Perceptrons and Localization of Attention's Mean-Field Landscape

Antonio Álvarez-López

UNIVERSIDAD AUTÓNOMA DE MADRID FAU DCN-AVH

Joint work with

Borjan Geshkovski (Inria Paris / Sorbonne Université) ◇ **Domènec Ruiz-Balet** (Universitat de Barcelona)

Outline

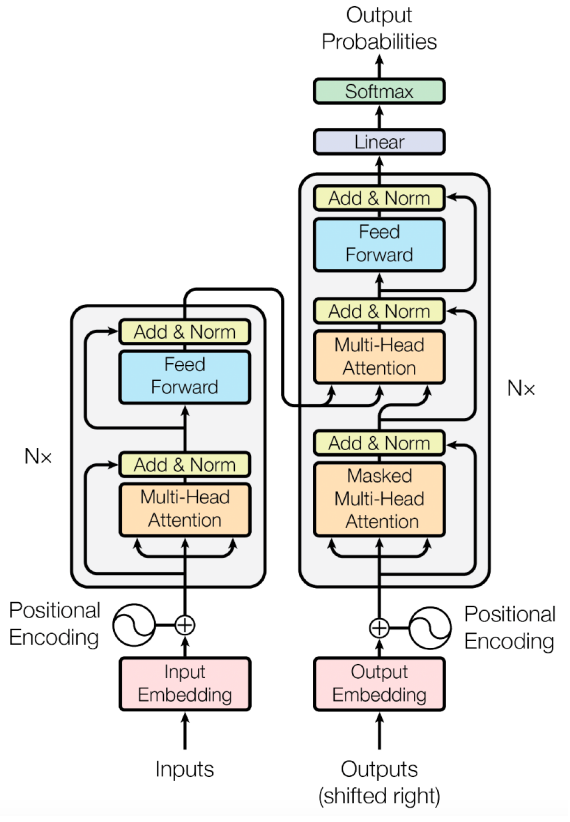
1. The transformer as a gradient flow
2. What's new: attention + perceptron
3. Atomicity
4. Anti-concentration

CHAPTER 1



The transformer as a gradient flow

Motivation



A. Vaswani et al. *Attention is all you need* (2017).

$>$ 250 000 citations as of 2026

Core mechanism:

Maps an input sequence of n tokens $(x_1, \ldots, x_n) \in (\mathbb{R}^d)^n$ to an output sequence $(y_1, \ldots, y_n) \in (\mathbb{R}^d)^n$ via stacked blocks of[1]

a) **Attention**

b) **Feed-forward networks**

c) **Layer normalization**

Today, they are the backbone of modern AI:

- **Language**: GPT, Claude, Gemini, LLaMA...
- **Vision**: ViT, DiT, SAM
- **Science**: AlphaFold, ESM

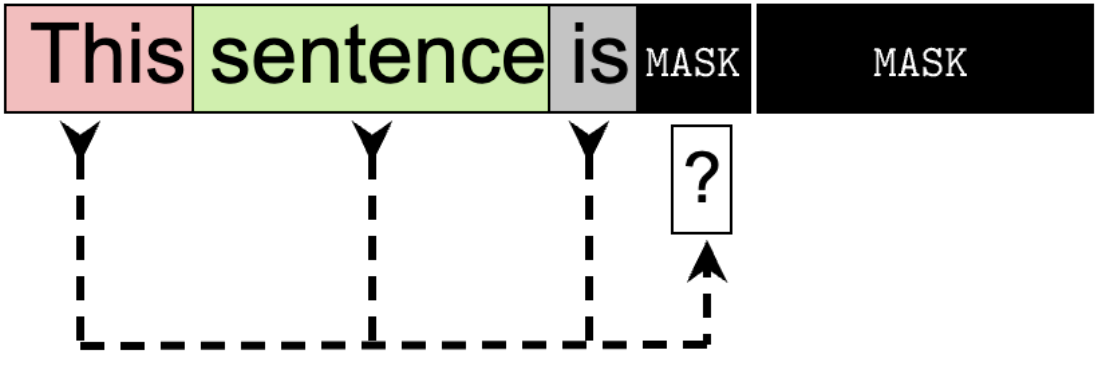
. . . yet the internal dynamics are still poorly understood.

[1] T. Lin et al. *A Survey of Transformers*. AI Open (2022). Comprehensive overview of variants.

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[1] **decoder**[3] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^{i} \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell \rangle}}{\sum_{k=1}^{i} e^{\langle B^\ell x_i^\ell, x_k^\ell \rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^{i} \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell \rangle}}{\sum_{k=1}^{i} e^{\langle B^\ell x_i^\ell, x_k^\ell \rangle}} x_j^\ell \right\|}}_{=: y_i^\ell \quad \text{(self-attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell) \right\|}}_{=: x_i^{\ell+1} \quad \text{(perceptron)}}$$



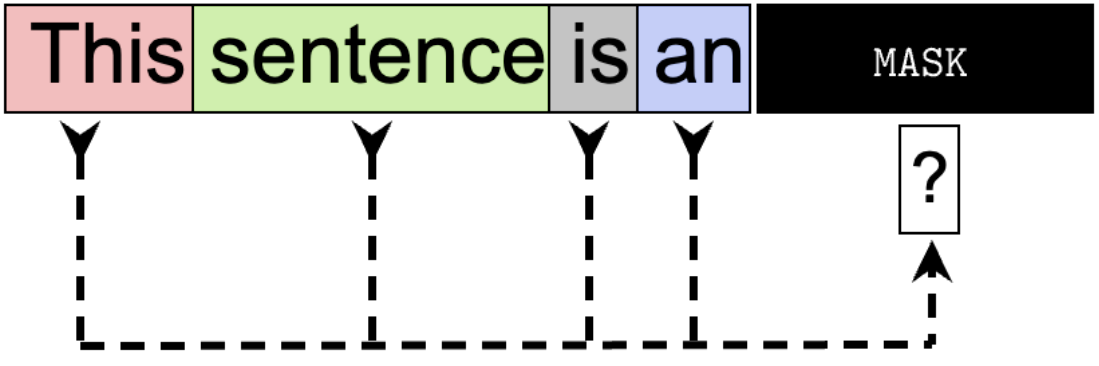
[2] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[3] OpenAI's GPT.

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[1] **decoder**[3] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^{i} \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell \rangle}}{\sum_{k=1}^{i} e^{\langle B^\ell x_i^\ell, x_k^\ell \rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^{i} \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell \rangle}}{\sum_{k=1}^{i} e^{\langle B^\ell x_i^\ell, x_k^\ell \rangle}} x_j^\ell \right\|}}_{=: y_i^\ell \quad \text{(self-attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell) \right\|}}_{=: x_i^{\ell+1} \quad \text{(perceptron)}}$$



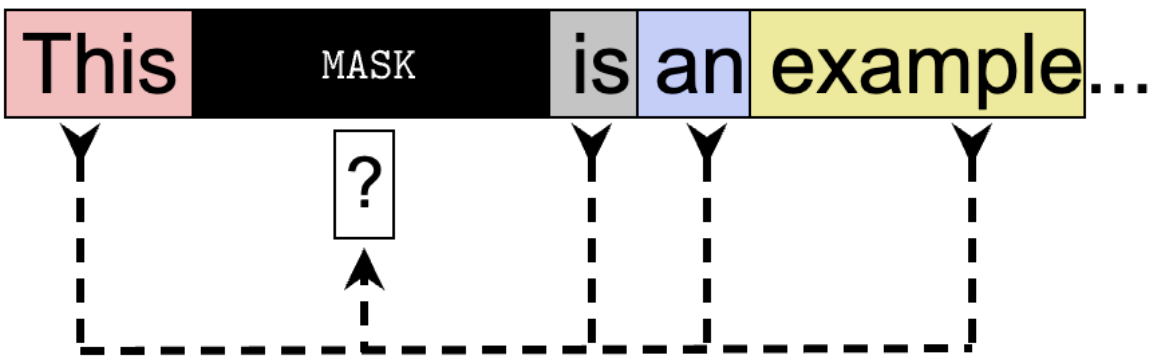
[2] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[3] OpenAI's GPT.

