

# Perceptrons and Localization of Attention's Mean-Field Landscape

*Machine Learning and PDEs Workshop*

**Antonio Álvarez-López**

Joint work with Borjan Geshkovski (Inria/Sorbonne) & Domènec Ruiz-Balet (UB)

Universidad Autónoma de Madrid  
FAU DCN-AvH

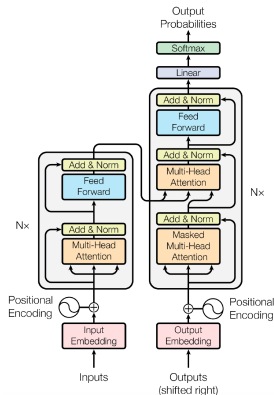
June 23, 2026

# Outline

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



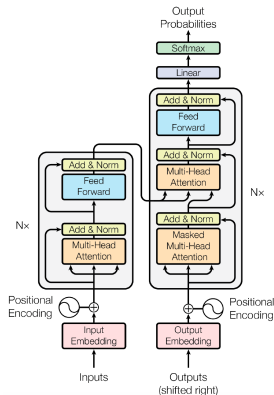
# Motivation



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

> 250 000 citations as of June 2026

# Motivation



## Core mechanism: Seq2Seq

Maps an input sequence of  $n$  tokens  $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$  to an output sequence  $(y_1, \dots, y_n) \in (\mathbb{R}^d)^n$  via stacked blocks of<sup>1</sup>

**Attention**

**Feed-forward networks**

**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.

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

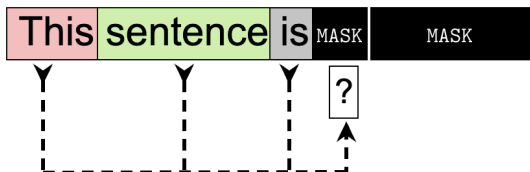
> 250 000 citations as of June 2026

<sup>2</sup> T. Lin et al. *A Survey of Transformers*. *AI Open* (2022). Comprehensive overview of variants.

# The forward dynamics of transformers

Roughly, the  $\ell$ -th layer of a post-LN<sup>3</sup> **decoder**<sup>4</sup> maps the  $i$ -th token  $x_i^\ell \in \mathbb{R}^d$  as:

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



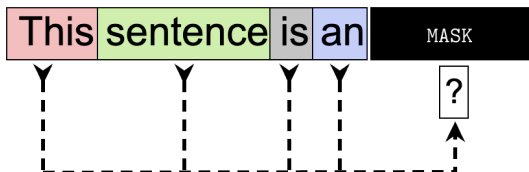
<sup>3</sup> N. Karagodin et al., *NeurIPS* (2025). Studies the effect of various normalization schemes on the interacting particle system.

<sup>4</sup> OpenAI's GPT.

# The forward dynamics of transformers

Roughly, the  $\ell$ -th layer of a post-LN<sup>3</sup> **decoder**<sup>4</sup> maps the  $i$ -th token  $x_i^\ell \in \mathbb{R}^d$  as:

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



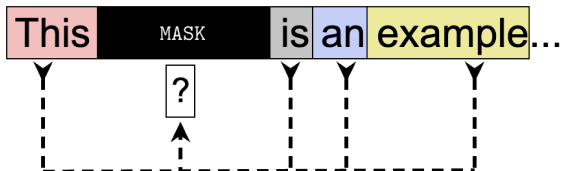
<sup>3</sup> N. Karagodin et al., *NeurIPS (2025)*. Studies the effect of various normalization schemes on the interacting particle system.

<sup>4</sup> OpenAI's GPT.

# The forward dynamics of transformers

Roughly, the  $\ell$ -th layer of a post-LN<sup>3</sup> **encoder**<sup>4</sup> maps the  $i$ -th token  $x_i^\ell \in \mathbb{R}^d$  as:

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



<sup>3</sup>N. Karagodin et al., *NeurIPS (2025)*. Studies the effect of various normalization schemes on the interacting particle system.

