritmos de entrenamiento basados en el Principio del Máximo de Pontryagin (incluyendo formulaciones de campo medio que capturan la naturaleza probabilística del aprendizaje) [EHL18; Li+18; Ben+19; Bon+23].

En términos computacionales, esta visión reinterpreta la retropropagación como la integración regresiva de una ecuación *adjunta*. Popularizado en [Che+18], este método permite, en principio, entrenar con un coste de memoria constante (independiente de la profundidad de la discretización) mediante la reconstrucción reversible de las trayectorias, aunque a menudo a costa de la estabilidad numérica. Una discusión teórica detallada sobre estas cuestiones puede encontrarse en [Mas+20].

Por último, cuando la entrada x es una variable aleatoria, la ecuación (5′) transporta su distribución de probabilidad a lo largo del tiempo. Esta observación conduce de manera natural a los *flujos de normalización continuos*, introducidos junto con las EDOs neuronales en [Che+18] y desarrollados

[0] El campo v_θ también puede incorporar términos físicos conocidos (modelos híbridos) o imponer estructuras matemáticas específicas. No obstante, nos centraremos en el caso canónico de una red neuronal estándar.

posteriormente en variantes escalables como [Gra+19; Onk+20], donde el objetivo es aprender dinámicas que transformen una distribución base simple en la distribución de los datos.

Más allá de los perceptrones

El término *red neuronal* no designa una arquitectura única. Numerosas familias de modelos exitosos surgen al sustituir las aplicaciones afines de (3′) por operadores estructurados, diseñados específicamente para cada modalidad de datos e incorporando sesgos inductivos adecuados.

Las *redes neuronales convolucionales* (CNNs), por ejemplo, reemplazan la conectividad total por operaciones de convolución con pesos compartidos, dando lugar a capas de la forma

$$x_{\ell+1} = \sigma\big(A_\ell * x_\ell + b_\ell\big), \qquad (\ell = 0, \dots, L-1), \tag{7′}$$

donde $*$ denota una convolución discreta (sobre índices espaciales y/o temporales). Al explotar la localidad y la equivarianza bajo traslaciones, las CNNs han sido centrales en visión por computador desde sus primeras implementaciones [Lec+98], convirtiéndose en el catalizador del auge del aprendizaje profundo tras sus éxitos decisivos en el reconocimiento de imágenes a gran escala [KSH12].

El tratamiento de datos secuenciales (texto, audio, series temporales) motivó el desarrollo de modelos recurrentes, arquitecturas explícitamente dinámicas donde una misma transformación se itera a lo largo del tiempo. Una *red neuronal recurrente* (RNN) básica se describe mediante

$$\begin{aligned} h_{\ell+1} &= \sigma\big(A_h h_\ell + A_x x_\ell + b\big), \\ y_\ell &= A_y h_\ell, \end{aligned} \qquad (\ell = 0, \dots, L-1), \tag{8′}$$

donde h_ℓ representa el estado oculto, x_ℓ la entrada en el paso ℓ, y las matrices de pesos permanecen constantes en el tiempo. El entrenamiento de RNNs clásicas en secuencias largas se ve dificultado por los problemas de gradiente evanescente o explosivo. Para mitigarlos, se introdujeron arquitecturas basadas en compuertas, como las *long short-term memory* (LSTM), que regulan el flujo de información mediante mecanismos multiplicativos [HS97; BSF94]. Cabe destacar que (8′) también admite formulaciones en tiempo continuo que se estudian de forma natural como sistemas de entrada-salida. Una referencia clásica donde se ha analizado la controlabilidad de dichos sistemas es [SS97].

Más recientemente, los *transformers* [Vas+17] se han erigido como el paradigma dominante para el modelado de secuencias. Un bloque *transformer* típico combina tres ingredientes: (i) un operador de autoatención que mezcla los distintos tokens mediante pesos dependientes de los datos, (ii) un perceptrón multicapa aplicado punto a punto de forma independiente a cada token, y (iii) conexiones residuales junto con capas de normalización (*LayerNorm*) [BKH16]. A nivel conceptual, el mecanismo de atención en la capa ℓ actualiza el i-ésimo *token* dentro de una secuencia de longitud n mediante la ecuación:

$$x^i_{\ell+1} = \sum_{j=1}^{n} \alpha_{ij}\, V(x^j_\ell), \qquad \text{donde} \quad \alpha_{ij} = \frac{\exp\big(s_\theta(x^i_\ell, x^j_\ell)\big)}{\sum_{k=1}^{n} \exp\big(s_\theta(x^i_\ell, x^k_\ell)\big)}, \tag{9′}$$

siendo s_θ una puntuación de similitud aprendida (típicamente un producto escalar tras aplicar proyecciones lineales) y V un mapa de valor aprendido (a menudo lineal). La atención multicabezal repite este mecanismo en varios subespacios aprendidos y recombina los resultados.

Si bien el mecanismo de autoatención captura las interacciones entre los elementos de la secuencia, las capas *feed-forward* aplicadas de forma independiente a cada *token* —o, alternativamente, las

capas convolucionales—son fundamentales para el poder expresivo del modelo. De hecho, las garantías teóricas que establecen la capacidad de aproximación universal de los *transformers* se apoyan de manera decisiva en estos componentes. Dada la multitud de variantes arquitectónicas existentes —que difieren en aspectos como el uso de atención enmascarada o sin enmascarar, la estructura de codificador o decodificador, la normalización previa o posterior a la capa, y la forma exacta del bloque *feed-forward* [PH22]—, en esta tesis no intentaremos modelar todas estas características simultáneamente. Por el contrario, abstraemos el efecto de la normalización de capa y nos centramos en la estructura central residual de atención y MLP presente, por ejemplo, en la arquitectura GPT-2 [Ope24].

Adoptando la perspectiva de profundidad continua introducida anteriormente en la ecuación (5′), resulta natural interpretar este bloque *transformer* simplificado desde un punto de vista dinámico. Si una capa residual con un paso de tamaño Δt aplica primero la autoatención y luego un MLP punto a punto, la actualización secuencial toma la forma:

$$y_\ell^i = x_\ell^i + \Delta t\, \mathrm{Att}_{\theta_\ell}^i(x_\ell^1, \dots, x_\ell^n), \qquad x_{\ell+1}^i = y_\ell^i + \Delta t\, f_{\theta_\ell}(y_\ell^i),$$

donde $\mathrm{Att}_{\theta_\ell}^i$ viene dada por la ecuación (9′) y f_{θ_ℓ} denota el MLP. Puesto que $y_\ell^i = x_\ell^i + O(\Delta t)$, una expansión de Taylor de primer orden nos da:

$$x_{\ell+1}^i = x_\ell^i + \Delta t\Big(f_{\theta_\ell}(x_\ell^i) + \mathrm{Att}_{\theta_\ell}^i(x_\ell^1, \dots, x_\ell^n)\Big) + O\big((\Delta t)^2\big).$$

Por lo tanto, aunque la autoatención y el MLP se componen secuencialmente a nivel discreto, sus contribuciones se vuelven aditivas en el límite de profundidad continua. Formalmente, esto conduce a un sistema de partículas interactuantes de la forma:

$$\dot{x}_i(t) = f_{\theta(t)}\big(x_i(t)\big) + \sum_{j=1}^{n} \frac{\exp\big(s_{\theta(t)}(x_i(t), x_j(t))\big)}{\sum_{k=1}^{n} \exp\big(s_{\theta(t)}(x_i(t), x_k(t))\big)}\, V(t) x_j(t), \qquad i \in [n]. \tag{10′}$$

Esta perspectiva es plenamente consistente con la interpretación dinámica de las capas *transformer* como discretizaciones numéricas de sistemas multipartícula [Lu+20; PH22].

Establecido este contexto teórico, a continuación presentamos un resumen de las contribuciones principales de esta tesis y un esquema estructurado del manuscrito.

Contributions

This thesis develops a unified mathematical viewpoint on deep learning models built from *flows*: maps obtained by evolving inputs through parameterized dynamical systems. The common thread is to interpret learning as the selection of time-dependent parameters—viewed as *control inputs*—so that the induced dynamics match a data-driven objective. This viewpoint turns representational questions into controllability questions, and connects expressivity, stability, and statistical performance to quantitative properties of the underlying dynamics.

Two learning paradigms. We focus on two canonical frameworks.

(i) Supervised learning. One observes labeled pairs (x_i, y_i) sampled from an unknown data-generating mechanism. The goal is to construct a predictor that fits the observations while performing well on unseen inputs. In this thesis, predictors are induced by terminal-time flow maps: training amounts to selecting controls so that the resulting flow realizes a desired input–output behavior. This makes *expressivity* a question of controllability of finite ensembles, while *generalization* becomes a question of how the same dynamics behave beyond finitely many points.

(ii) Generative modeling. One observes unlabeled samples from an unknown target distribution and aims to learn a model distribution that approximates it. In flow-based generative modeling, this is achieved by transporting a simple base law through an invertible transformation, such as the flow of a controlled dynamical system. Training can be viewed as an optimal control problem on probability measures, driven by a suitable measure of discrepancy.

The manuscript is structured in three parts, moving from the control of finitely many particles, to the control of probability laws, and finally to applications of the dynamical viewpoint—including distributional time series modeling and mean-field analyses of modern architectures.

Before these parts, we collect in **Chapter 1** the analytical tools used throughout the thesis, including controlled flows and their regularity (**Section 1.1**), measure transport and the continuity equation (**Section 1.2**), and a short neural-ODE toolbox for constructive proofs (**Section 1.3**).

Part I. Supervised Learning

Part I studies supervised learning through the controllability of controlled flows acting on point clouds. The central questions guiding this analysis are: *what can continuous-depth models fit, at what complexity cost, and how can controllability be leveraged to obtain meaningful population guarantees?* A key feature of our approach is that it is *constructive*: rather than relying solely on existence theorems, we explicitly synthesize controls and describe the geometric evolution they induce in the system.

Chapter 2: Data Classification

We reformulate *binary classification* with neural ODEs as a relaxed form of simultaneous controllability for N training points. Instead of prescribing an exact terminal target for each input, the

requirement is set-valued: points with different labels should be mapped into separated regions so that a fixed readout can discriminate them. Working in a minimal setting of a single-neuron flow, we measure *model complexity* by the number of *control switches*.

Our main contribution is a constructive *cluster-by-cluster* steering strategy that separates groups of size d (the ambient dimension) sequentially. This reduces the worst-case switch complexity from $O(N)$ to $O(N/d)$, showing that high dimensionality actively reduces the required temporal depth. To move beyond worst-case analysis, we also study the distribution of the number of switches; in particular, successful classification with constant (zero-switch) controls becomes increasingly likely as d grows. We complement the analysis with numerical experiments that validate theory.

Source. The results presented in **Chapter 2** are drawn from [ÁOZ25].

> A. Álvarez-López, R. Orive-Illera, and E. Zuazua. Cluster-Based Classification with Neural ODEs via Control. In: *Journal of Machine Learning* 4.2 (2025), pp. 128–156. DOI: 10.4208/jml.241114

Code. **Chapter 2** includes code from [Álv24].

> **Repository.**
> A. Álvarez-López. 2024. URL: https://github.com/antonioalvarezl/2024-WCS-NODEs

Chapter 3: Data Interpolation

We then turn to strong expressivity for *regression*: the *exact interpolation* of data points by a single flow map. In the control viewpoint used throughout the thesis, this becomes a *simultaneous controllability* problem where a single control must steer N prescribed initial states to N targets at the same terminal time. We provide explicit constructions for several neural ODE architectures, tracking their complexity in terms of *width* p (number of neurons) and *depth* (number of switches).

Our results yield an explicit width–depth tradeoff: increasing the width reduces the required number of switches, with bounds of the form $1+O(N/p)$. In overparameterized or high-dimensional regimes, we show that these constructions can be pushed to the autonomous limit, where interpolation becomes achievable even with constant controls. This phenomenon is shown to be robust across semi-autonomous, linear-in-control, and two-layer autonomous vector field variants. Finally, we relax the exact matching constraint to obtain approximation rates and probabilistic guarantees for approximate controllability.

Source. **Chapter 3** synthesizes results from three different works: [ÁSZ24], [ÁGR25], and [ÁLZ26a]. The exact distribution of content among these works is detailed within the chapter, just before presenting the results.

> A. Álvarez-López, A. H. Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural ODEs. In: *Neural Networks* 180 (2024), p. 106640. DOI: 10.1016/j.neunet.2024.106640

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. In: *Applied Mathematics & Optimization* 92.2 (2025), p. 33. DOI: 10.1007/s00245-025-10299-7

A. Álvarez-López, L. Liverani, and E. Zuazua. Constructive interpolation and generalization rates for neural ODEs: a control perspective. 2026. arXiv: 2606.00469

Chapter 4: Generalization

Exact interpolation alone does not imply statistical performance: a model can fit training data perfectly while performing poorly off-sample. In this chapter, we propose a route to generalization tailored to the controlled flow viewpoint. Rather than bounding the generalization gap uniformly over large parameter sets, we construct *data-dependent controls* designed to faithfully approximate the target function across its entire domain, thereby ensuring accurate predictions on unseen data.

The key idea is to replace pointwise matching by *cell-wise controllability*: we design flows that act coherently on the cells of a partition, thereby realizing structured nonparametric estimators. We identify a fundamental topological obstruction in this process: routing macroscopic cells without trajectory collisions makes purely autonomous models structurally insufficient, requiring time dependence.

This approach yields explicit bounds on the population risk for the resulting predictor, together with quantitative requirements on the control complexity needed to attain a prescribed accuracy. The resulting statements make the interplay between sample size, geometric resolution, and model complexity explicit, successfully identifying regimes in which the risk provably tends to zero.

Source. The results presented in **Chapter 4** are drawn from [ÁLZ26a].

A. Álvarez-López, L. Liverani, and E. Zuazua. Constructive interpolation and generalization rates for neural ODEs: a control perspective. 2026. arXiv: 2606.00469

Part II. Generative Modeling

Part II moves from fitting finite point clouds to controlling and approximating probability laws. The central representation is the *pushforward map* $(\Phi_T)_{\#}\mu$ induced by an *attainable flow*, together with its PDE formulation through the continuity equation. From a generative-modeling viewpoint, learning amounts to selecting a flow that transports a simple base distribution toward a target data distribution under a chosen discrepancy. The central questions guiding this part are: *which target measures can be approximated via controlled flows, under which discrepancies—with a focus on the Wasserstein and the total variation distances—, and with what quantitative complexity cost?*

Chapter 5: Controllability for Measure Transport

This chapter develops a general controllability theory for *flow-based generative models* at the level of probability measures. A central theme is that the relevant approximation mechanism depends strongly on the discrepancy used to compare measures.

For weak discrepancies such as the Wasserstein distance, we show that measure approximation is closely tied to the interpolation of finite point clouds up to relabeling. At the empirical level, this

relation is exact, while for general measures one needs suitable control on the geometric distortion of the interpolating flows. This leads to general lifting principles from particle interpolation to distributional approximation, and in particular to broad universality results for affine-invariant flow families in dimension $d \geq 2$.

For stronger discrepancies such as total variation, interpolation of point clouds is no longer sufficient. Restricting to absolutely continuous measures, we prove universality through a strategy inspired by Knothe–Rosenblatt-type constructions, which isolates the key geometric ingredients needed in higher dimension: reversibility, one-dimensional rearrangements, and shear-type coordinate mixing.

Finally, we provide explicit counterexamples showing that measure-level universality and interpolation of finite datasets are genuinely distinct notions: in general, the relation between them depends sharply on the discrepancy under consideration.

Source. The results presented in **Chapter 5** are drawn from [Álv+26] (to be submitted). This research was conducted during an FPU research stay at the National University of Singapore, in collaboration with the group of Professor Qianxiao Li.

> **In preparation.**
> A. Álvarez-López et al. Flow-based approximation and controllability of measures. 2026

Chapter 6: Continuous Normalizing Flows

This chapter focuses on *continuous normalizing flows* (CNFs), where the training objective is typically maximum likelihood. At the population level, this corresponds to minimizing the Kullback–Leibler divergence (relative entropy) between the target distribution and the transported base distribution. Relative entropy induces a substantially stronger notion of approximation than Wasserstein or total variation, and this shift requires additional structural control on the transported density. Our main contribution is an approximate controllability result in relative entropy for a CNF with piecewise-constant controls. Under natural assumptions on the relative tail decay of the target and the base, we explicitly construct controls that drive the transported density arbitrarily close to the target in relative entropy, and we track the induced complexity in terms of the number of switches.

Source. The results presented in **Chapter 6** are drawn from [ÁGR25].

> A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. In: *Applied Mathematics & Optimization* 92.2 (2025), p. 33. DOI: 10.1007/s00245-025-10299-7

Part III. Applications

Part III brings together two applications of the dynamical viewpoint: one data-driven and the other motivated by the theoretical understanding of a more complex architecture.

Chapter 7: Distributional Dynamics Modeling for Digital Health

We study a practical setting motivated by *continuous glucose monitoring* where the data are snapshot samples from a time-evolving distribution. We model the dynamics in continuous time, combining Gaussian mixture representations with shared parameters and a neural ODE that defines the time-evolving mixture weights on the simplex to create an interpretable estimator. The weights

are first fitted at observation times by minimizing the Gaussian-kernel Maximum Mean Discrepancy, exploiting closed forms for Gaussian mixtures, and are then lifted to a continuous-time trajectory by learning a neural ODE. This yields a continuous-time, interpretable representation of patient dynamics. We provide theoretical guarantees for this approach and validate its performance against competing models in low and moderate dimensions, demonstrating its practical value in longitudinal diabetes trials. Conceptually, this chapter frames the method as an application-oriented *lift* of the flow-based framework in **Part I** to distributional time series, aligned with the learning objective of **Part II**.

Source. The results presented in **Chapter 7** are drawn from [ÁM26].

> A. Álvarez-López and M. Matabuena. Continuous-Time Learning of Probability Distributions: A Case Study in a Digital Trial of Young Children with Type 1 Diabetes. 2026. eprint: 2603.24427

Code. **Chapter 7** includes code from [Álv25].

> **Repository.**
> A. Álvarez-López. 2025. URL: https://github.com/antonioalvarezl/2025-EM_MMD

Chapter 8: Mean-field Landscape of Transformers

We study a mean-field model for Transformer representations in which tokens are viewed as an *interacting particle system* evolving on a bounded latent geometry. In this framework, self-attention produces a state-dependent interaction, while the MLP block acts as an external drift that reshapes the energy landscape. Our results show that adding the perceptron qualitatively changes the stationary behavior of the dynamics, forcing stationary solutions toward singular (generically atomic) configurations under broad conditions, and we provide quantitative bounds describing the resulting discrete structure. We complement the analysis with numerical experiments illustrating the emergence of clusters and the scaling of support complexity. The chapter concludes with brief remarks on the *Mixture-of-Experts* architecture, highlighting how it can be interpreted dynamically and the analytical questions it raises.

Source. The results presented in **Chapter 8** are drawn from [ÁGR26].

> A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Perceptrons and localization of attention's mean-field landscape. Accepted as a Spotlight paper at the International Conference on Machine Learning (ICML) 2026. 2026. arXiv: 2601.21366

Code. **Chapter 8** includes code from [Álv26].

> **Repository.**
> A. Álvarez-López. 2026. URL: https://github.com/antonioalvarezl/2026-MLP-Attention-Energy

Contribuciones

Esta tesis desarrolla una perspectiva matemática unificada sobre modelos de aprendizaje profundo basados en *flujos*: aplicaciones construidas mediante la evolución de las entradas a través de sistemas dinámicos parametrizados. El hilo conductor consiste en interpretar el proceso de aprendizaje como la selección de parámetros dependientes del tiempo—vistos como *controles*—de manera que la dinámica inducida satisfaga un objetivo determinado por los datos. Esta perspectiva transforma problemas de representación en problemas de controlabilidad, y conecta conceptos como expresividad, estabilidad y rendimiento estadístico con propiedades cuantitativas de la dinámica subyacente.

Dos paradigmas de aprendizaje. Nos centramos en dos escenarios canónicos:

(i) Aprendizaje supervisado. Se observan pares etiquetados (x_i, y_i) muestreados de un proceso generador desconocido. El objetivo es construir un predictor que ajuste las observaciones y, simultáneamente, produzca predicciones precisas en entradas no vistas. En esta tesis, los predictores se definen como el mapa del flujo evaluado en el tiempo final. Así, entrenar equivale a seleccionar controles para que el flujo resultante reproduzca el comportamiento entrada–salida deseado. Esto convierte la *expresividad* en una cuestión de controlabilidad de conjuntos finitos, mientras que la *generalización* se relaciona con el comportamiento de dicha dinámica más allá de los puntos de entrenamiento.

(ii) Modelos generativos. Se observan muestras no etiquetadas de una distribución objetivo desconocida y se busca aprender una distribución modelo que la aproxime. En los modelos generativos basados en flujos, esto se logra transportando una distribución base sencilla mediante una transformación invertible, en nuestro caso, el flujo de un sistema dinámico controlado. El entrenamiento puede formularse como un problema de control óptimo sobre el espacio de medidas de probabilidad, guiado por la minimización de una discrepancia adecuada.

El manuscrito se estructura en tres partes, avanzando desde el control de un número finito de partículas, pasando por el control de leyes de probabilidad, hasta llegar a aplicaciones de esta perspectiva dinámica—incluyendo el modelado distribucional en series temporales y el análisis de arquitecturas modernas.

Como antesala a estos bloques, reunimos en el **Capítulo** 1 las herramientas analíticas fundamentales utilizadas a lo largo de la tesis. Este capítulo cubre la teoría de flujos controlados y su regularidad (**Sección** 1.1), el transporte de medidas y la ecuación de continuidad (**Sección** 1.2), y presenta un catálogo de dinámicas elementales de EDOs neuronales (**Sección** 1.3) que resultarán esenciales para las demostraciones constructivas posteriores.

Parte I. Aprendizaje Supervisado

Esta parte estudia el aprendizaje supervisado a través del análisis de controlabilidad de flujos que actúan sobre conjuntos finitos de puntos. Las preguntas centrales que guían este análisis son: *¿qué pueden ajustar los modelos de profundidad continua, cuál es el coste en términos de complejidad y*

cómo puede aprovecharse la controlabilidad para obtener garantías poblacionales significativas? Un rasgo distintivo de nuestro enfoque es su carácter *constructivo*: en lugar de depender únicamente de teoremas de existencia, sintetizamos controles explícitos y describimos la evolución geométrica que estos inducen en el sistema.

Capítulo 2: Clasificación de Datos

Abordamos la clasificación binaria con EDOs neuronales reformulándola como una forma *relajada* de controlabilidad simultánea para una nube finita de N puntos de entrenamiento. En lugar de imponer objetivos terminales exactos punto a punto, la restricción terminal es de tipo conjunto: basta con enviar los puntos de etiquetas distintas a regiones separadas para que una capa de salida fija pueda discriminarlos.

Trabajamos con un modelo mínimo de profundidad continua (un flujo controlado de una sola neurona con parámetros constantes a trozos), donde la complejidad se mide de forma natural por el número de *discontinuidades del control*. Nuestra primera contribución es una estrategia constructiva de separación por grupos: en vez de separar los puntos uno a uno, se separan secuencialmente grupos de tamaño d (la dimensión ambiente). Esto reduce la complejidad en el peor caso de $O(N)$ a $O(N/d)$ y pone de relieve una ventaja de la alta dimensión: en el régimen de peor caso, disponer de más dimensiones permite alcanzar la separación con menos cambios en la dinámica.

Para ir más allá de las cotas del peor caso, estudiamos también la *distribución* del número de conmutaciones bajo modelos de geometría aleatoria. Para una regla de control particularmente simple que impone colinealidad en los parámetros, demostramos que la clasificación con controles constantes (o con muy pocas conmutaciones) resulta cada vez más probable a medida que crece d, y complementamos el análisis con experimentos numéricos que ilustran esta transición.

Fuente. Los resultados expuestos en este capítulo forman parte de [ÁOZ25].

> A. Álvarez-López, R. Orive-Illera, and E. Zuazua. Cluster-Based Classification with Neural ODEs via Control. In: *Journal of Machine Learning* 4.2 (2025), pp. 128–156. DOI: 10.4208/jml.241114

Código. Este capítulo incluye código del repositorio [Álv24].

> **Repository.**
> A. Álvarez-López. 2024. URL: https://github.com/antonioalvarezl/2024-WCS-NODEs

Capítulo 3: Interpolación de Datos

Pasamos a estudiar la *expresividad fuerte* en regresión: la interpolación exacta de un conjunto finito de datos mediante un único mapa de flujo. En la perspectiva de flujo/control empleada a lo largo de la tesis, la interpolación se convierte en un problema de *controlabilidad simultánea*: un único control dependiente del tiempo debe transportar todas las entradas prescritas a sus salidas asignadas en un mismo tiempo terminal.

Nuestros resultados son constructivos: proporcionamos procedimientos explícitos de síntesis de control para varias arquitecturas de EDOs neuronales y analizamos cuantitativamente su complejidad en términos de profundidad (número de conmutaciones) y anchura (número de neuronas). Esto revela un intercambio explícito entre anchura y profundidad: aumentar la anchura reduce el número

de conmutaciones necesarias, obteniendo cotas de la forma $1 + O(N/p)$ para una anchura p. En regímenes suficientemente sobreparametrizados y de alta dimensión, las construcciones pueden llevarse al límite autónomo, demostrando que la interpolación es alcanzable incluso con controles constantes.

Mostramos además que este fenómeno es robusto en modelos de flujos controlados complementarios (incluyendo variantes semi-autónomas, lineales en el control y autónomas con campos vectoriales de dos capas). Finalmente, al relajar la exigencia de ajuste exacto en tiempo terminal, obtenemos tasas de aproximación y garantías probabilísticas para la controlabilidad *aproximada* bajo supuestos geométricos naturales sobre el conjunto de datos.

Fuente. Este capítulo sintetiza resultados de tres trabajos diferentes: [ÁSZ24], [ÁGR25], y [ÁLZ26a]. La distribución exacta del contenido entre estos trabajos se detalla en el propio capítulo, justo antes de presentar los resultados.

> A. Álvarez-López, A. H. Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural ODEs. In: *Neural Networks* 180 (2024), p. 106640. DOI: 10.1016/j.neunet.2024.106640

> A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. In: *Applied Mathematics & Optimization* 92.2 (2025), p. 33. DOI: 10.1007/s00245-025-10299-7

> A. Álvarez-López, L. Liverani, and E. Zuazua. Constructive interpolation and generalization rates for neural ODEs: a control perspective. 2026. arXiv: 2606.00469

Capítulo 4: Generalización

La interpolación exacta no garantiza por sí misma el rendimiento estadístico: un modelo puede ajustar perfectamente los datos de entrenamiento y, sin embargo, fracasar al predecir nuevas observaciones. En este capítulo, proponemos una vía hacia la generalización adaptada específicamente a la perspectiva de los flujos controlados. En lugar de buscar cotas uniformes para el error de generalización sobre grandes espacios de parámetros, optamos por construir *controles dependientes de los datos* diseñados para aproximar fielmente la función objetivo en todo su dominio, asegurando así predicciones precisas ante nuevos datos.

La idea central consiste en sustituir el ajuste punto a punto por la *controlabilidad por celdas*: diseñamos flujos que actúan de forma coherente sobre las regiones de una partición, implementando así estimadores no paramétricos estructurados. En este proceso, identificamos una obstrucción topológica fundamental: el transporte de celdas macroscópicas sin que se produzcan colisiones entre trayectorias hace que los modelos puramente autónomos sean insuficientes, lo que exige una dependencia temporal explícita.

Este enfoque permite obtener cotas explícitas del riesgo poblacional para el predictor resultante, junto con los requisitos cuantitativos de complejidad del control necesarios para alcanzar una precisión determinada. Estos resultados ponen de manifiesto la interacción entre el tamaño de la muestra, la resolución geométrica y la complejidad del modelo, identificando con éxito regímenes en los que se demuestra que el riesgo tiende a cero.

Fuente. Los resultados expuestos en este capítulo forman parte de [ÁLZ26a].

> A. Álvarez-López, L. Liverani, and E. Zuazua. Constructive interpolation and generalization rates for neural ODEs: a control perspective. 2026. arXiv: 2606.00469

Parte II. Modelado Generativo

La segunda parte pasa del ajuste de nubes finitas de puntos al control y aproximación de *leyes de probabilidad*. La representación central es la medida imagen $(\Phi_T)_{\#}\mu$ (*pushforward*) inducida por un flujo alcanzable, junto con su formulación mediante la *ecuación de continuidad*. Desde una perspectiva generativa, el aprendizaje equivale a seleccionar un flujo que transporte una distribución base sencilla hacia la distribución objetivo de los datos bajo una métrica de discrepancia seleccionada. Las preguntas centrales que guían esta parte son: *¿qué medidas objetivo pueden aproximarse mediante flujos controlados, bajo qué nociones de discrepancia—con especial énfasis en las distancias de Wasserstein y de variación total— y con qué coste cuantitativo de complejidad?*

Capítulo 5: Controlabilidad para el Transporte de Medidas

Este capítulo desarrolla una teoría general de controlabilidad para *modelos generativos basados en flujos* a nivel de medidas de probabilidad. Un tema central es que el mecanismo de aproximación relevante depende fuertemente de la discrepancia utilizada para comparar dichas medidas.

Para discrepancias débiles, como la distancia de Wasserstein, demostramos que la aproximación de medidas está estrechamente ligada a la interpolación de nubes finitas de puntos, salvo reetiquetado. A nivel empírico, esta relación es exacta, mientras que para medidas generales se requiere un control adecuado sobre la distorsión geométrica de los flujos interpoladores. Esto da lugar a principios generales de elevación (*lifting*) desde la interpolación de partículas hasta la aproximación distribucional y, en particular, a resultados amplios de universalidad para familias de flujos invariantes afines en dimensión $d \geq 2$.

Para discrepancias más fuertes, como la variación total, la interpolación de nubes de puntos ya no es suficiente. Restringiéndonos a medidas absolutamente continuas, probamos la universalidad mediante una estrategia inspirada en construcciones de tipo Knothe–Rosenblatt, la cual aísla los ingredientes geométricos clave necesarios en dimensiones superiores: reversibilidad, reordenamientos unidimensionales y mezcla de coordenadas de tipo cizalla (*shear*).

Finalmente, aportamos contraejemplos explícitos que muestran que la universalidad a nivel de medidas y la interpolación de conjuntos de datos finitos son nociones genuinamente distintas: en general, la relación entre ambas depende drásticamente de la discrepancia considerada.

Fuente. Los resultados expuestos en este capítulo forman parte de [Álv+26] (en preparación final). Esta investigación se llevó a cabo durante una estancia de investigación FPU en la Universidad Nacional de Singapur, en colaboración con el grupo del profesor Qianxiao Li.

> **In preparation.**
> A. Álvarez-López et al. Flow-based approximation and controllability of measures. 2026

Capítulo 6: Flujos de Normalización Continuos

Este capítulo se centra en los *flujos de normalización continuos* (CNFs), donde el objetivo de entrenamiento habitual es la maximización de la verosimilitud. A nivel poblacional, esto equivale a minimizar la divergencia de Kullback–Leibler (entropía relativa) entre la distribución objetivo y la distribución base transportada. La entropía relativa induce una noción de aproximación sustancialmente más fuerte que la distancia de Wasserstein o la variación total; este cambio exige un control estructural más estricto sobre la densidad transportada. Nuestra contribución principal es un resultado de controlabilidad aproximada en entropía relativa para una arquitectura de CNF con controles constantes a trozos. Bajo supuestos naturales sobre el decaimiento relativo de las colas de la medida objetivo y la base, construimos controles explícitos que llevan la densidad transportada arbitrariamente cerca del objetivo en entropía relativa, y analizamos la complejidad inducida en términos del número de conmutaciones.

Fuente. Los resultados expuestos en este capítulo forman parte de [ÁGR25].

> A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. In: *Applied Mathematics & Optimization* 92.2 (2025), p. 33. DOI: 10.1007/s00245-025-10299-7

Parte III. Aplicaciones

La tercera parte reúne dos aplicaciones de la perspectiva dinámica, una guiada por datos y otra motivada por la comprensión teórica de una arquitectura más compleja.

Capítulo 7: Modelado de Dinámicas Distribucionales para Salud Digital

Estudiamos un escenario práctico motivado por la monitorización continua de glucosa, donde los datos consisten en muestras puntuales de una distribución que evoluciona en el tiempo. Modelamos la dinámica en tiempo continuo combinando representaciones de mezclas gaussianas con parámetros compartidos y una EDO neuronal que define la evolución temporal de los pesos de la mezcla en el símplex, creando así un estimador interpretable. Los pesos se ajustan primero en los instantes de observación minimizando la MMD con núcleo gaussiano—aprovechando las formas cerradas para mezclas gaussianas—y posteriormente se elevan a una trayectoria continua mediante el aprendizaje de la EDO neuronal. Esto da lugar a una representación en tiempo continuo e interpretable de la dinámica del paciente. Proporcionamos garantías teóricas para este enfoque y validamos su desempeño frente a modelos competidores en dimensiones bajas y moderadas, demostrando su valor práctico en ensayos longitudinales de diabetes. Conceptualmente, este capítulo plantea el método como una extensión de la perspectiva basada en flujos del **Part I** a series temporales de distribuciones, conservando el objetivo de aprendizaje distribucional del **Part II** al tiempo que detalla las elecciones de modelado y los mecanismos de estabilización necesarios.

Fuente. Los resultados expuestos en este capítulo forman parte de [ÁM26].

> A. Álvarez-López and M. Matabuena. Continuous-Time Learning of Probability Distributions: A Case Study in a Digital Trial of Young Children with Type 1 Diabetes. 2026. eprint: 2603.24427

Código. Este capítulo incluye código del repositorio [Álv25].

Repository.
A. Álvarez-López. 2025. URL: https://github.com/antonioalvarezl/2025-EM_MMD

Capítulo 8: Análisis de Campo Medio de los Transformers

Estudiamos un modelo de campo medio para las representaciones de los *Transformers*, en el que los *tokens* se interpretan como un sistema de partículas interactuantes que evoluciona sobre una geometría latente acotada. En este marco, la autoatención produce una interacción dependiente del estado, mientras que el bloque MLP actúa como una deriva externa que reconfigura el paisaje energético. Nuestros resultados muestran que la adición del MLP altera cualitativamente el comportamiento asintótico de la dinámica, forzando las soluciones estacionarias hacia configuraciones singulares (frecuentemente atómicas) bajo condiciones muy generales. Asimismo, proporcionamos cotas cuantitativas que describen la estructura discreta resultante. Complementamos el análisis con experimentos numéricos que ilustran la aparición de clústeres y el escalado de la complejidad del soporte. El capítulo concluye con breves comentarios sobre la arquitectura de *Mixture-of-Experts*, señalando cómo se puede interpretar dinámicamente y las cuestiones analíticas que suscita.

Fuente. Los resultados expuestos en este capítulo forman parte de [ÁGR26].

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Perceptrons and localization of attention's mean-field landscape. Accepted as a Spotlight paper at the International Conference on Machine Learning (ICML) 2026. 2026. arXiv: 2601.21366

Código. Este capítulo incluye código del repositorio [Álv26].

Repository.
A. Álvarez-López. 2026. URL: https://github.com/antonioalvarezl/2026-MLP-Attention-Energy

Chapter 1

Preliminaries

1.1 Controlled flows

Throughout the thesis we adopt a *flow-based* viewpoint: models are represented by the terminal-time map of a *controlled dynamical system*. Fix a time horizon $T > 0$ and a state dimension $d \geq 1$. We consider dynamics of the form

$$\begin{cases} \dot{x}(t) = v_{\theta(t)}\big(x(t)\big), & t \in (0, T], \\ x(0) = x_0, \end{cases} \tag{CF}$$

where $\theta : [0, T] \to \Theta$ for some suitable parameter space Θ—whose structure and dimension depend on the specific model under consideration—and $v_\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a parameterized vector field. We refer to $t \mapsto \theta(t)$ as a *control input*, emphasizing that many arguments in **Parts I and II** amount to constructing (or quantifying the existence of) such functions that enforce a desired behavior.

Assumption 1 (Well-posedness). We assume that $v_\theta \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ for all $\theta \in \Theta$. Furthermore, for every measurable control $\theta(\cdot) \colon [0, T] \to \Theta$, there exist functions $\ell_\theta, m_\theta \in L^1(0, T)$ such that, for a.e. $t \in [0, T]$,

$$\begin{aligned} \|v_{\theta(t)}(x) - v_{\theta(t)}(x')\| &\leq \ell_\theta(t)\, \|x - x'\| \quad \text{for all } x, x' \in \mathbb{R}^d, \\ \|v_{\theta(t)}(0)\| &\leq m_\theta(t). \end{aligned}$$

This assumption is satisfied, for instance, if Θ is compact, and the map $\theta \mapsto v_\theta$ is continuous.

Under Assumption 1, the Carathéodory theory yields existence and uniqueness of solutions to (CF) for every initial condition $x_0 \in \mathbb{R}^d$. For $t \in [0, T]$, we denote by

$$\Phi_t^\theta : \mathbb{R}^d \to \mathbb{R}^d, \qquad \Phi_t^\theta(x_0) \coloneqq x(t),$$

the associated *flow map* at time t, where $x(\cdot)$ is the unique solution of (CF) starting from x_0.

Proposition 1.1.1. *Let $\theta : [0, T] \to \Theta$ be a measurable control. Under Assumption 1, for every $t \in [0, T]$ the map Φ_t^θ is a bi-Lipschitz homeomorphism of $\mathbb{R}^d$ satisfying*

$$\operatorname{Lip}(\Phi_t^\theta) \ \vee \ \operatorname{Lip}\big((\Phi_t^\theta)^{-1}\big) \ \leq \ \exp\left(\int_0^t \ell_\theta(s)\, \mathrm{d}s\right)$$

and moreover, for all $x_0 \in \mathbb{R}^d$,

$$\|\Phi_t^\theta(x_0)\| \leq e^{\int_0^t \ell_\theta(s)\, \mathrm{d}s} \|x_0\| + \int_0^t \exp\left(\int_s^t \ell_\theta(r)\, \mathrm{d}r\right) \|v_{\theta(s)}(0)\|\, \mathrm{d}s. \tag{1.1.1}$$

Proof of Proposition 1.1.1. Both estimates are standard consequences of Grönwall's lemma applied to $\|x(t) - x'(t)\|$ and $\|x(t)\|$, using the Lipschitz and growth bounds in Assumption 1. The inverse Lipschitz bound follows by applying the same argument to the backward flow. □

Remark 1.1.2. When $d = 1$, every flow map $\Phi_t^\theta : \mathbb{R} \to \mathbb{R}$ is strictly increasing. Indeed, since Φ_t^θ is injective by Proposition 1.1.1, the continuous map $t \mapsto \Phi_t^\theta(y_0) - \Phi_t^\theta(x_0)$ for $x_0 < y_0$ must remain positive to avoid vanishing, which would contradict injectivity.

Consequently, scalar systems only generate increasing flows and cannot approximate nonmonotone functions, nor interpolate point sets unless the target ordering is compatible. When such rearrangements are needed, we thus work in $d \geq 2$, where this topological obstruction disappears.

1.2 Measure transport

The same dynamics can be studied at the level of measures, which is the appropriate language for **Part II**. In this setting, the flow maps Φ_t^θ play the role of *transport maps* that push probability distributions forward in time.

Given a base probability measure $\mu_{\mathrm{B}} \in \mathcal{P}(\mathbb{R}^d)$ and a control $\theta(\cdot)$, we define the *pushforward measure*

$$\mu_t^\theta := (\Phi_t^\theta)_{\#}\mu_{\mathrm{B}} \in \mathcal{P}(\mathbb{R}^d), \qquad t \in [0, T]. \tag{1.2.1}$$

Equivalently, for every bounded continuous test function $\varphi \in \mathscr{C}_b(\mathbb{R}^d)$,

$$\int \varphi(x)\, \mathrm{d}\mu_t^\theta(x) = \int \varphi\big(\Phi_t^\theta(x)\big)\, \mathrm{d}\mu_{\mathrm{B}}(x).$$

If $v_{\theta(t)}$ is Lipschitz in space, the curve $t \mapsto \mu_t^\theta$ is the (unique) *weak solution* of the *continuity* (or *conservative transport*) *equation*

$$\partial_t \mu_t^\theta + \operatorname{div}\left(v_{\theta(t)}\, \mu_t^\theta\right) = 0, \qquad \mu_{|t=0}^\theta = \mu_{\mathrm{B}}. \tag{CE}$$

Indeed, the chain rule along the flow gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\varphi\big(t, \Phi_t^\theta(x)\big) = \big(\partial_t\varphi + v_{\theta(t)} \cdot \nabla\varphi\big)\big(t, \Phi_t^\theta(x)\big) \qquad \text{for any } \varphi \in \mathscr{C}_c^1([0,T) \times \mathbb{R}^d).$$

Integrating this identity with respect to $t \in [0, T)$ and $x \sim \mu_{\mathrm{B}}$, and applying the definition of the pushforward measure $\mu_t^\theta = (\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}$, directly yields the integral formulation of a weak solution. Finally, the uniqueness of this solution is a standard consequence of the uniqueness of the underlying characteristics (the flow) for Lipschitz vector fields.

Equation (CE) is the Eulerian counterpart of the Lagrangian dynamics (CF): while the ODE (CF) tracks the path of a single particle, the PDE (CE) describes the evolution of the density of a continuum of such particles transported by the flow.

Every $\mu \in \mathcal{P}(\mathbb{R}^d)$ admits a unique Lebesgue decomposition with respect to $\mathcal{L}^d$:

$$\mu = \rho\, \mathcal{L}^d + \mu^{\mathrm{s}}, \qquad \text{for} \quad \rho \in L^1(\mathbb{R}^d; \mathbb{R}_{\geq 0}) \quad \text{and} \quad \mu^{\mathrm{s}} \perp \mathcal{L}^d,$$

where $\nu \perp \mathcal{L}^d$ means ν is concentrated on a $\mathcal{L}^d$-null set. The singular part μ^{s} further splits uniquely into a singular-continuous part and a purely atomic part, so that

$$\mu = \rho\, \mathcal{L}^d + \mu^{\mathrm{sc}} + \sum_k \alpha_k\, \delta_{x_k},$$

where $\mu^{\mathrm{sc}} \perp \mathcal{L}^d$ and $\mu^{\mathrm{sc}}(\{x\}) = 0$ for every $x \in \mathbb{R}^d$, while the last term is purely atomic (finite or countable) with $x_k \in \mathbb{R}^d$, $\alpha_k > 0$ and $\sum_k \alpha_k \leq 1$.

Proposition 1.2.1. *Let $\mu_{\mathrm{B}} = \rho_{\mathrm{B}}\,\mathcal{L}^d + \mu_{\mathrm{B}}^{\mathrm{sc}} + \sum_k \alpha_k\,\delta_{x_k}$ and fix v_θ and $\theta(\cdot) : [0,T] \to \Theta$ satisfying Assumption 1. Then the solution of (CE) satisfies*

$$\mu_t^\theta = \rho_t^\theta\,\mathcal{L}^d \;+\; (\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}^{\mathrm{sc}} \;+\; \sum_k \alpha_k\,\delta_{\Phi_t^\theta(x_k)}, \qquad t \in [0,T],$$

where each component is preserved in the following sense:

(i) $(\Phi_t^\theta)_{\#}(\rho_{\mathrm{B}}\,\mathcal{L}^d) = \rho_t^\theta\,\mathcal{L}^d$, ***(ii)*** $(\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}^{\mathrm{sc}} \perp \mathcal{L}^d$, ***(iii)*** $(\Phi_t^\theta)_{\#}\big(\sum_k \alpha_k \delta_{x_k}\big) = \sum_k \alpha_k\,\delta_{\Phi_t^\theta(x_k)}$.

Proof of Proposition 1.2.1. All items follow from the fact that Φ_t^θ is a bi-Lipschitz homeomorphism by Proposition 1.1.1. In particular, Φ_t^θ preserves Lebesgue-null sets and does not create or destroy atoms, which implies preservation of absolute continuity and singularity with respect to $\mathcal{L}^d$. Moreover, non-atomicity is preserved since for every $y \in \mathbb{R}^d$,

$$(\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}^{\mathrm{sc}}(\{y\}) = \mu_{\mathrm{B}}^{\mathrm{sc}}\big(\{(\Phi_t^\theta)^{-1}(y)\}\big) = 0.$$

□

Density transport

Assume now that $\mu_{\mathrm{B}} = \rho_{\mathrm{B}}\,\mathcal{L}^d$. By Proposition 1.2.1, absolute continuity is preserved along the flow, so that $\mu_t^\theta = \rho_t^\theta\,\mathcal{L}^d$ for all $t \in [0,T]$. In this regime, (CE) reduces to

$$\partial_t \rho_t^\theta + \operatorname{div}\big(v_{\theta(t)}\,\rho_t^\theta\big) = 0, \qquad \rho_{|t=0}^\theta = \rho_{\mathrm{B}}, \tag{DCE}$$

which corresponds to the weak formulation of (CE) by substituting $\mathrm{d}\mu_t^\theta(x)$ with $\rho_t^\theta(x)\,\mathrm{d}x$. Since Φ_t^θ and $(\Phi_t^\theta)^{-1}$ are Lipschitz, they are differentiable a.e., and the change-of-variables formula yields

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}\big((\Phi_t^\theta)^{-1}(x)\big)\,\big|\det \nabla (\Phi_t^\theta)^{-1}(x)\big| \qquad \text{for a.e. } x \in \mathbb{R}^d. \tag{1.2.2}$$

Equivalently, using the identity

$$\frac{\mathrm{d}}{\mathrm{d}t}\log\big|\det \nabla \Phi_t^\theta(x)\big| = \operatorname{div} v_{\theta(t)}\big(\Phi_t^\theta(x)\big),$$

one obtains the Liouville formula along characteristics:

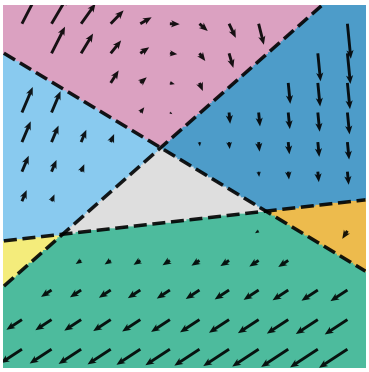
$$\rho_t^\theta(x) = \rho_{\mathrm{B}}\big((\Phi_t^\theta)^{-1}(x)\big)\,e^{-\int_0^t \operatorname{div} v_{\theta(s)}\left(\Phi_s^\theta\left((\Phi_t^\theta)^{-1}(x)\right)\right)\mathrm{d}s}. \tag{1.2.3}$$

Assuming that $\operatorname{div} v_\theta$ is uniformly bounded in space, this yields the L^p-bounds ($p \in [1,\infty]$):

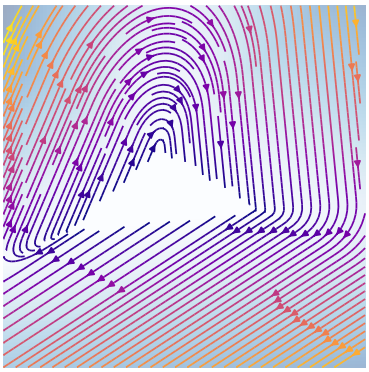
$$e^{-\frac{p-1}{p}\int_0^t \|\operatorname{div} v_{\theta(s)}\|_\infty\,\mathrm{d}s}\,\|\rho_{\mathrm{B}}\|_{L^p} \;\leq\; \|\rho_t^\theta\|_{L^p} \;\leq\; e^{\frac{p-1}{p}\int_0^t \|\operatorname{div} v_{\theta(s)}\|_\infty\,\mathrm{d}s}\,\|\rho_{\mathrm{B}}\|_{L^p}. \tag{1.2.4}$$

1.3 Neural ODE toolbox

The model (CF) becomes a *neural ODE* when the vector field $x \mapsto v_\theta(x)$ is parameterized by a neural network. In this thesis, we mostly use neural ODEs as a *constructive control* model: we design time-dependent parameters $t \mapsto \theta(t)$ so that the terminal map Φ_T^θ realizes a specific task.



(a) Polyhedral partition induced by the activation hyperplanes.



(b) Streamlines across the polyhedral cells.

Figure 1.1: Piecewise-affine geometry underlying vector fields of the form ($\mathcal{V}_p$) with $\sigma(s) = s_+$.

From the outset, it is convenient to precise about the control class. Unless stated otherwise, we work with essentially bounded controls

$$\theta(\cdot) \in L^\infty\big(0, T; \Theta\big).$$

A particularly relevant subclass for us consists of *piecewise-constant* with finitely many switches. These are defined by a partition $0 = t_0 < t_1 < \cdots < t_L < t_{L+1} = T$ such that $\theta(t)$ is constant on each interval (t_k, t_{k+1}). This matches the step-by-step structure of our constructions, where we concatenate simple flow maps over successive time windows.

If $\theta(t) = \theta_k$ on (t_k, t_{k+1}) and we denote by $\Phi_\tau^{\theta_k}$ the flow at time τ of $\dot{x} = v_{\theta_k}(x)$, then

$$\Phi_T^\theta = \Phi_{t_{L+1}-t_L}^{\theta_L} \circ \cdots \circ \Phi_{t_1-t_0}^{\theta_0}.$$

This is the continuous-time analogue of stacking residual blocks. In particular, in this control viewpoint the number of switches L plays the role of a macroscopic *depth*, while the architecture used to parameterize v_θ provides a notion of instantaneous *width*.

We typically parameterize v_θ by a one-hidden-layer network of width $p \geq 1$,

$$v_\theta(x) = W\sigma(Ax + b) = \sum_{i=1}^{p} w_i\, \sigma\big(a_i \cdot x + b_i\big), \tag{$\mathcal{V}_p$}$$

where $W \in \mathscr{M}_{d\times p}(\mathbb{R})$, $A \in \mathscr{M}_{p\times d}(\mathbb{R})$ and $b \in \mathbb{R}^p$ (in the formulation as a sum, $w_i \in \mathbb{R}^d$ is the i-th column of W, $a_i \in \mathbb{R}^d$ is the i-th row of A and $b_i \in \mathbb{R}$ is the i-th coordinate of b).

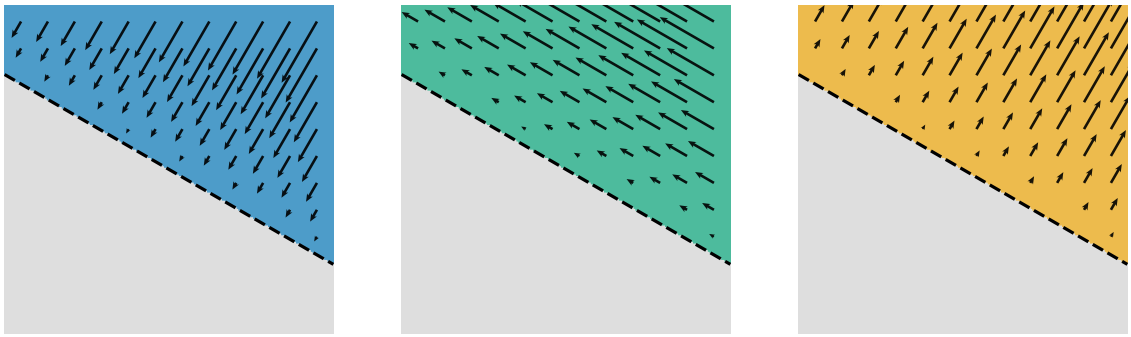
This class is geometrically transparent, yet highly expressive: the hyperplanes $\{a_i \cdot x + b_i = 0\}$ partition space into polyhedral cells. A recurrent choice for σ throughout **Parts I and II** is the *rectified linear unit* (ReLU),

$$\sigma(s) = s_+,$$

which renders v_θ piecewise affine. Despite its simplicity, this nonlinearity is enough to generate complex maps (cf. Figure 1.1).

Lemma 1.3.1. *Let v_θ be given by ($\mathcal{V}_p$), and assume that $\sigma : \mathbb{R} \to \mathbb{R}$ is ℓ_σ-Lipschitz. Then*

$$\mathrm{Lip}(v_\theta) \leq \ell_\sigma\, \|W\|_{\mathrm{op}}\, \|A\|_{\mathrm{op}}, \qquad \|v_\theta(0)\| \leq \|W\|_{\mathrm{op}}\, \big(\ell_\sigma \|b\| + \sqrt{p}\, |\sigma(0)|\big).$$



(a) **Compression** ($a \cdot w < 0$). Trajectories converge towards the hyperplane H.

(b) **Laminar shear** ($a \cdot w = 0$). Motion is parallel to the hyperplane H (constant activation).

(c) **Expansion** ($a \cdot w > 0$). Trajectories diverge away from the hyperplane H.

Figure 1.2: Basic dynamics with ($\mathcal{V}_1^+$) using $\sigma(s) = s_+$. The inactive half-space H^- remains frozen, while the active half-space H^+ exhibits three distinct regimes governed by the sign of $a \cdot w$.

In particular, if $t \mapsto (W(t), A(t), b(t))$ is an essentially bounded control, then Assumption 1 holds.

Proof of Lemma 1.3.1. By Lipschitz continuity of σ,

$$\|v_\theta(x) - v_\theta(y)\| \le \|W\|_{\mathrm{op}} \, \|\sigma(Ax+b) - \sigma(Ay+b)\| \le \|W\|_{\mathrm{op}} \, \ell_\sigma \, \|A(x-y)\| \le \ell_\sigma \, \|W\|_{\mathrm{op}} \, \|A\|_{\mathrm{op}} \, \|x-y\|.$$

Moreover, $\|\sigma(b)\| \le \sqrt{p}\,|\sigma(0)| + \ell_\sigma \|b\|$ and thus $\|v_\theta(0)\| = \|W\sigma(b)\| \le \|W\|_{\mathrm{op}} \|\sigma(b)\|$. □

Remark 1.3.2. For $\sigma(s) = s_+$ one has $\mathrm{Lip}(v_\theta) \le \|W\|_{\mathrm{op}} \|A\|_{\mathrm{op}}$ and $\|v_\theta(0)\| \le \|W\|_{\mathrm{op}} \|b\|$.

The single-neuron ansatz

The most basic building block is the field defined by a single perceptron and $\sigma(s) = s_+$:

$$v_\theta(x) = w \, (a \cdot x + b)_+, \tag{$\mathcal{V}_1^+$}$$

with $\theta = (w, a, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$. The affine map $x \mapsto a \cdot x + b$ defines the hyperplane

$$\mathsf{H} := \{x \in \mathbb{R}^d \mid \; a \cdot x + b = 0\},$$

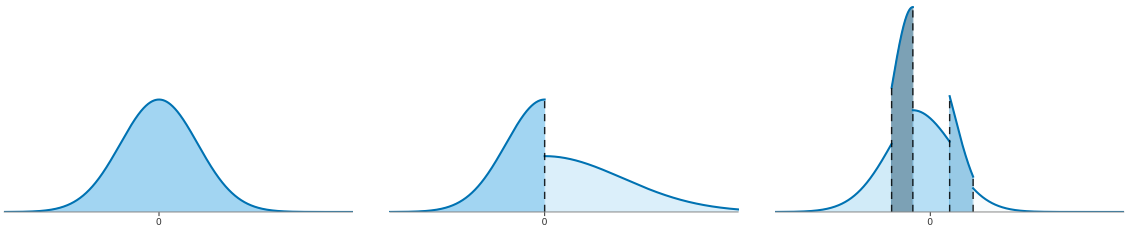
together with the half-spaces

$$\mathsf{H}^- = \{a \cdot x + b \le 0\}, \qquad \mathsf{H}^+ = \{a \cdot x + b > 0\}.$$

Thanks to ReLU, the dynamics becomes transparent: points within H^- are frozen (v_θ vanishes), while points in H^+ move along the direction w with a speed proportional to the signed distance to H (up to the factor $\|a\|$). Depending on the sign of $a \cdot w$, the active region H^+ undergoes compression, expansion, or a laminar motion (see Figure 1.2).

Proposition 1.3.3. *Fix $\theta = (w, a, b)$ and let Φ_t^θ be the flow of (CF)–($\mathcal{V}_1^+$). Then H^-, H, and H^+ are forward invariant. More precisely, for any $x_0 \in \mathbb{R}^d$:*

(i) *If $a \cdot x_0 + b \le 0$ then $\Phi_t^\theta(x_0) = x_0$ for all $t \ge 0$.*



(a) Initial Gaussian ρ_{B}. (b) One flow step. (c) Four flow steps.

Figure 1.3: Density evolution under single-neuron ReLU flows (Lemma 1.3.5). Starting from a standard Gaussian density, each flow step (DCE_1^+) introduces a kink at the corresponding hyperplane (dashed lines). Concatenating L steps thus yields a piecewise Gaussian density with up to L breakpoints. The shading of each Gaussian is inversely proportional to its variance.

(ii) *If $a \cdot x_0 + b > 0$, then*

$$\Phi_t^\theta(x_0) = x_0 + w\,(a \cdot x_0 + b) \times \begin{cases} \dfrac{e^{a \cdot w\, t} - 1}{a \cdot w}, & \text{if } a \cdot w \neq 0, \\ t, & \text{if } a \cdot w = 0. \end{cases} \tag{1.3.1}$$

In particular, the trajectory $\Phi_t^\theta(x_0)$ remains in the affine line $x_0 + \mathrm{span}\{w\}$ for all $t \geq 0$.

Proof of Proposition 1.3.3. If $a \cdot x_0 + b \leq 0$, then $v_\theta(x_0) = 0$ and the constant path $x(t) \equiv x_0$ solves (CF). By uniqueness of solutions, we obtain (i).

Now assume $x_0 \in \mathsf{H}^+$ and let $s(t) \coloneqq a \cdot x(t) + b$. As long as $s(t) > 0$, the dynamics $\dot{x} = ws(t)$ implies $\dot{s} = (a \cdot w)s$. Solving this linear ODE gives

$$a \cdot \Phi_t^\theta(x_0) + b = (a \cdot x_0 + b)e^{a \cdot w\, t}.$$

Since the exponential is always positive, $s(t)$ never changes sign; thus H^+ is forward invariant. Integrating $\dot{x}(t) = w(a \cdot x_0 + b)e^{a \cdot w\, t}$ yields (1.3.1). The fact that $\Phi_t^\theta(x_0)$ stays in the affine line $x_0 + \mathrm{span}\{w\}$ follows immediately from $a \cdot \Phi_t^\theta(x_0) + b = (a \cdot x_0 + b)\, e^{a \cdot w\, t}$. □

Remark 1.3.4. The explicit formula (1.3.1) will be used repeatedly to build controls by concatenation. Two elementary features are worth keeping in mind.

(a) Half-space gating. ReLU enables a *freeze-then-move* mechanism: by choosing H such that a given set is contained in H^-, we localize the action of v_θ to H^+, allowing for the selective movement of points.

(b) One-sided dilations. For $d = 1$, taking $a = 1$ and $b = -z$ gives $v_\theta(x) = w(x - z)_+$ and

$$\Phi_t(x) = x \quad \text{for } x \leq z, \qquad \Phi_t(x) = z + (x - z)e^{wt} \quad \text{for } x > z.$$

Concatenating such one-sided operations (with piecewise-constant controls) is a convenient way to implement monotone piecewise-affine maps on intervals; this mechanism will reappear in **Part II**.

So far the toolbox has been described in a Lagrangian language (moving points). We now switch to the Eulerian viewpoint and study the evolution of a scalar density under (DCE)–($\mathcal{V}_1^+$):

$$\partial_t \rho_t^\theta(x) + \partial_x \left(w(t)\ (a(t)\cdot x + b(t))_+\ \rho_t^\theta(x) \right) = 0, \qquad \rho^\theta_{|t=0} = \rho_{\mathrm{B}}. \tag{DCE$_1^+$}$$

The next lemma gives the explicit solution.

Lemma 1.3.5. *Take $T > 0$ and a probability density $\rho_{\mathrm{B}} \in L^1(\mathbb{R})$. Assume $\sigma(s) = s_+$ and let $\theta(t) \equiv (w, a, b) \in \mathbb{R}^3$ be constant with $a \neq 0$. Then the (unique) weak solution $\rho^\theta \in \mathscr{C}^0([0,T]; L^1(\mathbb{R}))$ of (DCE$_1^+$) is given for a.e. $x \in \mathbb{R}$ by*

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}(x)\, \mathbf{1}_{\{a\cdot+b\leq 0\}}(x) + e^{-wat}\, \rho_{\mathrm{B}}\left(-\frac{b}{a} + \left(x + \frac{b}{a} \right) e^{-wat} \right) \mathbf{1}_{\{a\cdot+b>0\}}(x), \qquad t \in [0,T].$$

Proof of Lemma 1.3.5. By Proposition 1.3.3, the flow Φ_t is the identity on $\{ax + b \leq 0\}$. On the region $\{ax + b > 0\}$, the characteristics solve $\dot{x} = w(ax + b)$, which yields

$$\Phi_t(x_0) = -\frac{b}{a} + \left(x_0 + \frac{b}{a} \right) e^{wat}, \qquad \text{for } x_0 \in \{x \mid ax + b > 0\}.$$

Since Φ_t is a $\mathscr{C}^1$-diffeomorphism on $\{ax + b > 0\}$ and the boundary $\{ax + b = 0\}$ has measure zero, (1.2.3) directly gives the expression for $\rho_t^\theta(x)$. The continuity $t \mapsto \rho_t^\theta$ in $L^1(\mathbb{R})$ is immediate from the explicit formula: the first term is independent of t, while the second term is obtained by applying to ρ_{B} a composition of a translation and a dilation whose parameters depend continuously on t. Since the action of translations and dilations is strongly continuous on $L^1(\mathbb{R})$, the curve $t \mapsto \rho_t^\theta$ is continuous, completing the proof. □

Remark 1.3.6. If $a = 0$, then $v_\theta(x) = w\,(b)_+$ is constant, and the solution is a translation:

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}\big(x - w\,(b)_+\, t\big), \qquad (t,x) \in [0,T] \times \mathbb{R}.$$

Part I

Supervised Learning

Overview

At the heart of machine learning lies *supervised learning* [LBH15], a framework that has been successfully applied across a wide number of domains [Lec+98; Ben+03]. Models based on continuous flows are now broadly used, with many coexisting variants differing in aspects such as parameterizations, readouts, and discretizations. A recurring methodological challenge is to evaluate and compare these models within a common mathematical framework.

From a learning perspective, a fundamental tension arises between two objectives: *expressivity*, understood as the capacity to fit the *training data*, and *generalization*, the ability to perform well on unseen inputs rather than merely memorizing the sample. Successful learning requires a delicate balance: the model must be rich enough to capture underlying patterns, yet sufficiently constrained to avoid overfitting. In **Part I**, we analyze this trade-off through the controllability properties of the induced flow maps.

Supervised learning with flow-based predictors

The standard supervised learning setup considers a *dataset* of size $N \geqslant 1$. This dataset consists of *inputs* x_i from an input space $\mathcal{X}$ and corresponding *labels* (or *targets*) y_i from a *label space* $\mathcal{Y}$:

$$\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}, \qquad \text{with } x_i \neq x_j \text{ for } i \neq j. \tag{I.1}$$

Unless stated otherwise, we assume a Euclidean input space, $\mathcal{X} = \mathbb{R}^d$. The choice of the label space dictates the learning task: we focus on $\mathcal{Y} = \mathbb{R}^d$ for *regression* and $\mathcal{Y} = \{-1, 1\}$ for *binary classification*. The distinctness assumption on the inputs is a harmless convenience to avoid conflicting labels. The goal is to use the information contained in $\mathcal{D}_N$ to infer the underlying *ground-truth* mapping $y(\cdot)\colon \mathcal{X} \to \mathcal{Y}$.

Throughout this thesis, we address this task using a *flow-based predictor*, obtained by composing a terminal-time flow map with a fixed *readout*. Concretely, letting $\Phi_t^\theta\colon \mathbb{R}^d \to \mathbb{R}^d$ denote the *flow* generated by the controlled dynamics in (CF), we consider predictors of the form

$$F_\theta \coloneqq g \circ \Phi_T^\theta\colon \mathbb{R}^d \to \mathbb{R}^m, \tag{I.2}$$

where the readout $g\colon \mathbb{R}^d \to \mathbb{R}^m$ is task-dependent (e.g., the identity for regression, a linear projection for classification). The trainable object is the control $\theta(\cdot)$, possibly under regularity constraints.

To quantify performance away from the observed data, we fix a reference probability measure $\mu \in \mathcal{P}(\mathcal{X})$ and define the *population (true) risk* as

$$\mathcal{R}(\theta) \coloneqq \int \ell\big(F_\theta(x), y(x)\big)\, \mathrm{d}\mu(x), \tag{I.3}$$

for a *loss function* $\ell\colon \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}_{\geqslant 0}$. Here, μ should be viewed as a reference *test distribution*: it specifies where and how we evaluate performance—in many settings, μ is interpreted as the (unknown) *data-generating law*. Implicit in (I.3) is the assumption that the labels are perfectly generated by the deterministic ground truth $y(\cdot)$, corresponding to the *noiseless* regime. (More

generally, one may consider $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ to accommodate stochastic labels capturing non-deterministic dependencies between inputs and targets.)

Since μ is unknown, $\mathcal{R}(\theta)$ cannot be computed, let alone minimized, from the *training set* $\mathcal{D}_N$. Instead, one introduces the *empirical measure* $\mu_N := \frac{1}{N}\sum_{i=1}^N \delta_{x_i} \in \mathcal{P}(\mathcal{X})$, and the *empirical risk*

$$\mathcal{R}_N(\theta) := \int \ell\big(F_\theta(x), y(x)\big)\, \mathrm{d}\mu_N(x) = \frac{1}{N}\sum_{i=1}^N \ell\big(F_\theta(x_i), y_i\big). \tag{I.4}$$

Training consists of selecting a *data-dependent control*. A standard choice is (regularized) *empirical risk minimization* over a class of *admissible controls* $\mathcal{U}$ (e.g., a subset of $L^\infty(0,T;\Theta)$):

$$\theta_N \in \arg\min_{\theta\in\mathcal{U}} \big\{\mathcal{R}_N(\theta) + \lambda\|\theta\|_{\mathcal{U}}\big\}, \tag{I.5}$$

for some penalty functional $\|\cdot\|_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}_{\geqslant 0}$ (which, despite the notation, need not be a formal norm) and some regularization parameter $\lambda > 0$. Here, $\|\cdot\|_{\mathcal{U}}$ encodes the preference for *simpler* controls; equivalently, one can view *regularization* as restricting $\mathcal{U}$ so as to control *model complexity*.

The proxy principle behind empirical risk minimization is captured by the bound:

$$\mathcal{R}(\theta) \leqslant \mathcal{R}_N(\theta) + \big|\mathcal{R}(\theta) - \mathcal{R}_N(\theta)\big|. \tag{I.6}$$

The first term measures *fit* on the observed sample, while the second is the *generalization gap*, i.e., the discrepancy between testing under μ and testing on the empirical proxy μ_N. One typically aims to make the fit term small by increasing expressivity, while keeping the gap small by controlling the complexity of $\mathcal{U}$.

Controllability as a bridge between expressivity and generalization

Classical generalization theory often proceeds by bounding $\sup_{\theta(\cdot)\in\mathcal{U}} |\mathcal{R}(\theta) - \mathcal{R}_N(\theta)|$ in probability. We return to this perspective in **Chapter 4**.

In **Part I** we adopt a different viewpoint: we treat $\theta(\cdot)$ as a control and study *controllability* properties of the induced family of flow maps. This makes it possible to (i) build explicit parameters with prescribed behavior and (ii) quantify the control complexity required to enforce that behavior.

Chapters 2, 3, and 4 can be read as a progression from weaker to stronger target constraints:

- ***Relaxed targets* (*classification*):** correct prediction requires terminal representations to fall into class-dependent decision regions, not to match pointwise targets.
- ***Exact targets* (*interpolation*):** regression-type objectives lead naturally to simultaneous pointwise interpolation constraints.
- **From points to cells (*generalization*):** to obtain population guarantees we control not only the training points but the behavior of the flow on cells of a partition, enabling risk estimates for specific predictors.

Roadmap of Part I

For navigation, **Part I** consists of the following three chapters.

- **Chapter 2**: Data Classification.
- **Chapter 3**: Data Interpolation.
- **Chapter 4**: Generalization.

Chapter 2

Data Classification

2.1 Introduction

In this chapter, we apply the supervised-learning framework introduced in **Part I (Overview)** to *data classification*, where the label space $\mathcal{Y}$ is finite, say $\mathcal{Y} = \{1, \dots, K\}$ for some integer $K \geqslant 2$. Many classical and modern methods fit within this paradigm (e.g. linear discriminants [Fis36], support vector machines [CV95] or random forests [Bre01]). Here we adopt the flow-based viewpoint: predictors are of the form (I.2), that is,

$$F_\theta = g \circ \Phi_T^\theta,$$

where Φ_T^θ is the terminal-time flow map induced by a controlled dynamics and g is a fixed readout.

Standard classification models typically identify labels $y \in \mathcal{Y}$ with canonical vectors in $\mathbb{R}^K$ (*one-hot encoding*): the y-th coordinate is 1 and all others are 0. Predictors such as F_θ optimize surrogate losses (typically, *cross-entropy*) to output a vector of scores (or logits) in $\mathbb{R}^K$. The predicted class is then defined as the index of the largest score.

Instead, we adopt a purely geometric viewpoint. What matters here is not probability calibration, but the existence of a representation for which a simple, fixed readout g linearly separates the classes. The dynamics acts as the mechanism to build such a representation. Since g is fixed a priori, the burden of classification falls entirely on Φ_T^θ: the flow must transport data points into disjoint, class-dependent decision regions at time T. This formalizes the *relaxed targets* principle introduced in **Part I (Overview)**: unlike regression, we do not require matching prescribed targets pointwise, only membership in the correct region.

Our main goal is to quantify the *control complexity* required for separation, measured by the *number of switches* L in the piecewise-constant model ($\mathcal{V}_1^+$). We frame this analysis in a *minimax* sense: we seek the minimum L sufficient to classify any dataset of a given size, regardless of its geometric arrangement. Our results highlight how a higher dimension d reduces the complexity needed for classification in the *worst-case scenario*.

Related work

The theoretical study of neural ODEs—and controlled flows in general—from a control theory perspective has gained significant attention in recent years, opening several research directions.

One line of research explores the approximation power of flows, using either geometric techniques based on Lie brackets [TG23; CLT20; Ela+23] or constructive methods for simultaneous control [LLS22; RZ23]. In [Che+25c], the authors conduct a detailed study of controllability and its connection with density in L^p spaces for $p < +\infty$. Universal approximation in the uniform norm is

established for a specific class of diffeomorphisms in [Ish+23]. We also highlight the work of [LJ18], which demonstrates density in $L^1(\mathbb{R}^d)$ for the family of ResNets with one neuron per hidden layer, corresponding to the forward Euler discretization of $(\mathcal{V}_1^+)$.

Another approach studies the training process as an optimal control problem. In [EHL18], population risk minimization in deep learning is formulated as a mean-field optimal control problem, and Hamilton–Jacobi–Bellman and Pontryagin optimality conditions are derived. The authors of [Li+18] propose a modification of the successive approximations method by augmenting the Hamiltonian to solve Pontryagin's maximum principle, resulting in an alternative training algorithm with rigorous error estimates. The study in [Bon+23] employs a continuity equation to analyze the mean-field dynamics, examining the existence and uniqueness of minimizers in the optimal control problem with L^2-regularization and establishing a mean-field maximum principle. In [IO24], this result is improved by introducing a kinetic regularization term in the loss function and proving the existence of minimizers. The work [GZ22] considers classical empirical risk minimization, deriving a manifestation of the turnpike property through specific regularization terms. Additionally, [EG23] reduces the complexity for data classification by introducing an L^1-norm penalty, which promotes temporal sparsity in the control. Complementing these deterministic perspectives, the stochastic nature of the training process has also been analyzed through the lens of partial differential equations. Notably, [BBI25] studies the behavior of stochastic gradient descent via Fokker-Planck equations, providing quantitative estimates on the escape times from suboptimal local minima and addressing asymptotic convergence for non-convex loss functions.

This chapter follows a third direction: estimating the complexity required to interpolate N points in the weak sense defined by linear classification. The only known study addressing this specific problem is [RZ23], where a bound of $O(N)$ switches using $(\mathcal{V}_1^+)$ is established. We build directly upon their results, reducing this bound to $O(N/d)$ switches and deriving a novel probabilistic guarantee.

2.2 Setup and main results

Let $\mathcal{D} \coloneqq \{(x_i, y_i)\}_i \subset \mathbb{R}^d \times \{0,1\}$ be a finite dataset such that $x_i \neq x_j$ for all $i \neq j$. We define the classes $\mathcal{R}$ (red circles) and $\mathcal{B}$ (blue crosses) by

$$\mathcal{R} = \{x_i \in \mathbb{R}^d \mid (x_i, 1) \in \mathcal{D}\}, \qquad \mathcal{B} = \{x_i \in \mathbb{R}^d \mid (x_i, 0) \in \mathcal{D}\}. \tag{2.2.1}$$

We work with the one-neuron model introduced in $(\mathcal{V}_1^+)$ and $\sigma(s) = s_+$, recalled below:

$$\dot{x}(t) = w(t)(a(t) \cdot x(t) + b(t))_+, \qquad \text{for } t \in [0,T]. \tag{$\mathcal{V}_1^+$}$$

We view $\theta(\cdot) = (w(\cdot), a(\cdot), b(\cdot))$ as a control and take

$$\theta \in L^\infty(0,T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}).$$

We further restrict to piecewise-constant controls, which reduces the learning problem to a finite-dimensional parameterization [Mas+20] and induces a layered structure akin to a discrete ResNet [RZ23; ÁSZ24]. We say that such a control has L discontinuities (or switches) if there exist times

$$0 =: t_0 < t_1 < \cdots < t_L < t_{L+1} \coloneqq T$$

such that θ is constant on each interval (t_{k-1}, t_k), for $k \in [L+1]$. The integer $L \geq 0$ is our notion of *temporal complexity*.

In our setting, the readout is fixed and binary. Classification is modeled by a separation requirement at terminal time: we seek a control θ such that

$$\Phi_T^\theta(\mathcal{R}) \subset \tau_{\mathcal{R}}, \qquad \Phi_T^\theta(\mathcal{B}) \subset \tau_{\mathcal{B}}, \tag{2.2.2}$$

where $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ is a pair of linearly separable regions in $\mathbb{R}^d$. We fix a coordinate index $\ell_\star \in [d]$ and take the target regions $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ to be half-spaces of the form

$$\left(\{x^{(\ell_\star)} > 1\}, \{x^{(\ell_\star)} \leqslant 1\}\right) \quad \text{or} \quad \left(\{x^{(\ell_\star)} \leqslant 1\}, \{x^{(\ell_\star)} > 1\}\right), \tag{2.2.3}$$

with decision boundary $\{x^{(\ell_\star)} = 1\}$. The specific threshold value 1 is only a convenient normalization; any affine threshold in the $\ell_\star$-th coordinate can be reduced to this form by translation and scaling. In Theorem 2.2.2, we will allow the algorithm to select a favorable coordinate direction among the canonical axes, which amounts to choosing $\ell_\star$ (and the corresponding half-spaces) a posteriori.

As discussed in **Section 1.3**, if θ is constant, then the dynamics acts relative to an affine hyperplane $\{a \cdot x + b = 0\}$: one half-space is transported along the direction w, while the complement remains fixed. For θ piecewise constant, the classification problem becomes geometric: we must construct a sequence of such hyperplanes whose associated transports achieve the separation goal (2.2.2), and the number of switches L counts how many hyperplanes we deploy.

The following questions naturally arise:

Question 1. What is the minimum number of switches L that is sufficient to guarantee classification of *any* finite dataset of a given size?

Question 2. What is the probability that a given dataset can be classified using exactly $L = k$ discontinuities, for a given $k \geqslant 0$?

Regarding Question 1, it was established in [RZ23] that $L = 3(|\mathcal{R}| \wedge |\mathcal{B}|)$ switches are sufficient to classify $(\mathcal{R}, \mathcal{B})$ using $(\mathcal{V}_1^+)$. The strategy in [RZ23] involves sequentially steering each point in $\mathcal{R} \cup \mathcal{B}$ individually to its corresponding target $\tau_{\mathcal{R}}$ or $\tau_{\mathcal{B}}$. Our minimax result refines this bound by relying on the concept of general position, a classical notion in combinatorial geometry [MFK94; Cov65; Son97]. We consider the following definition, illustrated in Figure 2.1:

Definition 2.2.1. A set $\mathcal{X} \subset \mathbb{R}^d$ is in *general position* (GP) if, for every $0 \leqslant m \leqslant d-1$, no affine subspace in $\mathbb{R}^d$ of dimension m contains more than $m+1$ points of $\mathcal{X}$.

Assuming this mild condition—easily achieved by slight perturbations of the dataset—we construct a family of pairwise parallel hyperplanes that partition the points of one class into subsets of size d. We can then define a simultaneous control method that classifies the points within these subsets together rather than individually, thereby reducing the upper bound for L:

Theorem 2.2.1. *Let $d \geqslant 2$. For any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1) in general position and any pair of target regions $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ as in (2.2.3), there exist $T > 0$ and a piecewise-constant control $\theta \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ with*

$$L = 4 \left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil - 1$$

discontinuities, such that the flow map of $(\mathcal{V}_1^+)$ satisfies $\Phi_T^\theta(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T^\theta(\mathcal{B}) \subset \tau_{\mathcal{B}}$.

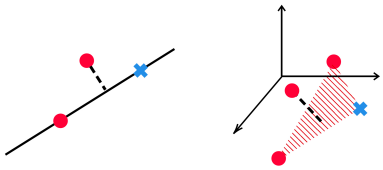


Figure 2.1: $\mathcal{X} \subset \mathbb{R}^2$ is in GP if no three points of $\mathcal{X}$ lie on the same line. $\mathcal{X} \subset \mathbb{R}^3$ is in GP if, additionally, no four points lie on the same plane.

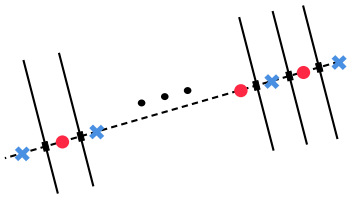


Figure 2.2: Configuration where the maximum number of $2N-1$ hyperplanes is required to separate the points.

The proof of this theorem is presented in **Section 2.4** along with the necessary tools.

Theorem 2.2.1 holds almost surely as long as the points are sampled from a nonsingular probability measure. However, in practice, numerous data points sharing the same label may initially be clustered together, enabling their simultaneous control with fewer parameters. This observation motivates the introduction of a probabilistic framework to address Question 2.

While the minimax analysis considers the hardest geometric configurations, we now investigate the complexity distribution when the points are random (pure noise). We focus on balanced datasets ($|\mathcal{R}| = |\mathcal{B}| = N$), which typically represent the most challenging combinatorial case for separation. To simplify the analysis and obtain exact distributions, we introduce a new control method that restricts $a(t)$ to be an optimally selected constant.

Theorem 2.2.2. *Fix $d \geqslant 2$ and $N \geqslant 1$. Consider any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1), with $|\mathcal{R}| = |\mathcal{B}| = N$. Assume that every point $x \in \mathcal{R} \cup \mathcal{B}$ is independently sampled from an absolutely continuous probability measure on $[0,1]^d$. Then, with probability 1, there exist $T > 0$, a pair of target regions $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ as in (2.2.3), and a control $\theta = (w, a, b) \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ with*

$$a(t) \in \{e_1, \dots, e_d\} \qquad \text{constant for all } t \in [0, T],$$

such that the flow map of $(\mathcal{V}_1^+)$ satisfies $\Phi_T^\theta(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T^\theta(\mathcal{B}) \subset \tau_{\mathcal{B}}$.

Furthermore, θ is piecewise constant, and its number of switches L satisfies, for $0 \leqslant k \leqslant 2N-2$,

$$\mathbb{P}(L \geqslant k) = \left(\sum_{p=\lceil \frac{k}{2}+1 \rceil}^{N} \binom{N-1}{p-1}^2 + \sum_{p=\lceil \frac{k+1}{2} \rceil}^{N-1} \binom{N-1}{p} \binom{N-1}{p-1} \right)^d 2^d \binom{2N}{N}^{-d}. \tag{2.2.4}$$

Remark 2.2.3. Stirling's approximation simplifies (2.2.4) for large values of d and N. If we set $k = 1$, the formula reduces to

$$\mathbb{P}(L \geqslant 1) = \left(1 - 2\binom{2N}{N}^{-1}\right)^d \sim \exp\left\{-\frac{\sqrt{\pi N}}{2^{2N-1}}\, d\right\}.$$

Assuming both d and N grow according to some relation $d = d(N)$, we can deduce that

$$\text{If} \quad \lim_{N\to\infty} \frac{2^{2N}}{d(N)\sqrt{N}} = 0, \qquad \text{then} \quad \lim_{N\to\infty} \mathbb{P}(L = 0) = 1.$$

This observation reveals an explicit trade-off between d and N that allows classifying all points employing a constant control—or equivalently, with an autonomous neural ODE.

The control method of Theorem 2.2.2 is designed such that the distribution (2.2.4) is derived from the number of hyperplanes required to separate the points by labels. The maximum value $L = 2N - 2$, corresponding to $2N - 1$ hyperplanes, also improves upon the bound of [RZ23].

In the following theorem, we return to the deterministic perspective to show that the maximum number of hyperplanes required to separate the two classes is indeed $2N - 1$. Moreover, we characterize the pathological configurations that attain this maximum, as illustrated in Figure 2.2.

Theorem 2.2.4. *Fix $d, N \geqslant 1$. For any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1) with $|\mathcal{R}| = |\mathcal{B}| = N$, the maximum number of hyperplanes in $\mathbb{R}^d$ required to separate the points by labels is $2N - 1$. Furthermore, this maximum is attained if and only if the points of $\mathcal{R}$ and $\mathcal{B}$ are collinear and alternating; i.e., there exist $u, u_0 \in \mathbb{R}^d$ and $-\infty < \lambda_1 < \cdots < \lambda_{2N} < +\infty$ such that*

$$u_0 + \lambda_{2k-1} u \in \mathcal{R} \qquad \text{and} \qquad u_0 + \lambda_{2k} u \in \mathcal{B} \qquad \text{for all } k \in [N].$$

We note that deriving the maximum is relatively straightforward; the main challenge lies in showing that it is attained only when all points are collinear and alternating. The proofs of Theorems 2.2.2 and 2.2.4 are presented in **Section 2.3**.

In **Section 2.5**, we complement the analysis with numerical experiments, using gradient-based training to estimate the empirical minimal complexity required for binary classification in the worst-case scenario, thus validating our theoretical bounds.

2.3 Classification via canonical separability

This section is centered on Theorems 2.2.2 and 2.2.4. We first translate the switching complexity of a restricted control strategy into a purely geometric quantity (the number of axis-orthogonal separating hyperplanes), and then compute its distribution in the i.i.d. setting. Several auxiliary combinatorial statements are stated as needed; their proofs are deferred to **Section 2.8**.

Suppose that all points of (2.2.1) are independently sampled from an absolutely continuous probability measure on $[0,1]^d$, with $|\mathcal{R}| = |\mathcal{B}| = N$ fixed, and that they satisfy the following condition:

$$x_i^{(\ell)} \neq x_j^{(\ell)} \qquad \text{for all } \ell \in [d] \quad \text{and} \quad i \neq j. \tag{2.3.1}$$

Note that (2.3.1) is fulfilled with probability 1. We now introduce our concept of *separability*:

Definition 2.3.1. A finite family $\mathcal{H}$ of affine $(d-1)$-dimensional hyperplanes in $\mathbb{R}^d$ *separates* $(\mathcal{R}, \mathcal{B})$ if it partitions the cube $[0,1]^d$ into polyhedra, each containing points of at most one of the two sets. We say that such $\mathcal{H}$ is a collection of *separating hyperplanes* for $(\mathcal{R}, \mathcal{B})$.

Note that some polyhedra may contain no points from either set. Now, for any pair $(\mathcal{R}, \mathcal{B})$ as in (2.2.1), and any collection of hyperplanes $\mathcal{H}$ in $\mathbb{R}^d$, let us consider

$$Z_{d,N}(\mathcal{R}, \mathcal{B}) := \min\big\{|\mathcal{H}| \mid \mathcal{H} \text{ separates } (\mathcal{R}, \mathcal{B})\big\}. \tag{2.3.2}$$

As defined, $Z_{d,N}$ is a random variable that maps each pair $(\mathcal{R}, \mathcal{B})$ to the minimum cardinality of any collection of separating hyperplanes for $(\mathcal{R}, \mathcal{B})$.

Theorem 2.2.2 is based on separating $(\mathcal{R}, \mathcal{B})$ using hyperplanes that are orthogonal to an optimally chosen canonical vector. We refer to this approach as *canonical k-separability*, drawing on the concept of *k-separability* introduced in [Duc06]. A finite set $\{x_i\} \subset \mathbb{R}^d$ with binary labels is

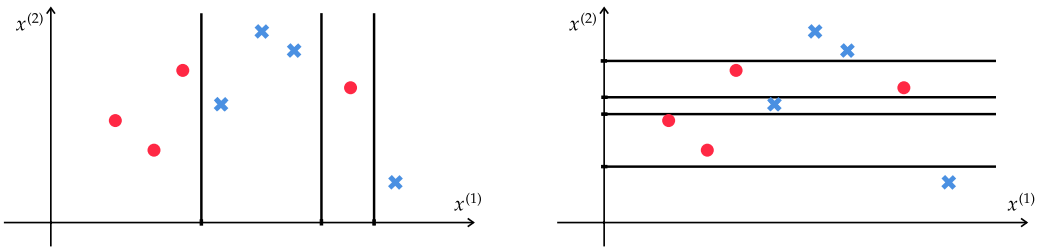


Figure 2.3: $Z^1_{d,N} = 3$ and $Z^2_{d,N} = 4$, computed by projecting data on the axes $x^{(1)}$ and $x^{(2)}$.

k-separable if there exists $w \in \mathbb{S}^{d-1}$ such that the projections $w \cdot x_i$ can be divided into k disjoint intervals, each containing only projections with the same label.

Canonical k-separability is a similar concept, but it constrains w to the canonical basis. We will determine the exact probability that $(\mathcal{R}, \mathcal{B})$ is k-separable under this constraint, which requires adapting the definition of $Z_{d,N}$ in (2.3.2).

For each $\ell \in [d]$, the usual projection $\pi^\ell : \mathbb{R}^d \to \mathbb{R}$ is defined by $\pi^\ell(x) = x^{(\ell)}$ for all $x \in \mathbb{R}^d$. Let Π^ℓ be the pointwise extension of π^ℓ, given by

$$\Pi^\ell(\mathcal{X}) = \{\pi^\ell(x_1), \dots, \pi^\ell(x_N)\} = \left\{x_1^{(\ell)}, \dots, x_N^{(\ell)}\right\} \tag{2.3.3}$$

for all $\mathcal{X} = \{x_i\}_{i=1}^N \subset \mathbb{R}^d$ satisfying condition (2.3.1), which ensures $|\Pi^\ell(\mathcal{X})| = N$.

We introduce the random variable $Z^\ell_{d,N}$ such that

$$Z^\ell_{d,N}(\mathcal{R}, \mathcal{B}) = Z_{1,N}\left(\Pi^\ell(\mathcal{R}), \Pi^\ell(\mathcal{B})\right) \tag{2.3.4}$$

for any pair $(\mathcal{R}, \mathcal{B})$ as in (2.2.1). The value $Z^\ell_{d,N}(\mathcal{R}, \mathcal{B})$ represents the minimum number of points required to separate the projections of both sets on the ℓ-th Cartesian axis. Those points determine a family of parallel hyperplanes in $\mathbb{R}^d$ given by the equations $x^{(\ell)} = h_k$ with $1 \leqslant k \leqslant Z^\ell_{d,N}(\mathcal{R}, \mathcal{B})$, which separate $(\mathcal{R}, \mathcal{B})$ in $\mathbb{R}^d$ (see Figure 2.3).

Given $d, N \geqslant 1$, we define

$$Z^\perp_{d,N} \coloneqq \min\left\{Z^1_{d,N}, \dots, Z^d_{d,N}\right\}. \tag{2.3.5}$$

The variable $Z^\perp_{d,N}$ represents the minimum cardinality of any family of separating hyperplanes, all of which are perpendicular to some Cartesian axis. For the example shown in Figure 2.3, we would compute $Z^\perp_{d,N}(\mathcal{R}, \mathcal{B}) = 3$.

Now, we prove Theorem 2.2.4 to determine and characterize the maximum value of $Z_{d,N}$:

Proof of Theorem 2.2.4. By relabeling the points, without loss of generality we can assume

$$\mathcal{R} = \{x_i\}_{1 \leqslant i \leqslant N} \qquad \text{and} \qquad \mathcal{B} = \{x_{N+i}\}_{1 \leqslant i \leqslant N}.$$

First, we prove that

$$\max_{(\mathcal{R},\mathcal{B})} Z_{d,N}(\mathcal{R}, \mathcal{B}) = 2N - 1. \tag{2.3.6}$$

Case $d = 1$. When the points of the pair $(\mathcal{R}, \mathcal{B})$ are alternating, i.e.,

$$x_1 < x_{N+1} < x_2 < \cdots < x_N < x_{2N} \quad \text{or} \quad x_{N+1} < x_1 < x_{N+2} < \cdots < x_{2N} < x_N, \tag{2.3.7}$$

we have $Z_{1,N}(\mathcal{R},\mathcal{B}) = 2N-1$. Conversely, if the points of $\mathcal{R}\cup\mathcal{B}$ are not alternating, then the number of line segments connecting consecutive points with different labels is less than $2N-1$, which proves (2.3.6) for $d=1$.

Case $d>1$. Since $Z_{d,N} \leqslant Z_{d,N}^{\perp}$ as defined in (2.3.5), we deduce that $\max Z_{d,N} \leqslant 2N-1$ by applying (2.3.6) coordinate-wise. To show that $\max Z_{d,N} \geqslant 2N-1$, assume the $2N$ points of $\mathcal{R}\cup\mathcal{B}$ are collinear and alternating. To separate $(\mathcal{R},\mathcal{B})$, each of the $2N-1$ line segments connecting two consecutive points of different labels must then be intersected by a hyperplane. Since any hyperplane that does not contain the entire line can intersect it in at most one point, we need at least $2N-1$ hyperplanes to separate $(\mathcal{R},\mathcal{B})$, as illustrated in Figure 2.2.

Necessity of collinearity and alternation. If the points of $\mathcal{R}\cup\mathcal{B}$ are collinear but not alternating, then the number of line segments connecting consecutive points with different labels is less than $2N-1$. Consequently, fewer than $2N-1$ hyperplanes suffice to separate them; that is, $Z_{d,N} < 2N-1$.

To show that collinearity is also a necessary condition to attain the maximum, we proceed by induction on N. For $N=1$, the claim is trivial. For $N=2$, if the four points are not collinear, their convex hull is at most a triangle, quadrilateral, or tetrahedron, which implies $Z_{d,N} \leqslant 2 < 3 = 2N-1$.

Let $N>2$ and suppose that $\mathcal{R}\cup\mathcal{B}$ is not contained in any line, yet $Z_{d,N}(\mathcal{R},\mathcal{B}) = 2N-1$. Since the dataset is finite, we can choose a line $r\subset\mathbb{R}^d$ with direction $v\in\mathbb{S}^{d-1}$ such that the orthogonal projection of $\mathcal{R}\cup\mathcal{B}$ onto r is injective. The number of points on r required to separate the projected sets is then $2N-1$. Indeed, it is at most $2N-1$ (as per (2.3.6) for $d=1$); on the other hand, these points correspond to a family of hyperplanes orthogonal to r that separate $(\mathcal{R},\mathcal{B})$, so it must be at least $Z_{d,N}(\mathcal{R},\mathcal{B}) = 2N-1$.

Therefore, we can choose hyperplanes $\mathsf{H}_1,\dots,\mathsf{H}_{2N-1}$ orthogonal to r that separate $(\mathcal{R},\mathcal{B})$, and moreover, the projected points $\{v\cdot x_i\}_{i=1}^{2N}\subset\mathbb{R}$ must be alternating as in (2.3.7). Without loss of generality, we may assume

$$v\cdot x_1 < h_1 < v\cdot x_{N+1} < \cdots < v\cdot x_N < h_{2N-1} < v\cdot x_{2N},$$

where $h_i\in\mathbb{R}$ are the scalars defining the orthogonal hyperplanes $\mathsf{H}_i = \{x\in\mathbb{R}^d \mid v\cdot x = h_i\}$, as illustrated in Figure 2.4 (left). We then define $\mathcal{R}' = \mathcal{R}\setminus\{x_1\}$ and $\mathcal{B}' = \mathcal{B}\setminus\{x_{N+1}\}$, each containing $N-1$ points.

Now, we show that $Z_{d,N-1}(\mathcal{R}',\mathcal{B}') = 2N-3$. Suppose that k hyperplanes $\mathsf{G}_1,\dots,\mathsf{G}_k$ in $\mathbb{R}^d$ suffice to separate $(\mathcal{R}',\mathcal{B}')$ for some $1\leqslant k\leqslant 2N-4$. Then the pair $(\mathcal{R},\mathcal{B})$ can be separated using $k+2$ hyperplanes $\mathsf{H}_1,\mathsf{H}_2,\mathsf{G}_1,\dots,\mathsf{G}_k$, as illustrated in Figure 2.4 (right). Since $k+2\leqslant 2N-2<2N-1$, this contradicts the fact that $Z_{d,N}(\mathcal{R},\mathcal{B}) = 2N-1$.

Now, we show that $Z_{d,N-1}(\mathcal{R}',\mathcal{B}') = 2N-3$. Suppose that k hyperplanes $\mathsf{G}_1,\dots,\mathsf{G}_k$ in $\mathbb{R}^d$ suffice to separate $(\mathcal{R}',\mathcal{B}')$ for some $1\leqslant k\leqslant 2N-4$. Then the pair $(\mathcal{R},\mathcal{B})$ can be separated using $k+2$ hyperplanes $\mathsf{H}_1,\mathsf{H}_2,\mathsf{G}_1,\dots,\mathsf{G}_k$, as illustrated in Figure 2.4 (right). Since $k+2\leqslant 2N-2<2N-1$, this contradicts the fact that $Z_{d,N}(\mathcal{R},\mathcal{B}) = 2N-1$.

The remainder of the proof consists of verifying that $(\mathcal{R},\mathcal{B})$ can be separated by at most $2N-2$ hyperplanes, which would lead to a contradiction. Since $Z_{d,N-1}(\mathcal{R}',\mathcal{B}') = 2N-3$, the inductive hypothesis on $N-1$ ensures that $\mathcal{R}'\cup\mathcal{B}'$ must be collinear along some line $r'\subset\mathbb{R}^d$. By assumption, the points in $\mathcal{R}\cup\mathcal{B}$ are not collinear, which implies that $x_1\notin r'$ or $x_{N+1}\notin r'$. Let $\mathsf{G}_1,\dots,\mathsf{G}_{2N-3}$ be hyperplanes that separate $(\mathcal{R}',\mathcal{B}')$, each intersecting r' transversely; that is, we can set $g_j = \mathsf{G}_j\cap r'\in\mathbb{R}^d$. We distinguish two cases:

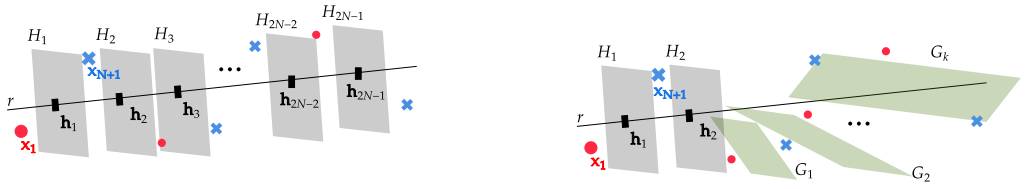


Figure 2.4: Figures supporting the argument presented in the proof of Theorem 2.2.4.

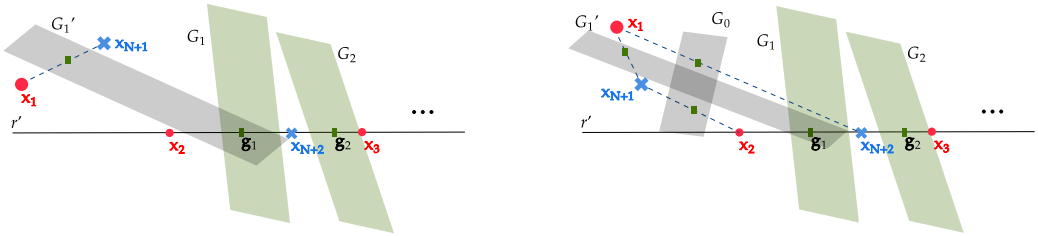


Figure 2.5: Figures supporting the argument presented in the proof of Theorem 2.2.4.

1. If x_1, x_{N+1}, and g_1 are not collinear, we can consider any hyperplane G_1' that intersects the open segment $(x_1, x_{N+1}) \coloneqq \{tx_1 + (1-t)x_{N+1} \mid 0 < t < 1\}$ transversely, and which also intersects the line r' transversely at g_1. If x_1 is on the same side of G_1' as x_2, the family $\{\mathsf{G}_1', \mathsf{G}_2, \ldots, \mathsf{G}_{2N-3}\}$ separates $(\mathcal{R}, \mathcal{B})$—as illustrated in Figure 2.5 (left)—implying $Z_{d,N}(\mathcal{R}, \mathcal{B}) \leqslant 2N-3$, which is a contradiction. Otherwise, we also consider any hyperplane G_0 that intersects both the open segment (x_2, x_{N+1}) and the open segment (x_1, x_{N+2}) transversely. Then, $\{\mathsf{G}_0, \mathsf{G}_1', \mathsf{G}_2, \ldots, \mathsf{G}_{2N-3}\}$ separates $(\mathcal{R}, \mathcal{B})$, as represented in Figure 2.5 (right), yielding $Z_{d,N}(\mathcal{R}, \mathcal{B}) \leqslant 2N-2$, which is also a contradiction.

2. If x_1, x_{N+1}, and g_1 are collinear, we can slightly translate G_1 and perturb its intersection with r' from g_1 to another point g_1' within the open segment (x_2, x_{N+2}). By doing so, we ensure that x_1, x_{N+1}, and g_1' are no longer collinear, thus reducing this to Case 1.

□

To obtain the probability distribution of $Z_{d,N}^{\perp}$, we first solve the case $d = 1$ employing combinatorial techniques:

Lemma 2.3.1. *For $N \geqslant 1$, let $Z_{1,N}$ be defined as in (2.3.2). For any $1 \leqslant k \leqslant 2N-1$, we have*

$$\mathbb{P}(Z_{1,N} = k) = \begin{cases} 2\binom{N-1}{p-1}^2 \binom{2N}{N}^{-1} & \text{if } k = 2p-1, \\ 2\binom{N-1}{p}\binom{N-1}{p-1}\binom{2N}{N}^{-1} & \text{if } k = 2p. \end{cases} \tag{2.3.8}$$

Remark 2.3.2. The probability mass function in (2.3.8) can be expressed in terms of the hyperge-

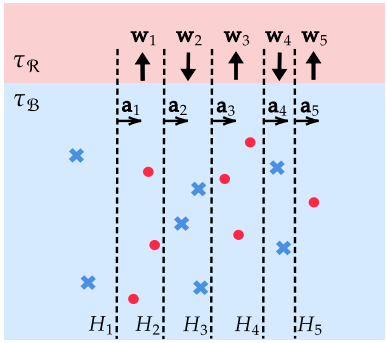


Figure 2.6: Schematic overview of the full control method used to prove Theorem 2.2.2. Parallel hyperplanes H_k separate the two classes into point clusters, which then move vertically in zigzag until all points are classified.

ometric distribution, whose mass function is defined by

$$H(k; M, K, n) = \frac{\binom{K}{k}\binom{M-K}{n-k}}{\binom{M}{n}},$$

for all $0 \leqslant M$, $0 \leqslant K, n \leqslant M$, and $(-M + K + n)_+ \leqslant k \leqslant K \wedge n$. Thus, for each $1 \leqslant k \leqslant 2N - 1$ we deduce

$$\mathbb{P}(Z_{1,N} = k) = \begin{cases} N(2N-1)^{-1} H(p-1; 2N-2, N-1, N-1), & \text{if } k = 2p-1, \\ (N-1)(2N-1)^{-1} H(p; 2N-2, N-1, N), & \text{if } k = 2p. \end{cases}$$

As a consequence of Lemma 2.3.1, we obtain the distribution of $Z^{\perp}_{d,N}$ for any $d \geqslant 1$:

Corollary 2.3.3. *Let $Z^{\perp}_{d,N}$ be defined as in (2.3.5), for some $d, N \geqslant 1$. For $1 \leqslant k \leqslant 2N - 1$, we have*

$$\mathbb{P}(Z^{\perp}_{d,N} \geqslant k) = \left(\sum_{p=\lceil \frac{k+1}{2} \rceil}^{N} \binom{N-1}{p-1}^2 + \sum_{p=\lceil \frac{k}{2} \rceil}^{N-1} \binom{N-1}{p} \binom{N-1}{p-1} \right)^d 2^d \binom{2N}{N}^{-d}. \tag{2.3.9}$$

The results derived in this section are now used to prove Theorem 2.2.2. For clarity, the entire control method is represented in Figure 2.6 and formalized in Algorithm 1.

Proof of Theorem 2.2.2. By the definition in (2.3.5), there exists a family of $z = Z^{\perp}_{d,N}(\mathcal{R}, \mathcal{B})$ hyperplanes, all orthogonal to e_ℓ for some $\ell \in [d]$, that separates $(\mathcal{R}, \mathcal{B})$. Moreover, $Z^{\perp}_{d,N}(\mathcal{R}, \mathcal{B})$ follows the distribution given by (2.3.9).

Without loss of generality, we assume $\ell = 1$. The separating family is then given by the equations $x^{(1)} = h_k$ for $k \in [z]$, with

$$0 < h_1 < \cdots < h_z < 1.$$

Set $h_0 := 0$ and $h_{z+1} := 1$. Fix $\ell_\star \in [d] \setminus \{1\}$ and define $\tau_{\mathcal{R}} = \{x^{(\ell_\star)} > 1\}$ and $\tau_{\mathcal{B}} = \{x^{(\ell_\star)} \leqslant 1\}$. We classify the points by clusters according to their $x^{(1)}$-coordinates in ascending order. We may

Algorithm 1 Classification of two N-point sets

Require: Two N-point sets $\mathcal{R}, \mathcal{B} \subset [0,1]^d$
Ensure: Classification of $\mathcal{R}$ and $\mathcal{B}$

Choose $\ell_{\mathrm{sep}} \in \arg\min_{\ell \in [d]} Z^{\ell}_{d,N}(\mathcal{R}, \mathcal{B})$ ▷ Axis achieving canonical separability
Fix any $\ell_\star \in [d] \setminus \{\ell_{\mathrm{sep}}\}$ ▷ Decision-coordinate index for the targets
$u \leftarrow e_{\ell_{\mathrm{sep}}}$
$\tau_{\mathcal{R}} \leftarrow \{x^{(\ell_\star)} > 1\}, \quad \tau_{\mathcal{B}} \leftarrow \{x^{(\ell_\star)} \le 1\}$
$\big((x_j, y_j)\big)_{j=1}^{2N} \leftarrow \mathrm{Reorder}(\mathcal{R} \cup \mathcal{B};\, \ell_{\mathrm{sep}})$ ▷ Sort by ascending ℓ_{sep}-th coordinate
$\mathcal{S} \leftarrow \{x_1\}, \quad \mathrm{anchor} \leftarrow 0$
for $j = 2, \dots, 2N$ **do**
  if $y_{j-1} = y_j$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{x_j\}$
  else if $x_{j-1} \in \mathcal{B}$ **and** $x_j \in \mathcal{R}$ **then**
    while $\mathcal{S} \not\subset \tau_{\mathcal{B}}$ **do**
      $\mathcal{R} \cup \mathcal{B} \leftarrow \mathrm{neuralODE}_{T=1}\big(\mathcal{R} \cup \mathcal{B};\, w = -e_{\ell_\star},\, a = u,\, b = \mathrm{anchor}\big)$
    end while
    $\mathcal{S} \leftarrow \{x_j\}, \quad \mathrm{anchor} \leftarrow -u \cdot x_{j-1}$
  else if $x_{j-1} \in \mathcal{R}$ **and** $x_j \in \mathcal{B}$ **then**
    while $\mathcal{S} \not\subset \tau_{\mathcal{R}}$ **do**
      $\mathcal{R} \cup \mathcal{B} \leftarrow \mathrm{neuralODE}_{T=1}\big(\mathcal{R} \cup \mathcal{B};\, w = e_{\ell_\star},\, a = u,\, b = \mathrm{anchor}\big)$
    end while
    $\mathcal{S} \leftarrow \{x_j\}, \quad \mathrm{anchor} \leftarrow -u \cdot x_{j-1}$
  end if
end for

assume that $\{x \in \mathcal{R} \cup \mathcal{B} \mid x^{(1)} < h_1\} \subset \mathcal{B}$; if this is not the case, we simply swap the roles of $\mathcal{R}$ and $\mathcal{B}$, and consequently the definitions of $\tau_{\mathcal{R}}$ and $\tau_{\mathcal{B}}$.

Taking $t_0 = 0$, we build the piecewise-constant control

$$(w, a, b)(t) = \sum_{k=1}^{z} (w_k, a_k, b_k)\, \mathbf{1}_{[t_{k-1}, t_k)}(t),$$
$$\text{where} \quad (w_k, a_k, b_k) = \big((-1)^{k+1} e_{\ell_\star},\, e_1,\, -h_k\big). \tag{2.3.10}$$

Each time horizon $t_k \geqslant t_{k-1}$ is chosen so that every point $x \in \mathcal{R} \cup \mathcal{B}$ with $h_k < x^{(1)} < h_{k+1}$ is mapped by the k-th segment of the flow to its corresponding target region $\tau_{\mathcal{R}}$ or $\tau_{\mathcal{B}}$. This is possible because the dataset is finite. Moreover, these movements do not affect points that have already been classified, since those satisfy $x^{(1)} < h_k$.

The described method correctly classifies all points according to their labels, and the number of switches is $L = z - 1$. Consequently, L inherits the probability distribution of $Z^{\perp}_{d,N} - 1$. □

Remark 2.3.4. Theorem 2.2.2 admits an interpretation in terms of the total variation seminorm, defined by

$$|\theta|_{\mathrm{TV}(0,T)} = \sup_{P} \sum_{i=1}^{|P|} \|\theta(t_i) - \theta(t_{i-1})\|,$$

where the supremum is taken over all finite partitions $P = \{0 = t_0 < t_1 < \cdots < t_{|P|} = T\}$. In particular, for piecewise-constant functions, $|\cdot|_{\mathrm{TV}(0,T)}$ equals the sum of the magnitudes of its jump

discontinuities. Thus, the control θ given by (2.3.10) satisfies

$$|\theta|_{\mathrm{TV}(0,T)} \leqslant L \cdot \max_{1 \leqslant k \leqslant L} \sqrt{\|w_{k+1} - w_k\|^2 + |b_{k+1} - b_k|^2} \leqslant \sqrt{5}L.$$

From here, we can upper-bound the tail probability of $|\theta|_{\mathrm{TV}(0,T)}$ via

$$\mathbb{P}\left(|\theta|_{\mathrm{TV}(0,T)} \geqslant \lambda\right) \leqslant \mathbb{P}\left(L \geqslant \frac{\lambda}{\sqrt{5}}\right) \qquad \text{for all } \lambda > 0,$$

and then apply the exact distribution derived in (2.2.4).

Remark 2.3.5. If we allow $a(t)$ to be any vector in $\mathbb{R}^d$, then the constraint (2.3.1) can be removed in Theorem 2.2.2. This ensures the existence of T, $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$, and θ for *every* configuration, rather than just almost surely. To accomplish this, one simply chooses a new vector basis in which (2.3.1) strictly holds, noting that the combinatorial conclusion of Lemma 2.3.1 remains valid in the new coordinates.

2.4 Classification by separability in general position

This section is centered on Theorem 2.2.1. The proof proceeds in two steps: (i) under the general-position hypothesis, we construct a separating family of *pairwise parallel* hyperplanes that isolates clusters of size d (see in particular Proposition 2.4.1); (ii) we convert this geometric separation into an explicit piecewise-constant control for a neural ODE with a particular activation, and then implement the localization with the original ReLU model. Technical lemmas that are not essential to the main narrative are proved later in **Section 2.8**.

Let $(\mathcal{R}, \mathcal{B})$ be as in (2.2.1), and assume for now that $|\mathcal{R}| = |\mathcal{B}| = N$. We recall Definition 2.2.1 and the random variable $Z_{d,N}$, defined in (2.3.2), to introduce the following quantity:

$$M(d, N) \coloneqq \max\left\{ Z_{d,N}(\mathcal{R}, \mathcal{B}) \mid (\mathcal{R}, \mathcal{B}) \text{ as in (2.2.1) in general position, } |\mathcal{R}| = |\mathcal{B}| = N \right\}.$$

Observe that for $d = 1$, any set of points is naturally in GP, so the result from Theorem 2.2.4 applies directly. Consequently, we have $M(1, N) = 2N - 1$ for all N, as stated in (2.3.6). However, when $d > 1$, the situation changes significantly. Assuming GP eliminates the pathological configurations described in Theorem 2.2.4, thereby reducing the maximum possible value of $Z_{d,N}$.

For example, assume that $d = N = 2$. Here, we find $M(2, 2) = 2$, which is strictly less than $2 \cdot 2 - 1 = 3$. To show this, consider connecting the two red points with a line r. Since all points are in GP, none of the blue points can lie on r. There are two possible cases:

1. If the blue points are on the same side of r, we can separate $\mathcal{R}$ from $\mathcal{B}$ with one line r' parallel to r, as shown in Figure 2.7 (left).
2. If the blue points are on different sides of r, we can separate $\mathcal{R}$ from $\mathcal{B}$ with two lines r' and r'' parallel to r, as shown in Figure 2.7 (right).

In the example of Figure 2.7 (left), if the separating lines were restricted to be perpendicular to the canonical axes, two lines would be needed to separate $(\mathcal{R}, \mathcal{B})$. Moreover, if the lines also had to be parallel—as in the framework of **Section 2.3**—then three lines would be required.

This geometric argument for the simple case $(d, N) = (2, 2)$ can be extended to derive a general bound for $M(d, N)$ that formalizes the improvement over Theorem 2.2.4 when the points are in GP. To this end, we cover $\mathcal{R}$ with (possibly overlapping) subsets of size d, and then separate $(\mathcal{R}, \mathcal{B})$ by isolating these subsets using hyperplanes under a transversality condition.

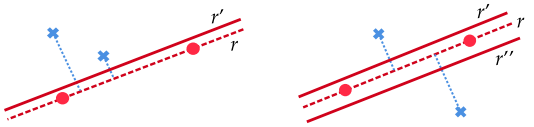


Figure 2.7: Separation of $(\mathcal{R}, \mathcal{B})$ in GP in $\mathbb{R}^2$ with $|\mathcal{R}| = |\mathcal{B}| = 2$ using at most two lines r', r''.

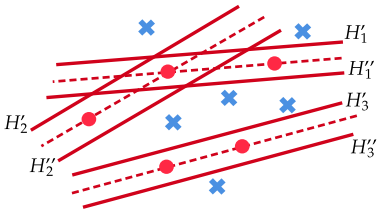


Figure 2.8: Hyperplanes constructed following the proof of Proposition 2.4.1, separating two unbalanced classes.

Proposition 2.4.1. *Let $d, N \geqslant 1$, fix $\ell_\star \in [d]$, and let $(\mathcal{R}, \mathcal{B})$ be as in (2.2.1) in general position, with $|\mathcal{R}| = |\mathcal{B}| = N$. Then, there exist hyperplanes $\mathsf{H}_1', \mathsf{H}_1'', \ldots, \mathsf{H}_{\lceil N/d \rceil}', \mathsf{H}_{\lceil N/d \rceil}''$ that separate $(\mathcal{R}, \mathcal{B})$. Moreover, for all $1 \leqslant j \leqslant \lceil N/d \rceil$, the following hold:*

1. *H_j' and H_j'' are parallel.*
2. *The region bounded by H_j' and H_j'' contains exactly $d \wedge |\mathcal{R}|$ points of $\mathcal{R}$ and no points of $\mathcal{B}$.*
3. *H_j' and H_j'' are not orthogonal to $e_{\ell_\star}$.*

In particular, for all $d, N \geqslant 1$, it follows that

$$M(d, N) \leqslant 2 \left\lceil \frac{N}{d} \right\rceil . \tag{2.4.1}$$

In the unbalanced case, where $|\mathcal{R}| \neq |\mathcal{B}|$, the method can be similarly applied to isolate the smaller set using pairwise parallel hyperplanes, as shown in Figure 2.8. This yields the following result:

Corollary 2.4.2. *Let $d \geqslant 1$. For any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1) in general position, there exists a family of $2 \left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil$ hyperplanes that separate the points by labels and satisfy the three conditions of Proposition 2.4.1, with $\mathcal{R}$ replaced by $\arg\min_{S \in \{\mathcal{R}, \mathcal{B}\}} |S|$.*

Corollary 2.4.2 enables the separation of any finite dataset in GP in $\mathbb{R}^d$ into clusters of size d. We now describe an inductive approach to classify these clusters using the dynamics of neural ODEs. First, let us consider the system

$$\dot{x}(t) = w(t)\, \sigma_{\mathrm{tr}}\big(a(t) \cdot x(t) + b(t)\big), \tag{$\mathcal{V}_1^{\mathrm{tr}}$}$$

where the activation function is defined as

$$\sigma_{\mathrm{tr}}(s) \coloneqq 1 \wedge s_+ = s_+ - (s-1)_+, \qquad \text{for } s \in \mathbb{R}. \tag{2.4.2}$$

Observe that the flow map associated with ($\mathcal{V}_1^{\mathrm{tr}}$) remains well-defined because σ_{tr} is Lipschitz continuous. The introduction of σ_{tr} is motivated by its properties, which facilitate the inductive argument. Specifically, in the half-space defined by $a(t) \cdot x + b(t) > 1$, the system ($\mathcal{V}_1^{\mathrm{tr}}$) simplifies to $\dot{x} = w(t)$. This property can be used to ensure that points already classified will remain so after each inductive step. For more details, see Figure 2.9, which provides geometric intuition for the proof, and the formal control method presented in Algorithm 2.