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[1] **encoder**[4] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell, x_k^\ell\rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell, x_k^\ell\rangle}} x_j^\ell \right\|}}_{=: y_i^\ell \quad \text{(self-attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell) \right\|}}_{=: x_i^{\ell+1} \quad \text{(perceptron)}}$$



[2] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[4] BERT model J. Devlin et al., NAACL (2019) (see [PH22, Algorithm 9] as well) and vision Transformers A. Dosovitskiy et al., ICLR (2021).

The forward dynamics of transformers

Roughly, the ℓ-th layer of a post-LN[1] **encoder**[4] maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

$$x_i^\ell \longmapsto \underbrace{\frac{x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell, x_k^\ell\rangle}} x_j^\ell}{\left\| x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell\rangle}}{\sum_{k=1}^n e^{\langle B^\ell x_i^\ell, x_k^\ell\rangle}} x_j^\ell \right\|}}_{=: y_i^\ell \quad \text{(self-attention)}} \longmapsto \underbrace{\frac{y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)}{\left\| y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)\right\|}}_{=: x_i^{\ell+1} \quad \text{(perceptron)}}$$

$\xrightarrow[\text{as } \Delta t \to 0]{\text{first-order}}$ **Interacting particle ODE on the sphere**[5]

$$\dot{x}_i(t) = \mathbf{P}^\perp_{x_i(t)} \left(V(t) \sum_{j=1}^n \frac{e^{\langle B(t)x_i(t), x_j(t)\rangle}}{\sum_{k=1}^n e^{\langle B(t)x_i(t), x_k(t)\rangle}} x_j(t) + W(t)\, \sigma(A(t)x_i(t) + b(t)) \right)$$

where $\mathbf{P}^\perp_x = I - xx^\top$ is the orthogonal projection onto $\mathrm{T}_x\mathbb{S}^{d-1}$.

layers $\leftrightarrow$ time | **tokens $\leftrightarrow$ particles** | **PostLayerNorm $\leftrightarrow$ sphere**

[2] N. Karagodin et al., NeurIPS (2025). Studies the effect of various normalization schemes on the interacting particle system.

[4] BERT model J. Devlin et al., NAACL (2019) (see [PH22, Algorithm 9] as well) and vision Transformers A. Dosovitskiy et al., ICLR (2021).

[5] Y. Lu et al., arXiv:1906.02762 (2019); S. Dutta et al., NeurIPS (2021). First to interpret stacked Transformer layers as a numerical ODE solver (Euler's method + Lie–Trotter splitting) and as an interacting particle system.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(V(t) \sum_{j=1}^{n} \frac{e^{\langle B(t)x_i(t),\, x_j(t)\rangle}}{\sum_k e^{\langle B(t)x_i(t),\, x_k(t)\rangle}} x_j(t) \;+\; W(t)\,\sigma(A(t)x_i(t) + b(t)) \right)$$

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \sum_{j=1}^{n} \frac{e^{\beta\, x_i(t)^\top x_j(t)}}{\sum_k e^{\beta\, x_i(t)^\top x_k(t)}}\, x_j(t) \;+\; W\, \sigma(A x_i(t)) \right)$$

1. Simplifications

- *Attention*: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).
- *Perceptron*: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

- *Attention*: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).
- *Perceptron*: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

- *Unnormalized attention*[6]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$
- *Conservative perceptron*: $W\sigma(A\,x) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \mathrm{diag}(\omega_i)$).

[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

- *Attention*: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).
- *Perceptron*: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

- *Unnormalized attention*[6]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \frac{1}{n}\, e^{\beta\, x_i^\top x_j}$
- *Conservative perceptron*: $W\sigma(A\,x) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \operatorname{diag}(\omega_i)$).

Example: $\mathrm{V}_\theta(x) = \sum_i \omega_i (a_i^\top x)_+^2$ for the ReLU perceptron.

[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \frac{1}{n} \sum_{j=1}^{n} e^{\beta\, x_i(t)^\top x_j(t)}\, x_j(t) \;+\; \frac{1}{2} \nabla \mathrm{V}_\theta(x_i(t)) \right)$$

1. Simplifications

- *Attention*: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).
- *Perceptron*: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

- *Unnormalized attention*[6]: remove softmax $\dfrac{e^{\beta\, x_i^\top x_j}}{\sum_k e^{\beta\, x_i^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, x_i^\top x_j}$
- *Conservative perceptron*: $W\sigma(A\,x) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \operatorname{diag}(\omega_i)$).

Note: Our main results will extend to B symmetric invertible, $b \neq 0$, and softmax attention.

[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

Mean-field variational formulation

$$\partial_t \mu_t \;+\; \mathrm{div}\left(\mu_t\, \mathbf{P}_x^{\perp}\left[\pm \int_{\mathbb{S}^{d-1}} e^{\beta\, x^\top y}\, y \,\mathrm{d}\mu_t(y) \;+\; \frac{1}{2}\nabla \mathrm{V}_\theta(x)\right]\right) \;=\; 0$$

1. Simplifications

- *Attention*: Isotropic and time-independent ($V(t) = \pm I$, $B(t) = \beta I$ with $\beta > 0$).
- *Perceptron*: Fix width $p = d$, constant weights $(W, A) \in (\mathbb{R}^{d\times d})^2$ and zero bias $b = 0$.

2. Gradient structure

- *Unnormalized attention*[6]: remove softmax $\dfrac{e^{\beta\, {x_i}^\top x_j}}{\sum_k e^{\beta\, {x_i}^\top x_k}} \longmapsto \dfrac{1}{n}\, e^{\beta\, {x_i}^\top x_j}$
- *Conservative perceptron*: $W\sigma(A\,x) = \frac{1}{2}\nabla \mathrm{V}_\theta(x)$ for a potential V_θ (iff $W = A^\top \mathrm{diag}(\omega_i)$).

Note: Our main results will extend to B symmetric invertible, $b \neq 0$, and softmax attention.

3. Mean-field limit

- $\mu_t^n := \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)} \xrightarrow[n\to\infty]{W_2} \mu_t$ unif. on $[0, T]$, and the limit solves a continuity equation.

[6] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. Bull. AMS (2025). A different Wasserstein gradient flow formulation is derived in M. Sander, P. Ablin, M. Blondel, G. Peyré. PMLR (2022) by replacing the softmax with doubly stochastic attention via Sinkhorn's algorithm.

A primer on Wasserstein gradient flows

For an energy functional $\mathsf{E} : \mathcal{P}(\mathbb{S}^{d-1}) \to \mathbb{R}$:

- **First variation:** $\frac{\delta \mathsf{E}}{\delta \mu} : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by $\lim_{\varepsilon \to 0} \frac{\mathsf{E}[\mu + \varepsilon \xi] - \mathsf{E}[\mu]}{\varepsilon} = \int \frac{\delta \mathsf{E}}{\delta \mu}[\mu](x) \, \mathrm{d}\xi(x)$.
- **Wasserstein gradient:** $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}}{\delta \mu}[\mu] \in \mathsf{T}(\mathbb{S}^{d-1})$.

Let our energy and its first variation be defined canonically:

$$\mathsf{E}_{\beta,\theta}[\mu] = \frac{1}{2\beta} \iint e^{\beta\, x^\top y} \, \mathrm{d}\mu(x) \mathrm{d}\mu(y) \; + \; \frac{1}{2} \int \mathsf{V}_\theta \, \mathrm{d}\mu, \qquad \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x) = \frac{1}{\beta} \int e^{\beta\, x^\top y} \, \mathrm{d}\mu(y) + \frac{1}{2} \mathsf{V}_\theta(x).$$

The mean-field PDE is a **Wasserstein gradient flow**

($+$ ascent / $-$ descent determined by $V(t) = \pm I$, with sign perturbations absorbed in V_θ)

$$\partial_t \mu_t \; \pm \; \mathrm{div} \left(\mu_t \, \underbrace{\mathbf{P}_x^\perp \left[\int_{\mathbb{S}^{d-1}} e^{\beta\, x^\top y} \, y \, \mathrm{d}\mu_t(y) \; + \; \frac{1}{2} \nabla \mathsf{V}_\theta(x) \right]}_{= \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu_t](x)} \right) \; = \; 0$$

[7] R. Jordan, D. Kinderlehrer, F. Otto. SIMA (1996). L. Ambrosio, N. Gigli, G. Savaré. Lectures in Mathematics (2008).