<sup>4</sup>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  $\ell$ -th layer of a post-LN<sup>3</sup> encoder maps the  $i$ -th token  $x_i^\ell \in \mathbb{R}^d$  as:

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

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

Interacting particle ODE on the sphere<sup>4</sup>

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

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

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

<sup>3</sup>N. Karagodin et al., *NeurIPS* (2025). Studies the effect of various normalization schemes on the interacting particle system.

<sup>4</sup>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 particle systems.

# Mean-field variational formulation

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

# Mean-field variational formulation

## 1. Simplifications

*Attention:* Isotropic and time-independent ( $V(t) = \pm I$ ,  $B(t) = \beta I$  with  $\beta > 0$ ).

*Perceptron:* constant weights ( $W, A \in (\mathbb{R}^{d \times d})^2$  and zero bias  $b = 0$ ).

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



# Mean-field variational formulation

## 1. Simplifications

*Attention*: Isotropic and time-independent ( $V(t) = \pm I$ ,  $B(t) = \beta I$  with  $\beta > 0$ ).

*Perceptron*: constant weights ( $W, A \in (\mathbb{R}^{d \times d})^2$  and zero bias  $b = 0$ ).

## 2. Gradient structure

*Unnormalized attention*<sup>5</sup>: remove softmax  $\frac{e^{\beta x_i \cdot x_j}}{\sum_k e^{\beta x_i \cdot x_k}} \mapsto \frac{1}{n} e^{\beta x_i \cdot x_j}$ .

*Conservative perceptron*:  $W\sigma(Ax) = \frac{1}{2} \nabla V_\theta(x)$  for a potential  $V_\theta$  (iff  $W = A^\top \text{diag}(\omega_i)$ ).

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

<sup>5</sup>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

## 1. Simplifications

*Attention:* Isotropic and time-independent ( $V(t) = \pm I$ ,  $B(t) = \beta I$  with  $\beta > 0$ ).

*Perceptron:* constant weights  $(W, A) \in (\mathbb{R}^{d \times d})^2$  and zero bias  $b = 0$ .

## 2. Gradient structure

*Unnormalized attention*<sup>5</sup>: remove softmax  $\frac{e^{\beta x_i \cdot x_j}}{\sum_k e^{\beta x_i \cdot x_k}} \mapsto \frac{1}{n} e^{\beta x_i \cdot x_j}$ .

*Conservative perceptron:*  $W\sigma(Ax) = \frac{1}{2} \nabla V_\theta(x)$  for a potential  $V_\theta$  (iff  $W = A^\top \text{diag}(\omega_i)$ ).

*Example:*  $V_\theta(x) = \sum_i \omega_i (a_i \cdot x)_+^2$  for the ReLU perceptron.

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

<sup>5</sup> 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

## 1. Simplifications

*Attention:* Isotropic and time-independent ( $V(t) = \pm I$ ,  $B(t) = \beta I$  with  $\beta > 0$ ).

*Perceptron:* constant weights ( $W, A \in (\mathbb{R}^{d \times d})^2$  and zero bias  $b = 0$ ).

## 2. Gradient structure

*Unnormalized attention*<sup>5</sup>: remove softmax  $\frac{e^{\beta x_i \cdot x_j}}{\sum_k e^{\beta x_i \cdot x_k}} \mapsto \frac{1}{n} e^{\beta x_i \cdot x_j}$ .

*Conservative perceptron:*  $W\sigma(Ax) = \frac{1}{2}\nabla V_\theta(x)$  for a potential  $V_\theta$  (iff  $W = A^\top \text{diag}(\omega_i)$ ).

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

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

<sup>5</sup>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

## 1. Simplifications

*Attention*: Isotropic and time-independent ( $V(t) = \pm I$ ,  $B(t) = \beta I$  with  $\beta > 0$ ).

*Perceptron*: constant weights  $(W, A) \in (\mathbb{R}^{d \times d})^2$  and zero bias  $b = 0$ .