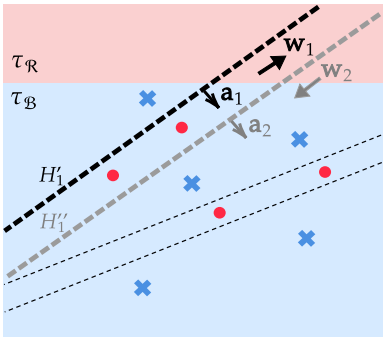


Figure 2.9: Schematic overview of an iteration in the control method used to prove Proposition 2.4.3. We aim to move the two red points that lie between the hyperplanes H_1' and H_1'' to the region $\tau_{\mathcal{R}}$ through a two-step process. First, we move all points within the half-space $\mathsf{H}_1'^{+}$, determined by a_1, in the direction of w_1. Second, we move all points within the half-space $\mathsf{H}_1''^{+}$, determined by a_2, in the direction of w_1. The controls are adjusted to ensure that the vector field in $\mathsf{H}_1''^{+}$ maintains a unit norm at all times. Consequently, the points in $\mathsf{H}_1''^{+}$ return exactly to their original positions after these two steps.

Proposition 2.4.3. *Let $d \geqslant 2$. For any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1) in general position, and any pair of target regions $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ as in (2.2.3), there exist $T > 0$ and a piecewise-constant control $\theta \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ with*

$$L = 2\left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil - 1 \qquad \textit{discontinuities,}$$

such that the flow map of $(\mathcal{V}_1^{\mathrm{tr}})$ satisfies $\Phi_T^\theta(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T^\theta(\mathcal{B}) \subset \tau_{\mathcal{B}}$.

The following lemma formalizes an idea from [RZ24, Section 2.2]. It shows that any flow of $(\mathcal{V}_1^{\mathrm{tr}})$ can be represented as the composition of two flows of $(\mathcal{V}_1^{+})$, provided w and a are orthogonal:

Lemma 2.4.4. *Let $\theta = (w, a, b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}$ be a constant control with $w \cdot a = 0$. Then, $\theta_1 = (w, a, b)$ and $\theta_2 = (-w, a, b - 1)$ satisfy*

$$\Phi_t\big(\Phi_t(\cdot\,; \theta_1)\,; \theta_2\big) = \Phi_t^{\mathrm{tr}}(\cdot\,; \theta) \qquad \textit{for all } t > 0,$$

where Φ_t and Φ_t^{tr} are the flow maps of $(\mathcal{V}_1^{+})$ and $(\mathcal{V}_1^{\mathrm{tr}})$, respectively.

We can now use Lemma 2.4.4, together with Proposition 2.4.3, to demonstrate Theorem 2.2.1.

Proof of Theorem 2.2.1. By virtue of Proposition 2.4.3, there exist $\bar{T} > 0$ and a piecewise-constant control $\bar{\theta} = (\bar{w}, \bar{a}, \bar{b})$ defined on $[0, \bar{T}]$, which presents $2\lceil(|\mathcal{R}| \wedge |\mathcal{B}|)/d\rceil - 1$ discontinuities, such that

$$\Phi_{\bar{T}}^{\mathrm{tr}}(\mathcal{R}; \bar{\theta}) \subset \tau_{\mathcal{R}} \qquad \text{and} \qquad \Phi_{\bar{T}}^{\mathrm{tr}}(\mathcal{B}; \bar{\theta}) \subset \tau_{\mathcal{B}},$$

where Φ_t^{tr} is the flow map of $(\mathcal{V}_1^{\mathrm{tr}})$. Moreover, $\bar{w}(t) \cdot \bar{a}(t) = 0$ for all $t \in [0, \bar{T}]$. Thus, by repeatedly applying Lemma 2.4.4 over each constant segment of $\bar{\theta}$, there exists a piecewise-constant control θ defined on $[0, 2\bar{T}]$ with $4\lceil(|\mathcal{R}| \wedge |\mathcal{B}|)/d\rceil - 1$ discontinuities, such that $\Phi_{2\bar{T}}(\cdot\,; \theta) = \Phi_{\bar{T}}^{\mathrm{tr}}(\cdot\,; \bar{\theta})$. We conclude by setting $T = 2\bar{T}$. □

Algorithm 2 Classification of two point sets via $(\mathcal{V}_1^{\mathrm{tr}})$

Require: Two finite sets $\mathcal{R}, \mathcal{B} \subset [0,1]^d$ in GP; Direction $e_{\ell_\star}$
Ensure: Classification of $\mathcal{R}$ and $\mathcal{B}$

$\mathcal{S} \leftarrow \arg\min_{S \in \{\mathcal{R}, \mathcal{B}\}} |S|, \quad \tau_{\mathcal{S}} \leftarrow \{x^{(\ell_\star)} > 1\}$
$\mathcal{C} \leftarrow \mathrm{cover}(\mathcal{S}, d, \lceil |\mathcal{S}|/d \rceil)$ ▷ Covering for $\mathcal{S}$ by $\lceil |\mathcal{S}|/d \rceil$ subsets $\mathcal{S}_j$ with $|\mathcal{S}_j| = d$
for $(j, \mathcal{S}_j)$ in enumerate$(\mathcal{C})$ **do**
  $\mathsf{H}_j(a_j, b_j) = \{x \mid a_j \cdot x + b_j = 0,\ \|a_j\| = 1\} \leftarrow \mathrm{Span}(\mathcal{S}_j)$
  if $|a_j \cdot e_{\ell_\star}| = 1$ **then**
    $a_j \leftarrow \mathrm{NormalizedPerturbation}(a_j)$ ▷ Achieves $|a_j \cdot e_{\ell_\star}| < 1,\ \|a_j\| = 1$
  end if
  $\delta_j \leftarrow \min\{\mathrm{dist}(x, \mathsf{H}_j) \mid x \in (\mathcal{R} \cup \mathcal{B}) \setminus \mathcal{S}_j\}$
  $b'_j \leftarrow b_j + 0.5\,\delta_j, \quad b''_j \leftarrow b_j - 0.5\,\delta_j$
  $w_j \leftarrow \mathrm{NormalizedProjection}_{\langle a_j \rangle^\perp}(e_{\ell_\star})$
  $d'_j = \min_{x \in \mathcal{S}_j} \sigma_{\mathrm{tr}}(a_j \cdot x + b'_j)$
  $d''_j = \min_{x \in (\mathcal{R} \cup \mathcal{B}) \cap \{x \mid a_j \cdot x + b''_j > 0\}} \sigma_{\mathrm{tr}}(a_j \cdot x + b''_j)$
  while $\mathcal{S}_j \not\subset \tau_{\mathcal{S}}$ **do**
    $\mathcal{R} \cup \mathcal{B} \leftarrow \mathrm{neuralODE}_{T=1}(\mathcal{R} \cup \mathcal{B};\ w = w_j,\ a = a_j/d'_j,\ b = b'_j/d'_j)$
    $\mathcal{R} \cup \mathcal{B} \leftarrow \mathrm{neuralODE}_{T=1}(\mathcal{R} \cup \mathcal{B};\ w = -w_j,\ a = a_j/d''_j,\ b = b''_j/d''_j)$
  end while
end for

Remark 2.4.5. As in Remark 2.3.4, we can extract from Theorem 2.2.1 a bound for the total variation of the controls as

$$|\theta|_{\mathrm{TV}(0,T)} \leqslant 2L\|\theta\|_\infty \leqslant 2\sqrt{1 + \|(a,b)\|_\infty^2} \left(4 \left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil - 1\right),$$

where $\|(a,b)\|_\infty$ depends on the minimum separation between points of $\mathcal{R} \cup \mathcal{B}$ and the maximum distance of any point to the origin.

The structure of pairwise parallel hyperplanes defined in Proposition 2.4.1 can be exploited for control when combined with certain architectures. To this end, we propose *triangular neural ODEs*

$$\dot{x}(t) = w(t)\,\sigma_\triangle\big(a(t) \cdot x(t) + b(t)\big), \tag{$\mathcal{V}_1^\triangle$}$$

where the activation $\sigma_\triangle$ is defined as

$$\sigma_\triangle(s) = (1 - |s - 1|)_+, \qquad \text{for } s \in \mathbb{R}. \tag{2.4.3}$$

The function $\sigma_\triangle$ remains globally Lipschitz, ensuring that the flow map is well-defined.

By replacing σ_{tr} with $\sigma_\triangle$, we can improve upon Proposition 2.4.3. Specifically, we can halve the number of values taken by the control θ. Adapting the proof is straightforward: Step 2 of the previous method can be omitted, as the support of $\sigma_\triangle$ is restricted to the interval $[0,2]$.

Corollary 2.4.6. *Let $d \geqslant 2$. For any dataset $(\mathcal{R}, \mathcal{B})$ as in (2.2.1) in general position, and any pair of target regions $(\tau_{\mathcal{R}}, \tau_{\mathcal{B}})$ as in (2.2.3), there exist $T > 0$ and a piecewise-constant control $\theta \in L^\infty(0, T; \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})$ with*

$$L = \left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil - 1 \qquad \textit{discontinuities,}$$

such that the flow map of $(\mathcal{V}_1^\triangle)$ satisfies $\Phi_T^\theta(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T^\theta(\mathcal{B}) \subset \tau_{\mathcal{B}}$.

Remark 2.4.7. Our separability-based methodology aligns with existing research in discrete geometry. In [BU95], the authors establish lower and upper bounds on the minimum number of hyperplanes required to individually separate N points in GP in $\mathbb{R}^d$. It is shown in [FMP91] that finding such a number for an arbitrary point set is an NP-complete problem. From a computational perspective, efficient algorithms for separability in low-dimensional spaces are proposed in [Ark+06; Hou93].

Remark 2.4.8. The VC dimension of a classification model is defined as the size of the largest set of points that can be shattered—meaning any arbitrary binary assignment of labels can be correctly classified [VC71]. We can interpret Theorem 2.2.1 in terms of the VC dimension of our model. Specifically, we can establish a lower bound: with $L \geqslant 1$ available discontinuities, our method can shatter any set of Ld points in $\mathbb{R}^d$.

2.5 Simulations

We present a computational test designed to evaluate the capacity of $(\mathcal{V}_1^+)$ for binary classification. The goal is to numerically estimate the minimal complexity it requires to classify an arbitrary dataset of a fixed size, and compare it with the complexity we established in Theorem 2.2.1. Unless otherwise specified, all numerical experiments in this work are implemented in PyTorch [PyT26; Pas+17]. For the numerical integration of the neural ODEs, we use the torchdiffeq library [Che21]. The code for the present experiment is available at [Álv24].

Let $\mathcal{D} = \{(x_i, y_i)\} \subset \mathbb{R}^d \times \{0, 1\}$ be a finite dataset and $\mathcal{R} = \{x_i \mid (x_i, 1) \in \mathcal{D}\}$, $\mathcal{B} = \{x_i \mid (x_i, 0) \in \mathcal{D}\}$ be such that $|\mathcal{R}| = |\mathcal{B}| = N$. For a fixed $L \geqslant 0$ and time $T > 0$, we aim to find a piecewise-constant control θ with L discontinuities such that the flow map of $(\mathcal{V}_1^+)$ satisfies

$$\Phi_T^\theta(\mathcal{R}) \subset \left\{x^{(1)} > 1\right\} \qquad \text{and} \qquad \Phi_T^\theta(\mathcal{B}) \subset \left\{x^{(1)} \leqslant 1\right\}. \tag{2.5.1}$$

We opt for σ_{tr} as in (2.4.2) and equation $(\mathcal{V}_1^{\text{tr}})$ to better visualize the results (see Figure 2.10). In this case, Proposition 2.4.3 establishes that $L = 2\lceil N/d \rceil - 1$ switches ensure any dataset in general position can be classified. This condition is generically satisfied when all points are sampled from a nonsingular probability measure, such as the uniform measure $\text{Unif}([0, 1]^d)$.

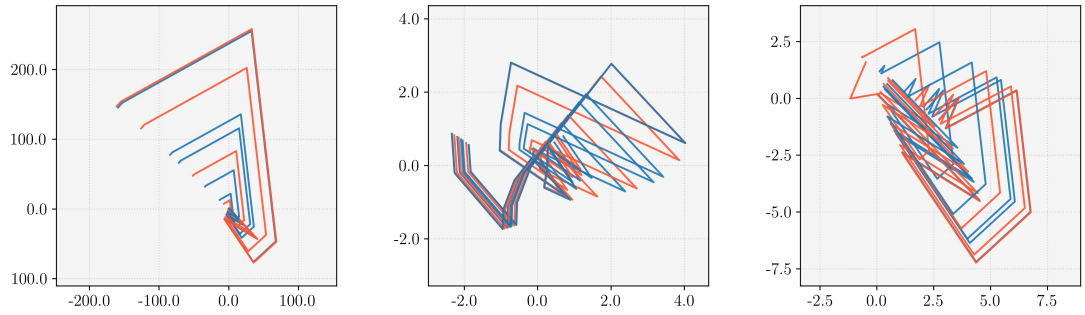


Figure 2.10: Trajectories for $\sigma = \text{ReLU}$ exhibit an exponential drift when $a \cdot x + b > 1$ (**left**). If $\sigma = \tanh$ (**middle**) or $\sigma = \sigma_{\text{tr}}$ (**right**), this drift is mitigated because $\|\sigma\|_\infty \leqslant 1$. Here, we have used $N = 5$, $L = 10$, and $T = 60$.

We design our experimental setup in the following way:

Data. We use ten random datasets, each consisting of $N = 30$ red points ($\mathcal{R}$) and $N = 30$ blue points ($\mathcal{B}$), sampled from $\mathrm{Unif}([0,1]^d)$ for various dimensions d ranging from 2 to $2N$.

Model. We consider equation ($\mathcal{V}_1^{\mathrm{tr}}$) with controls that are piecewise constant over $L+1$ time intervals of length $\Delta t = T(L+1)^{-1}$, where $T = 60$. To approximate the solution, we use an explicit Euler discretization scheme with a step size of $0.25\Delta t$. The control values are initialized using Kaiming uniform initialization; that is, each component is sampled from $\mathrm{Unif}([-\alpha, \alpha])$, where α is the inverse of the square root of the number of input units in the weight tensor.

Error. To enforce the correct classification of all points as defined by (2.5.1), we introduce the following margin-based loss function:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{R}|} \sum_{x_i \in \mathcal{R}} \mathrm{dist}\big(\Phi_T^\theta(x_i), \{x^{(1)} < 1.5\}\big)^2 + \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \mathrm{dist}\big(\Phi_T^\theta(x_i), \{x^{(1)} > 0.5\}\big)^2.$$

Training. Optimization is conducted using the Adam optimizer [KB15] with a learning rate of 0.01, which was determined to be optimal through a grid search over a range of values.

Procedure. First, for each d, we verify that setting $L = 2\lceil N/d \rceil - 1$ switches is sufficient to classify all ten datasets. Then, we gradually decrease the value of L until at least one of the datasets fails to be classified, according to the stopping criteria defined below.

Stopping criteria. If condition (2.5.1) is met, classification is successful and the training is halted. Conversely, we consider the following three failure criteria:

1. The maximum number of 70,000 epochs is reached.
2. *Slow convergence:* $\mathcal{L} \geqslant 0.15$ at 20,000 epochs, or $\mathcal{L} \geqslant 0.1$ at 40,000 epochs, or the minimum error does not decrease over 5,000 consecutive epochs.
3. *Local minima detection:* the maximum relative error over 50 consecutive epochs exceeds a threshold of 10^{-20}.

Evaluation. We conclude that a model with L switches lacks the practical capacity to classify the dataset if all 20 randomized initializations of the parameters fail (i.e., trigger a failure criterion) on any of the ten datasets.

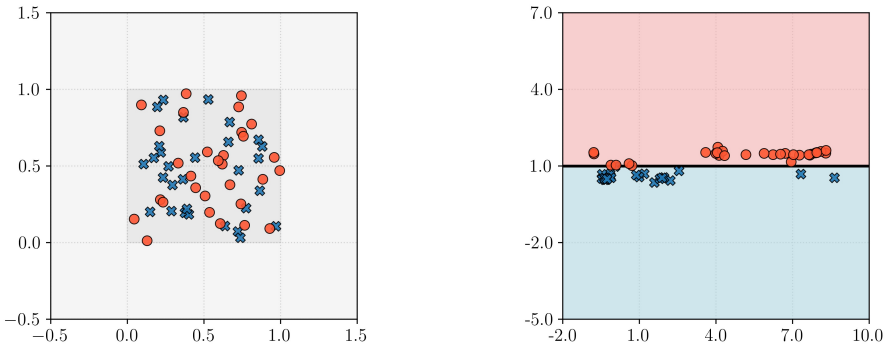


Figure 2.11: **Left:** Initial data comprising 30 red points and 30 blue points in $[0,1]^2$. **Right:** Final positions at time $T = 60$, having fixed $L = 37$ switches.

The experiments were conducted using specific seed settings to ensure reproducibility, and the results are shown in Figure 2.12. We observe that the theoretical bound $L = 2\lceil N/d \rceil - 1$ is typically very close to the optimal practical value L_* obtained through training. Specifically, for $d \in \{7, 10, 15, 20, 60\}$, we get exactly $L_* = 2\lceil N/d \rceil - 1$. Interestingly, for $d \in \{2, 30\}$, we find that $L_* > 2\lceil N/d \rceil - 1$. This indicates that while a separating control with fewer switches theoretically exists (Proposition 2.4.3), gradient descent struggles to find it due to the highly non-convex optimization landscape.

We also observed that training requires an increasing number of parameter initializations in lower dimensions. This tendency is particularly pronounced when $d = 2$, as shown by the purple bar with a hatched pattern in Figure 2.12. In that case, only five out of the ten datasets were successfully classified using $L = 37$ switches, and we were unable to find any higher value of L capable of classifying the remaining five datasets under pure gradient descent.

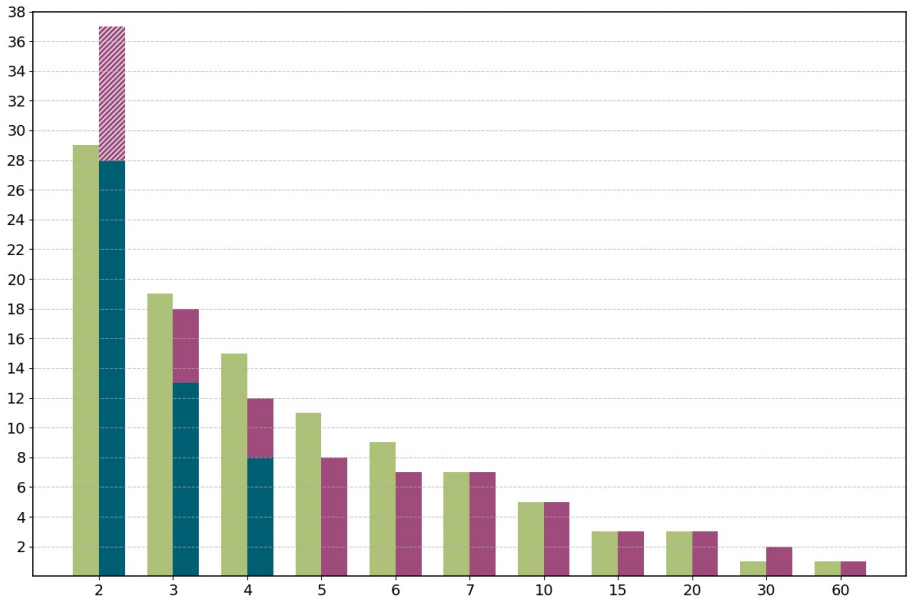


Figure 2.12: L versus d for fixed $N = 30$. **Green bars** represent the theoretical number of switches from Proposition 2.4.3, given by $2\lceil N/d \rceil - 1$. **Purple bars** show the minimum number of switches required by gradient-based training to successfully classify all datasets. **Blue bars** indicate the minimum number of switches found after data standardization.

Motivated by these optimization difficulties in low dimensions, we applied a preprocessing step to improve the separability of the data points. Specifically, for $d \in \{2, 3, 4\}$, we standardized the data by subtracting the mean and scaling to unit variance. This reduces the risk of the algorithm becoming trapped in poor local minima due to an overly dense point cloud. The improved results are represented by the blue bars in Figure 2.12.

In summary, we observed that the formula $L = 2\lceil N/d \rceil - 1$ given in Proposition 2.4.3 is a remarkably tight theoretical bound that closely approximates—and sometimes underestimates, due to optimization hardness—the practical complexity required by gradient descent, especially for large d with fixed N.

2.6 Discussion

We conclude with a discussion on connections and extensions of our results.

- **Multiclass classification.** While our results are stated for two classes, the control strategy extends naturally to $S > 2$ classes $\mathcal{R}_1, \dots, \mathcal{R}_S$. Assuming $\bigcup_{s=1}^{S} \mathcal{R}_s$ is in general position, one can apply a *one-vs-rest* reduction: first, use Proposition 2.4.3 to separate the smallest class from the others. Since our binary constructions can be arranged to keep the larger set fixed while moving only the smaller one, previously separated classes can be left unchanged in subsequent steps. This inductive process guarantees that, ultimately, every pair of distinct classes becomes linearly separable under Φ_T^θ.

- **Alternative activation functions.** Our probabilistic result (Theorem 2.2.2) relies only on the existence of a nontrivial interval on which σ vanishes, together with a region where transport is active. For such activations—often referred to as *well functions* [LLS22]—the specific choice of σ mainly affects the time scaling of each step, not the switching complexity.

  By contrast, the general-position construction (Theorem 2.2.1) relies on implementing localized transport via truncated activations and their representation through ReLU flows (cf. Lemma 2.4.4). Extending Theorem 2.2.1 therefore requires a substitute mechanism for local motion. Activations with compact support (such as $\sigma_\triangle$, used in Corollary 2.4.6) are natural candidates and may allow tighter local control and improved constants.

- **Classification vs. Separability.** A subtle but crucial point is the distinction between our dynamic classification and classical static separability. One might ask if finding a neural ODE control is equivalent to finding a sequence of separating hyperplanes. The answer is negative.

  First, the existence of separating hyperplanes does not automatically imply the existence of a valid control $w(t)$ for $(\mathcal{V}_1^+)$ that implements the transport; our constructions require specific arrangements (e.g., pairwise parallelism in Theorem 2.2.1) to ensure kinematic compatibility. Thus, for a random family of separating hyperplanes, finding such a $w(t)$ can be highly complex.

  Conversely, and perhaps more surprisingly, dynamic classification is often more powerful than static linear separation. As illustrated in Figure 2.13, $(\mathcal{V}_1^+)$ can successfully classify a dataset even if the instantaneous hyperplane $\{a(t) \cdot x + b(t) = 0\}$ does not fully separate the classes. The flow accumulates local deformations that eventually disentangle the data—a capability absent in static linear models.

  Our results address these questions by proving that, in the worst-case scenario (i.e., for any arbitrary dataset), neural ODEs can perform at least as well as linear classifiers. Furthermore, our numerical experiments confirm that, under such worst-case conditions, the practical performance of $(\mathcal{V}_1^+)$ trained via gradient descent closely matches our theoretical findings.

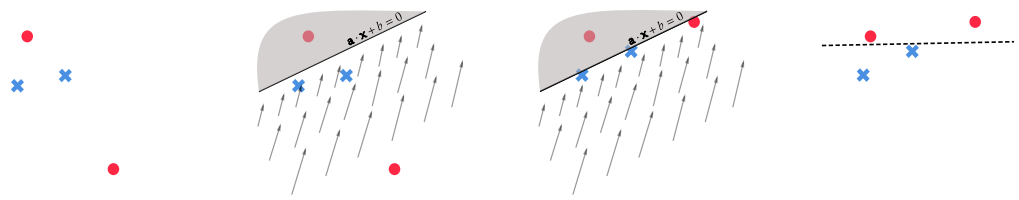


Figure 2.13: In the first picture, at least two hyperplanes are required to separate both classes (*static* classification). In the next three pictures, the same points evolve by $(\mathcal{V}_1^+)$ (with constant θ) and can eventually be separated using only one hyperplane (*dynamic* classification).

- **Connection to cross-entropy.** Our theorems guarantee separation into disjoint terminal regions, hence zero empirical classification error for the induced hard decision rule. In practice, training is often performed with *cross-entropy*, which rewards not just separation but also large margins. In the binary case, if $\hat{y}_i \in (0,1)$ denotes the predicted probability that $y_i = 1$ (e.g., $\hat{y}_i = (\mathrm{softmax} \circ P \circ \Phi_T(x_i))^{(1)}$), then the cross-entropy loss is

$$\mathcal{L}_{\mathrm{CE}}(x_i, y_i) = -y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i).$$

  Once a constructed flow places the training points strictly inside the correct open half-spaces, one can drive $\mathcal{L}_{\mathrm{CE}}$ arbitrarily small (on the training set) by amplifying the final margin—for instance, by scaling a final linear readout or by extending the time horizon—without increasing the number of switches.

- **Comparison with discrete architectures.** Our complexity bounds parallel developments in discrete deep learning. Early constructive results, which focused on sigmoid activation functions, showed that networks with one hidden layer could perform this task using $O(N)$ neurons [HH91]. With the rise of deep learning, researchers explored how adding layers could reduce the required complexity [ES15; Hua03], and existing results were extended to ReLU networks, as seen in [Zha+16]. For deep feed-forward networks, [YSJ19] solved the problem using $O(\sqrt{N})$ neurons, and [HZ25] develops a constructive methodology similar to ours.

  In the past decade, the focus has shifted toward modern architectures such as convolutional networks [NH18] or ResNets [HM17], which achieve the classification goal using $O(N \log N)$ hidden nodes, assuming a minimum distance between points. This bound was later improved to $O(N/d)$ in [YSJ19] under the assumption of general position. Notably, for conditional networks, the complexity can be further reduced to $O(\log N)$ neurons [Koy23]. Our continuous-time result $L \approx 4N/d$ provides a natural continuum counterpart to these findings, mathematically confirming that the efficiency gain from dimension d is a fundamental geometric property preserved in the continuous limit.

2.7 Perspectives

Several important directions remain open for future exploration. Below, we highlight a few of them:

1. **Sharpness of the bound.** A natural first question is whether the new bound we established for the minimal number of switches L required to classify any dataset in general position is sharp, or if it can be further improved.

   On the one hand, equality in (2.4.1) is achieved in simple cases, such as $d = N = 2$, as illustrated in Figure 2.7. However, our constructive algorithm does not fully exploit the entire capacity of the parameter space, as the piecewise-constant controls $a(t)$ and $w(t)$ repeat certain values across steps. Moreover, the algorithm is strictly designed to iteratively classify clusters of d points and does not attempt to isolate larger clusters, unlike Algorithm 2.

   If our theoretical bound for L is indeed sharp, a related open problem is to geometrically characterize the pathological configurations that attain this maximum, analogously to Theorem 2.2.4.

2. **Decision boundary.** For transparency, we fixed an axis-aligned decision boundary of the form $\{x^{(\ell_\star)} = 1\}$. Our constructions can be adapted to general affine hyperplanes, but this raises a genuine optimization question: if one is allowed to freely choose (or learn) the readout and the terminal boundary, what is the minimal achievable switching complexity? Furthermore, what geometric features of $(\mathcal{R}, \mathcal{B})$ determine an optimal boundary under simplicity constraints?

3. **Combination with self-attention.** Motivated by Transformer architectures and the continuous-depth perspective outlined in (10), one may consider augmenting the neural ODE with a self-attention interaction term, leading to an interacting-particle system such as

$$\dot{x}_i(t) = w(t)\sigma\big(a(t)\cdot x_i(t) + b(t)\big) + \sum_{j=1}^{N} \operatorname{softmax}\big(\beta\, x_i(t)^\top B(t)\, x_j(t)\big)\, V(t)\, x_j(t), \qquad i \in [N], \tag{2.7.1}$$

with additional controls $B(t), V(t) \in \mathscr{M}_{d\times d}(\mathbb{R})$ and a parameter $\beta > 0$, known as the inverse temperature. The attention component may induce clustering [Ges+23], potentially reducing the separation complexity required of the transport term. A natural open problem is to quantify, and constructively achieve, such reductions, in the spirit of [GRR24], or [AFZ25b] in the discrete hardmax regime (when $\beta \to \infty$).

2.8 Proofs

Proofs of Section 2.3

Proof of Lemma 2.3.1. Since all points in $\mathcal{R}\cup\mathcal{B}$ are i.i.d., each possible ordering is equally likely. Therefore, to determine $\mathbb{P}(Z_{1,N} = k)$, we count the fraction of orderings that have exactly k gaps between consecutive subsets of points with different labels. We compute the number of such favorable configurations by dividing them into two cases based on the parity of k:

Case 1: $k = 2p-1$, for $1 \leqslant p \leqslant N$. Any ordering of the points in $\mathcal{R}\cup\mathcal{B}$ that leads to $Z_{1,N}(\mathcal{R},\mathcal{B}) = 2p-1$ can be represented as the union of two partitions of $\mathcal{R}$ and $\mathcal{B}$, respectively constituted of subsets $R_j \subset \mathcal{R}$ and $B_j \subset \mathcal{B}$, for $j \in [p]$, with lengths $|R_j| = r_j$ and $|B_j| = b_j$, that satisfy

$$\sum_{j=1}^{p} r_j = \sum_{j=1}^{p} b_j = N. \tag{2.8.1}$$

There are $\binom{N-1}{p-1}$ possible partitions $\{R_1,\dots,R_p\}$ of $\mathcal{R}$, determined by the choice of $p-1$ gaps between consecutive elements of $\mathcal{R}$, among the total $N-1$ possibilities. Analogously, there are $\binom{N-1}{p-1}$ possible partitions $\{B_1,\dots,B_p\}$ of $\mathcal{B}$, to insert in the p gaps and therefore obtain $Z_{1,N}(\mathcal{R},\mathcal{B}) = k$. So there are $\binom{N-1}{p-1}^2$ possible configurations. On the other hand, we can repeat the argument but now considering the $\binom{N-1}{p-1}$ possible partitions of $\mathcal{B}$ and then inserting R_j into the p gaps. Both situations are symmetric; the only difference between them lies in whether the smallest point of $\mathcal{R}\cup\mathcal{B}$ (and thus the first subset of the two partitions) belongs to $\mathcal{R}$ or $\mathcal{B}$, as shown in Figure 2.14. Consequently, the total number of configurations for $(\mathcal{R},\mathcal{B})$ that yield $Z_{1,N}(\mathcal{R},\mathcal{B}) = 2p-1$ is $2\binom{N-1}{p-1}^2$.

Case 2: $k = 2p$, for $1 \leqslant p \leqslant N-1$. To obtain $Z_{1,N}(\mathcal{R},\mathcal{B}) = 2p$, we need an even number of gaps between subsets of the same color, so now we must consider the partitions of $\mathcal{R}$ and $\mathcal{B}$ that are respectively constituted of subsets $R_j, R_{p+1} \subset \mathcal{R}$ and $B_j \subset \mathcal{B}$ for $j \in [p]$ (or the reverse situation) with lengths $|R_j| = r_j$ and $|B_j| = b_j$, that satisfy

$$\sum_{j=1}^{p+1} r_j = \sum_{j=1}^{p} b_j = N. \tag{2.8.2}$$

Like in case 1, there exist $\binom{N-1}{p}$ possible partitions $\{R_1,\dots,R_{p+1}\}$ of $\mathcal{R}$ and $\binom{N-1}{p-1}$ possible partitions $\{B_1,\dots,B_p\}$ of $\mathcal{B}$. So there are $\binom{N-1}{p}\binom{N-1}{p-1}$ possibilities, each one depending on a choice



Figure 2.14: Examples for case 1 in the proof of Lemma 2.3.1.



Figure 2.15: Examples for case 2 in the proof of Lemma 2.3.1.

of $(r_1, \dots, r_{p+1}) \in \mathbb{N}^{p+1}$ and $(b_1, \dots, b_p) \in \mathbb{N}^p$ satisfying (2.8.2). We take also into account the symmetric situation, when the partitions of $\mathcal{R}$ and $\mathcal{B}$ are respectively of the form $\{R_1, \dots, R_p\}$ and $\{B_1, \dots, B_{p+1}\}$, shown in Figure 2.15. Therefore, it follows that the number of configurations for $(\mathcal{R}, \mathcal{B})$ that yield $Z_{1,N}(\mathcal{R}, \mathcal{B}) = 2p$ is $2\binom{N-1}{p}\binom{N-1}{p-1}$.

Finally, the total number of possibilities is exactly the number of ways to choose N points out of $2N$, which is given by the central binomial coefficient $\binom{2N}{N}$. Indeed,

$$2\sum_{p=1}^{N}\binom{N-1}{p-1}^2 + 2\sum_{p=1}^{N-1}\binom{N-1}{p}\binom{N-1}{p-1} = \binom{2N}{N}.$$

□

Proof of Corollary 2.3.3. Let $k \in [2N-1]$. By definition of $Z^{\perp}_{d,N}$, and the fact that $Z^{\ell}_{d,N}$ are independent and identically distributed to $Z_{1,N}$ for all ℓ, we can compute

$$\mathbb{P}\left(Z^{\perp}_{d,N} \geqslant k\right) = \mathbb{P}\left(\min_{1\leqslant \ell \leqslant d} Z^{\ell}_{d,N} \geqslant k\right) = \left(\mathbb{P}\left(Z_{1,N} \geqslant k\right)\right)^d.$$

We conclude the proof by applying Lemma 2.3.1 to deduce

$$\mathbb{P}\left(Z_{1,N} \geq k\right) = \left(\sum_{p=\lceil \frac{k+1}{2} \rceil}^{N}\binom{N-1}{p-1}^2 + \sum_{p=\lceil \frac{k}{2} \rceil}^{N-1}\binom{N-1}{p}\binom{N-1}{p-1}\right) 2\binom{2N}{N}^{-1}.$$

□

Proofs of Section 2.4

Proof of Proposition 2.4.1. If $d \leqslant N$, we can choose $\hat{R}_1, \dots, \hat{R}_{\lceil N/d \rceil} \subset \mathcal{R}$ such that $\mathcal{R} = \hat{R}_1 \cup \dots \cup \hat{R}_{\lceil N/d \rceil}$, with $|\hat{R}_j| = d$ for all j. Otherwise, build $\hat{R}_1$ by augmenting the set $\mathcal{R}$ with $d - N$ points chosen from $\mathbb{R}^d \setminus \mathcal{B}$ so that $\hat{R}_1 \cup \mathcal{B}$ remains in GP. Since $\mathcal{R} \cup \mathcal{B}$ is in GP, each $\hat{R}_j$ spans a unique hyperplane $\hat{\mathsf{H}}_j \subset \mathbb{R}^d$ satisfying

$$(\mathcal{R} \cup \mathcal{B} \setminus \hat{R}_j) \cap \hat{\mathsf{H}}_j = \emptyset \qquad \text{and} \qquad \hat{\mathsf{H}}_j \neq \hat{\mathsf{H}}_k \quad \text{for all } j \neq k. \tag{2.8.3}$$

For each j, choose two hyperplanes $\hat{\mathsf{H}}'_j$ and $\hat{\mathsf{H}}''_j$ that are parallel to $\hat{\mathsf{H}}_j$ and lie in the two distinct connected components of $\mathbb{R}^d \setminus \hat{\mathsf{H}}_j$. If these hyperplanes are chosen sufficiently close to $\hat{\mathsf{H}}_j$, then the region between $\hat{\mathsf{H}}'_j$ and $\hat{\mathsf{H}}''_j$ contains the subset $\hat{R}_j$ and no other points of $\mathcal{R} \cup \mathcal{B}$. Consequently, $(\mathcal{R}, \mathcal{B})$ is separated by the family of pairwise parallel hyperplanes

$$\hat{\mathsf{H}}'_1, \hat{\mathsf{H}}''_1, \dots, \hat{\mathsf{H}}'_{\lceil N/d \rceil}, \hat{\mathsf{H}}''_{\lceil N/d \rceil} \subset \mathbb{R}^d.$$

By construction, these hyperplanes meet the first two conditions of the statement; however, they may not satisfy the third condition, as some might be orthogonal to $e_{\ell_\star}$. Suppose that exactly $2p$ hyperplanes are orthogonal to $e_{\ell_\star}$ for some $1 \leq p \leq \lceil N/d \rceil$. Without loss of generality, assume these are

$$\hat{\mathsf{H}}'_1, \hat{\mathsf{H}}''_1, \dots, \hat{\mathsf{H}}'_p, \hat{\mathsf{H}}''_p.$$

Since these hyperplanes are parallel, and different by (2.8.3), we get $\hat{R}_i \bigcap \hat{R}_j = \emptyset$ for $i \neq j$.

Let $1 \leqslant j \leqslant p$. Because $\mathcal{R} \cup \mathcal{B}$ is finite, we can slightly adjust the direction vector of $\hat{\mathsf{H}}'_j$ and $\hat{\mathsf{H}}''_j$. This yields new parallel hyperplanes H'_j and H''_j in $\mathbb{R}^d$ that are no longer orthogonal to $e_{\ell_\star}$, yet still enclose exactly R_j, with no other points of $\mathcal{R} \cup \mathcal{B}$ in between. For $p < j \leqslant \lceil N/d \rceil$, set $\mathsf{H}'_j = \hat{\mathsf{H}}'_j$ and $\mathsf{H}''_j = \hat{\mathsf{H}}''_j$. This yields a new family

$$\mathsf{H}'_1, \mathsf{H}''_1, \dots, \mathsf{H}'_p, \mathsf{H}''_p, \mathsf{H}'_{p+1}, \mathsf{H}''_{p+1}, \dots, \mathsf{H}'_{\lceil N/d \rceil}, \mathsf{H}''_{\lceil N/d \rceil} \subset \mathbb{R}^d$$

that meets all the required conditions. Moreover, this proves (2.4.1).

□

Proof of Proposition 2.4.3. Fix $\ell_\star \in [d]$ and consider $\tau_{\mathcal{R}} = \{x^{(\ell_\star)} > 1\}$, $\tau_{\mathcal{B}} = \{x^{(\ell_\star)} \leqslant 1\}$. The strategy is to evolve $\mathcal{R}$ into the interior of $\tau_{\mathcal{R}}$ while keeping fixed $\mathcal{B}$. We first assume $|\mathcal{R}| = |\mathcal{B}| = N$ and will then extend to the general case $|\mathcal{R}| \neq |\mathcal{B}|$.

By Proposition 2.4.1, there exist $2\lceil N/d \rceil$ pairwise parallel hyperplanes separating $(\mathcal{R}, \mathcal{B})$. Assume these hyperplanes are defined by

$$\mathsf{H}'_j : u_j \cdot x + h'_j = 0, \qquad \text{and} \qquad \mathsf{H}''_j : u_j \cdot x + h''_j = 0, \qquad \text{for } j \in [\lceil N/d \rceil],$$

where $h'_j > h''_j$, and $u_j \in \mathbb{S}^{d-1}$ satisfies $|u_j \cdot e_{\ell_\star}| < 1$. Moreover, the region between each pair $(\mathsf{H}'_j, \mathsf{H}''_j)$ encloses a subset $R_j \subset \mathcal{R}$ such that $|R_j| = d \wedge |\mathcal{R}|$, and no other points of $\mathcal{R} \cup \mathcal{B}$ lie between these hyperplanes, namely,

$$(\mathsf{H}'^+_j \setminus \mathsf{H}''^+_j) \cap (\mathcal{R} \cup \mathcal{B}) = R_j \subset \mathcal{R} \qquad \text{with } |R_j| = d \wedge |\mathcal{R}|,$$

where $\mathsf{H}'^+_j = \{x \in \mathbb{R}^d \mid u_j \cdot x + h'_j > 0\}$ and $\mathsf{H}''^+_j = \{x \in \mathbb{R}^d \mid u_j \cdot x + h''_j > 0\}$. Note that $\mathsf{H}''^+_j \subset \mathsf{H}'^+_j$ because $h'_j > h''_j$. Taking $t_0 = 0$, we define

$$(w, a, b)(t) = \sum_{k=1}^{2\lceil N/d \rceil} (w_k, a_k, b_k)\, \mathbf{1}_{(t_{k-1}, t_k)}(t)$$

such that, for $1 \leqslant j \leqslant \lceil N/d \rceil$:

1. $(w_{2j-1}, a_{2j-1}, b_{2j-1}) = (v_j, u_j/d'_j, h'_j/d'_j)$, where
   - $v_j \in \mathbb{S}^{d-1}$ satisfies $v_j \cdot u_j = 0$ and $v_j \cdot e_{\ell_\star} > 0$ (for instance, take v_j to be the normalized projection of $e_{\ell_\star}$ onto the orthogonal subspace $\langle u_j \rangle^\perp$);
   - $d'_j = \min\left\{\sigma_{\mathrm{tr}}(u_j \cdot x + h'_j) \mid x \in R_j\right\} > 0$.

   Inside the half-space $\{x \in \mathbb{R}^d \mid u_j \cdot x + h'_j \geqslant d'_j\} \subset \mathsf{H}'^+_j$ and over (t_{2j-2}, t_{2j-1}), equation $(\mathcal{V}_1^{\mathrm{tr}})$ becomes $\dot{x} = v_j$ with $v_j \cdot e_{\ell_\star} > 0$. Since the dataset is finite and contained in this half-space, we can choose $t_{2j-1} > t_{2j-2}$ such that $\Phi_{t_{2j-1}}(x) \in \tau_{\mathcal{R}}$ for all $x \in R_j$.
2. $(w_{2j}, a_{2j}, b_{2j}) = (-v_j, u_j/d''_j, h''_j/d''_j)$, where

- $d_j'' = \min\left\{\sigma_{\mathrm{tr}}(u_j \cdot x + h_j'') \mid x \in (\mathcal{R} \cup \mathcal{B}) \cap \mathsf{H}_j''^{+}\right\} > 0.$

Inside the half-space $\{x \in \mathbb{R}^d \mid u_j \cdot x + h_j'' \geqslant d_j''\} \subset \mathsf{H}_j''^{+}$ and over (t_{2j-1}, t_{2j}), equation ($\mathcal{V}_1^{\mathrm{tr}}$) becomes $\dot{x} = -v_j$. Now, we set $t_{2j} = 2t_{2j-1} - t_{2j-2}$ so that

$$\Phi_{t_{2j}}(x) = \Phi_{t_{2j-2}}(x) \qquad \text{for all } x \in (\mathcal{R} \cup \mathcal{B}) \cap \mathsf{H}_j''^{+}.$$

All the while, we have

$$\Phi_{t_{2j}}(x) = \Phi_{t_{2j-1}}(x) \qquad \text{for all } x \in (\mathcal{R} \cup \mathcal{B}) \cap \mathsf{H}_j''^{-}.$$

For $T = t_{2\lceil N/d \rceil}$, we conclude $\Phi_T(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T(\mathcal{B}) = \mathcal{B} \subset \tau_{\mathcal{B}}$, with $L = 2\lceil N/d \rceil - 1$.

If $|\mathcal{R}| < |\mathcal{B}|$, we apply the same argument to obtain $\Phi_T(\mathcal{R}) \subset \tau_{\mathcal{R}}$ and $\Phi_T(\mathcal{B}) = \mathcal{B} \subset \tau_{\mathcal{B}}$ with $L = 2\lceil |\mathcal{R}|/d \rceil - 1$, by virtue of Corollary 2.4.2. If $|\mathcal{R}| > |\mathcal{B}|$, we swap the roles of $\mathcal{R}$ and $\mathcal{B}$ and the definitions of $\tau_{\mathcal{R}}$ and $\tau_{\mathcal{B}}$, with $L = 2\lceil |\mathcal{B}|/d \rceil - 1$. □

Proof of Lemma 2.4.4. Since $w \cdot a = 0$, it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}(a \cdot x(t) + b) = a \cdot \dot{x}(t) = a \cdot w\sigma(a \cdot x(t) + b) = 0.$$

Thus, $\sigma(a \cdot x(t) + b)$ is constant for all t. Suppose $a \cdot x + b \leqslant 1$. Then

$$\Phi_t^{\mathrm{tr}}(x\,;\,\theta) = x + tw(a \cdot x + b) = \Phi_t(x\,;\,\theta_1) = \Phi_t(\Phi_t(x\,;\,\theta_1)\,;\,\theta_2).$$

Otherwise, if $a \cdot x + b > 1$,

$$\begin{aligned}\Phi_t(\Phi_t(x\,;\,\theta_1)\,;\,\theta_2) &= \Phi_t(x + tw(a \cdot x + b)\,;\,\theta_2)\\ &= x + tw(a \cdot x + b) - tw(a \cdot (x + tw(a \cdot x + b)) + b - 1)\\ &= x + tw = \Phi_t^{\mathrm{tr}}(x\,;\,\theta).\end{aligned}$$

□

Chapter 3

Data Interpolation

3.1 Introduction

This chapter specializes the flow-based framework of **Part I (Overview)** to *regression*, i.e. $\mathcal{Y} = \mathbb{R}^d$ and the readout is $g = \mathsf{Id}$, so that the predictor in (I.2) becomes $F_\theta = \Phi_T^\theta$. Given a dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ such as (I.1), training amounts to selecting a control $\theta(\cdot)$ so that the flow map Φ_T^θ fits every sample (x_i, y_i). The strongest such requirement is *exact interpolation*:

$$\Phi_T^\theta(x_i) = y_i \qquad \text{for all } i \in [N]. \tag{3.1.1}$$

Interpolation will be used in two related but conceptually distinct ways.

On the one hand, for a fixed dataset, it serves as a natural *feasibility* benchmark for training: (3.1.1) asks whether the chosen family of controlled flows contains a zero-training-error solution, i.e. whether the data-fitting term in $\mathcal{R}_N$ can in principle be driven to zero.

On the other hand, beyond a single dataset, we turn (3.1.1) into an architecture-level benchmark by asking whether it can be enforced for arbitrary finite matchings of size N. In the controlled-flow viewpoint of **Chapter 1**, this is precisely a *simultaneous controllability* question: a single control trajectory must steer all initial conditions $x_1, \dots, x_N$ to their prescribed targets $y_1, \dots, y_N$ at the common time T.

To address the interpolation problem, the literature offers two complementary routes:

- *Geometric control*: one proves controllability of a family of vector fields using Lie-bracket criteria and then lifts this to multi-point interpolation (a prototypical tool is the Chow–Rashevsky theorem); see, e.g., [AS04; AS22; Che+25c].[1]
- *Constructive control*: one explicitly concatenates elementary dynamics (typically using piecewise-constant controls) to enforce (3.1.1); see, e.g., [RZ23; RAZ22; ÁOZ25].

As in **Chapter 2**, the present chapter follows the constructive route: we build explicit controls (rather than merely proving existence) and count the degrees of freedom required by the construction. This makes it possible to discuss *parsimony*: among architectures that can interpolate any dataset of size N, we seek those that achieve this with the smallest resource budget.

This brings us to the main objective of the chapter: developing a quantitative interpolation theory for the width-depth family $(\mathcal{V}_p^+)$. In our setting, *depth* is encoded in the temporal structure of the control (number of switches), while *width* is the number of neurons in the vector-field parametrization. Understanding this trade-off is crucial, as both numerical and theoretical studies

[1] Classical works also refer to this problem as *n-transitivity*; see [MV94].

indicate that increasing depth can improve representation and optimization behavior [Hua03; ES15; MLP17], an effect particularly visible during training [Zha+16; YSJ19].

Regarding the implementation of these controls, we keep the ReLU activation because it allows us to implement the elementary dynamical primitives summarized in Remark 1.3.4. These primitives are the basic building blocks of our constructions. All variants we consider inherit the same hyperplane-based geometric picture from **Chapter 1**, but differ in the parametrization of the vector field (e.g. linear-in-control, restricted time dependence, or additional compositional layers).

The formal problem setup is given in **Section 3.2**, the main results are stated in **Section 3.3**, and **Section 3.4** discusses implications and open directions.

3.2 Setup

Fix a state dimension $d \geq 2$ (see Remark 1.1.2 for the one-dimensional obstruction), and let

$$\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$$

be a finite dataset as in (I.1), with $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$.

Since every flow map $\Phi_T^\theta : \mathbb{R}^d \to \mathbb{R}^d$ is a homeomorphism, and hence injective, exact interpolation is impossible if two distinct inputs are assigned the same target. We therefore restrict (I.1) to datasets satisfying

$$x_i \neq x_j \quad \text{and} \quad y_i \neq y_j \qquad \text{for all } i \neq j. \tag{3.2.1}$$

Any dataset satisfying (3.2.1) will be called *admissible*. Throughout the chapter, we depict each pair (x_i, y_i) using a filled circle for the input point x_i and an open circle for the target y_i, matching their colors to indicate the correspondence.

Controllability

Fix an architecture, that is, a parameter set Θ and a control family $(v_\theta)_{\theta \in \Theta}$. Given an admissible dataset $\mathcal{D}_N$, the interpolation problem consists of finding a final time $T > 0$ and a control input $\theta(\cdot) : [0, T] \to \Theta$ such that the flow $\Phi_t^\theta : \mathbb{R}^d \to \mathbb{R}^d$ generated by $\dot{x}(t) = v_{\theta(t)}(x(t))$ satisfies:

$$\Phi_T^\theta(x_i) = y_i \qquad \text{for all } i \in [N]. \tag{3.2.2}$$

When this happens, we say that the dataset $\mathcal{D}_N$ is *simultaneously controllable* for the system.

As in **Chapter 2**, we adopt a worst-case notion of *expressivity*: the architecture should be able to enforce (3.2.2) for every admissible dataset.

Definition 3.2.1. We say that the system defined by a family of vector fields possesses the *simultaneous controllability* property, or *Universal Interpolation Property* (UIP), if for every $N \geq 1$, every admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ is simultaneously controllable.

Remark 3.2.1 (Time rescaling). In the definition above, the final time T is treated as a free parameter. However, this formulation is equivalent to a fixed-horizon one whenever the family of vector fields $\mathcal{V}$ is closed under positive scalar multiplication (i.e., $\lambda \mathcal{V} \subset \mathcal{V}$ for all $\lambda > 0$). Indeed, if a trajectory $x(t)$ solves the system on $[0, T]$, then for any prescribed horizon $T_* > 0$, the time-rescaled curve $\tilde{x}(s) \coloneqq x(sT/T_*)$ satisfies

$$\dot{\tilde{x}}(s) = \frac{T}{T_*} v_{\theta(sT/T_*)}\big(\tilde{x}(s)\big) \qquad \text{for } s \in [0, T_*].$$

Since $T/T_* > 0$, the scaled vector field remains valid within $\mathcal{V}$. Thus, any interpolation achieved at some time T can be exactly reproduced at any arbitrary time horizon T_*.

Architecture variants

We now specify the families of v_θ for which we analyze the UIP. They are variations of the models introduced in **Section 1.3** and can be viewed as different ways of trading time-dependence (*depth-in-time*) against spatial parametrization (*width*).

I. The one-layer ansatz. Consider the space ($\mathcal{V}_p$) imposing $\sigma(s) = s_+$,

$$v_{\theta(t)}(x) = \sum_{j=1}^{p} w_j(t)\,(a_j(t)\cdot x + b_j(t))_+\,, \qquad t \in [0,T], \tag{$\mathcal{V}_p^+$}$$

with control $\theta(\cdot) : t \mapsto (w_j(t), a_j(t), b_j(t))_{j=1}^p \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$. Unless stated otherwise, we restrict to piecewise-constant controls. If $\theta(\cdot)$ has L jumps, then the dynamics can be viewed as a depth-$(L+1)$ residual network upon time discretization, while p is the width.

Varying (p, L) in ($\mathcal{V}_p^+$) interpolates between two limiting regimes:

- ***Shallow* (autonomous) regime:** parameters are constant in time ($L = 0$), while p can be large:

  $$\dot{x}(t) = \sum_{j=1}^{p} w_j\,\big(a_j \cdot x(t) + b_j\big)_+, \qquad t \in [0,T]. \tag{$\mathcal{V}_{\mathrm{a}}^+$}$$

  The right-hand side is a d-dimensional one-hidden-layer ReLU network; its approximation properties are classical [Cyb89; Hor93; Les+93; Pin99]. At the discrete level, ($\mathcal{V}_{\mathrm{a}}^+$) corresponds to the ResNet depicted in Figure 3.1a.

- ***Narrow* (deep-in-time) regime:** here $p = 1$ is fixed, while L may grow:

  $$\dot{x}(t) = w(t)\,\big(a(t)\cdot x(t) + b(t)\big)_+, \qquad t \in [0,T]. \tag{$\mathcal{V}_1^+$}$$

  This is the architecture used in **Chapter 2** and studied, e.g., in [LJ18; RZ23]. Compared to ($\mathcal{V}_{\mathrm{a}}^+$), the dynamics are simpler, at the expense of L typically growing with N. Discretizations yield deep ResNets with one neuron per hidden layer (see Figure 3.1b).

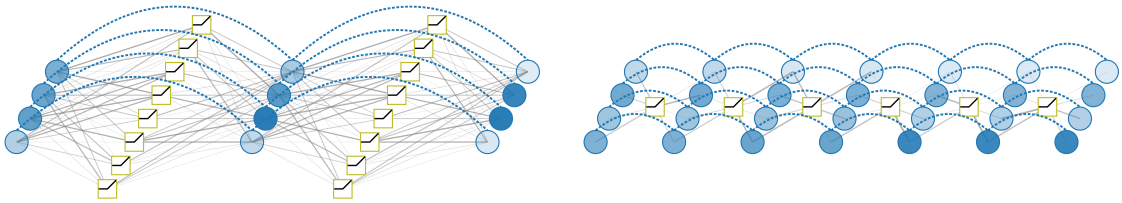
II. The semi-autonomous model. To approach the autonomous setting while retaining a minimal and explicit time dependence, we introduce

$$v_\theta(t,x) = \sum_{j=1}^{p} w_j\,(a_j \cdot x + c_j t + b_j)_+\,, \qquad t \in [0,T], \tag{$\mathcal{V}_{\mathrm{sa}}^+$}$$

with *constant* parameters $\theta \equiv (w_j, a_j, b_j, c_j)_{j=1}^p \in \big(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}\big)^p$. This model was investigated in [Li+26], where it was termed a *semi-autonomous neural ODE*. Geometrically, the activation hyperplanes translate at constant speeds c_j, while their orientations and flow directions are fixed. This retains the hyperplane-based intuition from **Chapter 1** while removing the need for switching.

III. The two-layer autonomous model. We consider a minimal deepening of ($\mathcal{V}_{\mathrm{a}}^+$) while preserving autonomy:

$$v_\theta(x) = \sum_{k=1}^{q} w_k \Big(\sum_{j=1}^{p} u_{kj}\,(a_{kj}\cdot x + b_{kj})_+ + c_k\Big)_+, \qquad x \in \mathbb{R}^d, \tag{$\mathcal{V}_{\mathrm{2a}}^+$}$$



(a) (Shallow ResNet): $L = 2$ residual blocks and hidden layer width $p = 8$, where all blocks share identical parameters.

(b) (Narrow ResNet): $L = 6$ residual blocks and $p = 1$, with layer-dependent parameters.

Figure 3.1: ResNets with $d = 4$ and $\sigma(s) = s_+$. Same color conventions as in Figure 2.

with constant parameters $w_k, a_{kj} \in \mathbb{R}^d$ and $c_k, u_{kj}, b_{kj} \in \mathbb{R}$. Compared to $(\mathcal{V}_{\mathrm{a}}^+)$, the *two-layer autonomous neural ODE* in $(\mathcal{V}_{\mathrm{2a}}^+)$ enriches the piecewise-affine geometry of the vector field without relying on time dependence.

IV. Linear-in-control model. Finally, we consider a neural ODE that is linear in the parameters:

$$v_{\theta(t)}(x) = W(t)\, x_+ + b(t), \qquad t \in [0, T], \tag{$\mathcal{V}_\ell^+$}$$

where the control is $\theta(\cdot) : t \mapsto (W(t), b(t)) \in \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d$. Here the width is fixed to $p = d$ for dimensional matching. The family of associated vector fields is convex and serves as a theoretical testbed [SF25; AS22].

Objectives

The rest of this chapter addresses interpolation from two complementary angles.

First, for each architecture above, we ask whether the UIP holds. A recurring theme is to quantify how far one can push toward autonomy while retaining exact controllability.

Second, we quantify the resources required by explicit constructions. For an architecture with parameter space $\Theta \subset \mathbb{R}^m$ and a piecewise-constant control with L jumps, we define

$$\kappa(\Theta, L) := (L + 1)\, m, \qquad m = \dim(\Theta). \tag{3.2.3}$$

This is the number of scalar values required to specify the control on the $(L + 1)$ constant time intervals. We treat switching times as fixed and do not include them in the count, since they do not affect existence of interpolating controls. When $\dim(\Theta)$ depends only on width, we write $\kappa(p, L)$ or $\kappa(p, q)$ accordingly. For instance, in $(\mathcal{V}_p^+)$ one has $m = p(2d+1)$ and thus $\kappa(p, L) = (L+1)p(2d+1)$.

Our goal is to identify architectures that possess the UIP while minimizing κ.

Sources

The results presented in the rest of this chapter are drawn from three works, distributed according to the architectural variants introduced above:

- **Model I** $(\mathcal{V}_p^+)$: based on [ÁSZ24].

- **Models II and III** ($\mathcal{V}^+_{\mathrm{sa}}$)–($\mathcal{V}^+_{\mathrm{2a}}$): based on [ÁLZ26b], an ongoing joint work.
- **Model IV** ($\mathcal{V}^+_{\ell}$): based on [ÁGR25].

3.3 Results

Table 3.1 provides a roadmap for the results. For each architecture we report a sufficient choice of width and (when relevant) number of switches L to interpolate an arbitrary admissible dataset $\mathcal{D}_N$, together with the resulting parameter budget κ from (3.2.3). Rows with no extra assumptions satisfy Definition 3.2.1 without additional hypotheses.

Table 3.1: Summary of sufficient resources for the UIP as per Definition 3.2.1. Complexity (κ) denotes the exact number of trainable scalar parameters, as per (3.2.3).

Model	Depth-Width	Complexity	Extra Assump.	Ref.
($\mathcal{V}^+_p$)	$p \geq 1 \mid L = 2\lceil \frac{N}{p} \rceil - 1$	$2(2d+1)\, p\lceil \frac{N}{p} \rceil$	–	Theorem 3.3.1
($\mathcal{V}^+_p$)	$p \geq 1 \mid L = \lceil \frac{N}{p} \rceil - 1$	$(2d+1)\, p\lceil \frac{N}{p} \rceil$	$d > N$	Corollary 3.3.5
($\mathcal{V}^+_{\mathrm{sa}}$)	$p = 2N \mid L = 0$	$4(d+1)N$	–	Theorem 3.3.7
($\mathcal{V}^+_{\ell}$)	$p = d \mid L = 4N + 2$	$(4N+3)\, d(d+1)$	–	Theorem 3.3.18
($\mathcal{V}^+_{\mathrm{a}}$)	$p = N \mid L = 0$	$(2d+1)N$	Assumption 2	Corollary 3.3.9
($\mathcal{V}^+_{\mathrm{2a}}$)	$p = 2d \mid q = N \mid L = 0$	$N(2d^2 + 5d + 1)$	(3.3.3)	Theorem 3.3.11

3.3.1 The one-layer ansatz

We start with ($\mathcal{V}^+_p$) and show that piecewise-constant controls can interpolate any admissible dataset. The result is quantitative in the sense that it makes explicit how the required number of switches depends on the width p.

Theorem 3.3.1. *Let $N \geq 1$, $d \geq 2$, and $T > 0$. For any admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ and any $p \geq 1$, there exists a piecewise-constant control θ such that the flow map Φ^θ_T generated by ($\mathcal{V}^+_p$) satisfies*

$$\Phi^\theta_T(x_i) = y_i, \qquad \text{for all } i \in [N].$$

Moreover, θ can be chosen with $L = 2\left\lceil \frac{N}{p} \right\rceil - 1$ jumps.

Remark 3.3.2 (Geometric intuition). The proof of Theorem 3.3.1, (deferred to **Subsection 3.5.1**) attains the UIP by explicitly concatenating the simple movements described in Remark 1.3.4.

A key difficulty with general superpositions of ReLU nonlinearities is that neurons can interact in a coupled way through the evolving activations. To decouple the dynamics, the construction imposes two geometric constraints on each time subinterval: all neurons share a common orientation a, and the flow directions are orthogonal to a (i.e. $\langle w_j, a \rangle = 0$). This forces the dynamics to act as a shear flow along parallel hyperplanes, preserving the activation values of the particles. As a result, the nonlinear interaction reduces to a family of independent linear transports within parallel strips.

The dataset is processed in batches of size p (matching the width). For each batch, the process consists of two phases:

1. **Parallel transport:** a shear flow simultaneously matches $d-1$ coordinates while keeping one pivot coordinate invariant;
2. **Reorientation:** a single control switch changes the orientation of the hyperplanes, after which the remaining pivot coordinate can be matched.

Iterating over $\lceil N/p \rceil$ batches yields $L = 2\lceil N/p \rceil - 1$.

Remark 3.3.3 (Alternative activations). The same strategy applies to any nonlinear $\sigma \in \mathscr{C}^{0,1}(\mathbb{R})$ possessing a *dead zone*,

$$\exists\, -\infty \le a < b \le +\infty \quad \text{with} \quad \sigma(z) = 0 \ \text{ for all } z \in (a,b). \tag{3.3.1}$$

For other activations (e.g. tanh or sigmoid), the above concatenation argument is not applicable. To our knowledge, no fully constructive method is currently available in such cases, although UIP has been proved via non-constructive geometric-control arguments [Che+25c].

Extending the construction to locally Lipschitz σ satisfying (3.3.1) (e.g. $\sigma^k(z) = (z \vee 0)^k$ for $k \ge 1$) is more delicate since finite-time blow-up can occur for general choices of $\theta(\cdot)$. Our specific shear-based controls avoid this issue: trajectories are piecewise affine in time, which guarantees global existence regardless of the growth of σ.

Theorem 3.3.1 exhibits a direct trade-off between width and the number of switches. The parameter budget (3.2.3) yields an explicit complexity estimate.

Corollary 3.3.4. *For the controls provided by Theorem 3.3.1, the complexity satisfies*

$$\kappa(p, L) = 2(2d+1)\, p \left\lceil \frac{N}{p} \right\rceil. \tag{3.3.2}$$

In particular, $\kappa(p,L) \ge 2(2d+1)N$, with equality if and only if p divides N (e.g. $p = 1$).

The two-phase strategy in Remark 3.3.2 implies that the switch count of Theorem 3.3.1 saturates at $L = 1$ once $p \ge N$, and therefore does not directly recover the fully autonomous regime ($L = 0$). In high dimension, this remaining switch can be eliminated by choosing a pivot direction adapted to the dataset.

Corollary 3.3.5. *Assume $d > N \ge 1$ and fix $T > 0$. For any admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ and any $p \ge 1$, there exists a piecewise-constant control θ such that the flow map Φ_T^θ generated by $(\mathcal{V}_p^+)$ satisfies*

$$\Phi_T^\theta(x_i) = y_i, \qquad \text{for all } i \in [N].$$

Moreover, θ can be chosen with $L = \lceil N/p \rceil - 1$ jumps.

In particular, if $p \ge N$, then $L = 0$ and $(\mathcal{V}_p^+)$ reduces to the autonomous model $(\mathcal{V}_a^+)$.

Essentially, the excess dimension allows one to fix a single pivot coordinate for the entire construction and thus bypass the reorientation step (see Figure 3.4 in **Section 3.5**).

Remark 3.3.6. The dimensional constraint $d > N$ can be circumvented by embedding the dataset into a higher-dimensional space, for instance by zero-padding the coordinates as in [DDT19]. This additional room eliminates topological obstructions and allows the same strategy to be applied to the lifted data.

Semi-autonomous regime as the constructive limit

In $(\mathcal{V}^+_{\mathrm{sa}})$, the directions (w_j, a_j) are time-independent, so the *reorientation* step of Remark 3.3.2 is no longer available. However, the explicit time dependence inside the activation, $(a_j \cdot x + c_j t + b_j)_+$, provides a different geometric degree of freedom: the activation hyperplanes translate at constant speeds c_j.

The next result shows that this is sufficient to recover exact controllability without switching.

Theorem 3.3.7. *For every $d \geq 2$ and $T > 0$, the system $(\mathcal{V}^+_{\mathrm{sa}})$ possesses the UIP. Moreover, given an admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$, it suffices to take $p = 2N$ neurons.*

3.3.2 Toward the autonomous limit

We now ask how far one can push the interpolation mechanism toward strict autonomy. We discuss three complementary routes: (i) assume a strong geometric separability property of the dataset; (ii) increase the compositional depth of the vector field while keeping it autonomous; (iii) relax exact interpolation to approximate controllability and leverage universal approximation.

(i) Probabilistic route.

When N is much larger than d, exact interpolation by an autonomous one-layer field is still possible under a strong geometric separability condition. The following hypothesis says that, after projection onto a suitable direction, each pair (x_i, y_i) lies in its own strip, separated from all other pairs.

Assumption 2 (Strip separability). There exist $a \in \mathbb{S}^{d-1}$, a permutation τ of $[N]$, and thresholds

$$-\infty < \beta_1 < \beta_2 < \cdots < \beta_{N+1} < +\infty$$

such that, for every $i \in [N]$,

$$a \cdot x_{\tau(i)} \in (\beta_i, \beta_{i+1}) \qquad \text{and} \qquad a \cdot y_{\tau(i)} \in (\beta_i, \beta_{i+1}).$$

Equivalently, the parallel hyperplanes $\{a \cdot x = \beta_i\}$ define strips (β_i, β_{i+1}) along the a-direction, and the i-th strip contains exactly the pair $(x_{\tau(i)}, y_{\tau(i)})$; see Figure 3.2a.

The following estimate quantifies how likely Assumption 2 is to hold for random data. As in (2.3.3), $\pi^k : \mathbb{R}^d \to \mathbb{R}$ denotes the projection onto the k-th coordinate.

Proposition 3.3.8. *Assume the setting of Assumption 2 with $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ sampled i.i.d. from a product measure μ (with i.i.d. coordinates). Then, for sufficiently large N,*

$$\mathbb{P}(\text{Assumption 2 holds}) \gtrsim 1 - \left[1 - \left(\frac{e}{2N}\right)^N\right]^d.$$

Under Assumption 2, one can construct a constant control such that the autonomous flow of $(\mathcal{V}^+_{\mathrm{a}})$ interpolates $\mathcal{D}_N$. This result serves as the neural-ODE analog of [Zha+16, Theorem 1]; see Figure 3.2b for a visualization.

Corollary 3.3.9. *Assume $d \geq 2$ and that the dataset $\mathcal{D}_N$ satisfies Assumption 2. Then, for any $T > 0$, there exists $\theta \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^N$ (i.e. taking $p = N$) such that the flow of $(\mathcal{V}^+_{\mathrm{a}})$ satisfies*

$$\Phi^\theta_T(x_i) = y_i \qquad \text{for all } i \in [N].$$



(a) Illustration of Assumption 2. The parallel hyperplanes orthogonal to a define strips, each containing exactly one pair (x_i, y_i).

(b) Trajectories of the autonomous flow ($\mathcal{V}_{\mathrm{a}}^{+}$) with width $p = N$ that successfully interpolate the dataset.

Figure 3.2: Geometric mechanism for exact interpolation in Corollary 3.3.9.

(ii) Localized fields by increased architectural depth.

We now return to the autonomous one-layer model ($\mathcal{V}_{\mathrm{a}}^{+}$). A major obstruction for a fully constructive UIP in this setting is the inability of one-layer ridge networks to localize in space: unless identically zero, they cannot have compact support in dimension $d \geq 2$. This rigidity is tied to the spectral degeneracy of ridge functions [Pin15]. We state and prove the following formal result for completeness.

Lemma 3.3.10. *Let $d \geq 2$ and let $\sigma \in \mathscr{C}^0(\mathbb{R})$ have at most polynomial growth. For any $p \geq 1$, define*

$$g(x) = \sum_{j=1}^{p} w_j \, \sigma(a_j \cdot x + b_j), \qquad (w_j, a_j, b_j)_{j=1}^{p} \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p.$$

If $g \not\equiv 0$, then $g \notin L^2(\mathbb{R}^d; \mathbb{R}^d)$. In particular, g cannot have compact support unless $g \equiv 0$.

While Lemma 3.3.10 does not rule out exact interpolation, it explains why explicit *pointwise* steering strategies are hard to implement with ($\mathcal{V}_{\mathrm{a}}^{+}$) alone. A minimal way to restore localization without reintroducing time dependence is to add one compositional layer. By composing ReLU units, the two-layer field ($\mathcal{V}_{\mathrm{2a}}^{+}$) can generate vector fields supported on convex polytopes (see Lemma 3.5.7 below), enabling the construction of localized controls.

We will further assume the stronger non-degeneracy condition

$$[x_i, y_i] \cap [x_j, y_j] = \emptyset \qquad \text{for all } i \neq j, \tag{3.3.3}$$

where $[a, b] \subset \mathbb{R}^d$ denotes the segment connecting $a, b \in \mathbb{R}^d$. In particular, (3.3.3) implies $x_i \neq y_j$ for all $i \neq j$.

Theorem 3.3.11. *Let $N \geq 1$, $d \geq 2$, and $T > 0$. Let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^d \times \mathbb{R}^d$ be an admissible dataset and assume in addition (3.3.3). Then there exists a (constant) control θ such that the flow map Φ_T^{θ} generated by ($\mathcal{V}_{\mathrm{2a}}^{+}$), with layer widths $p = 2d$ and $q = N$, satisfies*

$$\Phi_T^{\theta}(x_i) = y_i, \qquad \text{for all } i \in [N].$$

Remark 3.3.12 (Dimensional constraints and an improvement). The non-degeneracy condition (3.3.3) is heavily dependent on the ambient dimension. In $d \geq 3$, since one-dimensional objects generically do not intersect, this condition is satisfied for almost all data configurations. In

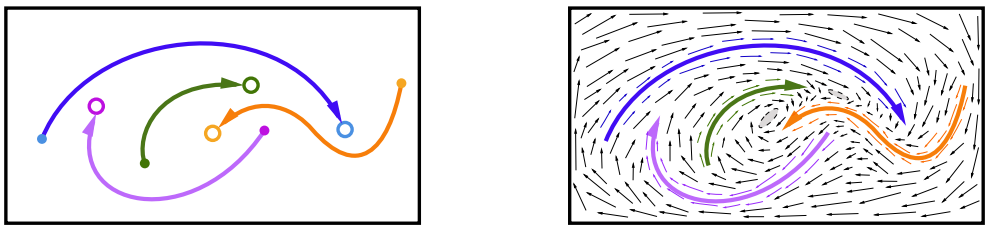


Figure 3.3: Construction of a smooth, compactly supported reference field $v_\star$ as in Proposition 3.3.13.

contrast, for $d = 2$, the condition is restrictive: straight-line segments connecting arbitrary pairs (x_i, y_i) easily intersect due to the lack of spatial degrees of freedom.

Our current construction (see the proof in **Subsection 3.3.2**) builds localized vector fields supported strictly around the reference segments $[x_i, y_i]$; thus, the validity of Theorem 3.3.18 in $d = 2$ is limited to configurations where these straight paths do not cross.

However, a more general strategy can establish the UIP for all $d \geqslant 2$ without requiring (3.3.3). Instead of a direct segment, each particle is guided through a sequence of intermediate waypoints chosen to avoid collisions. The global control is then assembled by concatenating *local connectors*: localized two-layer fields designed to steer the flow through specific turns at these waypoints. While this approach theoretically solves the intersection problem by routing around obstacles, due to the technical complexity and length of this construction, we consider it beyond the scope of this chapter and defer its full treatment to future work.

(iii) Relaxation to approximate controllability.

Outside special regimes such as Assumption 2 or added layers as in $(\mathcal{V}^+_{2a})$, proving the UIP for the autonomous one-layer case $(\mathcal{V}^+_a)$ is delicate. A convenient relaxation is to allow a small error $\varepsilon > 0$: rather than imposing $\Phi^\theta_T(x_i) = y_i$, we ask for

$$\max_{1 \leqslant i \leqslant N} \left\| \Phi^\theta_T(x_i) - y_i \right\| < \varepsilon.$$

The proof strategy has two stages. First, construct an *interpolating reference field* $v_\star$ with compact support whose time-T flow matches the data exactly. Second, approximate $v_\star$ uniformly on a compact domain containing the reference trajectories by a field of the form $(\mathcal{V}^+_a)$, and transfer the approximation error to the flow via a Grönwall stability estimate.

Proposition 3.3.13. *Fix $N \geq 1$, $d \geq 2$ and $T > 0$, and let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ be admissible. Let $R > 0$ be large enough so that $\{x_1, \dots, x_N, y_1, \dots, y_N\} \subset (-R, R)^d$. Then there is a compactly supported $v_\star \in C^\infty_c(\mathbb{R}^d; \mathbb{R}^d)$ such that the flow $\Phi^\star_t$ of $\dot{x} = v_\star(x)$ satisfies*

$$\Phi^\star_T(x_i) = y_i, \qquad \text{for all } i \in [N],$$

and each trajectory segment $\{\Phi^\star_t(x_i) \mid t \in [0, T]\}$ is contained in $(-R, R)^d$.

Because $v_\star$ is smooth, it admits quantitative approximation by one-hidden-layer ReLU vector fields on any compact set containing the reference trajectories. We invoke Barron-type approximation results, which give rates with a favorable dependence on the ambient dimension. Combining this with flow stability yields the following quantitative approximate controllability statement.

Theorem 3.3.14. *Fix $N \geq 1$, $d \geq 2$ and $T > 0$, and let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ be as in (I.1) with $\mathcal{Y} = \mathbb{R}^d$. Then for every $p \geq 1$ there exists $\theta = (w_j, a_j, b_j)_{j=1}^p \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R})^p$ such that*

$$v_\theta(x) = \sum_{j=1}^{p} w_j \left(a_j \cdot x + b_j\right)_+ \qquad (x \in \mathbb{R}^d),$$

generates a flow map $\Phi_T^\theta : \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$\max_{1 \leqslant i \leqslant N} \left\|\Phi_T^\theta(x_i) - y_i\right\| \;\leq\; C_{T,\mathcal{D}_N} \frac{\sqrt{d + \log(2+p)}}{p^{\frac{1}{2}+\frac{1}{d}}}, \tag{3.3.4}$$

where $C_{T,\mathcal{D}_N} > 0$ is independent of p. Moreover, $\mathrm{Lip}(v_\theta)$ is bounded uniformly in p.

Remark 3.3.15. Exact interpolation requires (3.2.1) because each map Φ_T^θ is injective. Once we relax to approximate interpolation, this obstruction disappears: given any dataset (allowing $y_i = y_j$) and any $\varepsilon > 0$, one can perturb the targets to distinct points y_i' with $\|y_i' - y_i\| < \varepsilon$, apply Theorem 3.3.14 to $\{(x_i, y_i')\}_{i=1}^N$, and then recover an ε-accurate approximation to the original targets by the triangle inequality.

Remark 3.3.16 (Non-smoothness). If one only assumes that the reference field is Lipschitz (without additional Barron control), one can instead invoke the results in [DHP21]–[Bac17], leading to analogous rates of the form $\max_{1 \leqslant i \leqslant N} \left\|\Phi_T^\theta(x_i) - y_i\right\| \lesssim (\log p)\, p^{-1/d}$ such as those in [ÁSZ24].

Remark 3.3.17 (Other approximation families). The proof of Theorem 3.3.14, which can be found in **Subsection 3.5.2**, only uses a quantitative density result with an explicit rate. Hence, the same strategy applies to other approximation families (polynomials, trigonometric systems, finite elements, wavelets), with the corresponding convergence rates.

3.3.3 The linear ansatz

We now consider the linear-in-control architecture ($\mathcal{V}_\ell^+$).

Theorem 3.3.18. *Let $N \geq 1$, $d \geq 2$, and $T > 0$. For any admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$, there exists a piecewise-constant control θ such that the flow map Φ_T^θ generated by ($\mathcal{V}_\ell^+$) satisfies*

$$\Phi_T^\theta(x_i) = y_i, \qquad \text{for all } i \in [N].$$

Moreover, θ can be chosen with at most $L = 4N + 2$ jumps.

The proof—see **Subsection 3.5.3**—is a modification of the construction used for Theorem 3.3.1.

3.4 Perspectives

The constructive controllability results and quantitative bounds obtained in this chapter naturally suggest several directions for further work.

1. **Minimizing the number of time jumps.** For the one-layer model ($\mathcal{V}_p^+$), Theorem 3.3.1 provides an explicit UIP construction with $L = 2\lceil N/p \rceil - 1$ switches (and $L = \lceil N/p \rceil - 1$ when $d > N$, cf. Corollary 3.3.5). A basic open problem is to determine the optimal switching

complexity: for fixed (d, p, N), what is the smallest number of jumps in time needed to interpolate *every* admissible dataset? Similar questions arise for the linear-in-control architecture $(\mathcal{V}_\ell^+)$ (Theorem 3.3.18).

2. **Interpolation in the one-layer, autonomous model.** Exact interpolation with $(\mathcal{V}_{\mathrm{a}}^+)$ was established here only in restricted regimes: high dimension $d > N$ (Corollary 3.3.5) or via separability under Assumption 2 (Corollary 3.3.9). It remains an open problem to find a construction for arbitrary $d \geq 2$ and arbitrary admissible datasets, or at least to replace Assumption 2 by a substantially weaker geometric condition.

   An intermediate step is to assume the existence of $a \in \mathbb{S}^{d-1}$ and a permutation τ of $[N]$ such that

   $$a \cdot x_{\tau(1)} < \cdots < a \cdot x_{\tau(N)} \qquad \text{and} \qquad a \cdot y_{\tau(1)} < \cdots < a \cdot y_{\tau(N)}.$$

   One could then attempt to apply the transversal separation mechanism used in the proof of Corollary 3.3.9, followed by a careful case-by-case analysis.

3. **Universal approximation.** Beyond interpolation of finitely many points, it is natural to ask for quantitative universal approximation results with explicit control complexity. In [RZ23], it is shown that $(\mathcal{V}_1^+)$ can approximate any simple compactly supported function by means of simultaneous control and space compression. Minimizing the number of time jumps required for this control, particularly by increasing width, is a non-trivial task that may require new constructive ideas.

4. **Switching hidden dimensions.** All models considered here keep the width p fixed along time. Allowing variations of p could, in principle, yield new trade-offs: shrinking p may reduce complexity, while enlarging it may create *working space* for separation or localization. Exploring effective methods to implement these transitions is an open mathematical question; for instance via (i) an augmented state space of maximal dimension together with controlled projections, or (ii) switching models where dimension changes occur at discrete jump times. For a recent work tackling architectures with variable width from a mean-field control perspective, we refer to [CFS25].

5. **Activation functions and constructive control beyond (3.3.1).** As discussed in Remark 3.3.3, the explicit control algorithms developed in this chapter rely on the presence of regions that are fixed by the flow. For more general activations, such fixed regions may disappear, and the dynamics tend to continuously disperse points. Developing constructive schemes comparable to those in this work would enable systematic complexity estimates (switching, width, time horizon) across different dynamics, and would inform the question of which activation functions are optimal from an expressivity viewpoint.

6. **The gap with optimal controls.** Our results are primarily constructive; consequently, a recurring issue is whether the obtained costs (switching, width) are close to optimal. Duality can be exploited to obtain sharper estimates of the gap between optimal and explicit controls for exact or approximate controllability. For instance, in [FZ00] a related duality-based analysis is carried out in the context of linear parabolic equations.

7. **Choosing an optimal reference.** In Theorem 3.3.14, the constant $C_{T,R,\mathcal{D}_N}$ depends on the reference field $v_\star$ and combines two distinct ingredients: (i) a flow stability estimate on $[0, T]$, which depends on the regularity of $v_\star$, and (ii) an approximation step, whose quality depends on how well $v_\star$ can be approximated on that same region.

   This motivates, for a fixed dataset and a fixed compact convex domain Ω containing every point

(e.g. $\Omega = [-R, R]^d$ as in Proposition 3.3.13), introducing the (non-empty) space of fields

$$\mathcal{V}_{\Omega,T}(\mathcal{D}_N) \coloneqq \left\{ v \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d) \;\middle|\; \Phi_T^v(x_i) = y_i \ \forall i \in [N], \text{ and } \Phi_t^v(x_i) \in \mathrm{Int}(\Omega) \ \forall t \in [0,T] \right\},$$

and the cost

$$\mathsf{L}_{\Omega,T}(\mathcal{D}_N) \coloneqq \inf_{v \in \mathcal{V}_{\Omega,T}(\mathcal{D}_N)} \mathrm{Lip}\left(v\big|_\Omega\right). \tag{3.4.1}$$

Understanding (or estimating) $\mathsf{L}_{\Omega,T}(\mathcal{D}_N)$ would lead to sharper stability constants and could ultimately improve the quantitative bounds in Theorem 3.3.14.

3.5 Proofs

3.5.1 Proofs of Subsection 3.3.1

Proof of Theorem 3.3.1

Recall that the piecewise-constant control is $\theta(\cdot) = \left(w_j(\cdot), a_j(\cdot), b_j(\cdot)\right)_{j=1}^p$. We begin with a generic-position reduction.

Lemma 3.5.1. *Let $N \geq 1$ and $d \geq 2$, and let $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ for $i \neq j$. There exists a (linear) change of coordinates in $\mathbb{R}^d$ such that, in the new coordinates,*

$$x_i^{(1)} \neq x_j^{(1)} \qquad \text{and} \qquad y_i^{(2)} \neq y_j^{(2)}, \qquad \text{for all } i \neq j. \tag{3.5.1}$$

Assume for the moment that (3.5.1) holds. Set

$$M \coloneqq \left\lceil \frac{N}{p} \right\rceil.$$

By time rescaling (cf. Remark 3.2.1), it suffices to construct a piecewise-constant control on a convenient horizon; we work on $[0, 2M]$ and produce a control that is constant on each unit interval. This yields the desired statement for any prescribed $T > 0$ without changing the number of switches.

Step 1. Using (3.5.1), relabel the data so that

$$x_1^{(1)} < x_2^{(1)} < \cdots < x_N^{(1)}.$$

For $r \in [M]$ define the index block

$$I_r \coloneqq \left\{(r-1)p + 1, \ldots, (rp) \wedge N\right\}, \qquad s_r \coloneqq |I_r| \leq p.$$

We construct θ on each interval $[r-1, r]$ so that, at time r, the points $\{\Phi_r^\theta(x_i)\}_{i \in I_r}$ satisfy

$$\left(\Phi_r^\theta(x_i)\right)^{(k)} = y_i^{(k)} \qquad \text{for all } i \in I_r, \text{ and } k \in \{2, \ldots, d\},$$

and points already treated in earlier blocks remain unchanged thereafter. (Points in future blocks may move; this is harmless.)

Fix $r \in [M]$. For readability, write $\tilde{x}_i \coloneqq \Phi_{r-1}^\theta(x_i)$ for the configuration at time $t = r-1$. On $[r-1, r]$ we choose

$$a_j = e_1, \qquad w_j^{(1)} = 0 \quad \text{for all } j \in [p],$$

and we set $w_j = 0$ for $j > s_r$. Next choose biases $b_1 > \cdots > b_{s_r}$ such that

$$-b_1 < \tilde{x}^{(1)}_{(r-1)p+1} < -b_2 < \cdots < \tilde{x}^{(1)}_{(r-1)p+s_r-1} < -b_{s_r} < \tilde{x}^{(1)}_{(r-1)p+s_r}, \tag{3.5.2}$$

and, if $r \geq 2$, also $\tilde{x}^{(1)}_{(r-1)p} < -b_1$. Then every point $\tilde{x}_i$ with $i \leq (r-1)p$ satisfies $\tilde{x}^{(1)}_i < -b_1$, hence all ReLU terms vanish there and such points are fixed on $[r-1, r]$.

For $j \in [s_r]$, the point $\tilde{x}_{(r-1)p+j}$ lies in the strip where exactly the first j neurons are active, hence along its trajectory

$$\dot{x} = \sum_{\ell=1}^{j} w_\ell \big(x^{(1)} + b_\ell\big), \qquad \text{and in particular} \qquad \dot{x}^{(1)} = 0.$$

Therefore $x^{(1)}$ is constant and, for each $k \in \{2, \dots, d\}$, integration over $[r-1, r]$ gives

$$\big(\Phi^\theta_r(x_{(r-1)p+j})\big)^{(k)} = \tilde{x}^{(k)}_{(r-1)p+j} + \sum_{\ell=1}^{j} w^{(k)}_\ell \big(\tilde{x}^{(1)}_{(r-1)p+j} + b_\ell\big).$$

Since $\tilde{x}^{(1)}_{(r-1)p+j} + b_j > 0$ by (3.5.2), we may choose $w^{(k)}_j$ recursively (for each $k \in \{2, \dots, d\}$ and $j \in [s_r]$) by

$$w^{(k)}_j := \frac{y^{(k)}_{(r-1)p+j} - \tilde{x}^{(k)}_{(r-1)p+j} - \sum_{\ell=1}^{j-1} w^{(k)}_\ell \big(\tilde{x}^{(1)}_{(r-1)p+j} + b_\ell\big)}{\tilde{x}^{(1)}_{(r-1)p+j} + b_j}. \tag{3.5.3}$$

This enforces $(\Phi^\theta_r(x_i))^{(k)} = y^{(k)}_i$ for all $i \in I_r$ and $k = 2, \dots, d$, while leaving all points with $i \leq (r-1)p$ unchanged. Iterating over $r \in [M]$, we obtain at time M:

$$\big(\Phi^\theta_M(x_i)\big)^{(k)} = y^{(k)}_i \qquad \text{for all } i \in [N] \text{ and } k \in \{2, \dots, d\}. \tag{3.5.4}$$

Step 2. At time M, (3.5.4) yields $(\Phi^\theta_M(x_i))^{(2)} = y^{(2)}_i$ for all $i \in [N]$; in particular these values are pairwise distinct by (3.5.1). Relabel the pairs (x_i, y_i) so that

$$\big(\Phi^\theta_M(x_1)\big)^{(2)} < \cdots < \big(\Phi^\theta_M(x_N)\big)^{(2)}.$$

Define blocks I_r as above, now with respect to this ordering.

For each $r \in [M]$, we design θ on the interval $[M+r-1, M+r]$ as follows. Write $\hat{x}_i := \Phi^\theta_{M+r-1}(x_i)$ for the current configuration. We take

$$a_j = e_2, \qquad w_j = \omega_j e_1 \quad \text{for } j \in [p],$$

with $\omega_j \in \mathbb{R}$, and again set $w_j = 0$ for $j > s_r$. Choose biases $b_1 > \cdots > b_{s_r}$ so that the analogue of (3.5.2) holds with the second coordinate $\hat{x}^{(2)}$ (in particular, all points from earlier blocks lie in $\hat{x}^{(2)} < -b_1$ and are therefore frozen on this interval). Since w_j is parallel to e_1, we have $\dot{x}^{(k)} = 0$ for $k \geq 2$, hence coordinates $2, \dots, d$ remain unchanged during Step 2. For $j \in [s_r]$ the point in position j of the current block satisfies, after integrating on $[M+r-1, M+r]$,

$$\big(\Phi^\theta_{M+r}(x_{(r-1)p+j})\big)^{(1)} = \hat{x}^{(1)}_{(r-1)p+j} + \sum_{\ell=1}^{j} \omega_\ell \big(\hat{x}^{(2)}_{(r-1)p+j} + b_\ell\big).$$

As before, $\hat{x}^{(2)}_{(r-1)p+j} + b_j > 0$, so we may choose ω_j recursively by

$$\omega_j \coloneqq \frac{y^{(1)}_{(r-1)p+j} - \hat{x}^{(1)}_{(r-1)p+j} - \sum_{\ell=1}^{j-1} \omega_\ell \big(\hat{x}^{(2)}_{(r-1)p+j} + b_\ell\big)}{\hat{x}^{(2)}_{(r-1)p+j} + b_j}. \tag{3.5.5}$$

This enforces $\big(\Phi^\theta_{M+r}(x_i)\big)^{(1)} = y^{(1)}_i$ for all $i \in I_r$, while keeping all previously matched points fixed. After $r = M$ we conclude that

$$\Phi^\theta_{2M}(x_i) = y_i \qquad \text{for all } i \in [N].$$

Switch count. The control is constant on each unit interval of $[0, 2M]$, hence it has $(2M) - 1 = 2\lceil N/p \rceil - 1$ discontinuities. This proves Theorem 3.3.1 under (3.5.1). The general case follows from Lemma 3.5.1.

Proof of Corollary 3.3.5

Proof of Lemma 3.5.1. For $N = 1$ the claim is trivial, so assume $N \geq 2$. For any $i \neq j$, let $v_{ij} \coloneqq x_i - x_j$ and $w_{ij} \coloneqq y_i - y_j$.

First, we choose $u_1 \in \mathbb{S}^{d-1}$ that satisfies two conditions:

1. $u_1 \cdot v_{ij} \neq 0$ for all pairs i, j. This ensures the projected points $x^{(1)}_i$ are distinct.
2. u_1 is not parallel to any w_{ij}.

The first condition excludes a finite union of hyperplanes (defined by $u \cdot v_{ij} = 0$). The second condition excludes a finite union of lines (the directions of w_{ij}). Since $d \geq 2$, the union of these sets has measure zero in $\mathbb{R}^d$, so such a u_1 exists.

Next, we look for $u_2 \in \mathbb{S}^{d-1}$ orthogonal to u_1. Consider the subspace $V = \{z \in \mathbb{R}^d \mid z \cdot u_1 = 0\}$, which has dimension $d - 1$. We need to choose $u_2 \in V$ such that $u_2 \cdot w_{ij} \neq 0$ for all pairs.

Fix a pair (i, j). Since we chose u_1 not parallel to w_{ij}, the vector w_{ij} is not orthogonal to V. Thus, the constraint $\{z \in V \mid z \cdot w_{ij} = 0\}$ defines a proper subspace of V. Since a finite union of proper subspaces cannot cover V, there exists a unit vector $u_2 \in V$ such that $u_2 \cdot w_{ij} \neq 0$ for all i, j.

Finally, we complete $\{u_1, u_2\}$ to an orthonormal basis $\{u_1, \ldots, u_d\}$. □

For $i \in [N]$ we set $\Delta_i \coloneqq x_i - y_i \in \mathbb{R}^d$. Since $d > N$, the family $\{\Delta_i\}_{i \in [N]}$ spans a subspace of dimension at most N, hence its orthogonal complement

$$V \coloneqq \bigcap_{i=1}^{N} \big\{v \in \mathbb{R}^d \mid v \cdot \Delta_i = 0\big\}$$

has dimension at least $d - N \geq 1$. Choose any unit vector $u_1 \in V \cap \mathbb{S}^{d-1}$ and complete it to an orthonormal basis $\{u_1, \ldots, u_d\}$ of $\mathbb{R}^d$. We use the associated coordinates $z^{(k)} \coloneqq z \cdot u_k$.

By construction, for every $i \in [N]$,

$$x^{(1)}_i - y^{(1)}_i \;=\; u_1 \cdot (x_i - y_i) \;=\; u_1 \cdot \Delta_i \;=\; 0, \tag{3.5.6}$$

The procedure is illustrated in Figure 3.4.



Figure 3.4: For $N = 2$, construction of a new first axis $x^{(1)} \mapsto \tilde{x}^{(1)}$ in $\mathbb{R}^3$, such that the first coordinates of the pairs (x_i, y_i) are matched. Theorem 3.3.1 is applied afterwards.

We now apply the construction of **Step 1** in the proof of Theorem 3.3.1 (working in the above coordinates). Namely, we design a piecewise-constant control for $(\mathcal{V}_p^+)$ that keeps the first coordinate fixed and matches coordinates $2, \dots, d$ in blocks of size p. After $M := \lceil N/p \rceil$ iterations, we obtain a time $t = M$ such that

$$\big(\Phi_M^\theta(x_i)\big)^{(k)} = y_i^{(k)} \qquad \text{for all } i \in [N] \text{ and } k \in \{2, \dots, d\},$$

while $(\Phi_M^\theta(x_i))^{(1)} = x_i^{(1)}$. Together with (3.5.6) this yields $\Phi_M^\theta(x_i) = y_i$ for all $i \in [N]$.

Finally, by time rescaling (cf. Remark 3.2.1), we transfer this construction to any prescribed horizon $T > 0$ without changing the number of switches. Since the control is constant on each of the M subintervals used in the block construction, the number of discontinuities is

$$L \;=\; M - 1 \;=\; \left\lceil \frac{N}{p} \right\rceil - 1.$$

In particular, if $p \geq N$ then $M = 1$ and thus $L = 0$. □

Proof of Theorem 3.3.7

We show that $(\mathcal{V}_{\mathrm{sa}}^+)$ is simultaneously controllable, and we give an explicit construction. In particular, to steer N points to prescribed targets we use at most $p = 2N$ neurons.

Geometrically, each triple (a_j, b_j, c_j) defines a moving hyperplane

$$a_j \cdot x(t) + c_j t + b_j = 0,$$

which translates with constant normal velocity c_j in the direction of a_j.

The following preliminary lemma provides exact controllability of $d - 1$ coordinates while keeping one coordinate fixed.

Lemma 3.5.2. *Let $N \geq 1$, $d \geq 2$, and let $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ be any finite dataset. For any $T > 0$ there exist parameters*

$$(w_i, a_i, b_i, c_i)_{i=1}^N \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^N$$

such that the flow map Φ_T^θ of $(\mathcal{V}_{\mathrm{sa}}^+)$ satisfies, for each $i \in [N]$,

$$\begin{aligned} \big[\Phi_T^\theta(x_i)\big]^{(1)} &= x_i^{(1)}, \\ \big[\Phi_T^\theta(x_i)\big]^{(k)} &= y_i^{(k)}, \qquad k = 2, \dots, d. \end{aligned}$$

Proof. Up to relabeling the points (and, if needed, choosing the first coordinate along a generic direction), we may assume

$$x_1^{(1)} < x_2^{(1)} < \cdots < x_N^{(1)}. \tag{3.5.7}$$

We set

$$a_i = e_1, \qquad w_i^{(1)} = 0, \qquad \text{for all } i \in [N].$$

Then ($\mathcal{V}_{\mathrm{sa}}^{+}$) becomes

$$\dot{x}(t) = \sum_{i=1}^{N} w_i \, (x^{(1)}(t) + c_i t + b_i)_+, \qquad t \in [0,T],$$

so, in components,

$$\dot{x}^{(1)}(t) = 0, \qquad \dot{x}^{(k)}(t) = \sum_{i=1}^{N} w_i^{(k)} \, (x^{(1)}(t) + c_i t + b_i)_+, \qquad k = 2, \dots, d. \tag{3.5.8}$$

In particular,

$$\left[\Phi_t^\theta(x_i)\right]^{(1)} = x_i^{(1)}, \qquad t \in [0,T],$$

so the first coordinate is frozen and the ordering (3.5.7) is preserved.

Step 1. Set

$$c_1 = \frac{x_1^{(1)} - 1 - x_N^{(1)}}{T}, \qquad b_1 = 1 - x_1^{(1)},$$

and for $i = 2, \dots, N$,

$$c_i = \frac{x_{i-1}^{(1)} - x_N^{(1)}}{T}, \qquad b_i = -x_{i-1}^{(1)}.$$

Then each moving hyperplane

$$\mathsf{H}_i(t): \quad x^{(1)} + c_i t + b_i = 0$$

satisfies:

- $\mathsf{H}_i(t)$ is orthogonal to e_1 for all t and translates with constant speed $|c_i|$ along e_1.
- For $i = 1$, $\mathsf{H}_1(0) : x^{(1)} = x_1^{(1)} - 1$ and $\mathsf{H}_1(T) : x^{(1)} = x_N^{(1)}$.
- For $i \geq 2$, $\mathsf{H}_i(0) : x^{(1)} = x_{i-1}^{(1)}$ and $\mathsf{H}_i(T) : x^{(1)} = x_N^{(1)}$.

By the choice of (b_i, c_i), the hyperplanes do not intersect on $[0,T)$ (they coalesce only at $t = T$).

Step 2. Fix $i \in [N]$. Since $x^{(1)}(t) \equiv x_i^{(1)}$, for each $k \geq 2$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\Phi_t^\theta(x_i)\right]^{(k)} = \sum_{j=1}^{N} w_j^{(k)} \, (x_i^{(1)} + c_j t + b_j)_+.$$

With the above choice of (c_j, b_j), if $j > i$ then $x_i^{(1)} + b_j \leq 0$ and $c_j \leq 0$, hence

$$x_i^{(1)} + c_j t + b_j \leq 0 \qquad \text{for all } t \in [0,T],$$

so neuron j is never active along the trajectory from x_i. Therefore,

$$\left[\Phi_T^\theta(x_i)\right]^{(k)} = x_i^{(k)} + \sum_{j=1}^{i} w_j^{(k)} \int_0^T (x_i^{(1)} + c_j t + b_j)_+ \, \mathrm{d}t.$$

Define

$$M_{i,j} \coloneqq \int_0^T (x_i^{(1)} + c_j t + b_j)_+ \, \mathrm{d}t.$$

By construction, $M_{i,j} > 0$ for all $1 \le j \le i \le N$, and in particular $M_{i,i} > 0$ for every i.

Step 3. Imposing $\left[\Phi_T^\theta(x_i)\right]^{(k)} = y_i^{(k)}$ for $k = 2, \dots, d$ yields, for each fixed k,

$$y_i^{(k)} - x_i^{(k)} = \sum_{j=1}^{i} w_j^{(k)} M_{i,j}, \qquad i \in [N],$$

which is a lower-triangular linear system in $(w_1^{(k)}, \dots, w_N^{(k)})$ with strictly positive diagonal. Hence it has a unique solution for each $k = 2, \dots, d$, completing the proof. □

Remark 3.5.3. In the proof of Lemma 3.5.2, all hyperplanes satisfy $\mathsf{H}_i(T) : x^{(1)} = x_N^{(1)}$. Moreover, since the first coordinate is constant and $x_i^{(1)} \le x_N^{(1)}$ for all i, we have

$$x_i^{(1)} + c_j t + b_j \le 0 \qquad \text{for all } t \ge T \text{ and all } i, j \in [N],$$

so the corresponding activations vanish for $t \ge T$. This *automatic shut-off* will be used below.

Proof of Theorem 3.3.7. Choose a unit vector so that the projections of the targets are pairwise distinct; up to a rigid change of coordinates we may assume this is the second coordinate, i.e.

$$y_i^{(2)} \ne y_j^{(2)} \quad \text{for } i \ne j.$$

We build $\theta \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^{2N}$ by concatenating two groups of N neurons which activate on disjoint time intervals.

Step 1. Apply Lemma 3.5.2 with horizon $T/2$. This provides parameters $(w_i, a_i, b_i, c_i)_{i=1}^N$ such that

$$\left[\Phi_{T/2}^\theta(x_i)\right]^{(1)} = x_i^{(1)}, \qquad \left[\Phi_{T/2}^\theta(x_i)\right]^{(k)} = y_i^{(k)}, \quad k = 2, \dots, d.$$

Moreover, by the construction (cf. Remark 3.5.3) these first N neurons are inactive for all $t \ge T/2$, i.e.

$$(a_i \cdot \Phi_t^\theta(x_j) + c_i t + b_i)_+ = 0, \qquad \text{for all } t \ge T/2 \text{ and all } i, j \in [N].$$

Define

$$m \coloneqq \left(\min_{1 \leqslant i \leqslant N} \min_{t \in [0, T/2]} \left[\Phi_t^\theta(x_i)\right]^{(2)} \right) \wedge \left(\min_{1 \leqslant i \leqslant N} y_i^{(2)} \right). \tag{3.5.9}$$

Step 2. Relabel the data points so that

$$y_1^{(2)} < y_2^{(2)} < \cdots < y_N^{(2)}. \tag{3.5.10}$$

At $t = T/2$ we already have $\left[\Phi_{T/2}^\theta(x_i)\right]^{(2)} = y_i^{(2)}$ and this coordinate will remain constant on the next interval $[T/2, T]$.

For the additional neurons indexed by $\ell = N+1, \dots, 2N$, set

$$a_\ell = -e_2, \qquad w_\ell^{(k)} = 0 \quad \text{for } k = 2, \dots, d,$$

so only the first coordinate is affected by these new neurons.

Choose (c_ℓ, b_ℓ) so that the corresponding moving hyperplanes

$$\mathsf{H}_\ell(t): \quad -x^{(2)} + c_\ell t + b_\ell = 0$$

all satisfy $\mathsf{H}_\ell(T/2): x^{(2)} = m$, and at $t = T$ reach prescribed levels:

for $\ell = N+1, \dots, 2N-1$ impose

$$c_\ell \frac{T}{2} + b_\ell = m, \qquad c_\ell\, T + b_\ell = y^{(2)}_{\ell+1-N},$$

and for $\ell = 2N$ impose

$$c_{2N} \frac{T}{2} + b_{2N} = m, \qquad c_{2N}\, T + b_{2N} = y^{(2)}_N + 1.$$

By definition of m in (3.5.9), these hyperplanes are never active on $[0, T/2]$, hence the second group of neurons does not affect Step 1.

Now fix $i \in [N]$ and consider the trajectory starting from x_i. On $[T/2, T]$ we have $x^{(2)}(t) \equiv y^{(2)}_i$ by **Step 1**, so the first coordinate satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\big[\Phi^\theta_t(x_i)\big]^{(1)} = \sum_{\ell=N+1}^{2N} w^{(1)}_\ell\, (-y^{(2)}_i + c_\ell t + b_\ell)_+.$$

Using the ordering (3.5.10) and the fact that $c_\ell t + b_\ell \le c_\ell T + b_\ell$ for $t \in [T/2, T]$, we get that, for $\ell = N+1, \dots, N+i-1$,

$$c_\ell t + b_\ell \le c_\ell T + b_\ell = y^{(2)}_{\ell+1-N} \le y^{(2)}_i,$$

so $(-y^{(2)}_i + c_\ell t + b_\ell)_+ \equiv 0$ on $[T/2, T]$. Hence only $\ell \ge N + i$ contribute and

$$\big[\Phi^\theta_T(x_i)\big]^{(1)} = x^{(1)}_i + \sum_{j=i}^{N} w^{(1)}_{N+j} \int_{T/2}^{T} (-y^{(2)}_i + c_{N+j} t + b_{N+j})_+ \,\mathrm{d}t.$$

Define, for $1 \le i \le j \le N$,

$$\widetilde{M}_{i,j} := \int_{T/2}^{T} (-y^{(2)}_i + c_{N+j} t + b_{N+j})_+ \,\mathrm{d}t.$$

By construction, $\widetilde{M}_{i,j} > 0$ for all $1 \le i \le j \le N$, and in particular $\widetilde{M}_{i,i} > 0$.

Imposing the remaining conditions $\big[\Phi^\theta_T(x_i)\big]^{(1)} = y^{(1)}_i$ yields the upper-triangular system

$$y^{(1)}_i - x^{(1)}_i = \sum_{j=i}^{N} w^{(1)}_{N+j}\, \widetilde{M}_{i,j}, \qquad i \in [N],$$

which has a unique solution since the diagonal entries $\widetilde{M}_{i,i}$ are strictly positive. This completes the construction with $p = 2N$ and proves the UIP.

□

Remark 3.5.4. Theorem 3.3.7 extends to any neural ODE of the form

$$\dot{x}(t) = \sum_{j=1}^{p} w_j \big(a_j \cdot x(t) + f_j(t) + b_j\big)_+, \qquad t \in [0,T], \tag{3.5.11}$$

provided each f_j is strictly monotone. Geometrically, the hyperplane velocities then vary over time rather than remain constant. We emphasize, however, that the choice $f_j(t) = c_j t$ is the simplest.

3.5.2 Proofs of Subsection 3.3.2

Proofs of Subsection 3.3.2

Proof of Proposition 3.3.8. We carry out a similar analysis to the one in **Section 2.3**, where we estimated the probability of requiring k hyperplanes to separate two sets of N points.

We seek a direction $a \in \mathbb{R}^d$ such that the projections of the pairs $\{(x_i, y_i)\}_{i=1}^N$ onto a form pairwise disjoint intervals (this ensures Assumption 2 holds). Let Z_k be the event that the k-th canonical basis vector e_k satisfies this condition.

Consider the event Z_1. Since the coordinates are i.i.d. with a continuous distribution, all $(2N)!$ relative orderings of the $2N$ projected points are equally likely. A favorable configuration occurs if and only if the elements of each pair appear adjacently (forming N disjoint blocks). There are $N!$ ways to order these blocks and 2^N internal permutations (swapping x_i, y_i) within them. Thus:

$$p_N := \mathbb{P}(Z_1) = \frac{N!\,2^N}{(2N)!}.$$

Using Stirling's approximation, we obtain

$$p_N \sim \frac{\sqrt{2\pi N}(N/e)^N 2^N}{\sqrt{4\pi N}(2N/e)^{2N}} = \frac{1}{\sqrt{2}} \left(\frac{e}{2N}\right)^N.$$

By the assumption of i.i.d. coordinates, the events $Z_1, \dots, Z_d$ are independent. The probability that Assumption 2 holds is at least the probability that one of these canonical projections works:

$$\mathbb{P}(\text{Assumption 2 holds}) \;\geq\; \mathbb{P}\left(\bigcup_{k=1}^{d} Z_k\right) = 1 - (1 - p_N)^d.$$

Substituting the asymptotic for p_N and absorbing the constant factor into $\gtrsim$ yields the claim. □

Proof of Corollary 3.3.9. Let $a \in \mathbb{S}^{d-1}$, τ, and $\beta_1 < \cdots < \beta_{N+1}$ be as in Assumption 2. After a rotation, we may assume $a = e_1$, and after relabeling we may assume $\tau = \mathsf{Id}$. Set $p = N$ and choose

$$a_i = e_1, \qquad b_i := -\beta_i, \qquad i \in [N].$$

Then $b_1 > \cdots > b_{N+1}$, and Assumption 2 reads

$$x_i^{(1)} \in (-b_i, -b_{i+1}) \qquad \text{and} \qquad y_i^{(1)} \in (-b_i, -b_{i+1}), \qquad i \in [N].$$

In particular, $x_i^{(1)} + b_i > 0$ and $y_i^{(1)} + b_i > 0$ for all i.

We define $w_1, \dots, w_N \in \mathbb{R}^d$ inductively so that the autonomous vector field

$$v_\theta(x) = \sum_{i=1}^{N} w_i \,(x^{(1)} + b_i)_+$$

satisfies $\Phi_T^\theta(x_i) = y_i$ for every $i \in [N]$.

Base step ($i = 1$). On the strip $(-b_1, -b_2)$, only the first neuron is active, so the dynamics for the trajectory starting at x_1 reduce to

$$\dot{x} = w_1 \, (x^{(1)} + b_1).$$

Since $x_1^{(1)} + b_1 > 0$ and $y_1^{(1)} + b_1 > 0$, Lemma 3.5.5 below yields a choice of w_1 such that the corresponding solution satisfies $x(T) = y_1$.

Induction step. Assume that $w_1, \dots, w_{n-1}$ are chosen so that $\Phi_T^\theta(x_i) = y_i$ for all $i \in [n-1]$. For any trajectory with $x^{(1)} < -b_n$, the term $(x^{(1)} + b_n)_+$ vanishes; hence, adding the n-th neuron does not affect the dynamics in the first $n-1$ strips. In particular, the already controlled pairs remain controlled.

To steer x_n to y_n, consider the drift generated by the first $n-1$ neurons,

$$D(x) \coloneqq \sum_{i=1}^{n-1} w_i \, (x^{(1)} + b_i)_+.$$

On the strip $(-b_n, -b_{n+1})$, we have $x^{(1)} > -b_n$ and therefore $(x^{(1)} + b_i)_+ = x^{(1)} + b_i$ for every $i \leq n-1$. Thus, for the trajectory starting at x_n (which we will keep inside this strip), the controlled dynamics take the form

$$\dot{x} = D(x) + w_n \, (x^{(1)} + b_n).$$

By Lemma 3.5.6 below, we can choose w_n so that the corresponding solution satisfies $x(T) = y_n$.

Strip invariance. For each n, while the trajectory stays in $(-b_n, -b_{n+1})$ the first coordinate solves a scalar affine ODE, hence it is monotone. Since both endpoints $x_n^{(1)}$ and $y_n^{(1)}$ lie in $(-b_n, -b_{n+1})$, the entire path $t \mapsto x^{(1)}(t)$ remains in that interval. Therefore, the active set of neurons does not change along the trajectory, and the above reduced equations are valid on $[0, T]$.

Proceeding up to $n = N$ defines $\theta = (w_i, a_i, b_i)_{i=1}^N$ and yields $\Phi_T^\theta(x_i) = y_i$ for all $i \in [N]$. □

Lemma 3.5.5. *Let $x_1, y_1 \in \mathbb{R}^d$ with $x_1 \neq y_1$. For any $T > 0$ and $b \in \mathbb{R}$ such that $x_1^{(1)} + b > 0$ and $y_1^{(1)} + b > 0$, there exists a unique $w \in \mathbb{R}^d$ such that the solution of*

$$\dot{x} = w \, (x^{(1)} + b)_+, \qquad x(0) = x_1,$$

satisfies $x(T) = y_1$ (and remains in the half-space $\{x^{(1)} + b > 0\}$).

Proof of Lemma 3.5.5. Since $x_1^{(1)} + b > 0$, we look for a solution that stays in $\{x^{(1)} + b > 0\}$; then $(x^{(1)} + b)_+ = x^{(1)} + b$ and the dynamics are $\dot{x} = w \, (x^{(1)} + b)$.

Set $z(t) \coloneqq x^{(1)}(t) + b$. Then $\dot{z}(t) = w^{(1)} z(t)$ and

$$z(t) = (x_1^{(1)} + b) e^{w^{(1)} t}.$$

Imposing $z(T) = y_1^{(1)} + b$ gives the unique choice

$$w^{(1)} = \frac{1}{T} \log \left(\frac{y_1^{(1)} + b}{x_1^{(1)} + b} \right).$$

In particular, $z(t) > 0$ for all $t \in [0,T]$, so the trajectory indeed stays in $\{x^{(1)} + b > 0\}$.

For $k \in \{2, \dots, d\}$ we have $\dot{x}^{(k)}(t) = w^{(k)} z(t)$, hence

$$x^{(k)}(T) = x_1^{(k)} + w^{(k)} \int_0^T z(t)\,\mathrm{d}t.$$

If $w^{(1)} \neq 0$, then

$$\int_0^T z(t)\,\mathrm{d}t = (x_1^{(1)} + b)\frac{e^{w^{(1)}T} - 1}{w^{(1)}} = \frac{y_1^{(1)} - x_1^{(1)}}{w^{(1)}},$$

and therefore

$$w^{(k)} = \frac{y_1^{(k)} - x_1^{(k)}}{y_1^{(1)} - x_1^{(1)}}\, w^{(1)}.$$

If $w^{(1)} = 0$ (equivalently $y_1^{(1)} = x_1^{(1)}$), then $z(t) \equiv x_1^{(1)} + b$ and

$$w^{(k)} = \frac{y_1^{(k)} - x_1^{(k)}}{T\,(x_1^{(1)} + b)}.$$

This determines w uniquely and yields $x(T) = y_1$. □

Lemma 3.5.6. *Let $n \in \{2, \dots, N\}$ and assume $w_1, \dots, w_{n-1} \in \mathbb{R}^d$ and $b_1 > \cdots > b_{n-1}$ are fixed. Let $x_n, y_n \in \mathbb{R}^d$ with $x_n \neq y_n$, and choose $b_n \in \mathbb{R}$ such that*

$$b_n < b_{n-1}, \qquad x_n^{(1)} + b_n > 0, \qquad y_n^{(1)} + b_n > 0.$$

Set

$$D(x) \coloneqq \sum_{i=1}^{n-1} w_i\,(x^{(1)} + b_i)_+.$$

Then, for any $T > 0$, there exists $w_n \in \mathbb{R}^d$ such that the solution of

$$\dot{x} = D(x) + w_n\,(x^{(1)} + b_n)_+, \qquad x(0) = x_n,$$

satisfies $x(T) = y_n$.

Proof of Lemma 3.5.6. Let

$$\mathsf{H}_n^+ \coloneqq \{x \in \mathbb{R}^d \mid\ x^{(1)} + b_n > 0\}.$$

We look for a solution that stays in H_n^+; this will be verified a posteriori. On H_n^+ we have $(x^{(1)} + b_n)_+ = x^{(1)} + b_n$, and since $b_i > b_n$ for all $i \leq n-1$, also $(x^{(1)} + b_i)_+ = x^{(1)} + b_i$. Therefore, on H_n^+,

$$D(x) = \sum_{i=1}^{n-1} w_i\,(x^{(1)} + b_i) = S_{n-1}\,x^{(1)} + C_{n-1}, \qquad S_{n-1} \coloneqq \sum_{i=1}^{n-1} w_i, \quad C_{n-1} \coloneqq \sum_{i=1}^{n-1} w_i b_i,$$

and the dynamics become

$$\dot{x} = (S_{n-1} + w_n)\,x^{(1)} + (C_{n-1} + w_n b_n). \tag{3.5.12}$$

Let $u(t) \coloneqq x^{(1)}(t)$. Fix $z \in \mathbb{R}$ and set $w_n^{(1)} = z$ (the other components of w_n will be chosen later). Then u solves the scalar affine ODE

$$\dot{u} = \alpha(z)\, u + \beta(z), \qquad u(0) = x_n^{(1)},$$

with

$$\alpha(z) \coloneqq S_{n-1}^{(1)} + z, \qquad \beta(z) \coloneqq C_{n-1}^{(1)} + z\, b_n.$$

Hence, for $\alpha(z) \neq 0$,

$$u_z(T) = e^{\alpha(z)T}\, x_n^{(1)} + \frac{\beta(z)}{\alpha(z)}\big(e^{\alpha(z)T} - 1\big), \tag{3.5.13}$$

while for $\alpha(z) = 0$ one has $u_z(T) = x_n^{(1)} + \beta(z)\, T$. Using $\lim_{\alpha \to 0} \frac{e^{\alpha T}-1}{\alpha} = T$, the expression (3.5.13) extends continuously to $\alpha = 0$. Therefore, $z \mapsto u_z(T)$ is continuous on $\mathbb{R}$.

We now compute its limits. Rewrite (3.5.13) as

$$u_z(T) = (x_n^{(1)} + b_n)e^{\alpha(z)T} - b_n + \big(C_{n-1}^{(1)} - b_n S_{n-1}^{(1)}\big)\, \frac{e^{\alpha(z)T} - 1}{\alpha(z)}.$$

As $z \to +\infty$, we have $\alpha(z) \to +\infty$ and the last term is $o(e^{\alpha(z)T})$, hence

$$u_z(T) \to +\infty \qquad \text{since } x_n^{(1)} + b_n > 0.$$

As $z \to -\infty$, we have $\alpha(z) \to -\infty$, hence $e^{\alpha(z)T} \to 0$ and $\frac{e^{\alpha(z)T}-1}{\alpha(z)} \to 0$, which gives

$$u_z(T) \to -b_n.$$

Since $y_n^{(1)} + b_n > 0$, we have $-b_n < y_n^{(1)}$, so

$$\lim_{z \to -\infty} \big(u_z(T) - y_n^{(1)}\big) = -b_n - y_n^{(1)} < 0, \qquad \lim_{z \to +\infty} \big(u_z(T) - y_n^{(1)}\big) = +\infty.$$

By the intermediate value theorem, there exists $z_* \in \mathbb{R}$ such that $u_{z_*}(T) = y_n^{(1)}$. From now on we fix $w_n^{(1)} \coloneqq z_*$ and denote the corresponding solution simply by $u(t)$.