CHAPTER 2



What's new: attention + perceptron

What's new: attention + perceptron

State of the art

For **pure attention**, the asymptotic picture is well understood:

- Ascent (attractive dynamics): Metastability emerges first, and later tokens collapse into a single Dirac mass (consensus/clustering/synchronization).
- Descent (repulsive dynamics): Particles equidistribute across the sphere (the uniform surface measure is the unique global minimizer).

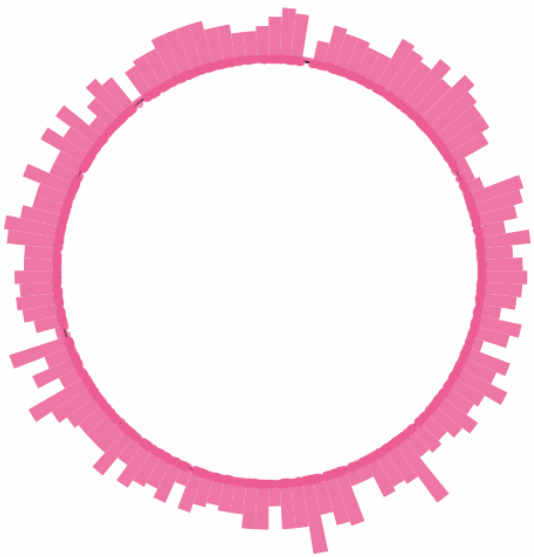
- Sander et al. PMLR (2022).
  Wasserstein gradient flows with doubly stochastic attention.
- Geshkovski et al. NeurIPS (2023).
  First clustering study for attention viewed as interacting particles.
- Castin, Ablin, Peyré. ICML/PMLR (2024).
  Sharp local-Lipschitz bounds for masked/unmasked attention, including the mean-field regime.
- Geshkovski et al. arXiv (2024).
  Metastability: long-lived multi-cluster configurations.
- Karagodin et al. NeurIPS (2024).
  Masked attention clusters (depending only on $\mathrm{spec}(V)$!) despite losing the mean-field gradient-flow structure.
- Yu et al. JMLR (2024).
  Interpretation via sparse rate reduction/compression.
- Abella et al. ICML/PMLR (2025).
  The essential trend towards consensus in attention.
- Alcalde, Fantuzzi, Zuazua. SIMODS (2025).
  Clustering in hardmax transformers.
- Bruno et al. ICLR (2025).
  Formalization of metastable clustered states.
- Bruno et al. NeurIPS (2025).
  Multi-scale analysis: structure formation vs coarsening.
- Burger et al. Phil. Trans. R. Soc. A (2025).
  Weighted gradient flow for general softmax attention.
- Castin, Ablin, Carrillo, Peyré. FoCM (2026).
  Mean-field limit for multi-head, L^2, Sinkhorn, sigmoid and masked attention, plus invariant Gaussian dynamics.
- Chen et al. arXiv (2025).
  Quantitative rates for mean-field clustering.
- Cowsik et al. Phys. Rev. E (2025).
  Signal propagation geometry and depth scaling.
- Geshkovski et al. Bull. AMS (2025).
  Framework: attention as interacting particles and gradient flows.
- Polyanskiy et al., arXiv (2025).
  Synchronization on $\mathbb{S}^1$ for all $\beta \geq -0.16$, but for no $\beta < -2/3$.
- Alcalde, Bungert, Riedl, Roith. arXiv (2026).
  Concentration bounds for mean-field transformers as $1/\beta \to 0+$.
- Alcalde, Fantuzzi, Zuazua. Math. Found. Mach. Learn. (2026).
  Hardmax attention + perceptron for exact sequence interpolation.
- Alcalde, Geshkovski, Ruiz-Balet. ICML (2026).
  Hardmax attention as Frank–Wolfe on the token convex hull.
- Criscitiello et al. SIADS (2026).
  Synchronization criteria on spheres.

What's new: attention + perceptron

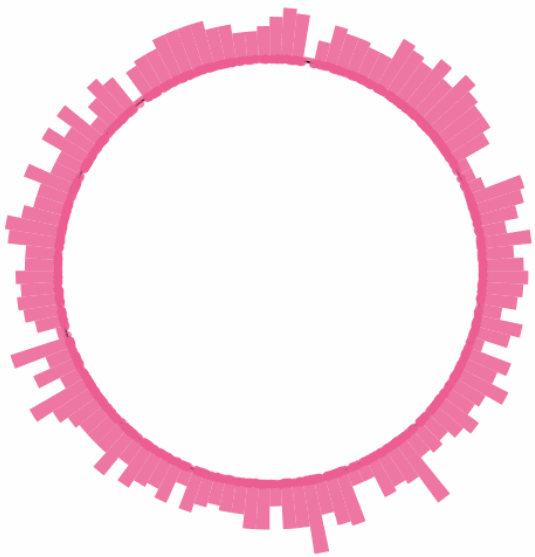
State of the art

For **pure attention**, the asymptotic picture is well understood:

- Ascent (attractive dynamics): Metastability emerges first, and later tokens collapse into a single Dirac mass (consensus/clustering/synchronization).
- Descent (repulsive dynamics): Particles equidistribute across the sphere (the uniform surface measure is the unique global minimizer).



Pure attention in $\mathbb{S}^1$ + gradient ascent.



Pure attention in $\mathbb{S}^1$ + gradient descent.

What's new: attention + perceptron

Our problem

Main question.

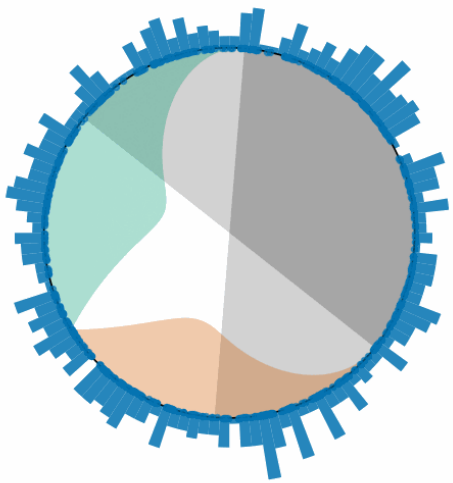
How does adding the perceptron modify the equilibria of pure self-attention?

What's new: attention + perceptron

Our problem

Main question.

How does adding the perceptron modify the equilibria of pure self-attention?

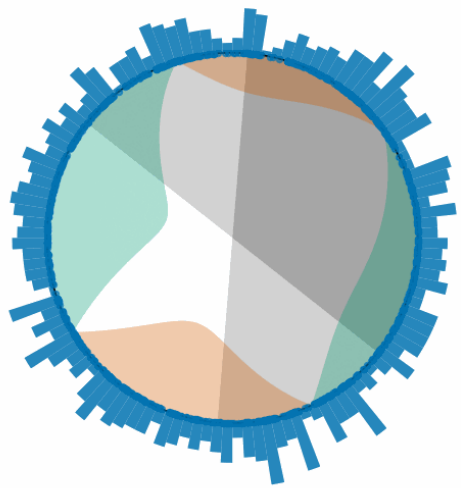


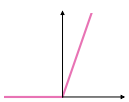
Radial profile: V_θ

$V_\theta(x) > 0$

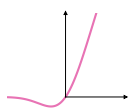
$V_\theta(x) < 0$

$(Ax)_i \leq 0$





ReLU perceptron in $\mathbb{S}^1$
+ gradient ascent.



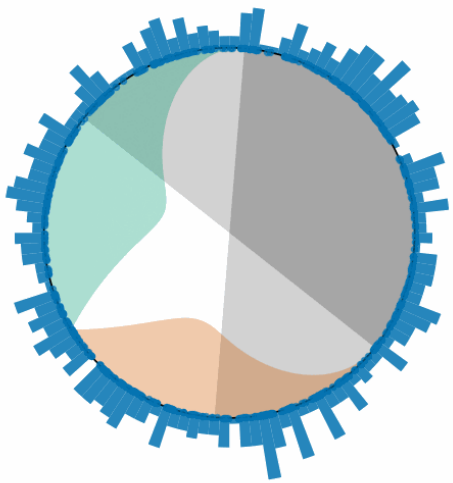
GELU perceptron in $\mathbb{S}^1$
+ gradient ascent.