## 2. Gradient structure

*Unnormalized attention*<sup>5</sup>: remove softmax  $\frac{e^{\beta x_i \cdot x_j}}{\sum_k e^{\beta x_i \cdot x_k}} \mapsto \frac{1}{n} e^{\beta x_i \cdot x_j}$ .

*Conservative perceptron*:  $W\sigma(Ax) = \frac{1}{2}\nabla V_\theta(x)$  for a potential  $V_\theta$  (iff  $W = A^\top \text{diag}(\omega_i)$ ).

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

## 3. Mean-field limit

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

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

<sup>5</sup> 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<sup>6</sup>

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

**First variation:**  $\frac{\delta E}{\delta \mu} : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$  defined by  $\lim_{\varepsilon \rightarrow 0} \frac{E[\mu + \varepsilon \xi] - E[\mu]}{\varepsilon} = \int \frac{\delta E}{\delta \mu}[\mu](x) d\xi(x)$ .

**Wasserstein gradient:**  $\nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu] \in T(\mathbb{S}^{d-1})$ .

$$\frac{d}{dt} E[\mu_t] = \int \frac{\delta E}{\delta \mu}[\mu_t] \partial_t \mu_t = - \int \frac{\delta E}{\delta \mu}[\mu_t] \operatorname{div} \left( \mu_t \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right) = \int \left\| \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right\|^2 d\mu_t \geq 0$$

<sup>6</sup>R. Jordan, D. Kinderlehrer, F. Otto. SIMA (1996). L. Ambrosio, N. Gigli, G. Savaré. Lectures in Mathematics (2008).

# A primer on Wasserstein gradient flows<sup>6</sup>

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

**First variation:**  $\frac{\delta E}{\delta \mu} : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$  defined by  $\lim_{\varepsilon \rightarrow 0} \frac{E[\mu + \varepsilon \xi] - E[\mu]}{\varepsilon} = \int \frac{\delta E}{\delta \mu}[\mu](x) d\xi(x)$ .

**Wasserstein gradient:**  $\nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu] \in T(\mathbb{S}^{d-1})$ .

$$\frac{d}{dt} E[\mu_t] = \int \frac{\delta E}{\delta \mu}[\mu_t] \partial_t \mu_t = - \int \frac{\delta E}{\delta \mu}[\mu_t] \operatorname{div} \left( \mu_t \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right) = \int \left\| \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right\|^2 d\mu_t \geq 0$$

Let our energy and its first variation be defined canonically:

$$E_{\beta, \theta}[\mu] = \frac{1}{2\beta} \iint e^{\beta x \cdot y} d\mu(x) d\mu(y) + \frac{1}{2} \int V_{\theta} d\mu, \quad \frac{\delta E_{\beta, \theta}}{\delta \mu}[\mu](x) = \frac{1}{\beta} \int e^{\beta x \cdot y} d\mu(y) + \frac{1}{2} V_{\theta}(x).$$

<sup>6</sup>R. Jordan, D. Kinderlehrer, F. Otto. SIMA (1996). L. Ambrosio, N. Gigli, G. Savaré. Lectures in Mathematics (2008).

# A primer on Wasserstein gradient flows<sup>6</sup>

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

**First variation:**  $\frac{\delta E}{\delta \mu} : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$  defined by  $\lim_{\varepsilon \rightarrow 0} \frac{E[\mu + \varepsilon \xi] - E[\mu]}{\varepsilon} = \int \frac{\delta E}{\delta \mu}[\mu](x) d\xi(x)$ .