The solution of $\dot{u} = \alpha u + \beta$ is

$$u(t) = -\frac{\beta}{\alpha} + \left(x_n^{(1)} + \frac{\beta}{\alpha}\right) e^{\alpha t} \quad (\alpha \neq 0), \qquad \text{or} \qquad u(t) = x_n^{(1)} + \beta t \quad (\alpha = 0),$$

so u is monotone on $[0, T]$ (indeed $u'(t)$ has constant sign). Since $u(0) = x_n^{(1)}$ and $u(T) = y_n^{(1)}$ both satisfy $u(\cdot) > -b_n$, it follows that

$$u(t) + b_n > 0 \qquad \text{for all } t \in [0, T].$$

In particular, the trajectory remains in H_n^+, so the reduction to (3.5.12) is consistent, and

$$I_n \coloneqq \int_0^T \big(u(t) + b_n\big)\, \mathrm{d}t > 0.$$

For each $k \in \{2, \dots, d\}$, taking the k-th component in (3.5.12) yields

$$\dot{x}^{(k)} = S_{n-1}^{(k)}\, u(t) + C_{n-1}^{(k)} + w_n^{(k)}\, (u(t) + b_n).$$

Integrating on $[0,T]$ we obtain

$$x^{(k)}(T) = x_n^{(k)} + \int_0^T \left(S_{n-1}^{(k)} u(t) + C_{n-1}^{(k)}\right) \mathrm{d}t + w_n^{(k)}\, I_n.$$

Since $I_n > 0$, we can choose

$$w_n^{(k)} \coloneqq \frac{y_n^{(k)} - x_n^{(k)} - \int_0^T \left(S_{n-1}^{(k)} u(t) + C_{n-1}^{(k)}\right) \mathrm{d}t}{I_n}, \qquad k = 2, \dots, d,$$

so that $x^{(k)}(T) = y_n^{(k)}$. With the already fixed $w_n^{(1)}$, this defines $w_n \in \mathbb{R}^d$ and yields $x(T) = y_n$. □

Proofs of Subsection 3.3.2

Proof of Lemma 3.3.10. For a single ridge function $x \mapsto \sigma(a \cdot x + b)$, its distributional Fourier transform is supported on the line $L = \{\lambda a \mid \lambda \in \mathbb{R}\} \subset \mathbb{R}^d$. By linearity of the Fourier transform,

$$\operatorname{supp}(\widehat{g}) \subseteq \bigcup_{j=1}^{p} L_j, \qquad L_j \coloneqq \{\lambda a_j \mid \lambda \in \mathbb{R}\}.$$

Since $d \geq 2$, the set $\bigcup_{j=1}^{p} L_j$ has null Lebesgue measure in $\mathbb{R}^d$. If $g \in L^2(\mathbb{R}^d)$, then $\widehat{g} \in L^2(\mathbb{R}^d)$ by Plancherel Theorem. An L^2 function supported on a set of measure zero must vanish almost everywhere, hence $\widehat{g} = 0$ a.e., so $g \equiv 0$, a contradiction. □

The proof of Theorem 3.3.11 requires the following preliminary lemma.

Lemma 3.5.7. *Fix $d \geq 2$. Let $v_1 : \mathbb{R}^d \to \mathbb{R}^d$ be of the form ($\mathcal{V}_{\mathrm{a}}^{+}$). Let $u_1, \dots, u_d \in \mathbb{R}^d$ be linearly independent, and let $\alpha_i < \beta_i$ be real numbers for each $i \in [d]$. For any $\delta_i > 0$ such that $\alpha_i + \delta_i < \beta_i - \delta_i$, define the sets*

$$K \coloneqq \{x \in \mathbb{R}^d \mid \alpha_i \leq u_i \cdot x \leq \beta_i \ \forall i \in [d]\} \qquad \text{and} \qquad \tilde{K} \coloneqq \{x \in \mathbb{R}^d \mid \alpha_i + \delta_i \leq u_i \cdot x \leq \beta_i - \delta_i \ \forall i \in [d]\}.$$

Then, there exists $v_2 : \mathbb{R}^d \to \mathbb{R}^d$ of the form ($\mathcal{V}_{2\mathrm{a}}^{+}$) such that $v_2 \equiv v_1$ on $\tilde{K}$ and $v_2 \equiv 0$ on $\mathbb{R}^d \setminus K$.

Proof of Lemma 3.5.7. Let $\delta_{\min} \coloneqq \min_{i \in [d]} \delta_i > 0$. We define the barrier function $B : \mathbb{R}^d \to \mathbb{R}$ by

$$B(x) \coloneqq \sum_{i=1}^{d} \Big((\alpha_i + \delta_i - u_i \cdot x)_+ + (u_i \cdot x - (\beta_i - \delta_i))_+ \Big).$$

By definition, if $x \in \tilde{K}$, all arguments inside the ReLUs are non-positive, yielding $B(x) = 0$. Conversely, if $x \notin K$, then for some index j, either $u_j \cdot x < \alpha_j$ or $u_j \cdot x > \beta_j$. In either case, the corresponding ReLU term is strictly greater than δ_j, ensuring that $B(x) > \delta_{\min} > 0$ on $\mathbb{R}^d \setminus K$.

Write $v_1(x) = \sum_{m=1}^{p} v_m L_m(x)$, where $L_m(x) \coloneqq (c_m \cdot x + d_m)_+$ and $v_m \in \mathbb{R}^d$. Since $u_1, \dots, u_d$ form a basis of $\mathbb{R}^d$, the function $B(x)$ grows linearly to $+\infty$ as $\|x\| \to \infty$ in any direction. Because each $L_m(x)$ grows at most linearly, the ratio $L_m(x)/B(x)$ is globally bounded on $\mathbb{R}^d \setminus K$. Thus, we can choose a scalar $\kappa > 0$ large enough so that

$$L_m(x) \leq \kappa B(x) \quad \text{for all } x \notin K \text{ and all } m \in [p].$$

Finally, define

$$v_2(x) \coloneqq \sum_{m=1}^{p} v_m \left(L_m(x) - \kappa B(x)\right)_+.$$

If $x \in \tilde{K}$, then $B(x) = 0$, which implies $v_2(x) = \sum_m v_m L_m(x) = v_1(x)$. If $x \notin K$, the bound $L_m(x) \leq \kappa B(x)$ guarantees that every ReLU is inactive, yielding $v_2(x) = 0$. Because $B(x)$ is a linear combination of ReLUs of affine functions, then v_2 is representable in the form ($\mathcal{V}_{2a}^+$). □

Remark 3.5.8. This result extends immediately to any pair of strictly nested convex polytopes: one replaces the defining inequalities of K in the construction of B with the affine constraints defining its facets.

We are now ready to prove Theorem 3.3.11.

Proof of Theorem 3.3.11. Fix $T > 0$ and an admissible dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$ satisfying (3.3.3). For each $i \in [N]$, let $u_i \coloneqq (y_i - x_i)/T$. The linear curve $\gamma_i(t) \coloneqq x_i + tu_i$ for $t \in [0, T]$ traces exactly the straight segment $S_i \coloneqq [x_i, y_i]$. Since these segments are compact and pairwise disjoint, the minimum distance between them is strictly positive:

$$\delta \coloneqq \min_{i \neq j} \operatorname{dist}(S_i, S_j) > 0.$$

We construct disjoint neighborhoods around each S_i as follows. For each i, fix an orthonormal basis $(e_{i,1}, \ldots, e_{i,d})$ with $e_{i,1} = \frac{y_i - x_i}{\|y_i - x_i\|}$. We define for each $\rho > 0$,

$$K_{i,\rho} \coloneqq \Big\{ x \in \mathbb{R}^d \mid -\rho \leq (x - x_i) \cdot e_{i,1} \leq \|y_i - x_i\| + \rho \quad \text{and} \quad |(x - x_i) \cdot e_{i,k}| \leq \rho \quad \text{for all } k \geq 2 \Big\}.$$

Set $r \coloneqq \delta/(4\sqrt{d})$. By definition, every $x \in S_i$ satisfies $(x - x_i) \cdot e_{i,1} = 0$ and $|(x - x_i) \cdot e_{i,k}| = 0$. This trivially yields $S_i \subset K_{i,r/2} \subset \operatorname{int}(K_{i,r})$.

Moreover, the maximum distance from any $x \in K_{i,r}$ to S_i is $\leqslant \sqrt{r^2 + (d-1)r^2} = r\sqrt{d} = \delta/4$, so $K_{i,r}$ lies entirely within the $\delta/4$-neighborhood of S_i. By the triangle inequality, for any $i \neq j$,

$$\operatorname{dist}(K_{i,r}, K_{j,r}) \geq \operatorname{dist}(S_i, S_j) - 2(\delta/4) \geq \delta - \delta/2 = \delta/2 > 0.$$

Hence, $K_{1,r}, \ldots, K_{N,r}$ are pairwise disjoint.

We now construct the vector field. For each $i \in [N]$, we apply Lemma 3.5.7 to the constant field $g_i(x) \equiv u_i$ (which trivially satisfies the form ($\mathcal{V}_a^+$)), setting the inner region to $K_{i,r/2}$ and the outer region to $K_{i,r}$. This yields a localized field v_i of the form ($\mathcal{V}_{2a}^+$) such that

$$v_i(x) = u_i \quad \text{for all } x \in K_{i,r/2}, \qquad v_i(x) = 0 \quad \text{for all } x \in \mathbb{R}^d \setminus K_{i,r}.$$

Define the global vector field $v_\theta(x) \coloneqq \sum_{i=1}^N v_i(x)$. Since the neighborhoods $K_{1,r}, \ldots, K_{N,r}$ are pairwise disjoint, for any fixed $i \in [N]$ we have

$$v_\theta(x) = v_i(x) + \sum_{j \neq i} \mathbf{0} = u_i \qquad \text{for all } x \in K_{i,r/2}.$$

Recall that the trace of the trajectory is $\gamma_i([0, T]) = S_i \subset K_{i,r/2}$. Therefore, the ODE dynamics exactly follow the constant velocity: $\dot{\gamma}_i(t) = v_\theta(\gamma_i(t)) = u_i$. Integrating this yields $\Phi_T^\theta(x_i) = \gamma_i(T) = y_i$. Since this holds for all $i \in [N]$, we deduce that Φ_T^θ perfectly interpolates the dataset.

Finally, we determine the architecture. By the explicit construction in Lemma 3.5.7, since $g_i(x) \equiv u_i$ relies on a single constant term, the localized field takes the form

$$v_i(x) = u_i\big(1 - \kappa_i B_i(x)\big)_+,$$

where $B_i(x)$ is the barrier function. Because $K_{i,r}$ is defined by bounding the d projections onto the basis $e_{i,k}$, the barrier B_i is a sum of exactly $2d$ ReLU terms (one for each upper and lower bound). Consequently, the global field v_θ exactly matches the ansatz $(\mathcal{V}_{2a}^+)$ with $p = 2d$ and $q = N$. □

Proofs of Subsection 3.3.2

First, we prove Proposition 3.3.13 with an inductive argument.

Proof of Proposition 3.3.13. For each $i \in [N]$ choose a smooth injective curve $\gamma_i : [0, T] \to (-R, R)^d$ with $\gamma_i(0) = x_i$ and $\gamma_i(T) = y_i$, and such that the traces $\Gamma_i := \gamma_i([0, T])$ are pairwise disjoint. (Since any connected open set in $\mathbb{R}^d$ is path-connected if $d \geq 2$, one can iteratively connect each pair by a continuous path that avoids the previous traces; finally smooth the result without changing endpoints.)

Since each γ_i is a smooth embedding, $u_i(\gamma_i(t)) := \dot\gamma_i(t)$ defines a smooth vector field on the submanifold Γ_i. We pick pairwise disjoint open neighborhoods U_i of Γ_i (contained in $(-R, R)^d$) and use a standard extension argument (e.g., via tubular neighborhoods and bump functions) to obtain smooth fields $v_i \in C_c^\infty(U_i; \mathbb{R}^d)$ such that

$$v_i(x) = u_i(x) = \dot\gamma_i(t) \quad \text{for all } x = \gamma_i(t) \in \Gamma_i.$$

Finally, defining $v_\star := \sum_{i=1}^N v_i$, we obtain a field in $C_c^\infty(\mathbb{R}^d)$ that restricts to $\dot\gamma_i$ on each trace. By uniqueness of solutions, the flow of $v_\star$ starting at x_i must follow γ_i, satisfying $\Phi_T^\star(x_i) = y_i$.

□

Proof of Theorem 3.3.14. Without loss of generality, we assume the targets are distinct. If some targets coincide, we may perturb them by an arbitrarily small amount; by the triangle inequality, this only affects the final constant.

Let $v_\star$ be the vector field given by Proposition 3.3.13, and let $(\Phi_t^\star)_{t\in[0,T]}$ denote its flow. Since the trajectories are bounded, there exists $R > 0$ such that

$$\bigcup_{t\in[0,T]} \Phi_t^\star(\{x_1, \dots, x_N\}) \subset [-R, R]^d.$$

We approximate $v_\star$ component-wise on $[-R, R]^d$. For $\ell \in [d]$, each component $v_\star^{(\ell)}$ satisfies the regularity assumptions of Theorem 3.5.9 (up to rescaling of R), which yields an affine term and a sum of m ReLU units (for any $m \geq 1$) satisfying (3.5.16).

Since $(\mathcal{V}_a^+)$ allows for signed output weights, using $a_0 \cdot x = \sum_{r=1}^d a_0^{(r)}((e_r \cdot x)_+ - ((-e_r) \cdot x)_+)$, the affine part can be exactly represented by $2d + 1$ ReLU units. Absorbing these, we obtain

$$s_\ell(x) = \sum_{k=1}^{m+2d+1} c_{\ell,k}\,(\alpha_{\ell,k} \cdot x + \beta_{\ell,k})_+,$$

where $\alpha_{\ell,k} \in \mathbb{R}^d$ and $c_{\ell,k}, \beta_{\ell,k} \in \mathbb{R}$. Therefore

$$\sup_{x\in[-R,R]^d} \left|v_\star^{(\ell)}(x) - s_\ell(x)\right| \le C_\ell \frac{\sqrt{d+\log(2+m)}}{m^{\frac{1}{2}+\frac{1}{d}}}, \qquad \mathrm{Lip}(s_\ell) \le L_\ell, \tag{3.5.14}$$

with $C_\ell, L_\ell > 0$ independent of m.

Let $m \coloneqq \lfloor p/d \rfloor - (2d+1)$ be the number of neurons allocated to the non-affine part of each component. (We assume p is large enough so $m \ge 1$; for small p, (3.3.4) holds trivially). The total number of neurons used is $d(m+2d+1) \le p$. We define

$$v_\theta(x) \coloneqq \sum_{\ell=1}^{d} s_\ell(x) e_\ell = \sum_{\ell=1}^{d} \sum_{k=1}^{m+2d+1} c_{\ell,k} e_\ell \left(\alpha_{\ell,k} \cdot x + \beta_{\ell,k}\right)_+.$$

By mapping the indices (ℓ, k) to a global index $j \in [p]$, this takes the form stated in the theorem:

$$v_\theta(x) = \sum_{j=1}^{p} w_j \left(a_j \cdot x + b_j\right)_+,$$

(padding with zero weights if strictly less than p). Thus (3.5.14) yields:

$$\sup_{x\in[-R,R]^d} \|v_\star(x) - v_\theta(x)\| \le \sqrt{d} \max_{1\leqslant\ell\leqslant d} \sup_{x\in[-R,R]^d} \left|v_\star^{(\ell)}(x) - s_\ell(x)\right| \le C_{R,v_\star} \frac{\sqrt{d+\log(2+p)}}{p^{\frac{1}{2}+\frac{1}{d}}}. \tag{3.5.15}$$

Moreover, $\mathrm{Lip}(v_\theta) \le \sqrt{d} \max_\ell L_\ell =: L$, independent of p.

Fix $x_0 \in \{x_1, \dots, x_N\}$ and define $e(t) \coloneqq \|\Phi_t^\star(x_0) - \Phi_t^\theta(x_0)\|$. Using the integral form of the solution:

$$\begin{aligned} e(t) &\le \int_0^t \|v_\star(\Phi_s^\star(x_0)) - v_\theta(\Phi_s^\star(x_0))\| \,\mathrm{d}s + \int_0^t \|v_\theta(\Phi_s^\star(x_0)) - v_\theta(\Phi_s^\theta(x_0))\| \,\mathrm{d}s \\ &\le t \sup_{x\in[-R,R]^d} \|v_\star(x) - v_\theta(x)\| \;+\; L \int_0^t e(s) \,\mathrm{d}s. \end{aligned}$$

Applying Grönwall's lemma yields $e(T) \le T e^{LT} \sup \|v_\star - v_\theta\|$. Since $\Phi_T^\star(x_i) = y_i$, combining this with (3.5.15) completes the proof. □

The proof of Theorem 3.3.14 relies on [KB18, Theorem 2]. We state it here for convenience.

Theorem 3.5.9. *Let $D = [-1,1]^d$. Suppose $f : D \to \mathbb{R}$ admits a Fourier representation*

$$f(x) = \int_{\mathbb{R}^d} e^{i\, x\cdot\omega}\, \hat{f}(\omega) \,\mathrm{d}\omega,$$

(where $\hat{f}$ is the Fourier transform of f) such that

$$v_{f,2} \;\coloneqq\; \int_{\mathbb{R}^d} \|\omega\|_1^2 \, |\hat{f}(\omega)| \,\mathrm{d}\omega \;<\; \infty.$$

Then for every $m \ge 1$ there exists a function $f_m : D \to \mathbb{R}$ of the form

$$f_m(x) = b_0 + a_0 \cdot x \;+\; \frac{v}{m} \sum_{k=1}^{m} b_k \left(a_k \cdot x - t_k\right)_+,$$

with parameters satisfying

$$b_k \in [-1,1], \qquad \|a_k\|_1 = 1, \qquad t_k \in [0,1], \qquad b_0 = f(0), \qquad a_0 = \nabla f(0), \qquad v \le 2\, v_{f,2},$$

and a universal constant $c > 0$ such that

$$\sup_{x \in D} |f(x) - f_m(x)| \;\le\; c\, v_{f,2}\, \frac{\sqrt{d + \log m}}{m^{\frac{1}{2}+\frac{1}{d}}}. \tag{3.5.16}$$

Moreover, the Lipschitz constant of f_m on D is bounded independently of m. Indeed, for $x, y \in D$,

$$|f_m(x) - f_m(y)| \le (\|a_0\| + v)\, \|x - y\|,$$

where we used that $\|a_k\| \le \|a_k\|_1 = 1$. Hence

$$\mathrm{Lip}(f_m) \le \|a_0\| + v \le \|\nabla f(0)\| + 2 v_{f,2},$$

which is independent of m (both terms only depend on f).

3.5.3 Proofs of Subsection 3.3.3

Proof of Theorem 3.3.18. Ultimately, $\theta = (W, b)$ takes the form

$$\theta(t) = \sum_{k=1}^{4N+3} \theta_k\, 1_{[T_{k-1}, T_k]}(t),$$

where $\theta_k = (W_k, b_k) \in \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d$ are constant and determined throughout the proof, whereas $T_k \coloneqq \frac{kT}{4N+3}$. For each $i \in [N]$, we write $x_i(t) \coloneqq \Phi_t^\theta(x_i)$.

Step 1.

We begin by taking $\theta_1 = (W_1, b_1)$ with

$$W_1 = 0, \qquad b_1 = \beta_1 \mathbf{1}$$

for $\beta_1 > 0$. Then for all $i \in [N]$,

$$x_i(t) = x_i(0) + (t - 0)\, \beta_1 \mathbf{1} \qquad \text{for } t \in [0, T_1].$$

Taking β_1 large enough, we find

$$x_i^{(k)}(T_1) > 0 \qquad \text{for all } (i,k) \in [N] \times [d].$$

We now take $\theta_2 = (W_2, b_2)$ with

$$W_2 = \alpha_1 \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ v_1 & \dots & v_k & \dots & v_d \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}, \qquad b_2 = 0,$$

where $\alpha_1 > 0$ and $v = (v_1, \dots, v_d) \in (\mathbb{R}_{>0})^d$ is such that $\langle v, x_i(T_1) - x_j(T_1)\rangle \neq 0$ for all $i \neq j$. Then, since $x_i(t) \in (\mathbb{R}_{>0})^d$ on $[T_1, T_2]$, we have $x_i(t)_+ = x_i(t)$ and

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(x_i^{(2)} - x_j^{(2)}\right)(t) = \alpha_1 \langle v,\, x_i(t) - x_j(t)\rangle \qquad \text{for } t \in [T_1, T_2].$$

In particular, for each fixed $i \neq j$, the right-hand side is nonzero at $t = T_1$, hence $x_i^{(2)} - x_j^{(2)}$ has a sign whenever $t - T_1 \ll 1$. Taking α_1 sufficiently small and rescaling time appropriately, we deduce

$$x_i^{(2)}(T_2) \neq x_j^{(2)}(T_2), \qquad \text{for all } i \neq j, \tag{3.5.17}$$

or, equivalently, $\Phi_{\theta_2}^{T_2}\left(\Phi_{\theta_1}^{T_1}(x_i)\right)_2 \neq \Phi_{\theta_2}^{T_2}\left(\Phi_{\theta_1}^{T_1}(x_j)\right)_2$.

Consider the backward equation

$$\begin{cases} \dot{y}(t) = \overline{W}(t)(y(t))_+ + \overline{b}(t), & t \in [T_{4N+1}, T], \\ y(T) = y_0, \end{cases}$$

and note that y also solves the forward equation ($\mathcal{V}_\ell^+$) for $\theta(t) = -\overline{\theta}(T-t)$.

Take $\theta_{4N+3} = (W_{4N+3}, b_{4N+3})$ with

$$W_{4N+3} = 0, \qquad b_{4N+3} = -\beta_2 \mathbf{1}$$

for $\beta_2 > 0$ large enough to ensure

$$y_i^{(k)} + \big(T - T_{4N+2}\big)\beta_2 > 0 \qquad \text{for all } (i,k) \in [N] \times [d].$$

Take $\theta_{4N+2} = (W_{4N+2}, b_{4N+2})$ with

$$W_{4N+2} = -\alpha_2 \begin{bmatrix} u_1 & \dots & u_k & \dots & u_d \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}, \qquad b_{4N+2} = 0,$$

where $\alpha_2 > 0$ and $u = (u_1, \dots, u_d) \in (\mathbb{R}_{>0})^d$ is such that $\langle u, y_i - y_j \rangle \neq 0$ for all $i \neq j$. Setting

$$\bar{y}_i \coloneqq \Phi_{T_{4N+2}-T_{4N+1}}^{-\theta_{4N+2}}\left(\Phi_{T_{4N+3}-T_{4N+2}}^{-\theta_{4N+3}}(y_i)\right), \qquad i \in [N],$$

and arguing just as above, by choosing α_2 sufficiently small it ensues

$$\bar{y}_i^{(1)} \neq \bar{y}_j^{(1)} \qquad \text{for all } i \neq j.$$

Step 2.

Due to (3.5.17), we can relabel all points as to have

$$x_1^{(2)}(T_2) < x_2^{(2)}(T_2) < \cdots < x_N^{(2)}(T_2). \tag{3.5.18}$$

We now show that

$$x_i^{(1)}(T_{2N+1}) = \bar{y}_i^{(1)} + c_1 \tag{3.5.19}$$

for some $c_1 \in \mathbb{R}$ and all $i \in [N]$. We argue by induction.

The base case $i = 1$ trivially holds with $\theta_3 \equiv (0,0)$ and $c_1 = x_1^{(1)}(T_3) - \bar{y}_1^{(1)}$, at time T_3. Of course, since the points don't move, the order (3.5.18) is propagated up to time T_3.

Assume (3.5.18) and (3.5.19) hold at time T_{2n-1} for all $i \in [n-1]$, for some $n \in \{2, \dots, N\}$. Take $\theta_{2n} = (W_{2n}, b_{2n})$ with

$$W_{2n} = 0, \qquad b_{2n} = -\alpha_3 \operatorname{sign}\big(x_{n-1}^{(2)}(T_{2n-1})\big)\, e_2,$$

where $\alpha_3 > 0$. Then for all $i \in [N]$, and for $t \in [T_{2n-1}, T_{2n}]$:

$$\begin{aligned} x_i^{(2)}(t) &= x_i^{(2)}(T_{2n-1}) - \alpha_3 \, (t - T_{2n-1}) \, \operatorname{sign}\big(x_{n-1}^{(2)}(T_{2n-1})\big), \\ x_i^{(k)}(t) &= x_i^{(k)}(T_{2n-1}), \qquad\qquad \text{for } k \neq 2. \end{aligned}$$

By virtue of the order (3.5.18) (which holds at time T_{2n-1} by heredity), we can choose α_3 so as to ensure

$$x_1^{(2)}(T_{2n}) < \cdots < x_{n-1}^{(2)}(T_{2n}) < 0 < x_n^{(2)}(T_{2n}). \tag{3.5.20}$$

All the while, $x_i^{(k)}(T_{2n}) = x_i^{(k)}(T_{2n-1})$ for all $k \neq 2$ and $i \in [N]$.

Now take $\theta_{2n+1} = (W_{2n+1}, b_{2n+1})$ with

$$W_{2n+1} = \frac{4N+3}{T \, x_n^{(2)}(T_{2n})} \begin{bmatrix} 0 & \omega & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}, \qquad b_{2n+1} = 0,$$

where

$$\omega = \bar{y}_n^{(1)} - x_n^{(1)}(T_{2n}) + c_1.$$

Because of (3.5.20), we see that $x_i^{(2)}(t) < 0$ for $i \leq n-1$ and $x_i^{(2)}(t) > 0$ for $i = n$ on $[T_{2n}, T_{2n+1}]$, hence only the point x_n is affected in its first coordinate. Therefore $x_n^{(1)}(T_{2n+1}) = \bar{y}_n^{(1)} + c_1$ holds, and we moreover have $x_i^{(1)}(T_{2n+1}) = x_i^{(1)}(T_{2n}) = \bar{y}_i^{(1)} + c_1$ for $i \in [n-1]$ by the induction hypothesis. All in all, we conclude that (3.5.19) holds.

Step 3.

We will now apply this argument simultaneously to all coordinates.

We relabel all points anew as to have

$$x_1^{(1)}(T_{2N+1}) < x_2^{(1)}(T_{2N+1}) < \cdots < x_N^{(1)}(T_{2N+1}). \tag{3.5.21}$$

We show that

$$x_i(T_{4N}) = \bar{y}_i + c \tag{3.5.22}$$

for some $c \in \mathbb{R}^d$ and all $i \in [N]$. We again argue by induction.

The base case trivially holds at time T_{2N+2} by taking $\theta_{2N+2} \equiv (0, 0)$ and

$$c^{(k)} = x_1^{(k)}(T_{2N+2}) - \bar{y}_1^{(k)} \qquad \text{for all } k \in [d].$$

Again, since the points don't move, the order (3.5.21) is propagated up to time T_{2N+2}.

Assume that (3.5.21) and (3.5.22) hold at time $T_{2N+2n-2}$ for all $i \in [n-1]$, for some $n \in \{2, \ldots, N\}$. Take $\theta_{2N+2n-1} = (W_{2N+2n-1}, b_{2N+2n-1})$ with

$$W_{2N+2n-1} = 0, \qquad b_{2N+2n-1} = -\alpha_4 \, \operatorname{sign}\big(x_{n-1}^{(1)}(T_{2N+2n-2})\big) \, e_1$$

for $\alpha_4 > 0$. Then for all $i \in [N]$, and for $t \in [T_{2N+2n-2}, T_{2N+2n-1}]$,

$$\begin{aligned} x_i^{(1)}(t) &= x_i^{(1)}(T_{2N+2n-2}) - \alpha_4 \, (t - T_{2N+2n-2}) \, \operatorname{sign}\big(x_{n-1}^{(1)}(T_{2N+2n-2})\big), \\ x_i^{(k)}(t) &= x_i^{(k)}(T_{2N+2n-2}), \qquad\qquad \text{for } k \neq 1. \end{aligned}$$

By virtue of (3.5.21) (which holds at time $T_{2N+2n-2}$ by heredity), we can choose α_4 so as to ensure

$$x_1^{(1)}(T_{2N+2n-1}) < \cdots < x_{n-1}^{(1)}(T_{2N+2n-1}) < 0 < x_n^{(1)}(T_{2N+2n-1}). \tag{3.5.23}$$

All the while $x_i^{(k)}(T_{2N+2n-1}) = x_i^{(k)}(T_{2N+2n-2})$ for all $k \neq 1$ and $i \in [N]$.

Now take $\theta_{2N+2n} = (W_{2N+2n}, b_{2N+2n})$ with

$$W_{2N+2n} = \frac{4N+3}{T\, x_n^{(1)}(T_{2N+2n-1})} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \omega_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{d-1} & 0 & \cdots & 0 \end{bmatrix}, \qquad b_{2N+2n} = 0,$$

where

$$\omega_{k-1} = \bar{y}_n^{(k)} - x_n^{(k)}(T_{2N+2n-1}) + c^{(k)}, \qquad k \in \{2, \dots, d\}.$$

Because of (3.5.23), for $i \leq n-1$ we have $x_i^{(1)}(t) < 0$ on $[T_{2N+2n-1}, T_{2N+2n}]$, hence $(x_i^{(1)}(t))_+ = 0$ and those points are not affected by θ_{2N+2n}. For $i = n$, we have $x_n^{(1)}(t) > 0$ on the same interval and $\dot{x}_n^{(1)} = 0$ (first row of W_{2N+2n} is zero), so the chosen W_{2N+2n} steers $x_n^{(k)}$ to $\bar{y}_n^{(k)} + c^{(k)}$ for all $k \in \{2, \dots, d\}$ in one step, while keeping $x_n^{(1)}$ fixed. Together with the induction hypothesis, this yields (3.5.22).

We can conclude the proof by taking $\theta_{4N+1} = (W_{4N+1}, b_{4N+1})$ with

$$W_{4N+1} = 0, \qquad b_{4N+1} = -\frac{4N+3}{T}\, c.$$

Then $x_i(T_{4N+1}) = \bar{y}_i$ for all $i \in [N]$. Finally, by the definition of $\bar{y}_i$ and the identity $\Phi_\Delta^{-\theta} = (\Phi_\Delta^{\theta})^{-1}$ for constant controls, applying θ_{4N+2} on $[T_{4N+1}, T_{4N+2}]$ and θ_{4N+3} on $[T_{4N+2}, T]$ yields $x_i(T) = y_i$ for all $i \in [N]$.

□

Chapter 4

Generalization

4.1 Introduction

As discussed in **Part I (Overview)**, the ultimate goal of supervised learning is to minimize the population risk $\mathcal{R}(\theta)$. Since the true data-generating distribution is unknown, models are trained to minimize the empirical risk $\mathcal{R}_N(\theta)$—computed from a finite dataset $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$—as a practical proxy. One expects that a sufficiently small training error will translate into good generalization on unseen data, a property that should ideally improve as N grows large.

In **Chapter 3**, we explored the absolute limits of empirical fitting. We established that several neural ODE families possess the property of simultaneous controllability (UIP), ensuring they are expressive enough to perfectly interpolate any admissible dataset and achieve exactly $\mathcal{R}_N(\theta) = 0$.

However, perfect data fitting without further capacity constraints does not guarantee generalization, but typically leads to severe overfitting. Rather than learning the underlying ground truth $y(\cdot)$, the flow simply memorizes the specific spatial configuration of $\mathcal{D}_N$. As shown below, this introduces geometric instabilities that degrade population performance, a phenomenon that becomes catastrophically pronounced in the large-N limit.

Interpolation does not imply generalization. To see why exact interpolation may fail to generalize, we analyze the generalization gap through the lens of optimal transport. Recalling the risk definitions from (I.3) and (I.4) and the predictor $F_\theta = g \circ \Phi_T^\theta$ from (I.2), this gap can be written as

$$\big|\mathcal{R}(\theta) - \mathcal{R}_N(\theta)\big| = \left| \int \ell\big(F_\theta(x), y(x)\big)\, \mathrm{d}\mu(x) - \int \ell\big(F_\theta(x), y(x)\big)\, \mathrm{d}\mu_N(x) \right|,$$

where $\mu_N := \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ is the empirical measure. While in general μ is simply a reference distribution and training points x_i can be arbitrary, we now specifically assume that the x_i are drawn i.i.d. from a data-generating measure $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

To quantify the discrepancy between μ and μ_N, we use the 1-Wasserstein distance

$$W_1(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|\, \mathrm{d}\pi(x, y), \qquad \text{for } \mu, \nu \in \mathcal{P}_1(\mathbb{R}^d), \tag{4.1.1}$$

where $\Pi(\mu, \nu)$ is the set of couplings—a formal treatment of Wasserstein spaces is deferred to **Chapter 5**; see (5.3.1)–(5.3.2). By the Kantorovich–Rubinstein duality (see [San15, Eq. (3.1)] or [Vil09, Thm. 5.10]), we can rewrite (4.1.1) as a supremum over 1-Lipschitz functions:

$$W_1(\mu, \nu) = \sup_{\mathrm{Lip}(f) \leqslant 1} \left(\int f\, \mathrm{d}\mu - \int f\, \mathrm{d}\nu \right). \tag{4.1.2}$$

For an illustrative bound, assume for simplicity that the loss ℓ, the readout g, and $y(\cdot)$ are sufficiently regular and that F_θ and y remain uniformly bounded on the relevant domain, so that $x \mapsto \ell(F_\theta(x), y(x))$ has Lipschitz constant controlled by $\mathrm{Lip}(\Phi_T^\theta)$ up to multiplicative constants. Rescaling (4.1.2) by this constant and evaluating at $\nu = \mu_N$ readily yields:

$$\left|\mathcal{R}(\theta) - \mathcal{R}_N(\theta)\right| \lesssim \mathrm{Lip}(\Phi_T^\theta)\, W_1(\mu, \mu_N). \tag{4.1.3}$$

It is well known that the convergence of μ_N to μ suffers from the *curse of dimensionality*: for continuous distributions, $\mathbb{E}[W_1(\mu, \mu_N)] \gtrsim N^{-1/d}$ [Dud69; FG15]. While faster rates are possible when μ concentrates on a low-dimensional set [WB19], the estimate (4.1.3) shows that flow stability is equally critical. For the interpolating controls from **Chapters 2** and **3**, this instability is the bottleneck. By Proposition 1.1.1 and Lemma 1.3.1, systems driven by vector fields of the form ($\mathcal{V}_p^+$) satisfy

$$\mathrm{Lip}(\Phi_T^\theta) \leqslant \exp\left(T \sum_{j=1}^{p} \|w_j(\cdot)\|_{L^\infty(0,T;\mathbb{R}^d)}\, \|a_j(\cdot)\|_{L^\infty(0,T;\mathbb{R}^d)} \right). \tag{4.1.4}$$

In the explicit constructions of Theorem 3.3.1, the control $\theta(\cdot)$ achieves exact interpolation by executing $O(N/p)$ switches. For fixed time horizon T and fixed width p, these constructions require parameter amplitudes that grow at least proportionally to the number of switches. Consequently, the exponent in (4.1.4) grows at least linearly with N, forcing $\mathrm{Lip}(\Phi_T^\theta)$ to explode exponentially in the large-N limit. This renders the bound (4.1.3) entirely vacuous, confirming that such controls must be viewed as *expressivity witnesses* rather than generalization guarantees.

Exact controllability yields a priori bounds. Avoiding this severe instability requires actively constraining the complexity of the learned flows (e.g., via fixed network width or parameter regularization). Suppose we are given an exact interpolating control $\theta_N^0 \in \mathcal{U}$ such that $\mathcal{R}_N(\theta_N^0) = 0$. While θ_N^0 itself might overfit, it provides a valuable theoretical certificate for regularized training. Recalling the empirical risk minimization framework from (I.5), consider a regularized predictor θ_N^λ selected via a penalty $\|\cdot\|_{\mathcal{U}}$ (not necessarily a norm) with parameter $\lambda > 0$. A straightforward comparison of the objective values evaluated at θ_N^λ and θ_N^0 immediately yields the *a priori* bounds

$$\|\theta_N^\lambda\|_{\mathcal{U}} \leqslant \|\theta_N^0\|_{\mathcal{U}}, \qquad \mathcal{R}_N(\theta_N^\lambda) \leqslant \lambda \|\theta_N^0\|_{\mathcal{U}}. \tag{4.1.5}$$

(More generally, any θ achieving $\mathcal{R}_N(\theta) \leqslant \varepsilon$ yields analogous bounds shifted by ε.) This highlights a concrete sense in which the expressivity results from **Chapter 3** inform generalization: they provide explicit certificates for the scale of the optimizer selected by regularization.

From point-level to cell-wise controllability. To overcome the limitations of exact data fitting and derive meaningful population-risk bounds, this chapter keeps the controllability approach. Standard generalization theories often rely on uniform-deviation analyses [Mar23; BG24], which bound the gap $\left|\mathcal{R}(\theta) - \mathcal{R}_N(\theta)\right|$ by computing global capacity measures—such as Rademacher complexities or covering numbers—over the restricted class $\mathcal{U}$. In heavily parameterized continuous-time flows, this worst-case approach often yields overly loose bounds.

We pursue a complementary strategy: instead of relying on global capacity bounds, we provide a data-dependent control $\theta_N = \theta_N(\mathcal{D}_N)$ whose induced flow map mimics a *nonparametric estimator*. Unlike parametric models restricted to a fixed global formula, nonparametric models allow their capacity to grow with the data, for instance by computing local spatial averages over shrinking neighborhoods.

The natural way to achieve this is by lifting the point-level interpolation to *cell-wise controllability*: we steer the flow not just at isolated points, but across entire disjoint regions. This yields bounds

that decompose into three components: (i) a statistical estimation term, (ii) a geometric boundary-layer term (inherent to approximating discontinuous proxy estimators with continuous flows), and (iii) a controllability (or expressivity) term capturing how accurately the flow can realize this proxy.

The remainder of the chapter executes this program. The next section formalizes the setup and states the main template bound. Subsequent sections estimate the statistical term for canonical partitions, such as histograms and Voronoi tessellations, and relate the controllability error to the inherent model complexity. Finally, all technical proofs are collected at the end of the chapter.

4.2 Setup and main results

We adopt the framework established in **Part I (Overview)**, specialized here to a regression problem on a compact domain. Fix $R > 0$ and $d \geqslant 2$, and let $I_R := [-R, R]^d$. Since we evaluate population performance solely via a test measure $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ supported on I_R, we can assume without loss of generality that our input space is precisely I_R. We set the label space to $\mathbb{R}^d$ and employ the standard squared loss $\ell(\hat{y}, y) = \|\hat{y} - y\|^2$.

Focusing on the *noiseless* regime, we assume an unknown ground-truth mapping $y(\cdot) \in L^2(\mu; \mathbb{R}^d)$. A training set of size N is thus given by

$$\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset I_R \times \mathbb{R}^d, \qquad \text{with} \quad y_i = y(x_i).$$

To derive quantitative rates, we assume throughout this chapter that the inputs $x_1, \dots, x_N$ are sampled i.i.d. from μ. This allows us to bound the expected population risk, where the expectation is taken with respect to the product measure $\mu^{\otimes N}$ on I_R^N (i.e., averaging over all possible realizations of the training set):

$$\mathbb{E}_{\mathcal{D}_N}\big[\mathcal{R}(\theta_N)\big] := \int_{I_R^N} \mathcal{R}\big(\theta_N(\mathcal{D}_N)\big)\, \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_N). \tag{4.2.1}$$

Nevertheless, because our analysis is fundamentally geometric, these arguments readily adapt to a deterministic-design setting by replacing probabilistic bounds with spatial descriptors of a fixed set of chosen inputs. Furthermore, while we restrict ourselves to deterministic labels, this framework naturally extends to joint distributions $\mu \in \mathcal{P}_{\mathrm{ac}}(I_R \times \mathbb{R}^d)$ to accommodate stochastic noise.

We use flow-based predictors with an identity readout, so that the predictor defined in (I.2) is $F_\theta = \Phi_T^\theta$, where Φ_T^θ is the flow generated by (CF). The *population risk* is then

$$\mathcal{R}(\theta) = \int_{I_R} \big\|\Phi_T^\theta(x) - y(x)\big\|^2 \, \mathrm{d}\mu(x) = \big\|\Phi_T^\theta - y\big\|_{L^2(\mu)}^2. \tag{4.2.2}$$

Our goal is to obtain a data-dependent control $\theta_N = \theta_N(\mathcal{D}_N)$ such that the resulting predictor achieves a small population risk $\mathcal{R}(\theta_N)$.

Main question. *Can one select, from data, a controlled flow whose population-risk decay matches the accuracy of classical nonparametric regression, up to controlled geometric and expressivity errors? In particular, how do the sample size N, the domain partitioning, and the model complexity interact quantitatively?*

Strategy

Instead of analyzing the empirical risk minimizer directly, our approach proceeds in two steps:

1. **Estimation:** Construct from $\mathcal{D}_N$ a piecewise-constant proxy $y_N(\cdot)$ that estimates $y(\cdot)$.
2. **Realization:** Design a control θ_N such that the induced flow $\Phi_T^{\theta_N}$ approximates y_N on the vast majority of the domain.

Concretely, let $\{P_k\}_{k=1}^K$ be a partition of I_R into convex cells. Let $\{r_k\}_{k=1}^K \subset \mathbb{R}^d$ be corresponding target values computed from $\mathcal{D}_N$ (e.g., via cell averages), and define the piecewise-constant estimator:

$$y_N(x) \coloneqq \sum_{k=1}^{K} r_k \, \mathbf{1}_{P_k}(x). \tag{4.2.3}$$

Given any control θ, we decompose the population error using

$$\left\|\Phi_T^\theta - y\right\|_{L^2(\mu)}^2 \;\leqslant\; 2\left\|y - y_N\right\|_{L^2(\mu)}^2 \;+\; 2\left\|y_N - \Phi_T^\theta\right\|_{L^2(\mu)}^2. \tag{4.2.4}$$

The first term is the *estimation error*, controlled via nonparametric statistics. The second is the *realization error*, quantifying how accurately the flow Φ_T^θ can reproduce the proxy y_N. Since Φ_T^θ is a homeomorphism, uniform approximation of jump discontinuities is topologically impossible. Thus, we only require Φ_T^θ to approximate y_N strictly within the interior of the cells, away from their boundaries.

For a margin $\delta > 0$, we define the trimmed cores and the boundary layer as:

$$P_k^\delta \coloneqq \big\{x \in P_k \mid \operatorname{dist}(x, \partial P_k) \geqslant \delta\big\}, \qquad \Omega_\delta \coloneqq I_R \setminus \bigcup_{k=1}^{K} P_k^\delta. \tag{4.2.5}$$

Since y_N is identically equal to r_k on each core P_k^δ, the realization error satisfies:

$$\begin{aligned}\left\|y_N - \Phi_T^\theta\right\|_{L^2(\mu)}^2 &\leqslant \sum_{k=1}^{K} \mu(P_k^\delta) \sup_{x \in P_k^\delta} \left\|\Phi_T^\theta(x) - r_k\right\|^2 + \int_{\Omega_\delta} \left\|y_N(x) - \Phi_T^\theta(x)\right\|^2 \mathrm{d}\mu(x) \\ &\leqslant \sum_{k=1}^{K} \mu(P_k^\delta) \sup_{x \in P_k^\delta} \left\|\Phi_T^\theta(x) - r_k\right\|^2 + 2\mu(\Omega_\delta)\Big(\|y_N\|_{L^\infty(I_R)}^2 + \|\Phi_T^\theta\|_{L^\infty(I_R)}^2\Big). \end{aligned} \tag{4.2.6}$$

.

Provided y_N and Φ_T^θ remain bounded, the contribution of Ω_δ scales directly with $\mu(\Omega_\delta)$. Modulo this mass, it suffices to enforce, for a small tolerance $\eta > 0$, a *cell-wise* target accuracy:

$$\sup_{x \in P_k^\delta} \left\|\Phi_T^\theta(x) - r_k\right\| < \eta, \qquad \text{for all } k \in [K]. \tag{4.2.7}$$

This condition is a region-wise analogue of exact point interpolation. Setting $A_k \coloneqq P_k^\delta$, the fact that these sets are disjoint compact convex sets motivates our controllability assumption for (CF):

Assumption 3. For any finite family of pairwise disjoint compact convex sets $\{A_k\}_{k=1}^K \subset \mathbb{R}^d$, any target vectors $\{r_k\}_{k=1}^K \subset \mathbb{R}^d$, and any $\eta > 0$, there exists a control $\theta(\cdot) : [0, T] \to \Theta$ such that

$$\Phi_T^\theta(A_k) \subset B(r_k, \eta) \qquad \text{for all } k.$$

The requirement in Assumption 3 is substantially stronger than point interpolation as studied in **Chapter 3**: the flow must route whole macroscopic regions, uniformly over all their points. For arbitrary targets, this cannot in general be expected from autonomous systems; some form of time

dependence is needed to decouple the motion of different cells. We defer a more detailed discussion of this obstruction, together with our architectural choice, to **Subsection 4.4.2**.

Accordingly, Assumption 3 can be verified in two ways:

(i) a *wide-network* approach, which relies on density results to approximate a reference field;

(ii) an *explicit switching* approach, extending the constructions of **Chapter 3** to interpolate cells instead of points.

While the latter is a natural geometric extension, it typically yields a flow with $\mathrm{Lip}(\Phi_T^\theta) \sim \eta^{-d^2}$ (similar to [RZ23, Theorem 5.2] in the context of $(\mathcal{V}_1^+)$). As $\eta \to 0$, this explosive growth induces severe instabilities, rendering the resulting predictor highly sensitive to spatial perturbations. Consequently, the present chapter focuses exclusively on route *(i)*.

Following this strategy, we establish a general template bound valid for any piecewise-constant estimator constructed on a convex partition, specializing it to two canonical choices:

- **Histogram partitions (Subsection 4.3.1):** These are defined by dividing the input domain into a fixed grid of hypercubes. They provide the cleanest bias-variance baseline, and in the presence of additive observation noise, histogram regression is known to be minimax-optimal.
- **Voronoi partitions (Subsection 4.3.2):** These are defined by assigning each spatial point to its nearest neighbor in the training sample [Aur91; BD15]. Generated directly by the sample points, Voronoi partitions yield faster geometric rates in the noiseless regime because the approximation error adapts to the local data density rather than a fixed grid resolution. This improves the statistical term, albeit at the cost of more complex cell geometries and a boundary-layer measure $\mu(\Omega_\delta)$ that scales with the complexity of the Voronoi skeleton.

Template bound

The following theorem synthesizes the main results developed throughout the chapter.

Main Theorem (Informal). *Assume $d \geqslant 2$ and that the controlled flow (CF) satisfies Assumption 3. Let $\mathcal{D}_N = \{(x_i, y(x_i))\}_{i=1}^N$ with $x_1, \dots, x_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{ac}(I_R)$, and $y(\cdot) \in \mathscr{C}^0(I_R; \mathbb{R}^d)$. Let y_N be a piecewise-constant estimator of y constructed on a convex partition $\{P_k\}$ of I_R, and let Ω_δ be the corresponding boundary layer of margin $\delta > 0$ as defined in (4.2.5).*

Then, for any tolerance $\eta > 0$ and any $\delta > 0$ small enough so that the trimmed cores are nonempty, there exists a data-dependent control $\theta_N = \theta_N(\mathcal{D}_N)$ such that the flow map $\Phi_T^{\theta_N}$ generated by (CF) satisfies:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|\Phi_T^{\theta_N} - y\right\|_{L^2(\mu)}^2\right] \lesssim \mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_N\right\|_{L^2(\mu)}^2\right] + \mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right] + \eta^2. \tag{4.2.8}$$

In particular, if $y(\cdot) \in \mathscr{C}^{0,\alpha}(I_R; \mathbb{R}^d)$ for some $\alpha \in (0, 1]$, μ has a density bounded above on I_R and δ is small enough, then there exists a control θ_N such that:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|\Phi_T^{\theta_N} - y\right\|_{L^2(\mu)}^2\right] \lesssim \begin{cases} N^{-\frac{2\alpha}{2\alpha+d}} + \delta N^{\frac{1}{2\alpha+d}} + \eta^2 & \text{(Histogram partition)}, \\ \left(\frac{\log N}{N}\right)^{2\alpha/d} + \delta N^{1/d} + \eta^2 & \text{(Voronoi partition)}, \end{cases} \tag{4.2.9}$$

where the Voronoi bound additionally requires the density of μ to be bounded below by a positive constant on I_R.

The bound (4.2.8) decomposes our generalization analysis into three independent tasks:

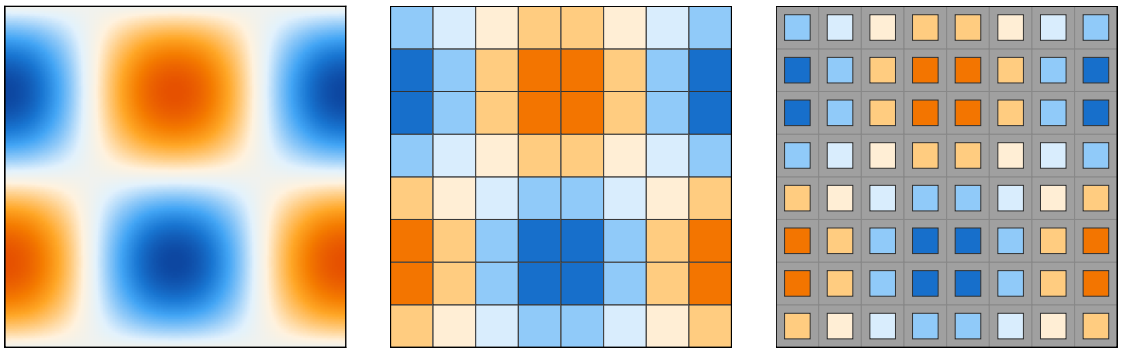


Figure 4.1: **Left:** Color map representing the ground truth $y(\cdot)$, showing a single coordinate of its range (the other coordinate is analogous). **Middle:** Piecewise-constant estimation y_N over a uniform grid. **Right:** Trimmed-core approximation.

- **Statistics:** Bound the error $\mathbb{E}_{\mathcal{D}_N}\left[\|y - y_N\|^2_{L^2(\mu)}\right]$ for concrete choices of partitions and proxies.
- **Geometry:** Control the boundary mass $\mu(\Omega_\delta)$ induced by the trimming.
- **Expressivity:** Relate the realizability tolerance η (from Assumption 3) to the model complexity.

4.3 Nonparametric estimation

We analyze the statistical term in (4.2.8) for the two canonical piecewise-constant estimators introduced earlier.

4.3.1 Histogram instantiation

We first analyze the *histogram* approach. Fix a resolution $h \in (0, 2R]$ and let

$$\mathcal{K} := \{0, \ldots, \lceil 2R/h \rceil - 1\}^d \tag{4.3.1}$$

be the corresponding multi-index set. For each $k \in \mathcal{K}$, we define the cube

$$Q_k := I_R \cap \prod_{j=1}^{d} \big[-R + k_j h,\, -R + (k_j + 1)h\big), \tag{4.3.2}$$

ensuring that the collection $\{Q_k\}_{k\in\mathcal{K}}$ forms a partition of I_R. The total number of cells is given by

$$K_h := |\mathcal{K}| = \left\lceil \frac{2R}{h} \right\rceil^d \lesssim h^{-d}, \tag{4.3.3}$$

and the diameter of each cell satisfies $\operatorname{diam}(Q_k) \leqslant \sqrt{d}\, h$.

We approximate the ground truth $y(\cdot)$ by averaging its values within each cell. This leads to the definition of both a population-level and an empirical estimator:

The *population average* $y_h \colon I_R \to \mathbb{R}^d$ is the best piecewise-constant approximation of y in $L^2(\mu)$:

$$y_h(x) := \sum_{k\in\mathcal{K}} s_k\, \mathbf{1}_{Q_k}(x), \tag{4.3.4}$$

where the coefficients $s_k \in \mathbb{R}^d$ are defined by

$$s_k \coloneqq \begin{cases} \dfrac{1}{\mu(Q_k)} \displaystyle\int_{Q_k} y(x)\,\mathrm{d}\mu(x), & \text{if } \mu(Q_k) > 0, \\ 0, & \text{if } \mu(Q_k) = 0. \end{cases} \tag{4.3.5}$$

The *empirical average* $y_{N,h} : I_R \to \mathbb{R}^d$, instead, is computed directly from the dataset $\mathcal{D}_N$:

$$y_{N,h}(x) \coloneqq \sum_{k \in \mathcal{K}} S_k \, \mathbf{1}_{Q_k}(x). \tag{4.3.6}$$

Letting $N_k \coloneqq \sum_{i=1}^N \mathbf{1}_{Q_k}(x_i)$ denote the number of samples falling into Q_k, the coefficients $S_k \in \mathbb{R}^d$ are now defined as

$$S_k \coloneqq \begin{cases} \dfrac{1}{N_k} \displaystyle\sum_{i : x_i \in Q_k} y(x_i), & \text{if } N_k \geqslant 1, \\ 0, & \text{if } N_k = 0. \end{cases} \tag{4.3.7}$$

We now establish the convergence rate for the empirical histogram estimator.

Proposition 4.3.1. *Let $\mathcal{D}_N = \{(x_i, y(x_i))\}_{i=1}^N$ with $x_1, \dots, x_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{ac}(I_R)$, and $y(\cdot) \in \mathscr{C}^0(I_R; \mathbb{R}^d)$. Then, for any $h \in (0, 2R]$, the estimator $y_{N,h}$ defined in (4.3.6) satisfies*

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_{N,h}\right\|^2_{L^2(\mu)}\right] \lesssim \omega_y(\sqrt{d}\,h)^2 + \frac{\|y\|^2_{L^\infty(I_R)}}{N h^d}, \tag{4.3.8}$$

where $\omega_y(t) \coloneqq \sup\{\|y(x) - y(x')\| \mid x, x' \in I_R,\ \|x - x'\| \leq t\}$ is the modulus of continuity of $y(\cdot)$.

The bound in (4.3.8) captures the classical bias-variance trade-off inherent to nonparametric estimation [Gyö+02]:

- **Bias:** The term $\omega_y(\sqrt{d}\,h)^2$ bounds the spatial approximation error, which vanishes as $h \to 0$.
- **Variance:** The term scaling as $1/(Nh^d)$ bounds the statistical estimation error. It diverges as $h \to 0$ because finer grids leave fewer samples per cell. Consequently, driving this error to zero requires N to grow strictly faster than the number of cells, ensuring that $Nh^d \to \infty$.

Minimizing the total expected error requires finding the optimal resolution h that balances these two competing terms.

The continuity assumption on $y(\cdot)$ can be relaxed at the expense of losing an algebraic decay rate for the bias term:

- **Bounded targets ($y \in L^\infty(I_R; \mathbb{R}^d)$):** The bound (4.3.8) remains valid if $\omega_y(\sqrt{d}\,h)^2$ is replaced by the exact bias $\|y - y_h\|^2_{L^2(\mu)}$. To see that this term vanishes as $h \to 0$, one can consider a sequence of nested grids (e.g., $h_m = 2R/2^m \to 0$). The population averages y_{h_m} are precisely the conditional expectations $\mathbb{E}[y \mid \mathcal{F}_m]$, where $\mathcal{F}_m$ is the σ-algebra generated by the m-th partition. Because $\mathcal{F}_m \subseteq \mathcal{F}_{m+1}$, this sequence forms a uniformly integrable martingale. By standard martingale convergence, e.g. [Dur19, Theorem 4.4.6], we have $y_{h_m} \to y$ in $L^2(\mu)$.
- **Square-integrable targets ($y \in L^2(\mu; \mathbb{R}^d)$):** The same convergence $\|y - y_h\|_{L^2(\mu)} \to 0$ holds (along nested refinements), but obtaining a variance bound of order $(Nh^d)^{-1}$ generally requires additional assumptions (e.g. boundedness or uniform control of cell masses). We do not pursue this level of generality here.

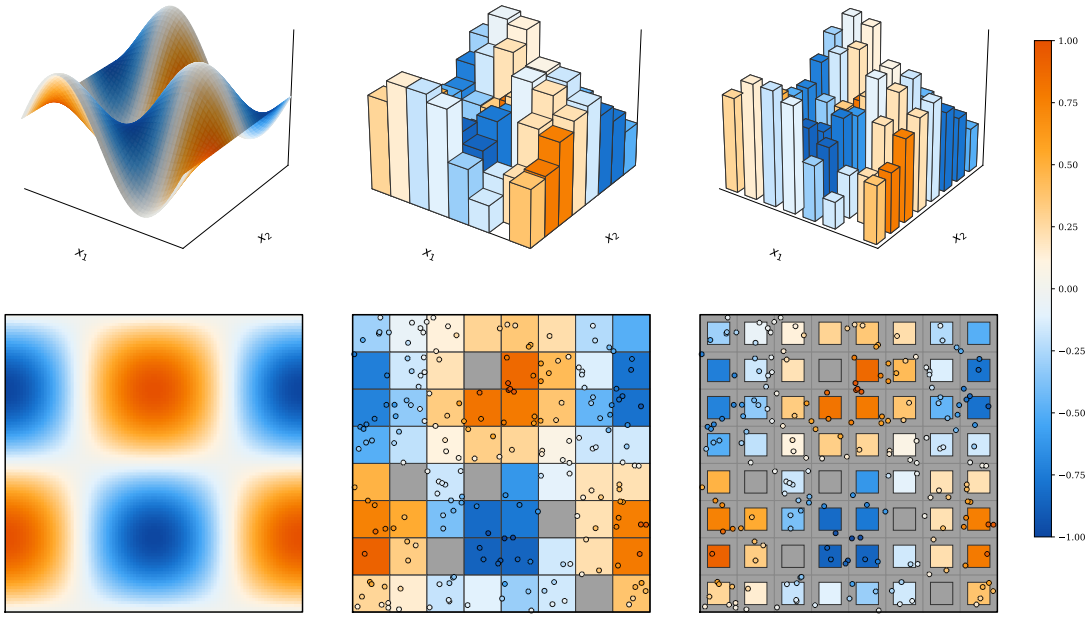


Figure 4.2: Illustration of the approximation procedure for Proposition 4.3.1. From left to right: the ground truth $y\colon [-1,1]^2 \to [-1,1]^2$, the empirical estimator $y_{N,h}$ built from N samples over the grid of size h, and the corresponding estimator on the trimmed cores. The plots in the bottom row show the corresponding domains and cell partitions.

On the other hand, by imposing stronger regularity on $y(\cdot)$, one can quantify the modulus of continuity and optimize h to achieve the sharpest convergence rate with respect to N. We focus on Hölder regularity $y(\cdot) \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$. Equivalently, the modulus of continuity satisfies:

$$\omega_y(t) \leqslant L_y t^\alpha \qquad \text{for some } \alpha \in (0,1] \text{ and } L_y > 0.$$

The following corollary derives the optimal statistical rate for any such ground truth.

Corollary 4.3.2. *Suppose $y(\cdot) \in \mathscr{C}^{0,\alpha}(I_R;\mathbb{R}^d)$ for some $\alpha \in (0,1]$. Then, the estimator y_h defined in (4.3.4) satisfies*

$$\|y - y_h\|^2_{L^2(\mu)} \lesssim h^{2\alpha}. \tag{4.3.9}$$

Moreover, $h = (2R) \wedge (N^{-\frac{1}{2\alpha+d}}) =: h_N$ yields the minimal convergence rate in (4.3.8),

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_{N,h_N}\right\|^2_{L^2(\mu)}\right] \lesssim N^{-\frac{2\alpha}{2\alpha+d}}. \tag{4.3.10}$$

We conclude our analysis of the histogram approach with a brief discussion on its sample complexity, optimality, and limitations.

Remark 4.3.3 (Sample complexity and the curse of dimensionality). The approximation by Riemann sums on a fixed grid of size h yields the bias rate in (4.3.9). To ensure a squared error

below a threshold $\varepsilon > 0$, the number of cells must scale as $K_h \gtrsim \varepsilon^{-\frac{d}{2\alpha}}$. Consequently, the sample complexity to guarantee a total expected risk below ε in (4.3.10) scales as:

$$N \gtrsim \varepsilon^{-\frac{2\alpha+d}{2\alpha}} = \varepsilon^{-1} \cdot \varepsilon^{-\frac{d}{2\alpha}}. \tag{4.3.11}$$

We observe that the sample complexity naturally decomposes into the product of the Monte Carlo rate ε^{-1} (associated with estimating a mean with fixed variance) and a geometric factor $\varepsilon^{-d/(2\alpha)}$. This reflects the curse of dimensionality: whereas numerical integration computes a single global average, function recovery requires estimating K_h distinct local averages simultaneously to resolve the spatial structure of $y(\cdot)$ in d dimensions.

Remark 4.3.4 (Kolmogorov width and minimax optimality). In terms of the number of cells $K_h \asymp h^{-d}$, the bias estimate (4.3.9) yields

$$\|y - y_h\|_{L^2(\mu)} \lesssim K_h^{-\alpha/d}.$$

This matches the optimal rate among all *linear* approximation spaces of dimension K_h, as captured by the *Kolmogorov n-width*

$$d_n(\mathcal{F}) := \inf_{\dim(V)=n} \sup_{f\in\mathcal{F}} \inf_{g\in V} \|f - g\|_{L^2(\mu)}.$$

For the isotropic $\mathscr{C}^{0,\alpha}$-Hölder ball $\mathcal{F}$, one has $d_n(\mathcal{F}) \asymp n^{-\alpha/d}$ [Pin85], an order attained by uniform piecewise-constant grids such as y_h.

Moreover, this exponent cannot be improved even by *adaptive* (nonlinear) methods with continuous parameter selection: the corresponding manifold n-widths exhibit the exact same decay rate on Besov (and hence Hölder) balls [DeV98, Eq. (9.4)]. Adaptivity can only yield strictly faster rates on smaller, spatially inhomogeneous target classes [Coh+01].

Remark 4.3.5 (Saturation of regularity). The histogram rate (4.3.10) saturates at Lipschitz continuity ($\alpha = 1$). Because the estimator $y_{N,h}$ is piecewise constant (an order-0 approximation), it cannot exploit higher-order derivatives. If the ground truth is smoother, e.g., $y(\cdot) \in \mathscr{C}^{m,\alpha}(I_R;\mathbb{R}^d)$ for $m \geq 1$, conceptually replacing the piecewise-constant proxy with a degree-m cellwise polynomial would improve the statistical bias to $O(h^{m+\alpha})$, matching the Kolmogorov n-width for such spaces.

However, transferring this higher-order rate to Φ_T^θ via (4.2.8) would require a complex, higher-order analogue of Assumption 3—approximating prescribed local polynomials up to tolerance η on disjoint trimmed cores. While theoretically feasible, we deliberately maintain the order-0 baseline: it matches the minimal regularity regime, interfaces directly with standard point-to-cell controllability mechanisms, and naturally extends the framework developed in **Chapter 3**.

Remark 4.3.6 (Optimality under noise). Despite the saturation limit, the rate $N^{-\frac{2\alpha}{2\alpha+d}}$ matches the celebrated minimax rate for nonparametric regression with additive noise [Sto82]. If samples are drawn from a joint distribution $\mu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that $y_i = y(x_i) + \varepsilon_i$, individual data points are unreliable. The optimal strategy is exactly what the histogram estimator does: locally averaging clusters of points to cancel out stochastic fluctuations.

In contrast to the noisy regime described in Remark 4.3.6, our setting is strictly *noiseless*: the training labels satisfy $y_i = y(x_i)$ exactly. In this scenario, spatial averaging is overly conservative because it effectively *distrusts* perfect data, wasting the precision of individual samples.

In the noiseless regime, enforcing *strict interpolation* of data bypasses the variance penalty entirely, yielding substantially faster rates. For instance, assuming $y(\cdot) \in \mathscr{C}^{k,\alpha}(I_R)$ with $k \geq 0$ integer, and $\alpha \in (0, 1]$, [Bau+17] shows that spline interpolants m_N achieve

$$\|y - m_N\|^2_{L^2(\mu)} \lesssim \|y - m_N\|^2_{L^\infty(I_R)} \lesssim \left(\frac{\log N}{N}\right)^{\frac{2(k+\alpha)}{d}}. \tag{4.3.12}$$

To exploit this noiseless potential while rigorously maintaining the piecewise structure required by Assumption 3, we turn to the simplest interpolation scheme: the nearest-neighbor partition.

4.3.2 Voronoi instantiation

Let $\{V_i\}_{i=1}^N$ be the partition of I_R into the *Voronoi cells* generated by the input set $\{x_i\}_{i=1}^N$:

$$V_i \coloneqq \{x \in I_R \mid \|x - x_i\| \leq \|x - x_j\| \text{ for all } j \neq i\}. \tag{4.3.13}$$

For any $x \in I_R$, let $x_{\mathrm{NN}}(x)$ denote its nearest neighbor inside the input set (up to boundary ties of measure zero):

$$x_{\mathrm{NN}}(x) \coloneqq \operatorname*{arg\,min}_{x_i \in \{x_1, \ldots, x_N\}} \|x - x_i\|. \tag{4.3.14}$$

Using this mapping, we define the *nearest-neighbor interpolant* y_N^V as the piecewise-constant function:

$$y_N^V(x) \coloneqq y\big(x_{\mathrm{NN}}(x)\big) = \sum_{i=1}^N y(x_i)\, \mathbf{1}_{V_i}(x). \tag{4.3.15}$$

The approximation error of y_N^V is governed purely by the *covering radius* of $\mathcal{D}_N$, which measures how well the training inputs cover the domain:

$$R_N \coloneqq \sup_{x \in I_R} \min_{1 \leq i \leq N} \|x - x_i\| = \sup_{x \in I_R} \|x - x_{\mathrm{NN}}(x)\|. \tag{4.3.16}$$

If $y(\cdot)$ is α-Hölder with constant $L_y > 0$, the pointwise error is bounded by the local distance to the sample:

$$\|y(x) - y_N^V(x)\| = \|y(x) - y(x_{\mathrm{NN}}(x))\| \leq L_y \|x - x_{\mathrm{NN}}(x)\|^\alpha \leq L_y R_N^\alpha. \tag{4.3.17}$$

To bound the covering radius R_N, assume that μ admits a density ρ on I_R satisfying:

$$\inf_{x \in I_R} \rho(x) > 0.$$

Under this assumption, the covering radius of i.i.d. samples satisfies a logarithmic scaling. Specifically, applying [Pen23, Corollary 4.5] (with connectivity parameter $k(n) \equiv 1$ and appropriately rescaling I_R), there exists a constant $C > 0$ and $N_0 \in \mathbb{N}$ such that, with probability 1, it holds:

$$R_N \leq C \left(\frac{\log N}{N}\right)^{1/d} \qquad \text{for all } N \geq N_0. \tag{4.3.18}$$

By substituting (4.3.18) into (4.3.17), we conclude that

$$\|y - y_N^V\|^2_{L^2(\mu)} \leq \|y - y_N^V\|^2_{L^\infty(I_R)} \lesssim \left(\frac{\log N}{N}\right)^{\frac{2\alpha}{d}}.$$

This rate is strictly faster than the corresponding histogram rate $O\big(N^{-\frac{2\alpha}{2\alpha+d}}\big)$ derived in (4.3.10).

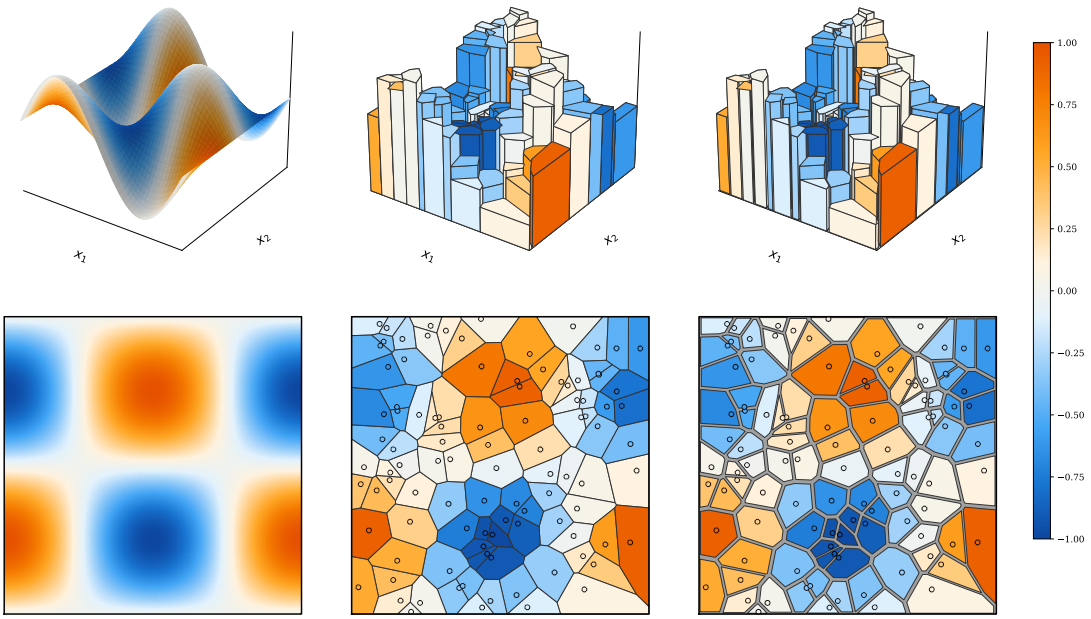


Figure 4.3: Illustration of the Voronoi realization. From left to right: the target $y(\cdot)$, the estimator y_N^V, and the trimmed proxy V_i^δ supported strictly on separated cores.

4.4 Flow realization and complexity

Recall the general risk decomposition from (4.2.8):

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \underbrace{\mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_N\right\|^2_{L^2(\mu)}\right]}_{\text{Statistics}} + \underbrace{\mathbb{E}_{\mathcal{D}_N}\left[\mu(\Omega_\delta)\right]}_{\text{Geometry}} + \underbrace{\eta^2}_{\text{Control}}, \tag{4.4.1}$$

where y_N is our piecewise-constant proxy (either the histogram $y_{N,h}$ or the interpolant y_N^V), and the geometry term implicitly assumes uniform $L^\infty(I_R)$-bounds on both y_N and the flow map.

Having analyzed the statistical term in (4.4.1), we now address the remaining two.

4.4.1 Geometric error

To realize (4.2.7) with a continuous flow, we trim the cells by a margin $\delta > 0$, as defined in (4.2.5). We now bound the removed mass $\mu(\Omega_\delta)$ for both canonical partitions considered.

Lemma 4.4.1. *For any $\mu \in \mathcal{P}_{ac}(I_R)$ with density bounded above, any partition $\{Q_k\}$ as in (4.3.2) with $h \in (0, 2R]$, and any margin $\delta \in (0, h/2)$, the boundary layer Ω_δ defined in (4.2.5) satisfies*

$$\mu(\Omega_\delta) \lesssim \frac{\delta}{h}. \tag{4.4.2}$$

For Voronoi partitions, Ω_δ corresponds to the δ-neighborhood of the *Voronoi skeleton* Σ_N, defined

as the union of all cell boundaries:

$$\Sigma_N := \bigcup_{i=1}^{N} \partial V_i. \tag{4.4.3}$$

Unlike the fixed grid, Σ_N depends on the random spatial configuration of the sample points. We bound its expected measure as follows:

Lemma 4.4.2. *For any integers $d \geq 2$ and $N \geq 1$, any $\mu \in \mathcal{P}_{ac}(I_R)$ with density bounded above and below by positive constants, and any sufficiently small margin $\delta > 0$, the boundary layer Ω_δ defined in (4.2.5) via the Voronoi cells $\{V_i\}_{i=1}^N$ generated by $\mathcal{D}_N$ (as in (4.3.13)) satisfies*

$$\mathbb{E}_{\mathcal{D}_N}\big[\mu(\Omega_\delta)\big] \lesssim \delta N^{1/d}. \tag{4.4.4}$$

Note that, by (4.3.13), Voronoi cells are intersections of half-spaces, and thus convex polytopes. Consequently, the corresponding trimmed cores V_i^δ are strictly disjoint, compact, and convex sets. This makes the Voronoi partition geometrically compatible with Assumption 3.

4.4.2 Dynamics and complexity

We now explain how Assumption 3 is realized and how the tolerance η depends on the model complexity.

At this point, the distinction between pointwise and cell-wise controllability becomes essential. For finite point clouds, autonomous systems may still (approximately) route individual particles to arbitrary target points, as isolated trajectories can bypass one another provided $d \geq 2$. In Assumption 3, however, each A_k is a macroscopic set, and the flow is required to uniformly map the entire region into a small ball $B(r_k, \eta)$.

This requirement creates a structural obstruction for autonomous systems. Indeed, if one target ball lies inside the initial domain of another cell, such a routing can only occur under strong compatibility conditions between the corresponding families of orbits. Since the targets r_k are completely arbitrary, no such compatibility can be guaranteed in general. Thus, unless $y(I_R) \cap I_R = \emptyset$, autonomous models—such as $(\mathcal{V}_{\mathrm{a}}^+)$ or $(\mathcal{V}_{\mathrm{2a}}^+)$—are not expected to satisfy Assumption 3 in full generality. This is discussed later in Remark 4.4.5. Time dependence therefore emerges as the natural mechanism for decoupling the motion of different cells.

Consequently, we use the *semi-autonomous neural ODE* defined in $(\mathcal{V}_{\mathrm{sa}}^+)$:

$$\dot{x}(t) = \sum_{j=1}^{p} w_j \left(a_j \cdot x(t) + c_j t + b_j\right)_+, \qquad t \in [0,T], \tag{$\mathcal{V}_{\mathrm{sa}}^+$}$$

with constant parameters $\theta = (w_j, a_j, b_j, c_j)_{j=1}^p \in (\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R})^p$. The following result quantifies how the network width p controls the realizability tolerance in Assumption 3, showing that steering finitely many disjoint, compact, convex cells up to an accuracy $\eta > 0$ requires a width $p \gtrsim \eta^{-2}$.

Proposition 4.4.3. *Let $R > 0$ be fixed, let $\{A_k\}_{k=1}^K$ be a finite collection of pairwise disjoint, compact, and convex subsets of $I_R = [-R,R]^d$, and let $\{\mathbf{r}_k\}_{k=1}^K \subset \mathbb{R}^d$ be corresponding target vectors. For every $\eta > 0$ and every integer $p \gtrsim \eta^{-2}$, there exists a control θ of width p such that the flow map Φ_T^θ of $(\mathcal{V}_{\mathrm{sa}}^+)$ satisfies:*

(i) $\Phi_T^\theta(A_k) \subset B(r_k, \eta)$ *for all $k \in [K]$;*

(ii) $\|\Phi_T^\theta\|_{L^\infty(I_R)} \le C$, *where the constant* $C > 0$ *is independent of* p *and* η.

Remark 4.4.4. The tolerance scaling $\eta \sim p^{-1/2}$, derived from standard Barron-type estimates, can be further improved to $\eta \sim p^{-\frac{1}{2}-\frac{3}{2d}}$ by exploiting the higher regularity of the underlying interpolating vector fields, as recently shown in [Sie25].

Remark 4.4.5. The structural obstruction to using autonomous systems is inherently tied to the condition $\Phi_T^\theta(A_k) \subset B(r_k, \eta)$ required by Assumption 3. This is stronger than what is needed for a purely $L^2(\mu)$-based approximation of y_N. In principle, a weaker, integral notion of realization could tolerate exceptional subsets of uncontrolled, yet negligible, mass inside the cells, potentially allowing for the use of autonomous systems. While we do not pursue this direction here, the strict set-inclusion is adopted because it enables the clean error decomposition in (4.2.6).

4.4.3 Main bounds

By combining the statistical, geometric, and control estimates derived above, we obtain explicit risk bounds for the predictor. At the abstract level, the realization term enters as η^2, where η is the controllability tolerance from Assumption 3. For $(\mathcal{V}_{\mathrm{sa}}^+)$ of width p, Proposition 4.4.3 yields $\eta \lesssim p^{-1/2}$, and therefore the control contribution becomes of order p^{-1}.

Theorem 4.4.6. *Let* $d \ge 2$ *and* $N \ge 1$ *be integers, let* $\mathcal{D}_N = \{(x_i, y(x_i))\}_{i=1}^N$ *with* $x_1, \dots, x_N$ *drawn i.i.d. from* $\mu \in \mathcal{P}_{ac}(I_R)$ *with density bounded above, and let* $y \in \mathscr{C}^0(I_R; \mathbb{R}^d)$. *For any resolution* $h \in (0, 2R]$, *margin* $\delta \in (0, h/4)$, *and integer* $p \geqslant 1$, *there exists a control* θ_N *of width* p *for the semi-autonomous vector field* v_θ *of the form* $(\mathcal{V}_{\mathrm{sa}}^+)$ *such that*

$$\mathbb{E}_{\mathcal{D}_N}\big[\mathcal{R}(\theta_N)\big] \;\lesssim\; \omega_y(\sqrt{d}\,h)^2 \;+\; \frac{\|y\|_{L^\infty(I_R)}^2}{Nh^d} \;+\; \frac{\delta}{h} \;+\; \frac{1}{p}. \tag{4.4.5}$$

Remark 4.4.7 (Parameter selection and rates). The decomposition in (4.4.5) highlights the trade-off between the four fundamental error sources of the problem: the statistical bias $\mathsf{B}(h) \sim \omega_y(h)^2$, the variance $\mathsf{V}(N, h) \sim 1/(Nh^d)$, the geometric gap $\mathsf{G}(h, \delta) \sim \delta/h$, and the finite-width realization error $\mathsf{C}(p) \sim 1/p$ (which for generic controlled flows is simply an arbitrary tolerance η^2 that can be made vanishingly small).

To ensure $\mathbb{E}[\mathcal{R}(\theta_N)] \lesssim \varepsilon$, we select the parameters sequentially:

(i) first, the grid resolution h must satisfy $\mathsf{B}(h) \lesssim \varepsilon$;
(ii) second, the sample size must scale as $N \gtrsim 1/(\varepsilon h^d)$;
(iii) third, the geometric margin must be small enough to yield $\delta \lesssim \varepsilon h$;
(iv) and finally, the network width must satisfy $p \gtrsim 1/\varepsilon$.

In the case where $y \in \mathscr{C}^{0,\alpha}(I_R)$, we have $\mathsf{B}(h) \lesssim h^{2\alpha}$, and the optimal balance is achieved by setting $h_N \asymp N^{-\frac{1}{2\alpha+d}}$. This choice recovers the minimax rate (4.3.10), provided that the remaining terms vanish at the same speed. For the geometric gap, this requires

$$\frac{\delta_N}{h_N} \lesssim h_N^{2\alpha} \quad \Longleftrightarrow \quad \delta_N \lesssim h_N^{1+2\alpha} \asymp N^{-\frac{1+2\alpha}{2\alpha+d}}.$$

This explicitly shows that as N increases and the grid becomes finer, the margin δ_N must shrink faster than the resolution h_N to ensure that the volume of the *uncontrolled* boundary layer remains negligible relative to the statistical error.

Finally, we state the corresponding result for the Voronoi estimator in the noiseless, Hölder regime.

Theorem 4.4.8. *Let $d \geq 2$ and $N \geq 1$ be integers, let $\mathcal{D}_N = \{(x_i, y(x_i))\}_{i=1}^N$ with $x_1, \dots, x_N$ drawn i.i.d. from $\mu \in \mathcal{P}_{ac}(I_R)$ with density bounded above and below by positive constants, and let $y \in \mathscr{C}^{0,\alpha}(I_R; \mathbb{R}^d)$. For any margin $\delta > 0$ and integer $p \geq 1$, there exists a control θ_N of width p for the semi-autonomous vector field v_θ of the form $(\mathcal{V}_{\mathrm{sa}}^+)$ such that*

$$\mathbb{E}_{\mathcal{D}_N}\big[\mathcal{R}(\theta_N)\big] \;\lesssim\; \left(\frac{\log N}{N}\right)^{2\alpha/d} \;+\; \delta N^{1/d} \;+\; \frac{1}{p}. \tag{4.4.6}$$

Remark 4.4.9 (Parameter selection for Voronoi). Equation (4.4.6) reveals a fundamental tension between the approximation error and the geometric boundary layer. To ensure that the total risk vanishes, the margin δ must shrink fast enough to counteract the growing complexity of the Voronoi skeleton (which scales as $N^{1/d}$). Specifically, to prevent the boundary layer from dominating the statistical rate, we require

$$\delta N^{1/d} \ll \left(\frac{\log N}{N}\right)^{2\alpha/d} \quad \Longleftrightarrow \quad \delta \ll \left(\frac{(\log N)^{2\alpha}}{N^{2\alpha+1}}\right)^{1/d}.$$

In the wide-network setting, one can always select such a δ without an architectural penalty, provided the width p is sufficiently large to satisfy the chosen tolerance $\eta \sim p^{-1/2} \lesssim (\log N/N)^{\alpha/d}$.

4.4.4 Discussion

To place our results in context, it is instructive to compare them with standard generalization bounds for neural ODEs. The prevailing literature typically relies on the classical risk decomposition

$$\mathcal{R}(\theta) \leqslant \mathcal{R}_N(\theta) + \sup_{\vartheta \in \mathcal{U}} \big|\mathcal{R}(\vartheta) - \mathcal{R}_N(\vartheta)\big|, \tag{4.4.7}$$

where the second term controls the generalization gap uniformly over a prescribed class $\mathcal{U}$.

For neural ODEs, the precise form of such a bound depends strongly on the chosen architecture and parameterization. In particular, for the autonomous neural ODE class

$$\dot{x}(t) = W\,\sigma(x(t)),$$

with W ranging over a bounded finite-dimensional set, [Mar23] proves a high-probability estimate for an empirical risk minimizer $\widehat{W}_N$ of the form

$$\mathcal{R}(\widehat{W}_N) \;\leq\; \mathcal{R}_N(\widehat{W}_N) \;+\; B(d+1)\sqrt{\frac{\log(R_W dN)}{N}} \;+\; B\sqrt{\frac{\log(1/\delta)}{N}}, \tag{4.4.8}$$

where R_W bounds the admissible parameter set and the constant B depends on the model and data bounds. Thus, for this finite-dimensional autonomous class, one recovers the familiar $N^{-1/2}$ scaling, up to logarithmic factors.

We emphasize, however, that bounds of the form (4.4.8) are model-dependent. In particular, they do not automatically apply to the semi-autonomous architecture $(\mathcal{V}_{\mathrm{sa}}^+)$ studied in this chapter. Still, since at fixed width p the parameter space of $(\mathcal{V}_{\mathrm{sa}}^+)$ remains finite-dimensional, one may naturally expect an analogous bound with the same rates (up to logarithmic factors and model-dependent constants). Establishing such a result rigorously lies beyond the scope of the present chapter.

Contrasting this with our approach reveals a clear, practical trade-off:

- **The curse of dimensionality.** By measuring complexity through the parameter space rather than through the geometry of the input domain, bounds such as (4.4.8) scale as $N^{-1/2}$ up to logarithmic factors, independently of the ambient dimension d. In contrast, because our construction relies on spatial partitioning, the statistical terms in our bounds necessarily exhibit the nonparametric dependence on d.
- **The expressivity–complexity tension.** The decomposition (4.4.7) exposes a structural dilemma: enlarging the admissible class $\mathcal{U}$ improves the approximation power and may reduce the empirical risk $\mathcal{R}_N$, but it simultaneously worsens the uniform generalization term. In the neural-ODE setting, this deterioration is reflected both in the size of the admissible parameter set (through quantities such as R_W) and in stability constants entering the bound. Thus, increased expressivity is paid for through a larger global complexity penalty.
- **Decoupling via controllability.** Our framework bypasses this tension by abandoning the uniform supremum in (4.4.7). Instead of controlling an entire hypothesis class, we obtain a single data-dependent flow map that realizes a nonparametric proxy. As a consequence, the network width only affects the realization term: in Theorems 4.4.6 and 4.4.8, increasing p improves the control accuracy through the term p^{-1}, while leaving the statistical estimation terms unchanged. In this sense, expressivity is decoupled from the sampling complexity.

In summary, while our approach is more sensitive to the input dimension, it shows a key conceptual point: highly expressive, continuous-depth models can safely emulate nonparametric estimators without incurring the severe variance penalties that uniform-capacity frameworks would predict.

4.5 Perspectives

The framework developed in this chapter views the neural ODE as a mechanism to learn a static map from inputs to outputs. Namely, if the data is generated by an unknown dynamical system with flow $\{y_t\}_{t\in[0,T]}$, our previous setting corresponds to observing only the terminal endpoint y_T.

However, in many time-series and physical modeling scenarios, one observes the entire trajectory for each initial condition $x \in I_R$. The dataset then consists of pairs (x_i, γ_i), where $\gamma_i(t) \coloneqq y_t(x_i)$ is observed continuously (or densely) over $t \in [0,T]$. To measure generalization in this dynamic setting, one naturally extends the static loss to a trajectory risk:

$$\mathcal{R}_{\mathrm{traj}}(\theta) \;\coloneqq\; \int_{I_R}\int_0^T \left\|\Phi_t^\theta(x) - y_t(x)\right\|^2 \mathrm{d}t\,\mathrm{d}\mu(x). \tag{4.5.1}$$

Lifting our cell-wise interpolation approach to this trajectory setting is conceptually straightforward. Given a convex partition of the input domain, one can construct a piecewise-constant *path proxy*. Because any valid flow must satisfy the initial condition $\Phi_0^\theta = \mathrm{Id}$, this proxy cannot be an absolute position; rather, it must be constructed as an *anchored* displacement. Within each cell Q_k, the proxy dictates that all points x follow a common learned displacement trajectory $b_k(t) \coloneqq \widehat{y}_k(t) - \widehat{y}_k(0)$.

The generalization analysis would then follow the exact same statistical–geometric–control decomposition as (4.4.1), carrying the time integral throughout the bounds. The statistical error would be governed by how well the cell-wise path averages approximate the true flow tubes, while the geometric boundary layer $\mu(\Omega_\delta)$ would remain unchanged. The fundamental modification lies in the control requirement: instead of merely steering cells to static target points at time T, the neural ODE must now satisfy a *dynamic tracking controllability*, keeping the predicted trajectories inside a thin time-dependent tube around the anchored proxy for all $t \in [0,T]$.

This shift from terminal-time steering to uniform-in-time tracking strongly reaffirms the need for time-dependent architectures. As previously discussed, Assumption 3 generally demands non-autonomous vector fields such as the semi-autonomous parameterization ($\mathcal{V}^{+}_{\mathrm{sa}}$). As a complementary benefit, this time-dependence equips the model to accurately track underlying data-generating dynamics that might themselves be non-autonomous (e.g., systems subject to time-varying external forces), a task where purely autonomous models would fail.

4.6 Proofs

4.6.1 Proofs of Section 4.3

Proof of Proposition 4.3.1. By the squared-sum inequality, the expected error can be decomposed into a bias and a variance component:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_{N,h}\right\|^2_{L^2(\mu)}\right] \leq 2\left\|y - y_h\right\|^2_{L^2(\mu)} + 2\,\mathbb{E}_{\mathcal{D}_N}\left[\left\|y_h - y_{N,h}\right\|^2_{L^2(\mu)}\right],$$

where y_h and $y_{N,h}$ are the population and empirical estimators defined in (4.3.4) and (4.3.6), respectively. We bound each term separately.

Bias. Fix a cell Q_k with $p_k \coloneqq \mu(Q_k) > 0$. The population coefficient s_k is the conditional expectation of $y(\cdot)$ on Q_k. By Jensen's inequality, for any $x \in Q_k$:

$$\|y(x) - s_k\| = \left\|\frac{1}{p_k}\int_{Q_k}\big(y(x) - y(x')\big)\,\mathrm{d}\mu(x')\right\| \leq \frac{1}{p_k}\int_{Q_k}\|y(x) - y(x')\|\,\mathrm{d}\mu(x').$$

For any $x, x' \in Q_k$, the distance satisfies

$$\|x - x'\| \leq \operatorname{diam}(Q_k) \leq \sqrt{d}\,h.$$

Consequently, $\|y(x) - y(x')\| \leq \omega_y(\sqrt{d}\,h)$, which implies $\|y(x) - s_k\| \leq \omega_y(\sqrt{d}\,h)$. Since cells with $p_k = 0$ do not contribute to the $L^2(\mu)$ norm, integrating over the partition yields:

$$\|y - y_h\|^2_{L^2(\mu)} = \sum_{k\in\mathcal{K}}\int_{Q_k}\|y(x) - s_k\|^2\,\mathrm{d}\mu(x) \leq \omega_y(\sqrt{d}\,h)^2\sum_{k\in\mathcal{K}}\mu(Q_k) = \omega_y(\sqrt{d}\,h)^2,$$

where $\mathcal{K}$ is the index set defined in (4.3.1).

Variance. Since $y_h|_{Q_k} = s_k$ and $y_{N,h}|_{Q_k} = S_k$, the variance term becomes:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y_h - y_{N,h}\right\|^2_{L^2(\mu)}\right] = \sum_{k\in\mathcal{K}} p_k\,\mathbb{E}_{\mathcal{D}_N}\left[\|s_k - S_k\|^2\right]. \tag{4.6.1}$$

Fix an index k and let $N_k \coloneqq \sum_{i=1}^{N}\mathbf{1}_{Q_k}(x_i)$ denote the number of samples falling into Q_k. Because the inputs are drawn i.i.d. from μ, N_k follows a binomial distribution, $N_k \sim \mathrm{Binom}(N, p_k)$.

Conditionally on $N_k = n \geq 1$, the coefficient S_k is the empirical average of n i.i.d. random variables $Y_1, \ldots, Y_n$ distributed as $y(X)$ for $X \sim \mu(\cdot \mid Q_k)$. By definition, $\mathbb{E}[Y_1] = s_k$. The variance of this empirical average is exactly:

$$\mathbb{E}\big[\|S_k - s_k\|^2 \mid N_k = n\big] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}(Y_j - s_k)\right\|^2\right] = \frac{1}{n}\,\mathbb{E}\big[\|Y_1 - s_k\|^2\big] \leq \frac{\|y\|^2_{L^\infty(I_R)}}{n}.$$

If $N_k = 0$, we defined $S_k = 0$, which gives $\|S_k - s_k\|^2 = \|s_k\|^2 \le \|y\|^2_{L^\infty(I_R)}$. Taking the unconditional expectation over the binomial variable N_k, we obtain:

$$\mathbb{E}_{\mathcal{D}_N}\left[\|S_k - s_k\|^2\right] \le \|y\|^2_{L^\infty(I_R)} \left(\mathbb{E}\left[\frac{\mathbf{1}_{\{N_k>0\}}}{N_k}\right] + \mathbb{P}(N_k = 0)\right).$$

We bound the two terms in the parenthesis. First, $\mathbb{P}(N_k = 0) = (1-p_k)^N \le e^{-Np_k}$. Second, utilizing the elementary inequality $\frac{1}{n} \le \frac{2}{n+1}$ for $n \ge 1$, we evaluate the expectation via an integral identity for the binomial generating function:

$$\mathbb{E}\left[\frac{\mathbf{1}_{\{N_k>0\}}}{N_k}\right] \le 2\,\mathbb{E}\left[\frac{1}{N_k+1}\right] = 2\int_0^1 \mathbb{E}[t^{N_k}]\,\mathrm{d}t = 2\int_0^1 (1-p_k+p_kt)^N\,\mathrm{d}t \le \frac{2}{p_k(N+1)}.$$

Substituting these estimates into (4.6.1) yields:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y_h - y_{N,h}\right\|^2_{L^2(\mu)}\right] \le \|y\|^2_{L^\infty(I_R)} \sum_{k\in\mathcal{K}} \left(\frac{2}{N+1} + p_k e^{-Np_k}\right).$$

Recall from (4.3.3) that the total number of cells is $K_h \lesssim h^{-d}$. Furthermore, observing that $\sup_{p\ge 0} pe^{-Np} = (eN)^{-1}$, we can bound the sum uniformly:

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y_h - y_{N,h}\right\|^2_{L^2(\mu)}\right] \le \|y\|^2_{L^\infty(I_R)} \left(\frac{2K_h}{N+1} + \frac{K_h}{eN}\right) \lesssim \frac{\|y\|^2_{L^\infty(I_R)} K_h}{N} \lesssim \frac{\|y\|^2_{L^\infty(I_R)}}{Nh^d}.$$

Combining the bias and variance upper bounds directly gives (4.3.8). □

Proof of Corollary 4.3.2. Let $L_y > 0$ be the Hölder constant of $y(\cdot)$ on I_R, so that

$$\|y(x) - y(x')\| \le L_y \|x - x'\|^\alpha \qquad \text{for all } x, x' \in I_R.$$

By definition, the modulus of continuity satisfies

$$\omega_y(t) \le L_y t^\alpha \qquad \text{for all } t \ge 0.$$

From the bias analysis in the proof of Proposition 4.3.1, we established $\|y - y_h\|^2_{L^2(\mu)} \le \omega_y(\sqrt{d}\,h)^2$. This implies:

$$\|y - y_h\|^2_{L^2(\mu)} \le L_y^2(\sqrt{d}\,h)^{2\alpha} = L_y^2 d^\alpha h^{2\alpha}.$$

Absorbing the constant $L_y^2 d^\alpha$ into the asymptotic notation $\lesssim$ directly yields the bias rate (4.3.9).

Next, substituting the bound $\omega_y(\sqrt{d}\,h)^2 \lesssim h^{2\alpha}$ into the general estimate (4.3.8), we know there exist constants $C_1, C_2 > 0$ (independent of N and h) such that

$$\mathbb{E}_{\mathcal{D}_N}\left[\left\|y - y_{N,h}\right\|^2_{L^2(\mu)}\right] \le C_1 h^{2\alpha} + \frac{C_2}{Nh^d}. \tag{4.6.2}$$

To minimize this upper bound, we equate $h^{2\alpha}$ and $N^{-1}h^{-d}$, which dictates the optimal scaling $h \asymp N^{-1/(2\alpha+d)}$. Taking into account that $h \le 2R$, we define

$$h_N := (2R) \wedge \left(N^{-\frac{1}{2\alpha+d}}\right).$$

Evaluating (4.6.2) at precisely $h = h_N$ yields:

$$C_1 h_N^{2\alpha} + \frac{C_2}{Nh_N^d} \lesssim N^{-\frac{2\alpha}{2\alpha+d}},$$

which proves (4.3.10) and concludes the proof. □

4.6.2 Proofs of Section 4.4

Proof of Lemma 4.4.1. Geometrically, Ω_δ is contained within the δ-neighborhood of the grid skeleton. For a single cell Q_k of side length h, the volume of its trimmed boundary rim is $|Q_k| - |Q_k^\delta| = h^d - (h-2\delta)^d \lesssim h^{d-1}\delta$. Summing over the $K_h \asymp h^{-d}$ cells comprising the partition, the total volume is bounded by $|\Omega_\delta| \lesssim (h^{-d})(h^{d-1}\delta) = \delta/h$. Because the probability density of μ is bounded from above by a constant, the measure directly satisfies $\mu(\Omega_\delta) \lesssim |\Omega_\delta| \lesssim \delta/h$. □

Proof of Lemma 4.4.2

We first isolate a preliminary estimate controlling the probability that the gap between the first two nearest-neighbor distances is small.

Lemma 4.6.1. *Assume $N \geq 2$, fix $x \in I_R$, and let $D_1(x) \leq D_2(x)$ be the distances from x to its nearest and second-nearest points within $\{x_1, \ldots, x_N\}$. Then, for every $\delta > 0$,*

$$\mathbb{P}_{\mathcal{D}_N}\big(D_2(x) - D_1(x) \leq 2\delta\big) \leq N(N-1)\int_0^\infty \big[F_x(r+2\delta) - F_x(r)\big]\big(1 - F_x(r)\big)^{N-2}\,\mathrm{d}F_x(r),$$

where $F_x(r) \coloneqq \mu\big(B(x,r)\big)$.

Proof. For each $i \in [N]$, set $Z_i \coloneqq \|x - x_i\|$. Reordering these distances as $Z_{[1]} \leq Z_{[2]} \leq \cdots \leq Z_{[N]}$, we clearly have $D_1(x) = Z_{[1]}$ and $D_2(x) = Z_{[2]}$.

By symmetry, we can bound the target probability by assuming x_1 is the nearest point and taking a union bound over all N possibilities:

$$\mathbb{P}\big(D_2(x) - D_1(x) \leq 2\delta\big) \leq N\,\mathbb{P}\big(Z_1 = Z_{[1]},\ Z_{[2]} \leq Z_1 + 2\delta\big).$$

Conditioning on the event $Z_1 = r$, the joint occurrence of $Z_1 = Z_{[1]}$ and $Z_{[2]} \leq r + 2\delta$ implies that at least one of the remaining $N-1$ distances must lie in the interval $[r, r+2\delta]$, while all others must be at least r. For a fixed index $j \in \{2, \ldots, N\}$, the probability of the event $r \leq Z_j \leq r + 2\delta$ is exactly $F_x(r+2\delta) - F_x(r)$, and for each $k \notin \{1, j\}$, the probability of $Z_k \geq r$ is $1 - F_x(r)$.

Since the sample points are independent, these individual probabilities multiply. Applying a union bound over the $N-1$ possible choices for the index j gives the conditional bound:

$$\mathbb{P}\big(Z_1 = Z_{[1]},\ Z_{[2]} \leq r + 2\delta \mid Z_1 = r\big) \leq (N-1)\big[F_x(r+2\delta) - F_x(r)\big]\big(1 - F_x(r)\big)^{N-2}.$$

Finally, integrating this conditional probability with respect to the law of Z_1, whose distribution function is F_x, yields the claimed inequality:

$$\mathbb{P}_{\mathcal{D}_N}\big(D_2(x) - D_1(x) \leq 2\delta\big) \leq N(N-1)\int_0^\infty \big[F_x(r+2\delta) - F_x(r)\big]\big(1 - F_x(r)\big)^{N-2}\,\mathrm{d}F_x(r).$$

□

Proof of Lemma 4.4.2. If $N = 1$, then $V_1 = I_R$, meaning $\Omega_\delta = \{x \in I_R \mid \mathrm{dist}(x, \partial I_R) < \delta\}$. Since μ has a bounded density, we deduce

$$\mathbb{E}_{\mathcal{D}_N}[\mu(\Omega_\delta)] = \mu(\Omega_\delta) \lesssim \delta \lesssim \delta N^{1/d}.$$

We may thus assume $N \geq 2$. By Fubini's theorem, introducing an independent test point $x \sim \mu$, we can write

$$\mathbb{E}_{\mathcal{D}_N}\big[\mu(\Omega_\delta)\big] = \int_{I_R} \mathbb{P}_{\mathcal{D}_N}\big(x \in \Omega_\delta\big)\, \mathrm{d}\mu(x) = \mathbb{P}\big(x \in \Omega_\delta\big).$$

Furthermore, since probabilities are bounded by 1, the claim holds trivially if $\delta N^{1/d} \geq 1$, so we may restrict our analysis to the regime where $\delta N^{1/d} \leq 1$.

Let $D_1(x)$ and $D_2(x)$ denote the distances from x to its nearest and second-nearest points in $\mathcal{D}_N$. If $x \in \Omega_\delta$ with $\operatorname{dist}(x, \partial I_R) > \delta$, then it must lie within distance δ of an internal Voronoi boundary. This implies there exists a point z with $\|x - z\| \leq \delta$ that is equidistant to at least two distinct sample points. By the triangle inequality,

$$D_2(x) - D_1(x) \leq 2\|x - z\| \leq 2\delta.$$

Consequently, we can bound the total probability by splitting it as

$$\mathbb{P}\big(x \in \Omega_\delta\big) \leq \mathbb{P}\big(\operatorname{dist}(x, \partial I_R) \leq \delta\big) + \mathbb{P}\big(D_2(x) - D_1(x) \leq 2\delta\big). \tag{4.6.3}$$

By assumption, μ admits a density bounded from above by some $\rho_{\max} < \infty$. Since the volume of the δ-strip near ∂I_R scales as $O(\delta)$, the first term satisfies $\mathbb{P}(\operatorname{dist}(x, \partial I_R) \leq \delta) \lesssim \delta$.

To bound the second term, we condition on x and define $F_x(r) \coloneqq \mu(B(x, r))$. Because the density is uniformly bounded by $\rho_{\max}$, the mass of the spherical annulus is controlled by its volume:

$$F_x(r + 2\delta) - F_x(r) \lesssim \delta(r + 2\delta)^{d-1}. \tag{4.6.4}$$

Applying Lemma 4.6.1 with (4.6.4) and the inequality $(r + 2\delta)^{d-1} \lesssim r^{d-1} + \delta^{d-1}$, we obtain:

$$\begin{aligned}\mathbb{P}\big(D_2(x) - D_1(x) \leq 2\delta \mid x\big) &\lesssim N\delta\,(N-1) \int_0^\infty (r^{d-1} + \delta^{d-1})\big(1 - F_x(r)\big)^{N-2}\, \mathrm{d}F_x(r) \\ &= N\delta \cdot I_x \;+\; N\delta^d\,(N-1) \int_0^\infty \big(1 - F_x(r)\big)^{N-2}\, \mathrm{d}F_x(r),\end{aligned}$$

where

$$I_x \coloneqq (N-1) \int_0^\infty r^{d-1}\big(1 - F_x(r)\big)^{N-2}\, \mathrm{d}F_x(r).$$

The second integral trivially evaluates to $1/(N-1)$, meaning the second term simplifies to $N\delta^d$.

It remains to estimate I_x uniformly in $x \in I_R$. Using Stieltjes integration by parts and noting that the boundary terms vanish, we obtain

$$I_x = (d-1) \int_0^\infty r^{d-2}\big(1 - F_x(r)\big)^{N-1}\, \mathrm{d}r.$$

To bound this, note that the density of μ is bounded from below by some $\rho_{\min} > 0$ on I_R. Thus, even near the boundary, $|B(x, r) \cap I_R| \gtrsim r^d$ uniformly in $x \in I_R$ for small $r > 0$. Therefore, there exist constants $c, r_0 > 0$ such that $F_x(r) \geq cr^d$ for all $r \leq r_0$, uniformly in x. Using $1 - z \leq e^{-z}$, we obtain

$$\big(1 - F_x(r)\big)^{N-1} \leq e^{-c(N-1)r^d} \qquad \text{for } 0 \leq r \leq r_0.$$

Moreover, if we let $D_R \coloneqq \operatorname{diam}(I_R) = 2R\sqrt{d}$, then $F_x(r) = 1$ for all $r \geq D_R$. Hence,

$$I_x \lesssim \int_0^{r_0} r^{d-2} e^{-c(N-1)r^d}\, \mathrm{d}r \;+\; \int_{r_0}^{D_R} r^{d-2}\big(1 - F_x(r_0)\big)^{N-1}\, \mathrm{d}r.$$

To evaluate the first term, we apply the change of variables $u = (N-1)r^d$, which gives

$$\int_0^{r_0} r^{d-2} e^{-c(N-1)r^d} \, \mathrm{d}r = \frac{1}{d(N-1)^{\frac{d-1}{d}}} \int_0^{(N-1)r_0^d} u^{-1/d} e^{-cu} \, \mathrm{d}u.$$

By extending the upper limit of integration to infinity, the remaining integral over u converges to a finite constant independent of N. Since $N - 1 \asymp N$, this first term is bounded by $O(N^{-(d-1)/d})$.

The second integral decays exponentially fast in N since its base $1 - F_x(r_0)$ is uniformly bounded away from 1. It is therefore asymptotically negligible and trivially absorbed by the polynomial rate $N^{-(d-1)/d}$. We thus conclude that $I_x \lesssim N^{-(d-1)/d}$ uniformly in $x \in I_R$.

Substituting this bound back into our conditional probability, we find:

$$\mathbb{P}\big(D_2(x) - D_1(x) \leq 2\delta \mid x\big) \lesssim N\delta\Big(N^{-(d-1)/d}\Big) + N\delta^d = \delta N^{1/d} + N\delta^d.$$

Since we operate in the regime $\delta N^{1/d} \leq 1$ with $d \geq 2$, it follows that $N\delta^d = (\delta N^{1/d})^d \leq \delta N^{1/d}$. The second probability term is thus absorbed by $\delta N^{1/d}$. Finally, taking the expectation over x and plugging both bounds back into (4.6.3), we conclude:

$$\mathbb{E}_{\mathcal{D}_N}\big[\mu(\Omega_\delta)\big] \lesssim \delta + \delta N^{1/d} \lesssim \delta N^{1/d},$$

which completes the proof. □

Proof of Proposition 4.4.3

We first provide two preliminary lemmas. The first establishes that we can find a smooth, compactly supported, time-varying vector field that simultaneously steers the disjoint sets A_k to their respective target balls.

Lemma 4.6.2. *Fix $d \geq 2$ and $K \geq 1$. Let $A_k \subset \mathbb{R}^d$ be pairwise disjoint compact convex sets, and let $r_k \in \mathbb{R}^d$ be distinct target points, for $k \in [K]$. Then, for every $\eta > 0$ and $T > 0$, there exists a time-varying vector field $v \in \mathscr{C}_c^\infty([0,T] \times \mathbb{R}^d; \mathbb{R}^d)$ whose flow map $\Phi_T : \mathbb{R}^d \to \mathbb{R}^d$ satisfies*

$$\Phi_T(A_k) \subset B(r_k, \eta) \qquad \text{for all } k \in [K].$$

Proof. For each $k \in [K]$, pick an arbitrary point $x_k \in A_k$. Since the sets A_k are pairwise disjoint and compact, there exists $\rho > 0$ such that the open neighborhoods $V_k := \{x \in \mathbb{R}^d \mid \operatorname{dist}(x, A_k) < \rho\}$ are pairwise disjoint. Because each A_k is convex, V_k is also convex.

Choose a smooth cutoff function $\chi_k \in \mathscr{C}_c^\infty(V_k)$ such that $0 \leq \chi_k \leq 1$ and $\chi_k \equiv 1$ on A_k, and define the vector field $v_k(x) := \chi_k(x)(x_k - x)$. By construction, $v_k \in \mathscr{C}_c^\infty(V_k; \mathbb{R}^d)$, and for any $x \in A_k$, it simplifies to $v_k(x) = x_k - x$. The induced flow on A_k is therefore $x(t) = x_k + e^{-t}(x(0) - x_k)$. Since A_k is convex and $x_k \in A_k$, this trajectory remains in A_k for all $t \geq 0$.

We define the global contraction field as $v^C(x) := \lambda \sum_{k=1}^K v_k(x)$. Since the supports V_k are disjoint, v^C is smooth and compactly supported. Given that the flow on each A_k accelerates to

$$x(t) = x_k + e^{-\lambda t}(x(0) - x_k),$$

for any chosen $\xi > 0$, we can pick λ large enough to ensure that $\Phi^c_{T/2}(A_k) \subset B(x_k, \xi)$ for all k. Since $d \geq 2$, an argument similar to the one used to prove Proposition 3.3.13 (see **Subsection 3.5.2**)

yields pairwise disjoint, smooth paths $\gamma_k : [0,T] \to \mathbb{R}^d$ for $k \in [K]$, such that $\gamma_k(t) = x_k$ for all $t \in [0, 2T/3]$ and $\gamma_k(T) = r_k$. By the compactness of $[0,T]$ and the finiteness of K, we have

$$m \coloneqq \min_{t\in[0,T]} \min_{i\neq j} \|\gamma_i(t) - \gamma_j(t)\| > 0.$$

Finally, we fix the radius ξ small enough such that $4\xi < m$ and $\xi < \eta$.

Let $\chi \in \mathscr{C}_c^\infty(\mathbb{R}^d)$ be a radial bump function with $\chi(y) = 1$ for $\|y\| \le 1$ and $\chi(y) = 0$ for $\|y\| \ge 2$. For $t \in [0,T]$, we define the transport field:

$$v_t^T(x) \coloneqq \sum_{k=1}^{K} \dot\gamma_k(t)\, \chi\Big(\frac{x - \gamma_k(t)}{\xi}\Big).$$

Because $4\xi < m$, the supports of these summands are pairwise disjoint for all t. Furthermore, if $\|x(t) - \gamma_k(t)\| \le \xi$, the field evaluates to $\dot x(t) = \dot\gamma_k(t)$, meaning the distance $x(t) - \gamma_k(t)$ remains constant. Thus, any point starting inside $B(x_k, \xi)$ at time $2T/3$ is transported along the curve γ_k. Note that $v_t^T(x) = 0$ for all $t \le 2T/3$ since the path γ is constant there.

To splice the two stages smoothly, we introduce a nondecreasing function $\varphi \in \mathscr{C}^\infty([0,T])$ satisfying $\varphi(t) = 0$ for $t \le T/3$ and $\varphi(t) = 1$ for $t \ge 2T/3$. The final vector field is constructed as:

$$v_t(x) \coloneqq (1 - \varphi(t))\, v^C(x) + \varphi(t)\, v_t^T(x).$$

Clearly, $v \in \mathscr{C}_c^\infty([0,T] \times \mathbb{R}^d; \mathbb{R}^d)$. By choosing the initial contraction rate λ sufficiently large, we guarantee that $\Phi_{T/3}(A_k) \subset B(x_k, \xi)$ for all k. During the transition window $[T/3, 2T/3]$, the curves γ_k are constant (hence $v^T = 0$), and the flow remains inside these balls. For $t \ge 2T/3$, the balls are transported along the paths γ_k. At the final time T, we achieve exactly

$$\Phi_T(A_k) \subset B(r_k, \xi) \subset B(r_k, \eta),$$

concluding the proof. □

By applying [Li+26, Theorem 2.3]—whose proof leverages [KB18, Theorem 2] over the extended spatio-temporal domain—one obtains an $O(p^{-1/2})$ approximation bound for the flows generated by $(\mathcal{V}_{\text{sa}}^+)$. We restate this result below for completeness.

Lemma 4.6.3 ([Li+26, Theorem 2.3]). *Let $v \in \mathscr{C}_c^\infty([0,T] \times \mathbb{R}^d; \mathbb{R}^d)$, fix $T > 0$, and let $\Omega \subset \mathbb{R}^d$ be a compact set. Let $(\Psi_t)_{t\in[0,T]}$ be the flow induced by $\dot x = v_t(x)$. Then, for every integer $p \ge 1$, there exists a control θ of width p such that the flow Φ_t^θ generated by the semi-autonomous field v_θ of the form $(\mathcal{V}_{\text{sa}}^+)$ satisfies*

$$\sup_{x\in\Omega} \sup_{t\in[0,T]} \|\Psi_t(x) - \Phi_t^\theta(x)\| \;\le\; \frac{C_{T,\Omega,v}}{\sqrt{p}},$$

where $C_{T,\Omega,v} > 0$ is independent of p.

Proof of Proposition 4.4.3. Fix $\eta > 0$ and assume without loss of generality that $\eta \le 1$ (otherwise, replace η by 1). By slightly perturbing the targets r_k if necessary (by a distance strictly less than $\eta/4$), we may assume they are pairwise distinct. Since the sets $\{A_k\}_{k=1}^K$ are compact and pairwise disjoint, we can apply Lemma 4.6.2 with targets r_k and tolerance $\eta/2$ to obtain a time-varying vector field $v \in \mathscr{C}_c^\infty([0,T] \times \mathbb{R}^d; \mathbb{R}^d)$ whose flow Ψ_t satisfies

$$\Psi_T(A_k) \subset B(r_k, \eta/2) \qquad \text{for all } k \in [K].$$

Next, applying Lemma 4.6.3 with $\Omega_0 = I_R$, there exists a control θ of width p such that the flow $(\Phi_t^\theta)_{t\in[0,T]}$ generated by v_θ of the form ($\mathcal{V}_{\text{sa}}^+$) satisfies

$$\|\Phi_T^\theta - \Psi_T\|_{L^\infty(I_R)} \;\le\; \frac{C}{\sqrt{p}},$$

for a constant $C > 0$ independent of p. Choosing $p \ge (2C/\eta)^2$ ensures that this error is at most $\eta/2$. Consequently, for any $x \in A_k$,

$$\|\Phi_T^\theta(x) - r_k\| \le \|\Phi_T^\theta(x) - \Psi_T(x)\| + \|\Psi_T(x) - r_k\| \le \frac{\eta}{2} + \frac{\eta}{2} = \eta,$$

which proves statement (i).