What's new: attention + perceptron

Our problem

Main question.

How does adding the perceptron modify the equilibria of pure self-attention?

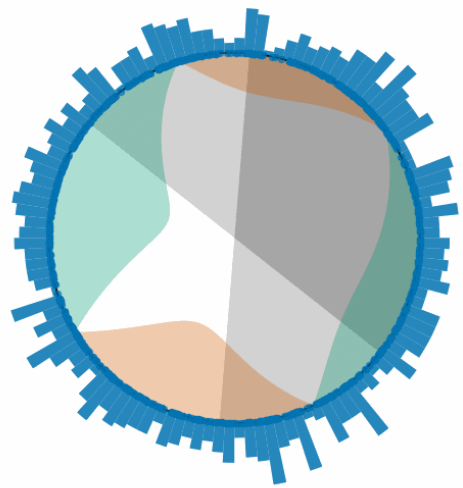


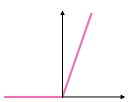
Radial profile: V_θ

$V_\theta(x) > 0$

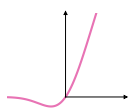
$V_\theta(x) < 0$

$(Ax)_i \leq 0$





ReLU perceptron in $\mathbb{S}^1$
+ gradient descent.



GELU perceptron in $\mathbb{S}^1$
+ gradient descent.

What's new: attention + perceptron

Equilibrium points

Stationary measures:

$$\operatorname{div}\left(\mu \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu]\right) = 0 \qquad \text{(weakly)}$$

What's new: attention + perceptron

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad (\mu\text{-a.e.})$$

What's new: attention + perceptron

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

What's new: attention + perceptron

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Second-order positive-definite (SOPD) critical points:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \;\geq\; 0 \qquad \forall\, \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}).$$

$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathsf{E}(\mu(t))$, with $\mu(t)$ the W_2-geodesic along $\xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla\phi \colon \phi \in C^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$

Strict SOPD: $\operatorname{Hess}_\mu \mathsf{E}(\xi,\xi) \geq \kappa \|\xi\|^2_{L^2(\mu)}$ for some $\kappa > 0$ independent of ξ.

What's new: attention + perceptron

Equilibrium points

Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Second-order positive-definite (SOPD) critical points:

$$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \;\geq\; 0 \qquad \forall\, \xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}).$$

$\operatorname{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathsf{E}(\mu(t))$, with $\mu(t)$ the W_2-geodesic along $\xi \in \mathsf{T}_\mu \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla\phi \colon \phi \in C^\infty(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$

Strict SOPD: $\operatorname{Hess}_\mu \mathsf{E}(\xi,\xi) \geq \kappa \|\xi\|^2_{L^2(\mu)}$ for some $\kappa > 0$ independent of ξ.

Why do we need both notions?

Stationarity yields enough rigidity when V_θ is piecewise polynomial (e.g., for $\sigma = \mathrm{ReLU}$).

Second-order analysis is vital when V_θ is real-analytic (e.g., for $\sigma = \mathrm{GELU}$), where stationarity is too soft and only stable configurations carry actual information.

CHAPTER 3



Atomicity

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Proof sketch.

1. The vector field $\nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu]$ is real-analytic on each cell $I \subset \mathbb{S}^1$ where $\operatorname{sign}(Ax)$ is constant.
2. By the identity theorem, if $\operatorname{supp}(\mu)$ were infinite in any I, the analytic extension of the stationary field associated with I would vanish globally.
3. Integrating this extended field forces $f^\mu(x) := \int e^{\beta x \cdot y} \, d\mu(y)$ to be a global quadratic polynomial.
4. By the lemma, μ has a polynomial density, implying full support ($\implies$ $\operatorname{supp}(\mu) = \mathbb{S}^1$).
5. Full support enforces stationarity everywhere. Comparing the global quadratic identities obtained from adjacent activation cells forces V_θ to be globally real-analytic $\implies$ *contradiction*.

Identity theorem (S. G. Krantz and H. R. Parks. A Primer of Real Analytic Functions (2002))

Let M be a connected real-analytic manifold and let $f : M \to \mathbb{R}$ be real-analytic. If the zero set of f contains a nonempty open subset, then $f \equiv 0$ on M. If $\dim M = 1$, it suffices that the zero set has an accumulation point in M.
B. S. Mityagin. Math. Notes (2020): We can relax "open subset" to "a subset of positive measure on M".

Atomicity on the circle (ReLU)

Stationarity is rigid enough for the ReLU to collapse the support.

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d = 2$, $\beta > 0$, $\sigma(s) = s_+$ and assume that V_θ is not real-analytic. Then every stationary measure $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finitely many atoms.

Proof sketch.

1. The vector field $\nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu]$ is real-analytic on each cell $I \subset \mathbb{S}^1$ where $\operatorname{sign}(Ax)$ is constant.
2. By the identity theorem, if $\operatorname{supp}(\mu)$ were infinite in any I, the analytic extension of the stationary field associated with I would vanish globally.
3. Integrating this extended field forces $f^\mu(x) := \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y)$ to be a global quadratic polynomial.
4. By the lemma, μ has a polynomial density, implying full support ($\implies$ $\operatorname{supp}(\mu) = \mathbb{S}^1$).
5. Full support enforces stationarity everywhere. Comparing the global quadratic identities obtained from adjacent activation cells forces V_θ to be globally real-analytic $\implies$ *contradiction*.

Lemma

For $\beta > 0$, the map $\mu \mapsto \left(f^\mu : x \mapsto \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y) \right)$ is injective on finite signed measures on $\mathbb{S}^{d-1}$, and if f^μ is a polynomial of degree $\leq k$, then μ has a density that is a polynomial of degree $\leq k$.

Tool: By the Funk–Hecke formula, spherical harmonics $Y_{j\ell}$ are eigenfunctions of the kernel: $\int e^{\beta x \cdot y} Y_{j\ell}(y) \mathrm{d}\sigma(y) = \lambda_j(\beta) Y_{j\ell}(x)$ with $\lambda_j(\beta) > 0$.

Atomicity on the circle (GELU)

Under real-analyticity, stationarity is not enough; strict stability collapses the support.

Theorem (GELU activation)

Let $d = 2$, $\beta > 0$, σ real-analytic. Every strict SOPD critical point $\mu \in \mathcal{P}(\mathbb{S}^1)$ of $\mathsf{E}_{\beta,\theta}$ is purely atomic with finitely many atoms.

Atomicity on the circle (GELU)

Under real-analyticity, stationarity is not enough; strict stability collapses the support.

Theorem (GELU activation)

Let $d = 2$, $\beta > 0$, σ real-analytic. Every strict SOPD critical point $\mu \in \mathcal{P}(\mathbb{S}^1)$ of $\mathsf{E}_{\beta,\theta}$ is purely atomic with finitely many atoms.

Proof sketch.

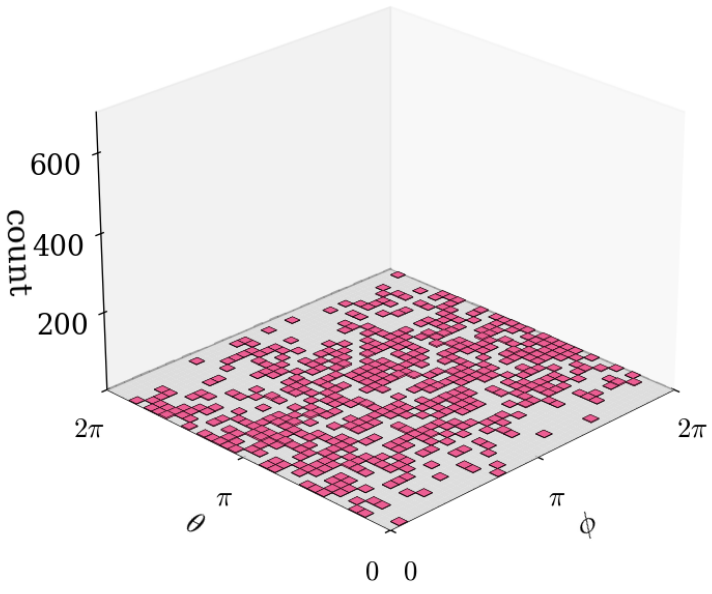
1. Again by the identity theorem, $\mathrm{supp}(\mu)$ being infinite would give $\nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu] = 0$ on $\mathbb{S}^1$, collapsing the Hessian to a bilinear form:

$$\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = -\frac{1}{\beta} \iint \mathsf{K}''_\beta(\theta-\phi)\, \xi(\theta)\xi(\phi)\mathrm{d}\mu(\theta)\mathrm{d}\mu(\phi)$$