**Wasserstein gradient:**  $\nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu] \in T(\mathbb{S}^{d-1})$ .

$$\frac{d}{dt} E[\mu_t] = \int \frac{\delta E}{\delta \mu}[\mu_t] \partial_t \mu_t = - \int \frac{\delta E}{\delta \mu}[\mu_t] \operatorname{div} \left( \mu_t \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right) = \int \left\| \nabla_{\mathbb{S}} \frac{\delta E}{\delta \mu}[\mu_t] \right\|^2 d\mu_t \geq 0$$

Let our energy and its first variation be defined canonically:

$$E_{\beta, \theta}[\mu] = \frac{1}{2\beta} \iint e^{\beta x \cdot y} d\mu(x) d\mu(y) + \frac{1}{2} \int V_{\theta} d\mu, \quad \frac{\delta E_{\beta, \theta}}{\delta \mu}[\mu](x) = \frac{1}{\beta} \int e^{\beta x \cdot y} d\mu(y) + \frac{1}{2} V_{\theta}(x).$$

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

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

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

<sup>6</sup> R. Jordan, D. Kinderlehrer, F. Otto. SIMA (1996). L. Ambrosio, N. Gigli, G. Savaré. Lectures in Mathematics (2008).

# State of the art

For **pure self-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).

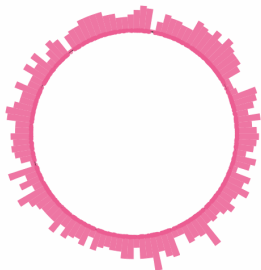


# State of the art

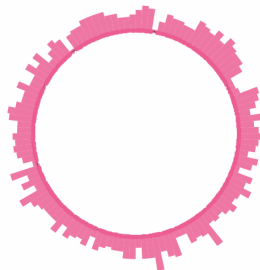
For **pure self-attention**, the asymptotic picture is well understood:

**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.

# State of the art

For **pure self-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 self-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  $\text{spec}(V)$ ) 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 self-attention.
- Castin, Ablin, Carrillo, Peyré. arXiv (2025).  
Mean-field limit for multi-head, L2, 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).  
 $W_2$  concentration bounds for mean-field transformers as  $1/\beta \rightarrow 0+$ .
- 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.

# Our problem

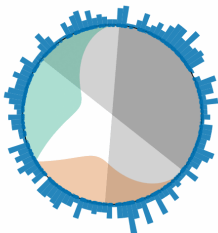
**Main question.**

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

# Our problem

## Main question.

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



Radial profile:  $V_\theta$

■  $V_\theta(x) > 0$

■  $V_\theta(x) < 0$

■  $(Ax)_i \leq 0$



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

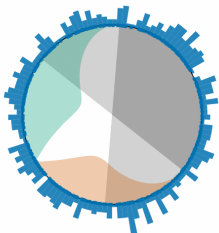


**GeLU** perceptron in  $\mathbb{S}^1$   
+ gradient ascent.

# Our problem

## Main question.

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



Radial profile:  $V_\theta$

- $V_\theta(x) > 0$
- $V_\theta(x) < 0$
- $(Ax)_i \leq 0$



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



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

# Equilibrium points

**Stationary measures:**

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

# Equilibrium points

**Stationary measures:**

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

# Equilibrium points

**Stationary measures:**

$$\nabla_{\mathbb{S}} \frac{\delta E_{\beta, \theta}}{\delta \mu} [\mu] = 0 \quad \text{on } \text{supp}(\mu).$$



# Equilibrium points

## Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta E_{\beta, \theta}}{\delta \mu} [\mu] = 0 \quad \text{on } \text{supp}(\mu).$$

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

$$\text{Hess}_{\mu} E_{\beta, \theta}(\xi, \xi) \geq 0 \quad \forall \xi \in T_{\mu} \mathcal{P}(\mathbb{S}^{d-1}).$$

$$\text{Hess}_{\mu} E_{\beta, \theta}(\xi, \xi) := \left. \frac{d^2}{dt^2} \right|_{t=0} E(\mu(t)), \text{ with } \mu(t) \text{ the } W_2\text{-geodesic along } \xi \in T_{\mu} \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla \phi : \phi \in C^{\infty}(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$$

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

# Equilibrium points

## Stationary measures:

$$\nabla_{\mathbb{S}} \frac{\delta E_{\beta, \theta}}{\delta \mu} [\mu] = 0 \quad \text{on } \text{supp}(\mu).$$