To prove (ii), observe that since v is compactly supported in space, $\Psi_T(I_R)$ is bounded. Therefore,

$$\|\Phi_T^\theta\|_{L^\infty(I_R)} \le \|\Psi_T\|_{L^\infty(I_R)} + \|\Phi_T^\theta - \Psi_T\|_{L^\infty(I_R)} \le \|\Psi_T\|_{L^\infty(I_R)} + \frac{\eta}{2}.$$

Because Ψ_T is the identity outside a fixed compact set (determined by T, I_R, and the configuration of the sets A_k and targets r_k) and $\eta \le 1$, then $\|\Psi_T\|_{L^\infty(I_R)}$ is uniformly bounded by a constant that depends on the problem geometry but is completely independent of p and η. This yields the required uniform bound, concluding the proof. □

Proofs of Theorems 4.4.6 and 4.4.8

Proof of Theorem 4.4.6. Fix $h \in (0, 2R]$ and $\delta \in (0, h/4)$. Let $M_y := \|y\|_{L^\infty(I_R)}$ and note that the estimator $y_{N,h}$ defined in (4.3.6) satisfies $\|y_{N,h}\|_{L^\infty(I_R)} \le M_y$. Thus,

$$\mathcal{R}(\theta_N) = \|y - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2 \le 2\|y - y_{N,h}\|_{L^2(\mu)}^2 + 2\|y_{N,h} - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2. \tag{4.6.5}$$

Taking the expectation $\mathbb{E}_{\mathcal{D}_N}$ and applying Proposition 4.3.1, we directly obtain:

$$\mathbb{E}_{\mathcal{D}_N}\big[\mathcal{R}(\theta_N)\big] \lesssim \omega_y(\sqrt{d}\,h)^2 + \frac{M_y^2}{N\,h^d} + \mathbb{E}_{\mathcal{D}_N}\big[\|y_{N,h} - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2\big]. \tag{4.6.6}$$

It remains to find the data-dependent control θ_N and bound the last term. Let

$$Q_k^\delta := \{x \in Q_k \mid \operatorname{dist}(x, \partial Q_k) \ge \delta\}$$

be the trimmed core of each Q_k, and let $\Omega_\delta := I_R \setminus \bigcup_{k\in\mathcal{K}} Q_k^\delta$ be the boundary layer. We split:

$$\begin{aligned}\|y_{N,h} - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2 &= \sum_{k\in\mathcal{K}} \int_{Q_k^\delta} \|y_{N,h}(x) - \Phi_T^{\theta_N}(x)\|^2 \,\mathrm{d}\mu(x) \;+\; \int_{\Omega_\delta} \|y_{N,h}(x) - \Phi_T^{\theta_N}(x)\|^2 \,\mathrm{d}\mu(x) \\ &\le \sum_{k\in\mathcal{K}} \mu(Q_k^\delta) \sup_{x\in Q_k^\delta} \|y_{N,h}(x) - \Phi_T^{\theta_N}(x)\|^2 \;+\; \int_{\Omega_\delta} \|y_{N,h} - \Phi_T^{\theta_N}\|^2 \,\mathrm{d}\mu. \end{aligned} \tag{4.6.7}$$

On each Q_k^δ, the estimator evaluates constantly to the average S_k defined in (4.3.7). Since the cores are mutually disjoint, $\sum_{k\in\mathcal{K}} \mu(Q_k^\delta) \le 1$, which trivially implies:

$$\sum_{k\in\mathcal{K}} \mu(Q_k^\delta) \sup_{x\in Q_k^\delta} \|y_{N,h}(x) - \Phi_T^{\theta_N}(x)\|^2 \le \max_{k\in\mathcal{K}} \sup_{x\in Q_k^\delta} \|\Phi_T^{\theta_N}(x) - S_k\|^2. \tag{4.6.8}$$

For the boundary layer Ω_δ, expanding the squared norm yields

$$\int_{\Omega_\delta} \|y_{N,h} - \Phi_T^{\theta_N}\|^2 \, \mathrm{d}\mu \leq 2\big(M_y^2 + \|\Phi_T^{\theta_N}\|_{L^\infty(I_R)}^2\big)\, \mu(\Omega_\delta). \tag{4.6.9}$$

By construction, the sets $\{Q_k^\delta\}_{k\in\mathcal{K}}$ are compact, convex, and pairwise disjoint. We apply Proposition 4.4.3 setting the targets to $r_k \coloneqq S_k$. Since the target vectors satisfy $\|S_k\| \leq M_y$, Proposition 4.4.3(ii) ensures the flow satisfies $\|\Phi_T^{\theta_N}\|_{L^\infty(I_R)} \leq C$, where the constant C may depend on M_y but is strictly independent of the architecture width p and the margin δ. Using this uniform bound alongside Lemma 4.4.1, we control the boundary integral:

$$\int_{\Omega_\delta} \|y_{N,h} - \Phi_T^{\theta_N}\|^2 \, \mathrm{d}\mu \lesssim (M_y^2 + C^2)\, \frac{\delta}{h} \lesssim \frac{\delta}{h}. \tag{4.6.10}$$

Combining (4.6.7)–(4.6.8)–(4.6.10) gives

$$\|y_{N,h} - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2 \lesssim \frac{\delta}{h} + \max_{k\in\mathcal{K}} \sup_{x\in Q_k^\delta} \|\Phi_T^{\theta_N}(x) - S_k\|^2. \tag{4.6.11}$$

Setting the target vectors to $r_k \coloneqq S_k$ and choosing a tolerance $\eta \asymp p^{-1/2}$, Proposition 4.4.3(i) guarantees the existence of a control θ_N of width p such that $\Phi_T^{\theta_N}(Q_k^\delta) \subset B(S_k, \eta)$ for all $k \in \mathcal{K}$. Consequently, the uniform error on the cores satisfies

$$\max_{k\in\mathcal{K}} \sup_{x\in Q_k^\delta} \|\Phi_T^{\theta_N}(x) - S_k\|^2 \leq \eta^2 \lesssim \frac{1}{p}.$$

Plugging this bound into (4.6.11) and substituting back into (4.6.6) directly yields (4.4.5).

□

Proof of Theorem 4.4.8. Recall from **Subsection 4.3.2** that $y_N^V(x) \coloneqq y(x_{\mathrm{NN}}(x))$ satisfies

$$\mathbb{E}_{\mathcal{D}_N}\Big[\|y - y_N^V\|_{L^2(\mu)}^2\Big] \lesssim \left(\frac{\log N}{N}\right)^{2\alpha/d}. \tag{4.6.12}$$

To bound the realization error, we fix a margin $\delta > 0$ and define the trimmed core of each Voronoi cell as $V_i^\delta \coloneqq \{x \in V_i \mid \mathrm{dist}(x, \partial V_i) \geq \delta\}$, and the boundary layer $\Omega_\delta \coloneqq I_R \setminus \bigcup_{i=1}^N V_i^\delta$.

Because V_i^δ are compact, convex and pairwise disjoint, we can apply Proposition 4.4.3 to the family $\{V_i^\delta\}_{i=1}^N$, assigning to each core the target vector $r_i \coloneqq y(x_i)$ with tolerance $\eta \coloneqq p^{-1/2}$. This yields a data-dependent control θ_N of width p such that

$$\sup_{x\in V_i^\delta} \|\Phi_T^{\theta_N}(x) - y(x_i)\| \leq \eta \qquad \text{for all } i \in [N].$$

Since $y_N^V(x) = y(x_i)$ on V_i^δ, this guarantees a uniform bound $\sup_{x\in I_R\setminus\Omega_\delta} \|\Phi_T^{\theta_N}(x) - y_N^V(x)\| \leq \eta$ safely outside the boundary layer.

Splitting the $L^2(\mu)$ realization error between the trimmed cores and the boundary layer, we obtain

$$\|\Phi_T^{\theta_N} - y_N^V\|_{L^2(\mu)}^2 \leq \eta^2 + \int_{\Omega_\delta} \|\Phi_T^{\theta_N} - y_N^V\|^2 \, \mathrm{d}\mu. \tag{4.6.13}$$

By Proposition 4.4.3(ii), the flow $\Phi_T^{\theta_N}$ is uniformly bounded independently of p and η. Furthermore, since y is continuous on the compact domain I_R, the interpolant y_N^V is also uniformly bounded. Using $\|a-b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the integrand of (4.6.13) is bounded by a constant, leading to

$$\|\Phi_T^{\theta_N} - y_N^V\|_{L^2(\mu)}^2 \lesssim \eta^2 + \mu(\Omega_\delta).$$

Taking expectations and applying Lemma 4.4.2 with $\eta^2 = 1/p$, we find

$$\mathbb{E}_{\mathcal{D}_N}\left[\|\Phi_T^{\theta_N} - y_N^V\|_{L^2(\mu)}^2\right] \lesssim \frac{1}{p} + \delta N^{1/d}. \tag{4.6.14}$$

Finally, we split the total population risk by adding and subtracting the interpolant y_N^V. This yields

$$\mathcal{R}(\theta_N) = \|y - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2 \lesssim \|y - y_N^V\|_{L^2(\mu)}^2 + \|y_N^V - \Phi_T^{\theta_N}\|_{L^2(\mu)}^2.$$

Taking the expectation and substituting (4.6.12)–(4.6.14), we obtain

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \left(\frac{\log N}{N}\right)^{2\alpha/d} + \delta N^{1/d} + \frac{1}{p},$$

which establishes (4.4.6). □

Part II

Generative Modeling

Overview

Generative modeling is the task of producing *new samples* distributed according to an (unknown) *data distribution* $\mu_\star \in \mathcal{P}(\mathbb{R}^d)$, where d is typically very large (e.g., representing image pixels, molecular coordinates, or token embeddings). In contrast with *supervised learning*, one does not observe labels: the available information is a finite dataset assumed to be i.i.d. draws from $\mu_\star$,

$$\{x_i\}_{i=1}^N \subset \mathbb{R}^d \qquad \text{or, equivalently,} \qquad \mu_N \coloneqq \frac{1}{N}\sum_{i=1}^N \delta_{x_i}.$$

The modeling goal is to construct a *sampler* whose outputs resemble the data beyond the training set, i.e. the output law must be close to $\mu_\star$. Modern dynamical generative models often frame this as a *probability transport problem*—a perspective deeply rooted in computational optimal transport [PC19]: one starts from a simple *base law* μ_{B} that is easy to sample in any dimension—most commonly $\mu_{\mathrm{B}} = \mathcal{N}(0, I_d)$—and learns some (deterministic or stochastic) dynamics that push μ_{B} forward to (an approximation of) $\mu_\star$. This is the principle of *flow-based generative modeling*.

Generative modeling by controlled transport dynamics.

Fix $T > 0$ and a family of vector fields $\{v_\theta\}_{\theta\in\Theta}$. Given a control input $\theta(\cdot)\colon [0,T] \to \Theta$, we consider the *controlled continuity equation*

$$\partial_t \mu_t^\theta + \operatorname{div}\left(v_{\theta(t)}\, \mu_t^\theta\right) = 0, \qquad \mu^\theta_{|t=0} = \mu_{\mathrm{B}}. \tag{CE}$$

As recalled in **Section 1.2**, the solution is $\mu_t^\theta = (\Phi_t^\theta)_\# \mu_{\mathrm{B}}$, where Φ_t^θ is the flow map of (CF). Accordingly, sampling from the model is implemented by the (Lagrangian) procedure

$$\text{draw } X_0 \sim \mu_{\mathrm{B}} \qquad \rightsquigarrow \qquad \text{solve } \dot{X}_t = v_{\theta(t)}(X_t) \qquad \rightsquigarrow \qquad \text{output } X_T, \tag{II.1}$$

so that the model distribution is μ_T^θ. In this dynamical language, learning a generator amounts to selecting $\theta(\cdot)$ so that μ_T^θ approximates $\mu_\star$ in a discrepancy chosen according to the specific task. Most generative pipelines assume $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Since the Lipschitz flows we consider preserve absolute continuity, we may adopt a density-based formulation $\mu_t^\theta = \rho_t^\theta\, \mathcal{L}^d$, governed by

$$\partial_t \rho_t^\theta + \operatorname{div}\left(v_{\theta(t)} \rho_t^\theta\right) = 0, \qquad \rho^\theta_{|t=0} = \rho_{\mathrm{B}}. \tag{DCE}$$

This setting is analytically convenient as it enables the use of identities (1.2.2)–(1.2.3) and the L^p bounds in (1.2.4).

A common template across flow-based generators

Many modern generators can be understood as different ways of *learning* a time-dependent transport field for (DCE). They mainly differ in the *training objective* used to fit the dynamics.

(i) ***Continuous normalizing flows*** (CNFs) work in the setting of (CE)–(II.1); see [Che+18]. The invertibility of Φ_T^θ enables exact likelihood evaluation via (1.2.3). In practice, the computationally

expensive divergence term is replaced by the Hutchinson trace estimator $\operatorname{div} v_\theta(x) = \mathbb{E}_\epsilon[\epsilon^\top \nabla v_\theta(x)\epsilon]$ with $\epsilon \sim \mathcal{N}(0, I_d)$, as proposed in [Gra+19]. Training is then performed by *maximum likelihood* (or its empirical counterpart over a dataset), which corresponds at the population level to minimizing the *Kullback–Leibler divergence*:

$$\min_{\theta(\cdot)} \ \mathrm{KL}\big(\mu_\star \,\|\, \mu_T^\theta\big) = \min_{\theta(\cdot)} \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu_T^\theta}(x)\right) \mathrm{d}\mu_\star(x). \tag{II.2}$$

(ii) ***Flow matching*** retains the generative mechanism of CNFs—sampling is still performed via (II.1)—but replaces likelihood training by regression on vector fields. Rather than learning the curve μ_t^θ implicitly, one defines it by interpolating between a noise sample $x_0 \sim \mu_{\mathrm{B}}$ and a data sample $x_1 \sim \mu_\star$ (typically via a straight line $x_t = (1-t)x_0 + tx_1$). While the global vector field generating the mixture of all such paths is intractable, the velocity u_t of a single path is trivial to compute (e.g., $u_t = x_1 - x_0$ for the linear case). *Conditional flow matching* trains $v_{\theta(t)}$ to match this simple individual velocity on average:

$$\min_{\theta(\cdot)} \ \mathbb{E}_{t,x_0,x_1}\Big[\big\|v_{\theta(t)}(x_t) - u_t(x_t \mid x_0, x_1)\big\|^2\Big]. \tag{II.3}$$

Theoretical results show that the optimal solution to (II.3) coincides with the vector field driving the marginal path μ_t^θ; see [Lip+23; Ton+24].

(iii) ***Diffusion models*** define a forward noising process (an SDE) that maps the data $\mu_\star$ to a noise law close to μ_{B}. At the density level, ρ_t evolves according to the *Fokker–Planck equation*:

$$\partial_t \rho_t = -\operatorname{div}\big(f(\cdot,t)\,\rho_t\big) + \tfrac{1}{2}g(t)^2 \Delta \rho_t. \tag{II.4}$$

Sampling is typically performed by simulating the *reverse-time SDE*, which requires knowing the *score function* $s(t,x) = \nabla_x \log \rho_t(x)$. In practice, one uses an approximation $s_\theta(t,x) \approx s(t,x)$ that is learned from data by (variants of) denoising score matching [Hyv05; Son+21].

The learned score also gives rise to a deterministic dynamics with the same one-time marginals as the diffusion process. Indeed, using the identity $\Delta\rho_t = \operatorname{div}(\rho_t \nabla \log \rho_t) = \operatorname{div}(\rho_t s)$, one can rewrite (II.4) as a conservative transport equation, the *probability flow ODE*

$$\dot{x}(t) = f(x(t),t) - \tfrac{1}{2}g(t)^2 s_\theta(x(t),t). \tag{II.5}$$

When the score is exact, this ODE has the same marginal densities as the diffusion process; in practice, one integrates the corresponding dynamics backward in time to generate samples from a noise initialization. This brings *score-based diffusion* models effectively within the scope of (DCE).

Mathematical focus of Part II.

The central question concerns *approximate controllability*: *given a target $\mu_\star$, can one choose $\theta(\cdot)$ so that μ_T^θ approximates $\mu_\star$, and how does the answer depend on the discrepancy used (e.g., Wasserstein distances, total variation, or relative entropy)?*

Chapter 5 addresses this problem in the setting of general controlled flows for Wasserstein and total variation. **Chapter 6** then specializes to neural-ODE parameterizations and develops constructive control designs aimed at decreasing the Kullback–Leibler divergence.

Roadmap of Part II

For navigation, **Part II** consists of the following two chapters.

- **Chapter 5**: Controllability for Measure Transport.
- **Chapter 6**: Continuous Normalizing Flows.

Chapter 5

Controllability for Measure Transport

5.1 Introduction

As discussed in **Part II (Overview)**, a number of modern generative methods can be unified under a common framework as *probability transport* problems: starting from a tractable *base* law μ_{B}, one seeks a dynamical system whose terminal-time map pushes μ_{B} forward to a *target* law $\mu_\star$ representing the data distribution. At the level of distributions, this task is equivalent to selecting an appropriate control in a continuity equation, such as (CE).

The objective of this chapter is to study this controllability problem at the level of measures, emphasizing its structural properties. Given a family $\mathcal{V}$ of controlled vector fields, we investigate which properties of $\mathcal{V}$ are sufficient to guarantee a *universal approximation property* (UAP) on the space of probability measures. We focus on the expressivity of $\mathcal{V}$ achieved purely through composition over time—a depth-oriented perspective aligned with [Che+25c]—as opposed to classical approximation theorems that attain density by increasing width (e.g., adding as many neurons as necessary). Furthermore, we explore the interplay between this UAP and the *simultaneous controllability* (recall the UIP) studied in **Chapter 3**, as well as how these results depend on the specific discrepancy chosen to compare measures.

This viewpoint differs from the classical function-level UAP [LLS22; Che+25c; LLS24; RZ23; DC25; Zha+20; Ish+23], where the question is whether attainable flow maps approximate arbitrary target maps. Here, the base law μ_{B} is fixed, and the question is whether its attainable pushforwards are dense in a suitable class of target measures for the chosen discrepancy.

A key structural difference is that measures are invariant under particle relabeling. This makes a *permutation-invariant* version of the UIP natural in the measure setting. For instance, in $d = 1$, every flow map is strictly increasing (see Remark 1.1.2)—a rigidity that obstructs function-level universality, but does not prevent the approximation of a target probability measure. Since empirical measures do not encode labels, one may match two point clouds up to permutation without changing the induced measure.

Formally, if $\mathsf{P} : \mathbb{R}^d \to \mathbb{R}^d$ is a μ_{B}-measure-preserving map (i.e., $\mathsf{P}_\# \mu_{\mathrm{B}} = \mu_{\mathrm{B}}$), then

$$(\Phi \circ \mathsf{P})_\# \mu_{\mathrm{B}} = \Phi_\#(\mathsf{P}_\# \mu_{\mathrm{B}}) = \Phi_\# \mu_{\mathrm{B}} \qquad \text{for every measurable } \Phi : \mathbb{R}^d \to \mathbb{R}^d.$$

Consequently, we are inherently free to reorder the mass of μ_{B} to accommodate the monotonicity of the flow Φ without changing the resulting transported measure. In the empirical setting, this corresponds to permuting particles; more generally, it corresponds to pre-composition with measure-preserving maps.

Related work. The measure-transport perspective adopted here is close in spirit to several recent developments. In particular, [RZ23] investigates controllability in Wasserstein-type metrics for the continuity equation driven by $(\mathcal{V}_1^+)$, while the companion work [RZ24] establishes L^1 (and thus TV) density for the same ansatz. Very recently, [GR26] obtained a constructive approximation result, simultaneously approximating a target diffeomorphism and the associated pushforward measure. In this chapter, we generalize these findings to the broader framework of controlled flows, identifying structural properties that yield such results.

Universal approximation for *invertible* neural architectures is addressed from a geometric perspective in [Ish+23]. There, a strong (Sobolev) notion of functional universality for diffeomorphisms is established and shown to imply distributional universality, including convergence in total variation under suitable assumptions. Conceptually, this relates to the shearing mechanisms we exploit in **Section 5.4**.

We also mention [Mar+24], which provides a nonparametric statistical framework for learning distributions via controlled flows, including quantitative results on approximation and stability.

5.2 Setup and main results

Fix $d \geq 1$. Let Θ be a parameter space and consider a (Lipschitz) *control family*

$$\mathcal{V} := \big\{ v_\theta \colon \mathbb{R}^d \to \mathbb{R}^d \mid \theta \in \Theta \big\} \subset \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d). \tag{5.2.1}$$

For each time horizon $T > 0$ and each control input $\theta \in L^\infty(0, T; \Theta)$, we consider the controlled dynamical system

$$\dot{x}(t) = v_{\theta(t)}\big(x(t)\big), \qquad x(0) = x_0. \tag{CF}$$

Under these regularity assumptions, (CF) induces a unique flow map $\Phi_t^\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ for every $t \in [0, T]$. Furthermore, each Φ_t^θ is a bi-Lipschitz homeomorphism whose Lipschitz constants satisfy the estimates derived in Proposition 1.1.1 from **Chapter 1**.

Lifting the dynamics to the space of probability measures, we fix a base law $\mu_{\mathrm{B}} \in \mathcal{P}(\mathbb{R}^d)$ and track its evolution via the pushforward:

$$\mu_t^\theta := (\Phi_t^\theta)_{\#}\mu_{\mathrm{B}}, \qquad t \in [0, T].$$

As discussed in **Section 1.2**, the resulting curve $t \mapsto \mu_t^\theta$ constitutes the unique weak solution to the continuity equation (CE) initialized at μ_{B}.

To emphasize the dependence on $\mathcal{V}$, for each $T > 0$ we introduce the sets of *attainable maps*

$$\begin{aligned} \mathcal{A}_{\mathcal{V},T} &:= \big\{ \Phi_T^\theta \colon \mathbb{R}^d \to \mathbb{R}^d \mid \theta \in L^\infty(0, T; \Theta),\ v_{\theta(t)} \in \mathcal{V} \text{ for a.e. } t \in [0, T] \big\}, \\ \mathcal{A}_{\mathcal{V}} &:= \bigcup_{T \geq 0} \mathcal{A}_{\mathcal{V},T}, \end{aligned}$$

and the corresponding sets of *attainable measures*:

$$\begin{aligned} \mathcal{A}_{\mathcal{V},T}(\mu_{\mathrm{B}}) &:= \big\{ \mu_T^\theta \in \mathcal{P}(\mathbb{R}^d) \mid \theta \in L^\infty(0, T; \Theta),\ v_{\theta(t)} \in \mathcal{V} \text{ for a.e. } t \in [0, T] \big\}, \\ \mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}}) &:= \bigcup_{T \geq 0} \mathcal{A}_{\mathcal{V},T}(\mu_{\mathrm{B}}). \end{aligned}$$

It is worth emphasizing that any map in $\mathcal{A}_{\mathcal{V}}$ or measure in $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ is attained in finite time for some specific control θ. In particular, these attainable sets do not include asymptotic limits as $T \to \infty$ (unless such limits are also reached in finite time).

Natural examples of control families are the neural ODE architectures studied in **Chapter 3**, all of which are *affine invariant* in the following sense.

Definition 5.2.1. We say that a control family $\mathcal{V}$ on $\mathbb{R}^d$ is *affine invariant* if

$$f \in \mathcal{V} \quad \Longrightarrow \quad \big(x \mapsto W f(Ax+b)\big) \in \mathcal{V} \quad \text{for all } W, A \in \mathscr{M}_{d\times d}(\mathbb{R}),\ b \in \mathbb{R}^d.$$

Remark 5.2.1. This structural property plays an important role in the expressivity of the family. As established in [Che+25c, Theorem 2.6], for $d \geq 2$, a system driven by an affine-invariant family possesses the UIP if and only if $\mathcal{V}$ contains at least one nonlinear vector field. This characterization also extends to $d = 1$, provided the interpolation points respect the strict order preservation of one-dimensional flows (i.e., $x_i < x_{i+1}$ implies $y_i < y_{i+1}$ for all i).

While the UIP guarantees exact controllability for finite sets of points, transitioning to the measure level requires topological care. For instance, since Lipschitz flows map atoms to atoms (recall Proposition 1.2.1), exactly reaching a continuous target from a discrete one is precluded. Depending on the chosen discrepancy, even approximate reachability may fail. Consequently, universality statements must be phrased relative to a specific discrepancy-dependent topology and restricted to target measures that are at least approximable by general Lipschitz transport.

To formalize this, we fix a *discrepancy* $\mathsf{D} : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ satisfying:

(i) $\mathsf{D}(\mu, \nu) = 0$ if and only if $\mu = \nu$; and

(ii) if $\mathsf{D}(\mu_n, \mu) \to 0$ as $n \to \infty$, then μ_n *converges weakly* to μ (denoted by $\mu_n \rightharpoonup \mu$), meaning

$$\int \varphi \, \mathrm{d}\mu_n \to \int \varphi \, \mathrm{d}\mu \qquad \text{for all } \varphi \in \mathscr{C}_b(\mathbb{R}^d).$$

This setting covers standard distances as well as extended-valued, non-symmetric information losses (e.g., the Kullback–Leibler divergence as in (II.2)). In this chapter, however, we will focus on the Wasserstein and total variation distances.

To ensure the approximation problem is well-posed, we introduce the following notion.

Definition 5.2.2. Given a discrepancy D, we say that the pair $(\mu_{\mathrm{B}}, \mu_\star)$ is D-*compatible* if

$$\mu_\star \in \overline{\mathcal{A}_{\mathscr{C}^{0,1}}(\mu_{\mathrm{B}})}^{\mathsf{D}}, \tag{5.2.2}$$

where $\mathcal{A}_{\mathscr{C}^{0,1}}(\mu_{\mathrm{B}})$ denotes the set of measures reachable from μ_{B} via globally Lipschitz vector fields. Equivalently, there exists a sequence of times $t_k > 0$ and flow maps $\Psi_{t_k} \in \mathcal{A}_{\mathscr{C}^{0,1}}$ such that

$$\mathsf{D}\big(\mu_\star, (\Psi_{t_k})_{\#}\mu_{\mathrm{B}}\big) \to 0 \qquad \text{as } k \to \infty.$$

Note that Definition 5.2.2 depends solely on the topology induced by D and the intrinsic structural properties of Lipschitz flows, independent of any specific choice of the control family. With this framework in place, we can rigorously define the notion of approximation relative to $\mathcal{V}$.

Definition 5.2.3. We say that the controlled system driven by $\mathcal{V}$ possesses the D-*universal approximation property* (D-UAP) if for every D-compatible pair $(\mu_{\mathrm{B}}, \mu_\star) \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ and every $\varepsilon > 0$, there exists $\nu \in \mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ such that

$$\mathsf{D}(\nu, \mu_\star) < \varepsilon. \tag{5.2.3}$$

Equivalently, the closure $\overline{\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})}^{\mathsf{D}}$ contains every D-compatible target measure.

If we restrict the class of admissible base measures to a specific subset $\mathcal{F}$, we say that the system possesses the D-*UAP on* $\mathcal{F}$ provided the approximation property holds for all base measures $\mu_{\mathrm{B}} \in \mathcal{F}$ and all their corresponding D-compatible targets $\mu_\star$. For instance, we refer to the D-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ when considering only $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

We now present our main results, organized into two sections based on the chosen discrepancy: **Section 5.3** for Wasserstein (W_p) and **Section 5.4** for the total variation distance (TV).

Wasserstein (Section 5.3). Our first result relates Wasserstein approximation to point interpolation when particle labels are allowed to be permuted. More precisely, Theorem 5.3.4 shows that the permutation-invariant relaxation of (an approximate version of) the UIP, denoted π-UIP, is equivalent to the W_p-UAP on empirical measures with equal weights and pairwise distinct atoms. At the level of general probability measures, the converse fails: a (divergence-free) system may possess the UIP (hence also the π-UIP) and still fail the W_p-UAP, see Proposition 5.3.5. Finally, in Theorem 5.3.8 we show that uniform control of the Lipschitz constants of the interpolating maps is exactly the additional ingredient needed to pass from the π-UIP to Wasserstein approximation for general measures.

We then provide a structural lifting principle from particle interpolation to measure approximation via the L^p-approximation of transport maps. Building on [Che+25c, Prop. 3.9], Proposition 5.3.11 yields the W_p-UAP under a shrinking condition, which in turn yields the affine-invariant universality criterion in Corollary 5.3.13.

Total variation (Section 5.4). In the total variation metric, empirical discretization fails to approximate non-atomic measures. Universality therefore requires controlling the evolution of densities rather than merely transporting mass. For dimension $d = 1$, this reduces to approximating monotone rearrangements in a strong topology, leading to Proposition 5.4.3 and its affine-invariant consequence, Corollary 5.4.5.

For dimensions $d \geq 2$, one must additionally capture *dependencies between coordinates*. This is achieved through *shear flows*, which provide a coordinate-mixing mechanism beyond purely component-wise transformations. Under assumptions (A0)–(A2), in Theorem 5.4.8 we establish the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ and verify these assumptions for standard neural ODE families in Corollary 5.4.9.

At a high level, the proof of Theorem 5.4.8 proceeds by reducing density approximation to a sequence of elementary steps: (i) coordinate-wise operations that realize the required one-dimensional rearrangements on suitable conditional distributions, and (ii) shear-induced mixing steps that propagate these adjustments across coordinates, thereby recovering genuinely multivariate target densities in TV.

To keep the narrative focused on the core conceptual statements, all proofs are deferred to **Section 5.6**. Proofs for the Wasserstein results are collected in **Subsection 5.6.1**, structured to mirror **Subsection 5.3.1** and **Subsection 5.3.2**, while proofs regarding total variation are gathered in **Subsection 5.6.2**.

5.3 Wasserstein distance

For $p \in [1, \infty)$, the *Wasserstein p-distance* between $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ is defined by

$$W_p(\mu, \nu) \coloneqq \inf_{\pi \in \Pi(\mu,\nu)} \left(\int \|x - y\|^p \, \mathrm{d}\pi(x, y) \right)^{1/p} \tag{5.3.1}$$

where $\Pi(\mu, \nu)$ denotes the set of couplings with marginals μ and ν:

$$\Pi(\mu, \nu) \coloneqq \big\{ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \mid \pi(\cdot, \mathbb{R}^d) = \mu, \ \pi(\mathbb{R}^d, \cdot) = \nu \big\}. \tag{5.3.2}$$

The distance $W_p(\mu, \nu)$ is finite if and only if $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, that is,

$$\int \|x\|^p \, \mathrm{d}\mu(x) + \int \|x\|^p \, \mathrm{d}\nu(x) < +\infty. \tag{5.3.3}$$

We now recall two well-known properties of the Wasserstein distance that will be used repeatedly:

1. For any $f \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ and any $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the pushforward operation expands the distance by at most the Lipschitz constant of f:

   $$W_p(f_{\#}\mu, f_{\#}\nu) \leq \mathrm{Lip}(f) \, W_p(\mu, \nu). \tag{5.3.4}$$

   This follows by pushing forward an optimal coupling between μ and ν through f.

2. Let $\mu_N = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}$ and $\nu_N = \frac{1}{N} \sum_{i=1}^N \delta_{y_i}$. For any $\pi \in S_N$, we have

   $$\frac{1}{N} \sum_{i=1}^N \delta_{(x_i, y_{\pi(i)})} \in \Pi(\mu_N, \nu_N),$$

   which yields an upper bound for $W_p^p(\mu_N, \nu_N)$. Minimizing over all permutations in S_N yields:

   $$W_p^p(\mu_N, \nu_N) = \min_{\pi \in S_N} \frac{1}{N} \sum_{i=1}^N \|x_i - y_{\pi(i)}\|^p. \tag{5.3.5}$$

With these preliminaries in place, we begin by recording two necessary conditions for W_p-compatibility (in the sense of Definition 5.2.2):

Proposition 5.3.1. *Fix $d \geqslant 1$ and $p \in [1, \infty)$, and assume that $(\mu_{\mathrm{B}}, \mu_\star) \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ is W_p-compatible in the sense of Definition 5.2.2. Then the following hold:*

(W1) $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_p(\mathbb{R}^d)$.

(W2) *If μ_{B} is supported on N points, then $\mu_\star$ is supported on at most N points. In particular, an atomless target measure cannot be W_p-compatible with a finitely supported base measure.*

Remark 5.3.2. By Proposition 1.2.1, each individual attainable measure preserves the atomic masses of μ_{B}. However, this need not remain true in the W_p-closure, since distinct atoms may merge in the limit. In standard generative modeling applications, the base measure μ_{B} is typically chosen to be atomless, so this obstruction is usually irrelevant.

The link between the interpolation of measures and the control of finite points relies on two fundamental concepts: the ability to approximate continuous distributions via discrete samples, and the flexibility to steer these samples without tracking individual particle identities.

The first ingredient is the approximation by empirical measures. For any $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, let $(X_i)_{i\geq 1}$ be a sequence of independent samples drawn from μ. Then, with probability 1,

$$\mu_N \coloneqq \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i} \qquad \text{satisfies} \qquad W_p(\mu_N, \mu) \to 0 \quad \text{as } N \to \infty. \tag{5.3.6}$$

This property is not unique to W_p; it holds naturally for any discrepancy D that metrizes weak convergence on a given class of measures (meaning $\mathsf{D}(\mu_n, \mu) \to 0$ if and only if $\mu_n \rightharpoonup \mu$).

To maintain narrative flow, the formal statement of this fact is deferred to **Subsection 5.6.1** (Lemma 5.6.2). It covers, for example, the bounded-Lipschitz and Prokhorov distances, as well as the maximum mean discrepancy with a Gaussian kernel [SFL11; Gre+12]. Since W_p metrizes weak convergence plus convergence of the p-th moment (equivalently, uniform p-integrability) on $\mathcal{P}_p(\mathbb{R}^d)$, the limit (5.3.6) holds, and explicit rates of convergence are available; see [FG15].

The second key concept adapts the UIP from **Chapter 3** to better suit the measure-theoretic setting by allowing for the relabeling of particles. Accordingly, we keep the same notion of admissibility as in **Chapter 3**: the source and target points are required to be pairwise distinct, but the target labels are now regarded as irrelevant up to permutation.

Definition 5.3.1. We say that the system associated with a control family $\mathcal{V}$ possesses the *approximate permutation-UIP* if for every $N \in \mathbb{N}$, every $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$, and every $\varepsilon > 0$, there exist $\Phi \in \mathcal{A}_{\mathcal{V}}$ and a permutation $\pi \in S_N$ such that

$$\max_{1\leqslant i\leqslant N} \|\Phi(x_i) - y_{\pi(i)}\| < \varepsilon. \tag{5.3.7}$$

When (5.3.7) holds with $\varepsilon = 0$, we say that the system possesses the *permutation-UIP* (π-UIP).

The π-UIP is the unlabeled counterpart of the UIP studied in **Chapter 3**: the source points remain labeled, while the target cloud is only prescribed up to permutation. In particular, the UIP trivially implies the π-UIP. The converse is subtler and depends on both the dimension and the structure of the control family.

- If $d = 1$, the two notions are fundamentally different. Since every one-dimensional flow map is strictly increasing (Remark 1.1.2), the UIP fails, whereas the π-UIP may still hold because one is free to relabel the target points to match the natural order of the input points.
- If $d \geq 2$, however, the two notions may coincide depending on the control family. For instance, an affine-invariant family possessing the π-UIP cannot consist solely of affine vector fields; if it did, all attainable maps would be affine, and therefore could not interpolate arbitrary finite point clouds even up to permutation. Consequently, the family must contain at least one nonlinear vector field, which, by Remark 5.2.1, guarantees that the system possesses the UIP.

For general autonomous control families of the form (5.2.1), we are not aware of any counterexample showing that the π-UIP is strictly weaker than the UIP for $d \geq 2$. However, the following example shows that this implication can fail outside the autonomous framework of this chapter.

Proposition 5.3.3. *Let $d \geq 2$. Let $\mathcal{V}$ be the class of those $v \in \mathscr{C}^0([0,1]; \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d))$ that can be decomposed along some direction $u \in \mathbb{S}^{d-1}$ as*

$$v_t(x) = g_t(\langle x, u\rangle)\, u + w_t(x),$$

where $g \in \mathscr{C}^0([0,1]; \mathscr{C}^{0,1}(\mathbb{R}))$*, and*

$$w \in \mathscr{C}^0([0,1]; \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)) \qquad \textit{satisfies} \qquad \langle w_t(x), u \rangle = 0 \qquad \textit{for all } (t,x).$$

If the admissible maps $\mathcal{A}_{\mathcal{V}}$ *are the time-*1 *flows generated by individual fields* $v \in \mathcal{V}$*, then the associated control system possesses the* π*-UIP, but not the UIP.*

The mechanism behind this counterexample is straightforward: fixing a single direction u throughout the entire evolution imposes a global one-dimensional monotonicity along that axis, which obstructs the UIP. We do not know how to reproduce this construction if $\mathcal{V}$ consists of autonomous fields. In that case, since admissible flows are obtained by concatenating time-independent dynamics, the distinguished direction u may change from one time interval to the next, thereby breaking the global order preservation.

5.3.1 Empirical approximation

The first natural question is whether the ability to transport entire probability measures guarantees the ability to move finite point clouds. The following result gives a precise answer: point controllability up to permutation is equivalent to the W_p-UAP on empirical measures with equal weights and pairwise distinct atoms.

More precisely, let

$$\mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d) := \left\{ \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i} \;\middle|\; N \geq 1,\ x_i \neq x_j \text{ for } i \neq j \right\}.$$

The point is that a W_p-compatible target for an empirical base in $\mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d)$ need not itself have the same number of distinct atoms. However, by compatibility, it can be approximated by reachable empirical measures with exactly the same number of pairwise distinct atoms, which is precisely the setting where the approximate π-UIP applies.

Theorem 5.3.4. *Fix* $p \in [1,\infty)$ *and* $d \geq 1$*. Let* $\mathcal{V}$ *be a Lipschitz control family on* $\mathbb{R}^d$*. Then the system possesses the approximate* π*-UIP iff the system possesses the* W_p*-UAP on* $\mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d)$*.*

In particular, the full W_p-UAP on $\mathcal{P}_p(\mathbb{R}^d)$ implies the approximate π-UIP. The converse, however, is false at the level of general probability measures: merely interpolating points does not guarantee measure-level approximation, as the flow may satisfy structural constraints at the level of measures that are invisible on finite point clouds.

This is illustrated by the next result. By Theorem 3.3.1, the control family (5.3.8) still possesses the UIP despite the zero-divergence constraint. Yet this same incompressibility prevents the system from having the W_p-UAP, since volume-preserving flows cannot create the concentration needed to match certain target measures.

For the sake of simplicity, we state the result for vector fields as in ($\mathcal{V}_1^+$) (single neuron), though the conclusion naturally holds for any fixed width.

Proposition 5.3.5. *Assume* $d \geq 2$*, fix* $p \in [1,\infty)$*, and consider the control family*

$$\mathcal{V} = \left\{ x \mapsto v_\theta(x) = w\,(a \cdot x + b)_+ \;\middle|\; \theta = (w,a,b) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R},\quad w \cdot a = 0 \right\}. \tag{5.3.8}$$

Then the associated system possesses the UIP (and hence the π*-UIP), but not the* W_p*-UAP.*

The geometric obstruction illustrated by Proposition 5.3.5 is not exclusive to (5.3.8); it applies generally to any divergence-free control family whose associated system possesses the π-UIP. Given that such a system can still interpolate any finite number of points, it is instructive to observe exactly how this lack of W_p-UAP manifests in practice.

As the following remark demonstrates, if we rely on the (approximate) π-UIP to match increasingly dense empirical measures, any lack of W_p-UAP inevitably forces the Lipschitz constant of the interpolating flow maps to blow up as the sample size increases.

Remark 5.3.6. Suppose the system does not possess the W_p-UAP, so there exists a W_p-compatible pair $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_p(\mathbb{R}^d)$ and $\delta > 0$ such that

$$W_p\big(\Phi_\# \mu_{\mathrm{B}}, \mu_\star\big) \geq \delta \qquad \text{for every } \Phi \in \mathcal{A}_{\mathcal{V}}.$$

Recalling (5.3.6), we may take empirical measures

$$\mu_{\mathrm{B}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i} \quad \text{and} \quad \mu_\star^N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i} \qquad \text{with} \qquad W_p(\mu_{\mathrm{B}}^N, \mu_{\mathrm{B}}) \to 0 \quad \text{and} \quad W_p(\mu_\star^N, \mu_\star) \to 0.$$

If μ_{B} and $\mu_\star$ are atomless (as in the proof of Proposition 5.3.5), then the corresponding point clouds are admissible with probability 1.

On the other hand, suppose the system possesses the approximate π-UIP: for each N and $\varepsilon_N > 0$, we can find $\Phi_N \in \mathcal{A}_{\mathcal{V}}$ and $\pi \in S_N$ such that $\max_i \|\Phi_N(x_i) - y_{\pi(i)}\| < \varepsilon_N$. By (5.3.5), we obtain

$$W_p^p\big((\Phi_N)_\# \mu_{\mathrm{B}}^N, \mu_\star^N\big) \leq \frac{1}{N}\sum_{i=1}^N \|\Phi_N(x_i) - y_{\pi(i)}\|^p < \varepsilon_N^p.$$

By choosing a sequence with $\varepsilon_N \to 0$, it necessarily follows that

$$\mathrm{Lip}(\Phi_N) \to \infty \qquad \text{as } N \to \infty.$$

Indeed, by the triangle inequality,

$$\delta \leq W_p\big((\Phi_N)_\# \mu_{\mathrm{B}}, \mu_\star\big) \leq W_p\big((\Phi_N)_\# \mu_{\mathrm{B}}, (\Phi_N)_\# \mu_{\mathrm{B}}^N\big) + W_p\big((\Phi_N)_\# \mu_{\mathrm{B}}^N, \mu_\star\big).$$

Using (5.3.4) and the triangle inequality again, we obtain

$$\delta \leq \mathrm{Lip}(\Phi_N)\, W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N) + W_p\big((\Phi_N)_\# \mu_{\mathrm{B}}^N, \mu_\star^N\big) + W_p(\mu_\star^N, \mu_\star).$$

Therefore,

$$\mathrm{Lip}(\Phi_N) \geq \frac{\delta - W_p\big((\Phi_N)_\# \mu_{\mathrm{B}}^N, \mu_\star^N\big) - W_p(\mu_\star^N, \mu_\star)}{W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N)}.$$

The numerator converges to $\delta > 0$, whereas the denominator tends to 0, hence $\mathrm{Lip}(\Phi_N) \to \infty$.

The remark above motivates a quantitative strengthening of the π-UIP that controls $\mathrm{Lip}(\Phi)$ in terms of the minimal Lipschitz cost of matching the two point clouds.

Definition 5.3.2. We say that the system associated with a control family $\mathcal{V}$ possesses the Γ-*quantitative π-UIP* if there exists a nondecreasing function $\Gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for every

$N \in \mathbb{N}$, every $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$, and every $\varepsilon > 0$, there exist $\Phi \in \mathcal{A}_\mathcal{V}$ and a permutation $\pi \in S_N$ such that

$$\max_{i \in [N]} \|\Phi(x_i) - y_{\pi(i)}\| < \varepsilon, \qquad \mathrm{Lip}(\Phi) \leq \Gamma\big(1 \vee \mathsf{L}(\mathcal{D}_N)\big),$$

where $\mathsf{L}(\mathcal{D}_N)$ is defined as

$$\mathsf{L}(\mathcal{D}_N) \coloneqq \min_{\pi \in S_N} \max_{i \neq j} \frac{\|y_{\pi(i)} - y_{\pi(j)}\|}{\|x_i - x_j\|}. \tag{5.3.9}$$

Remark 5.3.7. The quantity $\mathsf{L}(\mathcal{D}_N)$ is the smallest Lipschitz constant of a map defined on the finite set $\{x_1, \ldots, x_N\}$ whose image coincides with the set $\{y_1, \ldots, y_N\}$ (up to permutation). It captures exactly the geometric obstruction that forces large Lipschitz constants: if two source points are very close but must be mapped far apart, then $\mathsf{L}(\mathcal{D}_N)$ is large. Note the close relation with the quantity defined back in (3.4.1), here in the unlabeled setting.

Similarly, we can make the W_p-approximation quantitative in a Lipschitz sense:

Definition 5.3.3. Fix $p \in [1, \infty)$. We say that the system possesses the Γ-*quantitative* W_p-*UAP* if there exists a nondecreasing function $\Gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for every $\mu_\mathrm{B} \in \mathcal{P}_p(\mathbb{R}^d)$, every Lipschitz map $\Psi : \mathbb{R}^d \to \mathbb{R}^d$, and every $\varepsilon > 0$, there exists $\Phi \in \mathcal{A}_\mathcal{V}$ satisfying

$$W_p\big(\Phi_\# \mu_\mathrm{B},\, \Psi_\# \mu_\mathrm{B}\big) < \varepsilon, \qquad \mathrm{Lip}(\Phi) \leq \Gamma(\mathrm{Lip}(\Psi)).$$

As it turns out, these two quantitative properties—one defined on finite point clouds and the other on probability measures—are completely equivalent.

Theorem 5.3.8. *Fix $p \in [1, \infty)$. The system possesses the Γ-quantitative π-UIP if and only if it possesses the Γ-quantitative W_p-UAP. Moreover, one can take the same modulus Γ in both properties up to harmless modifications such as $\Gamma(L) \mapsto \Gamma(1 \vee L)$.*

As an immediate consequence of Theorem 5.3.8, any system possessing the Γ-quantitative π-UIP also possesses the standard W_p-UAP.

5.3.2 Approximation of transport maps

Finally, we study a stronger mechanism to lift the UIP to the W_p-UAP by using the approximation of transport maps in L^p as an intermediate step. Specifically, the following elementary bound links the functional and measure-theoretic frameworks:

Proposition 5.3.9. *Let $p \in [1, \infty)$, $\mu_\mathrm{B} \in \mathcal{P}_p(\mathbb{R}^d)$, and $\mathsf{R}, \mathsf{S} \in L^p(\mu_\mathrm{B}; \mathbb{R}^d)$. Then, we have*

$$W_p\big(\mathsf{R}_\# \mu_\mathrm{B},\, \mathsf{S}_\# \mu_\mathrm{B}\big) \;\leq\; \|\mathsf{R} - \mathsf{S}\|_{L^p(\mu_\mathrm{B})}.$$

Furthermore, for any Lipschitz control family $\mathcal{V}$, if $\mathcal{A}_\mathcal{V}$ is dense in $L^p(\mu_\mathrm{B}; \mathbb{R}^d)$ and $\mu_\mathrm{B} \in \mathcal{P}_{ac}(\mathbb{R}^d)$, then $\mathcal{A}_\mathcal{V}(\mu_\mathrm{B})$ is dense in $\mathcal{P}_p(\mathbb{R}^d)$ with respect to the W_p-distance.

The second statement relies on μ_B being non-atomic, which guarantees that every $\nu \in \mathcal{P}_p(\mathbb{R}^d)$ can be realized as a pushforward $\nu = \mathsf{S}_\# \mu_\mathrm{B}$ for some $\mathsf{S} \in L^p(\mu_\mathrm{B}; \mathbb{R}^d)$.

Establishing a direct connection between the UIP and this underlying L^p-approximation requires stronger structural assumptions in general. However, building on the theory developed in [LLS22;

Che+25c], the one-dimensional case is exceptionally well-behaved: in this setting, the UIP implies that $\mathcal{A}_{\mathcal{V}}$ is dense in $\mathscr{C}^0_{\mathrm{nd,loc}}(\mathbb{R})$, where we recall that the subscript "nd" restricts $\mathscr{C}^0_{\mathrm{loc}}(\mathbb{R})$ to the subset of nondecreasing functions. For completeness, we recall the argument in the following remark.

Remark 5.3.10. As explicitly detailed in [Che+25c, Section 3.3.1], for any $f \in \mathscr{C}^0_{\mathrm{nd}}(\mathbb{R})$ and any $\varepsilon > 0$, one can partition any given compact interval $[a, b]$ with a grid

$$a = z_0 < z_1 < \cdots < z_N = b \qquad \text{such that} \qquad f(z_k) - f(z_{k-1}) < \varepsilon.$$

By the UIP, there exists $\phi \in \mathcal{A}_{\mathcal{V}}$ satisfying $\phi(z_k) = f(z_k)$ for all k. Because both ϕ and f are nondecreasing and match at the endpoints of each subinterval, for any $z \in [z_{k-1}, z_k]$ we have

$$f(z_{k-1}) \le \phi(z) \le f(z_k) \qquad \text{and} \qquad f(z_{k-1}) \le f(z) \le f(z_k).$$

Since $f(z_k) - f(z_{k-1}) < \varepsilon$, this implies $|\phi(z) - f(z)| < \varepsilon$ everywhere, yielding:

$$\|\phi - f\|_{L^\infty(a,b)} < \varepsilon. \tag{5.3.10}$$

For $d \ge 2$, this geometric argument is no longer valid because multi-dimensional flows can twist and rotate space between the interpolated points. However, one way to recover the ability to approximate functions [LLS22; RZ23; TG23; Che+25c] consists of compressing the compact domain into a finite set of localized clusters and then leveraging the UIP to steer these clusters toward their target destinations.

The first step of this strategy essentially requires the system to generate mappings that can shrink the space. This capability is captured by the following definition:

Definition 5.3.4. We say that the control system possesses the *diagonal shrinking property* if $\mathcal{A}_{\mathcal{V}}$ can uniformly approximate any diagonal nondecreasing map on compact sets. That is, for every compact set $K \subset \mathbb{R}^d$ and every $h \in \mathscr{C}^0_{\mathrm{nd}}(\mathbb{R})$,

$$\Big(x \mapsto \big(h(x_1), \ldots, h(x_d)\big)\Big) \;\in\; \overline{\mathcal{A}_{\mathcal{V}}}^{\,L^\infty(K;\mathbb{R}^d)}.$$

With this definition in hand, we can now state the main approximation result:

Proposition 5.3.11. *Let $\mathcal{V}$ be a Lipschitz control family on $\mathbb{R}^d$, and assume the associated system possesses both the UIP and the diagonal shrinking property from Definition 5.3.4. Let moreover $\mu_{\mathrm{B}} \in \mathcal{P}_{ac}(\mathbb{R}^d) \cap \mathcal{P}_p(\mathbb{R}^d)$ be compactly supported. Then:*

(i) *The attainable set $\mathcal{A}_{\mathcal{V}}$ is dense in $L^p(\mu_{\mathrm{B}}; \mathbb{R}^d)$.*

(ii) *The attainable set $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ is dense in $\mathcal{P}_p(\mathbb{R}^d)$ with respect to the W_p-distance.*

Statement (i) is a direct consequence of [Che+25c, Proposition 3.9]—which establishes density in $L^p_{\mathrm{loc}}(\mathbb{R}^d; \mathbb{R}^d)$—applied to the compact support of μ_{B}. Statement (ii) then follows immediately by combining (i) with Proposition 5.3.9.

Remark 5.3.12. For a general (not necessarily compactly supported) base law $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_p(\mathbb{R}^d)$, density in $L^p_{\mathrm{loc}}(\mathbb{R}^d; \mathbb{R}^d)$ does not automatically imply density in $L^p(\mu_{\mathrm{B}}; \mathbb{R}^d)$. To see why,

consider a sequence of approximants Φ_n, a target map S, and a closed ball $B_R \subset \mathbb{R}^d$ of radius R. We can split

$$\|\Phi_n - \mathsf{S}\|^p_{L^p(\mu_{\mathrm{B}})} = \int_{B_R} \|\Phi_n - \mathsf{S}\|^p \,\mathrm{d}\mu_{\mathrm{B}} + \int_{B_R^c} \|\Phi_n - \mathsf{S}\|^p \,\mathrm{d}\mu_{\mathrm{B}}.$$

The first term is controlled by the local approximation, but the second requires tail control. A sufficient condition is that the sequence $\{\|\Phi_n\|^p\}_n$ is uniformly μ_{B}-integrable, meaning

$$\lim_{R\to\infty} \sup_n \int_{B_R^c} \|\Phi_n\|^p \,\mathrm{d}\mu_{\mathrm{B}} = 0.$$

If this uniform integrability holds, we can bound the tail of the error as

$$\int_{B_R^c} \|\Phi_n - \mathsf{S}\|^p \,\mathrm{d}\mu_{\mathrm{B}} \le 2^{p-1} \int_{B_R^c} \|\mathsf{S}\|^p \,\mathrm{d}\mu_{\mathrm{B}} + 2^{p-1} \int_{B_R^c} \|\Phi_n\|^p \,\mathrm{d}\mu_{\mathrm{B}}.$$

Since $\mathsf{S} \in L^p(\mu_{\mathrm{B}}; \mathbb{R}^d)$, the first term vanishes as $R \to \infty$, and by assumption, the second term vanishes uniformly in n. In this case, global $L^p(\mu_{\mathrm{B}}; \mathbb{R}^d)$ approximation holds, and Proposition 5.3.9 yields the W_p-density of $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$.

The assumptions of Proposition 5.3.11 turn out to be automatically satisfied by a broad class of control systems; in particular, by the neural ODE architectures studied in **Chapter 3**:

Corollary 5.3.13. *Let $d \ge 2$, and let $\mathcal{V} \subset \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ be an affine-invariant control family. Assume that $\mathcal{V}$ contains at least one field that is not affine. Then, for every $p \in [1, \infty)$ and every compactly supported $\mu_{\mathrm{B}} \in \mathcal{P}_{ac}(\mathbb{R}^d) \cap \mathcal{P}_p(\mathbb{R}^d)$, the set $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ is dense in $\mathcal{P}_p(\mathbb{R}^d)$ with respect to W_p.*

5.4 Total variation distance

For any pair of finite (nonnegative) measures μ and ν on $\mathbb{R}^d$, the *total variation (TV) distance* is defined by

$$|\mu - \nu|_{\mathsf{TV}} \coloneqq \sup_{A \in \mathcal{B}(\mathbb{R}^d)} |\mu(A) - \nu(A)|. \tag{5.4.1}$$

If μ and ν admit densities ρ_μ and ρ_ν with respect to Lebesgue measure, then

$$|\mu - \nu|_{\mathsf{TV}} = \tfrac{1}{2}\|\rho_\mu - \rho_\nu\|_{L^1(\mathbb{R}^d)} \;+\; \tfrac{1}{2}|\mu(\mathbb{R}^d) - \nu(\mathbb{R}^d)|. \tag{5.4.2}$$

In particular, if μ and ν are probability measures then the second term in (5.4.2) vanishes, yielding

$$|\mu - \nu|_{\mathsf{TV}} = \tfrac{1}{2}\|\rho_\mu - \rho_\nu\|_{L^1(\mathbb{R}^d)} \le 1.$$

There is also an equivalent characterization in terms of couplings, see [LP17, Proposition 4.7]:

$$|\mu - \nu|_{\mathsf{TV}} = \inf_{\pi \in \Pi(\mu,\nu)} \pi\big(\{(x,y) \in \mathbb{R}^d \times \mathbb{R}^d \mid \; x \ne y\}\big), \tag{5.4.3}$$

where the infimum is attained by some $\pi \in \Pi(\mu, \nu)$ such that $\pi(\{x \ne y\}) = |\mu - \nu|_{\mathsf{TV}}$ (recall (5.3.2)).

By Propositions 1.1.1 and 1.2.1, flow maps are homeomorphisms and preserve the Lebesgue decomposition structure. Consequently, any TV-compatible pair must satisfy the following structural constraint:

(TV1) Considering the Lebesgue decompositions

$$\mu_{\mathrm{B}} = \rho_{\mathrm{B}}\mathcal{L}^d + \mu_{\mathrm{B}}^{\mathrm{sc}} + \sum_k \alpha_k \delta_{x_k}, \qquad \mu_\star = \rho_\star \mathcal{L}^d + \mu_\star^{\mathrm{sc}} + \sum_j \beta_j \delta_{y_j},$$

the total masses of the absolutely continuous, singular continuous, and atomic parts must match exactly:

$$\int \rho_{\mathrm{B}} = \int \rho_\star, \qquad \mu_{\mathrm{B}}^{\mathrm{sc}}(\mathbb{R}^d) = \mu_\star^{\mathrm{sc}}(\mathbb{R}^d), \qquad \sum_k \alpha_k = \sum_j \beta_j.$$

Moreover, the multiset of atomic masses must match as well (up to relabeling).

Indeed, if (TV1) fails, there is a positive TV-gap between $\mu_\star$ and every attainable $\nu \in \mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$:

(i) if $\int(\rho_{\mathrm{B}} - \rho_\star) = m \neq 0$, then $|\mu_\star - \nu|_{\mathsf{TV}} \geq |m|/2$ for all ν; and

(ii) if $\mu_{\mathrm{B}}^{\mathrm{sc}} = 0$ but $\mu_\star^{\mathrm{sc}} \neq 0$, then $|\mu_\star - \nu|_{\mathsf{TV}} \geq \mu_\star^{\mathrm{sc}}(\mathbb{R}^d)/2$. A similar bound holds for atomic parts.

A natural restriction is to focus on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, since TV is exceedingly rigid on purely atomic measures. Indeed, it holds

$$|\delta_x - \delta_y|_{\mathsf{TV}} = 1 \qquad \text{for all } x \neq y.$$

Consequently, achieving TV-error < 1 forces the *exact* matching of the target atoms and their respective weights. In fact, on the class of purely atomic probability measures, the TV-UAP reduces essentially to the π-UIP.

For absolutely continuous measures, a natural starting point is to study the relation between the W_p-UAP and the TV-UAP. Even when restricting to $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, neither TV nor W_p controls the other metric in general, as the following example shows.

Example 5.4.1. The following two constructions illustrate the lack of mutual control.

(i) $W_p \to 0$ while TV stays maximal. On $[0,1]$, for an even integer n, define two probability densities that alternate on subintervals of length $1/n$:

$$\rho_n \coloneqq 2\sum_{k=0}^{n/2-1} \mathbf{1}_{[2k/n,(2k+1)/n]}, \qquad \eta_n \coloneqq 2\sum_{k=0}^{n/2-1} \mathbf{1}_{[(2k+1)/n,(2k+2)/n]}.$$

Let $\mu_n \coloneqq \rho_n \mathcal{L}^1$ and $\nu_n \coloneqq \eta_n \mathcal{L}^1$. Then $|\mu_n - \nu_n|_{\mathsf{TV}} = 1$, while the translation $T_n(x) = x + \frac{1}{n}$ transports μ_n onto ν_n, hence $W_p(\mu_n, \nu_n) \leq 1/n \to 0$.

(ii) TV $\to 0$ while W_p stays bounded. Let $\mu = \mathrm{Unif}([0,1]^d)$ and consider $\tau_R(x) = x + Re_1$ for any $R > 0$. Set $\eta_R = (\tau_R)_{\#}\mu$, choose $\varepsilon_n = 1/n$ and $R_n = n^{1/p} + 2$, and define

$$\nu_n = (1 - \varepsilon_n)\mu + \varepsilon_n \eta_{R_n} \;\in\; \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_p(\mathbb{R}^d).$$

Then $\operatorname{supp}\mu$ and $\operatorname{supp}\eta_{R_n}$ are disjoint for all n because $R_n > 1$, so $|\mu - \nu_n|_{\mathsf{TV}} = \varepsilon_n \to 0$. On the other hand, any coupling must move at least ε_n mass across a distance of $R_n - 1$, hence

$$W_p(\mu, \nu_n)^p \geq \varepsilon_n (R_n - 1)^p \geq 1.$$

The following standard lemma, proved in **Subsection 5.6.2**, shows that uniform control over the tails (e.g. assuming uniformly bounded supports) is exactly the missing ingredient needed to bridge this gap.

Lemma 5.4.2. *Fix $p \in [1, \infty)$, and let $\mathcal{F} \subset \mathcal{P}_p(\mathbb{R}^d)$ be a family with uniformly integrable p-th moments, meaning that*

$$\sup_{\mu \in \mathcal{F}} \int_{\|x\| > R} \|x\|^p \, \mathrm{d}\mu(x) \to 0 \qquad \text{as } R \to \infty,$$

Then, on $\mathcal{F}$, the TV*-UAP implies the W_p-UAP.*

The reason why the UIP is not the right notion for studying approximation in TV is that, unlike the Wasserstein case, if μ is non-atomic and μ_N is an empirical measure with N atoms, then $|\mu - \mu_N|_{\mathsf{TV}} = 1$ for all N. Moreover, if Φ is a homeomorphism,

$$|\Phi_{\#}\mu - \Phi_{\#}\nu|_{\mathsf{TV}} = |\mu - \nu|_{\mathsf{TV}}$$

(as proved later in Lemma 5.6.3). In particular, controlling $\mathrm{Lip}(\Phi)$ is irrelevant for TV-stability.

Therefore, a UIP (even with quantitative Lipschitz bounds) is not, by itself, a mechanism to obtain TV-UAP. To approximate in TV one must control how mass is distributed across regions, which motivates the mixing and shear constructions developed below in **Subsection 5.4.2**.

5.4.1 One-dimensional case: Monotone transport

The case $d = 1$ is exceptional because order preservation trivializes the geometry of transport. If $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R})$, there is a canonical transport map between them, the *monotone rearrangement*,

$$\mathsf{M}_{\mathrm{B}\to\star} := F_{\mu_\star}^{-1} \circ F_{\mu_{\mathrm{B}}}, \qquad F_\mu(x) := \mu((-\infty, x]). \tag{5.4.4}$$

Since one-dimensional flows are orientation-preserving (strictly increasing) homeomorphisms (see Remark 1.1.2), TV-universality in $d = 1$ reduces to a *functional* approximation problem: can we approximate monotone maps such as $\mathsf{M}_{\mathrm{B}\to\star}$ in a topology strong enough to control the pushed-forward densities?

Proposition 5.4.3. *Let $\mathcal{V}$ be a Lipschitz control family on $\mathbb{R}$ for which $\mathcal{A}_\mathcal{V}$ is dense in $W^{1,1}_{\mathrm{nd,loc}}(\mathbb{R})$, meaning that*

$$h \in \overline{\mathcal{A}_\mathcal{V}}^{W^{1,1}(I)} \qquad \text{for every nondecreasing } h \in W^{1,1}_{\mathrm{loc}}(\mathbb{R}) \text{ and compact interval } I \subset \mathbb{R}. \tag{5.4.5}$$

Then the control system generated by $\mathcal{V}$ possesses the TV*-UAP on $\mathcal{P}_{ac}(\mathbb{R})$.*

Remark 5.4.4. Condition (5.4.5) might appear weaker than the standard definition of sequential density in $W^{1,1}_{\mathrm{loc}}(\mathbb{R})$, which requires a single sequence $\{h_n\}_n \subset \mathcal{A}_\mathcal{V}$ converging to h on every compact interval simultaneously.

However, the two notions are equivalent. Indeed, one can construct the global approximating sequence by a diagonal argument: for each $m \in \mathbb{N}$, choose $g_m \in \mathcal{A}_\mathcal{V}$ such that

$$\|g_m - h\|_{W^{1,1}([-m,m])} < \frac{1}{m}.$$

Then, for any compact $I \subset \mathbb{R}$, choosing M large enough such that $I \subset [-M, M]$ guarantees convergence on I. Thus, the sequence $\{g_m\}_m$ converges to h in the topology of $W^{1,1}_{\mathrm{loc}}(\mathbb{R})$.

The proof of Proposition 5.4.3 proceeds by truncation and mollification: one first replaces $\mu_{\mathrm{B}}, \mu_\star$ by smooth, strictly positive approximations on a large compact interval, so that (5.4.4) becomes an increasing diffeomorphism, and then uses (5.4.5) to approximate this rearrangement in $W^{1,1}$, which controls the L^1-convergence of the pushed-forward densities.

We now analyze when Proposition 5.4.3 holds in the constrained setting of affine invariance.

Corollary 5.4.5. *Let $\mathcal{V} \subset \mathscr{C}^{0,1}(\mathbb{R})$ be an affine-invariant control family. If there exists $\sigma \in \mathcal{V}$ such that (i) $\sigma \in \mathscr{C}^1(\mathbb{R})$ and it is not an affine map, or (ii) $\sigma(s) = s_+$, then the corresponding control system possesses the* TV*-UAP on $\mathcal{P}_{ac}(\mathbb{R})$.*

Remark 5.4.6. Case *(ii)* is consistent with the L^1 (hence TV) approximate controllability results proved in [RZ24] in the context of ($\mathcal{V}_1^+$), here reinterpreted through the monotone-transport criterion of Proposition 5.4.3.

5.4.2 Multi-dimensional case: coordinate mixing

A natural multi-dimensional analogue of (5.4.5) would be to strengthen the diagonal shrinking property in Definition 5.3.4 from uniform approximation on compact sets to approximation in the stronger $W^{1,1}$ topology. While the capacity to approximate diagonal nondecreasing maps was the key mechanism to unlock the W_p-UAP (Proposition 5.3.11), it is insufficient to achieve the TV-UAP for $d \geq 2$.

The geometric reason is that diagonal transformations act independently on each axis and therefore cannot create statistical dependencies between variables. This obstruction can be cleanly formalized by considering entirely decoupled control families and the set of product measures,

$$\mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d) \coloneqq \{\otimes_{i=1}^d \mu_i \mid \mu_i \in \mathcal{P}(\mathbb{R})\}, \tag{5.4.6}$$

where each product measure is uniquely defined on the Borel σ-algebra by its values on rectangles $B = \prod_{i=1}^d B_i$ with $B_i \in \mathcal{B}(\mathbb{R})$, namely, $\otimes_{i=1}^d \mu_i(B) = \prod_{i=1}^d \mu_i(B_i)$.

Proposition 5.4.7. *Let $d \geq 2$ and let $\mathcal{V}$ be a Lipschitz control family that is* decoupled*, i.e., every $v \in \mathcal{V}$ satisfies*

$$v(x) = \big(v_1(x^{(1)}), \ldots, v_d(x^{(d)})\big). \tag{5.4.7}$$

Then, the control system generated by $\mathcal{V}$ does not possess the TV*-UAP on $\mathcal{P}(\mathbb{R}^d)$.*

The obstruction arises from two key facts demonstrated in the proof (see **Subsection 5.6.2**):

- **(i)** If $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$, then the reachable set $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ is contained in $\mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$;
- **(ii)** The set $\mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$ is closed in $(\mathcal{P}(\mathbb{R}^d), \mathsf{TV})$.

Together, these properties show that one cannot approximate a target measure with dependent coordinates starting from a product base measure.

Thus, TV-approximation is not merely about approximating one-dimensional transports; it requires a mechanism that mixes coordinates. In our setting, such a mechanism is provided by the ability to generate shear-type maps, as encoded in assumption (A2) below.

Theorem 5.4.8. *Let $d \geq 2$ and $\mathcal{A}_{\mathcal{V}}$ be the attainable set generated by a Lipschitz control family $\mathcal{V}$. Assume the following conditions:*

(A0) Reversibility: *If $\Phi \in \mathcal{A}_{\mathcal{V}}$, then $\Phi^{-1} \in \mathcal{A}_{\mathcal{V}}$.*

(A1) Marginal flows: *For every compact interval $I \subset \mathbb{R}$, every $\tau \in W^{1,1}_{\mathrm{nd,loc}}(\mathbb{R})$, every coordinate $j \in [d]$ and all $\varepsilon > 0$, there exist a function $\phi : \mathbb{R} \to \mathbb{R}$ and a map $\Phi \in \mathcal{A}_{\mathcal{V}}$ of the form*

$$\Phi = (x^{(1)}, \dots, x^{(j-1)}, \phi(x^{(j)}), x^{(j+1)}, \dots, x^{(d)}) \qquad \text{with } \|\phi - \tau\|_{W^{1,1}(I)} < \varepsilon.$$

(A2) Shear flows: *For every compact interval $J \subset \mathbb{R}$, every nondecreasing, piecewise affine $\sigma \in \mathscr{C}^0(\mathbb{R})$, any coordinates $i \neq j$ and all $\varepsilon > 0$, there exist a function $\psi : \mathbb{R} \to \mathbb{R}$ and a map $\Phi \in \mathcal{A}_{\mathcal{V}}$ of the form*

$$\Phi = (x^{(1)}, \dots, x^{(j-1)}, x^{(j)} + \psi(x^{(i)}), x^{(j+1)}, \dots, x^{(d)}) \quad \text{with } \|\psi - \sigma\|_{L^1(J)} < \varepsilon.$$

Then the system possesses the TV-*UAP on $\mathcal{P}_{ac}(\mathbb{R}^d)$.*

As in Corollary 5.4.5, we state a sufficient condition for neural ODEs to satisfy (A1)–(A2).

Corollary 5.4.9. *Let $d \geq 2$ and let $\sigma : \mathbb{R} \to \mathbb{R}$. Assume that $\mathcal{V} \subset \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ is an affine-invariant control family, closed under finite linear combinations, and such that*

$$x \longmapsto \sigma(ax^{(i)} + b)\, e_j \in \mathcal{V}, \qquad \text{for all } a, b \in \mathbb{R},\ i, j \in [d].$$

If σ is either $\mathscr{C}^1(\mathbb{R})$ and non-polynomial, or $\sigma(s) = s_+$, then the control system generated by $\mathcal{V}$ possesses the TV-*UAP on $\mathcal{P}_{ac}(\mathbb{R}^d)$.*

We conclude this section by discussing an application of Theorem 5.4.8.

Remark 5.4.10. The results above extend naturally to any discrepancy controlled by TV. Consider, for instance, the *Jensen-Shannon divergence* $\mathsf{JS}(\cdot,\cdot)$, defined for $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ by

$$\mathsf{JS}(\mu, \nu) := \frac{1}{2}\mathsf{KL}(\mu \| \lambda) + \frac{1}{2}\mathsf{KL}(\nu \| \lambda), \qquad \text{with } \lambda := \frac{\mu + \nu}{2},$$

where $\mathsf{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence (recall (II.2) in **Part II (Overview)**).

The Jensen-Shannon divergence is of particular relevance in generative modeling. For example, the original Generative Adversarial Network (GAN) objective is classically interpreted as minimizing the JS divergence between the data and the generated distributions. Relying on the well-known inequality

$$\mathsf{JS}(\mu, \nu) \leq (\log 2)\, |\mu - \nu|_{\mathsf{TV}}, \tag{5.4.8}$$

it immediately follows that under the assumptions of Theorem 5.4.8 (or Corollary 5.4.9), the system also possesses the JS-UAP.

Structure of the constructed flows

The proof of Theorem 5.4.8 adapts the strategy of [RZ24, Theorem 1]—compare also with [GR26]—, developed for $(\mathcal{V}_1^+)$, to the general controlled flow setting. The geometry of the construction can be understood as a discrete approximation of the Knothe–Rosenblatt (KR) rearrangement [Kno57], which is the canonical *triangular map* obtained by sequentially matching marginals and conditional distributions. Formally:

Definition 5.4.1. Let $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. The *Knothe–Rosenblatt rearrangement* from μ_{B} to $\mu_\star$ is the (a.e. unique) Borel map $\mathsf{KR} : \mathbb{R}^d \to \mathbb{R}^d$ such that:

(i) $\mathsf{KR}_{\#}\mu_{\mathrm{B}} = \mu_\star$;

(ii) KR is *triangular*, meaning that its k-th component depends only on the first k coordinates:

$$\mathsf{KR}(x) = \big(\mathsf{KR}_1(x^{(1)}),\ \mathsf{KR}_2(x^{(1)}, x^{(2)}),\ \ldots,\ \mathsf{KR}_d(x^{(1)}, \ldots, x^{(d)})\big); \tag{5.4.9}$$

(iii) for each $k \in [d]$, and for μ_{B}-a.e. $(x^{(1)}, \ldots, x^{(k-1)})$, the map

$$t \longmapsto \mathsf{KR}_k(x^{(1)}, \ldots, x^{(k-1)}, t)$$

is nondecreasing.

Remark 5.4.11. In practice, the map KR is constructed recursively through a sequence of one-dimensional monotone rearrangements. The first component KR_1 simply pushes the first marginal of μ_{B} onto the first marginal of $\mu_\star$. For any $k \geq 2$, assume the first $k-1$ components have already been evaluated, and denote the mapped variables by

$$\big(y^{(1)}, \ldots, y^{(k-1)}\big) = \big(\mathsf{KR}_1(x^{(1)}), \ldots, \mathsf{KR}_{k-1}(x^{(1)}, \ldots, x^{(k-1)})\big).$$

Then, the slice $t \mapsto \mathsf{KR}_k(x^{(1)}, \ldots, x^{(k-1)}, t)$ is uniquely determined as the monotone map that pushes the conditional law of $x^{(k)}$ given $(x^{(1)}, \ldots, x^{(k-1)})$ under μ_{B} onto the conditional law of $y^{(k)}$ given $(y^{(1)}, \ldots, y^{(k-1)})$ under $\mu_\star$.

We illustrate the construction of Theorem 5.4.8 for the case $d = 2$. The construction approximates the target triangular structure by composing four elementary steps, as depicted in Figure 5.1:

(S1) Marginal matching: First, we match the marginal distribution of $x^{(2)}$ using a coordinate-wise 1D map (via assumption (A1)). This fixes the last coordinate of the triangular map.

(S2) Discretization: Since we cannot realize a continuum of conditional maps simultaneously, we discretize the dependence on $x^{(2)}$ by dividing the domain into finitely many horizontal strips. Within each strip, the target density is approximated as constant with respect to $x^{(2)}$.

(S3) Shearing: We apply a shear transformation $(x^{(1)}, x^{(2)}) \mapsto (x^{(1)} + \sigma(x^{(2)}), x^{(2)})$ utilizing assumption (A2). This is the key geometric device: it spatially separates the horizontal strips so that their projections onto the $x^{(1)}$-axis become mutually disjoint.

(S4) Decoupled transport: Finally, a single global 1D map acts on the $x^{(1)}$-coordinate, again relying on assumption (A1). Because the strips are now spatially separated, this single global map can act independently on each strip, effectively implementing a collection of approximate conditional rearrangements simultaneously.

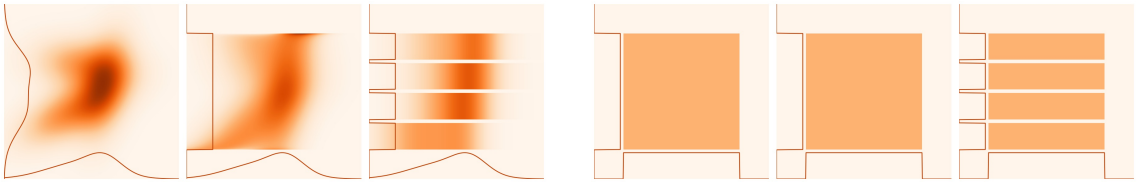
If $d \geq 3$, the proof iterates this logic via backward induction over the coordinates $(x^{(d)}, x^{(d-1)}, \ldots)$.

Remark 5.4.12. We use the KR viewpoint for geometric intuition. The proof of Theorem 5.4.8 does not claim that the constructed maps converge to the KR map in strong functional norms.

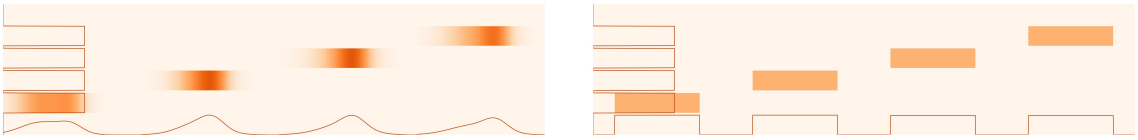
5.4.3 Relationship with point interpolation

Neither a system possessing the UIP (as per Definition 3.2.1) nor possessing the TV-UAP implies the other in general. The following two propositions provide complementary examples.

Proposition 5.4.13. *Assume $d \geq 2$ and consider the same control family $\mathcal{V}$ as in (5.3.8). Then the associated system possesses the UIP, but it does not possess the* TV-*UAP.*



(a) Marginal matching and discretization ((S1)–(S2)). The initial measure (**left**) and the target measure (**right**, chosen as $\mathrm{Unif}([0,1]^2)$) are divided into finitely many horizontal strips. Corresponding strips span the same vertical intervals and contain the same amount of mass.



(b) Shearing and conditional matching ((S3)–(S4)). A shear transformation (**left**) horizontally translates the strips so their projections onto the $x^{(1)}$-axis become mutually disjoint. Finally, a single global 1D map (**right**) can independently rearrange the mass within each separated strip.

Figure 5.1: Schematic illustration of the proof of Theorem 5.4.8 in 2D.

Proposition 5.4.14. *Let $d \geq 2$ and let $\mathcal{V}$ be the family of all $v \in \mathscr{C}^{0,1}(\mathbb{R}^d;\mathbb{R}^d)$ satisfying $v(0) = 0$. Then the associated system possesses the* TV*-UAP on $\mathcal{P}_{ac}(\mathbb{R}^d)$, but it does not possess the UIP.*

The two counterexamples rely on different mechanisms:

- In Proposition 5.3.5 and Proposition 5.4.13, we use the same control family (5.3.8), which possesses the UIP by Remark 3.3.2 and Theorem 3.3.1. On the other hand, divergence-free fields preserve the L^q norms of densities (by (1.2.4)), so the interpolation estimate between L^1 and L^r $(r > 2)$ used in the proof of Proposition 5.4.13 yields a uniform lower bound on the total variation distance whenever $\|\rho_{\mathrm{B}}\|_{L^2} \neq \|\rho_\star\|_{L^2}$.
- In Proposition 5.4.14, the obstruction is of a different nature: all vector fields vanish at the origin, so every attainable flow fixes 0, which rules out the UIP. On the other hand, this pointwise constraint is invisible at the level of absolutely continuous measures, and a local Dacorogna–Moser type argument [DM90] still yields the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. More precisely, one couples smooth approximations of ρ_{B} and $\rho_\star$ on a large annulus, arranged to coincide with the identity outside. The resulting flow fixes the origin (and, more generally, any prescribed closed discrete set) while still allowing TV-approximation by choosing the annulus sufficiently large.

Remark 5.4.15. In Proposition 5.4.14, the class $\{v \in \mathscr{C}^{0,1}(\mathbb{R}^d;\mathbb{R}^d) \mid v(0) = 0\}$ could likely be replaced by a much smaller parameterized family, for instance a neural family of the form

$$\mathcal{V}_{\sigma,0} = \Big\{ x \mapsto \sum_{j=1}^{m} w_j\big(\sigma(a_j \cdot x + b_j) - \sigma(b_j)\big) \;\Big|\; m \in \mathbb{N},\ w_j, a_j \in \mathbb{R}^d,\ b_j \in \mathbb{R} \Big\}.$$

for a fixed smooth non-polynomial Lipschitz activation σ. Indeed, every $v \in \mathcal{V}_{\sigma,0}$ satisfies $v(0) = 0$, so the associated flows still fix the origin and therefore fail the UIP. To turn this into a rigorous analogue of Proposition 5.4.14, it would suffice to prove a density statement in $W^{1,\infty}$ on compact sets for compactly supported vector fields vanishing near the origin.

5.5 Perspectives

The results above suggest some natural directions.

- **Necessity of mixing.** In TV, our assumptions (A0)–(A2) are sufficient, but their necessity is open. Is some form of shear-type mixing unavoidable for TV-universality in $d \geq 2$? Can one replace (A2) with a weaker, more intrinsic condition (e.g. a condition on the Lie algebra generated by $\mathcal{V}$) that still guarantees density of pushforward measures?
- **Quantitative rates.** The present results are qualitative (existence of controls and times). For applications in generative modeling, one would like quantitative bounds: given $\varepsilon > 0$, how large must the time horizon be and how do the bounds depend on dimension and on regularity of the target density? A related question is to translate such bounds into architectural complexity for neural parameterizations (width/depth, size of parameters).
- **Connections with optimal transport.** Our TV construction is inspired by a discretized triangular (Knothe–Rosenblatt) strategy. A broader question is to understand under which structural assumptions one can approximate canonical transports (KR rearrangement, Brenier map) in strong functional topologies by attainable flow maps under structural assumptions on $\mathcal{V}$. Recent progress in the setting of ($\mathcal{V}_1^+$) was obtained in [GR26].

5.6 Proofs

5.6.1 Proofs of Section 5.3: Wasserstein distance

Proof of Proposition 5.3.1. By W_p-compatibility, there exist attainable maps $\Psi_k \in \mathcal{A}_{\mathscr{C}^{0,1}}$ such that the pushforward measures $\nu_k := (\Psi_k)_{\#}\mu_{\mathrm{B}}$ satisfy

$$W_p(\nu_k, \mu_\star) \to 0 \qquad \text{as } k \to \infty.$$

Proof of (W1). For some $k \geq 1$, we have $W_p(\nu_k, \mu_\star) < \infty$, which guarantees that $\nu_k, \mu_\star \in \mathcal{P}_p(\mathbb{R}^d)$. Furthermore, by (1.1.1) we have $\|\Psi_k(x)\| \leq L\|x\| + \|\Psi_k(0)\|$ for some constant $L > 0$. Raising this inequality to the p-th power and integrating shows that Ψ_k (and analogously, its inverse) preserves the finiteness of p-th moments. Consequently,

$$\mu_{\mathrm{B}} \in \mathcal{P}_p(\mathbb{R}^d) \quad \Longleftrightarrow \quad \nu_k \in \mathcal{P}_p(\mathbb{R}^d),$$

which yields the result.

Proof of (W2). Assume $\mu_{\mathrm{B}} = \sum_{i=1}^N m_i \delta_{x_i}$ for some $m_i > 0$ with $\sum_{i=1}^N m_i = 1$. By Proposition 1.2.1, the attainable measures are of the form

$$\nu_k = \sum_{i=1}^N m_i \delta_{z_i^k} \qquad \text{where } z_i^k := \Psi_k(x_i).$$

Because $\nu_k \to \mu_\star$ in W_p, the p-th moments of the sequence $\{\nu_k\}_{k\geq 1}$ must be uniformly bounded:

$$\sup_k \int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}\nu_k(x) = \sup_k \sum_{j=1}^N m_j \|z_j^k\|^p < \infty.$$

Each term of the sum satisfies $m_i\|z_i^k\|^p \leq \sum_{j=1}^N m_j\|z_j^k\|^p$ with $m_i > 0$, so that the sequence $\{z_i^k\}_{k\geq 1}$ is uniformly bounded in $\mathbb{R}^d$. Passing to a subsequence if necessary, we may assume that for all i

$$z_i^k \to z_i \in \mathbb{R}^d \qquad \text{as } k \to \infty.$$

This implies $\nu_k \rightharpoonup \sum_{i=1}^N m_i \delta_{z_i}$. Since convergence in W_p implies weak convergence, the limit must be unique. We conclude that $\mu_\star = \sum_{i=1}^N m_i \delta_{z_i}$. If any limit points z_i coincide, their respective masses simply merge. In any case, the target measure $\mu_\star$ is supported on at most N points. □

To prove Proposition 5.3.3, we first establish a geometric observation: we can construct a finite configuration of points whose one-dimensional projections cannot be simultaneously order-preserved across all directions.

Lemma 5.6.1. *There exists an integer $N \geqslant 1$ large enough and points $x_1, \dots, x_N, y_1, \dots, y_N \in \mathbb{R}^d$ (with $d \geq 1$) such that, for every direction $u \in \mathbb{S}^{d-1}$, there exist i and j, with $1 \leq i < j \leq N$, satisfying*

$$\langle x_i, u\rangle < \langle x_j, u\rangle \qquad \text{and} \qquad \langle y_i, u\rangle > \langle y_j, u\rangle.$$

Proof of Lemma 5.6.1. Fix any unit vector $w \in \mathbb{S}^{d-1}$. For any $\rho > 0$, define the provisional points

$$x_1^{(w,\rho)} = -\rho w, \qquad x_2^{(w,\rho)} = \rho w, \qquad y_1^{(w,\rho)} = \rho w, \qquad y_2^{(w,\rho)} = -\rho w.$$

Observe that

$$\langle x_1^{(w,\rho)}, w\rangle = -\rho < \rho = \langle x_2^{(w,\rho)}, w\rangle \qquad \text{and} \qquad \langle y_1^{(w,\rho)}, w\rangle = \rho > -\rho = \langle y_2^{(w,\rho)}, w\rangle.$$

By the continuity of the inner product, there exists an open neighborhood $U_w \subset \mathbb{S}^{d-1}$ of w such that these strict inequalities hold for all $u \in U_w$. Since $\mathbb{S}^{d-1}$ is compact, we can extract an integer $M \geqslant 1$ and a finite subcovering $\{U_{w_k}\}_{k=1}^M$ from the open covering $\{U_w \mid w \in \mathbb{S}^{d-1}\}$.

Choose pairwise distinct positive numbers $\rho_1, \dots, \rho_M$, set $N \coloneqq 2M$, and define, for each $k \in [M]$,

$$x_{2k-1} \coloneqq -\rho_k w_k, \qquad x_{2k} \coloneqq \rho_k w_k, \qquad y_{2k-1} \coloneqq \rho_k w_k, \qquad y_{2k} \coloneqq -\rho_k w_k.$$

By construction, the points $x_1, \dots, x_N$ are pairwise distinct, and so are $y_1, \dots, y_N$.

For any $u \in \mathbb{S}^{d-1}$, there exists an index k such that $u \in U_{w_k}$. By construction, this implies

$$\langle x_{2k-1}, u\rangle < \langle x_{2k}, u\rangle \qquad \text{and} \qquad \langle y_{2k-1}, u\rangle > \langle y_{2k}, u\rangle.$$

Choosing $i = 2k-1$ and $j = 2k$ satisfies $1 \leq i < j \leq N$ and proves the claim. □

With this geometric obstruction in hand, the failure of the UIP follows naturally.

Proof of Proposition 5.3.3. We prove each of the two statements separately:

(No UIP). Fix any flow map $\Phi \in \mathcal{A}_\mathcal{V}$ generated by a field associated with a direction $u \in \mathbb{S}^{d-1}$. For any trajectory $x(\cdot)$ solving $\dot{x}(t) = v_t(x(t))$, its scalar projection $s(t) \coloneqq \langle x(t), u\rangle$ satisfies the one-dimensional ODE $\dot{s}(t) = g_t(s(t))$. Since g is Lipschitz, uniqueness holds for this scalar equation, meaning its time-1 flow on $\mathbb{R}$ is strictly increasing. Consequently, the flow map Φ preserves the strict order of projections along the direction u:

$$\langle x, u\rangle < \langle x', u\rangle \quad \Longrightarrow \quad \langle \Phi(x), u\rangle < \langle \Phi(x'), u\rangle \qquad \text{for all } x, x' \in \mathbb{R}^d.$$

By Lemma 5.6.1, there exists an $\{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$ such that for every $u \in \mathbb{S}^{d-1}$, there are indices $i < j$ for which $\langle x_i, u\rangle < \langle x_j, u\rangle$ but $\langle y_i, u\rangle > \langle y_j, u\rangle$. Due to the monotonicity established above, no map $\Phi \in \mathcal{A}_\mathcal{V}$ can satisfy $\Phi(x_i) = y_i$ for all $i \in [N]$. Therefore, the system does not possess the UIP.

(π-UIP). Let $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ be two sets of N distinct points in $\mathbb{R}^d$. We can choose a direction $u \in \mathbb{S}^{d-1}$ such that all projections $s_i \coloneqq \langle x_i, u\rangle$ and $r_j \coloneqq \langle y_j, u\rangle$ are pairwise distinct; this is always possible because the set of forbidden directions forms a finite union of proper subspheres in $\mathbb{S}^{d-1}$. Up to a permutation of the target points, we may assume that both sequences of projections share the same strict ordering:

$$s_1 < \cdots < s_N \qquad \text{and} \qquad r_1 < \cdots < r_N.$$

To simplify notation, we assume this relabeling has already been applied, so we write y_i instead of $y_{\pi(i)}$. We decompose $x_i = s_i u + z_i$ and $y_i = r_i u + w_i$, where z_i and w_i are orthogonal to u.

Step 1: Longitudinal transport. Since $s_i < s_{i+1}$ and $r_i < r_{i+1}$ for all i, there exists a strictly increasing Lipschitz homeomorphism $\phi : \mathbb{R} \to \mathbb{R}$ such that $\phi(s_i) = r_i$. We interpolate this map in time: choose $\eta \in \mathscr{C}^1([0,1])$ such that $\eta(0) = 0$ and $\eta(t) = 1$ for $t \in [1/2, 1]$. The isotopy $\phi_t \coloneqq (1-\eta(t))\,\mathrm{Id} + \eta(t)\phi$ is generated by the vector field

$$g_t(s) \coloneqq \partial_t \phi_t\big(\phi_t^{-1}(s)\big)$$

. We then define

$$v_t^{(1)}(x) \coloneqq g_t\big(\langle x, u\rangle\big)u.$$

By construction, $v^{(1)} \in \mathcal{V}$, its time-1 flow maps each x_i to an intermediate point $x_i' \coloneqq r_i u + z_i$, and $v_t^{(1)}(\cdot) \equiv 0$ for all $t \in [1/2, 1]$.

Step 2: Transverse transport. Next, we map each intermediate point x_i' to its final target y_i. Notice that both points lie in the affine hyperplane $\mathsf{H}_i \coloneqq \{x \in \mathbb{R}^d \mid \langle x, u\rangle = r_i\}$. Because the values r_i are strictly increasing, these N hyperplanes are separated by a positive distance. Thus, we can choose bounded, pairwise disjoint open neighborhoods $U_i \subset \mathbb{R}^d$ containing both x_i' and y_i.

For each $i \in [N]$, choose a compactly supported $W_i \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ with $\mathrm{supp}(W_i) \subset U_i$ whose time-1 flow maps x_i' to y_i. Since both points lie in H_i, this transport can be performed entirely inside H_i, so that $\langle W_i(x), u\rangle = 0$ for all $x \in \mathbb{R}^d$. Because the supports U_i are disjoint, we have

$$W(x) \coloneqq \sum_{i=1}^N W_i(x) \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d) \qquad \text{satisfying } \langle W(x), u\rangle = 0 \qquad \text{for all } x \in \mathbb{R}^d.$$

Step 3: Concatenation. Finally, to activate the transverse transport only after the longitudinal one has finished, choose a continuous cut-off function $\beta : [0,1] \to \mathbb{R}_{\geq 0}$ supported in $[1/2, 1]$ with $\int_0^1 \beta(t)\,\mathrm{d}t = 1$. Define

$$v_t^{(2)}(x) \coloneqq \beta(t)W(x), \qquad v_t(x) \coloneqq v_t^{(1)}(x) + v_t^{(2)}(x).$$

Since $\langle W(x), u\rangle = 0$, the field v has the form $v_t(x) = g_t\big(\langle x, u\rangle\big)u + w_t(x)$ with $w_t(x) = \beta(t)W(x)$ orthogonal to u, and therefore $v \in \mathcal{V}$. Moreover, the disjoint time supports of $v^{(1)}$ and $v^{(2)}$ guarantee that the time-1 flow of v first maps x_i to x_i' and then maps x_i' to y_i. Hence, there exists a flow map $\Phi \in \mathcal{A}_{\mathcal{V}}$ such that $\Phi(x_i) = y_i$ for all $i \in [N]$, and the system possesses the π-UIP. □

Proofs of Subsection 5.3.1: Empirical approximation

We begin by proving the lemma that ensures empirical approximation under weak topologies.

Lemma 5.6.2. *Let $\mathsf{D} : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ be a distance that metrizes weak convergence on the class of measures under consideration (i.e., $\mu_n \rightharpoonup \mu$ if and only if $\mathsf{D}(\mu_n, \mu) \to 0$). Then for every $\mu \in \mathcal{P}(\mathbb{R}^d)$ and empirical measures $\mu_N := \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ generated by independent samples $X_i \sim \mu$, we have $\mathsf{D}(\mu_N, \mu) \to 0$ almost surely as $N \to \infty$. In particular, for every $\Phi \in \mathscr{C}^0(\mathbb{R}^d; \mathbb{R}^d)$,*

$$\mathsf{D}(\Phi_{\#}\mu_N, \Phi_{\#}\mu) \to 0 \qquad \text{almost surely as } N \to \infty.$$

Proof of Lemma 5.6.2. Since D metrizes weak convergence on the relevant class, it suffices to prove that $\mu_N \rightharpoonup \mu$ almost surely. This follows from Varadarajan's theorem; see [Dud02, Theorem 11.4.1].

For the second statement, fix $\Phi \in \mathscr{C}^0(\mathbb{R}^d; \mathbb{R}^d)$. Since $\varphi \circ \Phi \in \mathscr{C}_b(\mathbb{R}^d)$ for any test function $\varphi \in \mathscr{C}_b(\mathbb{R}^d)$, the almost sure weak convergence $\mu_N \rightharpoonup \mu$ implies that, with probability one,

$$\int_{\mathbb{R}^d} \varphi(z)\,\mathrm{d}(\Phi_{\#}\mu_N)(z) = \int_{\mathbb{R}^d} \varphi\big(\Phi(x)\big)\,\mathrm{d}\mu_N(x) \longrightarrow \int_{\mathbb{R}^d} \varphi\big(\Phi(x)\big)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \varphi(z)\,\mathrm{d}(\Phi_{\#}\mu)(z).$$

Because D metrizes weak convergence, this immediately yields $\mathsf{D}(\Phi_{\#}\mu_N, \Phi_{\#}\mu) \to 0$ almost surely, concluding the proof. □

Proof of Theorem 5.3.4. We prove the two implications separately.

($\Leftarrow$).

Assume that the system possesses the W_p-UAP on $\mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d)$. Fix $\varepsilon > 0$ and sets of pairwise distinct points $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ in $\mathbb{R}^d$. The empirical measures

$$\mu_{\mathrm{B}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i} \qquad \text{and} \qquad \mu_\star^N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}$$

are W_p-compatible since the UIP for $\mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d)$ guarantees a flow mapping $x_i \mapsto y_i$ (in $d = 1$, after reordering the target points increasingly). Applying the W_p-UAP with tolerance $\varepsilon/N^{1/p}$ yields $\Phi \in \mathcal{A}_{\mathcal{V}}$ such that $W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N) < \varepsilon/N^{1/p}$.

Since Φ is injective, the points $\{\Phi(x_i)\}_{i=1}^N$ are pairwise distinct. By (5.3.5) and the norm inequality $\|z\|_\infty \leq \|z\|$, there exists a permutation $\pi \in S_N$ such that:

$$\max_{i\in[N]} \|\Phi(x_i) - y_{\pi(i)}\|_\infty \leq \left(\sum_{i=1}^N \|\Phi(x_i) - y_{\pi(i)}\|^p\right)^{1/p} = N^{1/p} W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N) < \varepsilon.$$

This proves the approximate π-UIP.

($\Rightarrow$) Conversely, assume the approximate π-UIP. Let

$$\mu_{\mathrm{B}} = \frac{1}{N}\sum_{i=1}^N \delta_{x_i} \in \mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d),$$

let $\mu_\star$ be a W_p-compatible target, and fix $\varepsilon > 0$. By Definition 5.2.2, there exists a sequence of flow maps $\Psi_k \in \mathcal{A}_{\mathscr{C}^{0,1}}$ such that $W_p((\Psi_k)_{\#}\mu_{\mathrm{B}}, \mu_\star) \to 0$. Choose k large enough so that

$$W_p((\Psi_k)_{\#}\mu_{\mathrm{B}}, \mu_\star) < \frac{\varepsilon}{2}.$$

Since Ψ_k is injective, the points $y_i \coloneqq \Psi_k(x_i)$ are pairwise distinct, and $(\Psi_k)_{\#}\mu_{\mathrm{B}} = \frac{1}{N}\sum_{i=1}^{N}\delta_{y_i}$. Applying the approximate π-UIP to $\{x_i\}_{i=1}^N$ and $\{y_i\}_{i=1}^N$ with tolerance $\varepsilon/2$, we find $\Phi \in \mathcal{A}_{\mathcal{V}}$ and $\pi \in S_N$ such that

$$\max_{i\in[N]} \|\Phi(x_i) - y_{\pi(i)}\| < \frac{\varepsilon}{2}.$$

The permutation π defines a transport plan between $\Phi_{\#}\mu_{\mathrm{B}}$ and $(\Psi_k)_{\#}\mu_{\mathrm{B}}$, hence

$$W_p^p(\Phi_{\#}\mu_{\mathrm{B}}, (\Psi_k)_{\#}\mu_{\mathrm{B}}) \le \frac{1}{N}\sum_{i=1}^{N}\|\Phi(x_i) - y_{\pi(i)}\|^p < \left(\frac{\varepsilon}{2}\right)^p.$$

Therefore,

$$W_p(\Phi_{\#}\mu_{\mathrm{B}}, (\Psi_k)_{\#}\mu_{\mathrm{B}}) < \frac{\varepsilon}{2}.$$

Finally, the triangle inequality yields

$$W_p(\Phi_{\#}\mu_{\mathrm{B}}, \mu_\star) \le W_p(\Phi_{\#}\mu_{\mathrm{B}}, (\Psi_k)_{\#}\mu_{\mathrm{B}}) + W_p((\Psi_k)_{\#}\mu_{\mathrm{B}}, \mu_\star) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,$$

which proves the W_p-UAP on $\mathcal{P}^{\neq}_{\mathrm{emp}}(\mathbb{R}^d)$. □

Proof of Proposition 5.3.5. We prove each of the two statements separately.

(UIP). For every $\theta = (w, a, b)$ with $w \cdot a = 0$, the vector field $v_\theta(x) = w\,(a \cdot x + b)_+$ satisfies

$$\operatorname{div} v_\theta(x) = (w \cdot a)\,\mathbf{1}_{\{a\cdot x + b > 0\}}(x) = 0 \qquad \text{for a.e. } x \in \mathbb{R}^d.$$

The associated system possesses the UIP by Remark 3.3.2 and Theorem 3.3.1.

(No W_p-UAP). Let $Q \coloneqq [0,1]^d$ and $H \coloneqq [0, 0.5] \times [0,1]^{d-1}$. Define $\mu_{\mathrm{B}} \coloneqq \mathbf{1}_Q\,\mathcal{L}^d$ and $\mu_\star \coloneqq 2\,\mathbf{1}_H\,\mathcal{L}^d$. Then $(\mu_{\mathrm{B}}, \mu_\star)$ is W_p-compatible in the sense of Definition 5.2.2. Indeed, $\mu_\star = \Psi_{\#}\mu_{\mathrm{B}}$ via the map $\Psi(x) = \big(x^{(1)}/2, x^{(2)}, \dots, x^{(d)}\big)$, which is the time-1 flow map of

$$\dot{x}(t) = -(\log 2)\, x^{(1)}(t) e_1, \qquad t \in [0,1].$$

Let $\Phi = \Phi_T^\theta \in \mathcal{A}_{\mathcal{V}}$ and write $\Phi_{\#}\mu_{\mathrm{B}} = \rho_\Phi\,\mathcal{L}^d$. Since by assumption we have $\operatorname{div} v_{\theta(t)} = 0$ a.e., (1.2.3) yields $\rho_\Phi = \mathbf{1}_Q \circ \Phi^{-1} = \mathbf{1}_{\Phi(Q)}$ a.e. in $\mathbb{R}^d$. In particular, $\rho_\Phi \in \{0,1\}$ a.e., and therefore

$$\Phi_{\#}\mu_{\mathrm{B}}(E) = \int_E \rho_\Phi \,\mathrm{d}x \le \mathcal{L}^d(E) \qquad \text{for all Borel } E \in \mathcal{B}(\mathbb{R}^d).$$

Applying this to $E = H$ gives $\Phi_{\#}\mu_{\mathrm{B}}(H) \le \mathcal{L}^d(H) = \frac{1}{2}$. On the other hand, $\mu_\star(H) = \int_H 2\,\mathrm{d}x = 1$.

Assume by contradiction that there exists a sequence $\{\Phi_n\}_n \subset \mathcal{A}_{\mathcal{V}}$ such that $W_p\big((\Phi_n)_{\#}\mu_{\mathrm{B}}, \mu_\star\big) \to 0$. Since W_p-convergence implies weak convergence, and H is a continuity set for $\mu_\star$, we obtain

$$(\Phi_n)_{\#}\mu_{\mathrm{B}}(H) \to \mu_\star(H) = 1 \qquad \text{as } n \to \infty,$$

which contradicts $(\Phi_n)_{\#}\mu_{\mathrm{B}}(H) \le \frac{1}{2}$ for every n. Thus $\mu_\star \notin \overline{\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})}^{W_p}$, and the system does not possess the W_p-UAP. □

Proof of Theorem 5.3.8. We prove the two implications separately.

(⇒) Assume that the system possesses the Γ-quantitative π-UIP. Fix $\mu_{\mathrm{B}} \in \mathcal{P}_p(\mathbb{R}^d)$, a Lipschitz map $\Psi : \mathbb{R}^d \to \mathbb{R}^d$, and $\varepsilon > 0$, and set $L \coloneqq \mathrm{Lip}(\Psi)$.

Relying on (5.3.6), we can find an empirical measure $\mu_{\mathrm{B}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$ with N large enough such that $W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N)$ is arbitrarily small. By perturbing the points slightly if necessary, we may assume that the atoms x_i are pairwise distinct without loss of generality. Define $y_i \coloneqq \Psi(x_i)$ for $i \in [N]$ and let $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N$. Since the set $\{x_1, \dots, x_N\}$ is finite and pairwise distinct, we have $m_N \coloneqq \min_{i \neq j} \|x_i - x_j\| > 0$.

Fix $\tau > 0$ such that $2\tau \leq m_N$. We can perturb the target points to choose pairwise distinct $\tilde{y}_i \in \mathbb{R}^d$ satisfying $\|\tilde{y}_i - y_i\| < \tau$ for all $i \in [N]$. For the perturbed dataset $\widetilde{\mathcal{D}}_N \coloneqq \{(x_i, \tilde{y}_i)\}_{i=1}^N$, a straightforward check yields $\mathsf{L}(\widetilde{\mathcal{D}}_N) \leq L + 2\tau/m_N \leq L + 1$.

By the Γ-quantitative π-UIP, for any $\eta > 0$ there exist $\Phi \in \mathcal{A}_{\mathcal{V}}$ and $\pi \in S_N$ such that $\max_{i \in [N]} \|\Phi(x_i) - \tilde{y}_{\pi(i)}\| < \eta$ and $\mathrm{Lip}(\Phi) \leq \Gamma(L+1)$.

Let $\tilde{\mu}_N \coloneqq \frac{1}{N}\sum_{i=1}^N \delta_{\tilde{y}_i}$. Then (5.3.5) directly implies

$$W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \tilde{\mu}_N) < \eta \qquad \text{and} \qquad W_p(\tilde{\mu}_N, \Psi_{\#}\mu_{\mathrm{B}}^N) < \tau.$$

By the triangle inequality, this yields $W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \Psi_{\#}\mu_{\mathrm{B}}^N) < \eta + \tau$.

Finally, combining this bound with the Lipschitz pushforward property (5.3.4), we bound the true objective:

$$\begin{aligned} W_p(\Phi_{\#}\mu_{\mathrm{B}}, \Psi_{\#}\mu_{\mathrm{B}}) &\leq W_p(\Phi_{\#}\mu_{\mathrm{B}}, \Phi_{\#}\mu_{\mathrm{B}}^N) + W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \Psi_{\#}\mu_{\mathrm{B}}^N) + W_p(\Psi_{\#}\mu_{\mathrm{B}}^N, \Psi_{\#}\mu_{\mathrm{B}}) \\ &\leq \mathrm{Lip}(\Phi) W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N) + (\eta + \tau) + \mathrm{Lip}(\Psi) W_p(\mu_{\mathrm{B}}^N, \mu_{\mathrm{B}}) \\ &\leq \big(\Gamma(L+1) + L\big) W_p(\mu_{\mathrm{B}}, \mu_{\mathrm{B}}^N) + \eta + \tau. \end{aligned}$$

By choosing N large enough, and τ, η sufficiently small, this distance can be made strictly less than ε. Thus, the system possesses the quantitative W_p-UAP with modulus $L \mapsto \Gamma(L+1)$.

($\Leftarrow$) Assume now that the Γ-quantitative W_p-UAP holds. Fix $N \in \mathbb{N}$, $\mathcal{D}_N = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^d$ with $x_i \neq x_j$ and $y_i \neq y_j$ if $i \neq j$, and $\delta > 0$.

Define the uniform empirical measures

$$\mu_{\mathrm{B}}^N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \qquad \mu_\star^N = \frac{1}{N}\sum_{i=1}^N \delta_{y_i}.$$

Choose a permutation $\pi \in S_N$ that attains the minimum in (5.3.9), and define the map on the discrete set as $\Psi_0(x_i) \coloneqq y_{\pi(i)}$ for $i \in [N]$. Then $\mathrm{Lip}(\Psi_0) = \mathsf{L}(\mathcal{D}_N)$. By Kirszbraun's theorem [ALM21, Theorem 1.2], Ψ_0 extends to a global map $\Psi : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\mathrm{Lip}(\Psi) = \mathsf{L}(\mathcal{D}_N), \qquad \Psi_{\#}\mu_{\mathrm{B}}^N = \mu_\star^N.$$

Applying the Γ-quantitative W_p-UAP to $(\mu_{\mathrm{B}}^N, \Psi)$ with $\varepsilon = \delta N^{-1/p}$, we obtain $\Phi \in \mathcal{A}_{\mathcal{V}}$ such that

$$W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N) < \delta N^{-1/p}, \qquad \mathrm{Lip}(\Phi) \leq \Gamma\big(\mathsf{L}(\mathcal{D}_N)\big).$$

Again, by (5.3.5),

$$W_p^p(\Phi_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N) = \min_{\sigma \in S_N} \frac{1}{N}\sum_{i=1}^N \|\Phi(x_i) - y_{\sigma(i)}\|^p,$$

which immediately implies

$$\min_{\sigma \in S_N} \max_{i \in [N]} \|\Phi(x_i) - y_{\sigma(i)}\| \leq N^{1/p} W_p(\Phi_{\#}\mu_{\mathrm{B}}^N, \mu_\star^N) < \delta.$$

This exactly verifies the definition of the Γ-quantitative π-UIP. □

Proofs of Subsection 5.3.2: Approximation of transport maps

Proof of Proposition 5.3.9. Consider the coupling $\pi := (\mathsf{R},\mathsf{S})_{\#}\mu_{\mathrm{B}} \in \Pi(\mathsf{R}_{\#}\mu_{\mathrm{B}}, \mathsf{S}_{\#}\mu_{\mathrm{B}})$; recall (5.3.2). By definition of W_p,

$$W_p^p\big(\mathsf{R}_{\#}\mu_{\mathrm{B}}, \mathsf{S}_{\#}\mu_{\mathrm{B}}\big) \le \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \,\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d} \|\mathsf{R}(z)-\mathsf{S}(z)\|^p \,\mathrm{d}\mu_{\mathrm{B}}(z). \tag{5.6.1}$$

For the density statement, fix $\nu \in \mathcal{P}_p(\mathbb{R}^d)$. Since $\mu_{\mathrm{B}} \ll \mathcal{L}^d$, there exists a Borel map $\mathsf{T}^\star$ such that $\mathsf{T}^\star_{\#}\mu_{\mathrm{B}} = \nu$ (for instance, the Brenier map). Since $\mathcal{A}_{\mathcal{V}}$ is dense in $L^p(\mu_{\mathrm{B}})$, we can approximate $\mathsf{T}^\star$ by a sequence in $\mathcal{A}_{\mathcal{V}}$, and the conclusion follows immediately from the inequality above. □

Proof of Corollary 5.3.13. Fix $p \in [1,\infty)$ and let $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_p(\mathbb{R}^d)$ be compactly supported. By Proposition 5.3.11, it suffices to verify that the system possesses both the UIP and the diagonal shrinking property.

UIP. Since $\mathcal{V}$ is affine invariant and contains a non-affine vector field, [Che+25c, Theorem 2.6] yields the UIP.

Diagonal shrinking property. Fix a compact set $K \subset \mathbb{R}^d$ and $h \in \mathscr{C}^0_{\mathrm{nd}}(\mathbb{R})$. Let $v \in \mathcal{V}$ be a non-affine vector field. Then there exist ξ, η, b such that the scalar function

$$f(t) := \eta \cdot v(t\xi + b)$$

is non-affine. For each $i \in [d]$, affine invariance implies that the vector field $x \mapsto f(x^{(i)})\, e_i$ belongs to $\mathcal{V}$ (take $A = \xi e_i^\top$ and $W = e_i \eta^\top$ in Definition 5.2.1). Hence the induced one-dimensional control family on the i-th coordinate is affine invariant and contains a non-affine element. By [Che+25c, Theorem 2.6], this scalar system possesses the UIP.

Applying the one-dimensional argument from (5.3.10) on the projection of K onto the i-th axis, for every $\varepsilon > 0$ we obtain an attainable scalar flow $\phi^{(i)}$ such that

$$\sup_{x\in K} \big|\phi^{(i)}(x^{(i)}) - h(x^{(i)})\big| < \varepsilon.$$

Lifting these scalar flows to $\mathbb{R}^d$ and concatenating them, we obtain $\Phi \in \mathcal{A}_{\mathcal{V}}$ such that

$$\sup_{x\in K} \Big\| \Phi(x) - \big(h(x^{(1)}), \dots, h(x^{(d)})\big) \Big\| \le \sqrt{d}\,\varepsilon.$$

Thus the diagonal shrinking property holds.

Consequently, both assumptions of Proposition 5.3.11 are satisfied. Applying this result yields the desired statement. □

5.6.2 Proofs of Section 5.4: Total variation distance

Preliminary lemmas

The following lemma records contraction in total variation under pushforward, and exact preservation for measurable bijections.

Lemma 5.6.3. *Let $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ be measurable and let μ, ν be finite signed measures on $\mathbb{R}^d$. Then*

$$\big|\, \mathsf{T}_{\#}\mu - \mathsf{T}_{\#}\nu \,\big|_{\mathsf{TV}} \;\le\; |\,\mu - \nu\,|_{\mathsf{TV}}.$$

If, in addition, T *is bijective and* T^{-1} *is measurable, then equality holds:*

$$\big|\,\mathsf{T}_{\#}\mu - \mathsf{T}_{\#}\nu\,\big|_{\mathsf{TV}} \;=\; |\,\mu-\nu\,|_{\mathsf{TV}}.$$

Proof of Lemma 5.6.3. For the contraction, we use the dual characterization of the total variation norm:

$$\begin{aligned}\big|\mathsf{T}_{\#}\mu - \mathsf{T}_{\#}\nu\big|_{\mathsf{TV}} &= \sup_{\|h\|_\infty\le 1}\Big|\int h\,\mathrm{d}(\mathsf{T}_{\#}\mu-\mathsf{T}_{\#}\nu)\Big| = \sup_{\|h\|_\infty\le 1}\Big|\int h\circ\mathsf{T}\,\mathrm{d}(\mu-\nu)\Big|\\ &\le \sup_{\|g\|_\infty\le 1}\Big|\int g\,\mathrm{d}(\mu-\nu)\Big| = |\mu-\nu|_{\mathsf{TV}},\end{aligned}$$

where the inequality follows because $h\circ\mathsf{T}$ is measurable and $\|h\circ\mathsf{T}\|_\infty\le 1$.

Assume now that T is bijective and both T and T^{-1} are measurable. Applying the contraction property proven above to the map T^{-1} and the signed measure $\mathsf{T}_{\#}\mu-\mathsf{T}_{\#}\nu$ yields

$$|\,\mu-\nu\,|_{\mathsf{TV}} = \big|\,\mathsf{T}^{-1}_{\#}(\mathsf{T}_{\#}\mu-\mathsf{T}_{\#}\nu)\,\big|_{\mathsf{TV}} \le \big|\,\mathsf{T}_{\#}\mu-\mathsf{T}_{\#}\nu\,\big|_{\mathsf{TV}}.$$

Combining this with the forward inequality gives the equality. □

Remark 5.6.4. A useful consequence of Lemma 5.6.3 is that the L^1-error introduced when perturbing an initial density is preserved along its evolution.

Indeed, fix

$$v\in L^\infty\big(0,T;\mathscr{C}^{0,1}(\mathbb{R}^d;\mathbb{R}^d)\big),\qquad \rho_{\mathrm{B}}\in L^1(\mathbb{R}^d),\qquad \eta\in L^1(\mathbb{R}^d),$$

and denote by ρ_t^η and ρ_t^0 the unique (weak) solutions of

$$\partial_t\rho_t+\operatorname{div}\big(\rho_t\,v(\cdot,t)\big)=0,\qquad \rho^\eta_{|t=0}=\rho_{\mathrm{B}}+\eta,\quad \rho^0_{|t=0}=\rho_{\mathrm{B}}.$$

Since the flow Φ_t is a bi-Lipschitz homeomorphism, the difference satisfies

$$(\rho_t^\eta-\rho_t^0)\,\mathcal{L}^d=(\Phi_t)_{\#}\big((\rho_{\mathrm{B}}+\eta)\mathcal{L}^d\big)-(\Phi_t)_{\#}\big(\rho_{\mathrm{B}}\mathcal{L}^d\big)=(\Phi_t)_{\#}\big(\eta\,\mathcal{L}^d\big).$$

Applying the equality case of Lemma 5.6.3 to the signed measure $\eta\,\mathcal{L}^d$ gives

$$\|\rho_t^\eta-\rho_t^0\|_{L^1(\mathbb{R}^d)}=\|\eta\|_{L^1(\mathbb{R}^d)}\qquad\text{for all } t\in[0,T].$$

We now state and prove the lemmas that were used in the proof of Corollary 5.4.5.

Lemma 5.6.5. *Let* $\mathcal{V}$ *be a Lipschitz control family dense in* $\mathscr{C}^1_{\mathrm{loc}}(\mathbb{R})$*. Then, for every increasing* $\mathscr{C}^1$*-diffeomorphism* $h:\mathbb{R}\to\mathbb{R}$*, every compact interval* I*, and every* $\varepsilon>0$*, there exists* $\Phi\in\mathcal{A}_{\mathcal{V}}$ *such that* $\|h-\Phi\|_{W^{1,1}(I)}<\varepsilon$*.*

Proof of Lemma 5.6.5. Let $I\subset\mathbb{R}$ be compact and define the linear homotopy

$$\Psi_t:=(1-t)\mathsf{Id}+th\qquad\text{for } t\in[0,1].$$

Since h is an increasing $\mathscr{C}^1$-diffeomorphism, each Ψ_t is again an increasing $\mathscr{C}^1$-diffeomorphism. Let $K:=\operatorname{conv}\big(I\cup h(I)\big)$, so that $\Psi_t(I)\subseteq K$ for all $t\in[0,1]$. Since h' is continuous and strictly positive, $\partial_x\Psi_t=(1-t)+th'$ is strictly positive on K. Consequently, both $\partial_x\Psi_t$ and $(\partial_x\Psi_t)^{-1}$ are uniformly bounded above and below on K, uniformly with respect to $t\in[0,1]$.

Define the time-dependent vector field generating $(\Psi_t)_{t\in[0,1]}$ by

$$u_t(y) \coloneqq \partial_t \Psi_t(\Psi_t^{-1}(y)) = (h - \mathsf{Id})\big(\Psi_t^{-1}(y)\big), \qquad y \in K.$$

Then $u \in \mathscr{C}\big([0,1];\mathscr{C}^1(K)\big)$ and $\sup_t \operatorname{Lip}(u_t|_K) < \infty$. Since the map $t \mapsto u_t$ is continuous with values in a Banach space, and by the density of $\mathcal{V}$ in $\mathscr{C}^1_{\mathrm{loc}}(\mathbb{R})$, for any $\delta > 0$ there exists a piecewise constant-in-time field w_t with values in $\mathcal{V}$ such that

$$\int_0^1 \|u_t - w_t\|_{\mathscr{C}^1(K)}\,\mathrm{d}t < \delta. \tag{5.6.2}$$

Let Φ_t be the flow of $\dot{x} = w_t(x)$ on $[0,1]$. A Grönwall estimate on $\Psi_t - \Phi_t$ yields

$$\sup_{t\in[0,1]} \|\Psi_t - \Phi_t\|_{L^\infty(I)} \le C_0\delta \tag{5.6.3}$$

for some constant $C_0 > 0$ depending only on $\sup_t \operatorname{Lip}(u_t|_K)$ and K.

For the derivatives, set $Y_t(x) \coloneqq \partial_x \Psi_t(x)$ and $Z_t(x) \coloneqq \partial_x \Phi_t(x)$. They satisfy

$$\partial_t Y_t(x) = u_t'(\Psi_t(x))\,Y_t(x), \qquad \partial_t Z_t(x) = w_t'(\Phi_t(x))\,Z_t(x).$$

Let $\Delta_t \coloneqq Y_t - Z_t$. Subtracting the equations gives

$$\partial_t \Delta_t = \big[u_t'(\Psi_t) - u_t'(\Phi_t)\big]Y_t + \big[u_t'(\Phi_t) - w_t'(\Phi_t)\big]Y_t + w_t'(\Phi_t)\Delta_t. \tag{5.6.4}$$

We bound the right-hand side of (5.6.4) term by term:

(i) Since u' is uniformly continuous, the first term is bounded by $C_Y\,\omega_{u'}\big(\|\Psi_t - \Phi_t\|_{L^\infty(I)}\big)$, where $C_Y \coloneqq \sup_{(t,x)\in[0,1]\times I} |Y_t(x)|$ and $\omega_{u'}$ is a modulus of continuity for u' on $[0,1]\times K$. By (5.6.3), we know that $\|\Psi_t - \Phi_t\|_{L^\infty(I)} \le C_0\delta$, so this first term is ultimately bounded by $C_Y\,\omega_{u'}(C_0\delta)$.

(ii) The second term is bounded by $C_Y\|u_t - w_t\|_{\mathscr{C}^1(K)}$, which is at most $C_Y\delta$.

(iii) In the third term, $|w_t'(\Phi_t)|$ is uniformly bounded on K by some $L > 0$, thanks to (5.6.2).

