Perceptrons and Localization of Attention's Mean-Field Landscape

Antonio Álvarez-López[1], Borjan Geshkovski[2], Domènec Ruiz-Balet[3]

Paper | Code

I. Overview

We study the forward pass of a Transformer as an interacting particle system on $\mathbb{S}^{d-1}$:

layers $\leftrightarrow$ time | tokens $\leftrightarrow$ particles | LayerNorm $\leftrightarrow$ sphere.

For suitable parameters, the mean-field limit is a Wasserstein gradient flow on $\mathcal{P}(\mathbb{S}^{d-1})$.

Question. How does adding the *perceptron block* modify the equilibria of *pure attention*?

Answer at a glance. The perceptron *reshapes* the energy landscape and *localizes* the (even diffuse!) equilibria of pure attention into singular, seed-independent distributions.

II. The variational reduction of a Transformer

Roughly, the ℓ-th layer of a post-LN encoder (e.g. BERT) maps the i-th token $x_i^\ell \in \mathbb{R}^d$ as:

DISCRETE DYNAMICAL SYSTEM

$$x_i^\ell \longmapsto \underbrace{\mathrm{LN}\left(x_i^\ell + \Delta t\, V^\ell \sum_{j=1}^n \frac{e^{\langle B^\ell x_i^\ell, x_j^\ell\rangle}}{\sum_k e^{\langle B^\ell x_i^\ell, x_k^\ell\rangle}} x_j^\ell\right)}_{=:\, y_i^\ell \text{ (Attention)}} \longmapsto \underbrace{\mathrm{LN}\left(y_i^\ell + \Delta t\, W^\ell \sigma(A^\ell y_i^\ell + b^\ell)\right)}_{=:\, x_i^{\ell+1} \text{ (Perceptron)}}$$

where $\mathrm{LN}(s) := s/\|s\|$, and σ is typically ReLU or GELU.

first-order as $\Delta t \to 0$

INTERACTING PARTICLE ODE

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

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

Deriving the gradient flow

1 **Simplifications.**
- ▷ Isotropic, time-independent attention: $V = \pm I_d$ and $B = \beta I_d \quad (\beta > 0)$.
- ▷ Time-independent perceptron and zero bias: $(W(t), A(t))$ constant and $b(t) = 0$.

2 **Gradient structure.**
- ▷ Unnormalized attention: Drop the softmax $e^{\langle Bx_i, x_j\rangle}/(\sum_k e^{\langle Bx_i, x_k\rangle}) \longmapsto e^{\beta x_i \cdot x_j}/n$.
- ▷ Conservative perceptron: $W\sigma(Ax) = \nabla \mathsf{V}_\theta(x)/2 \iff W = A^\top\Omega$ (Ω diagonal matrix).
  - ◦ *Example:* $\mathsf{V}_\theta(x) = \mathbf{1}^\top\Omega(Ax)_+^2$ for the ReLU perceptron.

3 **Mean-field/infinite-context limit.**
- ▷ Token distribution: $\mu_t^n = \frac{1}{n}\sum_i \delta_{x_i(t)} \xrightarrow[n\to\infty]{W_2} \mu_t$ uniformly on $[0,T]$, which solves:

WASSERSTEIN GRADIENT FLOW (ASCENT + / − DESCENT)

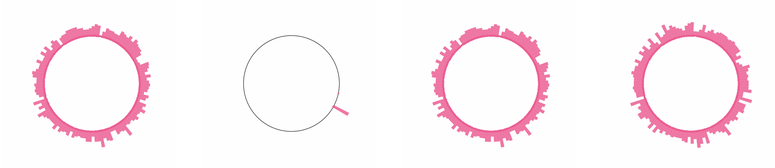
$$\partial_t \mu_t + \mathrm{div}\left(\mu_t\, \mathbf{P}^\perp_x\big[\pm \textstyle\int e^{\beta x\cdot y}\, y\, \mathrm{d}\mu_t(y) + \frac{1}{2}\nabla \mathsf{V}_\theta(x)\big]\right) = 0.$$