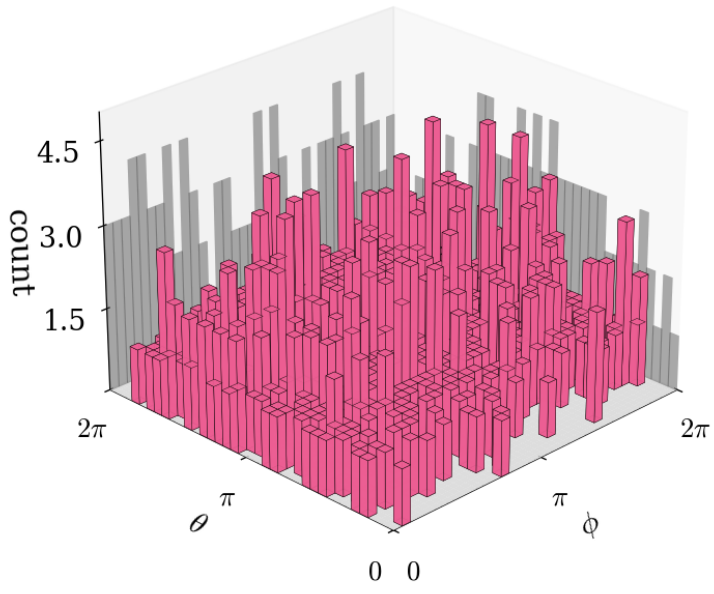
2. Around any accumulation point, we build unit-norm, zero-mean tangent vectors ξ_δ of support size δ.
3. Since $\int \xi_\delta \mathrm{d}\mu = 0$, we can center the kernel ($\mathsf{K}''_\beta \to \mathsf{K}''_\beta - \mathsf{K}''_\beta(0)$). The energy cost then vanishes as the support shrinks, contradicting the strict SOPD coercivity:

$$\left|\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi_\delta,\xi_\delta)\right| \lesssim \frac{1}{\beta} \sup_{|h|\le\delta} \left|\mathsf{K}''_\beta(h) - \mathsf{K}''_\beta(0)\right| \xrightarrow{\delta\to 0} 0$$

What happens if $d \geq 3$?

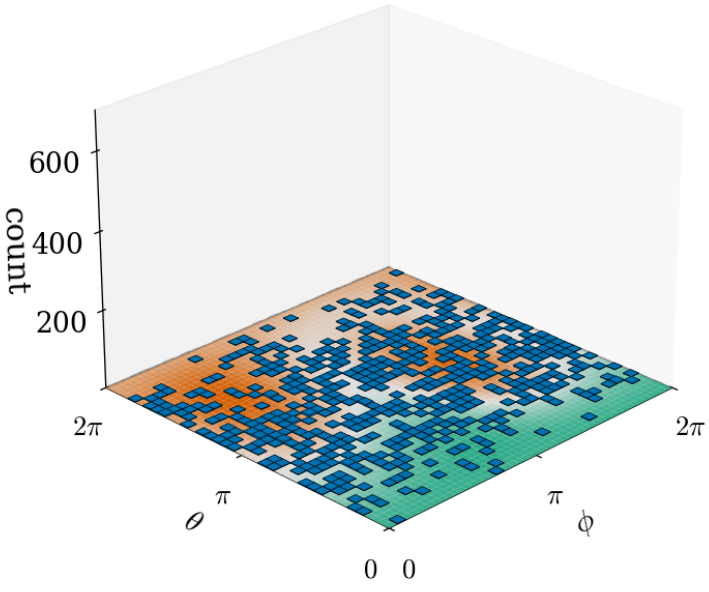


Pure attention in $\mathbb{S}^2$
+ gradient ascent.



Pure attention in $\mathbb{S}^2$
+ gradient descent.

What happens if $d \geq 3$?

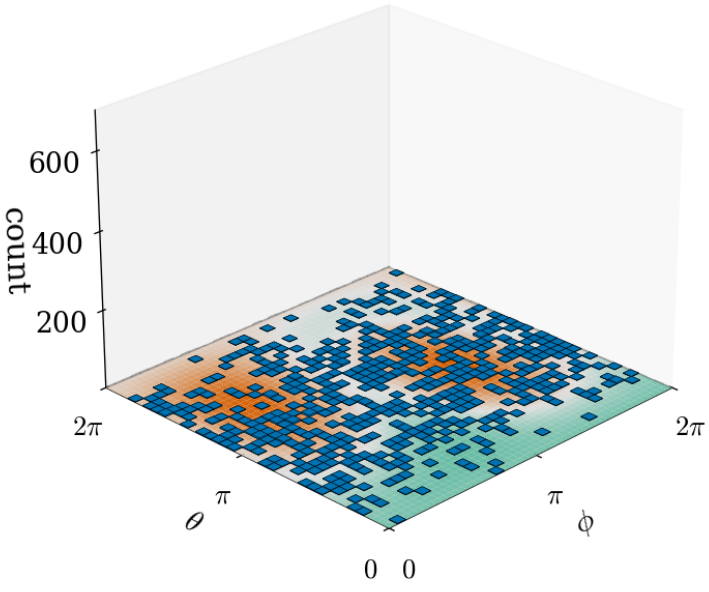


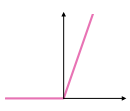
Surface profile: V_θ

$V_\theta(x) > 0$

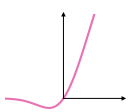
$V_\theta(x) < 0$

$(Ax)_i \leq 0$



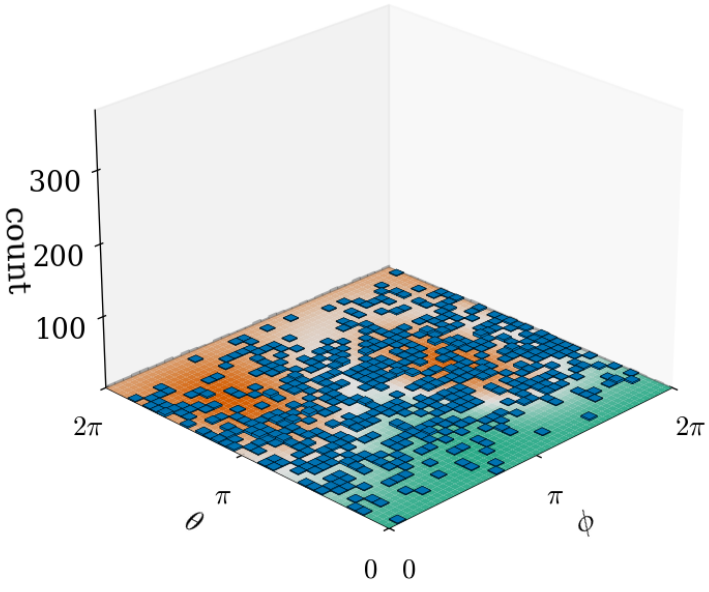


ReLU perceptron in $\mathbb{S}^2$
+ gradient ascent.



GELU perceptron in $\mathbb{S}^2$
+ gradient ascent.

What happens if $d \geq 3$?