All in all, taking $L^\infty(I)$ norms in (5.6.4) and applying Grönwall's lemma yields

$$\|\Delta_1\|_{L^\infty(I)} \le \eta(\delta)\,e^L,$$

where we have introduced

$$\eta(\delta) \coloneqq C_Y\delta + C_Y\omega_{u'}(C_0\delta) \to 0 \qquad \text{as } \delta \to 0.$$

Thus, $\|\partial_x h - \partial_x\Phi_1\|_{L^\infty(I)} \to 0$ as $\delta \to 0$, and therefore we also have $\|\partial_x h - \partial_x\Phi_1\|_{L^1(I)} \to 0$. Combining this with (5.6.3) gives $\|h - \Phi_1\|_{W^{1,1}(I)} < \varepsilon$ for δ sufficiently small. □

The following lemma provides a $W^{1,1}$ version of the Lie–Trotter splitting, allowing us to pass from the span of vector fields to the original family.

Lemma 5.6.6. *Let $v_1, \dots, v_p : \mathbb{R}^d \to \mathbb{R}^d$ be Lipschitz and $\mathscr{C}^1$, and set $v \coloneqq \sum_{i=1}^p v_i$. Let Φ_t be the flow of $\dot{x} = v(x)$. Fix $t > 0$ and define, for each $n \in \mathbb{N}$,*

$$S_{n,t} \coloneqq \big(\Phi^{v_p}_{t/n} \circ \cdots \circ \Phi^{v_1}_{t/n}\big) \circ \cdots \circ \big(\Phi^{v_p}_{t/n} \circ \cdots \circ \Phi^{v_1}_{t/n}\big) \qquad (n \text{ times}),$$

where $\Phi^{v_i}_s$ denotes the time-s flow of $\dot{x} = v_i(x)$. Then, for every compact set $K \subset \mathbb{R}^d$,

$$\|S_{n,t} - \Phi_t\|_{W^{1,1}(K)} \to 0 \qquad \text{as } n \to \infty.$$

Proof of Lemma 5.6.6. Fix $t > 0$, a compact $K \subset \mathbb{R}^d$, and set $h := t/n$. Since all vector fields are Lipschitz, there exists a compact $K' \supset K$ such that trajectories of both $\dot{x} = v(x)$ and the splitting approximations starting in K stay in K' up to time t.

Step 1: $\mathscr{C}^2$–fields. Assume first that $v_i \in \mathscr{C}^2$. Taylor estimates then give, uniformly on K',

$$\|\Phi_h^v - (\mathsf{Id} + hv)\|_{\mathscr{C}^1(K')} \le C_1 h^2, \qquad \|\Phi_h^{v_i} - (\mathsf{Id} + hv_i)\|_{\mathscr{C}^1(K')} \le C_1 h^2,$$

for some C_1 depending on the $\mathscr{C}^2(K')$ norms. Using $v = \sum_i v_i$ and composing,

$$\|\Phi_h^{v_p} \circ \cdots \circ \Phi_h^{v_1} - (\mathsf{Id} + hv)\|_{\mathscr{C}^1(K')} \le C_1' h^2$$

for some C_1', hence

$$\|\Phi_h^v - \Phi_h^{v_p} \circ \cdots \circ \Phi_h^{v_1}\|_{\mathscr{C}^1(K')} \le C_2 h^2.$$

Let $t_k = kh$ and $E_k := \|\Phi_{t_k} - S_{k,h}\|_{\mathscr{C}^1(K')}$. One-step stability gives

$$\begin{aligned} E_{k+1} &= \|\Phi_h \circ \Phi_{t_k} - S_{1,h} \circ S_{k,h}\|_{\mathscr{C}^1(K')} \\ &\le \|\Phi_h \circ \Phi_{t_k} - \Phi_h \circ S_{k,h}\|_{\mathscr{C}^1(K')} + \|\Phi_h \circ S_{k,h} - S_{1,h} \circ S_{k,h}\|_{\mathscr{C}^1(K')} \\ &\le \big(1 + h\|\nabla v\|_{L^\infty(K')}\big) E_k + C_2 h^2. \end{aligned}$$

By discrete Grönwall, and because $(1 + hL)^n \le e^{nhL} = e^{tL}$ for all $L > 0$,

$$E_n \le C_2\, h\, \frac{(1 + h\|\nabla v\|_{L^\infty(K')})^n - 1}{\|\nabla v\|_{L^\infty(K')}} \le C_2\, h\, \frac{e^{t\|\nabla v\|_{L^\infty(K')}} - 1}{\|\nabla v\|_{L^\infty(K')}}.$$

(If $\|\nabla v\|_{L^\infty(K')} = 0$ take the continuous limit of the bound.) This proves the claim for $\mathscr{C}^2$-fields, since convergence in $\mathscr{C}^1(K')$ implies convergence in $W^{1,1}(K)$.

Step 2: $\mathscr{C}^1$–fields. For v_i merely $\mathscr{C}^1$ and Lipschitz, approximate by $v_i^\varepsilon \in \mathscr{C}^2$ such that

$$\|v_i^\varepsilon - v_i\|_{\mathscr{C}^1(K')} \to 0 \qquad \text{as } \varepsilon \to 0,$$

and set $v^\varepsilon := \sum_i v_i^\varepsilon$. Denote by Φ_t^ε and $S_{n,t}^\varepsilon$ the corresponding flow and splitting map. Then

$$\|S_{n,t} - \Phi_t\|_{W^{1,1}(K)} \le \|S_{n,t} - S_{n,t}^\varepsilon\|_{W^{1,1}(K)} + \|S_{n,t}^\varepsilon - \Phi_t^\varepsilon\|_{W^{1,1}(K)} + \|\Phi_t^\varepsilon - \Phi_t\|_{W^{1,1}(K)}.$$

The first and third terms vanish as $\varepsilon \to 0$ by continuous dependence of flows on the vector field in the $\mathscr{C}^1(K')$ topology (for $S_{n,t}$ this follows by viewing it as the time–t flow of a piecewise constant in time vector field taking values in $\{v_1, \dots, v_p\}$). Fix $\eta > 0$; choose ε so small that both are $< \eta/3$ uniformly in n. For this fixed ε, Step 1 applies and yields n large such that the middle term is $< \eta/3$. Hence $\|S_{n,t} - \Phi_t\|_{W^{1,1}(K)} \to 0$ as $n \to \infty$.

□

Lemma 5.6.7. *Let $I \subset \mathbb{R}$ be a compact interval and $u \in W^{1,1}(I)$ a nondecreasing function. Then, for any $\varepsilon > 0$ there exists a strictly increasing, continuous piecewise affine function $\phi : I \to \mathbb{R}$ with finitely many pieces such that*

$$\|u - \phi\|_{W^{1,1}(I)} < \varepsilon.$$

Proof of Lemma 5.6.7. Write $I = [a, b]$ and set $L := b - a$. The case $L = 0$ is trivial, so assume $L > 0$. Since $u \in W^{1,1}(I)$, we have

$$u(x) = u(a) + \int_a^x u'(t)\,\mathrm{d}t \qquad \text{for all } x \in I,$$

and since u is nondecreasing, $u' \geq 0$ a.e. on I. Step functions are dense in $L^1(I)$, hence there exists a piecewise constant $g \geq 0$ with finitely many values such that

$$\|u' - g\|_{L^1(I)} < \frac{\varepsilon}{2(1+L)}.$$

Choose $\eta > 0$ so small that $\eta L(1+L) < \varepsilon/2$, and define

$$\phi(x) \coloneqq u(a) + \int_a^x \big(g(t) + \eta\big)\,\mathrm{d}t.$$

Then ϕ is continuous, piecewise affine with finitely many pieces, and

$$\phi'(x) = g(x) + \eta \geq \eta > 0 \qquad \text{a.e. on } I,$$

hence ϕ is strictly increasing. Moreover,

$$\|u' - \phi'\|_{L^1(I)} = \|u' - (g+\eta)\|_{L^1(I)} \leq \|u' - g\|_{L^1(I)} + \eta L.$$

Also, for every $x \in I$,

$$|u(x) - \phi(x)| = \left| \int_a^x \big(u'(t) - (g(t)+\eta)\big)\,\mathrm{d}t \right| \leq \|u' - \phi'\|_{L^1(I)},$$

so

$$\|u - \phi\|_{L^1(I)} \leq L\,\|u' - \phi'\|_{L^1(I)}.$$

Therefore,

$$\|u - \phi\|_{W^{1,1}(I)} \leq (1+L)\|u' - \phi'\|_{L^1(I)} \leq (1+L)\big(\|u' - g\|_{L^1(I)} + \eta L\big) < \varepsilon.$$

□

The following lemma demonstrates that scalar piecewise affine functions can be exactly represented by flows from the family $(\mathcal{V}_1^+)$.

Lemma 5.6.8. *Consider the family $(\mathcal{V}_1^+)$. Let I be a compact interval and let $h : I \to \mathbb{R}$ be strictly increasing and piecewise affine. Then there exists $\Phi \in \mathcal{A}_{\mathcal{V}}$ such that $\Phi(x) = h(x)$ for all $x \in I$.*

Proof of Lemma 5.6.8. Any h strictly increasing and piecewise affine on $I = [a, b]$ is determined by its value $h(a)$ and its slopes $m_0, \ldots, m_N \geq 0$ on the subintervals defined by

$$a = z_0 < z_1 < \cdots < z_{N+1} = b.$$

We construct the diffeomorphism Φ by composing a sequence of flows. The strategy is to match the function from left to right. First, we match the starting value $h(a)$. Consider the constant field $v_0(x) = h(a) - a$. Its flow at time $t = 1$ is the translation $\Phi_0(x) = x + (h(a) - a)$. Note that $\Phi_0(a) = h(a)$.

We now adjust the slopes iteratively. Suppose that after step k, we have constructed a map Φ_k that matches h exactly on the interval $[a, z_k]$ and has a constant slope S_k on the remaining interval $(z_k, b]$. We want to change the slope on the next segment $(z_k, z_{k+1}]$ from S_k to the target m_k, without altering the map on $[a, z_k]$.

Let $y_k \coloneqq h(z_k)$ be the image of the current knot. Consider the field $v(x) = \alpha_k(x - y_k)_+$ with $\alpha_k = \log(m_k/S_k)$. By Proposition 1.3.3, the flow Ψ generated by v at time $t = 1$ acts as:

$$\Psi(x) = \begin{cases} x & \text{if } x \le y_k, \\ y_k + e^{\alpha_k}(x - y_k) & \text{if } x > y_k. \end{cases}$$

Since $e^{\alpha_k} = m_k/S_k$, the composition $\Phi_{k+1} \coloneqq \Psi \circ \Phi_k$ preserves the map up to z_k (where $\Phi_k(x) \le y_k$) and updates the slope on $(z_k, b]$ to $S_k \cdot (m_k/S_k) = m_k$. Repeating this process for $k = 0, \dots, N$ yields the desired map Φ. □

Remark 5.6.9. While the sequential composition of flows generated by single neurons yields piecewise affine maps, the flow generated by a linear combination of neurons generally does not.

To see this, fix a time $t > 0$ and consider the system $\dot{x} = v(x)$ defined by $v(x) \coloneqq x_+ + (x - 1)_+$. This vector field partitions the dynamics into two distinct affine regimes:

$$\dot{x} = x \quad \text{for } x \in [0, 1], \qquad \text{and} \qquad \dot{x} = 2x - 1 \quad \text{for } x > 1.$$

Take an initial condition $x_0 \in (e^{-t}, 1)$. Initially, the trajectory follows $\dot{x} = x$, reaching $x = 1$ at time $\tau = -\ln x_0$. Because $x_0 > e^{-t}$, we are guaranteed that $\tau < t$.

For the remaining time interval of length $t - \tau$, the system enters the second regime and evolves according to $\dot{x} = 2x - 1$, starting from $x(\tau) = 1$. Integrating this equation yields the flow at time t:

$$\Phi_t(x_0) = \frac{1}{2} + \frac{1}{2}e^{2(t-\tau)}.$$

Substituting $e^{-2\tau} = e^{2\ln x_0} = x_0^2$, we obtain:

$$\Phi_t(x_0) = \frac{1}{2} + \frac{1}{2}e^{2t}x_0^2.$$

This strictly quadratic dependence on x_0 confirms that $x_0 \mapsto \Phi_t(x_0)$ is not piecewise affine.

Finally, we show the connection between TV and W_p.

Proof of Lemma 5.4.2. We first show that for every $R > 0$ and every $\mu, \nu \in \mathcal{F}$,

$$W_p(\mu, \nu)^p \le (2R)^p |\mu - \nu|_{\mathsf{TV}} + 2^p \left(\int_{\|x\| > R} \|x\|^p \, \mathrm{d}\mu(x) + \int_{\|x\| > R} \|x\|^p \, \mathrm{d}\nu(x) \right). \tag{5.6.5}$$

Let $\pi \in \Pi(\mu, \nu)$ be an optimal coupling for the total variation distance (recall (5.4.3)), thus satisfying $\pi(x \neq y) = |\mu - \nu|_{\mathsf{TV}}$. By (5.3.1), we have

$$W_p(\mu, \nu)^p \le \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \mathrm{d}\pi(x, y).$$

We split this integral over the truncated domain

$$A \coloneqq \big\{(x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mid \|x\| \le R, \ \|y\| \le R\big\}$$

and its complement A^c. Since $\|x - y\| \le 2R$ on A, and $\|x - y\| = 0$ on $\{x = y\}$, we have

$$\int_A \|x - y\|^p \, \mathrm{d}\pi(x, y) \le (2R)^p \pi(x \neq y) = (2R)^p |\mu - \nu|_{\mathsf{TV}}.$$

On A^c, the triangle inequality and the convexity of $t \mapsto t^p$ for $p \geq 1$ yield the elementary bound $\|x-y\|^p \leq 2^{p-1}\big(\|x\|^p + \|y\|^p\big)$. Therefore, the second contribution satisfies

$$\int_{A^c} \|x-y\|^p \,\mathrm{d}\pi(x,y) \leq 2^{p-1}\left(\int_{A^c} \|x\|^p \,\mathrm{d}\pi(x,y) + \int_{A^c} \|y\|^p \,\mathrm{d}\pi(x,y)\right).$$

To estimate the terms on the right-hand side, observe that for any $(x,y) \in A^c$, if $\|x\| \leq R$ then necessarily $\|y\| > R$, which implies $\|x\|^p \leq R^p < \|y\|^p$. Consequently, we can bound $\|x\|^p$ pointwise on A^c by

$$\|x\|^p \leq \|x\|^p \mathbf{1}_{\{\|x\|>R\}} + \|y\|^p \mathbf{1}_{\{\|y\|>R\}}.$$

Integrating this against π and passing to the marginals gives

$$\int_{A^c} \|x\|^p \,\mathrm{d}\pi(x,y) \leq \int_{\|x\|>R} \|x\|^p \,\mathrm{d}\mu(x) + \int_{\|y\|>R} \|y\|^p \,\mathrm{d}\nu(y).$$

By symmetry, the exact same bound applies to $\int_{A^c} \|y\|^p \,\mathrm{d}\pi(x,y)$. Substituting these two bounds back into the integral over A^c, and adding the estimate over A, yields the claimed inequality (5.6.5).

To conclude, let $\varepsilon > 0$. By the uniform integrability on $\mathcal{F}$, we can fix $R > 0$ large enough so that

$$2^p \int_{\|x\|>R} \|x\|^p \,\mathrm{d}\eta < \varepsilon^p/2 \qquad \text{for all } \eta \in \mathcal{F}.$$

Once R is fixed, the TV-UAP guarantees we can make $|\mu-\nu|_{\mathsf{TV}}$ small enough so that

$$(2R)^p |\mu-\nu|_{\mathsf{TV}} < \varepsilon^p/2.$$

Then (5.6.5) yields $W_p(\mu,\nu) < \varepsilon$, concluding the proof. □

Proofs of Subsection 5.4.1: One-dimensional case

The following lemma will be used in the proof of Proposition 5.4.3.

Lemma 5.6.10. *Let $\mu \in \mathcal{P}_{ac}(\mathbb{R})$ with density $\rho \in \mathscr{C}^1(\mathbb{R})$. Fix $R > 1$ and let $I_R := [-R,R]$. Let $h : \mathbb{R} \to \mathbb{R}$ be an increasing diffeomorphism and define*

$$J_R := h^{-1}(I_{R+1}), \qquad m_R := \inf_{x \in J_R} h'(x).$$

Assume that $h|_{J_R} \in \mathscr{C}^{1,1}(J_R)$ and $m_R > 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be an increasing bi-Lipschitz homeomorphism satisfying

$$\|g-h\|_{W^{1,1}(J_R)} \leq 1 \wedge C_R^{-1},$$

where C_R is the embedding constant for $W^{1,1}(J_R) \hookrightarrow \mathscr{C}^0(\overline{J_R})$. Then there exists a constant $C = C(h|_{J_R}, \rho, R, m_R)$ such that

$$|g_\#\mu - h_\#\mu|_{\mathsf{TV}} \;\leq\; C\,\|g-h\|_{W^{1,1}(J_R)} \;+\; |h_\#\mu|_{\mathsf{TV}(I_R^c)} \;+\; |h_\#\mu|_{\mathsf{TV}(I_{R-1}^c)}.$$

Proof of Lemma 5.6.10. We decompose the total variation on I_R and its complement:

$$|g_\#\mu - h_\#\mu|_{\mathsf{TV}} \leq |g_\#\mu - h_\#\mu|_{\mathsf{TV}(I_R)} + |g_\#\mu|_{\mathsf{TV}(I_R^c)} + |h_\#\mu|_{\mathsf{TV}(I_R^c)}.$$

Since by assumption $h|_{J_R} \in \mathscr{C}^{1,1}$ and g is bi-Lipschitz, both h^{-1} and g^{-1} are Lipschitz (hence absolutely continuous) on the relevant domains. Consequently, the pushforward densities exist a.e. and are given by (1.2.2). Moreover, for a.e. $y \in I_R$,

$$(h^{-1})'(y) = \frac{1}{h'(h^{-1}(y))}, \qquad (g^{-1})'(y) = \frac{1}{g'(g^{-1}(y))}. \tag{5.6.6}$$

Using $|ab - cd| \le |a-c| \cdot |b| + |c| \cdot |b-d|$, we bound the term on I_R:

$$\begin{aligned} |g_\# \mu - h_\# \mu|_{\mathsf{TV}(I_R)} =& \frac{1}{2} \int_{I_R} \left| \rho\big(g^{-1}(y)\big)\, (g^{-1})'(y) - \rho\big(h^{-1}(y)\big)\, (h^{-1})'(y) \right| \mathrm{d}y \\ \le& \frac{1}{2} \int_{I_R} \left| \rho\big(g^{-1}(y)\big) - \rho\big(h^{-1}(y)\big) \right| (g^{-1})'(y)\, \mathrm{d}y \\ &+ \frac{1}{2} \int_{I_R} \rho\big(h^{-1}(y)\big) \left| (g^{-1})'(y) - (h^{-1})'(y) \right| \mathrm{d}y =: \mathcal{I}_1 + \mathcal{I}_2. \end{aligned} \tag{5.6.7}$$

Step 1: Tail control. By Sobolev embedding on J_R,

$$\|g - h\|_{L^\infty(J_R)} \le C_R \|g - h\|_{W^{1,1}(J_R)} \le 1.$$

Write $J_R = [a, b]$. Since $h(J_R) = I_{R+1} = [-R-1, R+1]$, we have $h(a) = -R-1$ and $h(b) = R+1$. Using $|g - h| \le \|g - h\|_{L^\infty(J_R)} \le 1$ at the endpoints,

$$g(a) \le h(a) + 1 = -R, \qquad g(b) \ge h(b) - 1 = R.$$

Since g is increasing, $g(J_R) = [g(a), g(b)] \supset [-R, R] = I_R$, and for $x \notin J_R$ monotonicity implies $g(x) \notin I_R$. Hence

$$g^{-1}(I_R) \subset J_R. \tag{5.6.8}$$

Moreover, if $z \in h^{-1}(I_{R-1})$, then $h(z) \in I_{R-1}$ and $|g(z) - h(z)| \le 1$ implies $g(z) \in I_R$. Thus $h^{-1}(I_{R-1}) \subset g^{-1}(I_R)$, i.e. $g^{-1}(I_R^c) \subset h^{-1}(I_{R-1}^c)$, and therefore

$$|g_\# \mu|_{\mathsf{TV}(I_R^c)} \le |h_\# \mu|_{\mathsf{TV}(I_{R-1}^c)}. \tag{5.6.9}$$

Step 2: Bound for $\mathcal{I}_1$. Let $L_R \coloneqq \|\rho'\|_{L^\infty(J_R)}$. By (5.6.8), we can use the mean value theorem:

$$\mathcal{I}_1 \le \frac{1}{2} L_R \int_{I_R} |g^{-1}(y) - h^{-1}(y)|\, (g^{-1})'(y)\, \mathrm{d}y \le \frac{1}{2} L_R \, \|g^{-1} - h^{-1}\|_{L^\infty(I_R)}\, |g^{-1}(I_R)|. \tag{5.6.10}$$

By (5.6.8), we have $|g^{-1}(I_R)| \le |J_R|$. To bound $\|g^{-1} - h^{-1}\|_{L^\infty(I_R)}$, use that h^{-1} is $(1/m_R)$-Lipschitz on I_{R+1} because $h' \ge m_R$ on J_R. For $y \in I_R$, write $y = g(z)$ with $z = g^{-1}(y) \in J_R$:

$$|g^{-1}(y) - h^{-1}(y)| = |z - h^{-1}(g(z))| = |h^{-1}(h(z)) - h^{-1}(g(z))| \le \frac{1}{m_R} |h(z) - g(z)|.$$

Thus

$$\|g^{-1} - h^{-1}\|_{L^\infty(I_R)} \le \frac{\|g - h\|_{L^\infty(J_R)}}{m_R} \le \frac{C_R}{m_R} \|g - h\|_{W^{1,1}(J_R)}.$$

Substituting into (5.6.10) gives

$$\mathcal{I}_1 \le C_1(h|_{J_R}, \rho, R, m_R)\, \|g - h\|_{W^{1,1}(J_R)}. \tag{5.6.11}$$

Step 3: Bound for $\mathcal{I}_2$. Using (5.6.6), for a.e. $y \in I_R$,

$$\begin{aligned}|(g^{-1})'(y) - (h^{-1})'(y)| &= \left|\frac{1}{g'(g^{-1}(y))} - \frac{1}{h'(h^{-1}(y))}\right| \\ &\le \left|\frac{1}{g'(g^{-1}(y))} - \frac{1}{h'(g^{-1}(y))}\right| + \left|\frac{1}{h'(g^{-1}(y))} - \frac{1}{h'(h^{-1}(y))}\right|.\end{aligned}$$

For the first term, make the change of variables $y = g(x)$ (valid since g is bi-Lipschitz) and use $g^{-1}(I_R) \subset J_R$:

$$\int_{I_R} \left|\frac{1}{g'(g^{-1}(y))} - \frac{1}{h'(g^{-1}(y))}\right| \mathrm{d}y = \int_{g^{-1}(I_R)} \frac{|h'(x) - g'(x)|}{h'(x)} \,\mathrm{d}x \le \frac{1}{m_R} \|g' - h'\|_{L^1(J_R)}.$$

For the second term, since $h \in \mathscr{C}^{1,1}$ on the compact interval J_R, the map $x \mapsto 1/h'(x)$ is Lipschitz on J_R with constant $\mathrm{Lip}(h')/m_R^2$. Therefore

$$\int_{I_R} \left|\frac{1}{h'(g^{-1}(y))} - \frac{1}{h'(h^{-1}(y))}\right| \mathrm{d}y \le 2R \,\frac{\mathrm{Lip}(h')}{m_R^2} \,\|g^{-1} - h^{-1}\|_{L^\infty(I_R)}.$$

Using the bound on $\|g^{-1} - h^{-1}\|_{L^\infty(I_R)}$ from Step 2 gives

$$\mathcal{I}_2 \le C_2(h|_{J_R}, \rho, R, m_R) \,\|g - h\|_{W^{1,1}(J_R)}. \tag{5.6.12}$$

Combining (5.6.7)–(5.6.9)–(5.6.11)–(5.6.12), the result follows. □

Proof of Proposition 5.4.3. Let $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R})$ and denote their densities by ρ_{B} and $\rho_\star$. Fix a Gaussian mollifier η_δ and define smooth approximations $\mu_{\mathrm{B},\delta}$ and $\mu_{\star,\delta}$ by convolving the densities. These approximations satisfy

$$|\mu_{\mathrm{B},\delta} - \mu_{\mathrm{B}}|_{\mathsf{TV}} + |\mu_{\star,\delta} - \mu_\star|_{\mathsf{TV}} \longrightarrow 0 \qquad \text{as } \delta \to 0. \tag{5.6.13}$$

Since the convolved densities are strictly positive and smooth, their monotone rearrangement map M_δ (defined as in (5.4.4)) is a smooth increasing diffeomorphism of $\mathbb{R}$ satisfying $(\mathsf{M}_\delta)_{\#}\mu_{\mathrm{B},\delta} = \mu_{\star,\delta}$.

Fix $\varepsilon > 0$. First, choose $\delta > 0$ such that

$$|\mu_{\mathrm{B},\delta} - \mu_{\mathrm{B}}|_{\mathsf{TV}} + |\mu_{\star,\delta} - \mu_\star|_{\mathsf{TV}} < \frac{\varepsilon}{3}. \tag{5.6.14}$$

Next, choose $R > 1$ sufficiently large so that $I_R := [-R, R]$ captures most of the target mass:

$$|\mu_{\star,\delta}|_{\mathsf{TV}(I_R^c)} + |\mu_{\star,\delta}|_{\mathsf{TV}(I_{R-1}^c)} < \frac{\varepsilon}{3}. \tag{5.6.15}$$

Define $J_R := \mathsf{M}_\delta^{-1}(I_{R+1})$. Since M_δ is continuous and proper, J_R is a compact interval. Furthermore, since M_δ is smooth and strictly increasing,

$$m_R := \inf_{x \in J_R} \mathsf{M}_\delta'(x) > 0.$$

By assumption (5.4.5) applied with $h = \mathsf{M}_\delta$ and $I = J_R$, for any $\eta > 0$ there exists $\Phi \in \mathcal{A}_\mathcal{V}$ such that

$$\|\Phi - \mathsf{M}_\delta\|_{W^{1,1}(J_R)} \le \eta, \qquad \|\Phi - \mathsf{M}_\delta\|_{L^\infty(J_R)} \le C_R\,\eta,$$

where C_R is the embedding constant $W^{1,1}(J_R) \hookrightarrow \mathscr{C}^0(\overline{J_R})$. We may assume $\eta \le 1 \wedge C_R^{-1}$ (decreasing η if necessary). Note that Φ is a bi-Lipschitz increasing homeomorphism, so we can apply Lemma 5.6.10 with $\mu = \mu_{\mathrm{B},\delta}$, $h = \mathsf{M}_\delta$, and $g = \Phi$. Since $(\mathsf{M}_\delta)_\# \mu_{\mathrm{B},\delta} = \mu_{\star,\delta}$, the lemma gives

$$|\Phi_\# \mu_{\mathrm{B},\delta} - \mu_{\star,\delta}|_{\mathsf{TV}} \le C\eta + |\mu_{\star,\delta}|_{\mathsf{TV}(I_R^c)} + |\mu_{\star,\delta}|_{\mathsf{TV}(I_{R-1}^c)}, \tag{5.6.16}$$

for some constant C depending on $(\mathsf{M}_\delta|_{J_R}, \rho_{\mathrm{B}} * \eta_\delta, R, m_R)$.

Finally, using the triangle inequality and Lemma 5.6.3,

$$\begin{aligned} |\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}} &\le |\Phi_\# \mu_{\mathrm{B}} - \Phi_\# \mu_{\mathrm{B},\delta}|_{\mathsf{TV}} + |\Phi_\# \mu_{\mathrm{B},\delta} - \mu_{\star,\delta}|_{\mathsf{TV}} + |\mu_{\star,\delta} - \mu_\star|_{\mathsf{TV}} \\ &\le |\mu_{\mathrm{B}} - \mu_{\mathrm{B},\delta}|_{\mathsf{TV}} + |\Phi_\# \mu_{\mathrm{B},\delta} - \mu_{\star,\delta}|_{\mathsf{TV}} + |\mu_{\star,\delta} - \mu_\star|_{\mathsf{TV}}. \end{aligned}$$

Substituting (5.6.14), (5.6.15), and (5.6.16), and choosing η so that $C\eta < \varepsilon/3$, we conclude that $|\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}} < \varepsilon$. □

Proof of Corollary 5.4.5. By the affine invariance of $\mathcal{V}$, the condition $\sigma \in \mathcal{V}$ implies that $\mathcal{V}$ contains all vector fields of the form $x \mapsto \sigma(ax + b)$ for all $a, b \in \mathbb{R}$. Consequently, the linear span of $\mathcal{V}$ contains all one-hidden-layer neural networks with activation σ, i.e., functions of the form $x \mapsto \sum_{i=1}^m c_i \sigma(a_i x + b_i)$. We treat cases (i) and (ii) separately.

Case (i). Assume $\sigma \in \mathscr{C}^1(\mathbb{R})$ is not an affine map. Since $\sigma \in \mathscr{C}^{0,1}(\mathbb{R})$, any polynomial of degree 2 or higher would violate the global Lipschitz bound. Thus, being non-affine guarantees that σ is non-polynomial. By the classical universal approximation theorem [Pin99, Theorem 4.1], the span of such neural networks (and hence $\mathrm{span}(\mathcal{V})$) is dense in $\mathscr{C}^1_{\mathrm{loc}}(\mathbb{R})$.

Therefore, Lemma 5.6.5 applies to $\mathrm{span}(\mathcal{V})$ and yields that for every increasing $\mathscr{C}^1$-diffeomorphism h and compact interval I, there exists $\Psi \in \mathcal{A}_{\mathrm{span}(\mathcal{V})}$ such that $\|h - \Psi\|_{W^{1,1}(I)}$ is arbitrarily small.

Since increasing $\mathscr{C}^1$-diffeomorphisms are dense in $W^{1,1}_{\mathrm{nd,loc}}(\mathbb{R})$, it follows that $\mathcal{A}_{\mathrm{span}(\mathcal{V})}$ satisfies the density condition (5.4.5). Consequently, the control system generated by $\mathrm{span}(\mathcal{V})$ has the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R})$ by Proposition 5.4.3.

To conclude the proof for $\mathcal{V}$ itself, we use Lemma 5.6.6 (Lie–Trotter splitting) to approximate (in the $W^{1,1}(I)$ topology) the flow of any finite linear combination of fields by compositions of flows of individual fields in $\mathcal{V}$. In particular, this implies

$$\overline{\mathcal{A}_{\mathcal{V}}}^{W^{1,1}(I)} = \overline{\mathcal{A}_{\mathrm{span}(\mathcal{V})}}^{W^{1,1}(I)} \supseteq W^{1,1}_{\mathrm{nd}}(I),$$

so (5.4.5) holds for the original family $\mathcal{V}$.

Case (ii). Let $\sigma(s) = s_+$. Let $h \in W^{1,1}_{\mathrm{nd}}(I)$ for a compact interval I. By Lemma 5.6.7, there exists a sequence $\{A_n\}_n$ of functions that are increasing, continuous piecewise affine on I, satisfying $\|A_n - h\|_{W^{1,1}(I)} \to 0$. By Lemma 5.6.8, each function A_n coincides on I with the restriction of some flow map $\Phi_n \in \mathcal{A}_{\mathcal{V}}$. Hence (5.4.5) holds. □

Proofs of Subsection 5.4.2: Multi-dimensional case

Proof of Proposition 5.4.7. We proceed in two steps: first, we show that the forward flow preserves the product structure of measures; second, we establish that the set of product measures is closed under the total variation distance.

Step 1. Any flow map $\Phi \in \mathcal{A}_{\mathcal{V}}$ is generated by $\dot{x}(t) = v(x(t), \theta(t))$ for some chosen $\theta(\cdot)$. Since the dynamics decouple according to (5.4.7), this ODE splits completely into d independent scalar

equations $\dot{x}^{(i)}(t) = v_i(x^{(i)}(t), \theta(t))$ for $i \in [d]$. Thus, the flow map Φ has the product form

$$\Phi(x) = \big(\phi_1(x^{(1)}), \dots, \phi_d(x^{(d)})\big),$$

where each ϕ_i is the one-dimensional flow associated with the scalar equation $\dot{z}(t) = v_i\big(z(t), \theta(t)\big)$.

Let $\mu_{\mathrm{B}} = \mu_{\mathrm{B},1} \otimes \cdots \otimes \mu_{\mathrm{B},d} \in \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$, defined in (5.4.6). For any rectangle $B = \prod_{i=1}^d B_i$ with Borel sets $B_i \subset \mathbb{R}$, we have $\Phi^{-1}(B) = \prod_{i=1}^d \phi_i^{-1}(B_i)$. Thus,

$$(\Phi_{\#}\mu_{\mathrm{B}})(B) = \mu_{\mathrm{B}}\left(\prod_{i=1}^d \phi_i^{-1}(B_i)\right) = \prod_{i=1}^d \mu_{\mathrm{B},i}\left(\phi_i^{-1}(B_i)\right) = \prod_{i=1}^d \big((\phi_i)_{\#}\mu_{\mathrm{B},i}\big)(B_i).$$

Since rectangles form a π-system[1] that generates $\mathcal{B}(\mathbb{R}^d)$, and the collection of sets where two probability measures agree is a λ-system[2], the π–λ Theorem [Dur19, Theorem 2.1.6] ensures that

$$\Phi_{\#}\mu_{\mathrm{B}} = \big(\phi_1\big)_{\#}\mu_{\mathrm{B},1} \ \otimes \ \cdots \ \otimes \ \big(\phi_d\big)_{\#}\mu_{\mathrm{B},d}.$$

Step 2. Let $\nu^n = \nu_1^n \otimes \cdots \otimes \nu_d^n \to \nu$ in TV. For each coordinate projection $\pi^i(x) = x^{(i)}$, the TV-contraction property (Lemma 5.6.3) implies

$$|(\pi^i)_{\#}\nu^n - (\pi^i)_{\#}\nu|_{\mathsf{TV}} \le |\nu^n - \nu|_{\mathsf{TV}} \to 0$$

hence the marginals converge as $\nu_i^n \to \nu_i \coloneqq (\pi^i)_{\#}\nu$ in TV. For any rectangle $B = \prod_{i=1}^d B_i$, we have

$$\nu(B) = \lim_{n\to\infty} \nu^n(B) = \lim_{n\to\infty} \prod_{i=1}^d \nu_i^n(B_i) = \prod_{i=1}^d \nu_i(B_i).$$

This shows that ν and the product measure $\bigotimes_{i=1}^d \nu_i$ agree on the π-system of rectangles. By the same π–λ argument used in Step 1, it follows that $\nu = \bigotimes_{i=1}^d \nu_i$ on $\mathcal{B}(\mathbb{R}^d)$, and thus $\nu \in \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$.

Combining these results, if we choose a base measure $\mu_{\mathrm{B}} \in \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$, we obtain

$$\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}}) \subset \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d) \implies \overline{\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})}^{\mathsf{TV}} \subset \mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d).$$

Consequently, if the target measure $\mu_\star$ does not belong to $\mathcal{P}_{\mathrm{prod}}(\mathbb{R}^d)$ (i.e., its coordinates are not independent), it cannot be in the TV-closure of the attainable set, and the TV-UAP fails. □

Proof of Theorem 5.4.8. Fix $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ and $\varepsilon > 0$. It suffices to construct two maps $\Phi_1, \Phi_2 \in \mathcal{A}_{\mathcal{V}}$ and an intermediate measure ν such that

$$\big|\,(\Phi_1)_{\#}\mu_{\mathrm{B}} - \nu\,\big|_{\mathsf{TV}} < \frac{\varepsilon}{2} \quad \text{and} \quad \big|\,(\Phi_2)_{\#}\mu_\star - \nu\,\big|_{\mathsf{TV}} < \frac{\varepsilon}{2}.$$

Indeed, by (A0) we have $\Phi_2^{-1} \in \mathcal{A}_{\mathcal{V}}$, so that $\Phi \coloneqq \Phi_2^{-1} \circ \Phi_1 \in \mathcal{A}_{\mathcal{V}}$ (by time concatenation of controls), and using the equality case of Lemma 5.6.3 we conclude:

$$\begin{aligned}\big|\,\Phi_{\#}\mu_{\mathrm{B}} - \mu_\star\,\big|_{\mathsf{TV}} &= \big|\,(\Phi_2)_{\#}(\Phi_{\#}\mu_{\mathrm{B}}) - (\Phi_2)_{\#}\mu_\star\,\big|_{\mathsf{TV}} = \big|\,(\Phi_1)_{\#}\mu_{\mathrm{B}} - (\Phi_2)_{\#}\mu_\star\,\big|_{\mathsf{TV}} \\ &\le \big|\,(\Phi_1)_{\#}\mu_{\mathrm{B}} - \nu\,\big|_{\mathsf{TV}} + \big|\,\nu - (\Phi_2)_{\#}\mu_\star\,\big|_{\mathsf{TV}} < \varepsilon.\end{aligned}$$

[1] A collection $\mathcal{P}$ of subsets of Ω is called a π-system if $A, B \in \mathcal{P} \implies A \cap B \in \mathcal{P}$.

[2] A collection $\mathcal{L}$ of subsets of Ω is called a λ-system if: (i) $\Omega \in \mathcal{L}$; (ii) if $A, B \in \mathcal{L}$ and $A \subset B$, then $B \setminus A \in \mathcal{L}$; and (iii) if $A_1 \subset A_2 \subset \dots$ is an increasing sequence of sets in $\mathcal{L}$, then their union $\bigcup_{n=1}^\infty A_n$ is also in $\mathcal{L}$.

We now build Φ_1 and Φ_2, first treating the case $d = 2$ and then indicating the modifications required for $d > 2$. The strategy relies on several steps, illustrated in Figure 5.1. We recall that $\pi^i(x) = x^{(i)}$ denotes the projection onto the i-th coordinate.

(S1) Matching the second marginal. Consider the $x^{(2)}$-marginals

$$\mu_{\mathrm{B},2} \coloneqq (\pi^2)_{\#}\mu_{\mathrm{B}} \qquad \text{and} \qquad \mu_{\star,2} \coloneqq (\pi^2)_{\#}\mu_{\star}.$$

By (A1) with $j = 2$, the family $\mathcal{V}$ induces a 1D control system with reachable set

$$\mathcal{A}_{\mathcal{V}}(\mathcal{V}_2) \coloneqq \Big\{\phi : \mathbb{R} \to \mathbb{R} \;\mid\; \exists\, \Phi \in \mathcal{A}_{\mathcal{V}} \text{ such that } \; \Phi\big(x^{(1)}, x^{(2)}\big) = \big(x^{(1)}, \phi(x^{(2)})\big)\Big\},$$

satisfying (5.4.5). By Proposition 5.4.3, for any $\delta_1 > 0$ there exists $\phi_2 \in \mathcal{A}_{\mathcal{V}}(\mathcal{V}_2)$ such that

$$\big|\,(\phi_2)_{\#}\mu_{\mathrm{B},2} - \mu_{\star,2}\,\big|_{\mathsf{TV}} < \delta_1.$$

Let $\Phi^{\updownarrow} \in \mathcal{A}_{\mathcal{V}}$ be the flow map defined by $\Phi^{\updownarrow}\big(x^{(1)}, x^{(2)}\big) = \big(x^{(1)}, \phi_2(x^{(2)})\big)$, and set $\mu_{\mathrm{B}}^{(1)} \coloneqq (\Phi^{\updownarrow})_{\#}\mu_{\mathrm{B}}$. Then

$$\big|\,(\pi^2)_{\#}\mu_{\mathrm{B}}^{(1)} - \mu_{\star,2}\,\big|_{\mathsf{TV}} = \big|\,(\phi_2)_{\#}\mu_{\mathrm{B},2} - \mu_{\star,2}\,\big|_{\mathsf{TV}} < \delta_1.$$

(S2) Layerwise discretization. We approximate $\mu_{\mathrm{B}}^{(1)}$ and $\mu_\star$ on a large square by measures supported on horizontal layers. We ensure that, on each layer, the two density approximations share exactly the same mass and are independent of $x^{(2)}$.

Fix $R > 0$ and an integer $m \geq 2$; set $K_R \coloneqq [-R, R]^2$ and choose a gap size $\gamma \in \big(0, \frac{2R}{m-1}\big)$. Define the layer height and the vertical partition points by

$$h \coloneqq \frac{2R - (m-1)\gamma}{m}, \qquad a_k \coloneqq -R + (k-1)(h+\gamma), \qquad b_k \coloneqq a_k + h, \qquad \text{for } k \in [m].$$

We define the layers and gaps respectively as

$$L \coloneqq \bigcup_{k=1}^{m} L_k \qquad \text{and} \qquad G \coloneqq \bigcup_{k=1}^{m-1} G_k,$$

where

$$L_k \coloneqq [-R, R] \times (a_k, b_k) \qquad \text{and} \qquad G_k \coloneqq [-R, R] \times (b_k, a_{k+1}),$$

so that $K_R = L \sqcup G$. Write $\mu_{\mathrm{B}}^{(1)} = \rho_{\mathrm{B}}^{(1)}\,\mathcal{L}^2$ and $\mu_\star = \rho_\star\,\mathcal{L}^2$. For each $k \in [m]$, set the common target mass as

$$m_k \coloneqq \mu_{\mathrm{B}}^{(1)}(L_k) \wedge \mu_\star(L_k).$$

We define the finite measures $\mu_{\mathrm{B}}^{(2)}$ and $\mu_\star^{(2)}$ via their densities

$$\rho_{\mathrm{B}}^{(2)}(x) \coloneqq \begin{cases} \dfrac{m_k}{\mu_{\mathrm{B}}^{(1)}(L_k)}\,\dfrac{1}{h}\displaystyle\int_{a_k}^{b_k} \rho_{\mathrm{B}}^{(1)}(x^{(1)}, s)\,\mathrm{d}s, & \text{if } x \in L_k, \\ 0, & \text{if } x \notin L, \end{cases}$$

$$\rho_\star^{(2)}(x) \coloneqq \begin{cases} \dfrac{m_k}{\mu_\star(L_k)}\,\dfrac{1}{h}\displaystyle\int_{a_k}^{b_k} \rho_\star(x^{(1)}, s)\,\mathrm{d}s, & \text{if } x \in L_k, \\ 0, & \text{if } x \notin L, \end{cases}$$

with the convention $0/0 = 0$. By construction, $\int_{L_k} \rho_{\mathrm{B}}^{(2)} = m_k = \int_{L_k} \rho_\star^{(2)}$ for all k, and these densities depend only on $x^{(1)}$ within each layer.

Claim. Provided δ_1 (from **(S1)**) is sufficiently small, for each $\delta_2 > 0$ there exist R and a gap size γ small enough, and an integer m large enough, such that

$$\left| \mu_{\mathrm{B}}^{(2)} - \mu_{\mathrm{B}}^{(1)} \right|_{\mathsf{TV}} < \delta_2 \qquad \text{and} \qquad \left| \mu_\star^{(2)} - \mu_\star \right|_{\mathsf{TV}} < \delta_2.$$

Proof of Claim. First, fix $R > 0$ large enough so that

$$\mu_{\mathrm{B}}^{(1)}(\mathbb{R}^2 \setminus K_R) + \mu_\star(\mathbb{R}^2 \setminus K_R) < \frac{\delta_2}{16}.$$

Next, for any $m \geq 2$, we can choose γ small enough so that the mass in the gaps is negligible:

$$\mu_{\mathrm{B}}^{(1)}(G) + \mu_\star(G) < \frac{\delta_2}{16}.$$

Consider the auxiliary layerwise averages defined for $\rho \in \{\rho_{\mathrm{B}}^{(1)}, \rho_\star\}$ by

$$\hat{\rho}(x) \coloneqq \begin{cases} \dfrac{1}{h} \displaystyle\int_{a_k}^{b_k} \rho\big(x^{(1)}, s\big)\, \mathrm{d}s, & \text{if } x \in L_k, \\ 0, & \text{if } x \notin L. \end{cases}$$

Since the original densities are in $L^1(K_R)$, these averages—which are piecewise constant with respect to $x^{(2)}$—converge strongly to ρ in $L^1(L)$ as the mesh size $h \to 0$ (which is ensured by taking m large). Thus, for m sufficiently large, the associated measures $\hat{\mu}_{\mathrm{B}}$ and $\hat{\mu}_\star$ satisfy

$$\left|\mu_{\mathrm{B}}^{(1)} - \hat{\mu}_{\mathrm{B}}\right|_{\mathsf{TV}} \leq \frac{1}{2} \|\rho_{\mathrm{B}}^{(1)} - \hat{\rho}_{\mathrm{B}}\|_{L^1(L)} + \mu_{\mathrm{B}}^{(1)}(G) + \mu_{\mathrm{B}}^{(1)}(\mathbb{R}^2 \setminus K_R) < \frac{7\delta_2}{16}, \tag{5.6.17}$$

and the same bound holds for $\mu_\star$.

It remains to control the error introduced by the mass renormalization factors. On each L_k, we have $\rho_{\mathrm{B}}^{(2)} = \frac{m_k}{\mu_{\mathrm{B}}^{(1)}(L_k)}\, \hat{\rho}_{\mathrm{B}}$. Thus,

$$\int_{L_k} |\rho_{\mathrm{B}}^{(2)} - \hat{\rho}_{\mathrm{B}}| = \left| \frac{m_k}{\mu_{\mathrm{B}}^{(1)}(L_k)} - 1 \right| \int_{L_k} \hat{\rho}_{\mathrm{B}} = |m_k - \mu_{\mathrm{B}}^{(1)}(L_k)|.$$

Summing over k and recalling that $m_k = \mu_{\mathrm{B}}^{(1)}(L_k) \wedge \mu_\star(L_k)$, we obtain

$$\|\rho_{\mathrm{B}}^{(2)} - \hat{\rho}_{\mathrm{B}}\|_{L^1(L)} = \sum_{k=1}^m \big(\mu_{\mathrm{B}}^{(1)}(L_k) - \mu_\star(L_k)\big)_+ \leq \sum_{k=1}^m |\mu_{\mathrm{B}}^{(1)}(L_k) - \mu_\star(L_k)|.$$

A symmetric argument applies to $\rho_\star^{(2)}$. Since all (a_k, b_k) are disjoint, we have

$$\begin{aligned} \sum_{k=1}^m |\mu_{\mathrm{B}}^{(1)}(L_k) - \mu_\star(L_k)| \leq \sum_{k=1}^m \Big| (\pi^2)_{\#}\mu_{\mathrm{B}}^{(1)}((a_k, b_k)) - (\pi^2)_{\#}\mu_\star((a_k, b_k)) \Big| \\ + \mu_{\mathrm{B}}^{(1)}(\mathbb{R}^2 \setminus K_R) + \mu_\star(\mathbb{R}^2 \setminus K_R). \end{aligned}$$

Using the result from **(S1)** and the choice of R,

$$\sum_{k=1}^{m} |\mu_{\mathrm{B}}^{(1)}(L_k) - \mu_\star(L_k)| \le 2\big|(\pi^2)_\#\mu_{\mathrm{B}}^{(1)} - (\pi^2)_\#\mu_\star\big|_{\mathsf{TV}} + \frac{\delta_2}{16} < 2\delta_1 + \frac{\delta_2}{16}.$$

Therefore,

$$\big|\mu_{\mathrm{B}}^{(2)} - \hat{\mu}_{\mathrm{B}}\big|_{\mathsf{TV}} = \frac{1}{2}\|\rho_{\mathrm{B}}^{(2)} - \hat{\rho}_{\mathrm{B}}\|_{L^1(L)} < \delta_1 + \frac{\delta_2}{32}.$$

Finally, combining this with (5.6.17) via the triangle inequality,

$$\big|\mu_{\mathrm{B}}^{(2)} - \mu_{\mathrm{B}}^{(1)}\big|_{\mathsf{TV}} \le \big|\mu_{\mathrm{B}}^{(2)} - \hat{\mu}_{\mathrm{B}}\big|_{\mathsf{TV}} + \big|\hat{\mu}_{\mathrm{B}} - \mu_{\mathrm{B}}^{(1)}\big|_{\mathsf{TV}} < \delta_1 + \frac{\delta_2}{2}.$$

Analogously, $\big|\mu_\star^{(2)} - \mu_\star\big|_{\mathsf{TV}} < \delta_1 + \delta_2/2$. Choosing $\delta_1 \le \delta_2/2$ completes the proof of the claim. □

(S3) Shearing. We now separate the layers horizontally to ensure that their projections onto the first coordinate are pairwise disjoint.

Fix a separation $S > 0$. Define the shift values $S_k := (k-1)(2R+S)$ for $k \in [m]$. We introduce a continuous piecewise affine profile $\sigma : \mathbb{R} \to \mathbb{R}$ by

$$\sigma(x^{(2)}) := \begin{cases} 0, & \text{if } x^{(2)} < -R, \\ S_k, & \text{if } x^{(2)} \in [a_k, b_k], \\ S_k + \dfrac{S_{k+1} - S_k}{a_{k+1} - b_k}\big(x^{(2)} - b_k\big), & \text{if } x^{(2)} \in (b_k, a_{k+1}), \\ S_m, & \text{if } x^{(2)} > R. \end{cases}$$

The resulting function σ is piecewise affine and nondecreasing. Define the target sheared measures by

$$\mu_{\mathrm{B}}^{(3)} := \big(x^{(1)} + \sigma(x^{(2)}),\, x^{(2)}\big)_\#\mu_{\mathrm{B}}^{(2)}, \qquad \mu_\star^{(3)} := \big(x^{(1)} + \sigma(x^{(2)}),\, x^{(2)}\big)_\#\mu_\star^{(2)}.$$

Since $\sigma \equiv S_k$ on each layer L_k, the restriction of $\mu_{\mathrm{B}}^{(3)}$ (resp. $\mu_\star^{(3)}$) to the image of L_k corresponds to a rigid translation of the layer by S_k in the $x^{(1)}$-direction. Consequently, the sheared layers $\tilde{L}_k := ([-R, R] + S_k) \times (a_k, b_k)$ have pairwise disjoint projections onto the horizontal axis.

Claim. For every $\delta_3 > 0$ there exists $\Phi^{\mathrm{sh}}\big(x^{(1)}, x^{(2)}\big) = \big(x^{(1)} + \psi(x^{(2)}),\, x^{(2)}\big) \in \mathcal{A}_{\mathcal{V}}$ such that

$$\big|\,(\Phi^{\mathrm{sh}})_\#\mu_{\mathrm{B}}^{(2)} - \mu_{\mathrm{B}}^{(3)}\,\big|_{\mathsf{TV}} < \delta_3 \qquad \text{and} \qquad \big|\,(\Phi^{\mathrm{sh}})_\#\mu_\star^{(2)} - \mu_\star^{(3)}\,\big|_{\mathsf{TV}} < \delta_3.$$

Proof of Claim. We apply hypothesis (A2) with $J = [-R, R]$ and coordinates $i = 2$, $j = 1$. There exists a sequence of attainable maps

$$\Phi_n^{\mathrm{sh}}\big(x^{(1)}, x^{(2)}\big) = \big(x^{(1)} + \psi_n(x^{(2)}),\, x^{(2)}\big) \in \mathcal{A}_{\mathcal{V}}, \qquad n \in \mathbb{N},$$

such that $\|\psi_n - \sigma\|_{L^1([-R,R])} \to 0$ as $n \to \infty$.

We verify the convergence for $\mu_{\mathrm{B}}^{(2)}$ (the argument for $\mu_\star^{(2)}$ is identical). Recall from **(S2)** that its density depends only on $x^{(1)}$ within each layer L_k. Let us denote this profile by $\bar{\rho}_k(x^{(1)}) := \rho_{\mathrm{B}}^{(2)}\big(x^{(1)}, x^{(2)}\big)$ for $x \in L_k$. Since the density vanishes on the gaps G, the TV-distance reduces to

$$\big|\,(\Phi_n^{\mathrm{sh}})_\#\mu_{\mathrm{B}}^{(2)} - \mu_{\mathrm{B}}^{(3)}\,\big|_{\mathsf{TV}} = \frac{1}{2}\int_{\mathbb{R}^2} \Big|\rho_{\mathrm{B}}^{(2)}\big(x^{(1)} - \psi_n(x^{(2)}), x^{(2)}\big) - \rho_{\mathrm{B}}^{(2)}\big(x^{(1)} - \sigma(x^{(2)}), x^{(2)}\big)\Big|\,\mathrm{d}x$$

$$= \frac{1}{2}\sum_{k=1}^{m}\int_{a_k}^{b_k}\int_{\mathbb{R}} \left|\bar\rho_k\big(x^{(1)}-\psi_n(x^{(2)})\big)-\bar\rho_k\big(x^{(1)}-S_k\big)\right| \mathrm{d}x^{(1)}\,\mathrm{d}x^{(2)}.$$

For each k, define

$$g_k(t) \coloneqq \|\bar\rho_k(\cdot - t) - \bar\rho_k(\cdot - S_k)\|_{L^1(\mathbb{R})}.$$

The map $t \mapsto \bar\rho_k(\cdot - t)$ is continuous from $\mathbb{R}$ to $L^1(\mathbb{R})$; hence g_k is continuous and bounded by $2\|\bar\rho_k\|_{L^1}$. Since $\sigma(x^{(2)}) \equiv S_k$ for $x^{(2)} \in (a_k, b_k)$, the condition $\|\psi_n - \sigma\|_{L^1} \to 0$ implies that, up to extracting a subsequence (not relabeled), $\psi_n(x^{(2)}) \to S_k$ for almost every $x^{(2)} \in (a_k, b_k)$. Therefore,

$$g_k(\psi_n(x^{(2)})) \longrightarrow g_k(S_k) = 0 \qquad \text{for a.e. } x^{(2)} \in (a_k, b_k).$$

By the dominated convergence theorem,

$$\int_{a_k}^{b_k} g_k\big(\psi_n(x^{(2)})\big)\,\mathrm{d}x^{(2)} \longrightarrow 0 \qquad \text{as } n \to \infty.$$

Summing over k, we conclude that $\big|\,(\Phi_n^{\mathrm{sh}})_{\#}\mu_{\mathrm{B}}^{(2)} - \mu_{\mathrm{B}}^{(3)}\,\big|_{\mathsf{TV}} \to 0$. The same sequence Φ_n^{sh} drives the error to zero for $\mu_\star^{(2)}$. Thus, for n large enough, the claim holds.

(S4) Matching the first marginal. Recall from **(S3)** that $\mu_{\mathrm{B}}^{(3)}$ and $\mu_\star^{(3)}$ are supported on the strips $\tilde L_k \coloneqq I_k \times (a_k, b_k)$, where $I_k \coloneqq [-R + S_k, R + S_k]$ are pairwise disjoint thanks to the separation parameter S. On each strip, the densities are independent of $x^{(2)}$ and share the same mass m_k.

Fix a tolerance $\delta_4 > 0$ (to be chosen later). We will now define a transport map $\tau : \mathbb{R} \to \mathbb{R}$. Let $\mu_{\mathrm{B},1,k}$ and $\mu_{\star,1,k}$ be the $x^{(1)}$-marginals of the restrictions to $\tilde L_k$, for each k. Fix $\delta > 0$. By standard approximation, there exist densities $\rho_{\mathrm{B},k}^\delta, \rho_{\star,k}^\delta \in \mathscr{C}^{1,1}(I_k)$ with mass m_k such that

$$0 < c_\delta \le \rho_{\mathrm{B},k}^\delta(x),\ \rho_{\star,k}^\delta(x) \le C_\delta \qquad \text{for all } x \in I_k,$$

and the associated measures $\mu_{\mathrm{B},1,k}^\delta$ and $\mu_{\star,1,k}^\delta$ satisfy

$$\big|\mu_{\mathrm{B},1,k}^\delta - \mu_{\mathrm{B},1,k}\big|_{\mathsf{TV}} + \big|\mu_{\star,1,k}^\delta - \mu_{\star,1,k}\big|_{\mathsf{TV}} < \frac{\delta}{2m}.$$

If $m_k = 0$, we set $\mu_{\mathrm{B},1,k}^\delta = \mu_{\star,1,k}^\delta = 0$ and define $\mathsf{M}_{k,\delta} = \mathsf{Id}$ on I_k.

For $m_k > 0$, let $\mathsf{M}_{k,\delta} : I_k \to I_k$ be the monotone rearrangement from $\mu_{\mathrm{B},1,k}^\delta$ to $\mu_{\star,1,k}^\delta$. Due to the density bounds, $\mathsf{M}_{k,\delta}$ is a bi-Lipschitz increasing $\mathscr{C}^{1,1}$-diffeomorphism.

Define $\tau : \mathbb{R} \to \mathbb{R}$ by setting $\tau = \mathsf{M}_{k,\delta}$ on each I_k, and extend it on the gaps and tails by strictly increasing $\mathscr{C}^{1,1}$–pieces that match the values and derivatives of $\mathsf{M}_{k,\delta}$ at the endpoints. Then τ is an increasing bi-Lipschitz $\mathscr{C}^{1,1}$-diffeomorphism. Let

$$\mathsf{T}_\tau\big(x^{(1)}, x^{(2)}\big) \coloneqq \big(\tau(x^{(1)}), x^{(2)}\big).$$

By the product structure of the measures on the strips, for δ small we have

$$\big|\,(\mathsf{T}_\tau)_{\#}\mu_{\mathrm{B}}^{(3)} - \mu_\star^{(3)}\,\big|_{\mathsf{TV}} < \delta.$$

Indeed, write $\mu_{\mathrm{B}}^{(3)} = \sum_{k=1}^m \mu_{\mathrm{B},1,k} \otimes \lambda_k$ and $\mu_\star^{(3)} = \sum_{k=1}^m \mu_{\star,1,k} \otimes \lambda_k$, where $\lambda_k \coloneqq \frac{1}{h}\,\mathbf{1}_{(a_k,b_k)}\,\mathcal{L}^1$. Since the strips are disjoint,

$$\big|\,(\mathsf{T}_\tau)_{\#}\mu_{\mathrm{B}}^{(3)} - \mu_\star^{(3)}\,\big|_{\mathsf{TV}} = \sum_{k=1}^m \big|\,\tau_{\#}\mu_{\mathrm{B},1,k} - \mu_{\star,1,k}\,\big|_{\mathsf{TV}}.$$

Moreover, $(\tau|_{I_k})_{\#}\mu^{\delta}_{\mathrm{B},1,k} = \mu^{\delta}_{\star,1,k}$ by construction, hence by the triangle inequality,

$$\big|\,\tau_{\#}\mu_{\mathrm{B},1,k} - \mu_{\star,1,k}\,\big|_{\mathsf{TV}} \le \big|\,\mu_{\mathrm{B},1,k} - \mu^{\delta}_{\mathrm{B},1,k}\,\big|_{\mathsf{TV}} + \big|\,\mu^{\delta}_{\star,1,k} - \mu_{\star,1,k}\,\big|_{\mathsf{TV}}.$$

Summing over k yields the claim.

Now we approximate τ using (A1). Choose R_0 so large that all supports of $\mu^{\delta}_{\mathrm{B},1,k}$ and $\tau_{\#}\mu^{\delta}_{\mathrm{B},1,k}$ lie in $I_{R_0-1} \coloneqq [-R_0+1, R_0-1]$. Let

$$J_{R_0} \coloneqq \tau^{-1}(I_{R_0+1}).$$

Since τ^{-1} is Lipschitz, for every $x < y$ we have

$$y - x = |\tau^{-1}(\tau(y)) - \tau^{-1}(\tau(x))| \le \mathrm{Lip}(\tau^{-1})\,(\tau(y) - \tau(x)).$$

Hence $\frac{\tau(y)-\tau(x)}{y-x} \ge 1/\mathrm{Lip}(\tau^{-1})$. Passing to the limit as $y \to x$, and using that $\tau \in \mathscr{C}^1$, we get

$$\tau'(x) \ge \frac{1}{L} > 0 \qquad \text{for all } x \in \mathbb{R}.$$

Since τ' is continuous and J_{R_0} is compact, it follows that

$$m_{R_0} \coloneqq \min_{J_{R_0}} \tau' > 0.$$

By (A1) with $j = 1$, for any $\eta > 0$ there exists $\Phi^{\leftrightarrow}\big(x^{(1)}, x^{(2)}\big) = \big(\phi(x^{(1)}), x^{(2)}\big) \in \mathcal{A}_{\mathcal{V}}$ such that

$$\|\phi - \tau\|_{W^{1,1}(J_{R_0})} < \eta, \qquad \|\phi - \tau\|_{L^{\infty}(J_{R_0})} \le C_{R_0}\eta,$$

where C_{R_0} is the embedding constant $W^{1,1}(J_{R_0}) \hookrightarrow \mathscr{C}^0(\overline{J_{R_0}})$.

Applying Lemma 5.6.10 (on each I_k with $m_k > 0$) to the probability $m_k^{-1}\mu^{\delta}_{\mathrm{B},1,k}$, with $h = \tau$ and $g = \phi$, and noting that the tail terms vanish by the choice of R_0, we obtain constants C_k such that

$$\big|\,\phi_{\#}\mu^{\delta}_{\mathrm{B},1,k} - \tau_{\#}\mu^{\delta}_{\mathrm{B},1,k}\,\big|_{\mathsf{TV}} \le m_k\, C_k\, \|\phi - \tau\|_{W^{1,1}(J_{R_0})}.$$

Let $C_{\max} \coloneqq \max\{C_k \mid\ m_k > 0\}$. Choose η so that

$$\eta \le 1 \wedge C_{R_0}^{-1}, \qquad C_{\max}\eta < \frac{\delta_4}{4}, \qquad C_{R_0}\eta < \frac{S}{3}.$$

Then $\phi(I_k)$ remain pairwise disjoint. Using disjointness and summing over k,

$$\big|\,(\Phi^{\leftrightarrow})_{\#}\mu^{\delta}_{\mathrm{B}} - (\mathsf{T}_{\tau})_{\#}\mu^{\delta}_{\mathrm{B}}\,\big|_{\mathsf{TV}} = \sum_{k=1}^{m} \big|\,\phi_{\#}\mu^{\delta}_{\mathrm{B},1,k} - \tau_{\#}\mu^{\delta}_{\mathrm{B},1,k}\,\big|_{\mathsf{TV}} \le \Big(\sum_{k=1}^{m} m_k\Big) C_{\max}\eta < \frac{\delta_4}{4},$$

where $\mu^{\delta}_{\mathrm{B}} \coloneqq \sum_{k=1}^{m} \mu^{\delta}_{\mathrm{B},1,k} \otimes \lambda_k$.

Combining $\big|\,(\mathsf{T}_{\tau})_{\#}\mu^{(3)}_{\mathrm{B}} - \mu^{(3)}_{\star}\,\big|_{\mathsf{TV}} < \delta$, the TV-approximation of $\mu^{(3)}_{\mathrm{B}}$ by $\mu^{\delta}_{\mathrm{B}}$ (the sum of the bounds $< \delta/2$), and the estimate above, and choosing δ small enough, yields

$$\big|\,(\Phi^{\leftrightarrow})_{\#}\mu^{(3)}_{\mathrm{B}} - \mu^{(3)}_{\star}\,\big|_{\mathsf{TV}} < \delta_4.$$

Conclusion. Based on the construction in (S1)–(S4), we define

$$\Phi_1 \coloneqq \Phi^{\leftrightarrow} \circ \Phi^{\mathrm{sh}} \circ \Phi^{\updownarrow} \in \mathcal{A}_{\mathcal{V}}, \qquad \Phi_2 \coloneqq \Phi^{\mathrm{sh}} \in \mathcal{A}_{\mathcal{V}}.$$

Set the intermediate measure $\nu := \mu_\star^{(3)}$. Summing the errors from each step (and recalling from (S2) that the discretization error involves $\delta_1 + \delta_2/2$), we obtain

$$\left|(\Phi_2)_\#\mu_\star - \nu\right|_{\mathsf{TV}} \le \delta_1 + \delta_2 + \delta_3, \qquad \left|(\Phi_1)_\#\mu_{\mathrm{B}} - \nu\right|_{\mathsf{TV}} \le \delta_1 + \delta_2 + \delta_3 + \delta_4.$$

We first choose $\delta_2, \delta_3, \delta_4$ small enough, and then δ_1 sufficiently small (specifically $\delta_1 \le \delta_2/4$ as required in (S2)), so that the sum of all errors is less than ε. The reduction at the beginning of the proof then yields $\Phi = \Phi_2^{-1} \circ \Phi_1 \in \mathcal{A}_\mathcal{V}$ satisfying

$$\left|\Phi_\#\mu_{\mathrm{B}} - \mu_\star\right|_{\mathsf{TV}} < \varepsilon.$$

This completes the proof for $d = 2$.

Extension to arbitrary dimension.

We extend Theorem 5.4.8 to all $d \ge 2$ by iteratively eliminating the dependence of the densities on the coordinates $x^{(d)}, x^{(d-1)}, \dots, x^{(2)}$.

Fix $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ and $\varepsilon > 0$. As in the 2D case, we truncate measures to a large cube K_R and treat the mass outside as an additive error. It suffices to construct $\Phi_1, \Phi_2 \in \mathcal{A}_\mathcal{V}$ pushing the truncated μ_{B} and $\mu_\star$ to a common intermediate measure ν within tolerance $\varepsilon/2$. We describe the procedure for μ_{B} (the argument for $\mu_\star$ is identical).

We proceed by backward induction on the coordinate index j from d down to 2. Assume that at the start of step j the measures are supported on a finite union of disjoint blocks B_α whose projections onto the $x^{(j)}$-axis are pairwise disjoint, that the densities on each B_α depend only on the coordinates $(x^{(1)}, \dots, x^{(j)})$, and that μ_{B} and $\mu_\star$ assign the same total mass to each B_α. (For the base case $j = d$, this holds with a single block $B_1 = K_R$.)

Assuming the hypothesis holds at level j, we perform the following operations to reach level $j - 1$:

(S1') Marginal matching. For each block B_α, let μ_{B}^α and $\mu_\star^\alpha$ be the restrictions of the measures. We consider the monotone rearrangement between their $x^{(j)}$-marginals. Since the blocks have pairwise disjoint projections onto the $x^{(j)}$-axis, we can glue these local maps into a single nondecreasing function $\tau_j : \mathbb{R} \to \mathbb{R}$ (interpolating monotonically on the gaps). Invoking (A1) for the j-th coordinate, there exists

$$\left(x^{(1)}, \dots, \phi_j(x^{(j)}), \dots, x^{(d)}\right) \in \mathcal{A}_\mathcal{V}$$

which approximates τ_j in $W^{1,1}$ on the union of the projections. By Lemma 5.6.10, this matches the $x^{(j)}$-marginals on all blocks up to a small TV error.

(S2') Layering. We partition each block B_α into sufficiently thin slices along the $x^{(j)}$-direction. On each slice, we replace the density with its average with respect to $x^{(j)}$ (renormalizing to maintain the mass). This produces a new measure whose density is independent of $x^{(j)}$ on each slice. The error is controlled by the mesh size.

(S3') Shearing. Choose shifts S_k sufficiently large and distinct for each sub-layer and build a nondecreasing, piecewise affine σ_j with $\sigma_j \equiv S_k$ on the k-th sub-layer. By applying (A2) with $(i, j) = (j, j-1)$, we obtain a shear flow

$$\left(x^{(1)}, \dots, x^{(j-2)}, x^{(j-1)} + \psi_j(x^{(j)}), x^{(j)}, \dots, x^{(d)}\right) \in \mathcal{A}_\mathcal{V},$$

where $\psi_j \approx \sigma_j$ in L^1. Since the densities are now independent of $x^{(j)}$, the transport error is small (analogous to (S3) in 2D). The image of the slices under this shear will have pairwise disjoint projections onto the $x^{(j-1)}$-axis.

The resulting measure is supported on a new collection of blocks (the sheared slices) satisfying the inductive hypothesis for $j-1$. We iterate this process for $j = d, d-1, \dots, 2$.

Conclusion. After step $j = 2$, the measures are supported on a collection of *tubes* parallel to the $x^{(1)}$-axis with disjoint $x^{(1)}$-projections. On each tube, the densities depend only on $x^{(1)}$ and carry identical mass. This is exactly the situation of (S4) in the 2D proof. We apply a final flow acting on $x^{(1)}$ (using (A1) with $j = 1$) to match the densities. Summing all errors via Lemma 5.6.3 completes the proof.

Finally, we prove Corollary 5.4.9 to verify the realization by neural networks.

Proof of Corollary 5.4.9. It suffices to verify assumptions (A0), (A1), and (A2) of Theorem 5.4.8.

(A0). By affine invariance, if $v \in \mathcal{V}$, then $-v \in \mathcal{V}$. Since $\Phi_t^{-v} = (\Phi_t^v)^{-1}$, the reachable set is closed under inversion.

(A1). Fix $j \in [d]$ and a compact interval $I \subset \mathbb{R}$. By the hypothesis $x \mapsto \sigma(ax^{(i)} + b)\, e_j \in \mathcal{V}$ and affine invariance, the induced one-dimensional family of vector fields acting on the j-th coordinate contains the control family of Corollary 5.4.5. Hence Corollary 5.4.5 yields, for every $\tau \in W^{1,1}_{\mathrm{nd,loc}}(\mathbb{R})$ and every $\varepsilon > 0$, an attainable map

$$\Phi(x) = (x^{(1)}, \dots, x^{(j-1)}, \phi(x^{(j)}), x^{(j+1)}, \dots, x^{(d)})$$

with $\|\phi - \tau\|_{W^{1,1}(I)} < \varepsilon$, as required.

(A2). Fix $i \neq j$ and a compact interval $J \subset \mathbb{R}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous, nondecreasing and piecewise affine. By the standard universal approximation theorem for single-hidden-layer networks, for every $\varepsilon > 0$ there exist $m \in \mathbb{N}$ and coefficients $(a_k, b_k, c_k)_{k=1}^m$ such that

$$\psi(s) := \sum_{k=1}^{m} c_k\, \sigma(a_k s + b_k), \qquad s \in \mathbb{R},$$

satisfies $\|\psi - \sigma\|_{L^1(J)} < \varepsilon$. By assumption, each field $v_k(x) := c_k\, \sigma(a_k x^{(i)} + b_k)\, e_j$ belongs to $\mathcal{V}$. Its time-one flow is the shear

$$\Phi_1^{v_k}(x) = x + c_k\, \sigma(a_k x^{(i)} + b_k)\, e_j.$$

Composing these flows yields

$$\Phi_1^{v_m} \circ \cdots \circ \Phi_1^{v_1}(x) = x + \psi(x^{(i)})\, e_j,$$

which has the form required in assumption (A2) and satisfies $\|\psi - \sigma\|_{L^1(J)} < \varepsilon$. Therefore (A2) holds. Applying Theorem 5.4.8 completes the proof. □

Proofs of Subsection 5.4.3: Relationship with point interpolation

Proof of Proposition 5.4.13. The family $\mathcal{V}$ is the same as in Proposition 5.3.5, so the associated system possesses the UIP by Remark 3.3.2 and Theorem 3.3.1.

To prove that the TV-UAP fails, we will in fact establish a stronger obstruction: the approximation fails for any pair of absolutely continuous measures $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with densities $\rho_{\mathrm{B}}, \rho_\star$ satisfying $\|\rho_{\mathrm{B}}\|_{L^2} \neq \|\rho_\star\|_{L^2}$.

Since every $v \in \mathcal{V}$ is Lipschitz and satisfies $\operatorname{div} v = 0$ almost everywhere, the pushforward $\Phi_{\#}\mu_{\mathrm{B}}$ by any flow map $\Phi \in \mathcal{A}_{\mathcal{V}}$ has density $\rho_\Phi := \rho_{\mathrm{B}} \circ \Phi^{-1}$. Consequently, the L^p norms are preserved: $\|\rho_\Phi\|_{L^p} = \|\rho_{\mathrm{B}}\|_{L^p}$ for every $p \in [1, \infty]$ (recall (1.2.4)).

Assume first $\rho_{\mathrm{B}}, \rho_\star \in L^\infty(\mathbb{R}^d)$. Fix $\Phi \in \mathcal{A}_\mathcal{V}$. By the triangle inequality and the standard interpolation between L^1 and L^∞,

$$\begin{aligned}\big|\|\rho_\Phi\|_{L^2} - \|\rho_\star\|_{L^2}\big| \le \|\rho_\Phi - \rho_\star\|_{L^2} &\le \sqrt{\|\rho_\Phi - \rho_\star\|_{L^\infty}\,\|\rho_\Phi - \rho_\star\|_{L^1}} \\ &\le \sqrt{\big(\|\rho_\Phi\|_{L^\infty} + \|\rho_\star\|_{L^\infty}\big)\cdot 2\,|\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}}}.\end{aligned} \tag{5.6.18}$$

Hence, squaring and rearranging terms gives

$$|\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}} \ge \frac{\big|\|\rho_\Phi\|_{L^2} - \|\rho_\star\|_{L^2}\big|^2}{2\big(\|\rho_\Phi\|_{L^\infty} + \|\rho_\star\|_{L^\infty}\big)} = \frac{\big|\|\rho_{\mathrm{B}}\|_{L^2} - \|\rho_\star\|_{L^2}\big|^2}{2\big(\|\rho_{\mathrm{B}}\|_{L^\infty} + \|\rho_\star\|_{L^\infty}\big)}.$$

If $\|\rho_\star\|_{L^2} \neq \|\rho_{\mathrm{B}}\|_{L^2}$, this creates a strictly positive TV-gap that is uniform over all $\Phi \in \mathcal{A}_\mathcal{V}$, meaning the TV-UAP fails.

If instead $\rho_{\mathrm{B}}, \rho_\star \in L^r(\mathbb{R}^d)$ for some $r \in (2, \infty)$, we use the L^1–L^r interpolation bound in (5.6.18):

$$\|\rho_\Phi - \rho_\star\|_{L^2} \le \|\rho_\Phi - \rho_\star\|_{L^r}^{\theta} \|\rho_\Phi - \rho_\star\|_{L^1}^{1-\theta}, \qquad \theta = \frac{r}{2(r-1)} \in \left(\frac{1}{2}, 1\right).$$

Substituting the TV norm and matching terms as before yields

$$|\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}} \ge \frac{1}{2} \frac{\big|\|\rho_{\mathrm{B}}\|_{L^2} - \|\rho_\star\|_{L^2}\big|^{\frac{2(r-1)}{r-2}}}{\big(\|\rho_{\mathrm{B}}\|_{L^r} + \|\rho_\star\|_{L^r}\big)^{\frac{r}{r-2}}}.$$

Therefore, any pair $(\mu_{\mathrm{B}}, \mu_\star) \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \times \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ whose densities satisfy $\|\rho_{\mathrm{B}}\|_{L^2} \neq \|\rho_\star\|_{L^2}$ is separated by a uniform positive TV-gap under the action of $\mathcal{A}_\mathcal{V}$. In particular, taking the same $\mu_{\mathrm{B}}, \mu_\star$ as in the proof of Proposition 5.3.5, which are TV-compatible, we conclude that the system does not possess the TV-UAP. □

We now proceed to the proof of Proposition 5.4.14. To this end, we rely on the Dacorogna–Moser theorem, which we state here for completeness following [DM90, Theorem 1].

Theorem 5.6.11 (Dacorogna–Moser). *Let $k \ge 0$ be an integer, $0 < \alpha < 1$, and let $\Omega \subset \mathbb{R}^d$ be a bounded, connected open set with a $\mathscr{C}^{k+3,\alpha}$ boundary $\partial\Omega$. Let $\mu_1, \mu_2 \in \mathcal{P}_{\mathrm{ac}}(\overline{\Omega})$ be measures with densities $\rho_1, \rho_2 \in \mathscr{C}^{k,\alpha}(\overline{\Omega})$ such that $\rho_1, \rho_2 > 0$ on $\overline{\Omega}$.*

Then, there exists a diffeomorphism $\Phi : \overline{\Omega} \to \overline{\Omega}$ such that $\Phi, \Phi^{-1} \in \mathscr{C}^{k+1,\alpha}(\overline{\Omega}; \mathbb{R}^d)$, satisfying

$$\begin{cases} \rho_1(x) = \rho_2(\Phi(x)) \det \nabla\Phi(x), & x \in \Omega, \\ \Phi(x) = x, & x \in \partial\Omega, \end{cases} \tag{5.6.19}$$

or, equivalently, $\Phi_\# \mu_1 = \mu_2$.

Proof of Proposition 5.4.14. We prove each of the two statements separately:

(No UIP). By the choice of $\mathcal{V}$, every $\Phi \in \mathcal{A}_\mathcal{V}$ fixes the origin. Taking $N = 1$ with $x_1 = 0$ and $y_1 \neq 0$ shows that the UIP fails.

(TV-UAP). Fix $\varepsilon > 0$ and let $\mu_{\mathrm{B}}, \mu_\star \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ with densities ρ_{B} and $\rho_\star$. We will build $\Phi \in \mathcal{A}_\mathcal{V}$ such that $|\Phi_\# \mu_{\mathrm{B}} - \mu_\star|_{\mathsf{TV}} < \varepsilon$ using Theorem 5.6.11.

Choose an annulus $A \coloneqq \{x \in \mathbb{R}^d \mid \ r < \|x\| < R\}$ with $r > 0$ small and $R > 0$ large so that

$$\|\rho_i - \rho_i \mathbf{1}_A\|_{L^1(\mathbb{R}^d)} < \frac{\varepsilon}{10}, \qquad i \in \{\mathrm{B}, \star\}.$$

Set $\mu_i^A \coloneqq \rho_i \mathbf{1}_A \, \mathcal{L}^d$ and $m_i \coloneqq \mu_i^A(A) \in (0,1]$. Then $m_i \in (0,1]$ and

$$|m_\mathrm{B} - m_\star| \le \|\rho_\mathrm{B} - \rho_\mathrm{B}\mathbf{1}_A\|_{L^1} + \|\rho_\star - \rho_\star \mathbf{1}_A\|_{L^1} < \frac{\varepsilon}{5}.$$

For each $i \in \{\mathrm{B}, \star\}$, choose $\tilde{\rho}_i \in \mathscr{C}^\infty(\overline{A})$ such that

$$\tilde{\rho}_i > 0 \quad \text{on } \overline{A}, \qquad \int_A \tilde{\rho}_i \, \mathrm{d}x = 1, \qquad \left\| \tilde{\rho}_i - \frac{1}{m_i}\rho_i \mathbf{1}_A \right\|_{L^1(A)} < \frac{\varepsilon}{10}.$$

(For instance, mollify, add a tiny constant, then renormalize.) Denote $\tilde{\mu}_i \coloneqq \tilde{\rho}_i \, \mathcal{L}^d$, so that $\operatorname{supp} \tilde{\mu}_i = \overline{A}$. By Theorem 5.6.11 applied on A to $\tilde{\rho}_\mathrm{B}$ and $\tilde{\rho}_\star$, there exists a diffeomorphism $\psi : \overline{A} \to \overline{A}$ such that $\psi = \mathsf{Id}$ on ∂A and $\psi_\# \tilde{\mu}_\mathrm{B} = \tilde{\mu}_\star$.

Moreover, the Dacorogna–Moser construction yields ψ as the time-one map of a time-dependent vector field $(v_t)_{t\in[0,1]}$ on A, with $v_t = 0$ on ∂A for every t. Extending each v_t by 0 outside A, we obtain a globally Lipschitz vector field on $\mathbb{R}^d$, still denoted by v_t, such that $v_t(0) = 0$ for all $t \in [0,1]$. Since

$$\mathcal{V} = \{v \in \mathscr{C}^{0,1}(\mathbb{R}^d; \mathbb{R}^d) \mid v(0) = 0\},$$

it follows that $v_t \in \mathcal{V}$ for every t, and therefore the corresponding time-one flow Φ belongs to $\mathcal{A}_\mathcal{V}$. Since $\Phi = \mathsf{Id}$ on A^c,

$$|\Phi_\# \mu_\mathrm{B} - \mu_\star|_{\mathsf{TV}} \le |\mu_\mathrm{B} - \mu_\mathrm{B}^A|_{\mathsf{TV}} + |\Phi_\# \mu_\mathrm{B}^A - \mu_\star^A|_{\mathsf{TV}} + |\mu_\star^A - \mu_\star|_{\mathsf{TV}}.$$

Using Lemma 5.6.3 together with $\Phi_\#(m_\mathrm{B}\tilde{\mu}_\mathrm{B}) = m_\mathrm{B}\tilde{\mu}_\star$ and the triangle inequality, we get

$$\begin{aligned} |\Phi_\# \mu_\mathrm{B}^A - \mu_\star^A|_{\mathsf{TV}} &\le |\Phi_\# \mu_\mathrm{B}^A - \Phi_\#(m_\mathrm{B}\tilde{\mu}_\mathrm{B})|_{\mathsf{TV}} + |m_\mathrm{B}\tilde{\mu}_\star - \mu_\star^A|_{\mathsf{TV}} \\ &\le |\mu_\mathrm{B}^A - m_\mathrm{B}\tilde{\mu}_\mathrm{B}|_{\mathsf{TV}} + |m_\mathrm{B} - m_\star| + |m_\star\tilde{\mu}_\star - \mu_\star^A|_{\mathsf{TV}}. \end{aligned}$$

Collecting the bounds gives

$$\begin{aligned} |\Phi_\# \mu_\mathrm{B} - \mu_\star|_{\mathsf{TV}} \le& \frac{1}{2}\|\rho_\mathrm{B} - \rho_\mathrm{B}\mathbf{1}_A\|_{L^1} + \frac{1}{2}\|\rho_\mathrm{B}\mathbf{1}_A - m_\mathrm{B}\tilde{\rho}_\mathrm{B}\|_{L^1} + |m_\mathrm{B} - m_\star| \\ &+ \frac{1}{2}\|m_\star\tilde{\rho}_\star - \rho_\star\mathbf{1}_A\|_{L^1} + \frac{1}{2}\|\rho_\star\mathbf{1}_A - \rho_\star\|_{L^1} < \varepsilon. \end{aligned}$$

Thus the system possesses the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$.

□

Chapter 6

Continuous Normalizing Flows

6.1 Introduction

Besides serving as a useful abstraction of discrete neural networks, the formalism of neural ODEs has had tremendous impact on density estimation through the guise of *normalizing flows* [Gra+19]. This is a popular methodology in machine learning as evidenced by [Pap+21; KPB20; Lip+23; ABV23; GRV22; Poo+23] and the references therein, and goes back at least to [TV10; TT13; DKB15; TW16].

In *continuous normalizing flows*, one aims to estimate an unknown probability density $\rho_\star$ from available samples. This is achieved by fixing a simple initial probability density ρ_{B}, and matching the solution ρ_T of the continuity equation (CE)—recall (DCE) —to the known samples. Specifically, given $x_1, \dots, x_N$ sampled i.i.d. from the unknown density $\rho_\star$, one solves

$$\max_{\theta(\cdot)} \frac{1}{N}\sum_{i=1}^{N} \log \rho_T^\theta(x_i) = \max_{\theta(\cdot)} \frac{1}{N}\sum_{i=1}^{N} \left(\log \rho_{\mathrm{B}}((\Phi_T^\theta)^{-1}(x_i)) + \log\left|\det \nabla (\Phi_T^\theta)^{-1}(x_i)\right|\right). \tag{6.1.1}$$

(We shall keep this presentation formal, and do not detail what constraints are imposed on θ to ensure existence of maximizers.) This is the problem of *maximum (log-)likelihood* estimation on the data. The eponym "normalizing" then stems from the fact that ρ_{B} is most often a Gaussian probability density in practice.

If $\rho_\star$ were to be known, then (6.1.1) is the empirical version of

$$\max_{\theta(\cdot)} \int_{\mathbb{R}^d} \rho_\star(x)\, \log \rho_T^\theta(x)\, \mathrm{d}x. \tag{6.1.2}$$

Up to the constant $\int \rho_\star \log \rho_\star$, it holds:

$$(6.1.2) = \min_{\theta(\cdot)} \int \rho_\star(x) \log \frac{\rho_\star(x)}{\rho_T^\theta(x)}\, \mathrm{d}x. \tag{6.1.3}$$

This is the question of minimizing the *relative entropy* between $\rho_\star$ and ρ_T^θ, or equivalently, the *Kullback-Leibler (KL) divergence* between the measures $\mu_\star := \rho_\star \mathcal{L}^d$ and $\mu_T^\theta := (\Phi_T^\theta)_{\#}\mu_{\mathrm{B}}$. Recalling (II.2), the KL divergence is explicitly given by

$$\mathsf{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) := \int \rho_\star(x) \log \frac{\rho_\star(x)}{\rho_T^\theta(x)}\, \mathrm{d}x. \tag{6.1.4}$$

Minimizing this objective is a problem of general interest in Bayesian inference and sampling—our specific approach here is to parameterize it via densities given as solutions to (CE).

6.2 Setup and main results

Our focus in this chapter is the *feasibility* problem associated to (6.1.3)—namely, given an arbitrarily small $\varepsilon > 0$, we wish to find θ (depending on ε) such that

$$\int \rho_\star(x) \log \frac{\rho_\star(x)}{\rho_T^\theta(x)} \, \mathrm{d}x \leqslant \varepsilon.$$

This constitutes an *approximate controllability* or *approximate transportation* problem, which is nonlinear in the controls θ. Results for (CE) are only known in weaker[1] topologies—such as L^1, L^2 or Wasserstein distances [RZ23; RZ24; Mar+24; ÁSZ24; GRR24]—and, again in weaker topologies, also for continuity equations with vector fields that are linear in the controls θ [Ala+12; Rag24; DMR19; SF25].

Throughout, we restrict to velocity fields generated by $(\mathcal{V}_1^+)$. Accordingly, (CE) takes the form

$$\begin{cases} \partial_t \rho_t^\theta(x) + \operatorname{div}\left(w(t)\ (a(t)\cdot x + b(t))_+\ \rho_t^\theta(x)\right) = 0 & \text{in } [0,T]\times\mathbb{R}^d, \\ \rho_0^\theta(x) = \rho_{\mathrm{B}}(x) & \text{on } \mathbb{R}^d. \end{cases} \tag{DCE_1^+}$$

In addition to addressing the strong topology imposed by the relative entropy, we will require θ to be piecewise constant, with the aim of counting the number of discontinuities (switches). This is of interest because switches can be interpreted as analogous to discrete layers, as discussed throughout this thesis.

In the setting of such θ, given an initial density $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R}^d)$, the continuity equation (DCE_1^+) has a unique solution $\rho \in \mathscr{C}^0([0,T]; L^1 \cap L^\infty(\mathbb{R}^d))$ which is explicit and found by the method of characteristics (Lemma 6.5.3).

Our main result is as follows.

Theorem 6.2.1. *Suppose $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ for an arbitrary mean $m_{\mathrm{B}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\mathrm{B}} \in \mathscr{S}_d^+(\mathbb{R})$. Let $\rho_\star$ be any probability density on $\mathbb{R}^d$ such that $\rho_\star \log \rho_\star \in L^1(\mathbb{R}^d)$ and for some $M > 0$ and $\sigma_\bullet > 0$,*

$$\rho_\star(x) \leqslant \frac{1}{(2\pi\sigma_\bullet^2)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2\sigma_\bullet^2}} =: \rho_\bullet(x) \qquad \text{for all } \|x\| \geqslant M. \tag{6.2.1}$$

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta = (w, a, b) : [0,T] \to \mathbb{R}^{2d+1}$, piecewise constant with finitely many switches, such that the corresponding solution ρ_t^θ to (DCE_1^+) with data $\rho_{\mathrm{B}} = \frac{\mathrm{d}\mu_{\mathrm{B}}}{\mathrm{d}x}$ satisfies

$$\mathsf{KL}\left(\mu_\star \,\|\, \mu_T^\theta\right) = \int \rho_\star(x) \log \frac{\rho_\star(x)}{\rho_T^\theta(x)} \, \mathrm{d}x \leqslant \varepsilon,$$

where $\rho_\star = \frac{\mathrm{d}\mu_\star}{\mathrm{d}x}$ and $\mu_T^\theta = \rho_T^\theta \mathcal{L}^d$.

We defer a detailed discussion on the setup and extensions of the above result to **Section 6.3**, and instead briefly motivate the proof. Our starting point is [RZ24], in which L^1-approximate controllability is shown:

$$\|\rho_T^\theta - \rho_\star\|_{L^1(\mathbb{R}^d)} = \left|\mu_T^\theta - \mu_\star\right|_{\mathsf{TV}} \leqslant \varepsilon.$$

The proof is entirely constructive and hinges on performing exact transportation between piecewise-constant approximations of the initial and target densities, along with a continuity argument. This

[1] see (6.2.2) and Remark 6.3.5.

is achieved by moving and deforming each segment of the approximated initial density to match the approximated target density using a piecewise-constant θ. Consequently, the number of switches in $\theta(\cdot)$ corresponds to the number of segments in the approximation.

In higher dimensions, the proof extends by employing piecewise-constant approximations over grids, which result in an exponential number of cells (and thus segments) in the dimension d. The argument proceeds via induction on the dimension. Since the number of cells grows exponentially with d, the number of switches in θ similarly scales exponentially with d.

One way to make use of [RZ24] is to employ a functional inequality linking KL and TV. A well-known example is the (Csizár-Kullback-)*Pinsker inequality*:

$$|\mu_2 - \mu_1|_{\mathsf{TV}} \leqslant \sqrt{2 \cdot \mathsf{KL}(\mu_2 \,\|\, \mu_1)}, \tag{6.2.2}$$

which holds for arbitrary probability measures μ_1, μ_2. While the inequality sign in (6.2.2) is not in the desired direction so that we can straightforwardly apply it to conclude, it turns out that it can be reversed under stronger conditions on the tails of the measures (*Reverse Pinsker inequality*; see [Ver14; Sas15; SV16]):

Lemma 6.2.2 (Reverse Pinsker inequality). *Let μ_1, μ_2 be probability measures on $\mathbb{R}^d$ such that $\mu_2 \ll \mu_1$. Then*

$$\mathsf{KL}(\mu_2 \,\|\, \mu_1) \leqslant \frac{1}{2} \cdot \frac{\log \left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)}}{1 - \left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)}^{-1}} \cdot |\mu_2 - \mu_1|_{\mathsf{TV}}. \tag{6.2.3}$$

Recall that the Radon-Nikodym theorem only ensures that $\frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1}$ is measurable. Yet for (6.2.3) to be non-void we need

$$\left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)} < +\infty.$$

(Note that this quantity is always $\geqslant 1$ by definition.) Suppose instead that μ_1 and μ_2 are absolutely continuous with respect to the Lebesgue measure. We are asking for

$$\left\| \frac{\mathrm{d}\mu_2}{\mathrm{d}\mu_1} \right\|_{L^\infty(\mathbb{R}^d)} = \operatorname*{ess\,sup}_{x \in \mathbb{R}^d} \frac{\rho_2(x)}{\rho_1(x)} < +\infty.$$

Therefore, to convert the TV approximation result of [RZ24] to a KL one, we also ought to ensure that

$$\mathsf{L}(\rho_\star, \rho_T^\theta) \coloneqq \lim_{\|x\| \to \infty} \frac{\rho_\star(x)}{\rho_T^\theta(x)} < +\infty, \tag{6.2.4}$$

as well as $\rho_T^\theta(x) > 0$ for all $x \in \mathbb{R}^d$. Ensuring this condition is in essence the main novelty of our study. The complete proof can be found in **Subsection 6.5.2**.

At this point, we can also observe that KL is not symmetric with respect to its inputs, whence Theorem 6.2.1 does not directly entail a result when the densities are swapped. Interestingly, minimizing the reverse-KL is a problem of interest in its own right and is referred to as *variational inference* (see [WJ08; BKM17; RM15; JCP25] and the references therein for works on the topic). The following result turns out to be a simple modification of the proof of Theorem 6.2.1.

Theorem 6.2.3. *Suppose $\mu_{\mathrm{B}} = \mathcal{N}(m_{\mathrm{B}}, \Sigma_{\mathrm{B}})$ for an arbitrary mean $m_{\mathrm{B}} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{\mathrm{B}} \in \mathscr{S}_d^+(\mathbb{R})$. Let $\rho_\star$ be a strictly positive probability density such that $\rho_\star^{-1} \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$, $\rho_\star \log \rho_\star \in L^1(\mathbb{R}^d)$, and for some $M > 0$ and $\sigma_\bullet > 0$,*

$$\rho_\star(x) \geqslant \frac{1}{(2\pi\sigma_\bullet^2)^{\frac{d}{2}}} e^{-\frac{\|x\|^2}{2\sigma_\bullet^2}} =: \rho_\bullet(x) \qquad \text{for all } \|x\| \geqslant M. \tag{6.2.5}$$

Then for any $T > 0$ and $\varepsilon > 0$, there exists $\theta = (w, a, b) : [0, T] \to \mathbb{R}^{2d+1}$, piecewise constant with finitely many switches, such that the corresponding solution ρ_t^θ to (DCE_1^+) satisfies

$$\mathsf{KL}\left(\mu_T^\theta \,\|\, \mu_\star\right) = \int \rho_T^\theta(x) \log \frac{\rho_T^\theta(x)}{\rho_\star(x)} \,\mathrm{d}x \leqslant \varepsilon,$$

where $\rho_\star = \frac{\mathrm{d}\mu_\star}{\mathrm{d}x}$ and $\mu_T^\theta = \rho_T^\theta \mathcal{L}^d$.

We provide the proof in **Subsection 6.5.3**.

6.3 Remarks and extensions

We proceed to discuss the various assumptions made in Theorems 6.2.1 and 6.2.3. We address estimates on the number of switches (Remark 6.3.1), the assumption of a Gaussian initial density (Remark 6.3.2), the possible relaxation of condition (6.2.1) (Proposition 6.3.3), the choice of metric (Remark 6.3.5), and estimation from finitely many samples (Remark 6.3.6). Finally, in Remark 6.3.7, we discuss a variant of ($\mathcal{V}_1^+$) linear in θ, generalizing the model considered in ($\mathcal{V}_\ell^+$) for the ReLU.

Remark 6.3.1 (Number of switches). The controls θ from Theorem 6.2.1 have at most

$$L = \left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d + 10) + 2d$$

switches, where $R(\varepsilon), h(\varepsilon) > 0$ are such that

$$\int_{[-R(\varepsilon), R(\varepsilon)]^d} \rho_{\mathrm{B}}(x) \,\mathrm{d}x \wedge \int_{[-R(\varepsilon), R(\varepsilon)]^d} \rho_\star(x) \,\mathrm{d}x > 1 - \varepsilon$$

and

$$\int_{[-R(\varepsilon), R(\varepsilon)]^d} \left|\rho_{\mathrm{B}}^{h_\varepsilon}(x) - \rho_{\mathrm{B}}(x)\right| \,\mathrm{d}x \vee \int_{[-R(\varepsilon), R(\varepsilon)]^d} \left|\rho_\star^{h_\varepsilon}(x) - \rho_\star(x)\right| \,\mathrm{d}x < \varepsilon,$$

where $\rho_{\mathrm{B}}^{h_\varepsilon}$ and $\rho_\star^{h_\varepsilon}$ are piecewise-constant interpolants of ρ_{B} and $\rho_\star$ on a regular grid of the hypercube $[-R(\varepsilon), R(\varepsilon)]^d$ with cells of spacing $h(\varepsilon)$. The exponential growth with d of the number of jumps is due to [RZ24, Theorem 1], where the error scales inversely with the size of a multidimensional grid.

Similarly, in Theorem 6.2.3, the control θ has at most

$$L = \left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d + 10) + 2d,$$

switches, where $R(\varepsilon)$ and $h(\varepsilon)$ are as defined in Theorem 6.2.1.

Remark 6.3.2 (Beyond Gaussians). Theorems 6.2.1 and 6.2.3 can be generalized beyond the case of Gaussian initial densities. For Theorem 6.2.1, it suffices to replace (6.2.1) by the following three conditions:

- The densities ρ_{B} and $\rho_\star$ are strictly positive and moreover $\rho_{\mathrm{B}}^{-1}, \rho_\star \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$. Thus $\rho_{\mathrm{B}}(x) = e^{-U_{\mathrm{B}}(x)}$ and $\rho_\star(x) = e^{-U_\star(x)}$ for some functions $U_{\mathrm{B}}, U_\star : \mathbb{R}^d \to \mathbb{R}$.
- There exists $M_\star > 0$ such that

$$U_\star(x) \gtrsim \|x\| \qquad \text{for all } \|x\| > M_\star. \tag{6.3.1}$$

- Given $A \in \mathscr{M}_{d\times d}(\mathbb{R})$ with $\mathrm{spec}(A) \subset (0, +\infty)$, there exists $\lambda_A > 0$ such that the lower convex envelope, defined for all $x \in \mathbb{R}^d$ by

$$\operatorname{conv} U_\star(x) = \sup_{\substack{U \leqslant U_\star \\ U \text{convex}}} U(x)$$

satisfies

$$\lim_{\|x\|\to\infty} \frac{\operatorname{conv} U_\star(\lambda_A x)}{U_{\mathrm{B}}(Ax)} = +\infty. \tag{6.3.2}$$

For the reverse KL result from Theorem 6.2.3, one replaces the lower-tail hypothesis (6.2.5) by an analogous upper tail domination, and uses an upper convex envelope in place of $\operatorname{conv} U_\star$. In addition, strict positivity of $\rho_\star$ is necessary since ρ_T^θ has full support. We provide the corresponding proof adjustments in **Subsection 6.5.4**.

To build on Remark 6.3.2, one can inquire whether (DCE_1^+) can be used to couple two densities with tails of different nature. To this end, we suspect that the globally Lipschitz character of the vector field needs to be violated, in which case we could obtain an affirmative answer as illustrated by the following partial result.

Proposition 6.3.3. *Fix $p > 0$. Let $v(x) \coloneqq \mathbf{1}_{(1,+\infty)}\, x \log x$ and consider*

$$\begin{cases} \partial_t \rho_t(x) + \partial_x(v(x)\rho_t(x)) = 0 & \text{in } \mathbb{R}_{>0} \times \mathbb{R}, \\ \rho_0(x) = \rho_{\mathrm{B},p}(x) & \text{in } \mathbb{R}, \end{cases} \tag{6.3.3}$$

where $\rho_{\mathrm{B},p} \in \mathscr{C}^0(\mathbb{R})$ with $\rho_{\mathrm{B},p} \propto \exp(-|x|^p)$. Then, for every $q \in (0, p)$ there exists a time $T_q > 0$ such that the unique weak solution ρ to (6.3.3) satisfies

$$\lim_{x\to+\infty} \frac{\rho_{T_q}(x)}{\rho_{\mathrm{B},q}(x)} < +\infty.$$

Remark 6.3.4. The forward equation (6.3.3) can only make the tail heavier. If one instead considers

$$v(x) \coloneqq -\mathbf{1}_{(1,+\infty)}\, x \log x$$

(equivalently, runs the above flow backward-in-time), the same proof yields the analogous conclusion for any $q > p$ by taking $T_q = \log(q/p) > 0$.

We provide the proof in **Subsection 6.5.4**. Proposition 6.3.3 asserts that one can avoid assuming (6.2.1) in Theorem 6.2.1 by sacrificing the globally Lipschitz character of the nonlinearity and, in turn, use (DCE_1^+) to transport a Gaussian to a Laplace density, for example.

Remark 6.3.5 (Metric). We discuss extensions of Theorem 6.2.1 to other divergences which are common in statistics and information theory.

- Consider the squared *Hellinger distance*:

$$\mathsf{H}^2(\mu_1,\mu_2) \coloneqq \int \left(\sqrt{\mu_1(\,\mathrm{d}x)} - \sqrt{\mu_2(\,\mathrm{d}x)}\right)^2.$$

  One has

$$\mathsf{H}^2(\mu_1,\mu_2) \leqslant |\mu_1-\mu_2|_{\mathsf{TV}(\mathbb{R}^d)} \leqslant \sqrt{2}\cdot \mathsf{H}(\mu_1,\mu_2).$$

  Thus a direct application of [RZ24, Theorem 1] gives

$$\mathsf{H}^2\left(\mu_T^\theta, \mu_\star\right) \leqslant \varepsilon.$$

- More generally one could consider any *f-divergence* between $\mu_1 \ll \mu_2$. For a convex function $f : (0,+\infty) \to \mathbb{R}$ with $f(1) = 0$ and $\lim_{t\to 0^+} f(t) \in [0,+\infty]$, the f-divergence [PW24] is defined as

$$\mathsf{D}_f(\mu_1,\mu_2) = \int f\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)\mathrm{d}\mu_2 \tag{6.3.4}$$

  For instance, $f(t) = t\ln t$ corresponds to KL, $f(t) = |t-1|$ corresponds to TV, $f(t) = (\sqrt{t}-1)^2$ corresponds to H^2, and $f(t) = |t-1|^2$ corresponds to χ^2. From (6.3.4) one sees that $f_1 \lesssim f_2$ is equivalent to $\mathsf{D}_{f_1}(\mu_1,\mu_2) \lesssim \mathsf{D}_{f_2}(\mu_1,\mu_2)$ for all μ_1,μ_2 with $\mu_1 \ll \mu_2$. Thus by virtue of Theorem 6.2.1 we deduce that for any $\varepsilon > 0$ and f as above,

$$\text{if} \quad \sup_{t>0}\frac{f(t)}{t\ln t} < +\infty \qquad \text{then} \quad \mathsf{D}_f\left(\mu_\star \,\|\, \mu_T^\theta\right) \leqslant \varepsilon.$$

  The condition is satisfied by $f(t) = |t-1|$ and $f(t) = (\sqrt{t}-1)^2$.

- One could also consider the *Rényi entropy* [PW24]: for any $\lambda > 0$ with $\lambda \neq 1$, the Rényi entropy of order λ between $\mu_1 \ll \mu_2$, is defined as

$$\mathsf{D}_\lambda(\mu_1,\mu_2) \coloneqq \frac{1}{\lambda-1}\log\int_{\mathbb{R}^d}\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right)^\lambda \mathrm{d}\mu_2.$$

  By continuity $\mathsf{D}_1 = \mathsf{KL}$, and similarly $\mathsf{D}_\infty(\mu_1,\mu_2) = \log \operatorname{esssup}\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}$. The latter is known as worst-case regret. If μ_1 and μ_2 have continuous densities denoted ρ_1 and ρ_2 respectively, then $\mathsf{D}_\infty(\mu_1,\mu_2) \leqslant \varepsilon$ implies

$$\|\rho_1-\rho_2\|_{L^\infty(\mathbb{R}^d)} \lesssim \varepsilon. \tag{6.3.5}$$

  Our methods do not extend to the case of worst-case regret since we rely on the $L^1(\mathbb{R}^d)$-approximation of densities of [RZ24], which doesn't imply (6.3.5). On the other hand, for any fixed $\mu_1 \ll \mu_2$, the map $\lambda \in [0,+\infty] \mapsto \mathsf{D}_\lambda(\mu_1,\mu_2)$ is nondecreasing [EH14]. Theorem 6.2.1 thus implies that if $0 \leqslant \lambda \leqslant 1$, then

$$\mathsf{D}_\lambda\left(\mu_\star \,\|\, \mu_T^\theta\right) \leqslant \varepsilon.$$

Remark 6.3.6 (Statistics). In practice one only has access to n samples of the target density $\rho_\star$. Much like [RZ24], we can use Theorem 6.2.1 to obtain sample complexity bounds regarding the explicit construction of θ. To this end, one first employs a nonparametric estimator $\widehat{\rho}_{*,n}$ of $\rho_\star$ using the n samples (e.g., via a kernel density estimator), and then uses Theorem 6.2.1 to approximate $\widehat{\rho}_{*,n}$ to any desired accuracy ε. In doing so, sample complexity-wise, the specific construction of Theorem 6.2.1 can do as well as any nonparametric estimator of $\rho_\star$, see [Tsy08], but at the additional cost that is the number of switches L, which may be exponential in the dimension d.

Remark 6.3.7 (Linear vector field). The related work [SF25] considers the simpler vector field that is linear in θ:

$$v_\theta(x) = w\sigma(x) + b, \qquad \theta = (w, b) \in \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d, \tag{$\mathcal{V}_\ell$}$$

with $\sigma : \mathbb{R} \to \mathbb{R}$ assumed Lipschitz and smooth, and defined component-wise.

When $\sigma(s) = s_+$, the model ($\mathcal{V}_\ell$) cannot yield (6.2.4) in general, because $\sigma|_{\mathbb{R}_{\leqslant 0}} = 0$. Indeed, for $R > 0$ sufficiently large (depending on b) and any $x \in (\mathbb{R}_{<0})^d$ with $\|x\| \geqslant R$, the solution to (CE) with such v_θ is given by

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}\left(x - \int_0^t b(s)\,\mathrm{d}s\right).$$

Far from the origin, $\rho_t^\theta(x)$ is simply a translation of the initial density ρ_{B}. Consequently, the tails of the two distributions will not, in general, be of the same order.

However, much like ($\mathcal{V}_1^+$) [...], one can use ($\mathcal{V}_\ell$) to transport the empirical measure μ_0 supported on $\{x^{(i)}\}_{i=1}^N$ to the target μ_1 on $\{y^{(i)}\}_{i=1}^N$. In the case where σ is smooth, this can be done by using techniques from geometric control theory based on computing iterated Lie brackets, in the spirit of the celebrated Chow-Rashevskii theorem [AS22; CLT20; Sca23; TG23; Ela+22]. In the case of $\sigma(s) = s_+$, this result was established by us in Theorem 3.3.18.

6.4 Perspectives

Several directions suggested by the results and their proofs seem natural.

1. **Beyond neural ODEs.** The present constructions rely heavily on the fact that, under piecewise-constant controls, flow maps generated by ($\mathcal{V}_1^+$) become piecewise affine and thus preserve Gaussian tails. A first extension is to use general controlled flows as in **Chapter 5**, while still aiming for KL-approximation, which is stronger than Wasserstein and total variation. In that setting, Gaussian tails need not be preserved exactly; however, one may hope to control tail growth through quantitative stability estimates for the flow map.

2. **Necessary conditions.** While Theorems 6.2.1 and 6.2.3 provide sufficient conditions for approximate controllability in the KL topology, it is natural to analyze necessary conditions on the pair $(\mu_{\mathrm{B}}, \mu_\star)$.

   Approximation in KL divergence is strictly more demanding than in TV; recall (6.2.2). In particular, it requires the Radon-Nikodym derivative to exist for some admissible θ, which translates into rigid constraints on the supports. More precisely, for $\mathsf{KL}(\mu_\star\|\mu_T^\theta) < \infty$, one requires $\mu_\star \ll \mu_T^\theta$ for some θ, whereas for $\mathsf{KL}(\mu_T^\theta\|\mu_\star) < \infty$ one needs $\mu_T^\theta \ll \mu_\star$. In the usual setting where μ_{B} has full support (e.g., a Gaussian), μ_T^θ is also supported on $\mathbb{R}^d$. This forces $\mu_\star$ to have full support so as to attain $\mathsf{KL}(\mu_T^\theta\|\mu_\star) < \infty$.

3. **Optimal control.** The results in this chapter guarantee that $\mathsf{KL}(\mu_\star\|\mu_T)$ can be made arbitrarily small, but the constructive proofs rely on controls with high complexity (see Remark 6.3.1). This motivates the study of the corresponding *optimal control* problem: minimizing the KL divergence subject to the continuity equation constraint. In contrast with standard quadratic tracking costs, this choice leads to a singular adjoint structure. Formally, the first variation reads

   $$\frac{\delta}{\delta\mu}\mathsf{KL}(\mu_\star\|\mu) = -\frac{\mathrm{d}\mu_\star}{\mathrm{d}\mu},$$

   implying that the adjoint state is transported backward with a terminal datum given by the density ratio. Introducing a running tracking term might yield a more tractable optimality

system and connect the problem to classical PDE-constrained optimization (Pontryagin maximum principle, bang-bang structure, links to dynamic optimal transport).

4. **Greedy training.** From a computational perspective, the controls constructed in this chapter are piecewise constant. Consequently, the evolving measure itself admits a closed-form solution given by Lemma 6.5.3. The complexity for this controllability task can be high—recall Remark 6.3.1—but the local explicit structure suggests a greedy training strategy.

   Instead of optimizing a time-dependent field over the entire horizon (which typically requires expensive integration), one could train the model sequentially, solving a minimization problem for the KL divergence locally on each time subinterval. By exploiting the explicit density updates at each step, such a method would bypass the computational bottleneck of ODE solvers, offering a faster and mathematically interpretable alternative to end-to-end training of CNFs, aligned with the *switching* viewpoint developed throughout this thesis.

5. **Gradient flow.** Another analytic perspective is to study the intrinsic steepest descent of $\mu \mapsto \mathsf{KL}(\mu_\star \| \mu)$ with respect to the 2-Wasserstein metric. For $\mu_{\mathrm{B}} = \rho_{\mathrm{B}} \mathcal{L}^d$ and $\mu_\star = \rho_\star \mathcal{L}^d$, a formal computation shows that the density of μ follows the Wasserstein gradient flow:

$$\partial_t \rho_t - \operatorname{div}\big(\rho_\star \nabla \log \rho_t\big) = -\Delta \rho_\star, \qquad \rho_0 = \rho_{\mathrm{B}}$$

   or, in terms of the relative density $w_t = \rho_t / \rho_\star$,

$$\partial_t w_t = \frac{1}{\rho_\star} \operatorname{div}\big(\rho_\star \nabla \log w_t\big), \qquad w_0 = \frac{\rho_{\mathrm{B}}}{\rho_\star}.$$

   This describes a logarithmic diffusion equation with weight $\rho_\star$. The qualitative and asymptotic behavior of such flow appears less developed than the classical case of the reverse KL—whose Wasserstein gradient flow is the Fokker-Planck equation, see [JKO98]—and may provide an interesting bridge between CNFs and diffusion-type PDEs.

6.5 Proofs

6.5.1 Preliminary lemmas

We begin with a brief discussion on when $\mathsf{L}(\rho_\star, \rho_T^\theta) = \lim_{\|x\| \to \infty} \rho_\star(x)/\rho_T^\theta(x)$ can be made finite when $d = 1$, purely for illustrative purposes.

The following result provides an explicit formula for the evolved density under piecewise-constant controls; it follows directly by applying Lemma 1.3.5 recursively over each interval of the partition.

Lemma 6.5.1. *Take $T > 0$, a probability density $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R})$, and*

$$(w, a, b)(t) = \sum_{k=1}^{K} (w_k, a_k, b_k) \mathbf{1}_{[t_{k-1}, t_k)}(t) \qquad \text{for } t \in [0, T],$$

where $w_k, a_k, b_k \in \mathbb{R}$ with $a_k \neq 0$ and $0 = t_0 < t_1 < \ldots < t_K = T$. Then the unique solution to

$$\begin{cases} \partial_t \rho_t^\theta(x) + \partial_x \left(w(t)(a(t)x + b(t))_+ \rho_t^\theta(x) \right) = 0 & \text{on } [0, T] \times \mathbb{R}, \\ \rho_0^\theta(x) = \rho_{\mathrm{B}}(x) & \text{on } \mathbb{R} \end{cases}$$

is given by

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}(x) \mathbf{1}_{\{a_1 \cdot + b_1 \leqslant 0\}}(x) + e^{-w_1 a_1 t} \rho_{\mathrm{B}} \left(\left(x + \frac{b_1}{a_1} \right) e^{-w_1 a_1 t} - \frac{b_1}{a_1} \right) \mathbf{1}_{\{a_1 \cdot + b_1 \geqslant 0\}}(x)$$

for $t \in [0, t_1]$ *and*

$$\rho_t^\theta(x) = \rho_{t_{k-1}}^\theta(x)\mathbf{1}_{\{a_k\cdot+b_k\leqslant 0\}}(x) + e^{-w_k a_k(t-t_{k-1})}\rho_{t_{k-1}}^\theta\left(\left(x+\frac{b_k}{a_k}\right)e^{-w_k a_k(t-t_{k-1})} - \frac{b_k}{a_k}\right)\mathbf{1}_{\{a_k\cdot+b_k>0\}}(x)$$

for $t \in [t_{k-1}, t_k]$, *for all* $x \in \mathbb{R}$.

Remark 6.5.2. If $a_k = 0$, then

$$\rho_t^\theta(x) = \rho_{t_{k-1}}^\theta\big(x - w_k(b_k)_+(t-t_{k-1})\big) \qquad \text{for all } t \in [t_{k-1}, t_k],\ x \in \mathbb{R},$$

again by the method of characteristics.

From Lemma 6.5.1 we readily gather that

$$\rho_T^\theta(x) = \begin{cases} \alpha^+\rho_{\mathrm{B}}(\alpha^+ x + \beta^+), & \text{if } x \geqslant M, \\ \alpha^-\rho_{\mathrm{B}}(\alpha^- x + \beta^-), & \text{if } x \leqslant -M \end{cases}$$

for some $\alpha^\pm > 0$, $\beta^\pm \in \mathbb{R}$ and $M > 0$, all explicitly depending on (w, a, b)—in other words, one can choose $\alpha^\pm, \beta^\pm$ by choosing (w, a, b). Now consider ρ_{B} and $\rho_\star$ in Gibbs form:

$$\rho_{\mathrm{B}}(x) \propto e^{-U_{\mathrm{B}}(x)}, \qquad \rho_\star(x) \propto e^{-U_\star(x)},$$

with $U_{\mathrm{B}}, U_\star \in \mathscr{C}^0(\mathbb{R}; \mathbb{R}_{\geqslant 0})$. We see that

$$\frac{\rho_\star(x)}{\rho_T^\theta(x)} \propto \exp\big(U_{\mathrm{B}}(\alpha^\pm x + \beta^\pm) - U_\star(x)\big)$$

for all $\|x\| \geqslant M$. So $\mathsf{L}(\rho_\star, \rho_T^\theta)$ defined in (6.2.4) is finite if and only if

$$\lim_{\|x\|\to\infty} U_{\mathrm{B}}(\alpha^\pm x + \beta^\pm) - U_\star(x) \in [-\infty, +\infty). \tag{6.5.1}$$

One sees the constraints that (6.5.1) entails. For instance, one can consider as input the Laplace distribution—corresponding to $U_{\mathrm{B}}(x) = |x - m|/\sigma$ with parameters $m_{\mathrm{B}} \in \mathbb{R}$ and $\sigma_{\mathrm{B}} > 0$—, and as target the Gaussian—corresponding to $U_\star(x) = (x - m_\star)^2/2\sigma_\star^2$ with $m_\star \in \mathbb{R}$ and $\sigma_\star > 0$—, but not the converse.

Lemma 6.5.1 can be generalized to the case $d \geqslant 1$ without difficulty.