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

$$\text{Hess}_{\mu} E_{\beta, \theta}(\xi, \xi) \geq 0 \quad \forall \xi \in T_{\mu} \mathcal{P}(\mathbb{S}^{d-1}).$$

$$\text{Hess}_{\mu} E_{\beta, \theta}(\xi, \xi) := \left. \frac{d^2}{dt^2} \right|_{t=0} E(\mu(t)), \text{ with } \mu(t) \text{ the } W_2\text{-geodesic along } \xi \in T_{\mu} \mathcal{P}(\mathbb{S}^{d-1}) := \overline{\{\nabla \phi : \phi \in C^{\infty}(\mathbb{S}^{d-1})\}}^{L^2(\mu)}$$

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

## Why do we need both notions?

**Stationarity** yields enough rigidity when  $V_{\theta}$  is piecewise polynomial (e.g., for  $\sigma = \text{ReLU}$ ).

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

# Toolbox

## 1. 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  $f : M \rightarrow \mathbb{R}$  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  $M$ ”.

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

Let  $G : M \times \Lambda \rightarrow TM$  be a smooth family of vector fields on a manifold  $M$ . Assume that, at every  $(x, \lambda)$  such that  $G_\lambda(x) = 0$ ,

$$\operatorname{Im} D_x G + \operatorname{Im} D_\lambda G = T_x M.$$

Then, for a.e.  $\lambda$ , every zero of  $G_\lambda$  is nondegenerate (i.e.,  $D_x G_\lambda(x)$  is invertible).

## 3. Spherical convolution lemma

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

(ii) Moreover,  $f^\mu$  is even if and only if  $\mu(A) = \mu(-A)$  for all Borel  $A \subseteq \mathbb{S}^{d-1}$ .

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

# Atomicity on the circle (ReLU)

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

## Theorem (ReLU activation)

*Let  $d = 2$ ,  $\beta > 0$ ,  $\sigma(s) = s_+$  and assume that  $V_\theta$  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 (ReLU activation)

*Let  $d = 2$ ,  $\beta > 0$ ,  $\sigma(s) = s_+$  and assume that  $V_\theta$  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  $\nabla_s \frac{\delta E_{\beta, \theta}}{\delta \mu} [\mu]$  is real-analytic on each cell  $I \subset \mathbb{S}^1$  where  $\text{sign}(Ax)$  is constant.
2. By the **identity theorem**, if  $\text{supp}(\mu)$  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(i)**,  $\mu$  has a polynomial density, implying full support ( $\implies \text{supp}(\mu) = \mathbb{S}^1$ ).
5. Full support enforces stationarity everywhere. Comparing the global quadratic identities obtained from adjacent activation cells forces  $V_\theta$  to be globally real-analytic  $\implies$  *contradiction*.

# Atomicity on the circle (GeLU)

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

## Theorem (GeLU activation)

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

### *Proof sketch.*

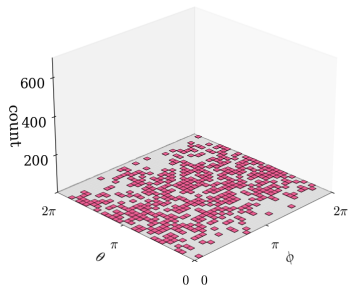
1. Again by the **identity theorem**,  $\text{supp}(\mu)$  being infinite would give  $\nabla_{\mathbb{S}} \frac{\delta E_{\beta,\theta}}{\delta \mu} [\mu] = 0$  on  $\mathbb{S}^1$ , collapsing the Hessian to a bilinear form:

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

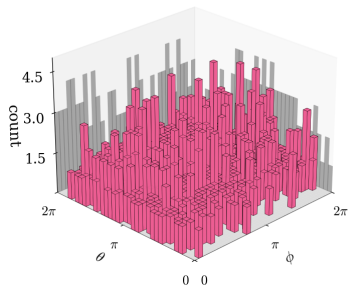
2. Around any accumulation point, we build unit-norm, zero-mean tangent vectors  $\xi_{\delta}$  of support size  $\delta$ .
3. Since  $\int \xi_{\delta} d\mu = 0$ , we can center the kernel ( $K''_{\beta} \rightarrow K''_{\beta} - K''_{\beta}(0)$ ). The energy cost then vanishes as the support shrinks,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

$$|\text{Hess}_{\mu} E_{\beta,\theta}(\xi_{\delta}, \xi_{\delta})| \lesssim \frac{1}{\beta} \sup_{|h| \leq \delta} |K''_{\beta}(h) - K''_{\beta}(0)| \xrightarrow{\delta \rightarrow 0} 0$$

# What happens if $d \geq 3$ ?



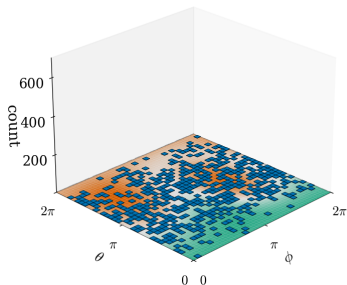
Pure attention in  $\mathbb{S}^2$   
+ gradient ascent.



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



# What happens if $d \geq 3$ ?

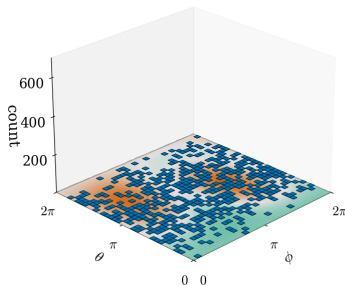


Surface profile:  $V_\theta$

- $V_\theta(x) > 0$
- $V_\theta(x) < 0$
- $(Ax)_i \leq 0$

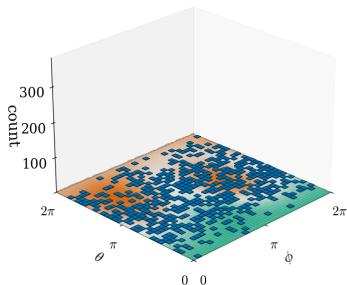


**ReLU** perceptron in  $\mathbb{S}^2$   
+ gradient ascent.



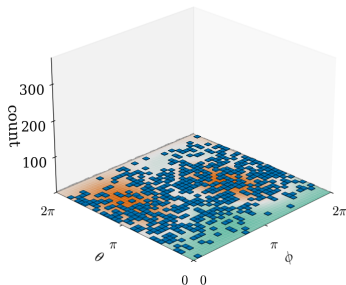
**GeLU** perceptron in  $\mathbb{S}^2$   
+ gradient ascent.

# What happens if $d \geq 3$ ?



Surface profile:  $V_\theta$

- $V_\theta(x) > 0$
- $V_\theta(x) < 0$
- $(Ax)_i \leq 0$



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



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

# What happens if $d \geq 3$ ?

## Theorem

Let  $d \geq 2$ . Suppose that  $\mu$  is either stationary for  $\sigma$  the ReLU and  $V_\theta$  is not real-analytic, or that  $\mu$  is a strict SOPD critical point for  $\sigma$  the GeLU. Then:

- (a) *Singular support:*  $\text{supp}(\mu)$  has zero surface measure.
- (b) *Generic atomicity:* For a dense set of  $(\beta, \theta)$ ,  $\mu$  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

Let  $d \geq 2$ . Suppose that  $\mu$  is either stationary for  $\sigma$  the ReLU and  $V_\theta$  is not real-analytic, or that  $\mu$  is a strict SOPD critical point for  $\sigma$  the GeLU. Then:

- (a) **Singular support:**  $\text{supp}(\mu)$  has zero surface measure.
- (b) **Generic atomicity:** For a dense set of  $(\beta, \theta)$ ,  $\mu$  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  $g_{\beta, \theta}(x) := \nabla_S \frac{\delta E_{\beta, \theta}}{\delta \mu} [\mu](x)$ . To apply **parametric transversality**, we perturb any  $a_j$  along  $h \in T_x \mathbb{S}^{d-1}$ :