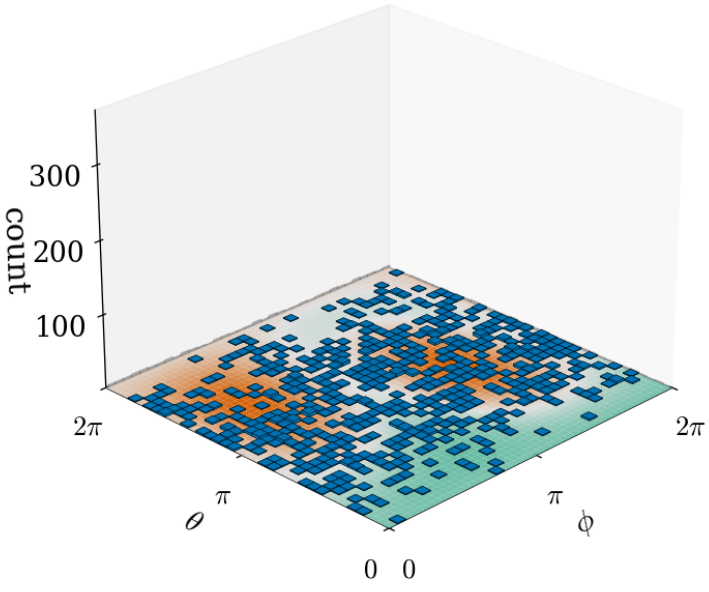


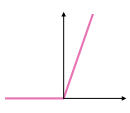
Surface profile: V_θ

$\mathrm{V}_\theta(x) > 0$

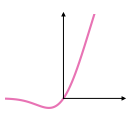
$\mathrm{V}_\theta(x) < 0$

$(Ax)_i \leq 0$





ReLU perceptron in $\mathbb{S}^2$
+ gradient descent.



GELU perceptron in $\mathbb{S}^2$
+ gradient descent.

What happens if $d \geq 3$?

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d \geq 2$. Suppose that μ is either stationary for σ the ReLU and V_θ is not real-analytic, or that μ is a strict SOPD critical point for σ the GELU. Then:

(a) *Singular support:* $\mathrm{supp}(\mu)$ *has zero surface measure.*

(b) *Generic atomicity: For a dense set of (β, θ), μ is purely atomic with finitely many atoms for GELU, and with countably many atoms inside $\{x : (Ax)_+ \neq 0\}$ for ReLU.*

What happens if $d \geq 3$?

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d \geq 2$. Suppose that μ is either stationary for σ the ReLU and V_θ is not real-analytic, or that μ is a strict SOPD critical point for σ the GELU. Then:

(a) *Singular support:* $\mathrm{supp}(\mu)$ *has zero surface measure.*

(b) *Generic atomicity: For a dense set of (β, θ), μ is purely atomic with finitely many atoms for GELU, and with countably many atoms inside $\{x : (Ax)_+ \neq 0\}$ for ReLU.*

Proof sketch. **(a)** Analogous to $d = 2$: a positive-measure support forces a global analytic identity.

(b) Set $g_{\beta,\theta}(x) := \nabla_{\mathbb{S}} \frac{\delta \mathsf{E}_{\beta,\theta}}{\delta \mu}[\mu](x)$. The proof is based on parametric transversality.

Parametric transversality theorem (J. M. Lee. Introduction to Smooth Manifolds (2012). Thm. 6.35)

Let G_λ be a smooth family of vector fields on M, smoothly depending on $\lambda \in \Lambda$. In local coordinates, assume that, whenever $G_\lambda(x) = 0$,

$$\mathrm{Im}\, D_x G_\lambda(x) + \mathrm{Im}\, D_\lambda G_\lambda(x) = T_x M.$$

Then, for a.e. λ, every zero of G_λ is nondegenerate (hence isolated by the inverse function theorem).

What happens if $d \geq 3$?

Theorem (A. Á-L, B. Geshkovski, D. Ruiz-Balet)

Let $d \geq 2$. Suppose that μ is either stationary for σ the ReLU and V_θ is not real-analytic, or that μ is a strict SOPD critical point for σ the GELU. Then:

(a) *Singular support:* $\mathrm{supp}(\mu)$ *has zero surface measure.*

(b) *Generic atomicity: For a dense set of (β, θ), μ is purely atomic with finitely many atoms for GELU, and with countably many atoms inside $\{x : (Ax)_+ \neq 0\}$ for ReLU.*

Proof sketch. **(a)** Analogous to $d = 2$: a positive-measure support forces a global analytic identity.

(b) Set $g_{\beta,\theta}(x) := \nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu}[\mu](x)$. The proof is based on parametric transversality.

We perturb any a_j along $h \in T_x\mathbb{S}^{d-1}$:

$$D_{a_j} g_{\beta,\theta}(x)[h] \;=\; \omega_j \sigma(a_j^\top x)\, h.$$

This is surjective whenever there is an active index ($\omega_j \sigma(a_j^\top x) \neq 0$). Thus, for generic parameters, the zeros are non-degenerate (hence isolated by the inverse function theorem).

For GELU, the zero set is finite by compactness. For ReLU, the argument is applied on every active stratum, yielding at most countably many zeros.

Softmax attention

- **Interacting particle ODE:**

$$\dot{x}_i(t) = \mathbf{P}^{\perp}_{x_i(t)} \left(\pm \sum_{j=1}^{n} \frac{e^{\beta\, x_i(t)\cdot x_j(t)}}{\sum_k e^{\beta\, x_i(t)\cdot x_k(t)}}\, x_j(t) \;+\; \frac{1}{2}\nabla V_\theta(x_i(t)) \right)$$

- **Mean-field limit:**

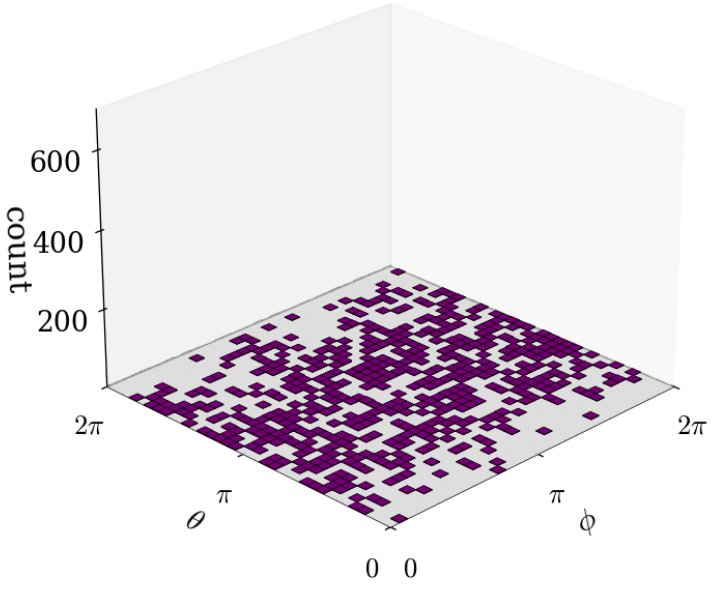
$$\partial_t \mu_t \;+\; \operatorname{div}\left(\mu_t\, \mathbf{P}^{\perp}_{x} \left[\pm\, \frac{1}{\beta} \frac{\nabla \int e^{\beta x\cdot y}\, \mathrm{d}\mu_t(y)}{\int e^{\beta x\cdot y}\, \mathrm{d}\mu_t(y)} \;+\; \frac{1}{2}\nabla V_\theta(x) \right] \right) \;=\; 0$$

The attention term is a Wasserstein gradient scaled by the positive weight $\frac{1}{\int e^{\beta x\cdot y}\, \mathrm{d}\mu(y)}$.