Lemma 6.5.3. *Take* $T > 0$, *a probability density* $\rho_{\mathrm{B}} \in \mathscr{C}^0(\mathbb{R}^d)$, *and*

$$(w, a, b)(t) = \sum_{k=1}^K (w_k, a_k, b_k)\mathbf{1}_{[t_{k-1}, t_k)}(t) \qquad \textit{for } t \in [0, T],$$

where $w_k, a_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, *and* $0 = t_0 < t_1 < \ldots < t_K = T$. *Then the unique solution to*

$$\begin{cases} \partial_t \rho_t^\theta(x) + \operatorname{div}\big(w(t)\,(a(t)\cdot x + b(t))_+\,\rho_t^\theta(x)\big) = 0 & \textit{on } [0,T]\times\mathbb{R}^d, \\ \rho_0^\theta(x) = \rho_{\mathrm{B}}(x) & \textit{on } \mathbb{R}^d \end{cases} \tag{6.5.2}$$

is given by

$$\rho_t^\theta(x) = \rho_{t_{k-1}}^\theta(x)\mathbf{1}_{\left\{\langle a_k,\cdot\rangle + b_k \leqslant 0\right\}}(x)$$

$$+ e^{-\langle w_k, a_k\rangle (t-t_{k-1})} \rho^\theta_{t_{k-1}} \left(\mathsf{A}_k(x)\right) \mathbf{1}_{\left\{\langle a_k, \cdot\rangle + b_k > 0\right\}}(x)$$

if $\langle w_k, a_k\rangle \neq 0$, *where*

$$\mathsf{A}_k(x) = e^{-(t-t_{k-1}) w_k a_k^\top} x - b_k \frac{1 - e^{-(t-t_{k-1})\langle w_k, a_k\rangle}}{\langle w_k, a_k\rangle} w_k,$$

and

$$\rho^\theta_t(x) = \rho^\theta_{t_{k-1}}(x) \mathbf{1}_{\left\{\langle a_k, \cdot\rangle + b_k \leqslant 0\right\}}(x) + \rho^\theta_{t_{k-1}}(\mathsf{B}_k(x)) \mathbf{1}_{\left\{\langle a_k, \cdot\rangle + b_k > 0\right\}}(x)$$

if $\langle w_k, a_k\rangle = 0$, *where*

$$\mathsf{B}_k(x) = e^{-(t-t_{k-1}) w_k a_k^\top} x - e^{-(t-t_{k-1}) w_k a_k^\top} (t - t_{k-1}) b_k w_k,$$

for all $t \in [t_{k-1}, t_k]$ *and for all* $x \in \mathbb{R}^d$.

Proof. We again start by $t \in [0, t_1)$. For $x \in \{\langle a_1, \cdot\rangle + b_1 \leqslant 0\}$, we clearly have $\rho^\theta_t(x) = \rho_{\mathrm{B}}(x)$. On $\{\langle a_1, \cdot\rangle + b_1 > 0\}$, the characteristics $x(t)$ are parametrized by solutions to

$$\dot{x}(t) = w_1 a_1^\top x(t) + w_1 b_1.$$

Moreover, setting $y(t) \coloneqq \langle a_1, x(t)\rangle + b_1$, we have $\dot{y}(t) = \langle a_1, w_1\rangle y(t)$, so the sign of $y(t)$ is preserved along trajectories. Since the divergence of the vector field on $\{\langle a_1, \cdot\rangle + b_1 > 0\}$ is $\langle w_1, a_1\rangle$, the density satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho^\theta_t(x(t)) = -\langle w_1, a_1\rangle \rho^\theta_t(x(t)).$$

Suppose $\langle w_1, a_1\rangle \neq 0$. Since

$$e^{t w_1 a_1^\top} = I_d + \frac{e^{t\langle w_1, a_1\rangle} - 1}{\langle w_1, a_1\rangle} w_1 a_1^\top,$$

the variation of constants formula yields

$$x(t) = e^{t w_1 a_1^\top} x(0) + \int_0^t e^{(t-s) w_1 a_1^\top} w_1 b_1 \,\mathrm{d}s = e^{t w_1 a_1^\top} x(0) + b_1 \frac{e^{t\langle w_1, a_1\rangle} - 1}{\langle w_1, a_1\rangle} w_1.$$

Inverting this relation gives

$$x(0) = e^{-t w_1 a_1^\top} x(t) - b_1 \frac{1 - e^{-t\langle w_1, a_1\rangle}}{\langle w_1, a_1\rangle} w_1.$$

Using $\rho^\theta_t(x(t)) = e^{-t\langle w_1, a_1\rangle} \rho_{\mathrm{B}}(x(0))$, we obtain

$$\rho^\theta_t(x) = e^{-t\langle w_1, a_1\rangle} \rho_{\mathrm{B}}\left(e^{-t w_1 a_1^\top} x - b_1 \frac{1 - e^{-t\langle w_1, a_1\rangle}}{\langle w_1, a_1\rangle} w_1\right)$$

for $x \in \{\langle a_1, \cdot\rangle + b_1 > 0\}$.

Now suppose $\langle w_1, a_1\rangle = 0$. Then the divergence is zero, so $\rho^\theta_t(x(t)) = \rho_{\mathrm{B}}(x(0))$ by (1.2.3). Since $w_1 a_1^\top$ is nilpotent, $e^{t w_1 a_1^\top} = I_d + t w_1 a_1^\top$. Thus

$$x(t) = e^{t w_1 a_1^\top} x(0) + t b_1 w_1,$$

and hence

$$x(0) = e^{-tw_1a_1^\top} x - e^{-tw_1a_1^\top} tb_1w_1.$$

Therefore

$$\rho_t^\theta(x) = \rho_{\mathrm{B}}\left(e^{-tw_1a_1^\top} x - e^{-tw_1a_1^\top} tb_1w_1\right)$$

for $x \in \{\langle a_1, \cdot\rangle + b_1 > 0\}$.

The computations are easily repeated for $k \geqslant 2$, replacing ρ_{B} by $\rho^\theta_{t_{k-1}}$ and t by $t - t_{k-1}$. □

The following comparison principle will be useful.

Lemma 6.5.4. *Let $d \geq 1$ and let ρ^θ and $\bar\rho^\theta$ be two solutions to (6.5.2) corresponding to the same θ, and to data $\rho_0^\theta(\cdot)$ and $\bar\rho_0^\theta(\cdot)$ respectively.*

Assume $\rho_0^\theta(x) < \bar\rho_0^\theta(x)$ for all $x \in \mathbb{R}^d$. Then $\rho_t^\theta(x) < \bar\rho_t^\theta(x)$ for all $t \in [0,T]$ and $x \in \mathbb{R}^d$.

Proof. Set $f(t,x) \coloneqq \bar\rho_t^\theta(x) - \rho_t^\theta(x)$ and $v_{\theta(t)}(x) \coloneqq w(t)(a(t)\cdot x + b(t))_+$. Along the characteristics $\dot x(t) = v_{\theta(t)}(x(t))$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} f(t,x(t)) &= \partial_t f(t,x(t)) + \nabla_x f(t,x(t))^\top v_{\theta(t)}(x(t)) \\ &= -f(t,x(t))\,\mathrm{div}\; v_{\theta(t)}(x(t)).\end{aligned}$$

Thus

$$f(t,x(t)) = f(0,x(0)) \exp\left(-\int_0^t \mathrm{div}\, v_{\theta(s)}(x(s))\,\mathrm{d}s\right)$$

for all $t \in [0,T]$ and $x(0) \in \mathbb{R}^d$. We may conclude. □

We also use the following elementary lemma.

Lemma 6.5.5. *Let ρ and $\bar\rho$ be two nonnegative and integrable functions that are proportional to Gaussian densities with covariance matrices Σ and $\bar\Sigma$ respectively. Suppose*

$$\left\langle (\bar\Sigma^{-1} - \Sigma^{-1})e_i, e_i\right\rangle < 0, \qquad \text{for some } i \in [d].$$

Then, for any $M > 0$, there exists $M(i) > 0$ such that

$$\bar\rho(x) > \rho(x) \qquad \text{in } \left\{x \in \mathbb{R}^d \mid |x^{(i)}| > M(i) \text{ and } |x^{(j)}| < M \text{ for } j \neq i\right\}.$$

If $\bar\Sigma^{-1} \prec \Sigma^{-1}$, then there exists some $\overline{M} > 0$ such that

$$\bar\rho(x) > \rho(x) \qquad \text{for all } \|x\| > \overline{M}.$$

Proof. Write

$$\frac{\bar\rho(x)}{\rho(x)} \propto e^{-\frac{1}{2}\left\langle (\bar\Sigma^{-1} - \Sigma^{-1})x, x\right\rangle + O(\|x\|)}.$$

The proof follows. □

6.5.2 Proof of Theorem 6.2.1

Proof of Theorem 6.2.1. Fix $\varepsilon_0 > 0$; $\theta = (w, a, b)$ takes the form

$$\theta(t) = \bar{\theta}(t)\mathbf{1}_{[0,\frac{T}{2}]}(t) + \sum_{k=1}^{2d} \theta_k \mathbf{1}_{[T_{k-1},T_k]}(t),$$

where $T_k \coloneqq \frac{T}{2} + \frac{kT}{4d}$, and $\bar{\theta}$ is piecewise constant with at most $L = \left\lceil \frac{2R(\varepsilon)}{h(\varepsilon)} \right\rceil^d (d+10)$ switches and such that

$$\left\| \rho^{\theta}_{T/2} - \rho_\star \right\|_{L^1(\mathbb{R}^d)} \leqslant \varepsilon_0. \tag{6.5.3}$$

Such a $\bar{\theta}$ exists by virtue of [RZ24, Theorem 1]. We seek to construct $\theta_1, \dots, \theta_{2d}$ so that

$$\mathsf{L}(\rho_\star, \rho^{\theta}_T) = \lim_{\|x\| \to +\infty} \frac{\rho_\star(x)}{\rho^{\theta}_T(x)} = 0.$$

Step 1.

Thanks to Lemma 6.5.3, there exists a finite $n \geqslant 1$, as well as polyhedra $P_i \subset \mathbb{R}^d$ and coefficients $(\alpha_i, A_i, \beta_i) \in \mathbb{R}_{>0} \times \mathscr{M}_{d\times d}(\mathbb{R}) \times \mathbb{R}^d$ with $\operatorname{spec}(A_i) \subset (0, +\infty)$ for $i \in [n]$, such that

$$\rho^{\theta}_{T/2}(x) = \sum_{i=1}^{n} \alpha_i \rho_{\mathrm{B}}(A_i x + \beta_i)\mathbf{1}_{P_i}(x) \qquad \text{for all } x \in \mathbb{R}^d.$$

Moreover, since the polyhedra P_i form a partition of $\mathbb{R}^d$ and each ρ_i is strictly positive, we have $\rho^{\theta}_{T/2} > 0$. Each function $\rho_i \coloneqq \alpha_i \rho_{\mathrm{B}}(A_i \cdot + \beta_i)$ is a Gaussian on $\mathbb{R}^d$, and the probability density $\rho_i / (\int \rho_i)$ has covariance matrix

$$\Sigma_i \coloneqq A_i^{-1} \Sigma_{\mathrm{B}} \left(A_i^{-1}\right)^\top.$$

Take $0 < \underline{\sigma}^2 < \min_{1\leqslant i\leqslant n} \operatorname{spec}(\Sigma_i)$ and define

$$\underline{\rho}(x) \coloneqq \alpha \exp\left(-\frac{\|x\|^2}{2\underline{\sigma}^2}\right) \tag{6.5.4}$$

for $x \in \mathbb{R}^d$, where $\alpha > 0$ is a constant to be specified. Then $\Sigma_i^{-1} \prec I_d/\underline{\sigma}^2$ for all $i \in [n]$, so by virtue of Lemma 6.5.5, there exists $M(i) > 0$ such that

$$\underline{\rho}(x) < \rho_i(x) \qquad \text{for all} \quad \|x\| > M(i).$$

Taking $\overline{M} \geqslant M \vee \max_{1\leqslant i\leqslant n} M(i)$ it follows that (6.2.1) and $\underline{\rho}(x) < \rho^{\theta}_{T/2}(x)$ hold for all $\|x\| > \overline{M}$. Choosing α small enough, we can deduce

$$\underline{\rho}(x) < \rho^{\theta}_{T/2}(x) \qquad \text{for all } x \in \mathbb{R}^d, \tag{6.5.5}$$

and $\overline{M}$ large enough to deduce

$$\int_{[-\overline{M},\overline{M}]^d} \rho_\star(x)\,\mathrm{d}x \wedge \int_{[-\overline{M},\overline{M}]^d} \rho^{\theta}_{T/2}(x)\,\mathrm{d}x > 1 - \varepsilon_0. \tag{6.5.6}$$

Step 2.

We will choose $\theta_1, \dots, \theta_{2d}$ so that

$$\rho_\bullet(x) < \underline{\rho}_T(x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d, \tag{6.5.7}$$

for some larger $\overline{\overline{M}} \geqslant \overline{M}$, where $\underline{\rho}_t(\cdot)$ is the unique solution to (DCE_1^+) in $[T/2, T]$ with data $\underline{\rho}_{T/2}(\cdot) = \underline{\rho}(\cdot)$. To this end, take

$$\begin{aligned} \theta_{2k-1} &\coloneqq (\omega e_k, e_k, -\overline{M}), \\ \theta_{2k} &\coloneqq (-\omega e_k, -e_k, -\overline{M}), \end{aligned}$$

where $\omega > 0$ is to be chosen, for $k \in [d]$. By virtue of Lemma 6.5.3, the solution reads

$$\underline{\rho}_t(x) = \underline{\rho}_{T_{2k-2}}(x)\mathbf{1}_{\left\{x^{(k)} \leqslant \overline{M}\right\}}(x) + e^{-\omega(t-T_{2k-2})}\underline{\rho}_{T_{2k-2}}\left(\mathsf{A}_{2k-2}(t,x)\right)\mathbf{1}_{\left\{x^{(k)} > \overline{M}\right\}}(x)$$

if $t \in [T_{2k-2}, T_{2k-1}]$, and

$$\underline{\rho}_t(x) = \underline{\rho}_{T_{2k-1}}(x)\mathbf{1}_{\left\{x^{(k)} \geqslant -\overline{M}\right\}}(x) + e^{-\omega(t-T_{2k-1})}\underline{\rho}_{T_{2k-1}}\left(\mathsf{A}_{2k-1}(t,x)\right)\mathbf{1}_{\left\{x^{(k)} < -\overline{M}\right\}}(x)$$

if $t \in [T_{2k-1}, T_{2k}]$, where

$$\mathsf{A}_j(t,x) = e^{-\omega(t-T_j)e_k e_k^\top} x + (-1)^j \overline{M}\left(1 - e^{-\omega(t-T_j)}\right) e_k,$$

for $j \in \{2k-2, 2k-1\}$, for all $k \in [d]$ and $x \in \mathbb{R}^d$. It ensues that

$$\underline{\rho}_T(x) = \underline{\rho}_{T/2}(x)\mathbf{1}_{[-\overline{M},\overline{M}]^d}(x) + \sum_{k=1}^{d}\sum_{\ell=0}^{1}\underline{\rho}_1^{k\ell}(x)\mathbf{1}_{\mathscr{S}_{k\ell}}(x) + \sum_{q\in\{0,1\}^d}\underline{\rho}_2^q(x)\mathbf{1}_{\mathscr{Q}_q}(x), \tag{6.5.8}$$

where $\underline{\rho}_1^{k\ell}$, $\underline{\rho}_2^q$ are Gaussian functions on $\mathbb{R}^d$, and $\mathscr{Q}_q$, $\mathscr{S}_{k\ell}$ are defined by

$$\mathscr{Q}_q \coloneqq \left\{x \in \mathbb{R}^d \mid (-1)^{q_j} x^{(j)} > \overline{M} \quad \text{for all } j \in [d]\right\}$$

and

$$\mathscr{S}_{k\ell} \coloneqq \left\{(-1)^\ell x^{(k)} > \overline{M}\right\} \setminus \bigcup_{q\in\{0,1\}^d} \mathscr{Q}_q.$$

The probability density $\underline{\rho}_1^{k\ell}/(\int \underline{\rho}_1^{k\ell})$ has covariance

$$\begin{aligned} \Sigma_{1,k\ell} &= \left(e^{-\frac{\omega T}{4d}e_k e_k^\top}\right)^{-1} \underline{\sigma}^2 I_d \left(\left(e^{-\frac{\omega T}{4d}e_k e_k^\top}\right)^{-1}\right)^\top \\ &= \underline{\sigma}^2 \left(I_d + (e^{-\frac{\omega T}{4d}} - 1)e_k e_k^\top\right)^{-1} \left(I_d + (e^{-\frac{\omega T}{4d}} - 1)e_k e_k^\top\right)^{-1} \\ &= \underline{\sigma}^2 \left(I_d + (e^{\frac{\omega T}{4d}} - 1)e_k e_k^\top\right) \left(I_d + (e^{\frac{\omega T}{4d}} - 1)e_k e_k^\top\right) \\ &= \underline{\sigma}^2 \left(I_d + (e^{\frac{\omega T}{2d}} - 1)e_k e_k^\top\right), \end{aligned}$$

Taking

$$\omega > \frac{2d}{T}\log\left(\frac{\sigma_\bullet^2}{\underline{\sigma}^2}\right), \tag{6.5.9}$$

we find for each $k \in [d]$, $\ell \in \{0,1\}$, that

$$\left\langle \left(\Sigma_{1,k\ell}^{-1} - \frac{1}{\sigma_\bullet^2} I_d \right) e_k, e_k \right\rangle = \frac{e^{-\frac{\omega T}{2d}}}{\underline{\sigma}^2} - \frac{1}{\sigma_\bullet^2} < 0.$$

A similar computation for the covariance of $\underline{\rho}_2^q / \int \underline{\rho}_2^q$ gives $\Sigma_{2,q} = \underline{\sigma}^2 e^{\frac{\omega T}{2d}} I_d$. We see that $\Sigma_{2,q}^{-1} \prec I_d/\sigma_\bullet^2$ for all $q \in \{0,1\}^d$ when ω satisfies (6.5.9). By Lemma 6.5.5 there exists some $M(0) \geqslant \overline{M}$ such that

$$\rho_\bullet(x) < \underline{\rho}_2^q(x) \qquad \text{for all } \|x\| > M(0),$$

for $q \in \{0,1\}^d$. Also by Lemma 6.5.5, for all $k \in [d]$ and $\ell \in \{0,1\}$ there exists $M(k) > 0$ such that

$$\rho_\bullet(x) < \underline{\rho}_1^{k\ell}(x) \qquad \text{in } \left\{ x \in \mathbb{R}^d \mid |x^{(k)}| > M(k) \text{ and } |x^{(j)}| < M(0) \text{ for } j \neq k \right\}.$$

Thanks to (6.5.8), taking $\overline{\overline{M}} \coloneqq \max_{0 \leqslant k \leqslant d} M(k)$ ensures (6.5.7).

Step 3.

Combining (6.2.1), (6.5.5), (6.5.7) and Lemma 6.5.4, we find

$$\rho_\star(x) < \rho_T^\theta(x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d.$$

In particular, this implies $\mathsf{L}(\rho_\star, \rho_T^\theta) = 0$, where L is defined in (6.2.4). We moreover have

$$\min_{x \in [-\overline{\overline{M}}, \overline{\overline{M}}]^d} \rho_T^\theta(x) > \min_{x \in [-\overline{\overline{M}}, \overline{\overline{M}}]^d} \underline{\rho}_T(x) > 0$$

by (6.5.4), (6.5.5) and (6.5.8), so we conclude that

$$\sup_{x \in \mathbb{R}^d} \frac{\rho_\star(x)}{\rho_T^\theta(x)} < +\infty.$$

On the other hand, $\rho_T^\theta(x) = \rho_{T/2}^\theta(x)$ for $x \in [-\overline{M}, \overline{M}]^d$, which, combined with (6.5.3) and (6.5.6), yields

$$\begin{aligned} \left\| \rho_T^\theta - \rho_\star \right\|_{L^1(\mathbb{R}^d)} &\leqslant \int_{[-\overline{M}, \overline{M}]^d} \left| \rho_{T/2}^\theta(x) - \rho_\star(x) \right| \mathrm{d}x \\ &\quad + \int_{\mathbb{R}^d \setminus [-\overline{M}, \overline{M}]^d} \rho_T^\theta(x) \, \mathrm{d}x \\ &\quad + \int_{\mathbb{R}^d \setminus [-\overline{M}, \overline{M}]^d} \rho_\star(x) \, \mathrm{d}x < 3\varepsilon_0. \end{aligned}$$

Taking ε_0 small enough and applying Lemma 6.2.2, we conclude. □

6.5.3 Proof of Theorem 6.2.3

Proof of Theorem 6.2.3. The proof is almost identical to that of Theorem 6.2.1—we only discuss the differences.

In Step 1, we take $\overline{\sigma}^2 > \max_{1 \leqslant i \leqslant n} \operatorname{spec}(\Sigma_i)$ instead of $\underline{\sigma}^2$, defining $\overline{\rho}(x)$ as $\underline{\rho}(x)$ in (6.5.4), as to have $\overline{\rho}(x) > \rho_{T/2}^\theta(x)$ for all $x \in \mathbb{R}^d$.

In Step 2, we take

$$\theta_{2k-1} = (-\omega e_k, e_k, -\overline{M}),$$
$$\theta_{2k} = (\omega e_k, -e_k, -\overline{M}),$$

for $k \in [d]$, with

$$\omega > \frac{2d}{T} \log\left(\frac{\overline{\sigma}^2}{\sigma_\bullet^2}\right).$$

Then the same argument as in Step 2 yields $\overline{\rho}_T(x) < \rho_\bullet(x)$ for all $x \in \mathbb{R}^d \setminus K$, where $K \coloneqq [-\overline{\overline{M}}, \overline{\overline{M}}]^d$ for some $\overline{\overline{M}} \geqslant \overline{M}$.

In Step 3, by Lemma 6.5.4 we have $\rho_T^\theta \leq \overline{\rho}_T$ on $\mathbb{R}^d$. Combining this with Step 2 and (6.2.5) (enlarging $\overline{\overline{M}}$ so that $\overline{\overline{M}} \geq M$) yields $\rho_T^\theta(x) \leq \rho_\bullet(x) \leq \rho_\star(x)$ for all $x \in \mathbb{R}^d \setminus K$.

On K, $\rho_T^\theta \in L^\infty(K)$ and $\rho_\star^{-1} \in L^\infty(K)$ by assumption, hence

$$\left\| \frac{\rho_T^\theta}{\rho_\star} \right\|_{L^\infty(K)} < +\infty.$$

Together with $\rho_T^\theta / \rho_\star \leq 1$ on K^c, we get

$$\left\| \frac{\rho_T^\theta}{\rho_\star} \right\|_{L^\infty(\mathbb{R}^d)} < +\infty.$$

Therefore Lemma 6.2.2 applies with $(\mu_2, \mu_1) = \left((\Phi_T^\theta)_\# \mu_{\mathrm{B}}, \mu_\star\right)$, and the proof concludes.

□

6.5.4 Proof of Proposition 6.3.3

Proof of Proposition 6.3.3. For $x \in \mathbb{R}$ and $t \in \mathbb{R}$, consider

$$x(t) = \begin{cases} x^{e^t} & \text{if } x \geqslant 1, \\ x & \text{if } x \leqslant 1. \end{cases}$$

This function solves

$$\begin{cases} \dot{x}(t) = v(x(t)) & t \in \mathbb{R}, \\ x(0) = x. \end{cases}$$

It is in fact the unique solution. Indeed, v is locally Lipschitz on $\mathbb{R}$, hence uniqueness holds on the maximal interval of existence, and the above formula defines $x(t)$ for all $t \in \mathbb{R}$.

One can thus define the solution $\rho_t(x)$ to (6.3.3) uniquely in the weak sense. We focus on the behavior of the solution $\rho_t(\cdot)$ on $\{x > 1\}$. By the method of characteristics, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t} \rho_t(x(t)) = -v'(x(t))\, \rho_t(x(t)) = -(\log x(t) + 1) \rho_t(x(t)),$$

which yields, for $x > 1$,

$$\log \rho_t(x(t)) - \log \rho_{\mathrm{B},p}(x) = -\int_0^t (\log(x(s)) + 1)\, \mathrm{d}s = -\left(\log(x)(e^t - 1) + t\right).$$

Therefore

$$\rho_t(x^{e^t}) = e^{-t}\rho_{\mathrm{B},p}(x)x^{-(e^t-1)}.$$

Taking $y = x^{e^t}$, we obtain $x = y^{\frac{1}{e^t}}$ and hence, for $y \geqslant 1$,

$$\rho_t(y) = e^{-t}\rho_{\mathrm{B},p}\left(y^{\frac{1}{e^t}}\right) y^{\frac{1}{e^t}-1}.$$

Since $v \equiv 0$ on $(-\infty, 1]$, we also have $\rho_t(x) = \rho_{\mathrm{B},p}(x)$ for $x \leqslant 1$. Summarizing,

$$\rho_t(x) = \rho_{\mathrm{B},p}(x)\mathbf{1}_{(-\infty,1)}(x) + e^{-t}\rho_{\mathrm{B},p}\left(x^{\frac{1}{e^t}}\right) x^{\frac{1}{e^t}-1}\mathbf{1}_{[1,\infty)}(x).$$

We thus see that

$$\lim_{x\to+\infty} \frac{\rho_t(x)}{\rho_{\mathrm{B},q}(x)} \propto e^{-t} \lim_{x\to+\infty} \frac{\exp\left(-x^{pe^{-t}}\right) x^{e^{-t}-1}}{\exp(-x^q)}.$$

It therefore suffices to take

$$T_q = \log\left(\frac{p}{q}\right) > 0,$$

which ensures $pe^{-T_q} = q$ and hence

$$\lim_{x\to+\infty} \frac{\rho_{T_q}(x)}{\rho_{\mathrm{B},q}(x)} < +\infty. \qquad \square$$

Beyond Gaussians

In this section we comment on how the extension discussed in Remark 6.3.2 can be accounted for in the above proofs. Fix $\varepsilon_0 > 0$. From [Aza13] we deduce the existence of a convex $U_\bullet \in \mathscr{C}^2(\mathbb{R}^d)$ with $U_\bullet \leqslant \operatorname{conv} U_\star$ such that

$$\|U_\bullet - \operatorname{conv} U_\star\|_{\mathscr{C}^0(\mathbb{R}^d)} < \varepsilon_0.$$

For any $A \in \mathscr{M}_{d\times d}(\mathbb{R})$ with $\operatorname{spec}(A) \subset (0, +\infty)$, the $\lambda_A > 0$ from (6.3.2) yields

$$\frac{U_\bullet(\lambda_A x)}{U_{\mathrm{B}}(Ax)} \geqslant \frac{\operatorname{conv} U_\star(\lambda_A x) - \varepsilon_0}{U_{\mathrm{B}}(Ax)},$$

and therefore

$$\lim_{\|x\|\to+\infty} \frac{U_\bullet(\lambda_A x)}{U_{\mathrm{B}}(Ax)} = +\infty.$$

We also record that $U_\bullet$ is coercive. Indeed, by (6.3.1) there exists $c_\star > 0$ such that $U_\star(x) \geqslant c_\star\|x\|$ for all $\|x\| > M_\star$. Setting

$$C_\star \coloneqq \sup_{\|x\|\leqslant M_\star} \left(c_\star\|x\| - U_\star(x)\right) < +\infty,$$

we obtain $U_\star(x) \geqslant c_\star\|x\| - C_\star$ for all $x \in \mathbb{R}^d$. Since $x \mapsto c_\star\|x\| - C_\star$ is convex and lies below $U_\star$, it also lies below $\operatorname{conv} U_\star$, hence

$$\operatorname{conv} U_\star(x) \geqslant c_\star\|x\| - C_\star \quad \text{and} \quad U_\bullet(x) \geqslant c_\star\|x\| - (C_\star + \varepsilon_0) \qquad \text{for all } x \in \mathbb{R}^d.$$

We now show that $e^{-U_\bullet}$ satisfies a comparison principle similar in spirit to Lemma 6.5.5.

Lemma 6.5.6. *Let $U : \mathbb{R}^d \to \mathbb{R}$ be convex and $U(x) \to +\infty$ as $\|x\| \to +\infty$. Fix $s \in (0, 1)$.*

1. *For any $k \in [d]$ and any bounded sets $B, B' \subset \mathbb{R}^d$,*

$$\lim_{t\to+\infty} \inf_{(y,y')\in B\times B'} \Big(U(y+te_k) - U(y'+ste_k)\Big) = +\infty. \tag{6.5.10}$$

2. *For any $\beta \in \mathbb{R}^d$,*

$$\lim_{\|x\|\to+\infty} \Big(U(x) - U(sx+\beta)\Big) = +\infty. \tag{6.5.11}$$

Proof. (1) Fix $y \in B$, $y' \in B'$ and define $z := (y' - sy)/(1-s)$. Then

$$y' + ste_k = s(y+te_k) + (1-s)z.$$

By convexity,

$$U(y'+ste_k) \leqslant s\,U(y+te_k) + (1-s)\,U(z),$$

hence

$$U(y+te_k) - U(y'+ste_k) \geqslant (1-s)\big(U(y+te_k) - U(z)\big).$$

As B, B' are bounded, the corresponding set of all z is bounded, so $U(z)$ is uniformly bounded above. On the other hand, $\inf_{y\in B} U(y+te_k) \to +\infty$ as $t \to +\infty$ by coercivity. Taking the infimum over y, y' gives (6.5.10).

(2) Write $sx + \beta = sx + (1-s)\,\frac{\beta}{1-s}$ and apply convexity:

$$U(sx+\beta) \leqslant s\,U(x) + (1-s)\,U\left(\frac{\beta}{1-s}\right),$$

so

$$U(x) - U(sx+\beta) \geqslant (1-s)\Big(U(x) - U\left(\frac{\beta}{1-s}\right)\Big) \to +\infty \qquad \text{as } \|x\| \to +\infty,$$

again by coercivity. □

It remains to outline the adjustments in the proof of Theorem 6.2.1 required to generalize the result. The generalization of Theorem 6.2.3 is completely analogous.

In Step 1, since $\beta_i \in \mathbb{R}^d$ and $\alpha_i > 0$ are fixed for all $i \in [n]$, thanks to (6.3.2) we can find $\underline{\lambda} > 0$ and $\overline{M} > 0$ such that

$$U_\bullet(\underline{\lambda}x) > \max_{1\leqslant i\leqslant n} (U_{\mathrm{B}}(A_i x + \beta_i) - \log\alpha_i) \qquad \text{for all } \|x\| > \overline{M}.$$

(Indeed, for each i apply (6.3.2) to A_i after the change of variables $x \mapsto x + A_i^{-1}\beta_i$. Then increase $\underline{\lambda}$ and use Lemma 6.5.6(2) to absorb the resulting translation inside $U_\bullet(\underline{\lambda}x)$, after enlarging $\overline{M}$ if needed.) We define $\underline{\rho}(x) = \alpha e^{-U_\bullet(\underline{\lambda}x)}$ for all $x \in \mathbb{R}^d$, which satisfies

$$\underline{\rho}(x) < \min_{1\leqslant i\leqslant n} \alpha_i \rho_{\mathrm{B}}(A_i x + \beta_i) \leqslant \rho^\theta_{T/2}(x) \qquad \text{for all} \quad \|x\| > \overline{M}.$$

Set

$$m_K := \operatorname*{ess\,inf}_{[-\overline{M},\overline{M}]^d} \rho^\theta_{T/2} > 0 \qquad \text{and} \qquad M_K := \|e^{-U_\bullet(\underline{\lambda}\cdot)}\|_{L^\infty([-\overline{M},\overline{M}]^d)} < \infty.$$

(Since $\rho^\theta_{T/2}$ is piecewise given by finitely many strictly positive continuous densities on a finite polyhedral partition, its essential infimum on the compact cube is positive.)

Choosing $\alpha \in (0, m_K/M_K)$ yields $\underline{\rho}(x) < \rho^\theta_{T/2}(x)$ for a.e. $x \in \mathbb{R}^d$. (Indeed, $\rho^\theta_{T/2}$ is a finite sum of terms $\alpha_i \rho_{\mathrm{B}}(A_i x + \beta_i)$ and $\rho_{\mathrm{B}}^{-1} \in L^\infty_{\mathrm{loc}}(\mathbb{R}^d)$.)

In Step 2, we take the same controls $\theta_1, \dots, \theta_{2d}$ that give (6.5.8), now with

$$\underline{\rho}_1^{k\ell}(x) = e^{-\frac{\omega T}{4d}} \underline{\rho}\left(\left(I_d + (e^{-\frac{\omega T}{4d}} - 1) e_k e_k^\top\right) x + \beta_{1,k\ell}\right), \qquad \underline{\rho}_2^q(x) = e^{-\frac{\omega T}{4}} \underline{\rho}\left(e^{-\frac{\omega T}{4d}} x + \beta_{2,q}\right)$$

for some $\beta_{1,k\ell}, \beta_{2,q} \in \mathbb{R}^d$. Take $\omega > 4d \log \underline{\lambda}/T$, so that

$$s \coloneqq \underline{\lambda}\, e^{-\frac{\omega T}{4d}} \in (0,1).$$

Then, by Lemma 6.5.6, the corner terms $\underline{\rho}_2^q$ satisfy

$$\frac{\underline{\rho}_2^q(x)}{e^{-U_\bullet(x)}} = e^{-\frac{\omega T}{4}} \alpha \exp\left(U_\bullet(x) - U_\bullet\big(sx + \underline{\lambda}\beta_{2,q}\big)\right) \to +\infty \qquad \text{as } \|x\| \to +\infty,$$

and hence $\underline{\rho}_2^q(x) > e^{-U_\bullet(x)}$ for all $\|x\|$ large enough, uniformly in $q \in \{0,1\}^d$. Similarly, on each strip $\mathscr{S}_{k\ell}$ the transverse coordinates are bounded, and the arguments of $U_\bullet$ in $\underline{\rho}_1^{k\ell}$ have the form $y' + s t e_k$ with y' bounded and $t \to +\infty$ along the unbounded coordinate; thus (6.5.10) yields

$$\underline{\rho}_1^{k\ell}(x) > e^{-U_\bullet(x)}$$

for $x \in \mathscr{S}_{k\ell}$ with $|x^{(k)}|$ large enough, uniformly in $k \in [d]$ and $\ell \in \{0,1\}$. Since there are finitely many $\underline{\rho}_1^{k\ell}$ and $\underline{\rho}_2^q$, there exists $\overline{\overline{M}} \geqslant \overline{M}$ such that

$$\underline{\rho}_T(x) > e^{-U_\bullet(x)} \qquad \text{for all } x \in \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d.$$

Finally, since $U_\bullet \leqslant U_\star$, we have $\rho_\star = e^{-U_\star} \leqslant e^{-U_\bullet}$, and by Lemma 6.5.4 we also have $\underline{\rho}_T(x) < \rho^\theta_T(x)$ for all $x \in \mathbb{R}^d$. Therefore,

$$\rho_\star(x) < \underline{\rho}_T(x) < \rho^\theta_T(x), \qquad \text{in } \mathbb{R}^d \setminus [-\overline{\overline{M}}, \overline{\overline{M}}]^d.$$

Part III

Applications

Overview

Part III presents two applications of the controlled-flow viewpoint developed in **Parts I** and **II**. While both chapters differ in scope—one is an application-driven modeling tool for time-indexed data, the other a theoretical study of a modern architecture where interaction and transport mechanisms are coupled—they share a common guiding principle: continuous-time flows provide both a flexible language to *build* models under practical constraints and an analytical microscope to *understand* qualitative behavior of complex dynamics.

A practical application: time-indexed distributional modeling

Chapter 7 addresses a learning task in which the object of interest is an evolving probability law inferred from samples observed sequentially in time.

This differs from the *static supervised-learning* setting in **Part I**: there one learns a predictor from a fixed dataset such as (I.1), and the integration interval in continuous-depth models represents an *artificial time* scale corresponding to network depth viewed as a continuum of layers. In **Chapter 7**, the data arrive as *snapshots over time* and the modeling goal is explicitly dynamical; accordingly, the integration variable is identified with the *physical time* scale of the measurements.

The chapter is application-driven and illustrates how the controlled-flow perspective informs modeling choices when one must balance robustness, interpretability, and implementability. The motivating example is *continuous glucose monitoring*, where one seeks clinically meaningful patient representations from noisy and irregular measurements. Although the target is distributional, the goal is not generative modeling in the sense of **Part II**, but rather to obtain an interpretable description of how a law changes over time so as to reveal patterns missed by scalar summaries. To this end, **Chapter 7** adopts a finite-dimensional surrogate for distributional dynamics: the evolving law is summarized by a low-dimensional state (mixture weights on the simplex) whose continuous-time evolution is modeled by a neural ODE.

A theoretical application: Transformers via mean field gradient flows

Chapter 8 uses continuous-time dynamics primarily as an analytical tool. At a mathematical level, a large language model specifies an autoregressive probability distribution over sequences of discrete symbols. A *token* is one such symbol drawn from a finite vocabulary (obtained by *tokenization*), and each token is mapped to a vector in $\mathbb{R}^d$ through an embedding layer [Mik+13]. Sequence order is incorporated by adding position-dependent features. This allows one to view the collection of token embeddings at a given depth as an interacting particle system.

A *Transformer* model alternates *self-attention* layers—which induce token interactions—with position-wise *feed-forward* blocks. For tractability, **Chapter 8** focuses on an idealized attention–perceptron backbone in continuous depth and omits masking/cross-attention effects, isolating the core interaction mechanisms. Within this idealization, depth is treated as time and token embeddings evolve as an interacting particle system; recall (10) in The Big Picture. Motivated by common normalization mechanisms (e.g. LayerNorm/RMSNorm), one further formulates the dynamics as taking place on $\mathbb{S}^{d-1}$.

In regimes relevant to very large context lengths, a mean-field limit becomes informative: the empirical distribution of token embeddings is described by a nonlinear continuity equation, which in variational regimes can be interpreted as a (weighted) *Wasserstein gradient flow* of an *interaction energy* and a potential term. The focus is on qualitative questions about the associated energy landscape—stationary states, their localization, and how coupling self-attention with a perceptron block (modeled via a neural ODE) reshapes equilibria. Although the same PDE language appears as in **Part II**, the objective here is different: the continuity equation is not used as a generative mechanism, but as a tool to understand architectural dynamics.

Roadmap of Part III

For navigation, **Part III** consists of the following two chapters.

- **Chapter 7**: Distributional Dynamics Modeling for Digital Health.
- **Chapter 8**: Mean-field Landscape of Transformers.

Chapter 7

Distributional Dynamics Modeling for Digital Health

7.1 Introduction

Learning a probability law from samples is a classical statistical problem, a central ingredient of modern generative modeling [BGC17; Pap+21], and increasingly important in scientific domains where data are functional or irregular. In these settings, the target is often not a single static distribution, but an evolving one: data arrive as snapshot samples from a non-stationary process, creating a need for frameworks that treat probability measures as the primary objects of analysis. This emerging paradigm, known as distributional data analysis [Gho+23; Sza+16], treats distributions as random objects in metric spaces—or aggregates them—for downstream unsupervised and supervised learning tasks.

Biomedical applications are the most prominent use case for this perspective. In digital health, measurements collected by continuous glucose monitoring (CGM) devices, accelerometers, or medical imaging modalities such as functional magnetic resonance imaging (fMRI) are now commonly represented through their empirical distributions, which serve as latent descriptions of the underlying physiological processes [Gho+25; GM24; Mat+25; Mat+22]. These high-frequency physiological time series are often noisy and difficult to align across individuals. Estimating the distribution of these series over specific time windows allows for the construction of patient representations that capture physiology with high precision [Mat+21; MP23; Gho+23]. Recent studies demonstrate that, when used properly, such representations can reveal clinically relevant patterns that traditional scalar biomarkers (like mean glucose or time-in-range) fail to detect [Kat+24; Mat+25; PKG25].

However, modeling the *time evolution* of these distributions poses a significant challenge, particularly when the goal is to obtain a representation that is stable, low-dimensional, and interpretable. Naively extending classical kernel density estimators (KDEs) [CD18] with a temporal dimension suffers from sensitivity to tuning parameters and the curse of dimensionality. Conversely, while deep generative models like normalizing flows [Pap+21] excel at density estimation, they typically operate as "black boxes", making it difficult to extract causal or physiological insights. On the other hand, statistical semiparametric approaches such as Generalized Additive Models for Location, Scale, and Shape (GAMLSS) alleviate some issues [RS05], yet most implementations are constrained to scalar responses and can impose rigid functional forms. Recent multilevel functional frameworks based on functional-quantile representations [MC26] offer interpretability but rely on linear dynamics, which limits their ability to capture complex relationships.

To fill this gap, we propose a time-dependent Gaussian mixture model in which a neural ODE governs the distribution shift smoothly over time. Neural ODEs have been applied to forecasting by

modeling latent states as continuous-time trajectories [Kid+20; RCD19; JB19], and this strategy has proven effective in biomedical applications [Qia+21].

Our modeling choice is to encode each snapshot law by a mixture with a shared dictionary of K-component parameters (means, covariances) and time-dependent mixture weights. In this framework, the "distributional state" is reduced to a weight vector evolving on the simplex Δ^{K-1}, so we obtain a continuous-time model that captures nonlinear relationships while remaining semantically transparent. This transparency is crucial in clinical settings: the shared means can be associated with clinically meaningful values (e.g., reference glycaemic ranges), and the resulting dynamics describe how probability mass flows between these states over time.

We formalize the mathematical problem in **Section 7.2**. In **Section 7.3**, we introduce our proposed two-stage estimator, followed by theoretical guarantees for the snapshot estimation stage in **Subsection 7.3.3**. We then evaluate the performance of our model against competitive baselines on synthetic data of low and moderate dimensions in **Section 7.4**. Finally, **Section 7.5** presents a case study on longitudinal diabetes data, demonstrating the clinical utility of the framework.

7.2 Setup

Let $T > 0$ and let $X_t \in \mathbb{R}^d$ denote a random variable observed at time $t \in [0, T]$. We denote its law by μ_t and (when it exists) its density by ρ_t. The time-indexed distribution function is

$$F_t(x) \;=\; \mathbb{P}\big(X_t \le x\big) \;=\; \int_{-\infty}^{x} \rho_t(r)\,\mathrm{d}r, \qquad x \in \mathbb{R}^d, \tag{7.2.1}$$

where the inequality and the integral are taken component-wise when $d > 1$.

Our target object is the curve $t \mapsto \mu_t \in \mathcal{P}(\mathbb{R}^d)$—equivalently, the density evolution $t \mapsto \rho_t$—from which F_t can be recovered via (7.2.1). In practice, we observe empirical data on a discrete time grid

$$\tau_m \coloneqq \{t_0, \dots, t_m\} \subset [0, T]. \tag{7.2.2}$$

At each $t_i \in \tau_m$, we record $N_i \ge 1$ observations:

$$X_{t_i,1}, \dots, X_{t_i,N_i} \sim \mu_{t_i}. \tag{7.2.3}$$

Two regimes are relevant, depending on the application:
(i) a *cross-sectional* setting, where observations are independent across time steps (and independent and identically distributed within each snapshot); and
(ii) a *longitudinal* setting characterized by temporal or causal dependence.

7.3 Model formulation

We seek an estimator of the curve $t \mapsto \mu_t$ that is continuous in time and provides a compact, interpretable summary of distributional shifts.

To this end, we approximate each snapshot law μ_t by a *Gaussian mixture* of K components with *shared* means and covariances, and *time-dependent* weights. The corresponding density is given by

$$\rho_{\theta(t)}(x) \;\coloneqq\; \sum_{s=1}^{K} \alpha_s(t)\,\mathcal{N}\big(x \mid m_s, \Sigma_s\big), \qquad \alpha(t) \coloneqq \big(\alpha_1(t), \dots, \alpha_K(t)\big) \in \Delta^{K-1},$$

where $m_s \in \mathbb{R}^d$ is the s-th component mean and $\Sigma_s \in \mathscr{S}_d^+(\mathbb{R})$ is the corresponding covariance matrix. The state of the system at time t is fully described by the parameter vector

$$\theta(t) \coloneqq \big(\alpha(t),\, m_1, \ldots, m_K,\, \Sigma_1, \ldots, \Sigma_K\big) \in \Delta^{K-1} \times (\mathbb{R}^d)^K \times (\mathscr{S}_d^+(\mathbb{R}))^K.$$

Note that only the mixing weights $\alpha(t)$ vary continuously in time within the simplex, while the component parameters (m_s, Σ_s) remain static.

This restricted architecture is theoretically sufficient to capture arbitrary distributional evolutions. A classical result of Wiener–Tauberian type [Wie32] implies that finite Gaussian mixtures are dense in the space of integrable functions. Consequently, for any target curve $(\rho_t)_{t\in[0,T]}$ satisfying mild tightness and continuity assumptions (see Proposition 7.3.1), and for any precision $\varepsilon > 0$, there exists an integer $K \geqslant 1$ and a weight trajectory $\alpha(\cdot)$ such that

$$\sup_{t\in[0,T]} \big\|\rho_t - \rho_{\theta(t)}\big\|_{L^1(\mathbb{R}^d)} < \varepsilon.$$

While asymptotic exactness requires $K \to \infty$, in practice we select a moderate K (cf. **Section 7.5**) to retain the interpretability of the model as a flow of mass between a few recognizable states.

Our estimator is built in two stages, detailed below.

7.3.1 Stage 1: Discrete-time MMD fitting

Maximum Mean Discrepancy.

To compare probability measures μ and ν on a common measurable space $\mathcal{X}$ (such as $\mathbb{R}^d$), we use the *maximum mean discrepancy* (MMD) [Gre+12; Mua+17]. This metric is constructed using a *reproducing kernel Hilbert space* (RKHS).

Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *positive definite kernel*—meaning it is symmetric and $\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$ for any finite sequences of $x_i \in \mathcal{X}$ and $c_i \in \mathbb{R}$. By the Moore–Aronszajn theorem [Aro50], the kernel k uniquely induces an RKHS—denoted by $\mathcal{H}$—of functions $f\colon \mathcal{X} \to \mathbb{R}$. This is a Hilbert space where, for every $x \in \mathcal{X}$, the point evaluation functional $f \mapsto f(x)$ is continuous. By the Riesz representation theorem, such continuity guarantees the existence of a unique *feature map* $k(\cdot, x) \in \mathcal{H}$ satisfying $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

Assuming the kernel is bounded, i.e., $\sup_{x\in\mathcal{X}} k(x,x) < \infty$, the linear functional $f \mapsto \mathbb{E}_\mu[f]$ is also continuous on $\mathcal{H}$ for any $\mu \in \mathcal{P}(\mathcal{X})$. Applying the Riesz representation theorem a second time, this functional can be uniquely represented by an element $m_\mu \in \mathcal{H}$, called the *kernel mean embedding* of μ. This element is given by the integral:

$$m_\mu \coloneqq \int k(\,\cdot\,, x)\, \mathrm{d}\mu(x) \in \mathcal{H},$$

which by construction satisfies $\langle f, m_\mu\rangle_{\mathcal{H}} = \mathbb{E}_\mu[f]$ for all $f \in \mathcal{H}$.

Formally, the MMD acts as a pseudometric on $\mathcal{P}(\mathcal{X})$, where the squared distance between two measures is evaluated as the squared $\mathcal{H}$-norm of the difference between their respective embeddings:

$$\mathrm{MMD}^2(\mu,\nu) \coloneqq \|m_\mu - m_\nu\|_{\mathcal{H}}^2 = \mathbb{E}\big[k(X,X')\big] + \mathbb{E}\big[k(Y,Y')\big] - 2\,\mathbb{E}\big[k(X,Y)\big]. \tag{7.3.1}$$

Here, the expectations are taken over independent random variables, such that X and X' are drawn independently from μ, and Y and Y' are drawn independently from ν.

If the map $\mu \mapsto m_\mu$ is injective—equivalently, if the kernel k is *characteristic*—the MMD defines a valid metric on $\mathcal{P}(\mathcal{X})$, satisfying $\mathrm{MMD}(\mu,\nu) = 0 \iff \mu = \nu$; see [SFL11; Sej+13].

We fix $\mathcal{X} = \mathbb{R}^d$ and employ the *Gaussian kernel*,

$$k(x, y) := \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right), \tag{7.3.2}$$

where $\sigma > 0$ is a bandwidth parameter. The Gaussian kernel is characteristic and bounded. In practice, the choice of σ strongly impacts the sensitivity of the metric to different scales. We use the *median heuristic* [GJK17], setting σ to the median of the pairwise Euclidean distances computed on the pooled dataset (aggregating samples across all time steps).

Intuition. The MMD measures how far apart μ and ν are in the high-dimensional feature space induced by k. Unlike likelihood-based methods, it does not require the target distribution to admit a density and is robust to finite-sample noise [AG24]. This makes it ideal for comparing the empirical distribution of snapshots $\{X_{t_i,j}\}_{j=1}^{N_i}$ from (7.2.3) with our continuous mixture model.

Discrete-time fitting.

For each $t_i \in \tau_m$, let μ_{t_i} denote the (unknown) target measure with density ρ_{t_i}. We define its empirical measure, built from the N_i samples $\{X_{t_i,j}\}_{j=1}^{N_i}$ introduced in (7.2.3), as

$$\widehat{\mu_{t_i}} := \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{X_{t_i,j}}.$$

We fit a Gaussian mixture model at each time step t_i, taking the form

$$\rho_{\theta_i}(x) = \sum_{s=1}^{K} \alpha_{i,s}\, \mathcal{N}(x \mid m_s, \Sigma_s), \qquad \alpha_i = (\alpha_{i,1}, \dots, \alpha_{i,K})^\top \in \Delta^{K-1}. \tag{7.3.3}$$

The optimal parameters are estimated by minimizing the MMD between the empirical distribution and our model. Formally, we minimize $\mathrm{MMD}^2(\widehat{\mu}_{t_i}, \mu_{\theta_i})$, where μ_{θ_i} is the probability measure with density ρ_{θ_i}.

At each $t_i \in \tau_m$, we use a Gaussian kernel k_i (such as (7.3.2)), with the bandwidth $\sigma_i > 0$ chosen via the median heuristic on the dataset $\{X_{t_i,j}\}_{j=1}^{N_i}$. By expanding the squared MMD, the objective function reduces to a convenient quadratic form in the mixing weights α_i:

$$\mathrm{MMD}^2(\widehat{\mu_{t_i}}, \mu_{\theta_i}) = \alpha_i^\top I_i\, \alpha_i - 2\, \alpha_i^\top J_i + C_i, \tag{7.3.4}$$

where C_i depends solely on the empirical data and can therefore be omitted from the minimization, while the matrix $I_i \in \mathscr{M}_{K\times K}(\mathbb{R})$ and the vector $J_i \in \mathbb{R}^K$ admit closed-form expressions for their entries:

$$(I_i)_{s,r} = \frac{\sigma_i^d}{\sqrt{\det(\Sigma_s + \Sigma_r + \sigma_i^2 I_d)}} \exp\Big(-\frac{1}{2}(m_s - m_r)^\top (\Sigma_s + \Sigma_r + \sigma_i^2 I_d)^{-1}(m_s - m_r)\Big), \tag{7.3.5}$$

$$(J_i)_s = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\sigma_i^d}{\sqrt{\det(\Sigma_s + \sigma_i^2 I_d)}} \exp\Big(-\frac{1}{2}(X_{t_i,j} - m_s)^\top (\Sigma_s + \sigma_i^2 I_d)^{-1}(X_{t_i,j} - m_s)\Big). \tag{7.3.6}$$

Optimization. We minimize (7.3.4) using the following alternating scheme:

1. Initialization. We run K-means clustering [Jai10] on the pooled dataset $\bigcup_{i,j} X_{t_i,j}$ to initialize the parameters. The means $\{m_s\}_{s=1}^K$ are the centroids of the identified clusters; $\{\Sigma_s\}_{s=1}^K$ are the

empirical covariance matrices of the points within each cluster; and the initial weights $\{\alpha_{i,s}\}_{s=1}^{K}$ correspond to the proportion of data points at time t_i assigned to cluster s.

2. Local update. For each t_i, keeping means and covariances fixed, we update the weights $\alpha_i^* \in \Delta^{K-1}$ via the quadratic program:

$$\alpha_i^* = \underset{\alpha \in \Delta^{K-1}}{\operatorname{argmin}} \left\{ \alpha^\top I_i \, \alpha - 2 \, \alpha^\top J_i + \sum_{s=1}^{K} \lambda_s \, \alpha_s^2 \right\}, \tag{7.3.7}$$

where $\lambda = (\lambda_s)_{s=1}^{K} \in \mathbb{R}^K$ is a vector of ridge hyperparameters that improves numerical conditioning and stabilizes the solution when Gaussian components become nearly collinear in the RKHS feature space.

3. Global update. We update the shared component parameters m_s and Σ_s iteratively via Adam gradient descent on the full MMD objective, keeping the current weights α_i^* fixed.

Output. Upon convergence of this scheme, we obtain a sequence of weights $\{\alpha_i^*\}_{i=0}^{m} \subset \Delta^{K-1}$ alongside a set of globally shared component means $\{m_s\}_{s=1}^{K}$ and covariances $\{\Sigma_s\}_{s=1}^{K}$. Consequently, the discrete-time distributional state at any time $t_i \in \tau_m$ is fully specified by

$$\rho_{\theta_i^*}(x) = \sum_{s=1}^{K} \alpha_{i,s}^* \, \mathcal{N}(x \mid m_s, \Sigma_s).$$

7.3.2 Stage 2: Continuous-time fitting

Once the discrete-time weights $\{\alpha_i^*\}_{i=0}^{m}$ are fitted, we use a continuous-time model for their evolution. We first solve an unconstrained controlled flow in $\mathbb{R}^K$,

$$\begin{cases} \dot{z}(t) &= v_\phi\big(z(t), t\big), \qquad t \in [0, T], \\ z(0) &= \alpha_0^*, \end{cases} \tag{7.3.8}$$

where $v_\phi : \mathbb{R}^K \times [0, T] \to \mathbb{R}^K$ is a multilayer perceptron. Following the discussion in **Chapter 4**—see in particular **Subsection 4.4.2** and **Section 4.5**—, we take v_ϕ to be a *semi-autonomous* neural ODE of the form ($\mathcal{V}_{\mathrm{sa}}^{+}$), as it provides a flexible (cell-wise controllable in the sense of Proposition 4.4.3) yet parameter-efficient way to model temporal evolutions. Then, we map $z(t)$ to valid mixture weights by simplex normalization:

$$\alpha(t) \;\coloneqq\; \frac{z(t)_+}{\langle \mathbf{1}, z(t)_+ \rangle} \in \Delta^{K-1}, \tag{7.3.9}$$

where $z_+ \coloneqq z \vee 0$. (With the convention that if $\langle \mathbf{1}, z(t)_+ \rangle = 0$, we replace $z(t)_+$ by $z(t)_+ + \varepsilon \mathbf{1}$ for a small $\varepsilon > 0$). Architecture and solver hyperparameters are reported in Table 7.1.

The parameters ϕ are optimized by matching the ODE predictions to the fitted weights α_i^*:

$$\mathcal{L}(\phi) = \sum_{i=0}^{m} \|\alpha(t_i; \phi) - \alpha_i^*\|^2 + \nu \|\phi\|^2, \tag{7.3.10}$$

where $\alpha(t_i; \phi)$ denotes the solution obtained by integrating (7.3.8) and applying (7.3.9) at time t_i, and $\nu \geqslant 0$ is a ridge hyperparameter.

Permutation symmetry. Because (m_s, Σ_s) are shared over time and kept fixed after the global fit, component labels are anchored. The neural ODE stage only evolves $\alpha_s(t)$, so trajectories cannot exchange labels.

Why MMD? The choice of the MMD metric in the first stage is motivated by several theoretical and practical advantages:
(i) *closed-form expressions:* with a Gaussian kernel, MMD between an empirical measure and a Gaussian mixture admits closed-form terms, enabling stable and fast updates; and
(ii) *robustness:* MMD is empirically robust under misspecification [CA22; AG24; Gao+21; Alq+23].

Why a neural ODE? Our object of interest is a *continuous-time* family of distributions $t \mapsto \rho_t$. In the case study, CGM measurements are irregular and not aligned across participants. Stage 1 provides snapshot estimates of the distributional state at a finite set of observation times. Stage 2 then seeks a continuous-time interpolation that is coherent in time and robust to estimation noise. This choice has several concrete advantages:
(i) *grid-free evaluation:* the density can be evaluated at arbitrary times without re-gridding or aligning subjects;
(ii) *adaptive computation:* the numerical integrator can allocate more steps where dynamics are fast and fewer where they are slow;
(iii) *parameter efficiency:* a single vector field represents transformations of effectively arbitrary *depth* in time; and
(iv) *smoothness and regularization:* the ODE imposes a low-dimensional dynamical prior that couples the discrete snapshots into a smooth trajectory, mitigating noise from finite-sample fitting.

Importantly, we do not use the neural ODE *solely* as a black-box sequence predictor of raw CGM traces (for which discrete architectures can be competitive [Wan24; Wu+23]). Rather, we leverage it to simultaneously model the smooth evolution of the distributional state, regularize the discrete estimates, and enable continuous-time forecasting.

7.3.3 Theoretical background

We provide theoretical guarantees for the discrete-time MMD fitting (Stage 1), with the proofs deferred to **Section 7.7**. Our first result establishes the representational capacity of the model, showing that a finite mixture with shared components is sufficient to approximate continuous temporal evolutions.

Proposition 7.3.1 (Universality). *Let $\{\rho_t\}_{t\in[0,T]} \subset L^1(\mathbb{R}^d)$ be probability densities. Assume:*

(1) For every $\eta > 0$, there exists $R < \infty$ such that

$$\int_{\{\|x\|>R\}} \rho_t(x)\,\mathrm{d}x < \eta \qquad \text{for all } t \in [0,T].$$

(2) The family of densities is uniformly translation-continuous:

$$\lim_{\|h\|\to 0} \sup_{t\in[0,T]} \|\rho_t(\cdot + h) - \rho_t(\cdot)\|_{L^1(\mathbb{R}^d)} = 0.$$

Then, for every $\varepsilon > 0$, there exist an integer $K \geqslant 1$, a shared variance $\sigma_0^2 > 0$, a set of component

means $\{m_s\}_{s=1}^K \subset \mathbb{R}^d$*, and a weight mapping* $\alpha : [0,T] \to \Delta^{K-1}$*, such that*

$$\sup_{t\in[0,T]} \left\| \rho_t(\cdot) - \sum_{s=1}^{K} \alpha_s(t)\, \mathcal{N}\left(\cdot \mid m_s, \sigma_0^2 I_d\right) \right\|_{L^1(\mathbb{R}^d)} < \varepsilon.$$

Furthermore, if the mapping $t \mapsto \rho_t$ *is continuous from* $[0,T]$ *into* $L^1(\mathbb{R}^d)$*, then the weight trajectory* $\alpha(\cdot)$ *can be chosen to be continuous.*

Proposition 7.3.1 formally justifies our shared-component architecture: one may approximate a time-varying target law by fixing the spatial parameters and letting only the weights absorb the temporal variation. Since our fitted model allows for general covariances (Σ_s) rather than a common isotropic variance $\sigma_0^2 I_d$, it is at least as expressive as the theoretical construction.

Having established representational capacity, our second result quantifies the statistical stability of the snapshot fitting procedure.

To formalize this, recall the matrix I_i and the empirical vector J_i introduced in (7.3.5) and (7.3.6). Observe that the entries of J_i are empirical averages $(J_i)_s = \frac{1}{N_i}\sum_{j=1}^{N_i} J_{i,s}(X_{t_i,j})$, where the underlying function $J_{i,s}\colon \mathbb{R}^d \to \mathbb{R}$ is the kernel-Gaussian convolution:

$$J_{i,s}(x) \;:=\; \int k_i(x,y)\,\mathcal{N}(y \mid m_s, \Sigma_s)\,\mathrm{d}y.$$

We define the population analogue $J_i^\circ \in \mathbb{R}^K$ with entries $(J_i^\circ)_s := \mathbb{E}[J_{i,s}(X_{t_i,1})]$, and the corresponding population minimizer:

$$\alpha_i^\circ \in \operatorname*{argmin}_{\alpha\in\Delta^{K-1}} \left\{ \alpha^\top I_i\, \alpha - 2\,\alpha^\top J_i^\circ + \sum_{s=1}^{K} \lambda_s \alpha_s^2 \right\}.$$

With these population quantities in place, we can state the concentration bounds.

Theorem 7.3.2 (Finite-sample stability). *Fix a time step* $t_i \in \tau_m$ *from the observation grid defined in* (7.2.2)*. Consider the quadratic program* (7.3.7) *whose empirical solution* α_i^* *is based on the independent observations* $(X_{t_i,j})_{j=1}^{N_i}$ *introduced in* (7.2.3)*.*

Assume the minimum ridge penalty satisfies $\lambda_{\min} := \min_{1\leqslant s\leqslant K} \lambda_s > 0$*, and that the kernel is bounded such that* $0 \le k_i(x,y) \le 1$*. Then, for any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, the estimation error at time* t_i *is bounded by*

$$\|\alpha_i^* - \alpha_i^\circ\| \;\le\; \frac{1}{\lambda_{\min}} \sqrt{\frac{K\log(2K/\delta)}{2N_i}}.$$

Consequently, applying a union bound yields a uniform guarantee across the entire observation grid: with probability at least $1-\delta$*,*

$$\max_{0\le i\le m} \|\alpha_i^* - \alpha_i^\circ\| \;\le\; \frac{1}{\lambda_{\min}} \sqrt{\frac{K\log\big(2K(m+1)/\delta\big)}{2N_*}}, \qquad \text{where } N_* = \min_{0\le i\le m} N_i.$$

Theorem 7.3.2 makes the role of the ridge terms explicit: $\lambda_{\min} > 0$ enforces strong convexity of the snapshot objective, yielding a Lipschitz dependence of the optimal weights on the empirical vector J_i. Since each coordinate of J_i is an average of bounded random variables, standard concentration bounds give the stated non-asymptotic rate. Notably, this convergence depends on the simplex dimension K rather than the ambient dimension d.

Remark 7.3.3. Fix the shared component parameters $(m_s, \Sigma_s)_{s=1}^K$ and write, for any $\alpha \in \Delta^{K-1}$,

$$\rho_\alpha(x) \coloneqq \sum_{s=1}^K \alpha_s \mathcal{N}(x \mid m_s, \Sigma_s), \qquad F_\alpha(x) \coloneqq \int_{(-\infty, x]} \rho_\alpha(r)\, \mathrm{d}r.$$

Since the mapping $\alpha \mapsto \rho_\alpha$ is affine, we have

$$\|\rho_{\alpha_i^*} - \rho_{\alpha_i^\circ}\|_{L^1(\mathbb{R}^d)} \le \|\alpha_i^* - \alpha_i^\circ\|_1 \le \sqrt{K}\, \|\alpha_i^* - \alpha_i^\circ\|,$$

and for any $x \in \mathbb{R}^d$,

$$|F_{\alpha_i^*}(x) - F_{\alpha_i^\circ}(x)| \le \|\alpha_i^* - \alpha_i^\circ\|_1.$$

Combining Theorem 7.3.2 with Remark 7.3.3 yields an explicit transfer of the $N_i^{-1/2}$ weight concentration to L^1-errors on densities and pointwise errors on cumulative distribution functions. At each snapshot, these downstream errors scale as $\mathcal{O}\big(K/(\lambda_{\min}\sqrt{N_i})\big)$.

7.4 Simulations

We benchmark finite-sample performance against state-of-the-art baselines. The code is available at [Álv25].

Unlike many competitors, our approach explicitly prioritizes interpretability through time-varying functions $\alpha_s(\cdot)$ for $s \in \{1, \dots, K\}$. The results show that this emphasis on interpretability does not compromise accuracy.

Data-generating process. Fix $T = 1$. The target density is a three-component Gaussian mixture with time-varying means and a common time-varying variance:

$$\rho_t(x) = \frac{1}{3} \sum_{s=1}^3 \mathcal{N}\big(x \mid m_s(t),\, \sigma^2(t)\, I_d\big), \qquad (t, x) \in [0, 1] \times \mathbb{R}^d, \tag{7.4.1}$$

where the variance evolves as $\sigma^2(t) = 1 + t$, and the means are

$$m_1(t) = -2 + 20t, \qquad m_2(t) = 16t, \qquad m_3(t) = 5 + 6t. \tag{7.4.2}$$

For $d \ge 2$, each $m_s(t) \in \mathbb{R}^d$ has identical coordinates given by (7.4.2). This data-generating process design captures both multimodal and unimodal regimes over time.

We evaluate the models on the regular grid $t_i = i/10$ for $i = 0, \dots, 10$ (yielding $m + 1 = 11$ time points) and generate $B = 100$ independent replicates. At each $t_i \in \tau_m$, we draw $N_i = n$ independent observations, with sample sizes

$$n \in \{20,\ 50,\ 100,\ 200,\ 300,\ 500\}.$$

This simulation scenario is intentionally more general than the working model in **Section 7.3**, because both the component means and the variance vary with time. It is included to evaluate how well the method approximates smoothly evolving distributions beyond the ideal shared-dictionary setting.

Competitors

We compare the proposed estimator against three baselines for estimating the time-indexed density ρ_t from the samples introduced in (7.2.3):
(i) a univariate generalized additive model for location, scale, and shape (GAMLSS) [RS05];
(ii) a time-conditional kernel density estimator (KDE) [Sil86; CD18; Tsy08]; and
(iii) a conditional masked autoregressive flow (MAF)[PPM17; Pap+21].

Hyperparameters for KDE and MAF are reported in Tables 7.2 and 7.3.

Generalized additive models for location, scale, and shape. We fit a univariate Gaussian distributional regression model with time-varying mean and variance,

$$X_t \sim \mathcal{N}\big(m(t),\, \sigma^2(t)\big),$$

where $m(\cdot)$ and $\log \sigma(\cdot)$ are modeled as smooth spline functions of t and estimated using the `gamlss` package in R [RS05]. We include this baseline only for $d = 1$, as multivariate extensions would require additional strong modeling assumptions on the dependence structure.

Time-conditional kernel density estimator. At each observed time $t_i \in \tau_m$, we estimate ρ_{t_i} using a KDE:

$$\rho_{t_i}^{\mathrm{KDE}}(x) = \frac{1}{N_i} \sum_{j=1}^{N_i} K_{h_i}\left(x - X_{t_i,j}\right), \qquad K_h(z) = h^{-d} K(z/h),$$

using a Gaussian kernel K and Scott's bandwidth rule for h_i [Sil86; CD18]. For intermediate times $t \in [t_i, t_{i+1}]$, we linearly interpolate the endpoint densities:

$$\rho_t^{\mathrm{KDE}}(x) \coloneqq (1-\lambda)\rho_{t_i}^{\mathrm{KDE}}(x) + \lambda \rho_{t_{i+1}}^{\mathrm{KDE}}(x), \qquad \lambda = \frac{t - t_i}{t_{i+1} - t_i}.$$

As a fully nonparametric method, KDE is inherently affected by the curse of dimensionality [Tsy08].

Masked autoregressive flow. We model ρ_t with a conditional normalizing flow $\rho_\theta(x \mid t)$ built from an invertible map $x = \Phi^\theta(u; t)$ with base noise $u \sim \mathcal{N}(0, I_d)$. The conditional density is given by the change-of-variables formula:

$$\rho_\theta(x \mid t) = \mathcal{N}\left((\Phi^\theta)^{-1}(x;t) \mid 0, I_d\right) \left|\det \nabla_x (\Phi^\theta)^{-1}(x;t)\right|.$$

In MAF [PPM17], the inverse map is autoregressive (implemented via masked networks), rendering the Jacobian triangular and the log-likelihood tractable. Parameters θ are learned via conditional maximum likelihood over the full sample $\{(t_i, X_{t_i,j})\}_{i,j}$ [Pap+21].

7.4.1 Results

We report results for a low-dimensional ($d = 1$) and a higher-dimensional setting ($d = 10$).

Low dimension ($d = 1$)

Figure 7.1 shows that our model is highly competitive with the baselines. While MAF can achieve slightly smaller L^2 errors in certain sample-size regimes, our approach remains accurate while simultaneously providing directly interpretable weight trajectories $\alpha(t)$.

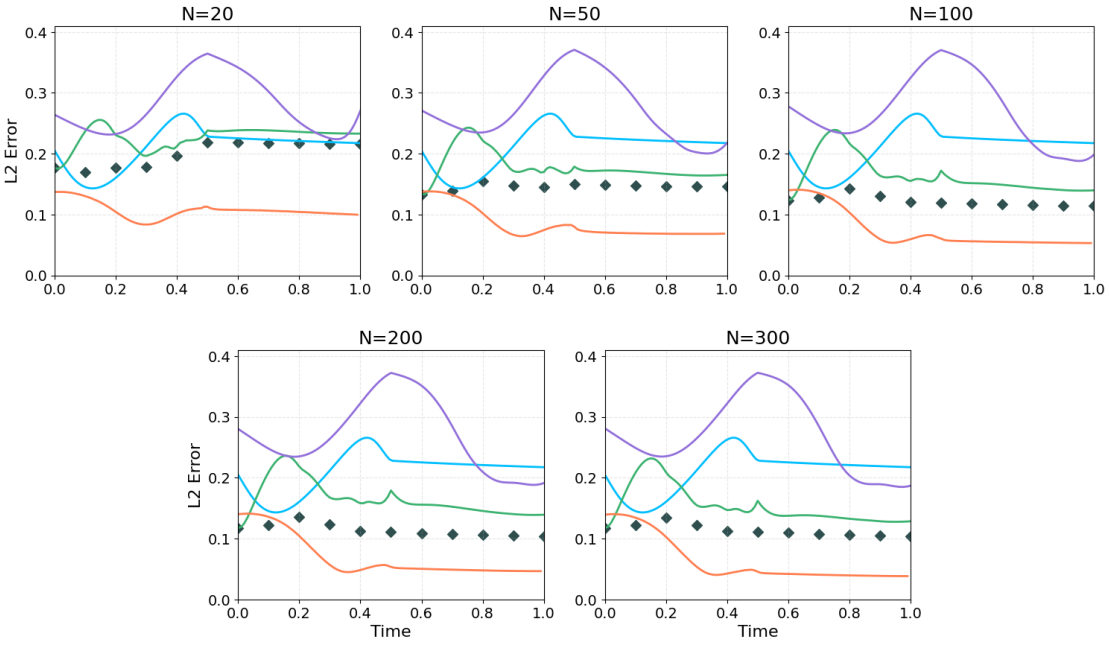


Figure 7.1: Pointwise L^2-error over time for $d = 1$. We compare our model, KDE, MAF, and GAMLSS. Curves represent averages over 100 independent runs. Errors corresponding to the discrete-time MMD stage are shown as ◆.

High dimension ($d = 10$)

Figure 7.2 demonstrates that our model outperforms the alternatives. The KDE performs the worst, which is consistent with its well-known sensitivity to high dimensionality [Tsy08].

7.4.2 Implementation details

We report the main hyperparameters used in our experiments. Unless stated otherwise, we reuse the same settings across simulations and the diabetes case study.

Table 7.1: Hyperparameters of the proposed two-stage estimator.

Step 1: per-time MMD **mixture fit**		**Step 2: time-series neural ODE fit**	
Parameter	**Value**	**Parameter**	**Value**
Mixture comp. (K)	10 (sim), 3 (case)	MLP	2 layers, 64, `tanh`
Kernel bandwidth	median heuristic	Integration time T	1.0
Ridge reg. (λ)	10^{-2}	Solver step size	0.01
Adam learning rate	10^{-3}	Adam learning rate	10^{-3}
Gradient iterations	50 / 20	Max epochs	2000
		Ridge reg. (ν)	10^{-10}

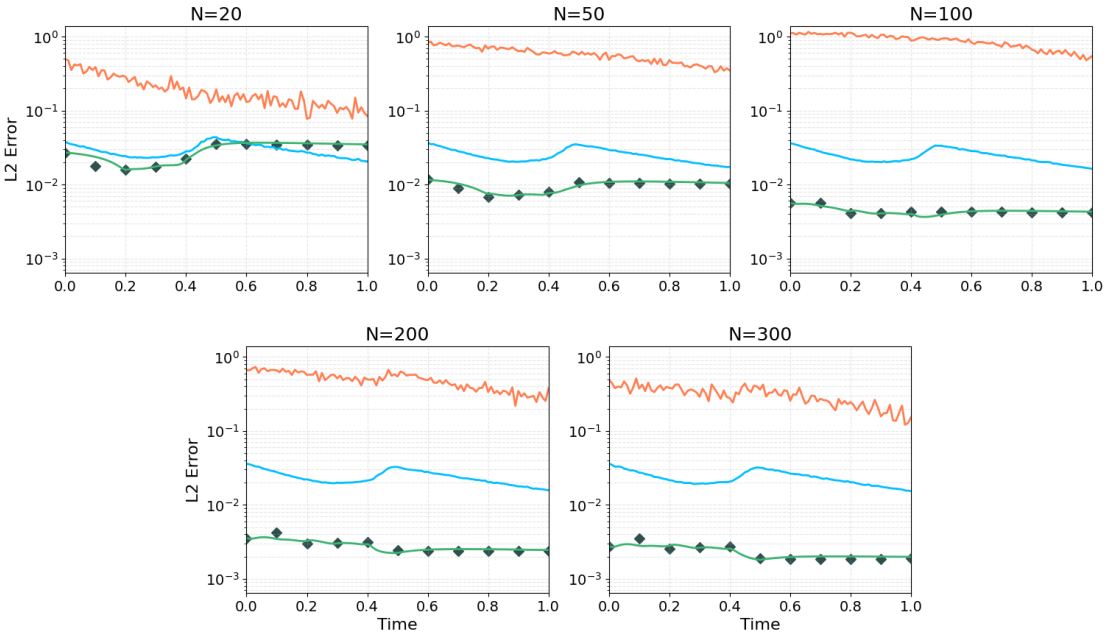


Figure 7.2: Pointwise L^2-error over time for $d = 10$. We compare our model, KDE and MAF. Curves represent averages over 100 independent runs. Errors corresponding to the discrete-time MMD stage are shown as ◆.

7.5 Medical case study

From a distributional data analysis perspective, this study aims to demonstrate the explainability advantages of our novel probabilistic framework for detecting distributional shifts in longitudinal clinical data.

Data description and scientific question

Our study is motivated by the Juvenile Diabetes Research Foundation (JDRF) Continuous Glucose Monitoring randomised controlled trial [Juv08; Juv09].[1] A total of 451 participants (adults and children) with type 1 diabetes mellitus (T1DM) were randomised to either a *treatment* arm—

[1] Publicly available at https://public.jaeb.org/datasets/diabetes.

Table 7.2: Normalizing-flow baseline (MAF): architecture and training hyperparameters.

Architecture			
Parameter	**Value**	**Parameter**	**Value**
Context embedding MLP	128	Blocks (Coupling/AR)	6
Flow transform hidden	64	Stacked blocks	2
Training			
Parameter	**Value**	**Parameter**	**Value**
Epochs	50	Batch size	64
Optimizer	Adam	Learning rate	10^{-3}

Table 7.3: Time-conditional KDE baseline: hyperparameters.

General Settings			
Parameter	**Value**	**Parameter**	**Value**
Context embedding MLP	128	Bandwidth selection	`scott`

receiving continuous glucose monitoring (CGM) data and education—or a *control* arm relying on standard self-monitoring without CGM feedback.

For data-quality reasons, we restrict our analysis to the $n = 443$ individuals with a maximum monitoring period of 26 weeks. Our primary clinical question is whether continuous glucose monitor (CGM) use (*treatment*) versus usual care (*control*) leads to different individual longitudinal glucose trajectories over the course of the trial. The continuous-time nature of our model is ideally suited to address this question because it allows us to characterize how the entire glucose distribution evolves over the full follow-up period.

Modeling the marginal density of glucose time series

Let $G_p(t)$ denote the CGM reading for participant $p \in [n]$ at (calendar) time $t \in [0, T]$. In the JDRF trial, CGM is recorded approximately every five minutes under free-living conditions; in practice, measurements may be missing or corrupted, so the number of available readings varies across participants and over time. After standard cleaning (discarding missing/failed sensor values), we aggregate the remaining measurements into consecutive non-overlapping *ten-day windows*. To facilitate comparison across subjects with different follow-up lengths, we rescale the study time to the unit interval $[0, 1]$ and denote by $0 = t_0 < t_1 < \cdots < t_m = 1$ the endpoints of the ten-day windows.

Within window $i \in [m]$, we collect the available glucose readings for participant p as

$$G_{p,i,1}, \ldots, G_{p,i,n_{p,i}},$$

where $n_{p,i}$ depends on missingness and device wear time. Following the *glucodensity* framework, which represents CGM profiles through the empirical distribution of glucose values over short windows (typically 10–14 days) [Mat+21; MC26], we assume that the glucose process is *approximately stationary within each window*. Consequently, the window observations $\{G_{p,i,\ell}\}_{\ell=1}^{n_{p,i}}$ can be treated as approximately i.i.d. draws from a window-specific marginal distribution with CDF

$$F_{t_i}^{(p)}(x) \;=\; \mathbb{P}\big(G_{p,i} \le x\big),$$

where $G_{p,i}$ denotes a glucose reading sampled within window i, and associated density $\rho_{t_i}^{(p)}$. Our goal is to estimate how these subject-specific distributions evolve across windows, and to compare the resulting distributional dynamics between treatment and control arms.

For computational and comparison purposes we assume that, for all $p \in [n]$, the marginal density can be well-approximated by a shared-anchor Gaussian mixture with participant-specific, time-dependent weights,

$$\rho_{\theta(t)}^{(p)}(x) = \sum_{s=1}^{K} \alpha_s^{(p)}(t)\, \mathcal{N}\big(x \mid m_s, \sigma_s^2\big). \qquad t \in [0, 1].$$

To ensure comparability, we impose that the component parameters $(m_s, \sigma_s^2)_{s\in[K]}$ are shared globally across all individuals; only the time-dependent mixing weights $\alpha_s^{(p)}(t)$ vary by participant.

In particular, Stage 1 yields discrete-time estimates of the weights at the window endpoints t_i, and Stage 2 provides a continuous-time trajectory $t \mapsto \boldsymbol{\alpha}^{(p)}(t)$ by smoothing/interpolating these snapshots.

Choice of the number of components. We set $K = 3$ to encode clinically meaningful phenotypes of glycaemic control (good, intermediate, poor). The same three components are used for all participants to enable cross-subject comparability; accordingly, $(m_s, \sigma_s^2)_{s=1,2,3}$ are shared across individuals, whereas only the time-varying mixture weights $\alpha_s^{(p)}(t)$ are subject-specific. Interpreting the trajectories of $\alpha_s^{(p)}(t)$ within these three reference groups yields a clinically transparent summary of distributional dynamics.

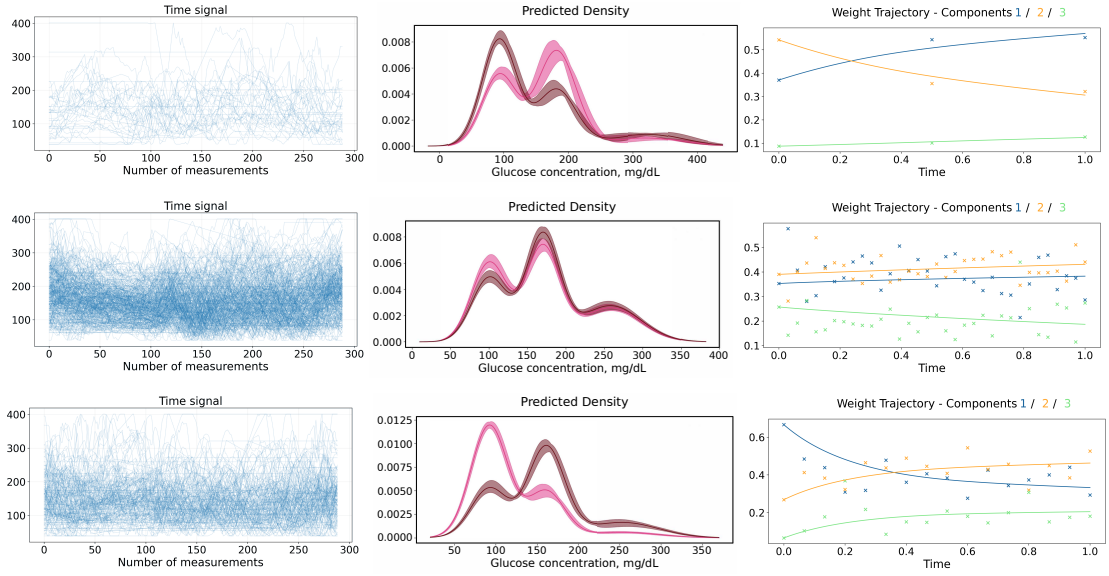


Figure 7.3: Three representative subjects (IDs 13, 62, and 377, one on each row). **Left:** the participant's full time series, as glucose concentration (mg/dL) vs. time slot. **Middle**: initial (pink) and final (dark red) MMD-fitted densities with 95% bootstrap confidence bands. **Right**: weight trajectories $\alpha_s^{(p)}(t)$ from the neural ODE (Step 2); endpoints match the middle-column fits.

Table 7.4: Calibrated global statistics for univariate and bivariate data.

Univariate ($d = 1$): Fitted means and scalar variances			
Parameter	$s = 1$	$s = 2$	$s = 3$
m_s	97.81	167.84	259.37
σ_s^2	537.72	501.04	1737.04
Bivariate ($d = 2$): Fitted means and covariance matrices			
Parameter	$s = 1$	$s = 2$	$s = 3$
m_s	(97.81, 0.58)	(167.66, 0.21)	(258.87, −0.059)
Σ_s	$\begin{bmatrix} 537.75 & -27.01 \\ -27.01 & 124.26 \end{bmatrix}$	$\begin{bmatrix} 501.01 & 1.41 \\ 1.41 & 63.62 \end{bmatrix}$	$\begin{bmatrix} 1737.04 & -18.21 \\ -18.21 & 208.75 \end{bmatrix}$

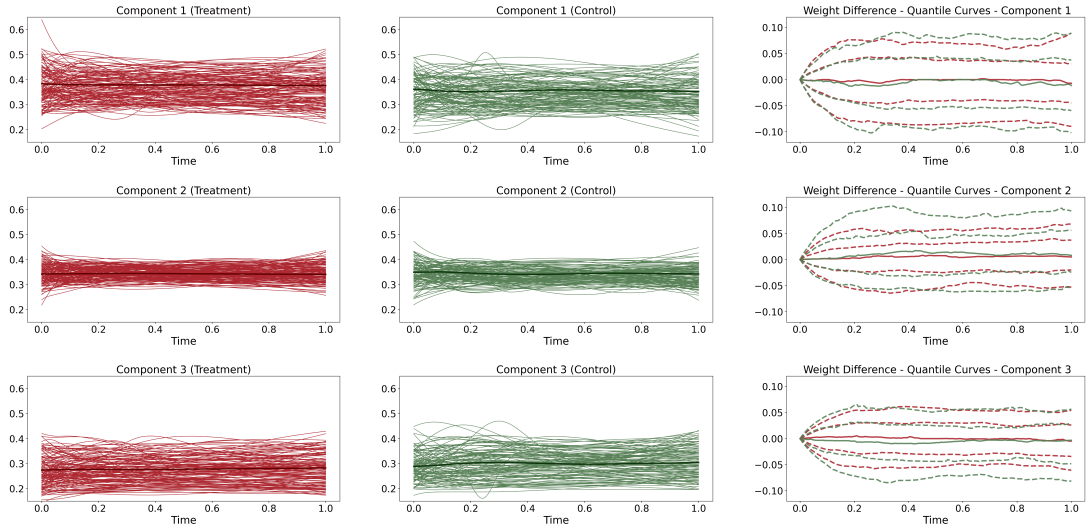


Figure 7.4: Component-wise weight dynamics by arm (components $s = 1, 2, 3$ in rows from top to bottom). **Columns 1-2:** individual curves $\alpha_s^{(p)}(t)$ (thin) and mean (thick) for treatment (red, **left**) and control (green, **right**) groups. **Column 3**: $(0.1, 0.25, 0.5, 0.75, 0.9)$-quantile trajectories of the process $Z_s^{(p)}$ defined in (7.5.1) across subjects for each of the two groups and by arm.

Table 7.4 reports the global mean and covariance of each Gaussian component. From a clinical perspective, these parameters were selected to represent three clearly differentiated glycaemic states: (i) well-controlled diabetes condition, (ii) suboptimal diabetes control, and (iii) poorly controlled diabetes. By computing, at every time, the proportion of participants falling within these three archetypes [CB94] (or phenotypes), we obtain a dynamic, standardized index that summarizes the longitudinal evolution of glycaemic profiles for every subject in the trial.

Figure 7.3 shows the three representative day-to-day CGM trajectories together with their estimated marginal densities, highlighting the intrinsic complexity of the continuous-time series. In the left panel we observe substantial intra-day variability, suggesting that an analysis at the day level is advisable. To make subjects with different visit schedules comparable, time was rescaled to the unit interval $[0, 1]$. We then focus on the subject-specific weight functions $\alpha_s^{(j)}(t)$.

For the three selected participants, the first subject shows an improvement (density shifts left and weight $\alpha_1^{(p)}(t)$ increases), the second remains relatively stable, and the third worsens (density shifts right and weight $\alpha_1^{(p)}(t)$ decreases). Importantly, some patients deteriorate over time even in the treatment arm, consistent with previous findings that only a subset of this cohort responds to the intervention. We also observe that certain participants exhibit apparent improvements in CGM metrics in the control group; however, these changes may reflect unrecorded external interventions—such as additional insulin dosing—rather than the effect of the study treatment. Moreover, despite randomization, baseline characteristics are not perfectly balanced between control and treatment groups (e.g. the treatment arm began with better glycaemic control) [MC26; Juv10].

Figure 7.4 contrasts the estimated mixture weights for the treatment and control arms. Participants in the treatment group devote a larger share of time to phenotype 1 across the entire follow-up, reflecting superior glycaemic control at all assessment points, including baseline. To investigate

individual temporal dynamics, we examined the centered trajectories.

$$Z_s^{(p)}(t) = \alpha_s^{(p)}(t) - \alpha_s^{(p)}(0), \qquad s \in [3] \quad \text{and} \quad p \in [n]. \tag{7.5.1}$$

For each arm, we plot the selected quantiles of the $Z_s^{(p)}(\cdot)$ process. For phenotype 1, the quantile curves are nearly identical—once good control is reached, further gains are limited. For phenotype 3 (the worst glycaemic control), roughly 10% of control participants spend at least 5% of their time in this state. Because these participants start from a poorer baseline, they exhibit larger absolute reductions-improvements that are inherently easier to achieve than those seen in the treatment arm. We summarize the main results below.

i) Roughly 10% of the treated participants increase their time in phenotype 1 within the first month. ii) Only a minority of patients, both improving and worsening, show time-varying changes in weights $\alpha_s^{(p)}$; improvements, when present, typically require at least two weeks and then stabilize or reverse after four months in the absence of intervention. iii) Continuous-time mixed-density models capture dynamic real-time fluctuations that static or discretized analyses miss. iv) Weight trajectories $\alpha_s^{(p)}$ provide a clinically interpretable metric to monitor individual response in CGM-based trials.

Bivariate analysis

We extend the univariate analysis by augmenting glucose levels with a local velocity feature computed within the same ten-day windows.

For each participant $p \in [n]$ and each ten-day window $i \in [m]$, we form the bivariate sample

$$(G_{p,i,\ell}, V_{p,i,\ell}), \qquad \ell \in [n_{p,i}],$$

where $G_{p,i,\ell}$ is the glucose reading and

$$V_{p,i,\ell} = \frac{G_{p,i,\ell} - G_{p,i,\ell-1}}{\Delta t}$$

is a finite-difference derivative computed from consecutive available measurements within the window, with sampling interval $\Delta t = 5$ minutes. (We fix $V_{p,i,1} = 0$ and, if one of the two consecutive readings is missing/corrupted, the corresponding difference is omitted.)

Bivariate mixture representation. Within each window, we model the joint density of (G, V) by the dynamic mixture

$$\rho_{\theta(t)}^{(p)}(g, v) = \sum_{s=1}^{K} \alpha_s^{(p)}(t)\, \mathcal{N}\left((g, v) \mid m_s, \Sigma_s\right), \qquad t \in [0, 1].$$

where $m_s \in \mathbb{R}^2$ and $\Sigma_s \in \mathscr{S}_2^+(\mathbb{R})$ are shared across all participants, and the weight trajectories $\alpha_s^{(p)}(t)$ capture subject-specific temporal dynamics. Again, we fix $K = 3$ and calibrate $\{m_s\}_{s=1,2,3}$ and $\{\Sigma_s\}_{s=1,2,3}$ using the same three representative patients, see Table 7.4.

Figure 7.6 presents the bivariate analogue of the analysis; it displays the estimated mixture weights for both arms and the quantile curves of $Z_s^{(p)}$, incorporating the temporal derivative of glucose.

7.6 Perspectives

The presented methodology admits several natural extensions.

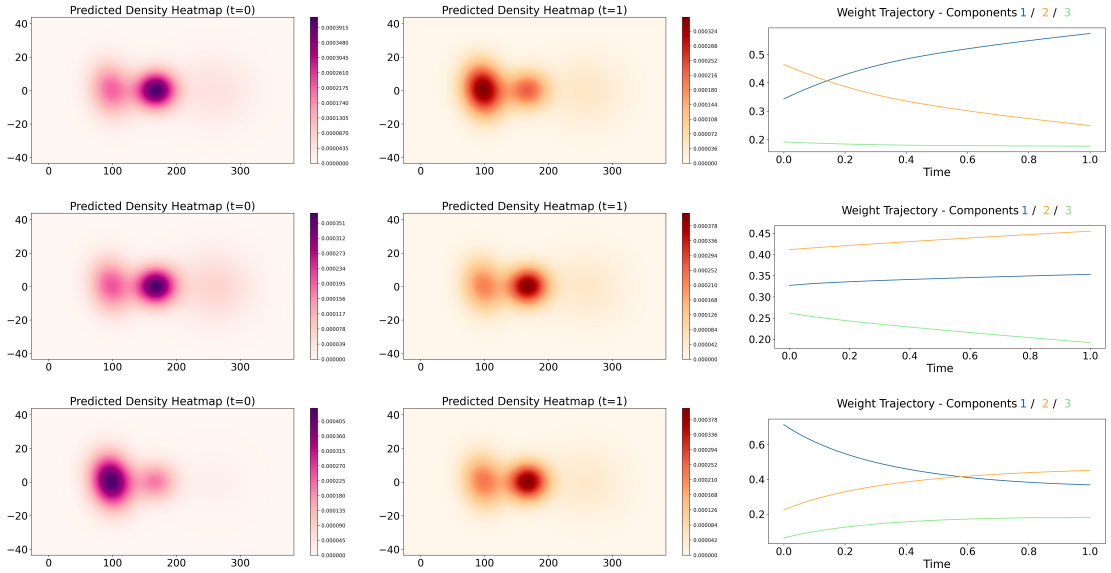


Figure 7.5: Three representative subjects (IDs 13, 62, and 377; one per row). **Left:** initial fitted bivariate density. **Middle:** final fitted bivariate density. **Right:** weight trajectories $\alpha_s^{(p)}(t)$ from the neural ODE model, whose endpoints match the middle-column fits.

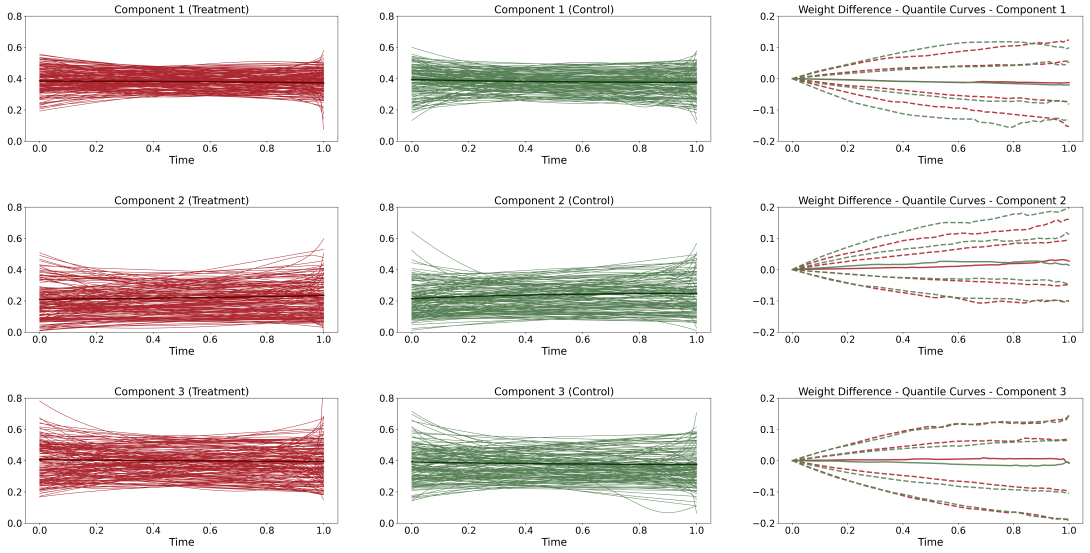


Figure 7.6: Component-wise weight dynamics on the full medical dataset after extending to a bivariate analysis by including the temporal derivative of glucose. **Columns 1–2:** individual curves $\alpha_s^{(p)}(t)$ (thin) and mean (thick) for treatment (left) and control (right). **Column 3:** $(0.1, 0.25, 0.5, 0.75, 0.9)$-quantile trajectories of $Z_s^{(p)} = \alpha_s^{(p)}(t) - \alpha_s^{(p)}(0)$ across subjects.

1. **Online and distributed variants.** In applications such as large cohort studies (e.g. *All of Us* [Inv19]), one would like to update $\alpha(t)$ and possibly the shared dictionary as new data arrive, without refitting the whole model. Because Stage 1 decomposes across time points and Stage 2 is a standard ODE training problem, both admit streaming and distributed implementations (mini-batching in time and across subjects).

2. **Uncertainty quantification for distribution-valued forecasts.** The output of the method is a curve $t \mapsto \mu_t$, so uncertainty can be reported at several levels: on scalar functionals (e.g. quantiles, time-in-range), on the weights $\alpha(t)$, or directly as a set of plausible distributions. A promising route is to combine the estimator with conformal prediction on metric spaces [LM24], using a distributional discrepancy (MMD or Wasserstein) as the conformity score.

3. **Richer longitudinal biomarkers.** Finally, the approach can be used for multivariate biomarkers (e.g. glucose together with its derivative, as in the bivariate analysis in Section **Section 7.5**, or there are even UCI studies with > 50 concurrent biomarkers). The shared-dictionary idea is particularly attractive in this regime, because it yields a common set of interpretable *anchors* and shifts the learning task to the low-dimensional simplex-valued trajectories.

4. **Beyond finite dimension.** In some domains the observations are curves, images, or trajectories and are more naturally modeled as elements of a separable Hilbert space. Since MMD is defined for general kernels, an extension is to use kernels on function spaces and mixtures of Gaussian measures defined by covariance operators, with numerical evaluation via spectral truncations. This direction connects with functional time-series modeling in Hilbert spaces [DM20] and with kernel mean embedding techniques [Mua+17].

5. **Closed-loop control of distributional dynamics.** In diabetes and related settings, the biomarker distribution can be steered by an external input, such as insulin delivery. A concrete extension is to couple the learned *natural* drift of $\alpha(t)$ with a control channel. This yields a distribution-level model predictive control problem where a clinically desirable reference distribution is tracked using an MMD-based objective, while respecting actuator limits and safety constraints.

7.7 Proofs

Proof of Proposition 7.3.1. Fix $\varepsilon > 0$. By assumption (1), we can choose a radius $R < \infty$ such that

$$\sup_{t\in[0,T]} \int_{\{\|x\|>R\}} \rho_t(x)\,\mathrm{d}x < \frac{\varepsilon}{4}.$$

Let $\varphi_{\sigma_0}(x) \coloneqq \mathcal{N}(x \mid 0, \sigma_0^2 I_d)$ be the Gaussian mollifier, and $g_t \coloneqq \rho_t * \varphi_{\sigma_0}$ be the smoothed densities. By standard properties of approximate identities and the uniform translation-continuity from assumption (2), we have

$$\sup_{t\in[0,T]} \|\rho_t - g_t\|_{L^1(\mathbb{R}^d)} < \frac{\varepsilon}{4}$$

for a sufficiently small $\sigma_0 > 0$. We fix such a σ_0.

Next, partition the closed ball $B_R \coloneqq \{x \in \mathbb{R}^d \mid \|x\| \le R\}$ into a finite number K of disjoint measurable sets $(A_s)_{s=1}^K$, each with a diameter bounded by $h > 0$, and pick a component mean $m_s \in A_s$ for each s. Define the unnormalized weights

$$\tilde{\alpha}_s(t) \coloneqq \int_{A_s} \rho_t(y)\,\mathrm{d}y.$$

The total mass contained in the ball is given by

$$Z_t \coloneqq \sum_{s=1}^{K} \tilde{\alpha}_s(t) = \int_{B_R} \rho_t(y)\,\mathrm{d}y,$$

which satisfies $Z_t \in [1-\varepsilon/4, 1]$. We then define the normalized mixing weights as $\alpha_s(t) \coloneqq \tilde{\alpha}_s(t)/Z_t$. We can split the smoothed density as follows:

$$g_t(x) = \sum_{s=1}^{K} \int_{A_s} \rho_t(y)\varphi_{\sigma_0}(x-y)\,\mathrm{d}y + \int_{B_R^c} \rho_t(y)\varphi_{\sigma_0}(x-y)\,\mathrm{d}y.$$

The second term has an $L^1(\mathbb{R}^d)$ norm of exactly $1 - Z_t \le \varepsilon/4$. To bound the first term, we use the Lipschitz property of the Gaussian mollifier:

$$\|\varphi_{\sigma_0}(\cdot - y) - \varphi_{\sigma_0}(\cdot - m_s)\|_{L^1(\mathbb{R}^d)} \le \|\nabla\varphi_{\sigma_0}\|_{L^1(\mathbb{R}^d)}\|y - m_s\| \le C_d\,\frac{h}{\sigma_0},$$

where C_d is a dimensional constant. This implies that

$$\sup_{t\in[0,T]} \left\| g_t - \sum_{s=1}^{K} \tilde{\alpha}_s(t)\varphi_{\sigma_0}(\cdot - m_s) \right\|_{L^1(\mathbb{R}^d)} \le \frac{\varepsilon}{4} + C_d\,\frac{h}{\sigma_0}. \tag{7.7.1}$$

Normalizing the weights from $\tilde{\alpha}_s(t)$ to $\alpha_s(t)$ introduces an additional $L^1(\mathbb{R}^d)$ error bounded by $|1 - Z_t| \le \varepsilon/4$, since $\|\varphi_{\sigma_0}\|_{L^1(\mathbb{R}^d)} = 1$.

By choosing $h \le \sigma_0\varepsilon(4C_d)^{-1}$, the sum of these three error bounds guarantees that (7.7.1) is strictly less than ε.

Finally, the continuity of the map $t \mapsto \alpha(t)$ follows directly from the continuity of $t \mapsto \rho_t$ in $L^1(\mathbb{R}^d)$ via the dominated convergence theorem. □

Proof of Theorem 7.3.2. Let $A_i \coloneqq I_i + \mathrm{diag}(\lambda)$. Since the matrix I_i is positive semi-definite ($I_i \succeq 0$) and the ridge penalties satisfy $\lambda_{\min} > 0$, we have $A_i \succeq \lambda_{\min} I_K$. The empirical and population minimizers over the simplex Δ^{K-1} must satisfy the standard variational inequality optimality conditions. Specifically, for all $\alpha \in \Delta^{K-1}$, we have

$$\langle 2A_i\alpha_i^* - 2J_i, \alpha - \alpha_i^* \rangle \ge 0, \qquad \text{and} \qquad \langle 2A_i\alpha_i^\circ - 2J_i^\circ, \alpha - \alpha_i^\circ \rangle \ge 0.$$

Taking $\alpha = \alpha_i^\circ$ in the first inequality and $\alpha = \alpha_i^*$ in the second, and adding them together, yields

$$\langle A_i(\alpha_i^* - \alpha_i^\circ), \alpha_i^* - \alpha_i^\circ \rangle \le \langle J_i - J_i^\circ, \alpha_i^* - \alpha_i^\circ \rangle.$$

Applying the strong convexity bound $A_i \succeq \lambda_{\min} I_K$ on the left side and the Cauchy–Schwarz inequality on the right side, we obtain

$$\lambda_{\min}\|\alpha_i^* - \alpha_i^\circ\|^2 \le \|J_i - J_i^\circ\|\|\alpha_i^* - \alpha_i^\circ\|,$$

which simplifies to

$$\|\alpha_i^* - \alpha_i^\circ\| \le \frac{1}{\lambda_{\min}}\|J_i - J_i^\circ\|.$$

Next, we bound the empirical deviation $\|J_i - J_i^\circ\|$. For each component $s \in \{1, \dots, K\}$, the random variable $J_{i,s}(X_{t_i,j})$ satisfies $0 \le J_{i,s}(X_{t_i,j}) \le 1$, because the kernel is bounded by $0 \le k_i(x, y) \le 1$

and the Gaussian density $\mathcal{N}(\cdot \mid m_s, \Sigma_s)$ integrates to 1. Consequently, the empirical entry $(J_i)_s$ is an average of N_i independent random variables taking values in $[0, 1]$, with expected value $(J_i^\circ)_s$. By Hoeffding's inequality and a union bound over all K components, we have, with probability at least $1 - \delta$,

$$\max_{1 \leqslant s \leqslant K} |(J_i)_s - (J_i^\circ)_s| \leq \sqrt{\frac{\log(2K/\delta)}{2N_i}}.$$

Converting this maximum absolute deviation to the Euclidean norm gives

$$\|J_i - J_i^\circ\| \leq \sqrt{K} \max_{1 \leqslant s \leqslant K} |(J_i)_s - (J_i^\circ)_s| \leq \sqrt{\frac{K \log(2K/\delta)}{2N_i}}.$$

Combining this with the Lipschitz bound derived from the optimality conditions yields the first result.

Finally, to obtain the uniform bound across the entire time grid, we apply a union bound over all observation times $i = 0, \ldots, m$. Adjusting the confidence level to $\delta/(m + 1)$ for each time step and bounding the sample size by $N_* = \min_{0 \leq i \leq m} N_i$ directly yields the second statement. □

Chapter 8

Mean-field Landscape of Transformers

8.1 Introduction

Transformers are the backbone of modern *large language models* and many state-of-the-art systems for vision and speech. Their distinctive operation is *self-attention*, which updates each token by aggregating information from the others through data-dependent weights [Vas+17]. In practice, this mechanism is composed across depth with *residual connections* and position-wise feed-forward blocks (MLPs), producing highly structured yet poorly understood representation dynamics.

For studying signal propagation, it is useful to cast the dynamics of token representations as an *interacting particle system* [San+22; Ges+25]. Motivated by the perspective of treating layer depth as a continuous time variable [Che+18] and by the fact that common *normalization schemes* (e.g., RMSNorm) keep embeddings on a compact manifold, one idealizes *token embeddings* as particles $x_i(t)$ evolving on the unit sphere $\mathbb{S}^{d-1}$ [Ges+25]. Self-attention then becomes a state-dependent coupling: each particle moves toward a weighted average of the others, with weights given by a softmax-type kernel of pairwise inner products. This interacting-particle viewpoint—which we adopt throughout—has gained traction in the past few years. It links Transformers to nonlinear consensus and collective-dynamics models [Lu+20; Dut+21; Ges+23], and it also interfaces naturally with statistical mechanics treatments of representation geometry [Cow+25; GG26; Tib+24; Cui+25]. While most of these works study idealized models, scaling laws derived by [Cow+25]—subsequently studied in [Che+25a; GG26]—have been used in training large language models (OLMO2 7B & 13B, [OLM+24]).

One object of interest in modern applications is very large context lengths. This is precisely where the *mean-field limit* of the interacting particle system becomes relevant: as the number of tokens n grows, the *empirical measure* of particles converges to a measure solving a nonlinear partial differential equation on the sphere whose velocity field is the attention interaction. This mean-field limit relies on a Dobrushin-type stability estimate in Wasserstein distance [Dob79]; its application to our specific setting—based on the Lipschitz dependence of the velocity field on both the state and the measure—is explicitly covered in [Ges+25, Remark 3.3]. A convenient advantage of the particle perspective is the transparent variational structure: the finite-n particle dynamics can be written as a preconditioned (or weighted-metric) gradient flow of an interaction energy, yielding *gradient descent* when the key-query-value-induced quadratic form is positive semidefinite and gradient *ascent* otherwise [Ges+25]. This gradient flow structure persists in the mean-field limit, where the limiting PDE is a (weighted) *Wasserstein gradient flow* [JKO98; Ott01]. This framework naturally connects to the study of equilibrium measures for nonlocal interaction energies, including the structural properties of their minimizers [CCP15; CFP17].

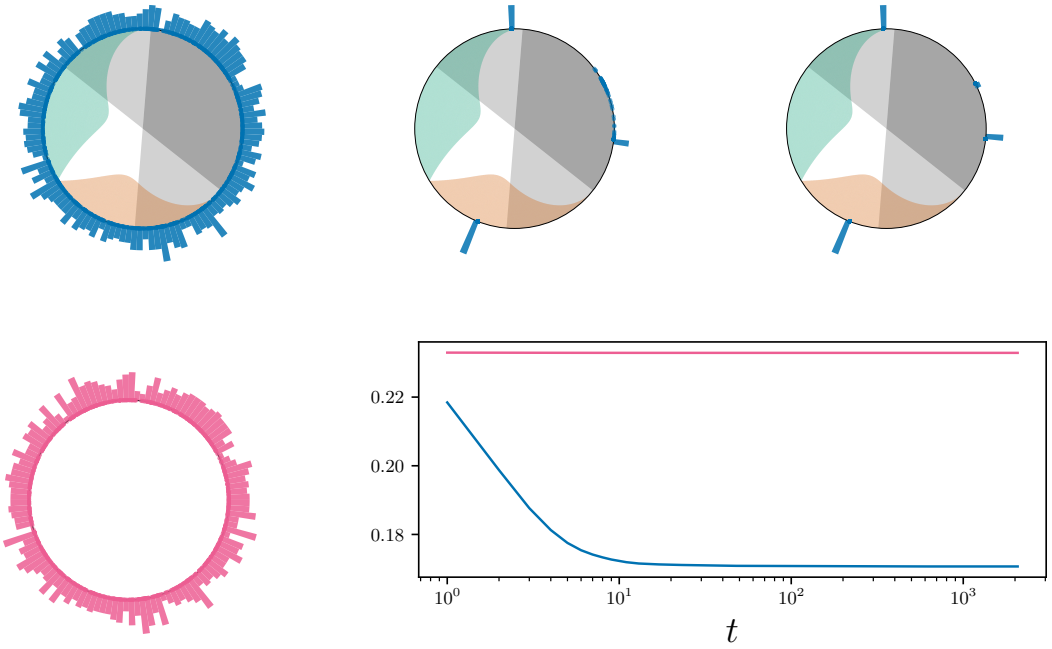


Figure 8.1: Gradient descent with ReLU perceptron in $d = 2$, starting from the uniform measure, with $\beta = 1$. **Top:** Particle histograms for the dynamics at initial, intermediate and final times. **Bottom left:** Final configuration for pure self-attention without a perceptron. **Bottom right:** The energy $\mathsf{E}_{\beta,\theta}$ with (blue) and without (pink) the perceptron. Background shading shows the perceptron landscape (green: > 0; orange: < 0). Areas where $a_j \cdot x + b_j \leqslant 0$ for some j are shown in lighter gray, while the dark gray regions correspond to the *dead zones* where the full potential vanishes.

This variational viewpoint has already enabled a rigorous analysis of the *ascent* (attractive) regime, where attention concentrates and synchronized clusters emerge, as well as of the *descent* (repulsive) regime, where particles tend to equi-distribute uniformly [Ges+23; Ges+25; Che+25b; Cri+24; PRY25; Cas+25; Yu+24; AFZ25a]. These results have been extended to settings with very general trained weights [AST25; Bur+25; KBH24; KM26], to decoder-only architectures with causal masking [KPR24; Bar+25], to finer landscape phenomena such as *metastability* [Ges+24; BPA25b; BPA25a; AGR25; Alt25], and to diffusive variants [Ger+25; BBR25; SS26; PS25].

A key architectural ingredient is still missing from most of this story: the *feed-forward perceptron block*. Practical evidence suggests that perceptrons qualitatively change the representation geometry by counteracting attention-driven collapse, inducing an order-chaos transition [Cow+25]. From the PDE viewpoint, this module acts as an external drift.

Main question. Our central question is therefore the following:

How does the perceptron block modify the stationary states of the mean-field Transformer?

This is the natural object to study for two reasons. First, stationary measures encode the equilibria of the Wasserstein gradient flow and thus describe its possible long-time organization at the level of

the energy landscape. Second, in the attention-only model, these equilibria display sharply different behaviors depending on the regime: attractive settings induce synchronized clusters, whereas repulsive settings permit diffuse, non-atomic states (with the uniform measure on the sphere acting as the global minimizer). Since the perceptron acts as an external field, the basic issue is whether it merely shifts the location of these equilibria or if it fundamentally changes their nature altogether.

Contributions. In this chapter, we demonstrate that the latter phenomenon occurs: the perceptron fundamentally reshapes the energy landscape. We show that, if the perceptron is conservative, the additional drift localizes the stationary states. Even when pure attention would naturally favor continuous equilibria, the coupled dynamics force the stationary measures to be singular, and generically atomic. In the circle, we prove that these measures have finite support.

Importantly, in the repulsive regime, this enforced discreteness does not imply a trivial collapse to a single point. Instead, we establish quantitative anti-concentration bounds ensuring that mass cannot accumulate arbitrarily at the interaction scale. Thus, the core challenge we address is not merely proving the existence or non-existence of equilibria, but rigorously determining the discrete geometric structure that emerges once the perceptron is coupled to self-attention.

8.2 Setup

Unless otherwise specified, all integrals are taken over $\mathbb{S}^{d-1}$, and we omit the domain from the notation for the sake of readability. We also recall that σ_d denotes the uniform probability measure (normalized surface measure) on $\mathbb{S}^{d-1}$.

8.2.1 Wasserstein gradient flows

We study the evolution of measures driven by a functional $\mathsf{E} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R}$. To this end, we first introduce the central concept of this chapter: the gradient flow in the space of probability measures endowed with the Wasserstein metric.

Definition 8.2.1. A curve $(\mu_t)_{t\geq 0} \subset \mathcal{P}(\mathbb{S}^{d-1})$ is a *Wasserstein gradient flow*—in the steepest descent convention—for E if it is locally absolutely continuous in the *Wasserstein-2 distance* metric $(\mathcal{P}(\mathbb{S}^{d-1}), W_2)$ (recall (5.3.1)) and there exists a velocity field $v_t \in L^2(\mu_t; \mathsf{T}\mathbb{S}^{d-1})$ such that

$$\partial_t \mu_t + \operatorname{div}_{\mathbb{S}}(\mu_t v_t) = 0, \tag{8.2.1}$$

in the sense of distributions, with velocity given by the gradient of the first variation,

$$v_t = -\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu_t]. \tag{8.2.2}$$

We recall that the *first variation* $\frac{\delta \mathsf{E}}{\delta \mu}[\mu]$ is defined (up to an additive constant) by the relation

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathsf{E}(\mu + \varepsilon\chi)\Big|_{\varepsilon=0} = \int_{\mathbb{S}^{d-1}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu] \, \mathrm{d}\chi,$$

for signed measures χ with $\chi(\mathbb{S}^{d-1}) = 0$ such that $\mu + \varepsilon\chi \in \mathcal{P}(\mathbb{S}^{d-1})$ for $|\varepsilon|$ small. In the Wasserstein geometry, the relevant perturbations are transport ones $\chi = -\operatorname{div}_{\mathbb{S}}(\mu\xi)$ with $\xi \in L^2(\mu; \mathsf{T}\mathbb{S}^{d-1})$, which identifies the Wasserstein gradient as above; see [AGS08].

8.2.2 Critical points

We are primarily interested in the asymptotic behavior of these flows, which relates to the critical points of E.

Definition 8.2.2. A measure $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ is a (Wasserstein) *critical point* of E, or *stationary solution* to (8.2.1), if it satisfies:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu](x) = 0 \qquad \text{for } \mu\text{-a.e. } x. \tag{8.2.3}$$

Remark 8.2.1. The pointwise condition (8.2.3) is equivalent to the standard notion of a stationary solution for (8.2.1), which requires μ to satisfy (in the weak sense)

$$\operatorname{div}_{\mathbb{S}} \left(\mu \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu] \right) = 0. \tag{8.2.4}$$

The implication (8.2.3) $\implies$ (8.2.4) is immediate. Conversely, if (8.2.4) holds, integrating it against the smooth test function $\varphi = \frac{\delta \mathsf{E}}{\delta \mu}[\mu]$ directly yields

$$\int \left\| \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu] \right\|^2 \mathrm{d}\mu = 0.$$

This strictly forces $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu](x) = 0$ for μ-a.e. x, recovering (8.2.3). Thus, for drifts of the form (8.2.2), the distributional and pointwise stationarity conditions coincide.

We are also interested in *second-order (Wasserstein) critical points*, namely critical points at which the Wasserstein Hessian is well-defined. Recall that

$$\mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla_{\mathbb{S}} \phi \mid \phi \in \mathscr{C}^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)} \tag{8.2.5}$$

is the *(Otto) tangent space* at μ, a Hilbert subspace of $L^2(\mu; \mathsf{T}\mathbb{S}^{d-1})$. Suppose μ lies in the regular (Riemannian) part of $(\mathcal{P}_2(\mathbb{S}^{d-1}), W_2)$ and E is twice differentiable in the W_2-sense at μ [Vil09, Ch. 15]. Then the *Wasserstein Hessian* at μ in the direction $\xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1})$ is defined by

$$\operatorname{Hess}_\mu \mathsf{E}(\xi, \xi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathsf{E}(\mu_t) \Big|_{t=0},$$

where (μ_t) is the W_2-geodesic emanating from μ with initial velocity field represented by $\xi = \nabla_{\mathbb{S}} \phi$. For small $|t|$, this geodesic is given explicitly by[1]

$$\mu_t = (T_t)_{\#} \mu, \qquad \text{with} \qquad T_t(x) = \exp_x \big(t \nabla_{\mathbb{S}} \phi(x) \big). \tag{8.2.6}$$

Definition 8.2.3. A second-order Wasserstein critical point μ is *second-order positive-definite* (SOPD) if

$$\operatorname{Hess}_\mu \mathsf{E}(\xi, \xi) \geq 0 \qquad \text{for all } \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}), \tag{8.2.7}$$

and it is *strictly* SOPD if, moreover, there is $\kappa > 0$ such that

$$\operatorname{Hess}_\mu \mathsf{E}(\xi, \xi) \geq \kappa \|\xi\|^2_{L^2(\mu)}. \tag{8.2.8}$$

A SOPD critical point is *degenerate* if it is not strictly SOPD, i.e. if there exists $\xi \neq 0$ with $\operatorname{Hess}_\mu \mathsf{E}(\xi, \xi) = 0$. If (8.2.7) fails (there exists ξ with $\operatorname{Hess}_\mu \mathsf{E}(\xi, \xi) < 0$), we call μ unstable, since there exists a direction of negative second variation.

[1] Here $\exp_x$ denotes the Riemannian exponential map on $\mathbb{S}^{d-1}$. In general (e.g., for singular μ), geodesics need not be uniquely determined by an $L^2(\mu)$ velocity field; in that case, one may instead work with directional second variations along Monge perturbations, but we will not need this level of generality here.