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

This is surjective whenever there is an active index ( $\omega_j \sigma(a_j \cdot x) \neq 0$ ). 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

## Stationary measures:

### Interacting particle ODE:

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

### Mean-field limit:

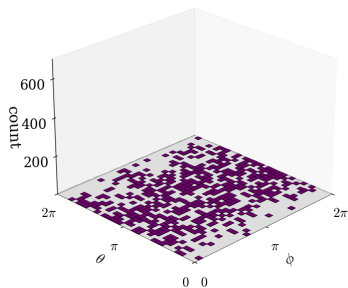
$$\partial_t \mu_t + \operatorname{div} \left( \mu_t \mathbf{P}_x^\perp \left[ \pm \frac{1}{\beta} \frac{\nabla_{\mathbb{S}} \int e^{\beta x \cdot y} d\mu_t(y)}{\int e^{\beta x \cdot y} d\mu_t(y)} + \frac{1}{2} \nabla_{\mathbb{S}} V_\theta(x) \right] \right) = 0$$

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

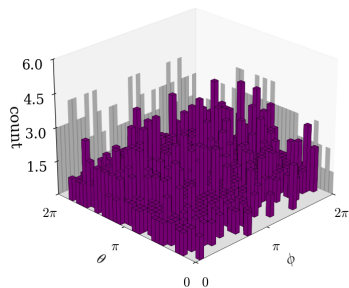
### Stationarity condition:

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

# Softmax attention

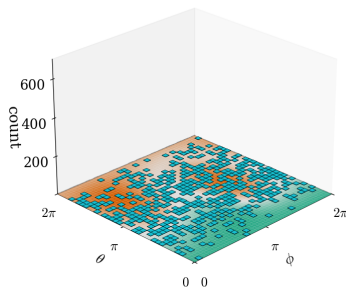


Pure attention in  $\mathbb{S}^2$   
+ gradient ascent.



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

# Softmax attention

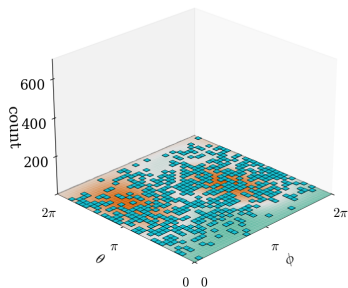


Surface profile:  $V_\theta$

- $V_\theta(x) > 0$
- $V_\theta(x) < 0$
- $(Ax)_i \leq 0$

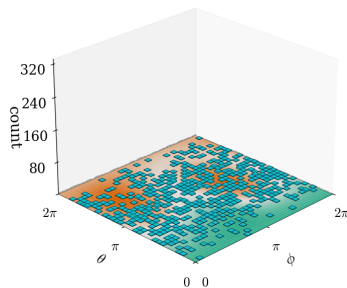


**ReLU** perceptron in  $\mathbb{S}^2$   
+ gradient ascent.



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

# Softmax attention

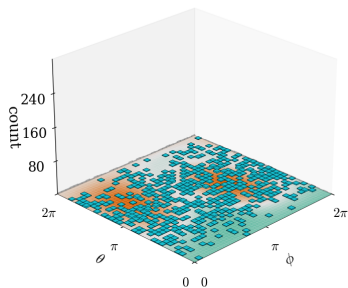


Surface profile:  $V_\theta$

- $V_\theta(x) > 0$
- $V_\theta(x) < 0$
- $(Ax)_i \leq 0$



**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  $J \subseteq \{1, \dots, d\}$ , 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  $d = 2$ ,  $\mu$  is purely atomic with finitely many atoms.
- (b) **Singular support:** For all  $d \geq 2$ ,  $\text{supp}(\mu)$  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  $J \subseteq \{1, \dots, d\}$ , 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  $d = 2$ ,  $\mu$  is purely atomic with finitely many atoms.
- (b) **Singular support:** For all  $d \geq 2$ ,  $\text{supp}(\mu)$  has zero surface measure.