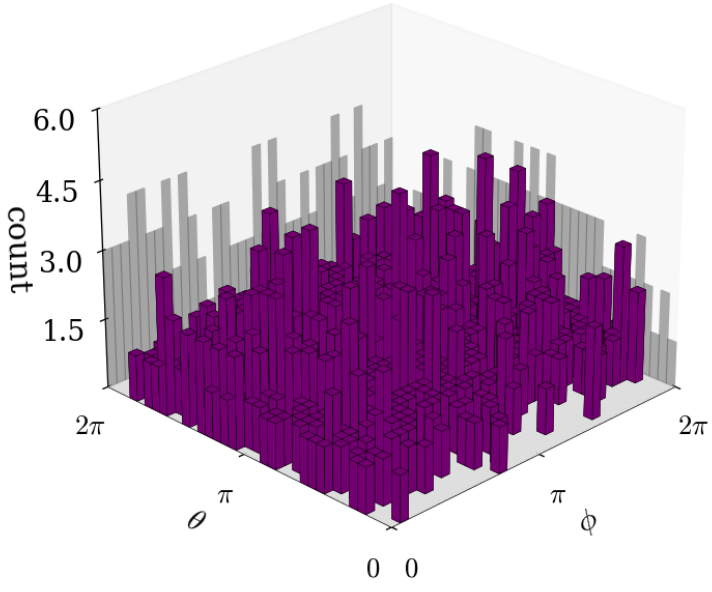
- **Stationarity condition:**

$$\pm \frac{1}{\beta} \nabla_{\mathbb{S}} \log \int e^{\beta x\cdot y}\, \mathrm{d}\mu(y) + \frac{1}{2}\nabla_{\mathbb{S}} V_\theta(x) = 0 \qquad \text{on } \operatorname{supp}(\mu).$$

Softmax attention

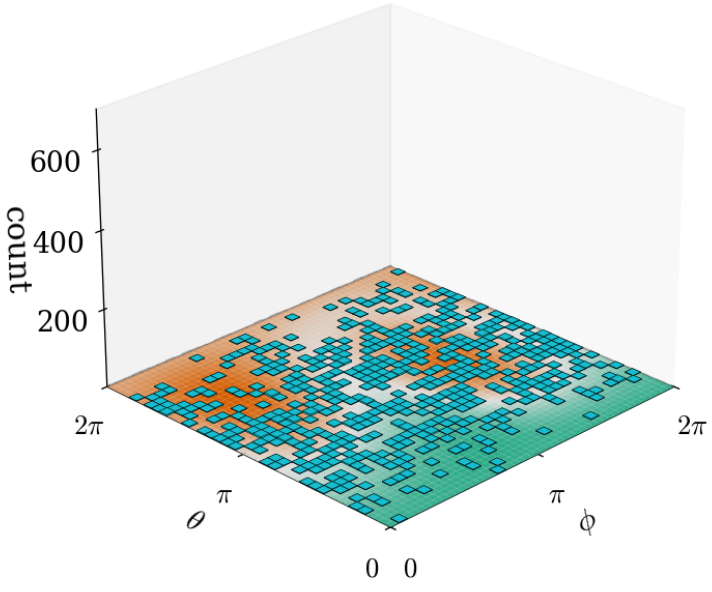


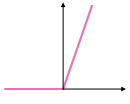
Pure attention in $\mathbb{S}^2$
+ gradient ascent.



Pure attention in $\mathbb{S}^2$
+ gradient descent.

Softmax attention

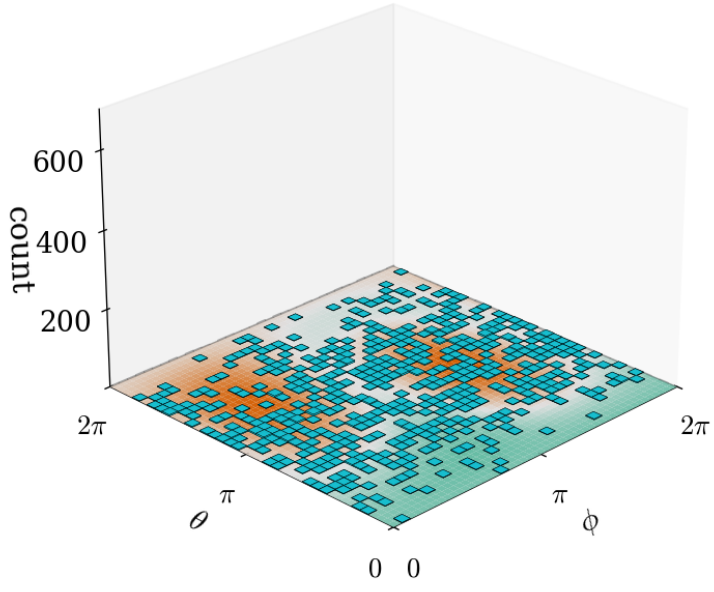


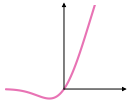


ReLU perceptron in $\mathbb{S}^2$
+ gradient ascent.

Surface profile: V_θ

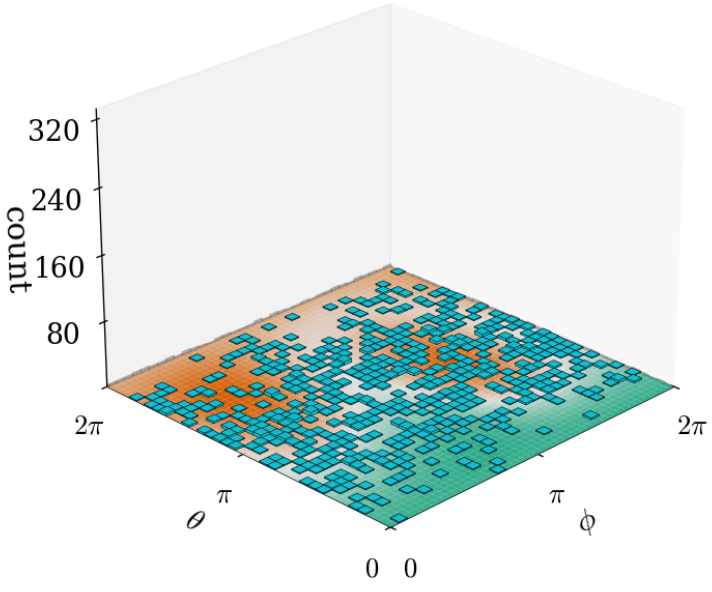
$\mathrm{V}_\theta(x) > 0$
$\mathrm{V}_\theta(x) < 0$
$(Ax)_i \leq 0$





GELU perceptron in $\mathbb{S}^2$
+ gradient ascent.

Softmax attention

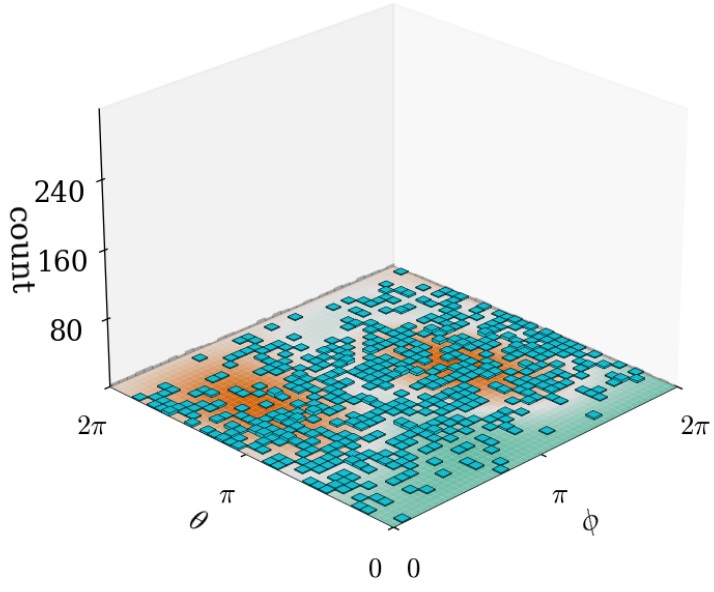


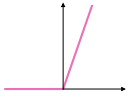
Surface profile: V_θ

$\mathrm{V}_\theta(x) > 0$

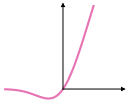
$\mathrm{V}_\theta(x) < 0$

$(Ax)_i \leq 0$





ReLU perceptron in $\mathbb{S}^2$
+ gradient descent.



GELU perceptron in $\mathbb{S}^2$
+ gradient descent.

Softmax attention

Theorem

Let $d \geq 2$, $\beta > 0$, $\sigma(s) = s_+$, and let $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ be a stationary measure. Assume that for any index subset $J \subseteq \{1, \ldots, d\}$, the matrix $M_J := \sum_{j \in J} \omega_j \, a_j a_j^\top - \sum_{j \notin J} \omega_j \, a_j a_j^\top$ is not a scalar multiple of the identity. Then:

(a) *Finiteness: If $d = 2$, μ is purely atomic with finitely many atoms.*

(b) *Singular support: For all $d \geq 2$, $\mathrm{supp}(\mu)$ has zero surface measure.*

Softmax attention

Theorem

Let $d \geq 2$, $\beta > 0$, $\sigma(s) = s_+$, and let $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ be a stationary measure. Assume that for any index subset $J \subseteq \{1, \ldots, d\}$, the matrix $M_J := \sum_{j \in J} \omega_j \, a_j a_j^\top - \sum_{j \notin J} \omega_j \, a_j a_j^\top$ is not a scalar multiple of the identity. Then:

(a) *Finiteness: If $d = 2$, μ is purely atomic with finitely many atoms.*

(b) *Singular support: For all $d \geq 2$, $\mathrm{supp}(\mu)$ has zero surface measure.*

Proof sketch. We prove **(a)**; statement **(b)** follows identically.