We focus on SOPD Wasserstein critical points because they ought to capture the appropriate notion of *stable* equilibria for Wasserstein gradient (descent) flows and because quantitative convergence estimates are typically driven by second-order information (e.g., Wasserstein Hessian/curvature or related PL-type inequalities) around such points [Vil09; Ott01; Che+25b].

In our coupled attention–perceptron setting, this perspective also highlights a surprising phenomenon: even in the repulsive (descent) regime, stable stationary states are forced to be discrete, so the mean-field flow can evolve from a smooth initialization toward a genuinely clustered configuration rather than a diffuse equilibrium.

8.2.3 Mean-field formulation of Transformers

Following [Ges+25], we model the layer-wise evolution of *token embeddings* as the flow of an *interacting particle system* on $\mathbb{S}^{d-1}$. This can be viewed as a geometric surrogate for the effect of layer normalization [BKH16; Xio+20]—more precisely, for the post-normalization scheme employed in the original Transformer [Vas+17]—which normalizes the token states as *time* (the layer index) evolves. At the mean-field level, the distribution of a particle under pure self-attention dynamics follows the Wasserstein gradient flow of

$$\mathsf{E}_\beta[\mu] := \frac{1}{2\beta} \iint e^{\beta x\cdot y} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \tag{8.2.9}$$

Its first variation is

$$\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x) = \frac{1}{\beta}\int e^{\beta x\cdot y}\,\mathrm{d}\mu(y), \tag{8.2.10}$$

and the Wasserstein gradient in (8.2.1) for the pure self-attention dynamics reads

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x) = \mathbf{P}_x^\perp \int e^{\beta x\cdot y} y \,\mathrm{d}\mu(y), \tag{8.2.11}$$

where we recall that $\mathbf{P}_x^\perp$ denotes the orthogonal projection onto the tangent space $\mathsf{T}_x\mathbb{S}^{d-1}$ at $x \in \mathbb{S}^{d-1}$.

In view of the formulation (10) introduced in The Big Picture, the perceptron component can be incorporated at the mean-field level as an additional drift field. This yields the continuity equation

$$\partial_t \mu_t \pm \operatorname{div}_{\mathbb{S}}\left(\mu_t\left(\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu_t] + v_\theta\right)\right) = 0, \tag{USA+P}$$

where v_θ is a velocity field on $\mathbb{S}^{d-1}$ given by the tangent projection of a perceptron of the form[2]

$$v_\theta(x) := \mathbf{P}_x^\perp \sum_{j=1}^{d} w_j \sigma(a_j \cdot x + b_j), \tag{$\mathcal{V}_d$}$$

where $w_j, a_j \in \mathbb{R}^d$ and $b_j \in \mathbb{R}$. Note that, for simplicity, we have fixed the width of v_θ to $p = d$ neurons. The general case is readily obtained by replacing d with p in the subsequent derivations.

We refer to the $+$ case in (USA+P) as *ascent*, and $-$ as *descent*. One typically uses

$$\sigma(s) = s_+ \quad \text{(ReLU)} \qquad \text{or} \qquad \sigma(s) = \tfrac{s}{2}\left(1 + \operatorname{erf}\left(\tfrac{s}{\sqrt{2}}\right)\right) \quad \text{(GeLU)}$$

[2]The tag (USA+P) reflects that the interaction field in (8.2.11) is often referred to as Unnormalized Self-Attention (USA) [Ges+25], while the "p" stands for the added perceptron drift.

(recall Figure 1 in The Big Picture). Our analysis distinguishes between these two cases. For simplicity, we henceforth omit biases b_j but discuss their inclusion later in Remark 8.3.6.

When the drift v_θ derives from a scalar potential, the coupled dynamics remain a Wasserstein gradient flow. For this to hold, we must impose that the outer weights w_j be proportional to the inner weights a_j. That is, there must exist $\omega_j \in \mathbb{R}$ such that

$$w_j = \omega_j a_j \qquad \text{for all } j \in [d]. \tag{8.2.12}$$

Under this condition, we can choose a primitive function $\Phi : \mathbb{R} \to \mathbb{R}$ such that $\Phi'(s) = 2\sigma(s)$, and define the scalar potential

$$\mathsf{V}_\theta(x) \coloneqq \sum_{j=1}^{d} \omega_j \Phi(a_j \cdot x), \tag{8.2.13}$$

which guarantees that $\frac{1}{2}\nabla_{\mathbb{S}}\mathsf{V}_\theta(x) = v_\theta(x)$. Henceforth, we focus on this case (assuming (8.2.12)) and denote the parameters by

$$\theta \coloneqq (\omega_j, a_j)_{j=1}^{d} \in (\mathbb{R}^{1+d})^d.$$

The full energy is then

$$\mathsf{E}_{\beta,\theta}[\mu] \coloneqq \frac{1}{2\beta} \iint e^{\beta x \cdot y} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) + \frac{1}{2}\int \mathsf{V}_\theta(x)\,\mathrm{d}\mu(x),$$

and (USA+P) is precisely its *Wasserstein gradient flow*, since

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x) = \int e^{\beta x \cdot y} \mathbf{P}_x^{\perp} y \,\mathrm{d}\mu(y) + v_\theta(x).$$

Whenever σ is continuous, the stationarity condition (8.2.3) extends to the entire support:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x) = 0 \qquad \text{for all } x \in \operatorname{supp}\mu. \tag{8.2.14}$$

Recall that the support of $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ is

$$\operatorname{supp}\mu \coloneqq \{x \in \mathbb{S}^{d-1} \mid \mu(U) > 0 \text{ for every open neighborhood } U \ni x\}. \tag{8.2.15}$$

Equivalently, $\operatorname{supp}\mu$ is the smallest closed subset $F \subset \mathbb{S}^{d-1}$ with $\mu(F) = 1$.

8.2.4 Practical considerations

While our toy model captures the core interaction dynamics, we highlight three main simplifications compared to Transformers used in practice.

Perceptrons. The drift given by the perceptron is a gradient field only under the specific weight symmetries assumed in (8.2.12). More general weights w_j could be interpreted as preconditioners, as discussed in [AGR25, Section 2.2].

Self-attention. Our idealized model abstracts away two main components of standard attention:

(i) Practical implementations replace $\beta x \cdot y$ with a general bilinear form $x^\top Q^\top K y$ involving the *query matrix* Q and the *key matrix* K. Furthermore, as introduced in (9) from The Big Picture and (2.7.1) from **Chapter 2**, the attention mechanism often transforms the output via a *value matrix* V. Moreover, the matrices (Q, K, V) are generally time-dependent (i.e., they vary across layers).

For technical clarity, we work in the time-independent and isotropic case analyzed in [Ges+25], where $Q^\top K = \beta I_d$ and $V = \pm I_d$. However, our main results can be extended to the setting where $Q^\top K$ is symmetric and invertible; see Remark 8.3.6 and the discussion after Lemma 8.7.1. The sign of V determines whether the dynamics follow gradient ascent or descent, but this does not change the stationary points.

(ii) In practice, attention scores are also normalized via a *softmax* function as

$$\nabla_{\mathbb{S}} \log \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y) = \frac{1}{\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x)} \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x).$$

This field corresponds to a *weighted Wasserstein gradient* $\mathbb{W}$ obtained by rescaling the standard variation by the strictly positive weight

$$\mathsf{w}[\mu](x) \coloneqq \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x). \tag{8.2.16}$$

The resulting dynamics are[3]

$$\partial_t \mu_t + \operatorname{div}_{\mathbb{S}} \left(\mu_t \left(\mathbb{W} \mathsf{E}_\beta[\mu_t] + v_\theta \right) \right) = 0, \tag{SA+P}$$

where the weighted Wasserstein gradient is defined by

$$\mathbb{W} \mathsf{E}[\mu] \coloneqq \frac{1}{\mathsf{w}[\mu]} \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu]. \tag{8.2.17}$$

These dynamics retain the structure of a gradient flow (under a conformally equivalent metric). Since $\mathsf{w} > 0$, the stationarity condition (8.2.14) becomes:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu] + \mathsf{w}[\mu] v_\theta = 0 \qquad \text{on } \operatorname{supp} \mu. \tag{8.2.18}$$

8.3 Results

We now state our main results, which characterize the stationary configurations of the mean-field dynamics defined by (USA+P).

8.3.1 Atomicity of critical points

We begin with the simplest case of the ReLU perceptron in $d = 2$. In this setting, the non-analyticity of the potential forces any stationary measure to have finite support.

Theorem 8.3.1. *Let $d = 2$ and $\beta > 0$, and fix $\sigma(s) = s_+$ in ($\mathcal{V}_d$). Assume that $\theta = (\omega_j, a_j)_j$ is such that V_θ is not real-analytic on $\mathbb{S}^1$. Then any $\mu \in \mathcal{P}(\mathbb{S}^1)$ satisfying the stationarity condition (8.2.14) is purely atomic and has finite support.*

Remark 8.3.2. The non-analyticity assumption on V_θ is a generic condition. For ReLU activations, the potential V_θ is piecewise quadratic on the respective activation regions. The excluded cases correspond precisely to parameter choices for which these piecewise expressions happen to perfectly match across all interfaces. In such pathological scenarios, the potential yields a globally real-analytic function (e.g., $\mathsf{V}_\theta \equiv 0$ or specific weight symmetries that effectively reduce the potential to a quadratic trigonometric polynomial).

[3] Similarly to our convention in ($\mathcal{V}_d$), the tag (SA+P) uses SA to denote standard (normalized) self-attention following the nomenclature of [Ges+25], with the "p" again referring to the perceptron.

When the activation σ is real-analytic (e.g., GeLU), simply looking at solutions of (8.2.14) is not enough due to the analyticity of the potential. However, the same conclusion holds for critical points at which the Wasserstein Hessian is well-defined and satisfies the strict positivity condition (8.2.8).

Theorem 8.3.3. *Let $d = 2$, $\beta > 0$ and $\theta = (\omega_j, a_j)_j$, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be real-analytic in $(\mathcal{V}_d)$. If $\mu \in \mathcal{P}(\mathbb{S}^1)$ is a strict SOPD Wasserstein critical point of $\mathsf{E}_{\beta,\theta}$, then μ is purely atomic and has finite support.*

Remark 8.3.4. By the contrapositive, Theorem 8.3.3 implies that if a SOPD Wasserstein critical point has infinite support, it must be degenerate. From a practical standpoint, this means that diffuse (non-atomic) stationary states cannot be strictly stable under real-analytic perceptron drifts.

In higher dimensions the landscape is more complicated. However, the following theorem establishes that stationary measures are singular under suitable structural assumptions, and that, for any fixed candidate measure, atomicity holds for a generic choice of parameters.

Theorem 8.3.5. *Let $d \geq 2$. Fix $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ and $\sigma : \mathbb{R} \to \mathbb{R}$ in $(\mathcal{V}_d)$.*

(i) *Fix $\beta > 0$. If μ is stationary for $\sigma(s) = s_+$ and $\theta = (\omega_j, a_j)_j$ is such that V_θ is not real-analytic, or if μ is a strict SOPD Wasserstein critical point for a real-analytic σ, then $\sigma_d(\operatorname{supp}\mu) = 0$. In particular, μ is singular with respect to σ_d.*

(ii) *If σ is real-analytic and $\sigma(s) \neq 0$ for $s \neq 0$, then there exists an open and dense parameter set $U_\mu \subset \mathbb{R}_{>0} \times (\mathbb{R}^{1+d})^d$ such that, if $(\beta, \theta) \in U_\mu$ and μ is stationary with respect to (β, θ), then μ is purely atomic with finite support.*

(iii) *If $\sigma(s) = s_+$, then there exists a dense parameter set $U_\mu \subset \mathbb{R}_{>0} \times (\mathbb{R}^{1+d})^d$ such that, if $(\beta, \theta) \in U_\mu$ and μ is stationary with respect to (β, θ), then the restriction of μ to the active regions $\bigcup_j \{a_j \cdot x > 0\}$ is purely atomic with at most countably many atoms.*

Remark 8.3.6. The proofs of Theorems 8.3.1 and 8.3.3 are deferred to **Subsection 8.7.2**, and the proof of Theorem 8.3.5 is given in **Subsection 8.7.3**. These results extend to:

(i) biases $a_j \cdot x + b_j$ inside the perceptron, provided at least one of the hyperplanes $\{a_j \cdot x + b_j = 0\}$ intersects the sphere $\mathbb{S}^{d-1}$ in more than one point (equivalently, $|b_j| < \|a_j\|$ for some j).

(ii) symmetric invertible B instead of βI_d in the interaction energy; see Remarks 8.7.2 and 8.7.7.

8.3.2 Anti-concentration bound

While stationary measures have a discrete nature in the repulsive case, the strict concavity of the kernel[4] $\varphi \mapsto e^{\beta \cos\varphi}$ on $(-\beta^{-\frac{1}{2}}, \beta^{-\frac{1}{2}})$ for $\beta \to \infty$ ensures that they cannot be too concentrated.

Theorem 8.3.7. *Let $d = 2$, $\beta > 0$, and $\theta = (\omega_j, a_j)_j$. Let σ be globally Lipschitz in $(\mathcal{V}_d)$, with $\operatorname{Lip}(\sigma) = 1$ and $\sigma(0) = 0$ for simplicity. Consider any SOPD Wasserstein critical point of the form*

$$\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i} \in \mathcal{P}(\mathbb{S}^1), \tag{8.3.1}$$

[4]This has also been used in [Ges+25; GRS24] in related but different contexts.

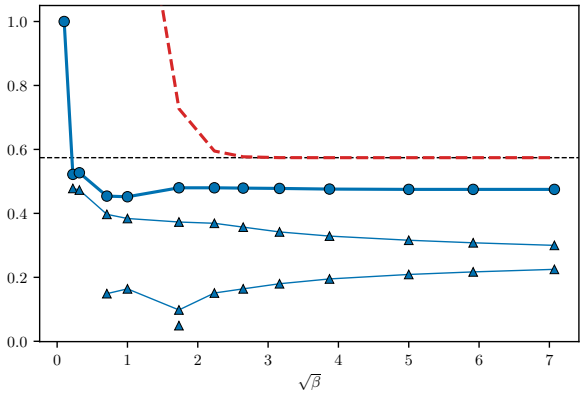


Figure 8.2: Cluster masses (in blue, the largest being the thickest) at final time as a function of $\sqrt{\beta}$ for gradient descent with GeLU perceptron, initialized with $N = 1000$ points of mass 10^{-3} (see **Section 8.4** for setup). The horizontal black and the red dashed lines represent the numerical term and the full upper bound in (8.3.3), respectively.

where $N \geq 2$, the weights $m_i > 0$ sum to 1, and the points $\varphi_i \in [0, 2\pi)$ are pairwise distinct.

Suppose there exists a subset $\mathcal{S} \subseteq [N]$ with $n = |\mathcal{S}| \geqslant 2$ satisfying the cluster condition

$$\max_{i,j \in \mathcal{S}} \min_{k \in \mathbb{Z}} |\varphi_i - \varphi_j + 2\pi k| \leq \frac{1}{2\sqrt{\beta}}. \tag{8.3.2}$$

Then, for β sufficiently large, the total mass of this cluster satisfies

$$\sum_{i \in \mathcal{S}} m_i \leq \frac{16}{16 + 3e^{11/8}} + O(e^{-\beta}) \approx 0.5742 + O(e^{-\beta}). \tag{8.3.3}$$

Moreover, if the parameters satisfy

$$\sum_j |\omega_j| \|a_j\|^2 < \frac{3}{16} e^{-1/8} \approx 0.16547, \tag{8.3.4}$$

then for every $\beta > 0$, the choice $\mathcal{S} = [N]$ is impossible in (8.3.2), that is, the entire support of μ cannot consist of a single cluster.

It follows that

Corollary 8.3.8. *Let β, θ, and σ be as in Theorem 8.3.7. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i} \in \mathcal{P}(\mathbb{S}^1)$ be any SOPD Wasserstein critical point, where $N \geq 1$, the points $\varphi_i \in [0, 2\pi)$ are pairwise distinct, and the weights $m_i > 0$ sum to 1, such that*

$$\operatorname{supp} \mu \subset \bigcup_{\ell=1}^{M} I_\ell, \qquad \text{where } 0 < L < 2\pi \text{ and } \max_{1 \leqslant \ell \leqslant M} |I_\ell| \leq L.$$

Let N_ε be the number of atoms with mass $\geq \varepsilon$ for $\varepsilon > 0$. Then, for β large enough,

$$N_\varepsilon \leq \frac{M}{\varepsilon} \left(1 + 2L\sqrt{\beta}\right) \left(\frac{16}{16 + 3e^{11/8}} + O(e^{-\beta})\right).$$

The proofs of Theorem 8.3.7 and Corollary 8.3.8 are deferred to **Subsection 8.7.4**.

We close this subsection with the natural question of *extremal points*.

Proposition 8.3.9. *Let $d \geq 2$, $\beta > 0$ and $\theta = (\omega_j, a_j)_j$, and let $\sigma \in \mathscr{C}^{0,1}(\mathbb{R})$. Then:*

(i) *The global maximizers of $\mathsf{E}_{\beta,\theta}$ are exactly $\mu = \delta_x$ with*

$$x \in \arg\max_{y \in \mathbb{S}^{d-1}} \sum_{j=1}^{d} \omega_j \Phi(a_j \cdot y).$$

(ii) *$\mathsf{E}_{\beta,\theta}$ has a unique global minimizer $\mu_* \in \mathcal{P}(\mathbb{S}^{d-1})$ which is invariant under rotations that fix every parameter a_j for which $\omega_j \neq 0$.*

The proof can be found in **Subsection 8.7.5**.

8.3.3 Maximizers for ReLU perceptrons

We now compute the maximizers from Proposition 8.3.9 for the case $\sigma(s) = s_+$. Define

$$\mathscr{Z} \coloneqq \{x \in \mathbb{S}^{d-1} \mid a_j \cdot x = 0 \text{ for some } j \in [d]\} \tag{8.3.5}$$

and let I be any connected component of $\mathbb{S}^{d-1} \setminus \mathscr{Z}$ with active index set

$$J_I \coloneqq \{j \in [d] \mid a_j \cdot x > 0 \text{ for all } x \in I\}. \tag{8.3.6}$$

The signs of $x \mapsto a_j \cdot x$ are fixed on I, hence

$$\mathsf{V}_\theta(x) = x^\top B_I \, x, \qquad B_I \coloneqq \sum_{j \in J_I} \omega_j \, a_j a_j^\top.$$

Thus, Proposition 8.3.9 reduces to solving and comparing a finite collection of constrained quadratic programs:

$$\max_{x \in \mathbb{S}^{d-1}} \mathsf{V}_\theta(x) = \max_{I} \; \max_{x \in \overline{I}} x^\top B_I \, x.$$

This yields finitely many candidates, one for each connected component of $\mathbb{S}^{d-1} \setminus \mathscr{Z}$. On each $\overline{I}$, either a principal eigenvector of B_I that lies in I maximizes $x^\top B_I x$, or else all maximizers lie on ∂I.

Across regions, the maximizers of V_θ may form a singleton, a finite set, or a continuum. In special cases, they can be described explicitly:

(i) Collinear directions. Suppose $a_j = \alpha_j \, a$ with $a \in \mathbb{R}^d$ and $\alpha_j \in \mathbb{R}$. Then

$$\mathsf{V}_\theta(x) = (a \cdot x)_+^2 \sum_{\alpha_j > 0} \omega_j \alpha_j^2 + (a \cdot x)_-^2 \sum_{\alpha_j < 0} \omega_j \alpha_j^2.$$

Hence

$$\max_{x \in \mathbb{S}^{d-1}} \mathsf{V}_\theta(x) = \|a\|^2 \max\Big\{ \sum_{\alpha_j > 0} \omega_j \alpha_j^2, \quad \sum_{\alpha_j < 0} \omega_j \alpha_j^2, \quad 0 \Big\}.$$

If $\max \mathsf{V}_\theta > 0$, the maximizer is $\delta_{a/\|a\|}$ or $\delta_{-a/\|a\|}$, according to which of the two sums is larger; both are maximizers in the tie case. Otherwise, any δ_x with $a \cdot x = 0$ is a maximizer.

(ii) Diagonal directions. If $a_j = \alpha_j e_j$ with $\alpha_j \in \mathbb{R}$ and $\omega_j \equiv 1$, then

$$\max_{x\in\mathbb{S}^{d-1}} \mathsf{V}_\theta(x) = \max_{x\in\mathbb{S}^{d-1}} \sum_{j=1}^{d} (\alpha_j x_j)_+^2 = \max_{1\leqslant j\leqslant d} |\alpha_j|^2,$$

and the maximizers are those δ_x with $x_k = 0$ if $|\alpha_k| < \max_j |\alpha_j|$ and $\alpha_k x_k \geq 0$ if $|\alpha_k| = \max_j |\alpha_j|$. If the maximum is attained at a unique index k, then $x = \operatorname{sign}(\alpha_k) e_k$.

(iii) Nonnegative. Assume $a_j \geq 0$ entry-wise and $\omega_j \equiv 1$. For any $x \in \mathbb{S}^{d-1}$,

$$\sum_{j=1}^{d} (a_j \cdot x)_+^2 \leq \sum_{j=1}^{d} (a_j \cdot x)^2 = \|Ax\|_2^2 \;\leq\; \|A\|_2^2,$$

where A is the matrix with rows a_j. Choose a right singular vector v_1 with $\|Av_1\|_2 = \|A\|_2$. Since $A^\top A$ has non-negative entries, the Perron–Frobenius theorem allows us to choose $v_1 \geq 0$, which implies $Av_1 \geq 0$. Thus $x = v_1$ saturates both inequalities, so $\delta_{v_1/\|v_1\|}$ is a maximizer (unique if the top singular value is simple).

8.3.4 The global minimizer: a toy setup

Minimizing $\mathsf{E}_{\beta,\theta}$ (i.e. running gradient descent in (USA+P)) balances two competing effects:

(i) the repulsive self-attention interaction E_β, which favors spread-out (in particular, antipodally symmetric) configurations;

(ii) the perceptron term $\frac{1}{2}\int \mathsf{V}_\theta \, \mathrm{d}\mu$, which biases mass toward minimizers of the potential V_θ.

We make this trade-off explicit in the degenerate case $d = 1$ with $\sigma(s) = s_+$. Here $\mathbb{S}^0 = \{-1, +1\}$ has trivial tangent spaces, so $\mathbf{P}_x^\perp \equiv 0$ and thus the field in (USA+P) vanishes; we simply minimize the energy over $\mathcal{P}(\mathbb{S}^0)$. Write

$$\mathcal{P}(\mathbb{S}^0) = \{\mu_\varepsilon = \varepsilon\delta_1 + (1-\varepsilon)\delta_{-1} \mid \; \varepsilon \in [0,1]\}.$$

For the attention term,

$$\frac{1}{2\beta} \iint e^{\beta x\cdot y} \, \mathrm{d}\mu_\varepsilon(x) \, \mathrm{d}\mu_\varepsilon(y) = \frac{1}{2\beta}\left[\varepsilon^2 e^\beta + 2\varepsilon(1-\varepsilon)e^{-\beta} + (1-\varepsilon)^2 e^\beta\right].$$

For the perceptron term, take one neuron with parameters $a \in \mathbb{R}$ and $\omega \in \mathbb{R}$, and the primitive $\Phi(s) = s_+^2$; then

$$\frac{1}{2}\int \mathsf{V}_\theta(x) \, \mathrm{d}\mu_\varepsilon(x) = \frac{1}{2}\Big[\varepsilon\, \omega\, (a)_+^2 + (1-\varepsilon)\, \omega\, (-a)_+^2\Big].$$

If $a < 0$, the roles of ± 1 are exchanged, so we may assume $a > 0$ without loss of generality. Then $(-a)_+^2 = 0$ and $(a)_+^2 = a^2$, hence

$$\mathsf{E}_{\beta,\theta}[\mu_\varepsilon] = \frac{1}{2\beta}\left[\varepsilon^2 e^\beta + 2\varepsilon(1-\varepsilon)e^{-\beta} + (1-\varepsilon)^2 e^\beta\right] + \frac{1}{2}\,\omega\,\varepsilon\,a^2.$$

Differentiating with respect to ε and setting the derivative to zero yields

$$\frac{e^\beta - e^{-\beta}}{\beta}(2\varepsilon - 1) + \frac{1}{2}\,\omega\, a^2 = 0, \qquad \Longrightarrow \qquad \varepsilon^* = \frac{1}{2} - \frac{\beta\,\omega\,a^2}{8\sinh\beta}.$$

Since the attention contribution is strictly convex in ε and the perceptron contribution is linear, $\mathsf{E}_{\beta,\theta}[\mu_\varepsilon]$ is strictly convex. Hence the constrained minimizer is the projection of ε^* onto $[0,1]$:

$$\varepsilon_{\min} = 1 \wedge \left(\tfrac{1}{2} - \frac{\beta\,\omega\,a^2}{8\sinh\beta}\right)_+ .$$

Regimes.

- **Repulsive bias ($\omega > 0$):** The perceptron term penalizes mass at $x = 1$, shifting the minimizer toward $\varepsilon^* < \frac{1}{2}$.
  - For *weak potentials* ($\beta\,\omega\,a^2 < 4\sinh\beta$), the minimizer is interior:
  $$\varepsilon_{\min} = \varepsilon^* \in \left(0, \tfrac{1}{2}\right).$$
  - For *strong potentials* ($\beta\,\omega\,a^2 \geq 4\sinh\beta$), the constraint saturates and the mass is pushed to the opposite pole:
  $$\varepsilon_{\min} = 0.$$
- **Attractive bias ($\omega < 0$):** The potential favors mass at $x = 1$, shifting the minimizer toward $\varepsilon^* > \frac{1}{2}$.
  - For *weak potentials* ($\beta\,|\omega|\,a^2 < 4\sinh\beta$), the minimizer is interior:
  $$\varepsilon_{\min} = \varepsilon^* \in \left(\tfrac{1}{2}, 1\right).$$
  - For *strong potentials* ($\beta\,|\omega|\,a^2 \geq 4\sinh\beta$), the constraint saturates and the mass concentrates at the target pole:
  $$\varepsilon_{\min} = 1.$$

8.3.5 Normalized self-attention

Recall from **Subsection 8.2.4** that replacing the standard attention mechanism with normalized self-attention (incorporating the *softmax* function) transforms the interaction into a *weighted Wasserstein gradient* flow. As established in (SA+P), the resulting dynamics modify the stationary condition to (8.2.18), where the perceptron drift is scaled by the strictly positive weight $\mathsf{w}[\mu]$.

The sparsity results obtained for the unnormalized case naturally extend to this setting, provided we impose a mild non-degeneracy condition on the perceptron weights. The proof is deferred to **Subsection 8.7.6**.

Proposition 8.3.10. *Let $d \geq 2$ and $\beta > 0$, let $\sigma(s) = s_+$ in (8.2.13), and let $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ be a stationary measure for (SA+P), meaning it satisfies the condition (8.2.18).*

Assume that $(\omega_j, a_j)_j$ satisfy the following non-degeneracy condition: for every index subset $J \subseteq [d]$,

$$\sum_{j\in J} \omega_j\, a_j a_j^\top - \sum_{j\notin J} \omega_j\, a_j a_j^\top \neq \lambda I_d \qquad \text{for all } \lambda \in \mathbb{R}. \tag{8.3.7}$$

Then $\sigma_d(\operatorname{supp}\mu) = 0$; in particular, μ is singular with respect to σ_d. Moreover, if $d = 2$, then μ is purely atomic with finitely many atoms.

8.4 Simulations

We complement our theoretical results with numerical simulations. The source code is available at [Álv26]. Our experiments are designed to:

(i) visualize the emergence of singular (typically atomic) stationary measures;

(ii) quantify the scaling of the number of clusters with respect to the inverse temperature β; and

(iii) assess the dependence of convergence time on the perceptron scale and the choice of attention normalization.

Methodology

Unless stated otherwise, we approximate the continuum dynamics of (USA+P)–(SA+P) by simulating the associated interacting particle system on $\mathbb{S}^1$. Specifically, we track the evolution of an empirical measure consisting of 1000 particles, which are initialized uniformly on $[0, 2\pi)$ using a fixed random seed. The induced system of coupled ordinary differential equations is then integrated via an explicit Euler scheme with time step $\Delta t = 0.1$.

We sweep the inverse temperatures

$$\beta \in \{0.01, 0.05, 0.1, 0.5, 1, 3, 5, 7, 10, 15, 25, 35, 50\}.$$

The perceptron weights are sampled from a standard normal distribution and fixed across all runs.

For a given β, motivated by Lemma 8.7.3 we identify clusters on $\mathbb{S}^1$ by grouping particles whose pairwise geodesic distance satisfies:

$$\operatorname{dist}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}} \wedge \frac{\pi}{4}. \tag{8.4.1}$$

The simulation stops when the number of clusters remains constant over a window of 5 snapshots (with a snapshot every 10 steps) and the maximum velocity is $\max_i |\dot{\varphi}_i| \leq 10^{-4}$. We remark that the system may exhibit metastability [Ges+24; BPA25b; BPA25a], where the dynamics enter a slow regime that satisfies the convergence criteria over long windows despite residual motion at larger timescales.

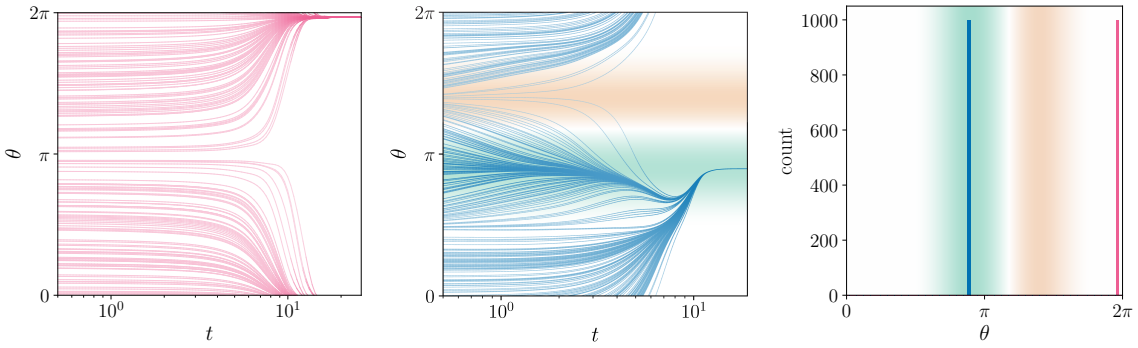
Emergence of clusters

We first consider gradient ascent on $\mathbb{S}^1$ with a ReLU perceptron. Figure 8.3 illustrates three representative runs, revealing two distinct phases:

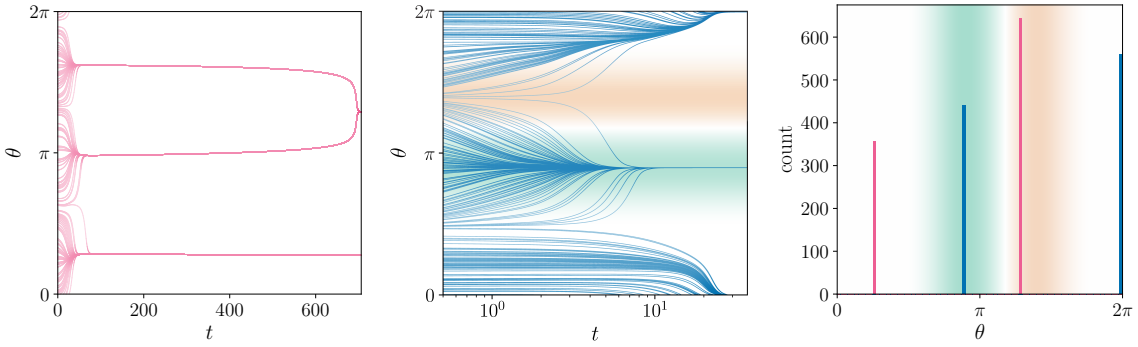
(i) **Aggregation**: Driven by the interaction energy, particles rapidly collapse into local basins of attraction.

(ii) **Locking**: The drift v_θ traps these clusters near local maxima of the potential V_θ, effectively acting as a *magnet*.

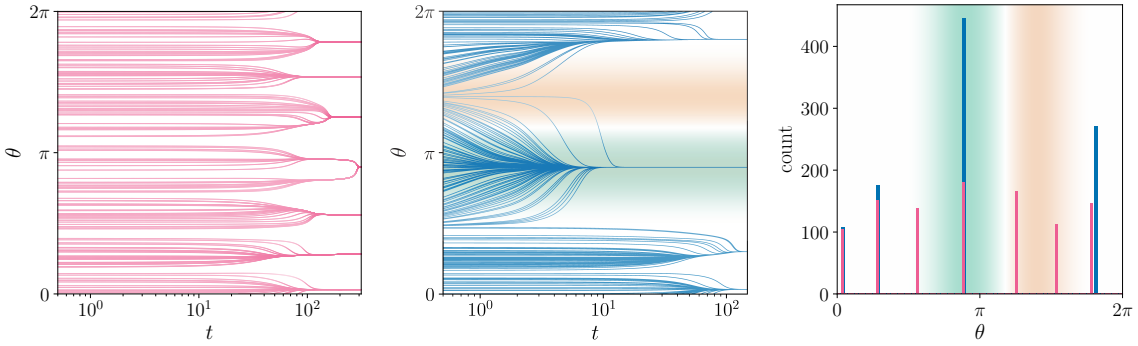
This protocol is repeated on $\mathbb{S}^2$. To visualize the results, we map each particle $x_i(t) \in \mathbb{S}^2$ to spherical coordinates $(\varphi_i(t), \psi_i(t))$ and compute a two-dimensional histogram on a uniform grid.



(a) $\beta = 0.05$

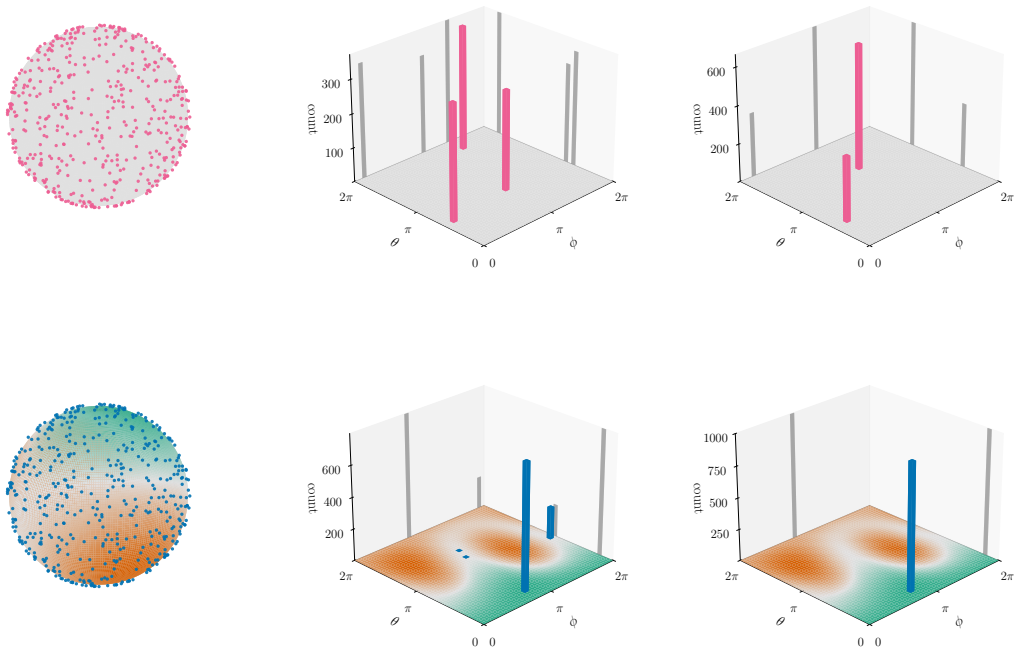


(b) $\beta = 5$



(c) $\beta = 50$

Figure 8.3: Gradient ascent on $\mathbb{S}^1$ with ReLU perceptron. **Left:** Pure self-attention. **Middle:** Self-attention with a ReLU perceptron. **Right:** Measure at final time. Background shading represents the potential landscape (green: positive; orange: negative values).



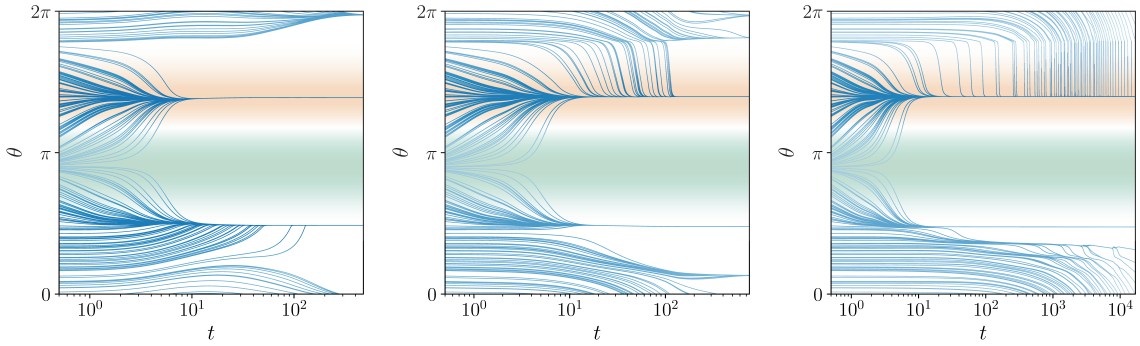
(a) Initial (b) Intermediate (c) Final

Figure 8.4: Gradient ascent on $\mathbb{S}^2$ with $\beta = 1$. **Top row:** Pure self-attention. **Bottom row:** Self-attention with ReLU perceptron. Animation available at https://github.com/antonioalvarezl/2026-MLP-Attention-Energy/blob/main/examples/USAS2.gif.

Gradient descent

We now examine the minimization (descent) dynamics, corresponding to (USA+P) with a negative sign. In this regime, pure self-attention is repulsive and promotes particle dispersion, whereas the perceptron drift enforces singular stationary configurations. As established in Proposition 8.3.9, this system admits a unique global minimizer μ_*, and the results of Theorem 8.3.5 characterize its stationary points.

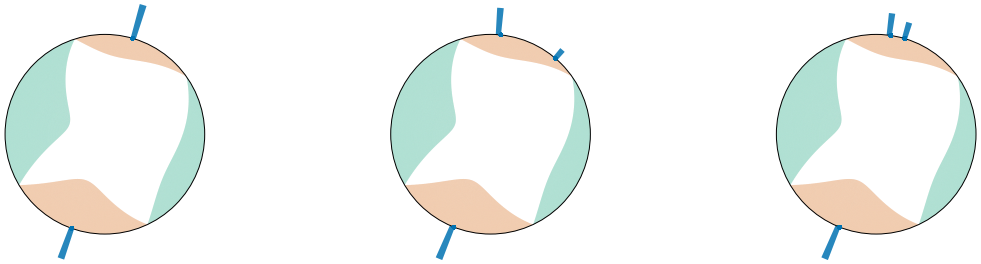
For the ReLU perceptron $\sigma(s) = s_+$, Figure 8.5 shows trajectories for various β. A significant fraction of the mass remains trapped in *dead zones* where the ReLU activation is identically zero. Numerically, this results in a mixed stationary pattern: atomic clusters in active regions coexisting with frozen mass in the regions where the potential vanishes, consistent with Theorem 8.3.5.



(a) $\beta = 0.05$ (b) $\beta = 5$ (c) $\beta = 50$

Figure 8.5: Gradient descent on $\mathbb{S}^1$ with ReLU perceptron.

We repeat the same experiment using GeLU. Figure 8.6 shows representative histograms at numeric convergence. The analogous setup on $\mathbb{S}^2$ appears in Figure 8.7.



(a) $\beta = 0.05$ (b) $\beta = 5$ (c) $\beta = 50$

Figure 8.6: Histograms at final time for gradient descent with GeLU perceptron.

Softmax normalization

We now study the dynamics governed by (SA+P) across the settings previously considered on $\mathbb{S}^1$ and $\mathbb{S}^2$. Figure 8.8 illustrates the particle trajectories on $\mathbb{S}^1$ for three representative values of β.

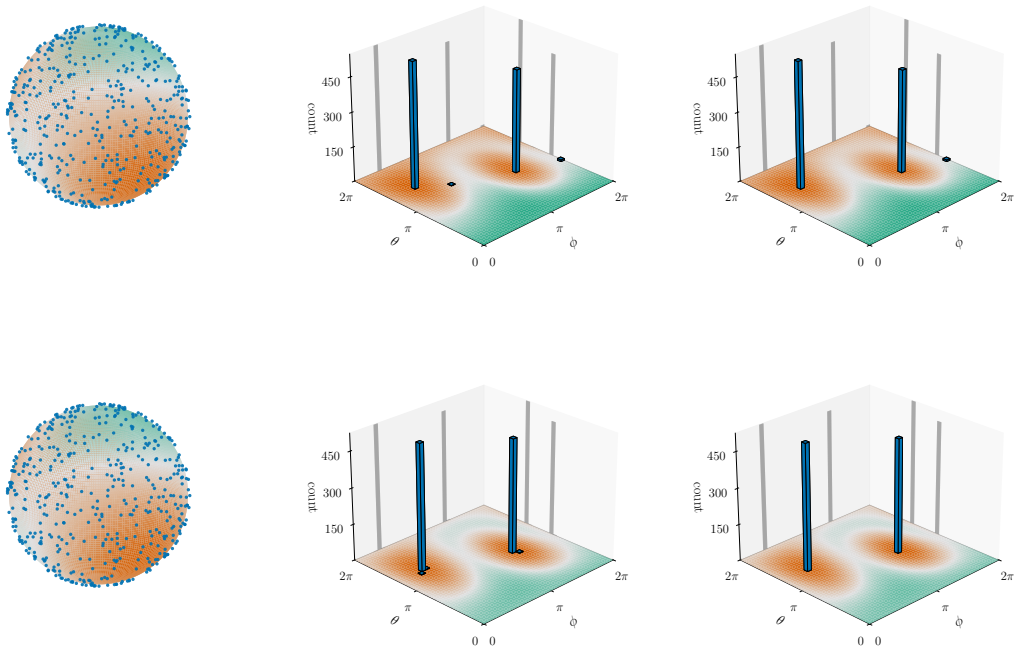


Figure 8.7: Gradient descent on $\mathbb{S}^2$ with $\beta = 1$. **Top row:** ReLU perceptron. **Bottom row:** GeLU perceptron. Animation available at https://github.com/antonioalvarezl/2026-MLP-Attention-Energy/blob/main/examples/USAdS2.gif.

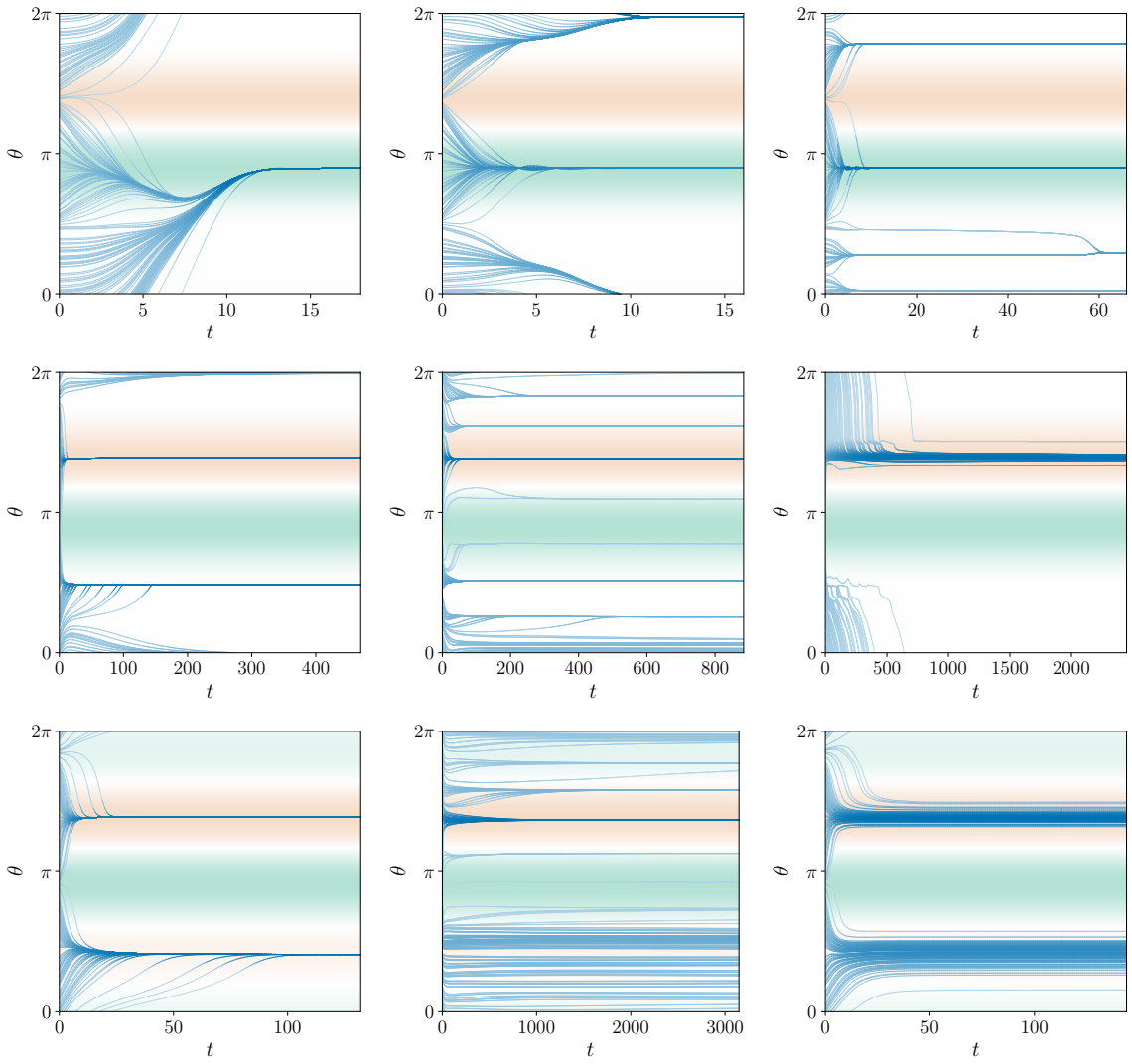


Figure 8.8: Trajectories following softmax-normalized attention. **Top row:** Gradient ascent with ReLU perceptron. **Middle row:** Gradient descent with ReLU perceptron. **Bottom row:** Gradient descent with GeLU perceptron.

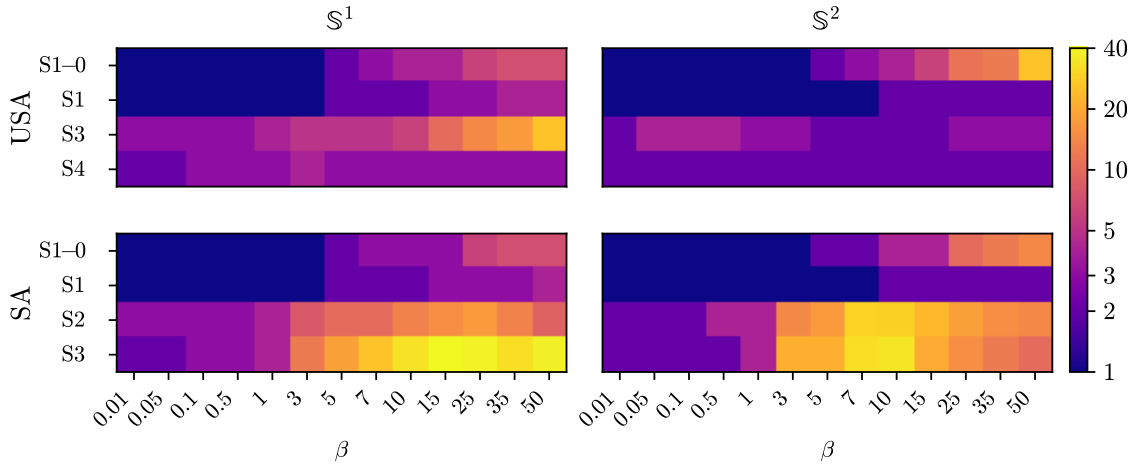


Figure 8.9: Number of clusters at numeric convergence across setups. S1/S1-0: Ascent with/without ReLU perceptron. S2: Descent with ReLU perceptron. S3: Descent with GeLU perceptron.

Scaling of support size

We count the number of clusters (atoms) at numeric convergence for every setup considered across the swept values of β. Results are summarized in Figure 8.9.

For pure self-attention (setup S1-0), the cluster counts reported in Figure 8.9 correspond to the number of atoms in the metastable configuration, rather than the long-time limit. Indeed, these dynamics are known to collapse the initial measure to a single Dirac mass [Ges+25; Che+25b; GRR24].

The same caveat applies when adding a perceptron (setup S1): for the fixed θ used throughout, the perceptron potential exhibits a single local maximum (cf. Figure 8.3), so $\mathsf{E}_{\beta,\theta}$ has a unique maximizer by Proposition 8.3.9. Nevertheless, on computational time horizons we still observe metastable configurations with more than one atom.

Convergence time vs. perceptron scale

To assess how the convergence rate depends on the relative magnitude of the perceptron drift versus the attention mechanism, we fix $\beta = 1$ and vary the common weight scale $|\omega_j|$ for a fixed perceptron realization. We record the time required for the system to reach a stationary state under both unnormalized (USA+P) and softmax-normalized (SA+P) attention using a GeLU perceptron. The results are presented in Figure 8.10.

Sensitivity to initial configuration

To evaluate the robustness of the stationary states, we perform 10 independent simulations at $\beta = 10$ with GeLU activation, using identical weights θ but resampled uniform initializations. Figure 8.11 compares three settings. In the absence of a perceptron, the cluster locations are highly seed-dependent. In contrast, the perceptron drift breaks this symmetry, anchoring the clusters to specific locations that are largely independent of the initial configuration.

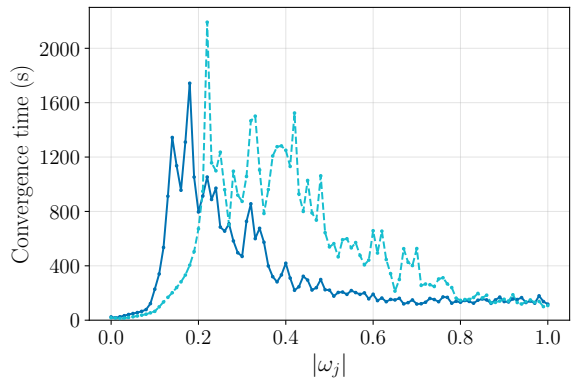


Figure 8.10: Convergence time versus the perceptron scale for fixed $\beta = 1$ under unnormalized attention (blue) and softmax-normalized attention (light blue). The horizontal axis denotes the fixed absolute value shared by all ω_j: here we first sample $\omega_j \sim \mathcal{N}(0, 1)$ and then rescale each component j so that $|\omega_j|$ equals the value on the axis (with the sign retained).

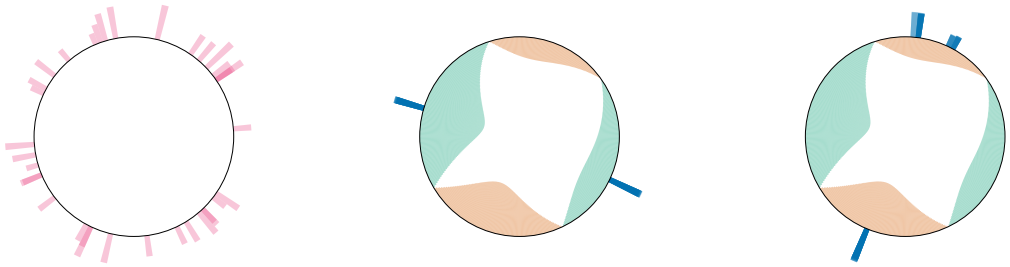


Figure 8.11: Superposition of stationary measures from 10 independent uniform initializations with $\beta = 10$ and $\sigma = \mathrm{GeLU}$. **Left:** Pure self-attention descent yields seed-dependent cluster locations. **Middle:** The perceptron drift in ascent breaks the symmetry and selects seed-independent clusters. **Right:** In descent, the perceptron still anchors a seed-independent stationary support, which is less concentrated and may include spurious atoms.

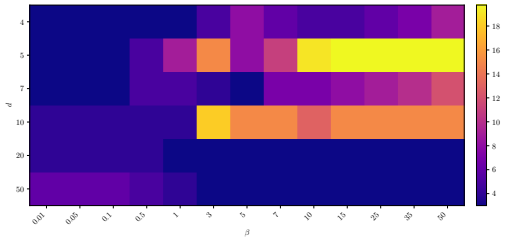


Figure 8.12: Number of clusters at numeric convergence across dimensions $d \in \{4, 5, 7, 10, 20, 50\}$ for gradient descent using unnormalized self-attention with a ReLU perceptron.

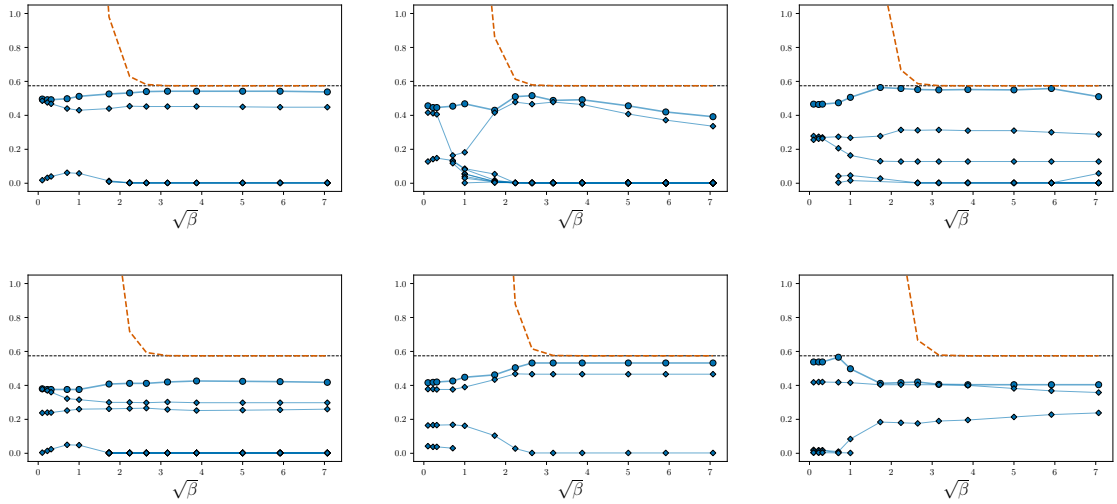


Figure 8.13: Cluster masses (in blue, the largest being the thickest) at final time for varying $\sqrt{\beta}$. The horizontal and red dashed lines represent the numerical term and the full upper bound in (8.3.3), respectively. **Top row:** Gradient descent with ReLU ($d = 4, 5, 7$). **Bottom row:** Same setup for higher dimensions ($d = 10, 20, 50$). (See **Section 8.4** for setup).

Higher dimensions

For general dimension d, we update the cluster identification rule in (8.4.1) to

$$\operatorname{dist}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}} \wedge \frac{\pi}{2d}.$$

This heuristic choice accounts for the concentration of measure on high-dimensional spheres, where random points tend to be nearly orthogonal. Figure 8.12 and Figure 8.13 summarize the cluster counts and their respective masses across dimensions $d \in \{4, 5, 7, 10, 20, 50\}$. Notably, we observe empirically that the cluster masses in these higher dimensions remain strictly below the same numerical upper bound derived for $d = 2$ in Theorem 8.3.7.

8.5 Outlook: Mixture-of-Experts

Mixture-of-Experts (MoE) models were introduced in the early 1990s as a way to combine several predictors through an input-dependent gate [Jac+91; JJ94].[5] Their modern resurgence in large-scale language models is typically motivated by *conditional computation*: one increases the total number of parameters by having many expert MLPs, while only activating a small number of experts per token. This is the principle behind sparse Transformer MoEs such as the sparsely-gated layer of [Sha+17] and subsequent large-scale implementations [Lep+21; FZS22].

Formally, an MoE layer consists of *expert* maps $E_i : \mathbb{R}^d \to \mathbb{R}^d$ for $i \in [m]$ and a *gating network* $G \coloneqq (g_1, \dots, g_m)$ producing nonnegative weights $g_i : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. The output is

$$\mathrm{MoE}(x) \coloneqq \sum_{i=1}^{m} g_i(x)\, E_i(x). \tag{8.5.1}$$

Unlike classical MoEs, which often use dense weights (all experts contribute) [Jac+91; JJ94], in modern MoEs (used in the Transformer architecture) the gating network is typically *sparse*: $G(x)$ activates for each token $x \in \mathbb{R}^d$ a fixed number $k \in [m]$ of experts—the *granularity* of the MoE layer. More precisely, a score vector $s(x) \in \mathbb{R}^m$ determines the active index set $S(x) \subset [m]$ of size k, and $G(x)$ then assigns non-zero weights only to the experts in $S(x)$. In the sparsely-gated layer of [Sha+17], a common template is

$$G(x) = \mathrm{softmax}\left(\text{Top-}k\big(s(x) + \varepsilon(x)\big)\right), \tag{8.5.2}$$

where ε is (optional) noise and Top-k masks all but the k largest coordinates, i.e., sets them to $-\infty$ so they vanish upon applying the softmax. Equation (8.5.2) corresponds to masking before the softmax; other implementations apply a Top-k mask after a softmax and then renormalize.

In this convention, the active index set $S(x)$ is chosen hard—via the Top-k operation—while the weights on the selected experts satisfy $\sum_i g_i(x) = 1$. Alternatively, one could instead simplify the gate by taking uniform weights on the active set, e.g.,

$$g_i(x) = \frac{1}{k} \mathbf{1}_{S(x)}(i)$$

(equivalently, up to a rescaling of the experts, $g_i(x) \in \{0, 1\}$ with exactly k ones, as in [BR25]).

A widely used choice for the scores is *linear routing*, where one sets

$$s_i(x) \coloneqq r_i \cdot x \tag{8.5.3}$$

for some learnable *routing vectors* $r_1, \dots, r_m \in \mathbb{R}^d$. Then $G(x)$ activates in (8.5.2) those indices corresponding to the k largest scores (via Top-k). Under this scheme, we define the *routing cells*

$$\mathcal{U}_S \coloneqq \Big\{ x \in \mathbb{R}^d \mid \min_{i \in S} r_i \cdot x > \max_{j \notin S} r_j \cdot x \Big\}, \qquad \text{for } S \in \mathcal{S}_k \coloneqq \{T \subset [m] \mid |T| = k\}. \tag{8.5.4}$$

The active index set $S(x)$ is constant for every $x \in \mathcal{U}_S$. Thus, away from the tie set where some scores coincide, hard Top-k routing admits an explicit piecewise representation on $\mathbb{R}^d$. If one uses uniform weights on the active set, then

$$g_i(x) = \frac{1}{k}\, \mathbf{1}_{S(x)}(i) = \frac{1}{k} \sum_{\substack{S \in \mathcal{S}_k \\ i \in S}} \mathbf{1}_{\mathcal{U}_S}(x), \qquad x \in \mathbb{R}^d,$$

[5] The published record [Jac+91] notes that the idea was presented earlier by Jacobs and Hinton (Connectionist Summer School, 1988).

and therefore

$$\mathrm{MoE}(x) = \frac{1}{k} \sum_{S \in \mathcal{S}_k} \mathbf{1}_{\mathcal{U}_S}(x) \sum_{i \in S} E_i(x), \qquad x \in \mathbb{R}^d. \tag{8.5.5}$$

This decomposition makes explicit how the routing mechanism isolates the activation of cells in the left sum, decoupling it from the aggregation of the selected experts in the right sum.

Back to the sphere

When incorporating MoEs into the Transformer, these layers typically replace the perceptron blocks rather than the self-attention mechanism, a design choice motivated by both efficiency and architectural integrity [Sha+17].

Returning to the analytical framework of this chapter, the relevant object is the drift induced by the MoE layer on $\mathbb{S}^{d-1}$. Consider a collection of tangent expert fields,

$$v_{\theta_i} : \mathbb{S}^{d-1} \to \mathsf{T}\mathbb{S}^{d-1}, \qquad i \in [m],$$

and let $g_i : \mathbb{S}^{d-1} \to \mathbb{R}_{\geq 0}$ denote the restriction of the gating weights to the sphere. The induced drift is

$$v_{\boldsymbol{\theta},G}(x) \coloneqq \sum_{i=1}^{m} g_i(x)\, v_{\theta_i}(x), \qquad x \in \mathbb{S}^{d-1}, \tag{8.5.6}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)$.

Assume the scores are generated by (8.5.3) with $r_i \in \mathbb{S}^{d-1}$ for every i. Since $r_i \cdot x = \cos\big(\mathrm{dist}_{\mathbb{S}}(r_i, x)\big)$, maximizing $r_i \cdot x$ is equivalent to minimizing the geodesic distance to r_i. Thus, away from ties, each region

$$\mathcal{U}_S \cap \mathbb{S}^{d-1} = \Big\{ x \in \mathbb{S}^{d-1} \mid \min_{i \in S} r_i \cdot x > \max_{j \notin S} r_j \cdot x \Big\} \tag{8.5.7}$$

contains exactly those points where the set of k nearest sites (in geodesic distance) is S. This is precisely an order-k spherical *Voronoi diagram* (see Figure 8.14 and [SH75; Lee82; SS00; CPH22]; see also [CE87] for efficient Euclidean constructions).

For a granularity of $k = 1$, one recovers the *Voronoi cells* studied in **Subsection 4.3.2** of **Chapter 4**, now restricted to $\mathbb{S}^{d-1}$:

$$\mathcal{U}_{\{i\}} \cap \mathbb{S}^{d-1} = \big\{ x \in \mathbb{S}^{d-1} \mid r_i \cdot x > r_j \cdot x \quad \text{for all } j \neq i \big\}.$$

For general $k \geq 1$, the defining inequalities in (8.5.7) show that each order-k cell is an intersection of hemispheres:

$$\mathcal{U}_S \cap \mathbb{S}^{d-1} = \bigcap_{\substack{i \in S \\ j \notin S}} \big\{ x \in \mathbb{S}^{d-1} \mid (r_i - r_j) \cdot x > 0 \big\},$$

and is therefore a spherically convex polyhedral region.

A conservative routing mechanism

We now return to the variational viewpoint. Assume each expert is a gradient in the sense of (8.2.13): for each $i \in [m]$, there exists $\mathsf{V}_{\theta_i} : \mathbb{S}^{d-1} \to \mathbb{R}$ such that $v_{\theta_i}(x) = \frac{1}{2} \nabla_{\mathbb{S}} \mathsf{V}_{\theta_i}(x)$. Then (8.5.6) becomes

$$v_{\boldsymbol{\theta},G}(x) = \frac{1}{2} \sum_{i=1}^{m} g_i(x)\, \nabla_{\mathbb{S}} \mathsf{V}_{\theta_i}(x), \qquad x \in \mathbb{S}^{d-1}. \tag{8.5.8}$$

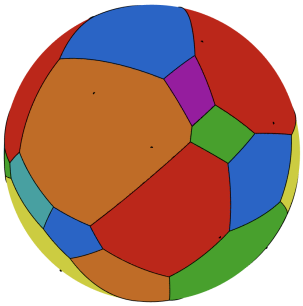


Figure 8.14: Order-k spherical routing diagram for $m = 10$ routing vectors (black dots) and granularity $k = 2$ in $\mathbb{S}^2$.

A natural question is whether there exists a global scalar potential $\mathsf{V}_{\boldsymbol{\theta},G}$ such that $v_{\boldsymbol{\theta},G} = \frac{1}{2}\nabla_{\mathbb{S}}\mathsf{V}_{\boldsymbol{\theta},G}$.

For Top-k routing, the gates g_i are discontinuous across the boundaries of the cells $\mathcal{U}_S$. Consequently, $v_{\boldsymbol{\theta},G}$ is only piecewise smooth and cannot be the gradient of a global $\mathscr{C}^1$-potential unless the piecewise potentials match exactly across boundaries—a very strong compatibility requirement.

This limitation motivates considering conservative softenings designed to be variational by construction. One such option is to couple the routing directly to the expert potentials themselves. Fix $k \in [m]$ and $\mathcal{S}_k = \{S \subset [m] \mid |S| = k\}$. For $\tau > 0$, define the weights

$$q_S^\tau(x) \coloneqq \frac{\exp\bigl(\tau^{-1}\sum_{i\in S}\mathsf{V}_{\theta_i}(x)\bigr)}{\sum_{T\in\mathcal{S}_k}\exp\bigl(\tau^{-1}\sum_{j\in T}\mathsf{V}_{\theta_j}(x)\bigr)}, \qquad S \in \mathcal{S}_k, \tag{8.5.9}$$

and the associated drift

$$v_{\boldsymbol{\theta}}^\tau(x) \coloneqq \frac{1}{2}\sum_{S\in\mathcal{S}_k} q_S^\tau(x)\sum_{i\in S}\nabla_{\mathbb{S}}\mathsf{V}_{\theta_i}(x). \tag{8.5.10}$$

This mechanism is conservative: defining the log-sum-exp potential

$$\mathsf{V}_{\boldsymbol{\theta}}^\tau(x) \coloneqq \tau\log\Bigl(\sum_{S\in\mathcal{S}_k}\exp\bigl(\tau^{-1}\sum_{i\in S}\mathsf{V}_{\theta_i}(x)\bigr)\Bigr), \tag{8.5.11}$$

a direct differentiation gives

$$\frac{1}{2}\nabla_{\mathbb{S}}\mathsf{V}_{\boldsymbol{\theta}}^\tau(x) = \frac{1}{2}\sum_{S\in\mathcal{S}_k} q_S^\tau(x)\sum_{i\in S}\nabla_{\mathbb{S}}\mathsf{V}_{\theta_i}(x) = v_{\boldsymbol{\theta}}^\tau.$$

As $\tau \to 0^+$, the log-sum-exp in (8.5.11) converges pointwise to $\max_{S\in\mathcal{S}_k}\sum_{i\in S}\mathsf{V}_{\theta_i}(x)$, so this formulation can be viewed as an entropic regularization of Top-k selection driven by expert energies. The construction is conceptual rather than algorithmic: for general k, it involves $\binom{m}{k}$ subsets.

Under potential-based routing, the MoE block contributes the external potential $\mathsf{V}_{\boldsymbol{\theta}}^\tau$ to the energy functional. In particular, if $\mathsf{E}_\beta[\mu]$ denotes the (attention) interaction energy used throughout the chapter, one may consider

$$\mathsf{E}_{\beta,\boldsymbol{\theta}}^\tau[\mu] \coloneqq \mathsf{E}_\beta[\mu] + \frac{1}{2}\int \mathsf{V}_{\boldsymbol{\theta}}^\tau(x)\,\mathrm{d}\mu(x), \tag{8.5.12}$$

which keeps the dynamics within the same Wasserstein-gradient-flow framework.

8.6 Perspectives

Our results on the mean-field landscape of Transformers suggest some natural directions.

- **Non-conservative perceptron.** Our stationary analysis relies crucially on the fact that the perceptron drift is conservative, so that the dynamics admit an energy functional and stationary states satisfy the pointwise condition (8.2.3). Generally, the MLP does not from a scalar potential, and the relevant stationary equation is just (8.2.4). Understanding whether atomicity, clustering, or metastability persist in this genuinely non-variational regime is an open problem.
- **Long-time convergence.** The present chapter classifies stationary states, but does not prove convergence of the mean-field PDE toward them. A natural next step is to identify dynamical conditions ensuring convergence to stationary measures.

The MoE discussion above suggests several directions where the mean-field and variational viewpoints of this chapter could be extended:

- **Hard routing as a nonsmooth limit.** Top-k MoE yields a piecewise-smooth drift on $\mathbb{S}^{d-1}$ with discontinuities across routing boundaries. A natural goal is to develop a well-posed mean-field theory for the limiting continuity equation with discontinuous velocity fields, and to relate it to a nonsmooth variational structure.
- **Low-temperature asymptotics.** For conservative soft routing (e.g., log-sum-exp couplings), one can ask for rigorous convergence results as $\tau \to 0^+$: do the corresponding Wasserstein gradient flows converge to meaningful limiting dynamics, and how do boundary layers near routing walls affect transient clustering and metastability?
- **Interaction between MoE and discrete equilibria.** Our results show that a perceptron potential enforces singular and often atomic equilibria. It is natural to quantify how an MoE multi-well potential changes the number, masses, and separation scales of atoms, and whether analogues of Theorem 8.3.7 and Corollary 8.3.8 persist with explicit dependence on (m, k, τ).
- **Routing coupled to the population.** In practice, MoE gates are trained with balancing terms depending on the global token distribution. At the mean-field level, this corresponds to gates $g_i = g_i(x, \mu)$, introducing a new nonlocal dependence beyond attention. Identifying when such couplings admit an energy functional and a Wasserstein-gradient-flow structure would connect architectural constraints directly to variational principles.

8.7 Proofs

8.7.1 Preliminaries

To analyze the dynamics on the sphere, we consider the spherical harmonics, defined as the restrictions of homogeneous harmonic polynomials in $\mathbb{R}^d$ to $\mathbb{S}^{d-1}$. Together, these spherical harmonics form an orthonormal basis for $L^2(\mathbb{S}^{d-1}, \sigma_d)$. Furthermore, the harmonics of a specific degree j are the eigenfunctions of the Laplace–Beltrami operator $\Delta_{\mathbb{S}^{d-1}}$ with eigenvalues $-j(j+d-2)$.

Let $\{Y_{j\ell}\}_{1 \leqslant \ell \leqslant N_j}$ denote this basis for the subspace of degree j, where the dimension N_j is given by

$$N_j \coloneqq \binom{d+j-1}{j} - \binom{d+j-3}{j-2}.$$

Under the standard convention that $\binom{n}{k} = 0$ for $k < 0$, we have $N_0 = 1$ and $N_1 = d$.

In the context of self-attention, the interaction kernel $e^{\beta x\cdot y}$ acts as an integral operator on the sphere. By the Funk–Hecke formula [Mül66], the spherical harmonics are precisely its eigenfunctions:

$$\int_{\mathbb{S}^{d-1}} e^{\beta x\cdot y}\, Y_{j\ell}(x)\,\mathrm{d}\sigma_d(x) = \lambda_j(\beta)\, Y_{j\ell}(y). \tag{8.7.1}$$

For $\beta > 0$, the eigenvalues $\lambda_j(\beta)$ are strictly positive and explicitly given by

$$\lambda_j(\beta) \coloneqq \Gamma\left(\frac{d}{2}\right)\left(\frac{2}{\beta}\right)^{\frac{d}{2}-1} I_{j+\frac{d}{2}-1}(\beta), \qquad j \geqslant 0,$$

where I_α is the modified Bessel function of the first kind of order α, characterized by the series expansion

$$I_\alpha(z) \coloneqq \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m+\alpha+1)}\left(\frac{z}{2}\right)^{2m+\alpha} \qquad (z \in \mathbb{R}).$$

This spectral decomposition allows us to characterize the evolution of the particle system by tracking how the attention mechanism acts on each frequency j of the distribution.

Lemma 8.7.1. *For any $\beta > 0$, the map $\mu \in \mathcal{P}(\mathbb{S}^{d-1}) \mapsto f^\mu \in \mathscr{C}^\infty(\mathbb{S}^{d-1})$, defined by*

$$f^\mu(x) \coloneqq \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot y}\,\mathrm{d}\mu(y),$$

is injective. Moreover, if $f^\mu = \sum_{j=0}^{\infty}\sum_{\ell=1}^{N_j} \widehat{f}^\mu_{j\ell}\, Y_{j\ell}$ is its spherical harmonic expansion, the following properties hold:

(i) *μ is absolutely continuous with respect to σ_d with a density $\rho \coloneqq \frac{\mathrm{d}\mu}{\mathrm{d}\sigma_d} \in L^2(\sigma_d)$ if and only if*

$$\sum_{j=0}^{\infty}\sum_{\ell=1}^{N_j}\left|\frac{\widehat{f}^\mu_{j\ell}}{\lambda_j(\beta)}\right|^2 < \infty.$$

In this case, the density is explicitly given in $L^2(\sigma_d)$ by

$$\rho(x) = \sum_{j=0}^{\infty}\sum_{\ell=1}^{N_j}\frac{\widehat{f}^\mu_{j\ell}}{\lambda_j(\beta)} Y_{j\ell}(x). \tag{8.7.2}$$

Consequently, f^μ is a spherical polynomial of degree at most k if and only if μ has a density that is a spherical polynomial of degree at most k.

(ii) *f^μ is an even function ($f^\mu(x) = f^\mu(-x)$) if and only if μ is antipodally symmetric, meaning $\mu(A) = \mu(-A)$ for every Borel set $A \subset \mathbb{S}^{d-1}$.*

Proof. We divide the proof into three parts.

Injectivity. Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{S}^{d-1})$ and assume

$$\int e^{\beta x\cdot y}\,\mathrm{d}\mu_1(y) = \int e^{\beta x\cdot y}\,\mathrm{d}\mu_2(y) \quad \text{for all } x \in \mathbb{S}^{d-1}.$$

Set $\nu \coloneqq \mu_1 - \mu_2$ and define $f^\nu(x) \coloneqq \int e^{\beta x\cdot y}\,\mathrm{d}\nu(y)$; then $f^\nu \equiv 0$. For each (j,ℓ), using Fubini's theorem and (8.7.1), we compute the Fourier coefficients of f^ν:

$$0 = \widehat{f}^\nu_{j\ell} = \int f^\nu(x) Y_{j\ell}(x)\,\mathrm{d}\sigma_d(x) = \int\left(\int e^{\beta x\cdot y}\, Y_{j\ell}(x)\,\mathrm{d}\sigma_d(x)\right)\mathrm{d}\nu(y)$$

$$= \lambda_j(\beta) \int Y_{j\ell}(y)\, \mathrm{d}\nu(y). \tag{8.7.3}$$

Since $\lambda_j(\beta) > 0$, it follows that $m_{j\ell}(\nu) := \int Y_{j\ell}\, \mathrm{d}\nu = 0$ for all j, ℓ. Hence, for any finite spherical polynomial $P = \sum a_{j\ell} Y_{j\ell}$, we have

$$\int P\, \mathrm{d}\nu = \sum a_{j\ell}\, m_{j\ell}(\nu) = 0.$$

By the Stone–Weierstrass theorem, spherical polynomials are uniformly dense in $\mathscr{C}^0(\mathbb{S}^{d-1})$. Since ν is a finite signed measure, uniform approximation implies $\int h\, \mathrm{d}\nu = 0$ for all $h \in \mathscr{C}^0(\mathbb{S}^{d-1})$. By the uniqueness of the Riesz representation theorem, $\nu = 0$, and thus $\mu_1 = \mu_2$.

For later use, note that applying the computation of (8.7.3) to μ instead of ν yields the relation

$$\widehat{f}^{\mu}_{j\ell} = \lambda_j(\beta)\, m_{j\ell}, \qquad \text{where} \quad m_{j\ell} := \int Y_{j\ell}\, \mathrm{d}\mu. \tag{8.7.4}$$

Proof of (i). Assume $\sum_{j,\ell} \left|\widehat{f}^{\mu}_{j\ell}/\lambda_j(\beta)\right|^2 < \infty$. Since $\{Y_{j\ell}\}_{j,\ell}$ is an orthonormal basis of $L^2(\sigma_d)$, there exists a unique function $g \in L^2(\sigma_d)$ such that

$$g(x) := \sum_{j=0}^{\infty} \sum_{\ell=1}^{N_j} \frac{\widehat{f}^{\mu}_{j\ell}}{\lambda_j(\beta)}\, Y_{j\ell}(x)$$

with convergence in $L^2(\sigma_d)$. Since $g \in L^2(\sigma_d)$ and σ_d is finite, we have $g \in L^1(\sigma_d)$, so $g\,\sigma_d$ defines a finite signed measure on $\mathbb{S}^{d-1}$. For any finite spherical polynomial $P = \sum a_{j\ell} Y_{j\ell}$, using (8.7.4),

$$\int P\, \mathrm{d}\mu = \sum a_{j\ell} m_{j\ell} = \sum a_{j\ell} \frac{\widehat{f}^{\mu}_{j\ell}}{\lambda_j(\beta)} = \int P\, g\, \mathrm{d}\sigma_d.$$

Since spherical polynomials are dense in $\mathscr{C}^0(\mathbb{S}^{d-1})$, this identity extends to all $h \in \mathscr{C}^0(\mathbb{S}^{d-1})$. By the uniqueness of the Riesz representation theorem, $\mathrm{d}\mu = g\, \mathrm{d}\sigma_d$, which yields (8.7.2). Conversely, if μ has a density $\rho \in L^2(\sigma_d)$ so that $\mathrm{d}\mu = \rho\, \mathrm{d}\sigma_d$, then $m_{j\ell} = \langle \rho, Y_{j\ell} \rangle_{L^2}$. Parseval's identity then gives

$$\sum_{j,\ell} \left| \frac{\widehat{f}^{\mu}_{j\ell}}{\lambda_j(\beta)} \right|^2 = \sum_{j,\ell} |m_{j\ell}|^2 = \|\rho\|^2_{L^2(\sigma_d)} < \infty.$$

Finally, f^{μ} is a spherical polynomial of degree at most k if and only if $\widehat{f}^{\mu}_{j\ell} = 0$ for $j > k$, which, by (8.7.2), is equivalent to the density ρ being a spherical polynomial of degree at most k.

Proof of (ii). If μ is antipodally symmetric, then for every $x \in \mathbb{S}^{d-1}$,

$$f^{\mu}(-x) = \int_{\mathbb{S}^{d-1}} e^{\beta(-x)\cdot y}\, \mathrm{d}\mu(y) = \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot(-y)}\, \mathrm{d}\mu(y) = \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot y}\, \mathrm{d}\mu(y) = f^{\mu}(x),$$

so f^{μ} is an even function. Conversely, assume $f^{\mu}(-x) = f^{\mu}(x)$ for all $x \in \mathbb{S}^{d-1}$. Define the antipodal pushforward $\mu^- := (-\mathrm{Id})_{\#}\mu$. Then, for every $x \in \mathbb{S}^{d-1}$,

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot y}\, \mathrm{d}\mu^-(y) &= \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot(-y)}\, \mathrm{d}\mu(y) = \int_{\mathbb{S}^{d-1}} e^{\beta(-x)\cdot y}\, \mathrm{d}\mu(y) \\ &= f^{\mu}(-x) = f^{\mu}(x) = \int_{\mathbb{S}^{d-1}} e^{\beta x\cdot y}\, \mathrm{d}\mu(y). \end{aligned}$$

By the injectivity proved above, we conclude that $\mu^- = \mu$, meaning μ is antipodally symmetric. □

Remark 8.7.2. Fix a symmetric invertible matrix $B \in \mathscr{M}_{d\times d}(\mathbb{R})$ and define, for $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$,

$$f_B^\mu(x) \coloneqq \int e^{x^\top By}\,\mathrm{d}\mu(y), \qquad x \in \mathbb{S}^{d-1}.$$

Then the same conclusions remain valid for f_B^μ, although the injectivity proof from Lemma 8.7.1 requires a different argument.

Injectivity. Let $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{S}^{d-1})$ and assume

$$\int e^{x^\top By}\,\mathrm{d}\mu_1(y) = \int e^{x^\top By}\,\mathrm{d}\mu_2(y) \qquad \text{for all } x \in \mathbb{S}^{d-1}. \tag{8.7.5}$$

Set $\nu \coloneqq \mu_1 - \mu_2$ and define

$$F_\nu(z) \coloneqq \int_{\mathbb{S}^{d-1}} e^{z\cdot y}\,\mathrm{d}\nu(y), \qquad z \in \mathbb{R}^d.$$

Then (8.7.5) implies that $F_\nu \equiv 0$ on

$$\partial\Omega \coloneqq \{B^\top x \mid \|x\| = 1\},$$

the boundary of the ellipsoid $\Omega \coloneqq \{B^\top x \mid \|x\| < 1\}$. Moreover, differentiating under the integral sign shows that $(\Delta - 1)F_\nu = 0$ on $\mathbb{R}^d$, since for any fixed $y \in \mathbb{S}^{d-1}$, the map $z \mapsto e^{z\cdot y}$ satisfies

$$\Delta(e^{z\cdot y}) = \|y\|^2 e^{z\cdot y} = e^{z\cdot y}.$$

By uniqueness for the Dirichlet problem for $\Delta - 1$ on Ω, we obtain $F_\nu \equiv 0$ in Ω. Hence F_ν vanishes in a neighborhood of 0, and therefore on all of $\mathbb{R}^d$ by real-analyticity. Expanding at 0, we get $\int P(y)\,\mathrm{d}\nu(y) = 0$ for every polynomial P. Since the restrictions of polynomials to $\mathbb{S}^{d-1}$ are uniformly dense in $\mathscr{C}^0(\mathbb{S}^{d-1})$, it follows that $\nu = 0$, and thus $\mu_1 = \mu_2$.

Parity. The proof that f_B^μ is even if and only if μ is antipodally symmetric is exactly the same as in Lemma 8.7.1**(ii)**, using the injectivity of f_B^μ established above.

Lemma 8.7.3. *Fix $d = 2$ and $\beta > 0$. Let*

$$\mathsf{K}_\beta(\varphi) \coloneqq e^{\beta\cos\varphi}, \qquad \varphi \in (-\pi, \pi],$$

and define

$$\varphi_c(\beta) \coloneqq \arccos\left(\frac{\sqrt{1+4\beta^2}-1}{2\beta}\right) \in \left(0, \frac{\pi}{2}\right).$$

Then

$$\varphi_c(\beta) = \beta^{-1/2} + O(\beta^{-3/2}) \quad \text{as } \beta\to\infty, \qquad \text{and} \qquad \varphi_c(\beta) \to \frac{\pi}{2} \quad \text{as } \beta \to 0^+.$$

Moreover, K_β is strictly concave on $(-\varphi_c(\beta), \varphi_c(\beta))$.

For every $\lambda \in (0,1)$ there exists $\beta_0 = \beta_0(\lambda) > 0$ such that

$$\sup_{|\varphi|\le\lambda\varphi_c(\beta)} \mathsf{K}_\beta''(\varphi) \le -e^{-\lambda^2/2}\,\frac{1-\lambda^2}{2}\,\beta e^\beta \qquad \text{for all } \beta \ge \beta_0(\lambda), \tag{8.7.6}$$

and there exists $\beta_1 > 0$ such that

$$\max_{\varphi\in[\varphi_c(\beta),\pi]} \mathsf{K}_\beta''(\varphi) \le 2\beta e^{\beta-3/2} \qquad \text{for all } \beta \ge \beta_1. \tag{8.7.7}$$

Proof. A direct computation gives

$$\mathsf{K}''_\beta(\varphi) = e^{\beta\cos\varphi}\big(\beta^2\sin^2\varphi - \beta\cos\varphi\big). \tag{8.7.8}$$

Hence $\mathsf{K}''_\beta(\varphi) = 0$ if and only if

$$\beta\cos^2\varphi + \cos\varphi - \beta = 0.$$

The positive root is

$$\cos\varphi = \frac{\sqrt{1+4\beta^2}-1}{2\beta}, \tag{8.7.9}$$

while the other root is < -1. Therefore $\varphi_c(\beta)$ is well defined and belongs to $(0,\pi/2)$. From (8.7.9) we obtain

$$\frac{\sqrt{1+4\beta^2}-1}{2\beta} = 1 - \frac{1}{2\beta} + O(\beta^{-2}) \qquad \text{as } \beta\to\infty,$$

and since $\cos\varphi = 1 - \varphi^2/2 + O(\varphi^4)$ as $\varphi\to 0$, it follows that

$$\varphi_c(\beta) = \beta^{-1/2} + O(\beta^{-3/2}) \qquad \text{as } \beta\to\infty.$$

Similarly,

$$\frac{\sqrt{1+4\beta^2}-1}{2\beta} = \beta + O(\beta^3) \qquad \text{as } \beta\to 0^+,$$

hence $\cos\varphi_c(\beta)\to 0$ and therefore $\varphi_c(\beta)\to\pi/2$ as $\beta\to 0^+$.

Since $\mathsf{K}''_\beta(0) = -\beta e^\beta < 0$, and $\varphi_c(\beta)$ is the first positive zero of K''_β, we conclude that K_β is strictly concave on $(-\varphi_c(\beta),\varphi_c(\beta))$.

To prove (8.7.6), we first show that K''_β is strictly increasing on $[0,\varphi_c(\beta)]$. Differentiating (8.7.8), we find

$$\mathsf{K}'''_\beta(\varphi) = -\beta\sin\varphi\,\mathsf{K}''_\beta(\varphi) + e^{\beta\cos\varphi}\beta\sin\varphi(2\beta\cos\varphi+1). \tag{8.7.10}$$

For $0<\varphi<\varphi_c(\beta)$, we have $\sin\varphi>0$, $\cos\varphi>0$, and $\mathsf{K}''_\beta(\varphi)<0$, so both terms on the right-hand side of (8.7.10) are positive. Thus

$$\mathsf{K}'''_\beta(\varphi) > 0 \qquad \text{for } 0<\varphi<\varphi_c(\beta).$$

Since K''_β is even, it follows that

$$\sup_{|\varphi|\le\lambda\varphi_c(\beta)} \mathsf{K}''_\beta(\varphi) = \mathsf{K}''_\beta(\lambda\varphi_c(\beta)).$$

Using the expansion of $\varphi_c(\beta)$ obtained above, we have

$$\cos(\lambda\varphi_c(\beta)) = 1 - \frac{\lambda^2}{2\beta} + O(\beta^{-2}), \qquad \sin^2(\lambda\varphi_c(\beta)) = \frac{\lambda^2}{\beta} + O(\beta^{-2}).$$

Substituting into (8.7.8) gives

$$\begin{aligned}
\mathsf{K}''_\beta(\lambda\varphi_c(\beta)) &= e^{\beta\cos(\lambda\varphi_c(\beta))}\big(\beta^2\sin^2(\lambda\varphi_c(\beta)) - \beta\cos(\lambda\varphi_c(\beta))\big)\\
&= e^{\beta-\lambda^2/2}\big(1+O(\beta^{-1})\big)\left(\beta(\lambda^2-1) + \frac{\lambda^2}{2} + O(1)\right)\\
&= -(1-\lambda^2)\,e^{\beta-\lambda^2/2}\,\beta\big(1+O(\beta^{-1})\big).
\end{aligned}$$

Therefore, for β sufficiently large,

$$\mathsf{K}_\beta''(\lambda\varphi_c(\beta)) \le -e^{-\lambda^2/2}\,\frac{1-\lambda^2}{2}\,\beta e^\beta,$$

which proves (8.7.6).

We now prove (8.7.7). Rewriting (8.7.10), we obtain

$$\mathsf{K}_\beta'''(\varphi) = \beta e^{\beta\cos\varphi} q_\beta(\cos\varphi)\sin\varphi,$$

where

$$q_\beta(t) \coloneqq \beta^2 t^2 + 3\beta t - \beta^2 + 1.$$

Thus the critical points of K_β'' in $(0,\pi)$ are precisely the points where $q_\beta(\cos\varphi) = 0$. Now

$$q_\beta(-1) = 1 - 3\beta < 0 \qquad \text{for } \beta > \tfrac{1}{3},$$

while, using that $\mathsf{K}_\beta''(\varphi_c(\beta)) = 0$,

$$q_\beta(\cos\varphi_c(\beta)) = 2\beta\cos\varphi_c(\beta) + 1 > 0.$$

The roots of q_β are

$$\frac{-3 \pm \sqrt{4\beta^2+5}}{2\beta}.$$

Since $\frac{-3-\sqrt{4\beta^2+5}}{2\beta} < -1$, the unique root in the interval $(-1, \cos\varphi_c(\beta))$ is

$$t_* = \frac{-3+\sqrt{4\beta^2+5}}{2\beta}. \tag{8.7.11}$$

Let $\varphi_* \in (\varphi_c(\beta), \pi)$ be defined by $\cos\varphi_* = t_*$. Then

$$\mathsf{K}_\beta'''(\varphi) > 0 \quad \text{for } \varphi \in (\varphi_c(\beta), \varphi_*), \qquad \mathsf{K}_\beta'''(\varphi) < 0 \quad \text{for } \varphi \in (\varphi_*, \pi),$$

so K_β'' attains its maximum on $[\varphi_c(\beta), \pi]$ at φ_*. Since $q_\beta(t_*) = 0$, we can express the quadratic term $\beta^2(1-t_*^2)$ as

$$\beta^2(1-t_*^2) = 3\beta t_* + 1.$$

Substituting this relation into (8.7.8), we obtain

$$\mathsf{K}_\beta''(\varphi_*) = e^{\beta t_*}\big(2\beta t_* + 1\big).$$

Finally, taking the Taylor expansion of $\sqrt{4\beta^2+5} = 2\beta\sqrt{1+\frac{5}{4\beta^2}}$ in (8.7.11)[6] gives

$$t_* = 1 - \frac{3}{2\beta} + \frac{5}{8\beta^2} + O(\beta^{-4}),$$

so

$$\begin{aligned}\mathsf{K}_\beta''(\varphi_*) &= e^{\beta-\frac{3}{2}}\left(1+\frac{5}{8\beta}+O(\beta^{-2})\right)2\beta\left(1-\frac{1}{\beta}+\frac{5}{8\beta^2}+O(\beta^{-4})\right)\\ &= 2\beta e^{\beta-\frac{3}{2}}\left(1-\frac{3}{8\beta}+O(\beta^{-2})\right).\end{aligned}$$

In particular, for β sufficiently large,

$$\mathsf{K}_\beta''(\varphi_*) \le 2\beta e^{\beta-\frac{3}{2}},$$

which proves (8.7.7). □

[6] Recall that $\sqrt{1+x} = 1 + \frac{x}{2} - \frac{x^2}{8} + O(x^3)$ as $x \to 0$.

8.7.2 Proof of Theorem 8.3.1 and Theorem 8.3.3

Proof of Theorem 8.3.1. Set $x(\varphi) = (\cos\varphi, \sin\varphi)$ and $\tau(\varphi) = (-\sin\varphi, \cos\varphi)$. For $\varphi \in [0, 2\pi)$ define

$$f_\beta(\varphi) \coloneqq \int e^{\beta\cos(\psi-\varphi)} \sin(\psi-\varphi)\, \mathrm{d}\mu(\psi),$$
$$f_\theta(\varphi) \coloneqq \omega_1\, (a_1 \cdot x(\varphi))_+\, a_1 \cdot \tau(\varphi) + \omega_2\, (a_2 \cdot x(\varphi))_+\, a_2 \cdot \tau(\varphi),$$

and set $g \coloneqq f_\beta + f_\theta$, corresponding to the gradient $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu] + v_\theta$ expressed in angular coordinates.

Let $\mathscr{Z} \coloneqq \{\varphi \in [0, 2\pi) \mid a_j \cdot x(\varphi) = 0 \text{ for some } j\}$. Clearly, the set $\mathscr{Z}$ is finite. On any connected component $I \subset [0, 2\pi) \setminus \mathscr{Z}$, the active set

$$J_I \coloneqq \{j \mid \ a_j \cdot x(\varphi) > 0 \text{ for all } \varphi \in I\}$$

is constant, and for $\varphi \in I$ we have

$$(a_j \cdot x(\varphi))_+ = \mathbf{1}_{J_I}(j)\, a_j \cdot x(\varphi), \qquad \frac{\mathrm{d}}{\mathrm{d}\varphi} (a_j \cdot x(\varphi))_+ = \mathbf{1}_{J_I}(j)\, a_j \cdot \tau(\varphi).$$

Thus f_θ is real-analytic on each I.

On the other hand, since

$$\mathsf{K}_\beta(\psi) \coloneqq e^{\beta\cos\psi} = I_0(\beta) + 2\sum_{n\geq 1} I_n(\beta)\cos(n\psi),$$

its Fourier series is absolutely convergent. Hence the convolution with the measure μ,

$$(\mathsf{K}_\beta * \mu)(\varphi) \coloneqq \int \mathsf{K}_\beta(\varphi - \psi)\, \mathrm{d}\mu(\psi),$$

is real–analytic on $\mathbb{S}^1$, and $f_\beta(\varphi) = \frac{1}{\beta}\frac{\mathrm{d}}{\mathrm{d}\varphi}(\mathsf{K}_\beta * \mu)(\varphi)$ is real–analytic as well.

All in all, we deduce that g is continuous on $[0, 2\pi)$ and real–analytic on each I.

For each component I, from (8.2.14) we get

$$\operatorname{supp}\mu \cap I \subset \{\varphi \in I \mid g(\varphi) = 0\}. \tag{8.7.12}$$

Assume, for a contradiction, that $\operatorname{supp}\mu$ is infinite. Then $\operatorname{supp}\mu \cap I$ must be infinite for at least one of the finitely many components I. Define the global real-analytic function

$$\widetilde{g}(\varphi) \coloneqq f_\beta(\varphi) + \sum_{j\in J_I} \omega_j \langle a_j, x(\varphi)\rangle \langle a_j, \tau(\varphi)\rangle,$$

which satisfies $\widetilde{g}(\varphi) = g(\varphi)$ for all $\varphi \in I$. Hence, by (8.7.12), the zero set of $\widetilde{g}$ contains infinitely many points in I and thus has an accumulation point in $\mathbb{S}^1$. By the identity theorem for real-analytic functions, it follows that $\widetilde{g} \equiv 0$ on $\mathbb{S}^1$. Equivalently,

$$f_\beta(\varphi) = -\sum_{j\in J_I} \omega_j \langle a_j, x(\varphi)\rangle \langle a_j, \tau(\varphi)\rangle \quad \text{for all } \varphi \in \mathbb{S}^1.$$

Integrating in φ, and recalling $f_\beta = \frac{1}{\beta}(\mathsf{K}_\beta * \mu)'$, we find that $g = H'$ on I, for

$$H(\varphi) \coloneqq \frac{1}{\beta}(\mathsf{K}_\beta * \mu)(\varphi) + \sum_j \frac{\omega_j}{2}\, (a_j \cdot x(\varphi))_+^2.$$

Since $\widetilde{g} \equiv 0$ on $\mathbb{S}^1$, we obtain the global identity

$$(\mathsf{K}_\beta * \mu)(\varphi) = C_I - \beta \sum_{j \in J_I} \frac{\omega_j}{2} \, (a_j \cdot x(\varphi))^2 \qquad (\varphi \in [0, 2\pi)) \tag{8.7.13}$$

for some constant C_I.

By Lemma 8.7.1**(i)**, (8.7.13) implies that μ is absolutely continuous with an L^2-density which is a trigonometric polynomial of degree ≤ 2. As a nonnegative real–analytic function on $\mathbb{S}^1$, this density cannot vanish on an open arc unless it is identically zero; hence $\operatorname{supp} \mu = \mathbb{S}^1$.

Let I' be a component adjacent to I, and let φ_* be the common boundary point. Since $\operatorname{supp} \mu = \mathbb{S}^1$, we have $\operatorname{supp} \mu \cap I'$ infinite, and repeating the argument used on I gives another global identity

$$(\mathsf{K}_\beta * \mu)(\varphi) = C_{I'} - \beta \sum_{j \in J_{I'}} \frac{\omega_j}{2} \, (a_j \cdot x(\varphi))^2, \qquad \varphi \in [0, 2\pi). \tag{8.7.14}$$

Let $D \coloneqq (J_I \cup J_{I'}) \setminus (J_I \cap J_{I'})$. Clearly $D \neq \emptyset$, since $J_I \neq J_{I'}$ because at least one index changes sign when crossing φ_*. Subtracting (8.7.14) from (8.7.13) we obtain

$$\frac{\beta}{2} \sum_{j \in D} \gamma_j \, \omega_j \, (a_j \cdot x(\varphi))^2 \;\equiv\; C_I - C_{I'} \qquad \varphi \in [0, 2\pi), \tag{8.7.15}$$

where $\gamma_j \in \{\pm 1\}$ records whether j is added or removed when passing from I to I'. For every $j \in D$ we have $a_j \cdot x(\varphi_*) = 0$, hence evaluating (8.7.15) at $\varphi = \varphi_*$ yields $C_I = C_{I'}$ and therefore

$$\sum_{j \in D} \gamma_j \, \omega_j \, (a_j \cdot x(\varphi))^2 \;\equiv\; 0 \qquad , \varphi \in [0, 2\pi). \tag{8.7.16}$$

The left-hand side of (8.7.16) is exactly the difference of $(\mathsf{V}_\theta \circ x)(\varphi)$ on I and I', which therefore vanishes identically on $[0, 2\pi)$. Hence $(\mathsf{V}_\theta \circ x)(\varphi)$ coincides on $\overline{I \cup I'}$ with a single trigonometric polynomial of degree at most 2.

Since this argument applies to every adjacent pair of components, it follows by connectivity that $\mathsf{V}_\theta \circ x$ is given in the full $\mathbb{S}^1$ by a single trigonometric polynomial of degree at most 2. In particular, V_θ is real–analytic on $\mathbb{S}^1$, contradicting our assumption. This shows that $\operatorname{supp} \mu$ is finite, and combining this with (8.7.12) yields that μ is purely atomic with finitely many atoms. □

Remark 8.7.4. The conclusion of Theorem 8.3.1 holds for any σ globally Lipschitz, piecewise polynomial (e.g. the leaky ReLU), provided V_θ is not real-analytic.

The proof applies mutatis mutandis. On each component I, the field f_θ now becomes a trigonometric polynomial of some finite degree. Consequently, if the support accumulates in I, analytic continuation implies that the global identity (8.7.13) holds with a polynomial of some finite degree. Lemma 8.7.1**(i)** then guarantees a global polynomial density, forcing $\operatorname{supp} \mu = \mathbb{S}^1$. The subtraction argument via (8.7.15) yields the contradiction exactly as before.

The same extension applies to Theorem 8.3.5.

Proof of Theorem 8.3.3. Let $\mu \in \mathcal{P}(\mathbb{S}^1)$ be a strict SOPD Wasserstein critical point of $\mathsf{E}_{\beta,\theta}$ as per (8.2.8). In particular, $\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}$ is well-defined at μ. Write $x(\varphi) = (\cos\varphi, \sin\varphi)$ and $\mathsf{K}_\beta(\psi) = e^{\beta \cos\psi}$, and define

$$H(\varphi) \coloneqq \frac{1}{\beta}(\mathsf{K}_\beta * \mu)(\varphi) + \frac{1}{2} \, (\mathsf{V}_\theta \circ x)\,(\varphi).$$

Since σ is real-analytic, the potential $\varphi \mapsto (\mathsf{V}_\theta \circ x)(\varphi)$ is real-analytic on $\mathbb{S}^1$. The convolution term $(\mathsf{K}_\beta * \mu)$ is also real-analytic. Hence H is real-analytic on $\mathbb{S}^1$.

Assume for contradiction that $\operatorname{supp}\mu$ is infinite. Then $\operatorname{supp}\mu$ has an accumulation point in $\mathbb{S}^1$. The first-order condition (8.2.14) forces H' to vanish on $\operatorname{supp}\mu$. Since H' is real-analytic and vanishes on a set with an accumulation point, the identity theorem implies $H' \equiv 0$ on $\mathbb{S}^1$. Consequently, $H'' \equiv 0$ on $\mathbb{S}^1$, which gives the global cancellation:

$$\frac{1}{\beta}\,(\mathsf{K}''_\beta * \mu)(\varphi) + \frac{1}{2}\,\partial^2_\varphi\,(\mathsf{V}_\theta \circ x)\,(\varphi) \;=\; 0 \qquad (\varphi \in \mathbb{S}^1). \tag{8.7.17}$$

To compute the second variation, let $\xi \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^1)$. Since the Wasserstein Hessian is well-defined at μ, the second-variation formula applies:

$$\begin{aligned}\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = \frac{1}{2\beta}\iint \mathsf{K}''_\beta(\varphi-\psi)\big(\xi(\varphi)-\xi(\psi)\big)^2\,\mathrm{d}\mu(\varphi)\,\mathrm{d}\mu(\psi)\\ + \frac{1}{2}\int \partial^2_\varphi\,(\mathsf{V}_\theta\circ x)\,(\varphi)\,\xi(\varphi)^2\,\mathrm{d}\mu(\varphi).\end{aligned} \tag{8.7.18}$$

Expanding the square using the symmetry of K''_β, we rewrite this as:

$$\begin{aligned}\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = \int\left[\frac{1}{\beta}(\mathsf{K}''_\beta * \mu)(\varphi) + \frac{1}{2}\partial^2_\varphi\,(\mathsf{V}_\theta\circ x)\,(\varphi)\right]\xi(\varphi)^2\,\mathrm{d}\mu(\varphi)\\ - \frac{1}{\beta}\iint \mathsf{K}''_\beta(\varphi-\psi)\,\xi(\varphi)\xi(\psi)\,\mathrm{d}\mu(\varphi)\,\mathrm{d}\mu(\psi).\end{aligned} \tag{8.7.19}$$

By (8.7.17), the term in brackets vanishes identically. Thus, the Hessian reduces to a bilinear form:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = -\frac{1}{\beta}\iint \mathsf{K}''_\beta(\varphi-\psi)\,\xi(\varphi)\xi(\psi)\,\mathrm{d}\mu(\varphi)\,\mathrm{d}\mu(\psi). \tag{8.7.20}$$

We construct a sequence of tangent vectors along which (8.7.20) is arbitrarily small, thereby contradicting strict positivity.

Fix a small arc $J_\delta \subset \mathbb{S}^1$ centered at an accumulation point of $\operatorname{supp}\mu$, with diameter $\le \delta$, so that $\mu(J_\delta) > 0$. Pick two disjoint subarcs $I_1, I_2 \subset J_\delta$ with $\mu(I_i) > 0$. Choose $\phi_i \in \mathscr{C}^\infty_c(I_i)$ non-constant, view it as a smooth function on $\mathbb{S}^1$ by extending it by zero outside I_i, and set $\eta_i := \phi_i'$. Then $\eta_i \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^1)$ (recall (8.2.5)) and η_i is supported in I_i. Moreover, by choosing ϕ_i so that ϕ_i' does not vanish on a neighborhood of some point in $\operatorname{supp}\mu \cap I_i$, we ensure $\|\eta_i\|_{L^2(\mu)} > 0$.

Set $m_i := \int \eta_i\,\mathrm{d}\mu$. If $(m_1, m_2) \neq (0,0)$, define

$$\xi_0 := m_2\,\eta_1 - m_1\,\eta_2,$$

so that $\int \xi_0\,\mathrm{d}\mu = m_2 m_1 - m_1 m_2 = 0$. If $m_1 = m_2 = 0$, simply set $\xi_0 := \eta_1$ (which already satisfies $\int \xi_0\,\mathrm{d}\mu = 0$). In either case, since η_1 and η_2 have disjoint supports and ξ_0 is not identically zero, we have $\xi_0 \not\equiv 0$.

Normalize $\xi_\delta := \xi_0/\|\xi_0\|_{L^2(\mu)}$. Then

$$\xi_\delta \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^1), \qquad \operatorname{supp}\xi_\delta \subset J_\delta \qquad \text{and} \qquad \|\xi_\delta\|_{L^2(\mu)} = 1.$$

Moreover, by construction, $\int \xi_\delta\,\mathrm{d}\mu = 0$. Evaluating (8.7.20) at $\xi = \xi_\delta$:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi_\delta,\xi_\delta) = -\frac{1}{\beta}\iint_{J_\delta\times J_\delta} \mathsf{K}''_\beta(\varphi-\psi)\,\xi_\delta(\varphi)\xi_\delta(\psi)\,\mathrm{d}\mu(\varphi)\,\mathrm{d}\mu(\psi).$$

Since $\int \xi_\delta \, \mathrm{d}\mu = 0$, adding a constant to the kernel does not change the integral. We replace $\mathsf{K}''_\beta(\varphi - \psi)$ by $\mathsf{K}''_\beta(\varphi - \psi) - \mathsf{K}''_\beta(0)$. Let $\omega(\delta) \coloneqq \sup_{|h| \le \delta} |\mathsf{K}''_\beta(h) - \mathsf{K}''_\beta(0)|$ for $\delta \in [0, \pi)$. Then,

$$\begin{aligned}\left|\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi_\delta, \xi_\delta)\right| &\le \frac{1}{\beta} \iint |\mathsf{K}''_\beta(\varphi - \psi) - \mathsf{K}''_\beta(0)| \cdot |\xi_\delta(\varphi)| \cdot |\xi_\delta(\psi)| \, \mathrm{d}\mu \, \mathrm{d}\mu \\ &\le \frac{\omega(\delta)}{\beta} \|\xi_\delta\|_{L^1}^2.\end{aligned}$$

Using Cauchy-Schwarz, $\|\xi_\delta\|_{L^1}^2 \le \|\xi_\delta\|_{L^2}^2 = 1$. Thus, the Hessian is bounded by $\omega(\delta)/\beta$, which tends to 0 as $\delta \to 0$. This implies

$$\inf \left\{ \mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi, \xi) \mid \ \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^1), \ \|\xi\|_{L^2(\mu)} = 1 \right\} = 0,$$

contradicting (8.2.8). Therefore $\operatorname{supp} \mu$ must be finite, and μ is purely atomic with finitely many atoms.

□

8.7.3 Proof of Theorem 8.3.5

Proof of Theorem 8.3.5. We split the proof into two parts.

Part I. Non-absolute continuity

We reuse the definitions of $(\mathscr{Z}, I, J_I)$ introduced in (8.3.5)–(8.3.6).

Let $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ solve (8.2.14). Assume first that $\sigma(s) = s_+$ and V_θ is not real-analytic on $\mathbb{S}^{d-1}$. On I the signs of $x \mapsto a_j \cdot x$ are fixed, hence

$$\mathsf{V}_\theta(x) = \sum_{j=1}^{d} \omega_j (a_j \cdot x)_+^2 = \sum_{j \in J_I} \omega_j (a_j \cdot x)^2, \qquad x \in I. \tag{8.7.21}$$

By expanding

$$e^{\beta x \cdot y} = \sum_{m \ge 0} \frac{\beta^m}{m!} (x \cdot y)^m$$

and using the Weierstrass M-test, we deduce that $\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu]$ is real-analytic on $\mathbb{S}^{d-1}$. Thus, $H \coloneqq \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu]$ and $g \coloneqq \nabla_{\mathbb{S}} H$ are both real-analytic on every I.

For each I, from (8.2.14) we have

$$\operatorname{supp} \mu \cap I \subset \{x \in I \mid g(x) = 0\}. \tag{8.7.22}$$

Assume for contradiction that $\sigma_d(\operatorname{supp} \mu) > 0$. Since $\sigma_d(\mathscr{Z}) = 0$, we deduce that $\sigma_d(\operatorname{supp} \mu \cap I) > 0$ for some component I. Then by (8.7.22), the zero set of g has positive measure in I, hence $g \equiv 0$ on I by analyticity [Mit20]. Thus H is constant on I and, using (8.7.21),

$$\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x) = C_I - \frac{1}{2} \sum_{j \in J_I} \omega_j \, (a_j \cdot x)^2, \qquad x \in I, \tag{8.7.23}$$

for some $C_I \in \mathbb{R}$. Both sides are real-analytic on $\mathbb{S}^{d-1}$ and I is open, so by the identity theorem this equality holds for all $x \in \mathbb{S}^{d-1}$.

By Lemma 8.7.1**(i)**, it follows that μ is absolutely continuous with a density ρ which is a spherical polynomial of degree ≤ 2; in particular, ρ is real-analytic and not identically zero. Since the zero set of a nontrivial real-analytic function on $\mathbb{S}^{d-1}$ has null σ_d-measure, it follows that $\operatorname{supp}\mu = \mathbb{S}^{d-1}$.

Since $\operatorname{supp}\mu = \mathbb{S}^{d-1}$, the previous argument applies to every component I', yielding global identities similar to (8.7.23). Exactly as in the last part of the proof of Theorem 8.3.1, combining these relations for adjacent cells forces V_θ to be a quadratic polynomial on all of $\mathbb{S}^{d-1}$, hence real-analytic. This contradicts the assumption that V_θ is not real-analytic. Therefore $\sigma_d(\operatorname{supp}\mu) = 0$.

Now assume that σ is real-analytic and that μ is a strict SOPD Wasserstein critical point. If $\sigma_d(\operatorname{supp}\mu) > 0$, then for some component I one has $\sigma_d(\operatorname{supp}\mu \cap I) > 0$. Since g is real-analytic on I and vanishes on $\operatorname{supp}\mu \cap I$, it follows that $g \equiv 0$ on I, hence H is constant on I and therefore globally constant by analyticity. The same localization argument as in the proof of Theorem 8.3.3 then yields tangent perturbations along which the Hessian becomes arbitrarily small, contradicting (8.2.8).

Part II. Atomicity

Fix $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ and define

$$g_{\beta,\theta}(x) \coloneqq \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x) = \int e^{\beta x\cdot y} \mathbf{P}_x^{\perp} y \, \mathrm{d}\mu(y) + \sum_{j=1}^{d} \omega_j \sigma(a_j \cdot x)\, \mathbf{P}_x^{\perp} a_j.$$

If μ is stationary, then $\operatorname{supp}\mu \subset \{x \in \mathbb{S}^{d-1} \mid \ g_{\beta,\theta}(x) = 0\}$, by (8.2.14).

Our goal is to show that, for a dense subset of parameters, all zeros of $g_{\beta,\theta}$ are nondegenerate (and thus isolated). To achieve this, we invoke the parametric transversality theorem [Lee12]. This theorem guarantees the result provided that the global evaluation map $(x, \beta, \theta) \mapsto g_{\beta,\theta}(x)$ is transverse to the zero section.

Explicitly, we need to verify that for every (x, β, θ) such that $g_{\beta,\theta}(x) = 0$, the differential map surjects onto the tangent space:

$$\operatorname{Im} D_{(\beta,\theta)} g_{\beta,\theta}(x) + \operatorname{Im} D_x g_{\beta,\theta}(x) = \mathsf{T}_x \mathbb{S}^{d-1}. \tag{8.7.24}$$

If (8.7.24) holds, the theorem implies that for a residual set of parameters (β, θ), the zeros of $x \mapsto g_{\beta,\theta}(x)$ are nondegenerate.

Assume first that σ is real-analytic and $\sigma(s) \neq 0$ for all $s \neq 0$. The map $(x, \beta, \theta) \mapsto g_{\beta,\theta}(x)$ is then real-analytic on $\mathbb{S}^{d-1} \times \mathbb{R}_{>0} \times (\mathbb{R}^{1+d})^d$. We restrict to the open dense parameter set where a_j are linearly independent and $\omega_j \neq 0$ for all j. On this subset, for every $x \in \mathbb{S}^{d-1}$ there exists $j \in [d]$ such that $a_j \cdot x \neq 0$, thus $\omega_j \sigma(a_j \cdot x) \neq 0$ thanks to the assumption on σ.

The differential with respect to a_j evaluated at $h \in \mathbb{R}^d$ is given by

$$D_{a_j} g_{\beta,\theta}(x)[h] = \omega_j \big(\sigma'(a_j \cdot x)\,(x \cdot h)\mathbf{P}_x^{\perp} a_j + \sigma(a_j \cdot x)\, \mathbf{P}_x^{\perp} h\big).$$

By restricting to $h \in \mathsf{T}_x \mathbb{S}^{d-1}$ so that $\mathbf{P}_x^{\perp} h = h$ and $x \cdot h = 0$, we deduce that

$$D_{a_j} g_{\beta,\theta}(x)[h] = \omega_j \sigma(a_j \cdot x) h.$$

Since $\omega_j \sigma(a_j \cdot x) \neq 0$, the linear map $D_{a_j} g_{\beta,\theta}(x)$ acts as a non-zero scalar multiplication on $\mathsf{T}_x \mathbb{S}^{d-1}$, meaning it is onto $\mathsf{T}_x \mathbb{S}^{d-1}$. In particular, (8.7.24) holds.

By the parametric transversality theorem, the set of parameters for which all zeros of $g_{\beta,\theta}$ are nondegenerate is dense. Moreover, since $\mathbb{S}^{d-1}$ is a compact manifold, this set of parameters is also open; see [Hir76, Thm 2.2.7]. Finally, since nondegenerate zeros are isolated by the inverse function theorem, the compactness of $\mathbb{S}^{d-1}$ guarantees that $g_{\beta,\theta}$ has only finitely many zeros.

Now assume $\sigma(s) = s_+$. Then $x \mapsto g_{\beta,\theta}(x)$ is smooth on each component I of $\mathbb{S}^{d-1} \setminus \mathscr{Z}$. As before, assume a_j linearly independent and $\omega_j \neq 0$. Fix a component I with $J_I \neq \varnothing$ and let $x \in I$ with $g_{\beta,\theta}(x) = 0$. Choose any $j \in J_I$. Then for any $h \in \mathsf{T}_x\mathbb{S}^{d-1}$, the differential with respect to a_j is given by

$$D_{a_j} g_{\beta,\theta}(x)[h] = \omega_j \, \sigma(a_j \cdot x)\, h = \omega_j \, (a_j \cdot x)\, h.$$

Since $a_j \cdot x > 0$ on I and $\omega_j \neq 0$, this map is a scaled isometry, hence surjective onto $\mathsf{T}_x\mathbb{S}^{d-1}$. By the parametric transversality theorem, for a dense set of parameters, the zeros in I are nondegenerate and thus form a discrete set.

To handle the boundaries, we consider the submanifolds F defined by the intersection of some hyperplanes $\{a_k \cdot x = 0\}$ inside $\mathcal{A} \coloneqq \bigcup_j \{a_j \cdot x > 0\}$. Since the restriction of $g_{\beta,\theta}$ to F takes values in the larger space $\mathsf{T}\mathbb{S}^{d-1}$, we consider the projected field $g^F(x) \coloneqq \mathbf{P}^{\perp}_{\mathsf{T}_x F} g_{\beta,\theta}(x)$ for $x \in F$. Note that if $g_{\beta,\theta}(x) = 0$, then necessarily $g^F(x) = 0$. We apply transversality to the map $g^F : F \to \mathsf{T}F$. Let $x \in F$ be a zero of g^F. Since $x \in \mathcal{A}$, there exists j such that $a_j \cdot x > 0$. For any $h \in \mathsf{T}_x F$, the differential with respect to a_j acts as

$$D_{a_j} g^F(x)[h] = \mathbf{P}^{\perp}_{\mathsf{T}_x F}\big(\omega_j (a_j \cdot x) h\big) = \omega_j (a_j \cdot x) h.$$

This map surjects onto $\mathsf{T}_x F$. Thus, for a dense set of parameters, the zeros of g^F are isolated in F. By inclusion, the zeros of $g_{\beta,\theta}$ in F are also isolated.