- **Energy:** $\mathsf{E}_{\beta,\theta}[\mu] = \frac{1}{2\beta}\iint e^{\beta x\cdot y}\mathrm{d}\mu(x)\mathrm{d}\mu(y) \pm \frac{1}{2}\int \mathsf{V}_\theta(x)\mathrm{d}\mu(x).$
- **Stationary:** $\nabla\frac{\delta \mathsf{E}_{\beta,\theta}}{\delta\mu}[\mu] = 0 \quad \text{on } \mathrm{supp}\,\mu \qquad \left[\frac{\delta \mathsf{E}_{\beta,\theta}}{\delta\mu}[\mu](x) = \frac{1}{\beta}\int e^{\beta x\cdot y}\mathrm{d}\mu(y) \pm \frac{1}{2}\mathsf{V}_\theta(x)\right].$
- **Stable:** $\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \geq 0 \qquad \forall\, \xi \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^{d-1}).$
  - **Strictly stable:** $\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) \geq \kappa\|\xi\|^2_{L^2(\mu)}$ with $\kappa > 0$ indep. of ξ.

III. Baseline: pure attention

For pure attention, the asymptotics are well understood [2, 1]:

- Ascent (attractive): Metastability emerges first, then particles collapse to a single Dirac.
- Descent (repulsive): Particles equidistribute (global minimizer is the uniform measure).



(a) Ascent: initial (b) Ascent: final (c) Descent: initial (d) Descent: final

IV. The perceptron localizes the landscape

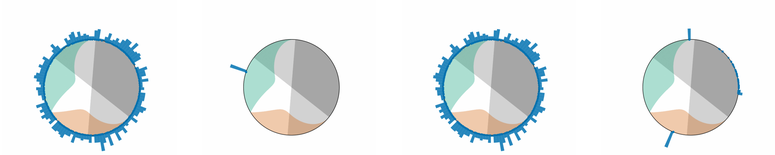
Note: Theorems 1–3 extend to symmetric invertible matrices B, non-zero biases $b \neq 0$ and softmax attention.

Theorem 1 (*"Stationarity is rigid enough for the ReLU on the circle"*)

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

Proof sketch. **1.** On each activation region, stationarity is real-analytic. If $\mathrm{supp}(\mu)$ were infinite, the identity theorem would force $\int e^{\beta x\cdot y}\mathrm{d}\mu(y)$ to be a global polynomial.
2. An invertibility lemma then forces μ to have a polynomial density on $\mathbb{S}^1$.
3. This propagates real-analyticity globally to V_θ, yielding a contradiction.

Radial profile: V_θ — $\mathsf{V}_\theta(x) > 0$ | $\mathsf{V}_\theta(x) < 0$ | $(Ax)_i \leq 0$



(e) ReLU (Ascent: initial) (f) ReLU (Ascent: final) (g) ReLU (Descent: initial) (h) ReLU (Descent: final)

Theorem 2 (*"Under real-analyticity, stationarity is not enough; strict stability collapses the support"*)

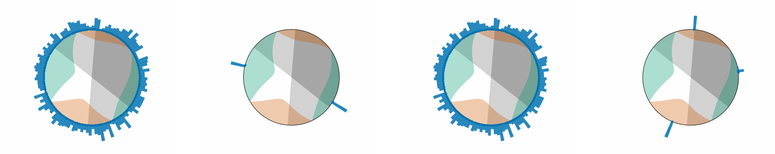
Let $d = 2$, $\beta > 0$ and V_θ real-analytic (e.g. for σ the GELU). Every strictly stable Wasserstein critical point $\mu \in \mathcal{P}(\mathbb{S}^1)$ is purely atomic with finite support.