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

1. Let  $f^\mu(x) := \int e^{\beta x \cdot y} d\mu(y) > 0$ . If  $\text{supp}(\mu)$  is infinite on a cell  $I$  with index set  $J$ , the identity theorem yields a global quadratic equality for the logarithm:

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

2. Since the RHS is even,  $f^\mu$  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  $x^\top M_J x \equiv \text{constant}$  on the sphere, forcing  $M_J = cI \implies \text{contradiction}$ .

# A toy model: Attention vs. Perceptron

## The 0-dimensional sphere

Let  $\mathbb{S}^0 = \{-1, +1\}$  and  $\mu = \varepsilon\delta_{+1} + (1 - \varepsilon)\delta_{-1}$ .

*Which force dominates? It depends on the ratio  $\theta/\beta$ .*

The minimizer can be computed explicitly:

► **Launch interactive simulation** ◀

# Quantitative anti-concentration

## Theorem

Let  $d = 2$ ,  $\text{Lip}(\sigma) = 1$  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  $S \subseteq \{1, \dots, N\}$  with  $|S| \geq 2$  satisfies

$$\max_{i,j \in S} \text{dist}_S(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

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

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

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

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

**Kernel bound:**  $e^{\beta \cos \theta}$  is strictly concave for  $|\theta| \lesssim \beta^{-1/2}$ . Thus,  $\mathcal{T}_{\text{in}} < -e^{\beta-1/8} \sum_{i \in S} m_i$ .

**Inter-cluster:**  $\mathcal{T}_{\text{out}} > 0$  but bounded by  $e^{\beta-3/2} \sum_{k \notin S} m_k$ .

**Perceptron:**  $\mathcal{T}_{\theta} > 0$  and bounded using the Lipschitz constant of  $\sigma$ .

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

# Quantitative anti-concentration

## Theorem

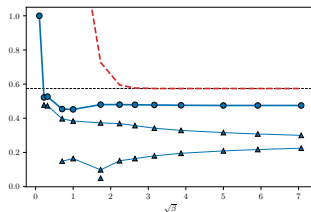
Let  $d = 2$ ,  $\text{Lip}(\sigma) = 1$  with  $\sigma(0) = 0$ . Let  $\mu = \sum_{i=1}^N m_i \delta_{\varphi_i}$  be a SOPD critical point with  $N \geq 2$ . If a sub-cluster  $\mathcal{S} \subseteq \{1, \dots, N\}$  with  $|\mathcal{S}| \geq 2$  satisfies

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

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

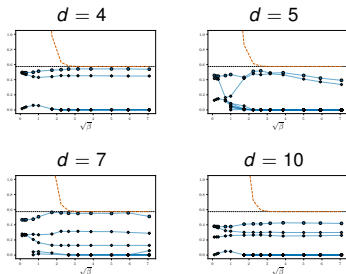
Let  $d = 2$ ,  $\text{Lip}(\sigma) = 1$  with  $\sigma(0) = 0$ . Let  $\mu = \sum_{i=1}^N m_i \delta_{\varphi_i}$  be a SOPD critical point with  $N \geq 2$ . If a sub-cluster  $S \subseteq \{1, \dots, N\}$  with  $|S| \geq 2$  satisfies

$$\max_{i,j \in S} \text{dist}_S(\varphi_i, \varphi_j) \leq \frac{1}{2\sqrt{\beta}},$$

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

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

*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:

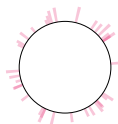
- (A) Stationary measures are **singular** and *generically atomic* (in  $d = 2$ , finitely supported).
- (B) In descent, atomicity  $\neq$  trivial collapse: we get quantitative **anti-concentration bounds**.
- (C) **Localization**: the perceptron *anchors* clusters to specific, seed-independent locations.

# Conclusions

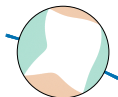
A (conservative) perceptron *fundamentally reshapes* the energy landscape:

- (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