Since $\mathscr{Z}$ consists of finitely many hyperplanes, $\mathcal{A}$ is the finite union of such I and F. Consequently, the set of zeros of $g_{\beta,\theta}$ in $\mathcal{A}$ is a finite union of discrete sets. If μ is stationary, then $\operatorname{supp}\mu \cap \mathcal{A}$ is countable. □

Remark 8.7.5. Theorem 8.3.5**(ii)** does not strictly preclude the existence of non-atomic stationary measures; rather, it implies that such a measure μ can only be stationary if the parameters lie in the exceptional (non-generic) set associated with μ. A concrete example is the linear activation $\sigma(s) = s$, which yields a quadratic potential $\mathsf{V}_\theta(x) = \sum \omega_j (a_j \cdot x)^2$. If a stationary measure μ has support with non-empty interior, the stationarity condition forces the attention field to be locally (and thus globally) quadratic. By Lemma 8.7.1**(i)**, this implies μ admits a smooth density (a spherical polynomial of degree ≤ 2). Since such a measure is not finitely supported, the parameters allowing it must necessarily lie in the non-generic set specific to μ.

8.7.4 Proofs of Theorem 8.3.7 and Corollary 8.3.8

Proof of Theorem 8.3.7. We prove a more general estimate for an arbitrary parameter $\lambda \in (0,1)$, from which the theorem follows by setting $\lambda = 1/2$.

Restricting the energy functional $\mathsf{E}_{\beta,\theta}$ to atomic measures of the form $\nu = \sum_{i=1}^N m_i \delta_{\psi_i}$ (where m_i are fixed as in (8.3.1)), it becomes a function of the angular coordinates $\boldsymbol{\psi} = (\psi_1, \dots, \psi_N) \in (\mathbb{R}/2\pi\mathbb{Z})^N$:

$$\mathsf{E}_{\beta,\theta}(\boldsymbol{\psi}) = \frac{1}{2\beta} \sum_{i,j=1}^N m_i m_j \mathsf{K}_\beta(\psi_i - \psi_j) + \frac{1}{2} \sum_{i=1}^N m_i (\mathsf{V}_\theta \circ x)(\psi_i).$$

Since μ is a SOPD Wasserstein critical point, it satisfies (8.2.7) and (8.2.14). Restricting the energy functional $\mathsf{E}_{\beta,\theta}$ to atomic measures of the form $\nu = \sum_{i=1}^N m_i \delta_{\psi_i}$ with fixed masses, the second

variation at $\mu = \sum_{i=1}^N m_i \delta_{\varphi_i}$ is represented by the Euclidean Hessian with respect to the angular variables $\boldsymbol{\psi}$, evaluated at $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_N)$. In particular, for every $\xi \in \mathbb{R}^N$,

$$\nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\psi})\big|_{\boldsymbol{\psi}=\boldsymbol{\varphi}}[\xi,\xi] \geq 0. \tag{8.7.25}$$

Thus, the matrix $\nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\varphi})$, with entries defined by

$$\frac{\partial^2 \mathsf{E}_{\beta,\theta}}{\partial\psi_i \partial\psi_j}(\boldsymbol{\varphi}) = \begin{cases} -\frac{1}{\beta} m_i m_j \, \mathsf{K}_\beta''(\varphi_i - \varphi_j), & \text{for } i \neq j, \\ \frac{1}{\beta} m_i \sum_{k\neq i} m_k \mathsf{K}_\beta''(\varphi_i - \varphi_k) + \frac{1}{2} m_i \, \partial_\varphi^2 (\mathsf{V}_\theta \circ x)(\varphi_i), & \text{for } i = j, \end{cases}$$

is positive semi-definite.

Fix $\lambda \in (0,1)$, and let $\mathcal{S} \subseteq [N]$ be a subset with $n = |\mathcal{S}| \geq 2$ satisfying

$$\min_{k\in\mathbb{Z}} |\varphi_i - \varphi_j + 2\pi k| \leq \frac{\lambda}{\sqrt{\beta}} \qquad \text{for all } i,j \in \mathcal{S}. \tag{8.7.26}$$

After relabeling the atoms, we may assume without loss of generality that $\mathcal{S} = [n]$. We evaluate the quadratic form $\nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\varphi})[\xi,\xi]$ along the vector $\xi \in \mathbb{R}^N$ defined by

$$\xi_i = \frac{1}{m_i} \quad (i < n), \qquad \xi_n = -\frac{n-1}{m_n}, \qquad \xi_i = 0 \quad (i > n).$$

Note that $\sum_{i=1}^n m_i \xi_i = 0$. We split

$$\nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\varphi})[\xi,\xi] = \mathcal{T}_{\text{in}} + \mathcal{T}_{\text{out}} + \mathcal{T}_\theta, \tag{8.7.27}$$

where

$$\mathcal{T}_{\text{in}} := \frac{1}{\beta} \sum_{1\leq i<j\leq n} m_i m_j \, \mathsf{K}_\beta''(\varphi_i - \varphi_j)\,(\xi_i - \xi_j)^2,$$

$$\mathcal{T}_{\text{out}} := \frac{1}{\beta} \sum_{i=1}^n \sum_{k>n} m_i m_k \, \mathsf{K}_\beta''(\varphi_i - \varphi_k)\,\xi_i^2,$$

$$\mathcal{T}_\theta := \frac{1}{2} \sum_{i=1}^n m_i \, \partial_\varphi^2 (\mathsf{V}_\theta \circ x)(\varphi_i)\,\xi_i^2.$$

Using the above choice of ξ_i and the identity

$$\sum_{1\leq i<j\leq n} m_i m_j (\xi_i - \xi_j)^2 = \sum_{i=1}^n m_i \sum_{j=1}^n m_j \xi_j^2 - \left(\sum_{i=1}^n m_i \xi_i \right)^2,$$

we can estimate

$$\mathcal{T}_{\text{in}} \leq \frac{1}{\beta} \max_{1\leq i<j\leq n} \mathsf{K}_\beta''(\varphi_i - \varphi_j) \sum_{i=1}^n m_i \sum_{j=1}^n m_j \xi_j^2.$$

Using (8.7.26) and Lemma 8.7.3, we may assume β is large enough that $\frac{\lambda}{\sqrt{\beta}} \leq \varphi_c(\beta)$ and K_β'' is even and increasing on $[0, \varphi_c(\beta)]$. Therefore

$$\max_{1\leq i<j\leq n} \mathsf{K}_\beta''(\varphi_i - \varphi_j) \leq \mathsf{K}_\beta'' \left(\frac{\lambda}{\sqrt{\beta}} \right).$$

Using (8.7.8) and the Taylor expansions $\cos(\lambda/\sqrt{\beta}) = 1 - \lambda^2/(2\beta) + O(\beta^{-2})$ and $\sin^2(\lambda/\sqrt{\beta}) = \lambda^2/\beta + O(\beta^{-2})$, we obtain

$$\mathsf{K}''_\beta\left(\frac{\lambda}{\sqrt{\beta}}\right) = -(1-\lambda^2)\,\beta\, e^{\beta-\lambda^2/2}\,(1+O(\beta^{-1})).$$

In particular, for all $\beta \geq \beta_0(\lambda)$ (possibly increasing β_0), $\mathsf{K}''_\beta(\lambda/\sqrt{\beta}) \leq -\frac{1}{2}(1-\lambda^2)\,\beta\, e^{\beta-\lambda^2/2}$, hence

$$\mathcal{T}_{\text{in}} \leq -\frac{e^{\beta-\frac{\lambda^2}{2}}(1-\lambda^2)}{2}\sum_{i=1}^n m_i \sum_{j=1}^n m_j \xi_j^2. \tag{8.7.28}$$

Second, from (8.7.7) we obtain, for large enough β,

$$\mathcal{T}_{\text{out}} \leq 2e^{\beta-\frac{3}{2}}\sum_{k>n} m_k \sum_{i=1}^n m_i \xi_i^2. \tag{8.7.29}$$

Finally, since σ is globally Lipschitz and $\|x\| = \|x'\| = \|x''\| = 1$, we can compute for a.e. φ:

$$\begin{aligned}\frac{1}{2}\partial_\varphi^2(\mathsf{V}_\theta \circ x)(\varphi) &= \sum_j \omega_j \left(\sigma'\left(a_j \cdot x(\varphi)\right)\left(a_j \cdot x'(\varphi)\right)^2 + \sigma\left(a_j \cdot x(\varphi)\right) a_j \cdot x''(\varphi)\right)\\ &\leq \sum_j |\omega_j| \left(\|\sigma\|_{\mathscr{C}^{0,1}}\|a_j\|^2 + (|\sigma(0)| + \|\sigma\|_{\mathscr{C}^{0,1}}\|a_j\|)\,\|a_j\|\right) =: C_\theta.\end{aligned}$$

Therefore

$$\mathcal{T}_\theta \leq C_\theta \sum_{i=1}^n m_i \xi_i^2, \tag{8.7.30}$$

Combining (8.7.27), (8.7.28), (8.7.29), and (8.7.30), we obtain an upper bound for the Hessian:

$$\nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\varphi})[\xi,\xi] \leq \left[\tfrac{e^{\beta-\frac{\lambda^2}{2}}(\lambda^2-1)}{2}\sum_{i=1}^n m_i + 2e^{\beta-\frac{3}{2}}\sum_{k>n} m_k + C_\theta\right]\sum_{j=1}^n m_j\xi_j^2.$$

For the second order condition (8.7.25) to hold, the right-hand-side cannot be negative. Therefore:

$$\frac{e^{\beta-\frac{\lambda^2}{2}}(1-\lambda^2)}{2}\sum_{i=1}^n m_i \leq 2e^{\beta-\frac{3}{2}}\left(1-\sum_{i=1}^n m_i\right) + C_\theta.$$

Solving for $\sum_{i=1}^n m_i$, we get

$$\sum_{i=1}^n m_i \leq \frac{2e^{\beta-\frac{3}{2}} + C_\theta}{2e^{\beta-\frac{3}{2}} + \frac{e^{\beta-\frac{\lambda^2}{2}}(1-\lambda^2)}{2}} = \frac{4}{4 + e^{\frac{3}{2}-\frac{\lambda^2}{2}}(1-\lambda^2)} + O(e^{-\beta}).$$

Setting $\lambda = 1/2$, this bound yields (8.3.3).

For the second part, assume that μ is itself a single cluster satisfying the pairwise distance condition (so $n = N$ and $\sum_{k>n} m_k = 0$). Then the upper bound for the Hessian gives:

$$0 \leq \nabla^2 \mathsf{E}_{\beta,\theta}(\boldsymbol{\varphi})[\xi,\xi] \leq \left[-\frac{e^{\beta-\frac{\lambda^2}{2}}(1-\lambda^2)}{2}\cdot 1 + C_\theta\right]\sum_{j=1}^n m_j\xi_j^2,$$

hence necessarily

$$C_\theta \;\geq\; \frac{e^{\beta-\frac{\lambda^2}{2}}(1-\lambda^2)}{2}.$$

With $\lambda = 1/2$ this reads

$$C_\theta \geq \frac{3}{8}e^{\beta-\frac{1}{8}}.$$

Under the assumptions $\|\sigma\|_{\mathscr{C}^{0,1}} = 1$ and $\sigma(0) = 0$, we have $C_\theta = 2\sum_j |\omega_j| \cdot \|a_j\|^2$. Thus, if

$$\sum_j |\omega_j| \, \|a_j\|^2 < \frac{3}{16}e^{-\frac{1}{8}},$$

then $C_\theta < \frac{3}{8}e^{-\frac{1}{8}} \leq \frac{3}{8}e^{\beta-\frac{1}{8}}$ for all $\beta > 0$. This excludes the case $\mathcal{S} = [N]$. □

Proof of Corollary 8.3.8. Let Λ_β denote the exact upper bound obtained in the proof of Theorem 8.3.7 for a cluster of diameter $\frac{1}{2\sqrt{\beta}}$, i.e.,

$$\Lambda_\beta \coloneqq \frac{2e^{\beta-\frac{3}{2}} + C_\theta}{2e^{\beta-\frac{3}{2}} + \frac{3}{8}e^{\beta-\frac{1}{8}}} = \frac{16}{16+3e^{11/8}} + O(e^{-\beta})$$

We define a covering of the support of μ. For each of the arcs I_ℓ of length $|I_\ell| \leq L$, we partition it into K_ℓ disjoint sub-intervals $J_{\ell,1}, \ldots, J_{\ell,K_\ell}$, each of length at most $\frac{1}{2\sqrt{\beta}}$. The minimal number of such intervals required is

$$K_\ell \coloneqq 1 \vee \left\lceil \frac{|I_\ell|}{1/(2\sqrt{\beta})} \right\rceil \leq 1 + 2|I_\ell|\sqrt{\beta} \leq 1 + 2L\sqrt{\beta}.$$

Consider any such sub-interval $J = J_{\ell,k}$. Let

$$n_\varepsilon(J) \coloneqq |\{i \mid \varphi_i \in J,\ m_i \geq \varepsilon\}|.$$

We distinguish cases based on the number of atoms in J:

- If J contains at least two atoms, they form a cluster satisfying the pairwise distance condition of Theorem 8.3.7 (since the diameter of J is $\frac{1}{2\sqrt{\beta}}$). Thus, $\mu(J) \leq \Lambda_\beta$.
- If J contains 0 or 1 atom, then obviously $n_\varepsilon(J) \leq 1$.

Combining these, if $\varepsilon \leq \Lambda_\beta$, we have $1 \leq \Lambda_\beta/\varepsilon$. Therefore, in the first case $n_\varepsilon(J) \cdot \varepsilon \leq \mu(J) \leq \Lambda_\beta$, and in the second case $n_\varepsilon(J) \leq 1 \leq \Lambda_\beta/\varepsilon$. Hence, for any sub-interval, $n_\varepsilon(J) \leq \Lambda_\beta/\varepsilon$.

Since every atom with mass $\geq \varepsilon$ belongs to at least one sub-interval $J_{\ell,k}$, we have

$$N_\varepsilon \leq \sum_{\ell=1}^{M}\sum_{k=1}^{K_\ell} n_\varepsilon(J_{\ell,k}) \leq \sum_{\ell=1}^{M} K_\ell \frac{\Lambda_\beta}{\varepsilon}.$$

Substituting the bound for K_ℓ, we obtain:

$$N_\varepsilon \leq M(1+2L\sqrt{\beta})\frac{\Lambda_\beta}{\varepsilon}.$$

If $\varepsilon > \Lambda_\beta$, we use the trivial bound $N_\varepsilon \leq 1/\varepsilon$. For β large enough so that $M(1+2L\sqrt{\beta})\Lambda_\beta \geq 1$, the bound in the statement is larger than $1/\varepsilon$ and thus remains valid. □

8.7.5 Proof of Proposition 8.3.9

Proof of Proposition 8.3.9. Since $e^{\beta x\cdot y} \le e^{\beta}$ on $\mathbb{S}^{d-1}$, we have for any μ,

$$\mathsf{E}_{\beta,\theta}[\mu] = \frac{1}{2\beta}\iint e^{\beta x\cdot y}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) + \frac{1}{2}\int \mathsf{V}_\theta\,\mathrm{d}\mu \le \frac{e^{\beta}}{2\beta} + \frac{1}{2}\max_{x\in\mathbb{S}^{d-1}} \mathsf{V}_\theta(x),$$

and equality holds iff $\mu = \delta_x$ with $x \in \arg\max_{y\in\mathbb{S}^{d-1}} \sum_{j=1}^{d} \omega_j\,\Phi(a_j\cdot y)$. This proves the characterization of global maximizers.

Existence of the minimizer follows from compactness of $\mathcal{P}(\mathbb{S}^{d-1})$ and continuity of the integrands. To see uniqueness, note that the kernel $e^{\beta x\cdot y}$ is conditionally strictly positive definite: for any finite signed measure ν with $\nu(\mathbb{S}^{d-1}) = 0$,

$$\iint e^{\beta x\cdot y}\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y) = \sum_{k\ge 1}\lambda_k(\beta)\sum_{\ell=1}^{N_k}\Big(\int Y_{k\ell}\,\mathrm{d}\nu\Big)^2 > 0 \quad \text{if } \nu \neq 0.$$

Thus E_β is strictly convex on $\mathcal{P}(\mathbb{S}^{d-1})$, see [BMV22, Prop. 2.1 & Thm. 4.1]. Adding the linear term $\frac{1}{2}\int \mathsf{V}_\theta\,\mathrm{d}\mu$ preserves strict convexity; thus the minimizer is unique.

Finally, $\mathsf{E}_\beta[R_\#\mu] = \mathsf{E}_\beta[\mu]$ for all $R \in O(d)$, and if moreover $Ra_j = a_j$ for all j, then $\mathsf{V}_\theta(Rx) = \sum_j \omega_j\,\Phi(a_j\cdot Rx) = \sum_j \omega_j\,\Phi(Ra_j\cdot x) = \mathsf{V}_\theta(x)$. Hence $\mathsf{E}_{\beta,\theta}[R_\#\mu] = \mathsf{E}_{\beta,\theta}[\mu]$ for all such R, and by uniqueness $R_\#\mu_* = \mu_*$.

□

Remark 8.7.6. It is clear that the minimizer μ_* is SOPD. Under additional curvature on the potential, one can ensure strictness. For $d = 2$, the identity (8.7.18) yields for all $\xi \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^1)$:

$$\mathrm{Hess}_\mu\mathsf{E}_{\beta,\theta}(\xi,\xi) \ge \left(\frac{1}{2}\inf_{\varphi\in\operatorname{supp}\mu}\partial_\varphi^2(\mathsf{V}_\theta\circ x)(\varphi) \;-\; \frac{\|\mathsf{K}_\beta''\|_\infty}{\beta}\right)\|\xi\|_{L^2(\mu)}^2,$$

where we have used that

$$\iint(\xi(\varphi)-\xi(\psi))^2\,\mathrm{d}\mu(\varphi)\,\mathrm{d}\mu(\psi) = 2\|\xi\|_{L^2(\mu)}^2 - 2\left(\int\xi\,\mathrm{d}\mu\right)^2 \le 2\|\xi\|_{L^2(\mu)}^2.$$

Consequently, a sufficient condition for μ to be strict SOPD is: there exists $\kappa > 0$ such that

$$\partial_\varphi^2\left(\mathsf{V}_\theta\circ x\right)(\varphi) \ge \frac{2\|\mathsf{K}_\beta''\|_\infty}{\beta} + 2\kappa \qquad (\varphi\in\operatorname{supp}\mu). \tag{8.7.31}$$

For the kernel $\mathsf{K}_\beta(\psi) = e^{\beta\cos\psi}$ one has $\mathsf{K}_\beta''(\psi) = \left(\beta^2\sin^2\psi - \beta\cos\psi\right)e^{\beta\cos\psi}$, hence $\|\mathsf{K}_\beta''\|_\infty = \beta e^{\beta}$. Thus, for V_θ as in (8.2.13), condition (8.7.31) simplifies to

$$\sum_{j=1}^{2}\omega_j\left(\sigma'\left(a_j\cdot x(\varphi)\right)\left(a_j\cdot x'(\varphi)\right)^2 + \sigma\left(a_j\cdot x(\varphi)\right)a_j\cdot x''(\varphi)\right) \ge e^{\beta} + \kappa,$$

for $\varphi\in\operatorname{supp}\mu$.

An analogous sufficient condition holds for $d \ge 3$, replacing $\partial_\varphi^2\left(\mathsf{V}_\theta\circ x\right)$ with the minimum eigenvalue of $\nabla^2\mathsf{V}_\theta$, and $\|\mathsf{K}_\beta''\|_\infty$ with the constant $C_{\beta,d} < \infty$ such that $\mathrm{Hess}_\mu\mathsf{E}_\beta(\xi,\xi) \ge -C_{\beta,d}\|\xi\|_{L^2(\mu)}^2$ for all μ and $\xi\in\mathsf{T}_\mu\mathcal{P}(\mathbb{S}^{d-1})$.

8.7.6 Proof of Proposition 8.3.10

Proof of Proposition 8.3.10. Because $\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu]$ is positive and real-analytic in $\mathbb{S}^{d-1}$, its logarithm is well-defined and real-analytic. On the other hand, on each connected component I of the set $\mathbb{S}^{d-1} \setminus \mathscr{Z}$ we have

$$\mathsf{V}_\theta(x) = \sum_{j=1}^{d} \omega_j (a_j \cdot x)_+^2 = \sum_{j \in J_I} \omega_j (a_j \cdot x)^2 \quad \text{for } x \in I,$$

where $\mathscr{Z}$ and J_I are defined in (8.3.5) and (8.3.6). Therefore

$$H_{\log} := \log\left(\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu]\right) + \frac{1}{2}\mathsf{V}_\theta, \qquad g_{\log} := \nabla_{\mathbb{S}} H_{\log}$$

are real-analytic on I. By (8.2.18), we get $\operatorname{supp}\mu \subset \{g_{\log} = 0\}$.

(a) Suppose, for contradiction, that $\sigma_d(\operatorname{supp}\mu \cap I) > 0$ for some component I. As $g_{\log}$ is real–analytic on I and vanishes on a set of positive measure, we have $g_{\log} \equiv 0$ on I, hence $H_{\log}$ is constant on I, i.e., there exists $C_I \in \mathbb{R}$ such that

$$\log\left(\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x)\right) = C_I - \frac{1}{2}\sum_{j \in J_I} \omega_j \, (a_j \cdot x)^2, \qquad x \in I.$$

Define the global real–analytic vector field

$$\widetilde{g}_{\log}(x) := \nabla_{\mathbb{S}} \left(\log\left(\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x)\right) + \frac{1}{2}\sum_{j \in J_I} \omega_j (a_j \cdot x)^2 \right).$$

We have $\widetilde{g}_{\log} = g_{\log} \equiv 0$ on I, so the zero set of $\widetilde{g}_{\log}$ has an accumulation point in $\mathbb{S}^{d-1}$. By the identity theorem for real–analytic functions on $\mathbb{S}^{d-1}$, we get $\widetilde{g}_{\log} \equiv 0$ on $\mathbb{S}^{d-1}$. Therefore

$$\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x) = \exp\left(C_I - \frac{1}{2}\sum_{j \in J_I} \omega_j \, (a_j \cdot x)^2\right), \qquad x \in \mathbb{S}^{d-1}. \tag{8.7.32}$$

Since the right-hand side of (8.7.32) is even, it follows that μ is antipodally symmetric by Lemma 8.7.1**(ii)**. Thus $\operatorname{supp}\mu$ meets the antipodal component I' of I, whose active set is $J_{I'} = [d] \setminus J_I$.

Repeating the argument on I' yields another global identity

$$\frac{\delta \mathsf{E}_\beta}{\delta \mu}[\mu](x) = \exp\left(C_{I'} - \frac{1}{2}\sum_{j \in J_{I'}} \omega_j \, (a_j \cdot x)^2\right), \qquad x \in \mathbb{S}^{d-1}.$$

Dividing the two expressions we obtain, for all $x \in \mathbb{S}^{d-1}$,

$$\sum_{j \in J_I} \omega_j \, (a_j \cdot x)^2 - \sum_{j \notin J_I} \omega_j \, (a_j \cdot x)^2 \equiv 2(C_I - C_{I'}). \tag{8.7.33}$$

Equivalently,

$$x^\top M x \equiv \text{const} \quad (x \in \mathbb{S}^{d-1}), \qquad \text{for } M := \sum_{j \in J_I} \omega_j \, a_j a_j^\top - \sum_{j \notin J_I} \omega_j \, a_j a_j^\top.$$

A quadratic form that is constant on the unit sphere is a scalar multiple of $\|x\|^2$, which contradicts (8.3.7). Therefore $\sigma_d(\operatorname{supp}\mu \cap I) = 0$ for every I. Since $\sigma_d(\partial I) = 0$, we conclude $\sigma_d(\operatorname{supp}\mu) = 0$.

(b) Now let $d = 2$ and suppose that $\operatorname{supp}\mu$ is infinite, so that $\operatorname{supp}\mu \cap I$ is infinite for some component I. Since $g_{\log}$ is continuous on $\mathbb{S}^1$ and real–analytic on I, and $g_{\log} = 0$ on $\operatorname{supp}\mu$, the zero set has an accumulation point in $\mathbb{S}^1$. The identity theorem applied to $\widetilde{g}_{\log}$ as above yields the global identity (8.7.32), and repeating the argument on the antipodal component I' gives (8.7.33), the same contradiction as in part (a). □

Remark 8.7.7. The indirect argument used in the proof of Proposition 8.3.10 also applies to (SA+P) with a symmetric nonsingular matrix $B \in \mathscr{M}_{d\times d}(\mathbb{R})$, i.e., by replacing E_β with

$$\mathsf{E}_B[\mu] \coloneqq \frac{1}{2}\iint e^{x^\top By}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y), \qquad \frac{\delta \mathsf{E}_B}{\delta\mu}[\mu](x) = \int e^{x^\top By}\,\mathrm{d}\mu(y).$$

Indeed, $\frac{\delta \mathsf{E}_B}{\delta\mu}[\mu] > 0$ and it is real-analytic on $\mathbb{S}^{d-1}$, so the definition $H_{\log} = \log\left(\frac{\delta \mathsf{E}_B}{\delta\mu}[\mu]\right) + \frac{1}{2}\mathsf{V}_\theta$ and the inclusion $\operatorname{supp}\mu \subset \{\nabla_{\mathbb{S}} H_{\log} = 0\}$ on each component I remain unchanged. If $\sigma_d(\operatorname{supp}\mu \cap I) > 0$, the same analyticity argument forces the global identity

$$\frac{\delta \mathsf{E}_B}{\delta\mu}[\mu](x) = \exp\left(C_I - \frac{1}{2}\sum_{j\in J_I}\omega_j(a_j\cdot x)^2\right) \qquad (x \in \mathbb{S}^{d-1}). \tag{8.7.34}$$

Hence, $\frac{\delta \mathsf{E}_B}{\delta\mu}[\mu]$ is even, and by Remark 8.7.2, this implies that μ is antipodally symmetric, so $\sigma_d(\operatorname{supp}\mu \cap I') > 0$ for the antipodal component I'. Repeating the step on I' yields (8.7.33) and leads to the same contradiction.

We note that a similar antipodal argument extends Theorems 8.3.1 and 8.3.5 to E_B, simply by removing the exponential from (8.7.34) and proceeding similarly.

Conclusions

The work presented in this doctoral thesis addresses various aspects of machine learning using models based on controlled flows, analyzed from the perspective of control theory. The primary motivation for this approach stems from neural ordinary differential equations (neural ODEs), understood as dynamical systems where the vector field is a neural network. In recent years, these models have driven numerous breakthroughs and opened new research avenues in deep learning, notably by bridging the gap between machine learning and established mathematical disciplines such as optimal control.

Throughout this thesis, the concept of controllability has been a central pillar, serving as a guarantee of feasibility across the various tasks considered. The material is organized into three main parts:

In the first part (**Part I**), we focus on supervised learning, tackling classification and regression tasks, and establishing theoretical bounds for out-of-distribution performance (i.e., guarantees for model extrapolation to new data points).

In the second part (**Part II**), we explore generative modeling. We study the lifting of discrete input data to probability measures, minimizing various information losses culminating in relative entropy, which imposes a particularly strong topology.

Finally, the third part (**Part III**) offers tangible perspectives on transitioning these concepts to a more practical domain. On an applied level, we explore time series modeling for the dynamics of probability distributions within the context of digital health. On a theoretical level, we analyze the dynamical effect of incorporating a perceptron into a self-attention block.

Below, we detail the specific conclusions drawn from each of the individual chapters presented.

Part I. Supervised Learning

Chapter 2: Data Classification

In this chapter, we analyzed the capacity of neural ODEs to classify finite datasets (in a minimax sense) through the lens of simultaneous control and geometric separability.

Our main result establishes that, for the single-neuron model ($\mathcal{V}_1^+$) with piecewise-constant controls, any dataset $(\mathcal{R}, \mathcal{B})$ in $\mathbb{R}^d$ can be classified using

$$L = 4\left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil$$

control discontinuities (equivalently, $L{+}1$ constant-control segments). The only assumption is that the points are in general position (GP), a condition satisfied almost surely for a finite sample drawn from an absolutely continuous measure $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. This improves the previous bound of $L = 3(|\mathcal{R}| \wedge |\mathcal{B}|)$ by leveraging the enhanced separability that point clouds exhibit in higher dimensions. Technically, we show that within an arbitrary finite configuration in GP, any subset of d points can be isolated between two parallel hyperplanes, and we turn this geometric fact into an

explicit control construction. The numerical experiments reported in the chapter are consistent with the sharpness of the resulting N/d scaling in high dimensions.

Our second result shows that worst-case complexity is typically not required:

(i) we characterize the pathological configurations for which the maximal canonical complexity $L = 2N - 2$ occurs;
(ii) in the canonical-separability setting of **Section 2.3**, we derive the exact distribution of the required number of switches L when the points are i.i.d. and both classes have a fixed size $N \geq 1$. This analysis again exhibits a high-dimensional advantage: as d increases, the probability that some coordinate projection yields a low-alternation pattern increases. In particular, we deduce that if $d \gtrsim 2^{2N}/\sqrt{N}$, classification using constant controls ($L = 0$) is possible with high probability.

Overall, these results can be viewed as a manifestation of a *blessing of dimensionality* [Don00]: the phenomenon where machine learning problems become more tractable in high dimensions. Here, higher dimensionality improves the geometric freedom of the control, reducing the switching complexity needed for guaranteed separation.

Chapter 3: Data Interpolation

In this chapter, we analyzed exact and approximate interpolation via neural ODE flows in $\mathbb{R}^d$, establishing the *universal interpolation property* (UIP, see Definition 3.2.1) as a worst-case expressivity benchmark. For the one-layer ansatz ($\mathcal{V}_p^+$), Theorem 3.3.1 provides a constructive proof of the UIP, exhibiting an explicit width-depth-in-time trade-off: any admissible dataset of size N can be interpolated by a piecewise-constant control with $L = 2\lceil N/p \rceil - 1$ switches. This control strategy simplifies significantly in the high-dimensional regime $d > N$, recovering the fully autonomous case as soon as $p \geqslant N$ (Corollary 3.3.5).

Pushing toward autonomy, we identified two complementary mechanisms. First, by eliminating switching while keeping a minimal time dependence inside the activation, the semi-autonomous model ($\mathcal{V}_{\mathrm{sa}}^+$) ensures the UIP with a linear width of $p = 2N$ (Theorem 3.3.7). In the strictly autonomous setting, exact controllability is achieved either under geometric separability assumptions (Assumption 2–Corollary 3.3.9) or by increasing architectural depth: the two-layer model ($\mathcal{V}_{2\mathrm{a}}^+$) leverages spatial localization to satisfy the UIP (Theorem 3.3.11).

Finally, as exact control with ($\mathcal{V}_{\mathrm{a}}^+$) is out of reach, Theorem 3.3.14 guarantees approximate interpolation with error rates of Barron-type. These avoid the curse of dimensionality by ensuring that the error decays with the network width p as $\lesssim p^{-1/2}$, a rate that is independent of the ambient dimension d. We also analyzed the linear-in-control architecture ($\mathcal{V}_{\ell}^+$), deriving exact controls with $O(N)$ switches (Theorem 3.3.18). To summarize these findings, Table 3.1 compiles the results for each model, detailing the complexities and any geometric assumptions required by our approach.

Chapter 4: Generalization

This chapter studies generalization for flow-based predictors. We first show that the exact interpolation of finite datasets (as seen in **Chapters 2** and **3**) often produces highly unstable flows with exploding Lipschitz constants, rendering standard generalization bounds ineffective. To overcome this, we shift from interpolating finite datasets to a cell-wise approach: instead of steering isolated data points, we construct a piecewise-constant nonparametric proxy of the target on a compact domain and steer entire trimmed spatial cells through the flow.

This strategy relies on the cell-wise controllability formalized in Assumption 3. Since this requirement is too restrictive for purely autonomous architectures, we leverage the semi-autonomous model ($\mathcal{V}_{\mathrm{sa}}^+$).

For this architecture, Proposition 4.4.3 bounds the realization tolerance by $O(p^{-1/2})$, contributing an $O(p^{-1})$ control error to the population risk.

The core template bound in (4.2.8) explicitly decomposes the expected risk into statistical approximation, geometric boundary-layer (trimming), and expressivity (controllability) terms. For histogram partitions, this yields the estimate (Theorem 4.4.6):

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \omega_y(\sqrt{d}\,h)^2 + \frac{\|y\|^2_{L^\infty(I_R)}}{Nh^d} + \frac{\delta}{h} + \frac{1}{p}.$$

For α-Hölder targets, optimizing the partition size h recovers the classical nonparametric rate $O(N^{-2\alpha/(2\alpha+d)})$, penalized by the geometric (δ/h) and controllability $(1/p)$ costs. Note that, unlike in the interpolation setting, the curse of dimensionality reappears here due to the exponential scaling of the number of cells needed to cover the compact domain. In the noiseless regime, replacing histograms with a Voronoi nearest-neighbor proxy improves the statistical term at the expense of a more delicate geometric contribution (Theorem 4.4.8):

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \left(\frac{\log N}{N}\right)^{2\alpha/d} + \delta N^{1/d} + \frac{1}{p}.$$

Ultimately, we show that controlled flows can match classical nonparametric regression rates through a clear statistics–geometry–expressivity trade-off.

Part II. Generative Modeling

Chapter 5: Controllability for Measure Transport

In this chapter, we developed a control-theoretic framework for approximation at the level of probability measures by controlled flow maps. Given a base law μ_{B} and a discrepancy D between probability measures, we introduced the notions of D-compatibility and D-Universal Approximation Property (D-UAP) in Definitions 5.2.2 and 5.2.3, thereby formalizing when the attainable set of pushforwards $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ is dense, in the topology induced by D, among all target measures that are intrinsically approximable by Lipschitz transport.

For Wasserstein distances, we identified the finite-dimensional controllability notion underlying measure approximation: not the UIP itself, but its permutation-invariant counterpart, the π-UIP (Definition 5.3.1), reflecting the fact that empirical measures do not depend on particle labels. More precisely, we proved that the approximate π-UIP is equivalent to the W_p-UAP on the class of empirical measures with equal weights and pairwise distinct atoms (Theorem 5.3.4). Thus, at the empirical level, Wasserstein approximation is exactly the unlabeled interpolation problem. We then showed that passing from this discrete picture to general measures requires quantitative control on the interpolating flows: in Theorem 5.3.8, the Γ-quantitative π-UIP (Definition 5.3.2) is proved equivalent to the Γ-quantitative W_p-UAP (Definition 5.3.3). We also established a structural route to Wasserstein universality via approximation of transport maps in $L^p(\mu_{\mathrm{B}})$ (Propositions 5.3.9 and 5.3.11), yielding in particular the affine-invariant universality criterion of Corollary 5.3.13.

For the total variation distance, the situation is markedly different. Since TV is rigid on atomic measures and invariant under homeomorphisms, pointwise interpolation alone is not a relevant approximation mechanism. In dimension $d = 1$, we showed that TV-universality on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R})$ reduces to strong approximation of monotone rearrangements (Proposition 5.4.3), leading to the affine-invariant criterion in Corollary 5.4.5. In dimensions $d \geq 2$, we first identified a structural obstruction: decoupled control families preserve product structure and therefore cannot achieve the TV-UAP (Proposition 5.4.7). The key positive ingredient is instead coordinate mixing: under

reversibility, marginal flows, and shear flows (Item (A0)–Item (A2)), we proved the TV-UAP on $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ (Theorem 5.4.8) through a constructive strategy inspired by the Knothe–Rosenblatt rearrangement, and verified these assumptions for standard affine-invariant neural ODE families in Corollary 5.4.9.

Finally, we clarified the relation between pointwise controllability and measure-level universality through explicit counterexamples. In the Wasserstein setting, Proposition 5.3.5 shows that even the UIP does not imply the W_p-UAP for general measures. In total variation, Propositions 5.4.13 and 5.4.14 show that neither the UIP nor the TV-UAP implies the other in general. Altogether, these results show that universality for generative flows depends fundamentally on the discrepancy and on the geometric structure of the control family.

Chapter 6: Continuous Normalizing Flows

In this chapter, we studied the feasibility problem associated with maximum likelihood in continuous normalizing flows, namely, the existence of a structured control θ such that the terminal law $\mu_T^\theta = (\Phi_T^\theta)_{\#}\mu_{\mathrm{B}}$ approximates a target law $\mu_\star$ in the forward or reverse Kullback–Leibler divergence.

Compared to **Chapter 5**, the main contribution here is a mechanism to upgrade TV-approximation to KL-approximation by carefully controlling density ratios. Under an upper-tail condition on the target density (specifically, a Gaussian upper bound outside a compact set), Theorem 6.2.1 yields, for arbitrary $T > 0$ and $\varepsilon > 0$, a piecewise-constant control θ with finitely many switches such that

$$\mathsf{KL}\big(\mu_\star \,\|\, \mu_T^\theta\big) < \varepsilon.$$

Under a corresponding lower-tail condition (a Gaussian lower bound) and strict positivity of $\rho_\star$, Theorem 6.2.3 provides an analogous statement for the reverse divergence $\mathsf{KL}\big(\mu_T^\theta \,\|\, \mu_\star\big)$. In both cases, the core idea is to combine:

(i) the constructive TV-approximation from [RZ24], with
(ii) a reverse Pinsker inequality, whose coefficient depends on $\|\rho_1/\rho_2\|_{L^\infty}$, and
(iii) a constructive tail control ensuring that this essential supremum remains finite.

Part III. Applications

Chapter 7: Distributional Dynamics Modeling for Digital Health

In this chapter, we proposed a simple and interpretable approach to learn distributional dynamics from snapshot samples. The model represents each μ_t by a Gaussian mixture with shared means and covariances, and time-dependent weights $\alpha(t) \in \Delta^{K-1}$. The resulting estimator combines:

(i) a per-time maximum mean discrepancy (MMD) fit with closed-form Gaussian terms and a ridge-stabilized weight update, and
(ii) a continuous-time smoothing stage in which $\alpha(t)$ is obtained by integrating a neural ODE and projecting back to the simplex.

On synthetic benchmarks, the method is competitive with flexible nonparametric alternatives and can outperform them in higher dimensions, while providing weight trajectories that directly quantify distributional shifts. In the CGM trial, the learned trajectories separate clinically meaningful phenotypes of glycaemic control and reveal heterogeneous within-arm responses that are difficult to detect from static summaries. Finally, the approximation and stability statements in Proposition 7.3.1 and Theorem 7.3.2 justify the shared-dictionary modeling choice and clarify the finite-sample behavior of the MMD snapshot estimator.

Chapter 8: Mean-field Landscape of Transformers

In this chapter, we studied mean-field attention on the sphere $\mathbb{S}^{d-1}$ as a Wasserstein gradient flow for an energy combining an (unnormalized) attention interaction with a perceptron-induced potential, where $\beta > 0$ denotes the inverse temperature (equivalently, the interaction strength) of the attention kernel. This formulation is motivated both by the interacting-particle viewpoint and by normalization mechanisms that constrain embeddings to a compact manifold. Our first main result establishes the discrete nature of stationary solutions. When the perceptron-induced potential is built from the ReLU activation and $d = 2$, every stationary measure has finite support (Theorem 8.3.1). The reason is that the attention contribution to the first variation is real-analytic on $\mathbb{S}^1$, whereas the ReLU term generically destroys this analyticity; as a result, the stationarity condition can only hold on finitely many support points. When the perceptron uses a real-analytic activation such as GeLU, stationarity alone no longer forces finiteness, but the same conclusion does hold for strictly stable stationary measures in the Wasserstein sense (Theorem 8.3.3). In $d \geq 2$, stationary measures are necessarily singular with respect to the uniform surface measure, and for generic parameters they are atomic as well (Theorem 8.3.5).

In the descent (thus repulsive) regime, we prove an anti-concentration bound: cluster masses at the natural angular scale $\beta^{-1/2}$ are uniformly controlled by a numerical constant (Theorem 8.3.7). Consequently, the limiting measure can be atomic without collapsing into a single atom, in sharp contrast to the attractive regime. We also bound the support complexity, showing that the number of heavy atoms scales at most like $\sqrt{\beta}$ (Corollary 8.3.8). Furthermore, we characterize the extremal points of the energy (Proposition 8.3.9): the global maximizers are Dirac masses supported at maximizers of the perceptron potential, whereas the global minimizer is unique. For ReLU perceptrons, finding these maximizers reduces to a finite family of quadratic programs, yielding explicit solutions in several regimes, and in a one-dimensional toy setting we compute the minimizer explicitly, which makes the competition between repulsive attention and the perceptron bias completely transparent.

Numerical experiments in **Section 8.4** illustrate this clustering behavior across dimensions and inverse temperatures, empirically validating the dependence of the number of atoms on β and on the model parameters in qualitative agreement with our bounds.

We also showed that these sparsity conclusions extend beyond unnormalized attention. For normalized self-attention, the softmax induces a weighted Wasserstein gradient flow rather than the standard one, but under a mild non-degeneracy condition on the perceptron parameters we still obtain singular stationary measures in every dimension, and finite atomicity in $d = 2$ (Proposition 8.3.10).

Finally, as an outlook, we discussed Mixture-of-Experts (MoE) layers on $\mathbb{S}^{d-1}$, relating hard linear top-k routing to order-k spherical Voronoi diagrams. We also proposed a potential-based soft routing that preserves the Wasserstein gradient flow structure.

Conclusiones

El trabajo expuesto en esta tesis doctoral aborda diversos aspectos del machine learning utilizando modelos basados en flujos controlados, analizados desde la perspectiva de la teoría de control. La motivación principal de este enfoque radica en las ecuaciones diferenciales ordinarias neuronales, entendidas como sistemas dinámicos donde el campo vectorial está dado por una red neuronal. En los últimos años, estos modelos han impulsado numerosos avances y abierto nuevas líneas de investigación en el campo del deep learning, destacando especialmente por tender puentes fundamentales con áreas matemáticas consolidadas como el control óptimo.

A lo largo de esta tesis, el concepto de controlabilidad ha jugado un papel central, actuando como una garantía de viabilidad para las diversas tareas consideradas. El material se ha estructurado en tres grandes bloques:

En la Parte I, nos centramos en el aprendizaje supervisado, abordando tareas de clasificación y regresión, y estableciendo cotas teóricas para el rendimiento del modelo fuera del conjunto de entrenamiento (es decir, garantías de extrapolación a nuevos puntos).

En la Parte II, exploramos el modelado generativo (generative modeling). Para ello, estudiamos el levantamiento (lift) de datos discretos de entrada a medidas de probabilidad, minimizando diversas pérdidas de información hasta culminar con la entropía relativa, la cual impone una topología de convergencia particularmente fuerte.

Por último, la Parte III ofrece perspectivas tangibles sobre cómo trasladar estos conceptos a un terreno más práctico. A nivel de aplicaciones, exploramos el modelado de series temporales para la dinámica de distribuciones de probabilidad en el ámbito de la salud digital. A nivel teórico, analizamos el efecto que tiene la incorporación de un perceptrón sobre la dinámica de un bloque de autoatención.

A continuación, se detallan las conclusiones específicas correspondientes a cada uno de los capítulos expuestos.

Parte I. Aprendizaje Supervisado

Capítulo 2: Clasificación de Datos

En este capítulo, analizamos la capacidad de las EDOs neuronales para clasificar conjuntos de datos finitos (en un sentido minimax) bajo la perspectiva del control simultáneo y la separabilidad geométrica.

Nuestro resultado principal establece que, para el modelo de una sola neurona ($\mathcal{V}_1^+$) con controles constantes a trozos, cualquier conjunto de datos $(\mathcal{R}, \mathcal{B})$ en $\mathbb{R}^d$ puede ser clasificado utilizando

$$L = 4 \left\lceil \frac{|\mathcal{R}| \wedge |\mathcal{B}|}{d} \right\rceil$$

discontinuidades de control (o equivalentemente, $L+1$ segmentos de control constante). La única suposición es que los puntos se encuentren en posición general, una condición que se satisface casi

con seguridad para una muestra finita extraída de una medida absolutamente continua $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Esto mejora la cota previa de $L = 3(|\mathcal{R}| \wedge |\mathcal{B}|)$ al aprovechar la mayor separabilidad que exhiben las nubes de puntos en dimensiones altas. A nivel técnico, demostramos que dentro de una configuración finita arbitraria en posición general, cualquier subconjunto de d puntos puede aislarse entre dos hiperplanos paralelos, y transformamos este hecho geométrico en una construcción de control explícita. Los experimentos numéricos presentados en este capítulo corroboran que el escalado resultante de N/d es óptimo en altas dimensiones.

Nuestro segundo resultado demuestra que la complejidad del peor de los casos rara vez se materializa en la práctica:

(i) caracterizamos las configuraciones patológicas para las cuales se alcanza la complejidad canónica máxima $L = 2N - 2$;
(ii) en el marco de separabilidad canónica de la Sección 2.3, derivamos la distribución exacta del número necesario de conmutaciones L cuando los puntos son i.i.d. y ambas clases tienen un tamaño fijo $N \geq 1$. Este análisis vuelve a evidenciar una ventaja en altas dimensiones: a medida que d aumenta, también lo hace la probabilidad de que alguna proyección coordenada produzca un patrón de baja alternancia. En particular, deducimos que si $d \gtrsim 2^{2N}/\sqrt{N}$, la clasificación utilizando controles constantes ($L = 0$) es posible con alta probabilidad.

En conjunto, estos resultados pueden interpretarse como una manifestación de la *bendición de la dimensionalidad* [Don00]: el fenómeno por el cual los problemas de aprendizaje automático se vuelven más tratables en dimensiones altas. En este contexto, una mayor dimensionalidad incrementa la libertad geométrica del control, reduciendo significativamente la complejidad de conmutación necesaria para garantizar la separación.

Capítulo 3: Interpolación de Datos

En este capítulo, analizamos la interpolación exacta y aproximada mediante flujos de EDOs neuronales en $\mathbb{R}^d$, estableciendo la *propiedad de interpolación universal* (UIP, Definición 3.2.1) como un punto de referencia de expresividad en el peor de los casos. Para el *ansatz* de una sola capa ($\mathcal{V}_p^+$), el Teorema 3.3.1 proporciona una prueba constructiva de la UIP, exhibiendo un compromiso explícito entre anchura y profundidad en el tiempo: cualquier conjunto de datos admisible de tamaño N puede ser interpolado mediante un control constante a trozos con $L = 2\lceil N/p \rceil - 1$ conmutaciones. Esta estrategia de control se simplifica significativamente en el régimen de alta dimensionalidad $d > N$, recuperando el caso totalmente autónomo tan pronto como $p \geqslant N$ (Corolario 3.3.5).

Avanzando hacia la autonomía, identificamos dos mecanismos complementarios. En primer lugar, al eliminar las conmutaciones manteniendo una dependencia temporal explícita mínima dentro de la activación, el modelo semiautónomo ($\mathcal{V}_{\mathrm{sa}}^+$) garantiza la UIP con una anchura lineal $p = 2N$ (Teorema 3.3.7). En el marco estrictamente autónomo, la controlabilidad exacta se alcanza bien bajo asunciones de separabilidad geométrica (Hipótesis 2–Corolario 3.3.9), o bien incrementando la profundidad arquitectónica: el modelo de dos capas ($\mathcal{V}_{\mathrm{2a}}^+$) aprovecha la localización espacial para satisfacer la UIP (Teorema 3.3.11).

Por último, dado que el control exacto con ($\mathcal{V}_{\mathrm{a}}^+$) es inalcanzable, el Teorema 3.3.14 garantiza la interpolación aproximada con tasas de error explícitas de tipo Barron. Estas tasas evitan la maldición de la dimensionalidad al asegurar que el error decaiga con la anchura de la red p como $\lesssim p^{-1/2}$, un ritmo de convergencia independiente de la dimensión espacial d. También analizamos la arquitectura lineal en el control ($\mathcal{V}_{\ell}^+$), derivando controles exactos con $O(N)$ conmutaciones (Teorema 3.3.18). Para resumir estos hallazgos, la Tabla 3.1 recopila los resultados de cada modelo, detallando las complejidades y cualquier suposición geométrica requerida por nuestro enfoque.

Capítulo 4: Generalización

Este capítulo estudia la capacidad de generalización de los predictores basados en flujos. En primer lugar, vemos que la interpolación exacta de conjuntos de datos finitos (como se vio en los Capítulos 2 y 3) a menudo produce flujos altamente inestables con constantes de Lipschitz que explotan, lo que invalida las cotas de generalización estándar. Para superar esto, pasamos de interpolar conjuntos de datos finitos a un enfoque por celdas: en lugar de transportar puntos de datos aislados, construimos un estimador no paramétrico (un *proxy* constante a trozos del objetivo sobre un dominio compacto) y transportamos celdas espaciales enteras recortadas a través del flujo.

Esta estrategia se basa en la controlabilidad por celdas formalizada en el Supuesto 3. Dado que este requisito es demasiado restrictivo para arquitecturas puramente autónomas, aprovechamos el modelo semiautónomo ($\mathcal{V}^{+}_{\mathrm{sa}}$). Para esta arquitectura, la Proposición 4.4.3 acota la tolerancia de realización por $O(p^{-1/2})$, lo que aporta un error de control de $O(p^{-1})$ al riesgo poblacional.

La cota estructural principal en (4.2.8) descompone explícitamente el riesgo esperado en términos de aproximación estadística, capa límite geométrica (recorte) y expresividad (controlabilidad). Para particiones basadas en histogramas, esto produce la estimación explícita (Teorema 4.4.6):

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \omega_y(\sqrt{d}\,h)^2 + \frac{\|y\|^2_{L^\infty(I_R)}}{Nh^d} + \frac{\delta}{h} + \frac{1}{p}.$$

Para objetivos α-Hölder, optimizar el tamaño de la partición h recupera la tasa no paramétrica clásica $O(N^{-2\alpha/(2\alpha+d)})$, penalizada explícitamente por los costes geométricos (δ/h) y de expresividad ($1/p$). Nótese que, a diferencia del escenario de interpolación, aquí reaparece la maldición de la dimensionalidad debido al escalado exponencial del número de celdas necesarias para cubrir el dominio compacto. En el régimen sin ruido, sustituir los histogramas por un estimador de vecinos más cercanos de Voronoi mejora el término estadístico a costa de una contribución geométrica más delicada (Teorema 4.4.8):

$$\mathbb{E}_{\mathcal{D}_N}\left[\mathcal{R}(\theta_N)\right] \lesssim \left(\frac{\log N}{N}\right)^{2\alpha/d} + \delta N^{1/d} + \frac{1}{p}.$$

En última instancia, mostramos que los flujos controlados pueden igualar las tasas clásicas de regresión no paramétrica a través de un claro compromiso entre estadística, geometría y expresividad.

Parte II. Modelado Generativo

Capítulo 5: Controlabilidad para el Transporte de Medidas

En este capítulo, hemos desarrollado un marco teórico de control para la aproximación a nivel de medidas de probabilidad mediante mapas de flujo controlados. Dada una ley base μ_{B} y una discrepancia D entre medidas de probabilidad, introdujimos las nociones de D-compatibilidad y la Propiedad de Aproximación Universal respecto a D (D-UAP) en las Definiciones 5.2.2 y 5.2.3, formalizando así cuándo el conjunto alcanzable de medidas imagen $\mathcal{A}_{\mathcal{V}}(\mu_{\mathrm{B}})$ es denso, en la topología inducida por D, dentro de todas las medidas objetivo que son intrínsecamente aproximables mediante transporte Lipschitz.

Para las distancias de Wasserstein, identificamos la noción de controlabilidad en dimensión finita que subyace a la aproximación de medidas: no la UIP en sí, sino su variante invariante por permutaciones, la π-UIP (Definición 5.3.1), lo cual refleja el hecho de que las medidas empíricas no dependen de las etiquetas de las partículas. Más concretamente, demostramos que la π-UIP aproximada es equivalente a la W_p-UAP sobre la clase de medidas empíricas con pesos iguales y

átomos mutuamente distintos (Teorema 5.3.4). Por lo tanto, a nivel empírico, la aproximación en Wasserstein es exactamente el problema de interpolación sin etiquetas. A continuación, mostramos que pasar de este escenario discreto a medidas generales requiere un control cuantitativo sobre los flujos interpoladores: en el Teorema 5.3.8, se demuestra que la π-UIP Γ-cuantitativa (Definición 5.3.2) es equivalente a la W_p-UAP Γ-cuantitativa (Definición 5.3.3). Asimismo, establecimos una vía estructural hacia la universalidad en Wasserstein mediante la aproximación de mapas de transporte en $L^p(\mu_{\mathrm{B}})$ (Proposiciones 5.3.9 y 5.3.11), obteniendo en particular el criterio de universalidad afín-invariante del Corolario 5.3.13.

Para la distancia de variación total, la situación es notablemente distinta. Puesto que la TV es rígida sobre medidas atómicas e invariante bajo homeomorfismos, la interpolación puntual por sí sola no constituye un mecanismo de aproximación relevante. En dimensión $d = 1$, mostramos que la universalidad en TV sobre $\mathcal{P}_{\mathrm{ac}}(\mathbb{R})$ se reduce a la aproximación fuerte de reordenamientos monótonos (Proposición 5.4.3), lo que conduce al criterio afín-invariante en el Corolario 5.4.5. En dimensiones $d \geq 2$, primero identificamos una obstrucción estructural: las familias de control desacopladas preservan la estructura producto y, por consiguiente, no pueden alcanzar la TV-UAP (Proposición 5.4.7). En cambio, el ingrediente positivo clave es la mezcla de coordenadas: asumiendo reversibilidad, flujos marginales y flujos de cizalla (Condiciones (A0)–(A2)), demostramos la TV-UAP sobre $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ (Teorema 5.4.8) a través de una estrategia constructiva inspirada en el reordenamiento de Knothe–Rosenblatt, y verificamos estas hipótesis para familias estándar de EDOs neuronales afín-invariantes en el Corolario 5.4.9.

Finalmente, clarificamos la relación entre la controlabilidad puntual y la universalidad a nivel de medidas mediante contraejemplos explícitos. En el contexto de Wasserstein, la Proposición 5.3.5 muestra que ni siquiera la UIP implica la W_p-UAP para medidas generales. En variación total, las Proposiciones 5.4.13 y 5.4.14 demuestran que, en general, ni la UIP ni la TV-UAP implican a la otra. En conjunto, estos resultados evidencian que la universalidad de los flujos generativos depende fundamentalmente de la discrepancia elegida y de la estructura geométrica de la familia de control.

Capítulo 6: Flujos de Normalización Continuos

En este capítulo, estudiamos el problema de viabilidad asociado a la máxima verosimilitud en flujos normalizadores continuos; en concreto, la existencia de un control estructurado θ tal que la ley terminal $\mu_T^\theta = (\Phi_T^\theta)_{\#}\mu_{\mathrm{B}}$ aproxime una ley objetivo $\mu_\star$ en la divergencia de Kullback–Leibler directa o inversa.

En comparación con el **Capítulo 5**, la principal contribución aquí es un mecanismo para elevar la aproximación en TV a una aproximación en KL mediante el estricto control de los cocientes de densidad. Bajo una condición de decaimiento en las colas superiores de la densidad objetivo (específicamente, una cota superior gaussiana fuera de un conjunto compacto), el Teorema 6.2.1 garantiza, para $T > 0$ y $\varepsilon > 0$ arbitrarios, un control constante a trozos θ con un número finito de conmutaciones tal que

$$\mathsf{KL}\big(\mu_\star \,\|\, \mu_T^\theta\big) < \varepsilon.$$

Bajo una condición análoga para las colas inferiores (una cota inferior gaussiana) y la positividad estricta de $\rho_\star$, el Teorema 6.2.3 proporciona un resultado equivalente para la divergencia inversa $\mathsf{KL}\big(\mu_T^\theta \,\|\, \mu_\star\big)$. En ambos casos, la idea central consiste en combinar:

(i) la aproximación constructiva en variación total de [RZ24], con
(ii) una desigualdad de Pinsker inversa, cuyo coeficiente depende de $\|\rho_1/\rho_2\|_{L^\infty}$, y
(iii) un control constructivo de las colas que asegura que este supremo esencial se mantenga finito.

Parte III. Aplicaciones

Capítulo 7: Modelado de Dinámicas Distribucionales para Salud Digital

En este capítulo, propusimos un enfoque simple e interpretable para aprender dinámicas distribucionales a partir de muestras instantáneas. El modelo representa cada μ_t mediante una mezcla gaussiana con medias y covarianzas compartidas, y pesos dependientes del tiempo $\alpha(t) \in \Delta^{K-1}$. El estimador resultante combina:

(i) un ajuste de minimización de la discrepancia media máxima (MMD) por instante de tiempo con términos gaussianos de forma cerrada y una actualización de pesos estabilizada mediante penalización *ridge*, y
(ii) una etapa de suavizado en tiempo continuo en la cual $\alpha(t)$ se obtiene integrando una EDO neuronal y proyectando de vuelta sobre el símplex.

En los bancos de pruebas sintéticos, el método es competitivo frente a alternativas no paramétricas flexibles y puede superarlas en dimensiones altas, a la vez que proporciona trayectorias de pesos que cuantifican directamente los cambios distribucionales. En el ensayo de monitorización continua de glucosa, las trayectorias aprendidas separan fenotipos de control glucémico clínicamente significativos y revelan respuestas heterogéneas dentro de un mismo grupo que son difíciles de detectar a partir de resúmenes estáticos. Finalmente, los resultados de aproximación y estabilidad de la Proposición 7.3.1 y el Teorema 7.3.2 justifican la elección de un modelo de diccionario compartido y aclaran el comportamiento en muestras finitas del estimador MMD.

Capítulo 8: Paisaje de Campo Medio de los Transformers

En este capítulo, estudiamos la atención de campo medio sobre la esfera $\mathbb{S}^{d-1}$ como un flujo de gradiente de Wasserstein para una energía que combina una interacción de atención (no normalizada) con un potencial inducido por un perceptrón, donde $\beta > 0$ denota la temperatura inversa (equivalentemente, la fuerza de interacción) del núcleo de atención. Esta formulación está motivada tanto por la perspectiva de partículas interactuantes como por los mecanismos de normalización que restringen los *embeddings* a una variedad compacta. Nuestro primer resultado principal establece la naturaleza discreta de las soluciones estacionarias. Cuando el potencial inducido por el perceptrón se construye a partir de la activación ReLU y $d = 2$, toda medida estacionaria tiene soporte finito (Teorema 8.3.1). La razón es que la contribución de la atención a la primera variación es analítica real sobre $\mathbb{S}^1$, mientras que el término ReLU destruye genéricamente esta analiticidad; como resultado, la condición de estacionariedad solo puede cumplirse en un número finito de puntos de soporte. Cuando el perceptrón utiliza una activación analítica real como GeLU, la estacionariedad por sí sola ya no fuerza la finitud, pero la misma conclusión se mantiene para medidas estacionarias estrictamente estables en el sentido de Wasserstein (Teorema 8.3.3). En $d \geq 2$, las medidas estacionarias son necesariamente singulares con respecto a la medida de superficie uniforme y, para parámetros genéricos, son también puramente atómicas (Teorema 8.3.5).

En el régimen de descenso (por tanto, repulsivo), demostramos una cota de anticoncentración: las masas de los clústeres a la escala angular natural $\beta^{-1/2}$ están controladas uniformemente por una constante numérica (Teorema 8.3.7). En consecuencia, la medida límite puede ser atómica sin colapsar en un único átomo, en claro contraste con el régimen atractivo. También acotamos la complejidad del soporte, demostrando que el número de átomos pesados escala, a lo sumo, como $\sqrt{\beta}$ (Corolario 8.3.8). Además, caracterizamos los puntos extremales de la energía (Proposición 8.3.9): los maximizadores globales son masas de Dirac soportadas en los maximizadores del potencial del perceptrón, mientras que el minimizador global es único. Para los perceptrones ReLU, encontrar estos maximizadores se reduce a una familia finita de programas cuadráticos, lo que proporciona

soluciones explícitas en varios regímenes; además, en un escenario de juguete unidimensional calculamos el minimizador explícitamente, lo que hace completamente transparente la competición entre la atención repulsiva y el sesgo del perceptrón.

Los experimentos numéricos en la Sección 8.4 ilustran este comportamiento de agrupamiento en distintas dimensiones y temperaturas inversas, validando empíricamente la dependencia del número de átomos con respecto a β y a los parámetros del modelo, en concordancia cualitativa con nuestras cotas.

También demostramos que estas conclusiones de dispersión (*sparsity*) se extienden más allá de la atención no normalizada. Para la autoatención (*self-attention*) normalizada, la función *softmax* induce un flujo de gradiente de Wasserstein ponderado en lugar del estándar, pero bajo una condición moderada de no degeneración sobre los parámetros del perceptrón, seguimos obteniendo medidas estacionarias singulares en cualquier dimensión, y atomicidad finita en $d = 2$ (Proposición 8.3.10).

Por último, como perspectiva de futuro, analizamos las capas de Mezcla de Expertos (MoE) en $\mathbb{S}^{d-1}$, relacionando el enrutamiento rígido (*hard routing*) lineal *top-k* con los diagramas de Voronoi esféricos de orden k. También propusimos un enrutamiento suave (*soft routing*) basado en potenciales que preserva la estructura del flujo de gradiente de Wasserstein.

Bibliography

[AST25] Á. R. Abella, J. P. Silvestre, and P. Tabuada. Consensus Is All You Get: The Role of Attention in Transformers. In: *Proceedings of the International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. PMLR, 2025, pp. 174–184.

[AS22] A. Agrachev and A. Sarychev. Control on the manifolds of mappings with a view to the deep learning. In: *Journal of Dynamical and Control Systems* 28.4 (2022), pp. 989–1008.

[AS04] A. A. Agrachev and Y. Sachkov. *Control Theory from the Geometric Viewpoint*. Vol. 87. Encyclopaedia of Mathematical Sciences. Springer, 2004.

[Ala+12] F. Alabau-Boussouira, R. Brockett, O. Glass, J. le Rousseau, and E. Zuazua. Notes on the control of the Liouville equation. In: Springer, 2012, pp. 101–129.

[ABV23] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. In: *Journal of Machine Learning Research* 26 (2023), 209:1–209:80.

[AFZ25a] A. Alcalde, G. Fantuzzi, and E. Zuazua. Clustering in Pure-Attention Hardmax Transformers and Its Role in Sentiment Analysis. In: *SIAM Journal on Mathematics of Data Science* 7.3 (2025), pp. 1367–1393. DOI: 10.1137/24M167086X.

[AFZ25b] A. Alcalde, G. Fantuzzi, and E. Zuazua. Exact Sequence Interpolation with Transformers. 2025. arXiv: 2502.02270.

[AGR25] A. Alcalde, B. Geshkovski, and D. Ruiz-Balet. Attention's forward pass and Frank-Wolfe. 2025. arXiv: 2508.09628.

[Alq+23] P. Alquier, B.-E. Chérief-Abdellatif, A. Derumigny, and J.-D. Fermanian. Estimation of copulas via maximum mean discrepancy. In: *Journal of the American Statistical Association* 118.543 (2023), pp. 1997–2012.

[AG24] P. Alquier and M. Gerber. Universal robust regression via maximum mean discrepancy. In: *Biometrika* 111.1 (2024), pp. 71–92.

[Alt25] C. Altafini. Multistability of Self-Attention Dynamics in Transformers. 2025. arXiv: 2511.11553.

[Álv+26] A. Álvarez-López, J. Cheng, Q. Li, and L. Zhang. Flow-based approximation and controllability of measures. 2026.

[ÁGR25] A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Constructive approximate transport maps with normalizing flows. In: *Applied Mathematics & Optimization* 92.2 (2025), p. 33. DOI: 10.1007/s00245-025-10299-7.

[ÁGR26] A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Perceptrons and localization of attention's mean-field landscape. Accepted as a Spotlight paper at the International Conference on Machine Learning (ICML) 2026. 2026. arXiv: 2601.21366.

[ÁLZ26a] A. Álvarez-López, L. Liverani, and E. Zuazua. Constructive interpolation and generalization rates for neural ODEs: a control perspective. 2026. arXiv: 2606.00469.

[ÁLZ26b] A. Álvarez-López, L. Liverani, and E. Zuazua. Exact interpolation and nonparametric generalization rates for neural ODEs: a controllability perspective. 2026.

[ÁM26] A. Álvarez-López and M. Matabuena. Continuous-Time Learning of Probability Distributions: A Case Study in a Digital Trial of Young Children with Type 1 Diabetes. 2026. eprint: 2603.24427.

[ÁOZ25] A. Álvarez-López, R. Orive-Illera, and E. Zuazua. Cluster-Based Classification with Neural ODEs via Control. In: *Journal of Machine Learning* 4.2 (2025), pp. 128–156. DOI: 10.4208/jml.241114.

[ÁSZ24] A. Álvarez-López, A. H. Slimane, and E. Zuazua. Interplay between depth and width for interpolation in neural ODEs. In: *Neural Networks* 180 (2024), p. 106640. DOI: 10.1016/j.neunet.2024.106640.

[AGS08] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. 2nd ed. Lectures in Mathematics ETH Zürich. Birkhäuser Basel, 2008. DOI: 10.1007/978-3-7643-8722-8.

[Ark+06] E. M. Arkin, F. Hurtado, J. S. B. Mitchell, C. Seara, and S. S. Skiena. Some lower bounds on geometric separability problems. In: *Internat. J. Comput. Geom. Appl.* 16.1 (2006), pp. 1–26.

[Aro50] N. Aronszajn. Theory of Reproducing Kernels. In: *Transactions of the American Mathematical Society* 68.3 (1950), pp. 337–404. DOI: 10.1090/S0002-9947-1950-0051437-7.

[Aur91] F. Aurenhammer. Voronoi Diagrams—A Survey of a Fundamental Geometric Data Structure. In: *ACM Computing Surveys* 23.3 (1991), pp. 345–405. DOI: 10.1145/116873.116880.

[Aza13] D. Azagra. Global and fine approximation of convex functions. In: *Proceedings of the London Mathematical Society* 107.4 (2013), pp. 799–824. DOI: 10.1112/plms/pds099.

[ALM21] D. Azagra, E. Le Gruyer, and C. Mudarra. Kirszbraun's Theorem via an Explicit Formula. In: *Canadian Mathematical Bulletin* 64.1 (2021), pp. 142–153. DOI: 10.4153/S0008439520000314.

[BKH16] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. 2016. arXiv: 1607.06450.

[Bac17] F. Bach. Breaking the Curse of Dimensionality with Convex Neural Networks. In: *Journal of Machine Learning Research* 18.19 (2017), pp. 1–53.

[Bac24] F. Bach. *Learning Theory from First Principles*. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2024.

[BBR25] K. Balasubramanian, S. Banerjee, and P. Rigollet. On the Structure of Stationary Solutions to McKean–Vlasov Equations with Applications to Noisy Transformers. 2025. arXiv: 2510.20094.

[Bar+25] F. Barbero, A. Arroyo, X. Gu, C. Perivolaropoulos, P. Veličković, R. Pascanu, and M. M. Bronstein. Why do LLMs attend to the first token? In: *Conference on Language Modeling*. 2025. arXiv: 2504.02732.

[BBI25] D. Barbieri, M. Bonforte, and P. Ibarrondo. Is Stochastic Gradient Descent Effective? A PDE Perspective on Machine Learning processes. 2025. arXiv: 2501.08425.

[Bau+17] B. Bauer, L. Devroye, M. Kohler, A. Krzyżak, and H. Walk. Nonparametric estimation of a function from noiseless observations at random points. In: *Journal of Multivariate Analysis* 160 (2017), pp. 93–104. DOI: 10.1016/j.jmva.2017.05.010.

[Ben+03] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. In: *Journal of Machine Learning Research* 3 (2003), pp. 1137–1155.

[BGC17] Y. Bengio, I. Goodfellow, and A. Courville. *Deep learning*. Vol. 1. MIT Press Cambridge, MA, USA, 2017.

[BSF94] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. In: *IEEE Transactions on Neural Networks* 5.2 (1994), pp. 157–166. DOI: 10.1109/72.279181.

[Ben+19] M. Benning, E. Celledoni, M. J. Ehrhardt, B. Owren, and C.-B. Schönlieb. Deep learning as optimal control problems: models and numerical methods. 2019. arXiv: 1904.05657.

[Ber17] D. P. Bertsekas. *Dynamic Programming and Optimal Control*. 4th ed. Vol. 1. Athena Scientific, 2017.

[BD15] G. Biau and L. Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015. DOI: 10.1007/978-3-319-25388-6.

[BMV22] D. Bilyk, R. W. Matzke, and O. Vlasiuk. Positive Definiteness and the Stolarsky Invariance Principle. In: *Journal of Mathematical Analysis and Applications* 513.2 (2022), p. 126220. DOI: 10.1016/j.jmaa.2022.126220.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York, NY: Springer, 2006. DOI: 10.1007/978-0-387-45528-0.

[BKM17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. In: *Journal of the American statistical Association* 112.518 (2017), pp. 859–877.

[BG24] L. Bleistein and A. Guilloux. On the Generalization and Approximation Capacities of Neural Controlled Differential Equations. In: *International Conference on Learning Representations*. 2024. DOI: 10.48550/arXiv.2305.16791.

[BR25] E. Boix-Adsera and P. Rigollet. The Power of Fine-Grained Experts: Granularity Boosts Expressivity in Mixture of Experts. 2025. arXiv: 2505.06839.

[BU95] R. P. Boland and J. Urrutia. Separating collections of points in Euclidean spaces. In: *Inform. Process. Lett.* 53.4 (1995), pp. 177–183.

[Bon+23] B. Bonnet, C. Cipriani, M. Fornasier, and H. Huang. A measure theoretical approach to the mean-field maximum principle for training NeurODEs. In: *Nonlinear Analysis* 227 (2023), p. 113161. DOI: 10.1016/j.na.2022.113161.

[Bre01] L. Breiman. Random Forests. In: *Machine Learning* 45.1 (2001), pp. 5–32.

[BPA25a] G. Bruno, F. Pasqualotto, and A. Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In: *Advances in Neural Information Processing Systems*. 2025.

[BPA25b] G. Bruno, F. Pasqualotto, and A. Agazzi. Emergence of Meta-Stable Clustering in Mean-Field Transformer Models. In: *International Conference on Learning Representations*. 2025. DOI: 10.48550/arXiv.2410.23228.

[Bur+25] M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 383.2298 (2025), p. 20240233. DOI: 10.1098/rsta.2024.0233.

[CCP15] J. A. Cañizo, J. A. Carrillo, and F. S. Patacchini. Existence of Compactly Supported Global Minimisers for the Interaction Energy. In: *Archive for Rational Mechanics and Analysis* 217.3 (2015), pp. 1197–1217. DOI: 10.1007/s00205-015-0852-3.

[CFP17] J. A. Carrillo, A. Figalli, and F. S. Patacchini. Geometry of minimizers for the interaction energy with mildly repulsive potentials. In: *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* 34.5 (2017), pp. 1299–1308. DOI: 10.1016/j.anihpc.2016.10.004.

[Cas+25] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré. A Unified Perspective on the Dynamics of Deep Transformers. 2025. arXiv: 2501.18322.

[CD18] J. E. Chacón and T. Duong. *Multivariate kernel smoothing and its applications*. CRC Press, 2018.

[CE87] B. Chazelle and H. Edelsbrunner. An Improved Algorithm for Constructing kth-Order Voronoi Diagrams. In: *IEEE Transactions on Computers* 36.11 (1987), pp. 1349–1354. DOI: 10.1109/TC.1987.5009474.

[Che+18] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural Ordinary Differential Equations. In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018, pp. 6572–6583.

[Che+25a] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Critical attention scaling in long-context transformers. In: (2025). arXiv: 2510.05554.

[Che+25b] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Quantitative Clustering in Mean-Field Transformer Models. 2025. arXiv: 2504.14697.

[Che+25c] J. Cheng, Q. Li, T. Lin, and Z. Shen. Interpolation, Approximation, and Controllability of Deep Neural Networks. In: *SIAM Journal on Control and Optimization* 63.1 (2025), pp. 625–649. DOI: 10.1137/23M1599744.

[CA22] B.-E. Chérief-Abdellatif and P. Alquier. Finite-sample properties of parametric MMD estimation: Robustness to misspecification and dependence. In: *Bernoulli* 28.1 (2022), pp. 181–213.

[Cho40] W.-L. Chow. Über Systeme von linearen partiellen Differentialgleichungen erster Ordnung. In: *Mathematische Annalen* 117 (1940), pp. 98–105. DOI: 10.1007/BF01450011.

[CFS25] C. Cipriani, M. Fornasier, and A. Scagliotti. From NeurODEs to AutoencODEs: A mean-field control framework for width-varying neural networks. In: *European Journal of Applied Mathematics* 36.2 (2025), pp. 188–230. DOI: 10.1017/S0956792524000032.

[CPH22] M. Claverol, A. d. L. H. Parrilla, and C. Huemer. Voronoi Diagrams of Arbitrary Order on the Sphere. 2022. arXiv: 2207.13788.

[Coh+01] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree Approximation and Optimal Encoding. In: *Applied and Computational Harmonic Analysis* 11.2 (2001), pp. 192–226. DOI: 10.1006/acha.2001.0336.

[Cor07] J.-M. Coron. *Control and Nonlinearity*. Vol. 136. Mathematical Surveys and Monographs. Providence, RI: American Mathematical Society, 2007.

[CV95] C. Cortes and V. Vapnik. Support-vector networks. In: *Machine Learning* 20.3 (1995), pp. 273–297.

[Cov65] T. M. Cover. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. In: *IEEE Transactions on Electronic Computers* EC-14.3 (1965), pp. 326–334. DOI: 10.1109/PGEC.1965.264137.

[Cow+25] A. Cowsik, T. Nebabu, X.-L. Qi, and S. Ganguli. Geometric Dynamics of Signal Propagation Predict Trainability of Transformers. In: *Physical Review E* 112.5 (2025), p. 055301.

[Cri+24] C. Criscitiello, Q. Rebjock, A. D. McRae, and N. Boumal. Synchronization on circles and spheres with nonlinear interactions. 2024. arXiv: 2405.18273.

[CLT20] C. Cuchiero, M. Larsson, and J. Teichmann. Deep neural networks, generic universal interpolation, and controlled ODEs. In: *SIAM J. Math. Data Sci.* 2.3 (2020), pp. 901–919.

[Cui+25] H. Cui, F. Behrens, F. Krzakala, and L. Zdeborová. A Phase Transition between Positional and Semantic Learning in a Solvable Model of Dot-Product Attention. In: *Journal of Statistical Mechanics: Theory and Experiment* 2025.7 (2025), p. 074001. DOI: 10.1088/1742-5468/ade137.

[CB94] A. Cutler and L. Breiman. Archetypal analysis. In: *Technometrics* 36.4 (1994), pp. 338–347.

[Cyb89] G. Cybenko. Approximation by superpositions of a sigmoidal function. In: *Mathematics of Control, Signals and Systems* 2.4 (1989), pp. 303–314. DOI: 10.1007/BF02551274.

[DM90] B. Dacorogna and J. Moser. On a partial differential equation involving the Jacobian determinant. In: *Annales de l'Institut Henri Poincaré C, Analyse non linéaire* 7.1 (1990), pp. 1–26. DOI: 10.1016/S0294-1449(16)30981-0.

[Dec86] R. Dechter. Learning while searching in constraint-satisfaction-problems. In: *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence*. AAAI'86. Philadelphia, Pennsylvania: AAAI Press, 1986, pp. 178–183.

[DHP21] R. DeVore, B. Hanin, and G. Petrova. Neural network approximation. In: *Acta Numerica* 30 (2021), pp. 327–444.

[DeV98] R. A. DeVore. Nonlinear Approximation. In: *Acta Numerica* 7 (1998), pp. 51–150. DOI: 10.1017/S0962492900002816.

[DKB15] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear independent components estimation. In: *International Conference on Learning Representations*. 2015. DOI: 10.48550/arXiv.1410.8516.

[Dob79] R. L. Dobrushin. Vlasov equations. In: *Funktsional. Anal. i Prilozhen.* 13.2 (1979), pp. 48–58.

[Don00] D. Donoho. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality. In: *AMS Math Challenges Lecture* (2000), pp. 1–32.

[Dor22] J. R. Dorronsoro Ibero. Tecnología, computación e inteligencia. In: *Encuentros multidisciplinares* 24.72 (2022). Lección inaugural de la UAM (curso académico 2006–2007), publicada en la sección "Conferencias célebres", pp. 1–24.

[DC25] Y. Duan and Y. Cai. A Minimal Control Family of Dynamical Systems for Universal Approximation. In: *IEEE Transactions on Automatic Control* 70.11 (2025), pp. 7233–7244. DOI: 10.1109/TAC.2025.3569123.

[DM20] P. Dubey and H.-G. Müller. Functional models for time-varying random objects. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.2 (2020), pp. 275–327.

[Duc06] W. Duch. K-separability. In: *Artificial Neural Networks - ICANN 2006*. Vol. 4131. Lecture Notes in Computer Science. 2006, pp. 188–197.

[Dud69] R. M. Dudley. The Speed of Mean Glivenko-Cantelli Convergence. In: *The Annals of Mathematical Statistics* 40.1 (1969), pp. 40–50. DOI: 10.1214/aoms/1177697802.

[Dud02] R. M. Dudley. *Real Analysis and Probability*. Vol. 74. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511755347.

[DDT19] E. Dupont, A. Doucet, and Y. W. Teh. Augmented Neural ODEs. In: *Advances in Neural Information Processing Systems*. 2019, pp. 3140–3150.

[DMR19] M. Duprez, M. Morancey, and F. Rossi. Approximate and exact controllability of the continuity equation with a localized vector field. In: *SIAM Journal on Control and Optimization* 57.2 (2019), pp. 1284–1311.

[Dur19] R. Durrett. *Probability. Theory and Examples*. 5th ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. DOI: 10.1017/9781108591034.

[Dut+21] S. Dutta, T. Gautam, S. Chakrabarti, and T. Chakraborty. Redesigning the Transformer Architecture with Insights from Multi-Particle Dynamical Systems. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 5531–5544.

[E17] W. E. A proposal on machine learning via dynamical systems. In: *Communications in Mathematics and Statistics* 5.1 (2017), pp. 1–11. DOI: 10.1007/s40304-017-0103-z.

[EHL18] W. E, J. Han, and Q. Li. A mean-field optimal control formulation of deep learning. In: *Research in the Mathematical Sciences* 6.1 (2018), p. 10. DOI: 10.1007/s40687-018-0172-y.

[Ela+22] K. Elamvazhuthi, B. Gharesifard, A. L. Bertozzi, and S. Osher. Neural ode control for trajectory approximation of continuity equation. In: *IEEE Control Systems Letters* 6 (2022), pp. 3152–3157.

[Ela+23] K. Elamvazhuthi, X. Zhang, S. Oymak, and F. Pasqualetti. Learning on Manifolds: Universal Approximations Properties using Geometric Controllability Conditions for Neural ODEs. In: *Learning for Dynamics and Control Conference*. Proceedings of Machine Learning Research. 2023, pp. 1–11.

[ES15] R. Eldan and O. Shamir. The Power of Depth for Feedforward Neural Networks. In: *JMLR: Workshop and Conference Proceedings* 49 (2015), pp. 1–34.

[EH14] T. van Erven and P. Harremos. Rényi divergence and Kullback-Leibler divergence. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820.

[EG23] C. Esteve-Yagüe and B. Geshkovski. Sparsity in long-time control of neural ODEs. In: *Systems & Control Letters* 172 (2023), p. 105452. DOI: https://doi.org/10.1016/j.sysconle.2022.105452.

[FZS22] W. Fedus, B. Zoph, and N. Shazeer. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. In: *Journal of Machine Learning Research* 23.120 (2022), pp. 1–39.

[FZ00] E. Fernández-Cara and E. Zuazua. The cost of approximate controllability for heat equations: the linear case. In: *Advances in Differential Equations* 5.4-6 (2000), pp. 465–514.

[FZ03] E. Fernández-Cara and E. Zuazua. Control theory: History, mathematical achievements and perspectives. In: *Bol. Soc. Esp. Mat. Apl* 26 (2003), pp. 79–140.

[Fis36] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. In: *Annals of Eugenics* 7.7 (1936), pp. 179–188.

[FG15] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. In: *Probability Theory and Related Fields* 162.3-4 (2015), pp. 707–738. DOI: 10.1007/s00440-014-0562-9.

[Fra20] A. L. Fradkov. Early History of Machine Learning. In: *IFAC-PapersOnLine* 53.2 (2020). 21st IFAC World Congress, pp. 1385–1390. DOI: https://doi.org/10.1016/j.ifacol.2020.12.1888.

[FMP91] R. Freimer, J. S. B. Mitchell, and C. Piatko. *On the Complexity of Shattering Using Arrangements*. Technical Report. Cornell University, 1991.

[GRV22] M. Gabrié, G. M. Rotskoff, and E. Vanden-Eijnden. Adaptive Monte Carlo augmented with normalizing flows. In: *Proceedings of the National Academy of Sciences* 119.10 (2022), e2109420119.

[Gao+21] R. Gao, F. Liu, J. Zhang, B. Han, T. Liu, G. Niu, and M. Sugiyama. Maximum Mean Discrepancy Test is Aware of Adversarial Attacks. In: *Proceedings of the International Conference on Machine Learning*. ML Research Press, 2021, pp. 3564–3575.

[GJK17] D. Garreau, W. Jitkrittum, and M. Kanagawa. Large sample analysis of the median heuristic. In: (2017). arXiv: 1707.07269.

[Ger+25] N. Gerber, R. Gvalani, M. Hairer, G. Pavliotis, and A. Schlichting. Formation of clusters and coarsening in weakly interacting diffusions. In: (2025). arXiv: 2510.17629.

[Ges+24] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet. Dynamic metastability in the self-attention model. 2024. arXiv: 2410.06833.

[Ges+23] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. The emergence of clusters in self-attention dynamics. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023, pp. 57026–57037.

[Ges+25] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. A Mathematical Perspective on Transformers. In: *Bulletin of the American Mathematical Society* 62.3 (2025), pp. 427–479.

[GRR24] B. Geshkovski, P. Rigollet, and D. Ruiz-Balet. Measure-to-measure interpolation using Transformers. 2024. arXiv: 2411.04551.

[GRS24] B. Geshkovski, P. Rigollet, and Y. Sun. On the number of modes of Gaussian kernel density estimators. In: (2024). arXiv: 2412.09080.

[GR26] B. Geshkovski and D. Ruiz-Balet. Constructive conditional normalizing flows. 2026. arXiv: 2602.08606 [math.OC].

[GZ22] B. Geshkovski and E. Zuazua. Turnpike in optimal control of PDEs, ResNets, and beyond. In: *Acta Numerica* 31 (2022), pp. 135–263. DOI: 10.1017/S0962492922000046.

[Gho+25] R. Ghosal, S. K. Ghosh, J. A. Schrack, and V. Zipunnikov. Distributional outcome regression via quantile functions and its application to modelling continuously monitored heart rate and physical activity. In: *Journal of the American Statistical Association* just-accepted (2025), pp. 1–20.

[GM24] R. Ghosal and M. Matabuena. Multivariate Scalar on Multidimensional Distribution Regression With Application to Modeling the Association Between Physical Activity and Cognitive Functions. In: *Biometrical Journal* 66.7 (2024), e202400042.

[Gho+23] R. Ghosal, V. R. Varma, D. Volfson, I. Hillel, J. Urbanek, J. M. Hausdorff, A. Watts, and V. Zipunnikov. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in Alzheimer's disease. In: *Biostatistics* 24.3 (2023), pp. 539–561.

[GG26] A. Giorlandino and S. Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. In: *International Conference on Learning Representations*. 2026. DOI: 10.48550/arXiv.2505.24333.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[Gra+19] W. Grathwohl, R. T. Q. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In: *International Conference on Learning Representations*. 2019. DOI: 10.48550/arXiv.1810.01367.

[Gre+12] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. In: *Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[Gyö+02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002. DOI: 10.1007/b97848.

[HR18] E. Haber and L. Ruthotto. Stable architectures for deep neural networks. In: *Inverse Problems* 34.1 (2018), p. 014004. DOI: 10.1088/1361-6420/aa9a90. arXiv: 1705.03341.

[HM17] M. Hardt and T. Ma. Identity Matters in Deep Learning. In: *International Conference on Learning Representations*. 2017. DOI: 10.48550/arXiv.1611.04231.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York, NY: Springer, 2009. DOI: 10.1007/978-0-387-84858-7.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In: *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[HG16] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). 2016. arXiv: 1606.08415.

[HZ25] M. Hernández and E. Zuazua. Constructive Universal Approximation and Finite Sample Memorization by Narrow Deep ReLU Networks. 2025. arXiv: 2409.06555 [stat.ML].

[Hir76] M. W. Hirsch. *Differential Topology*. Vol. 33. Graduate Texts in Mathematics. Springer-Verlag, 1976.

[HS97] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[Hor93] K. Hornik. Some new results on neural network approximation. In: *Neural Networks* 6.8 (1993), pp. 1069–1072. DOI: https://doi.org/10.1016/S0893-6080(09)80018-X.

[Hou93] M. F. Houle. Algorithms for weak and wide separation of sets. In: *Discrete Applied Mathematics* 45.2 (1993), pp. 139–159. DOI: https://doi.org/10.1016/0166-218X(93)90057-U.

[Hua03] G.-b. Huang. Learning capability and storage capacity of two-hidden-layer feedforward networks. In: *IEEE Transactions on Neural Networks* 14.2 (2003), pp. 274–281.

[HH91] S.-C. Huang and Y.-F. Huang. Bounds on the number of hidden neurons in multilayer perceptrons. In: *IEEE Transactions on Neural Networks* 2.1 (1991), pp. 47–55.

[Hyv05] A. Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. In: *Journal of Machine Learning Research* 6.24 (2005), pp. 695–709.

[Inv19] A. O. U. R. P. Investigators. The "All of Us" Research Program. In: *New England Journal of Medicine* 381.7 (2019), pp. 668–676. DOI: 10.1056/NEJMsr1809937.

[Ish+23] I. Ishikawa, T. Teshima, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Universal Approximation Property of Invertible Neural Networks. In: *Journal of Machine Learning Research* 24.287 (2023), pp. 1–68.

[IO24] N. Isobe and M. Okumura. Variational Formulations of ODE-Net as a Mean-Field Optimal Control Problem and Existence Results. In: *Journal of Machine Learning* 3.4 (2024), pp. 413–444. DOI: https://doi.org/10.4208/jml.231210.

[Jac+91] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive Mixtures of Local Experts. In: *Neural Computation* 3.1 (1991), pp. 79–87. DOI: 10.1162/neco.1991.3.1.79.

[Jai10] A. K. Jain. Data clustering: 50 years beyond K-means. In: *Pattern recognition letters* 31.8 (2010), pp. 651–666.

[JB19] J. Jia and A. R. Benson. Neural Jump Stochastic Differential Equations. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[JCP25] Y. Jiang, S. Chewi, and A.-A. Pooladian. Algorithms for Mean-Field Variational Inference Via Polyhedral Optimization in the Wasserstein Space. In: *Foundations of Computational Mathematics* (2025). DOI: 10.1007/s10208-025-09721-x.

[JJ94] M. I. Jordan and R. A. Jacobs. Hierarchical Mixtures of Experts and the EM Algorithm. In: *Neural Computation* 6.2 (1994), pp. 181–214. DOI: 10.1162/neco.1994.6.2.181.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. The Variational Formulation of the Fokker–Planck Equation. In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17. DOI: 10.1137/S0036141096303359. eprint: https://doi.org/10.1137/S0036141096303359.

[JS72] V. Jurdjevic and H. J. Sussmann. Control Systems on Lie Groups. In: *Journal of Differential Equations* 12.2 (1972), pp. 313–329. DOI: 10.1016/0022-0396(72)90035-6.

[Juv08] Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Continuous glucose monitoring and intensive treatment of type 1 diabetes. In: *New England Journal of Medicine* 359.14 (2008), pp. 1464–1476. DOI: 10.1056/NEJMoa0805017.

[Juv09] Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. The effect of continuous glucose monitoring in well-controlled type 1 diabetes. In: *Diabetes care* 32.8 (2009), pp. 1378–1383.

[Juv10] Juvenile Diabetes Research Foundation Continuous Glucose Monitoring Study Group. Effectiveness of continuous glucose monitoring in a clinical care environment: evidence

from the Juvenile Diabetes Research Foundation continuous glucose monitoring (JDRF-CGM) trial. In: *Diabetes Care* 33.1 (2010), pp. 17–22.

[Kal60] R. Kalman. On the general theory of control systems. In: *IFAC Proceedings Volumes* 1.1 (1960), pp. 491–502. DOI: https://doi.org/10.1016/S1474-6670(17)70094-8.

[KPR24] N. Karagodin, Y. Polyanskiy, and P. Rigollet. Clustering in causal attention masking. In: *Advances in Neural Information Processing Systems*. 2024.

[KM26] M. Karbevski and A. Mijoski. Key and Value Weights Are Probably All You Need: On the Necessity of the Query, Key, Value Weight Triplet in Self-Attention Transformers. In: *ICLR 2026 Workshop on Deep Generative Models: Theory, Principle, and Efficacy (DeLTa)*. 2026. DOI: 10.48550/arXiv.2510.23912.

[Kat+24] S. Katta, H. Parikh, C. Rudin, and A. Volfovsky. Interpretable Causal Inference for Analyzing Wearable, Sensor, and Distributional Data. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3340–3348.

[Kha02] H. K. Khalil. *Nonlinear Systems*. 3rd ed. Prentice Hall, 2002.

[Kid+20] P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural Controlled Differential Equations for Irregular Time Series. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 6696–6707.

[KB15] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations*. 2015. DOI: 10.48550/arXiv.1412.6980.

[KB18] J. Klusowski and A. Barron. Approximation by combinations of ReLU and squared ReLU ridge functions with l1 and l0 controls. In: *IEEE Transactions on Information Theory* 64.12 (2018), pp. 7649–7656. DOI: 10.1109/TIT.2018.2874447.

[Kno57] H. Knothe. Contributions to the Theory of Convex Bodies. In: *Michigan Mathematical Journal* 4.1 (1957), pp. 39–52. DOI: 10.1307/mmj/1028990175.

[KPB20] I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 3964–3979.

[KBH24] H. Koubbi, M. Boussard, and L. Hernandez. The impact of lora on the emergence of clusters in transformers. In: (2024). arXiv: 2402.15415.

[Koy23] E. Koyuncu. Memorization Capacity of Neural Networks with Conditional Computation. In: *International Conference on Learning Representations*. 2023. DOI: 0.48550/arXiv.2303.11247.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates Inc., 2012, pp. 1097–1105.

[Lec+98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. In: *Nature* 521 (2015), pp. 436–44.

[Lee82] d.-T. Lee. On k-Nearest Neighbor Voronoi Diagrams in the Plane. In: *IEEE Transactions on Computers* 31.6 (1982), pp. 478–487. DOI: 10.1109/TC.1982.1676031.

[Lee12] J. M. Lee. Sard's Theorem. In: *Introduction to Smooth Manifolds*. New York, NY: Springer New York, 2012, pp. 125–149. DOI: 10.1007/978-1-4419-9982-5_6.

[Lep+21] D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In: *International Conference on Learning Representations*. 2021. DOI: 10.48550/arXiv.2006.16668.

[Les+93] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. In: *Neural Networks* 6.6 (1993), pp. 861–867. DOI: 10.1016/S0893-6080(05)80131-5.

[LP17] D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. 2nd ed. Providence, RI: American Mathematical Society, 2017. DOI: 10.1090/mbk/107.

[Li+18] Q. Li, L. Chen, C. Tai, and W. E. Maximum Principle Based Algorithms for Deep Learning. In: *Journal of Machine Learning Research* 18.165 (2018), pp. 1–29. arXiv: 1710.09513.

[LLS22] Q. Li, T. Lin, and Z. Shen. Deep learning via dynamical systems: An approximation perspective. In: *Journal of the European Mathematical Society* 25.5 (2022), pp. 1671–1709. DOI: 10.4171/jems/1221.

[LLS24] Q. Li, T. Lin, and Z. Shen. Deep Neural Network Approximation of Invariant Functions through Dynamical Systems. In: *Journal of Machine Learning Research* 25.278 (2024), pp. 1–57.

[Li+26] Z. Li, K. Liu, L. Liverani, and E. Zuazua. Universal Approximation of Dynamical Systems by Semiautonomous Neural ODEs and Applications. In: *SIAM Journal on Numerical Analysis* 64 (2026), pp. 193–223. DOI: 10.1137/24M1679690.

[LJ18] H. Lin and S. Jegelka. ResNet with One-Neuron Hidden Layers is a Universal Approximator. In: *Advances in Neural Information Processing Systems*. 2018, pp. 6172–6181.

[Lip+23] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In: *International Conference on Learning Representations*. 2023. DOI: 10.48550/arXiv.2210.02747.

[Lu+20] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu. Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View. In: *International Conference on Learning Representations*. 2020. DOI: 10.48550/arXiv.1906.02762.

[LM24] G. Lugosi and M. Matabuena. Uncertainty quantification in metric spaces. In: (2024). arXiv: 2405.05110.

[Mar23] P. Marion. Generalization Bounds for Neural Ordinary Differential Equations and Deep Residual Networks. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., 2023, pp. 48918–48938.

[Mar+24] Y. Marzouk, Z. Ren, S. Wang, and J. Zech. Distribution learning via neural differential equations: a nonparametric statistical perspective. In: *Journal of Machine Learning Research* 25.232 (2024), pp. 1–61. arXiv: 2309.01043.

[Mas+20] S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama. Dissecting Neural ODEs. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 3952–3963.

[MC26] M. Matabuena and C. M. Crainiceanu. Multilevel Functional Distributional Models with Applications to Continuous Glucose Monitoring in Diabetes Clinical Trials. In: *The Annals of Applied Statistics* 20.1 (2026), pp. 476–495. DOI: 10.1214/26-AOAS2139.

[Mat+22] M. Matabuena, P. Félix, Z. A. A. Hammouri, J. Mota, and B. del Pozo Cruz. Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES. In: *Aging Clinical and Experimental Research* 34.12 (2022), pp. 3107–3114. DOI: 10.1007/s40520-022-02260-3.

[Mat+25] M. Matabuena, R. Ghosal, J. E. Aguilar, A. Keshet, R. Wagner, C. Fernández Merino, J. Sánchez Castro, V. Zipunnikov, J.-P. Onnela, and F. Gude. Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics. In: *Scientific Reports* 15.1 (2025), p. 33662. DOI: 10.1038/s41598-025-18119-2.

[MP23] M. Matabuena and A. Petersen. Distributional data analysis of accelerometer data from the NHANES database using nonparametric survey regression models. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 72.2 (2023), pp. 294–313.

[Mat+21] M. Matabuena, A. Petersen, J. C. Vidal, and F. Gude. Glucodensities: A new representation of glucose profiles using distributional data analysis. In: *Statistical methods in medical research* 30.6 (2021), pp. 1445–1464.

[MP43] W. S. McCulloch and W. Pitts. A Logical Calculus of the Ideas Immanent in Nervous Activity. In: *The Bulletin of Mathematical Biophysics* 5.4 (1943), pp. 115–133. DOI: 10.1007/BF02478259.

[MLP17] H. Mhaskar, Q. Liao, and T. Poggio. When and why are deep networks better than shallow ones? In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[MV94] P. W. Michor and C. Vizman. n-transitivity of certain diffeomorphism groups. In: *Acta Mathematica Universitatis Comenianae. New Series* 63.2 (1994), pp. 221–225.

[Mik+13] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In: *International Conference on Learning Representations*. 2013. DOI: 10.48550/arXiv.1301.3781.

[Mit97] T. M. Mitchell. *Machine Learning*. New York: McGraw-Hill, 1997.

[Mit20] B. S. Mityagin. The Zero Set of a Real Analytic Function. In: *Mathematical Notes* 107.3 (2020), pp. 529–530. DOI: 10.1134/S0001434620030189.

[Mua+17] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf. Kernel mean embedding of distributions: A review and beyond. In: *Foundations and Trends in Machine Learning* 10.1-2 (2017), pp. 1–141.

[Mül66] C. Müller. Funk—Hecke formula. In: *Spherical Harmonics*. Vol. 17. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer, 1966, pp. 18–20. DOI: 10.1007/BFb0094783.

[MFK94] D. Mumford, J. Fogarty, and F. Kirwan. *Geometric Invariant Theory*. Springer Berlin, 1994.

[NH18] Q. Nguyen and M. Hein. Optimization Landscape and Expressivity of Deep CNNs. In: *Proceedings of the International Conference on Machine Learning*. PMLR, 2018, pp. 3730–3739.

[OLM+24] T. OLMo, P. Walsh, L. Soldaini, D. Groeneveld, K. Lo, S. Arora, A. Bhagia, Y. Gu, S. Huang, M. Jordan, et al. 2 OLMo 2 Furious. In: *arXiv preprint arXiv:2501.00656* (2024). arXiv: 2501.00656.

[Onk+20] D. Onken, S. W. Fung, X. Li, and L. Ruthotto. OT-Flow: Fast and Accurate Continuous Normalizing Flows via Optimal Transport. In: *AAAI Conference on Artificial Intelligence*. 2020.

[Ott01] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. In: *Communications in Partial Differential Equations* 26.1-2 (2001), pp. 101–174.

[Pap+21] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64. arXiv: 1912.02762.

[PPM17] G. Papamakarios, T. Pavlakou, and I. Murray. Masked Autoregressive Flow for Density Estimation. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017.

[PKG25] J. Park, N. Kok, and I. Gaynanova. Beyond fixed thresholds: optimizing summaries of wearable device data via piecewise linearization of quantile functions. In: *arXiv preprint arXiv:2501.11777* (2025). arXiv: 2501.11777.

[Pas+17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic Differentiation in PyTorch. In: *NIPS 2017 Workshop on Autodiff*. 2017.

[PS25] M. A. Peletier and A. Shalova. Nonlinear diffusion limit of non-local interactions on a sphere. In: (2025). arXiv: 2512.03185.

[Pen23] M. D. Penrose. Random Euclidean coverage from within. In: *Probability Theory and Related Fields* 185 (2023), pp. 747–814. DOI: 10.1007/s00440-022-01182-5.

[PC19] G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. In: *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[PH22] M. Phuong and M. Hutter. Formal Algorithms for Transformers. In: *CoRR* abs/2207.09238 (2022). DOI: 10.48550/arXiv.2207.09238. arXiv: 2207.09238.

[Pin85] A. Pinkus. *n-Widths in Approximation Theory*. Vol. 7. Ergebnisse der Mathematik und ihrer Grenzgebiete (3). Berlin, Heidelberg: Springer, 1985. DOI: 10.1007/978-3-642-69894-1.

[Pin99] A. Pinkus. Approximation theory of the MLP model in neural networks. In: *Acta Numerica* 8 (1999), pp. 143–195. DOI: 10.1017/S0962492900002919.

[Pin15] A. Pinkus. *Ridge Functions*. Cambridge University Press, 2015.

[PRY25] Y. Polyanskiy, P. Rigollet, and A. Yao. Synchronization of Mean-Field Models on the Circle. 2025. arXiv: 2507.22857.

[PW24] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.

[Pon+62] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The Mathematical Theory of Optimal Processes*. Interscience Publishers, 1962.

[Poo+23] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen. Multisample flow matching: straightening flows with minibatch couplings. In: *Proceedings of the International Conference on Machine Learning*. 2023, pp. 28100–28127.

[Qia+21] Z. Qian, W. Zame, L. Fleuren, P. Elbers, and M. van der Schaar. Integrating expert ODEs into neural ODEs: pharmacology and disease progression. In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 11364–11383.

[Rag24] M. Raginsky. Some Remarks on Controllability of the Liouville Equation. In: (2024). arXiv: 2404.14683.

[RM15] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In: *Proceedings of the International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. PMLR. 2015, pp. 1530–1538.

[RS05] R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 54.3 (2005), pp. 507–554. DOI: 10.1111/j.1467-9876.2005.00510.x.

[Ros58] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. In: *Psychological review* 65.6 (1958), p. 386.

[RCD19] Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud. Latent Ordinary Differential Equations for Irregularly-Sampled Time Series. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[RAZ22] D. Ruiz-Balet, E. Affili, and E. Zuazua. Interpolation and approximation via Momentum ResNets and Neural ODEs. In: *Systems & Control Letters* 162 (2022), p. 105182.

[RZ23] D. Ruiz-Balet and E. Zuazua. Neural ODE Control for Classification, Approximation, and Transport. In: *SIAM Review* 65.3 (2023), pp. 735–773. DOI: 10.1137/21M1411433.

[RZ24] D. Ruiz-Balet and E. Zuazua. Control of neural transport for normalising flows. In: *Journal de Mathématiques Pures et Appliquées* 181 (2024), pp. 58–90. DOI: 10.1016/j.matpur.2023.10.005.

[RHW86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0.

[Sam59] A. L. Samuel. Some Studies in Machine Learning Using the Game of Checkers. In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229. DOI: 10.1147/rd.33.0210.

[San+22] M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. Sinkformers: Transformers with Doubly Stochastic Attention. In: *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3515–3530.

[San15] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Vol. 87. Progress in Nonlinear Differential Equations and Their Applications. Birkhäuser, 2015. DOI: 10.1007/978-3-319-20828-2.

[Sas15] I. Sason. On Reverse Pinsker Inequalities. In: (2015). arXiv: 1503.07118.

[SV16] I. Sason and S. Verdú. f-divergence Inequalities. In: *IEEE Transactions on Information Theory* 62.11 (2016), pp. 5973–6006.

[Sca23] A. Scagliotti. Deep Learning approximation of diffeomorphisms via linear-control systems. In: *Mathematical Control and Related Fields* 13.3 (2023), pp. 1226–1257.

[SF25] A. Scagliotti and S. Farinelli. Normalizing flows as approximations of optimal transport maps via linear-control neural ODEs. In: *Nonlinear Analysis* 257 (2025), p. 113811. DOI: 10.1016/j.na.2025.113811.

[SS00] D. Schmitt and J.-C. Spehner. Order-k Voronoi Diagrams, k-Sections, and k-Sets. In: *Discrete and Computational Geometry*. Springer Berlin Heidelberg, 2000, pp. 290–304.

[Sej+13] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. In: *The Annals of Statistics* (2013), pp. 2263–2291.

[SS26] A. Shalova and A. Schlichting. Solutions of stationary McKean–Vlasov equation on a high-dimensional sphere and other Riemannian manifolds. In: *Advances in Nonlinear Analysis* 15.1 (2026), p. 20250141. DOI: 10.1515/anona-2025-0141.

[SH75] M. I. Shamos and D. Hoey. Closest-Point Problems. In: *Proceedings of the 16th Annual Symposium on Foundations of Computer Science (FOCS 1975)*. IEEE, 1975, pp. 151–162. DOI: 10.1109/SFCS.1975.8.

[Sha+17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In: *International Conference on Learning Representations*. 2017. DOI: 10.48550/arXiv.1701.06538.

[Sie25] J. W. Siegel. Optimal Approximation of Zonoids and Uniform Approximation by Shallow Neural Networks. In: *Constructive Approximation* 62.2 (2025), pp. 441–469. DOI: 10.1007/s00365-025-09712-9.

[Sil86] B. W. Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986.

[Son+21] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In: *International Conference on Learning Representations*. 2021. DOI: 10.48550/arXiv.2011.13456.

[SS97] E. Sontag and H. Sussmann. Complete controllability of continuous-time recurrent neural networks. In: *Systems & Control Letters* 30.4 (1997), pp. 177–183. DOI: https://doi.org/10.1016/S0167-6911(97)00002-9.

[Son97] E. D. Sontag. Shattering All Sets of 'k' Points in "General Position" Requires (k — 1)/2 Parameters. In: *Neural Computation* 9.2 (1997), pp. 337–348. DOI: 10.1162/neco.1997.9.2.337.

[Son98] E. D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. 2nd ed. Springer, 1998.

[SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. In: *Journal of Machine Learning Research* 12.7 (2011).

[Sto82] C. J. Stone. Optimal global rates of convergence for nonparametric regression. In: *The annals of statistics* (1982), pp. 1040–1053.

[SB18] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press, 2018.

[Sza+16] Z. Szabó, B. K. Sriperumbudur, B. Póczos, and A. Gretton. Learning theory for distribution regression. In: *Journal of Machine Learning Research* 17.152 (2016), pp. 1–40.

[TT13] E. G. Tabak and C. V. Turner. A family of nonparametric density estimation algorithms. In: *Communications on Pure and Applied Mathematics* 66.2 (2013), pp. 145–164.

[TV10] E. G. Tabak and E. Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. In: *Communications in Mathematical Sciences* 8.1 (2010), pp. 217–233.

[TG23] P. Tabuada and B. Gharesifard. Universal Approximation Power of Deep Residual Neural Networks Through the Lens of Control. In: *IEEE Transactions on Automatic Control* 68.5 (2023), pp. 2715–2728. DOI: 10.1109/TAC.2022.3190051.

[Tib+24] L. Tiberi, F. Mignacco, K. Irie, and H. Sompolinsky. Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers. In: *Advances in Neural Information Processing Systems*. 2024.

[TW16] J. M. Tomczak and M. Welling. Improving variational auto-encoders using Householder flow. In: (2016). arXiv: 1611.09630.

[Ton+24] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. In: *Transactions on Machine Learning Research* (2024).

[Tré05] E. Trélat. *Contrôle optimal : théorie & applications*. Vuibert, 2005.

[Tsy08] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. 1st. Springer Publishing Company, Incorporated, 2008.

[VC71] V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. In: *Theory of Probability and its Applications* 16.2 (1971), pp. 264–280.

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates Inc., 2017, pp. 6000–6010.

[Ver14] S. Verdú. Total variation distance and the distribution of relative information. In: *2014 information theory and applications workshop (ITA)*. IEEE. 2014, pp. 1–3.

[Vil09] C. Villani. *Optimal Transport. Old and New*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer, 2009. DOI: 10.1007/978-3-540-71050-9.

[WJ08] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305.

[Wan24] S. Wang. Timemixer: Decomposable multiscale mixing for time series forecasting. In: *International Conference on Learning Representations*. 2024. DOI: 10.48550/arXiv.2405.14616.

[WB19] J. Weed and F. Bach. Sharp Asymptotic and Finite-Sample Rates of Convergence of Empirical Measures in Wasserstein Distance. In: *Bernoulli* 25.4 A (2019), pp. 2620–2648. DOI: 10.3150/18-BEJ1065.

[Wie32] N. Wiener. Tauberian Theorems. In: *Annals of Mathematics* 33.1 (1932), pp. 1–100.

[Wu+23] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In: *International Conference on Learning Representations*. 2023. DOI: 10.48550/arXiv.2210.02186.

[Xio+20] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu. On Layer Normalization in the Transformer Architecture. In: *Proceedings of the International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 10524–10533.

[Yu+24] Y. Yu, S. Buchanan, D. Pai, T. Chu, Z. Wu, S. Tong, H. Bai, Y. Zhai, B. D. Haeffele, and Y. Ma. White-Box Transformers via Sparse Rate Reduction: Compression Is All There Is? In: *Journal of Machine Learning Research* 25.300 (2024), pp. 1–128.

[YSJ19] C. Yun, S. Sra, and A. Jadbabaie. Small ReLU Networks Are Powerful Memorizers: A Tight Analysis of Memorization Capacity. In: *Advances in Neural Information Processing Systems*. Curran Associates Inc., 2019, pp. 15558–15569.

[Zha+16] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In: *Communications of the ACM* 64 (2016).

[Zha+20] H. Zhang, X. Gao, J. Unterman, and T. Arodz. Approximation Capabilities of Neural ODEs and Invertible Residual Networks. In: *Proceedings of the International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 11086–11095.

[Zua22] E. Zuazua. Control and Machine Learning. In: *SIAM News* 55.8 (Oct. 3, 2022).

Software and code

[Álv24] A. Álvarez-López. 2024. URL: https://github.com/antonioalvarezl/2024-WCS-NODEs.

[Álv25] A. Álvarez-López. 2025. URL: https://github.com/antonioalvarezl/2025-EM_MMD.

[Álv26] A. Álvarez-López. 2026. URL: https://github.com/antonioalvarezl/2026-MLP-Attention-Energy.

[Che21] R. T. Q. Chen. torchdiffeq. 2021. URL: https://github.com/rtqichen/torchdiffeq.

[Ope24] OpenAI. GPT-2. 2024. URL: https://github.com/openai/gpt-2.

[PyT26] PyTorch Team. PyTorch. 2026. URL: https://pytorch.org/.

Index

D-compatible, **125**
D-universal approximation property, **125**
f-divergence, **172**

Activation function, **2**
Adaptive, **3**
Adjoint equation, **6**
Admissible control, **40**
Admissible dataset, **66**
Affine invariant, **125**
Approximate controllability, 122, 168
Approximate transport, **168**
Artificial intelligence, **1**
Attainable flow, 19
Attainable map, **124**
Attainable measure, **124**

Backpropagation, **4**
Base distribution, 6, 121
Binary classification, 17, 39

Canonical k-separability, **45**
Cell-wise controllability, **19**, **96**
Characteristic kernel, **191**
Classification, 40
Conditional flow matching, **122**
Conservative transport equation, **30**
Constructive control, **31**, **65**
Continuity equation, **30**
Continuous glucose monitoring, 20, **187**
Continuous normalizing flow, 6, 20, **121**, 167
Continuous-depth model, **5**
Control complexity, **41**
Control family, **124**
Control input, 17, **29**
Control switch, 18
Control theory, **5**
Controllability, **6**
Controllable, **6**
Controlled continuity equation, **121**
Controlled dynamical system, 29
Convolutional layer, **6**
Convolutional neural network, **6**
Covering radius, **104**
Cross-entropy, 41, 59

Data classification, **41**
Data distribution, 6, 121
Data-dependent control, 19, **40**
Data-generating law, **39**
Dataset, **39**
Decoupled control family, **136**
Deep feedforward network, **4**
Deep learning, **2**
Degenerate second-order positive-definite, **212**
Depth, **4**, 18, 32
Diagonal shrinking property, **132**
Diffusion model, **122**
Discrepancy, **125**
Dynamical system, **5**

Empirical average, **101**
Empirical measure, **40**, 209
Empirical risk, **40**
Empirical risk minimization, **40**
Estimation error, **98**
Euler discretization, **5**
Exact interpolation, 18, **65**
Exact target, **40**
Expert, **230**
Exploding gradient, 4
Expressivity, 17, **39**, 66
Extremal point, **218**

Feasibility, **65**, 168
Feature map, 191
Feed-forward, 187
First variation, **211**
Fit, **40**
Flow, **17**, 39
Flow map, 5, **29**
Flow matching, **122**
Flow-based generative modeling, 19, **121**
Flow-based predictor, **39**
Fokker–Planck equation, **122**

Gating network, **230**
Gaussian error linear unit, 3
Gaussian kernel, **192**
Gaussian mixture, **190**
General position, **43**

Generalization, 17, **39**, 40
Generalization gap, **40**
Generative modeling, **2**, 121
Geometric control, **65**
Gradient ascent/descent, 209
Gradient descent, 4
Granularity, **230**
Ground truth, **39**

Hellinger distance, **172**
Hidden state, **7**
Histogram, **100**

Input, **5**
Input data, **39**
Interacting particle system, 21, 209, 213
Interaction energy, 188
Interpolation, 40

Jensen-Shannon divergence, **137**

Kernel mean embedding, **191**
Key matrix, **214**
Knothe–Rosenblatt rearrangement, **137**
Kolmogorov width, **103**
Kullback–Leibler divergence, 122
Kullback-Leibler divergence, **167**

Label, **39**
Label space, **39**
Large language model, **209**
Linear routing, **230**
Long short-term memory, **7**
Loss function, **39**

Machine learning, **1**
Marginal flows, **137**
Maximum likelihood, 122, **167**
Maximum mean discrepancy, **191**
Mean-field limit, **209**
Median heuristic, **192**
Metastability, 210
Minimax, **41**
Mixture of Experts, **230**
Mixture-of-Experts, 21
Model complexity, 18, 40
Modelado generativo basado en flujos, 26
Monotone rearrangement, **135**
Multi-head attention, **7**
Multilayer perceptron, **4**

Narrow neural ODE, **67**
Nearest-neighbor interpolant, **104**
Neural network, **2**
Neural ODE, **5**, 31
Noiseless, 39, 97
Nonparametric estimator, **96**
Normalization layer, 7, 209
Normalizing flow, 167
Number of switches, **41**

One-hot encoding, **41**
Optimal control, **6**, 173
Ordinary differential equation, **5**
Otto tangent space, **212**

Parameterized vector field, 5
Parsimony, **65**
Perceptron, **2**
Perceptron block, **210**
Permutation UIP, **128**
Permutation-invariant, 123
Physical time, 187
Piecewise-constant, **32**
Pinsker inequality, **169**
Population average, **100**
Population risk, **39**, 97
Positive definite kernel, **191**
Probability flow ODE, **122**
Probability transport, 123
Probability transport problem, **121**
Pushforward map, 19
Pushforward measure, **30**

Query matrix, **214**

Reachability, **6**
Readout, **39**
Realization error, **98**
Rectified linear unit, 3, **32**
Recurrent neural network, **6**
Regression, **18**, **39**, 65
Regularization, **40**
Reinforcement learning, **2**
Relative entropy, **167**
Relaxed target, **40**
Reproducing kernel Hilbert space, **191**
Residual connection, 209
Residual network, **5**
Reverse Pinsker inequality, 169
Reversibility, **137**
Ridge function, **3**
Routing cell, **230**

Routing vector, **230**
Rényi entropy, **172**

Sampler, **121**
Score function, **122**
Score-based diffusion, 122
Second-order positive-definite, **212**
Second-order Wasserstein critical point, **212**
Self-attention, 7, 187, **209**
Semi-autonomous neural ODE, **67**, 106
Separability, **45**
Separating hyperplanes, **45**
Sequence modeling, **7**
Sequential data, 6
Shallow network, **3**
Shallow neural ODE, **67**
Shear flow, 126, **137**
Sigmoidal, 4
Similarity score, **7**
Simultaneous controllability, 18, **65**, 66
Simultaneously controllable, **66**
Single-hidden-layer network, 3
Skip connection, **5**
Softmax, **215**, 220
Stable equilibrium, 213
Static supervised learning, **187**
Stationary solution, **212**
Strict second-order positive-definite, **212**
Supervised learning, **2**, 39, 121

Target, **39**
Temporal complexity, **42**
Test distribution, **39**
Token, **187**
Token embedding, 209, 213
Tokenization, **187**
Total variation distance, **133**
Training, 4, **40**
Training data, **39**
Training set, **40**
Transformer, **7**, 187, **209**
Triangular map, **137**, **138**
Triangular neural ODE, **54**
Two-layer autonomous neural ODE, **68**

Universal approximation property, 123
Universal approximation theorem, **3**
Universal interpolation property, **66**, 252
Unsupervised learning, **2**

Value matrix, **214**
Vanishing gradient, 4
Variational inference, **169**
Vector field, 5
Voronoi cell, **104**, 231
Voronoi diagram, **231**
Voronoi skeleton, 105

Wasserstein critical point, **212**
Wasserstein distance, **127**
Wasserstein gradient flow, 188, 209, **211**, 214
Wasserstein Hessian, **212**
Wasserstein-2 distance, **211**
Weak solution, **30**
Weighted Wasserstein gradient, **215**, 220
Well function, **58**
Width, **3**, 18, 32
Worst-case scenario, 41