Proof sketch. **1.** Infinite support + real-analyticity $\implies$ Wasserstein Hessian at μ becomes:

$$\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = -\frac{1}{\beta}\iint \mathsf{K}''_\beta(\theta-\phi)\,\xi(\theta)\xi(\phi)\mathrm{d}\mu(\theta)\mathrm{d}\mu(\phi) \qquad \text{for all } \xi \in \mathsf{T}_\mu\mathcal{P}(\mathbb{S}^1).$$

2. Around any accumulation point, build zero-mean, unit-norm ξ_δ (support size δ) s.t.

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



(i) GELU (Ascent: initial) (j) GELU (Ascent: final) (k) GELU (Descent: initial) (l) GELU (Descent: final)

IV. (continued)

Theorem 3 (*"Singular support and generic atomicity in every dimension"*)

(a) Let μ satisfy the assumptions of either Theorem 1 or Theorem 2, with $d = 2$ replaced by $d \geq 2$. Then $\mathrm{supp}(\mu)$ has zero surface measure.
(b) Let $d \geq 2$ and fix μ. For a dense set of (β,θ), if μ is stationary for (β,θ) then it is purely atomic with finite support for GELU, and purely atomic on $\{x : (Ax)_+ \neq 0\}$ for ReLU.

Proof sketch. (a): Analogous to $d = 2$ (identity theorem for multivariate functions).
(b): For GELU and an open dense set of (β,θ), the stationarity equation has full-rank differential. By parametric transversality, all zeros of the field are isolated. By compactness, they are finite. For ReLU, this argument applies on every active stratum defined by $(Ax)_+$.

V. Anti-concentration (repulsive regime)

Theorem 4 (*"Local concavity prevents a trivial collapse"*)

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

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

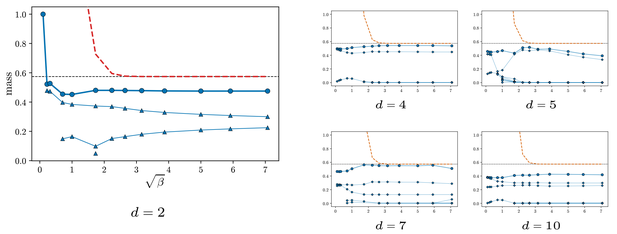
then, for $\beta > 0$ large enough,

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

Proof sketch. Test against ξ supported on $\mathcal{S}$ with zero mean, and split the Hessian into:

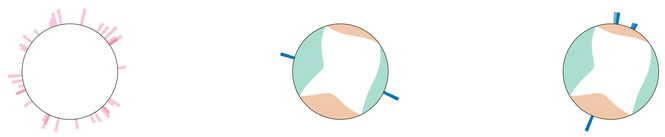
$$\mathrm{Hess}_\mu \mathsf{E}_{\beta,\theta}(\xi,\xi) = \underbrace{\mathcal{T}_{\text{in}}}_{\text{intra-cluster}} + \underbrace{\mathcal{T}_{\text{out}}}_{\text{inter-cluster}} + \underbrace{\mathcal{T}_\theta}_{\text{perceptron}} \;\geq\; 0.$$

The kernel $e^{\beta\cos\varphi}$ is concave for $|\varphi| \leq (2\sqrt{\beta})^{-1}$, thus $\mathcal{T}_{\text{in}} < 0$, while $\mathcal{T}_{\text{out}}, \mathcal{T}_\theta$ are bounded.



Gradient descent (+GELU)
$N = 1000$ particles.
Largest cluster
Smaller clusters
Red dashed: full bound.
Black line: limit 0.5742.

VI. Localization is robust across initializations



(m) Pure att. (ascent) · 10 seeds (n) + GELU perceptron (ascent) (o) + GELU perceptron (descent)

References

[1] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. *The emergence of clusters in self-attention dynamics.* NeurIPS, 2023.
[2] B. Geshkovski, C. Letrouit, Y. Polyanskiy, P. Rigollet. *A mathematical perspective on Transformers.* Bull. Amer. Math. Soc. **62**(3):427–479, 2025.

[1]Universidad Autónoma de Madrid, Spain. [2]Inria & Laboratoire Jacques-Louis Lions, Sorbonne Université, France. [3]Universitat de Barcelona, Spain.

ICML 2026

antonio.alvarezl@uam.es