1. Let $f^\mu(x) := \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y)$. If $\mathrm{supp}(\mu)$ is infinite on a cell I with index set J, the identity theorem yields:

$$\log f^\mu(x) = C_J + \tfrac{\beta}{2} x^\top Q_J x \qquad \text{on } \mathbb{S}^{d-1}.$$

2. Since the RHS is even, f^μ must be even. By lemma(ii), the antipodal cell $-I$ also has infinite support.
3. Repeating step 1 on $-I$ yields another global equality. Subtracting both reveals that $x^\top M_J x \equiv$ constant on the sphere, forcing $M_J = cI \implies$ *contradiction*.

Lemma

The map $\mu \mapsto \left(f^\mu : x \mapsto \int e^{\beta x \cdot y} \, \mathrm{d}\mu(y) \right)$ is even if and only if $\mu(A) = \mu(-A)$ for all Borel $A \subseteq \mathbb{S}^{d-1}$.

CHAPTER 4

Anti-concentration

Quantitative anti-concentration

Theorem

Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Proof sketch. Test against ξ supported on $\mathcal{S}$ with zero mean ($\sum_{i \in \mathcal{S}} m_i \xi_i = 0$), and split the Wasserstein Hessian into three components:

$$\mathrm{Hess}_{\mu} \mathsf{E}_{\beta,\theta}(\xi, \xi) = \underbrace{\mathcal{T}_{\mathrm{in}}}_{\text{intra-cluster}} + \underbrace{\mathcal{T}_{\mathrm{out}}}_{\text{inter-cluster}} + \underbrace{\mathcal{T}_{\theta}}_{\text{perceptron}} \geq 0.$$

- Intra-cluster: The kernel $e^{\beta \cos\theta}$ is strictly concave for $|\theta| \lesssim \beta^{-1/2}$. Thus, $\mathcal{T}_{\mathrm{in}} < -e^{\beta - 1/8} \sum_{i \in \mathcal{S}} m_i$.
- Inter-cluster: $\mathcal{T}_{\mathrm{out}} > 0$ but bounded by $e^{\beta - 3/2} \sum_{k \notin \mathcal{S}} m_k$.
- Perceptron: $\mathcal{T}_{\theta} > 0$ and bounded using the Lipschitz constant of σ.

Enforcing $\mathcal{T}_{\mathrm{in}} + \mathcal{T}_{\mathrm{out}} + \mathcal{T}_{\theta} \geq 0$ limits the mass $\sum_{i \in \mathcal{S}} m_i$.

Quantitative anti-concentration

Theorem

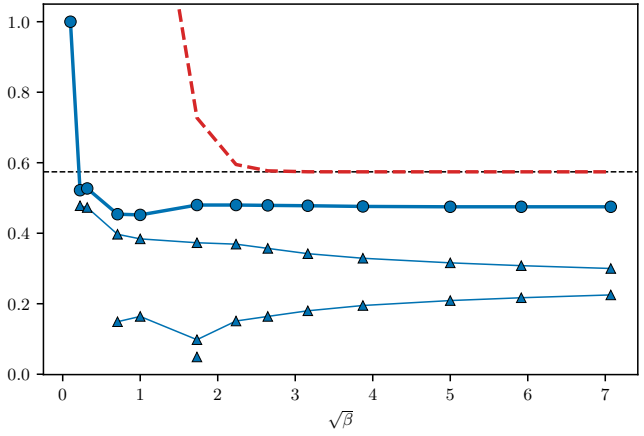
Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \;\leq\; \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \;\leq\; 0.5742 + O(e^{-\beta})$$

Verified empirically in $d = 2$



Gradient descent + GELU

$N = 1000$ particles.

Largest cluster

Other clusters

Red: full bound.

Black: mass $= 0.5742$.

Quantitative anti-concentration

Theorem

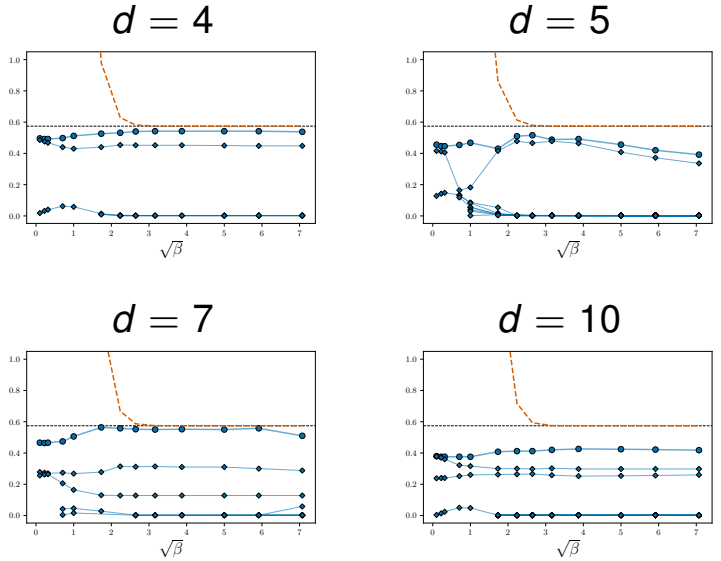
Let $d = 2$, $\mathrm{Lip}(\sigma) = 1$ with $\sigma(0) = 0$. Let $\mu = \sum_{i=1}^{N} m_i \delta_{\varphi_i}$ be a SOPD critical point with $N \geq 2$. If a sub-cluster $\mathcal{S} \subseteq \{1, \dots, N\}$ with $|\mathcal{S}| \geq 2$ satisfies

$$\max_{i,j \in \mathcal{S}} \mathrm{dist}_{\mathbb{S}}(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

then, for $\beta > 0$ large enough,

$$\sum_{i \in \mathcal{S}} m_i \leq 0.5742 + O(e^{-\beta})$$

Even in higher dimensions! (empirically)



Gradient descent + GELU

$N = 1000$ particles.

Largest cluster

Other clusters

Red: full bound.

Black: mass $= 0.5742$.

Conclusions

A (conservative) perceptron *fundamentally reshapes* the energy landscape:

(A) Stationary measures are singular and *generically atomic* (in $d = 2$, finitely supported).

(B) In descent, atomicity $\neq$ trivial collapse: we get quantitative anti-concentration bounds.

(C) Localization: the perceptron *anchors* clusters to specific, seed-independent locations.

Conclusions

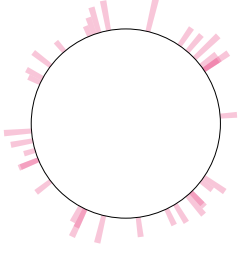
A (conservative) perceptron *fundamentally reshapes* the energy landscape:

(A) Stationary measures are singular and *generically atomic* (in $d = 2$, finitely supported).

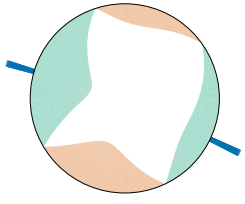
(B) In descent, atomicity $\neq$ trivial collapse: we get quantitative anti-concentration bounds.

(C) Localization: the perceptron *anchors* clusters to specific, seed-independent locations.

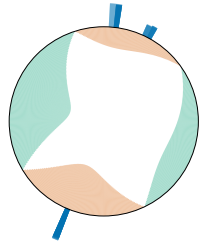
Overlay of stationary measures from 10 independent initializations (GELU, $\beta = 10$)



Pure Att. (Ascent)
Seed-dependent



+ Perceptron (Ascent)
Anchored clusters



+ Perceptron (Descent)
Anchored clusters

Thank you for your attention!



Scan for the paper

A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet.

Perceptrons and localization of attention's mean field landscape.

International Conference on Machine Learning (ICML) 2026 · Spotlight paper

Part of this work was carried out during a research stay supported by the CSIC iMOVE 2024 programme.