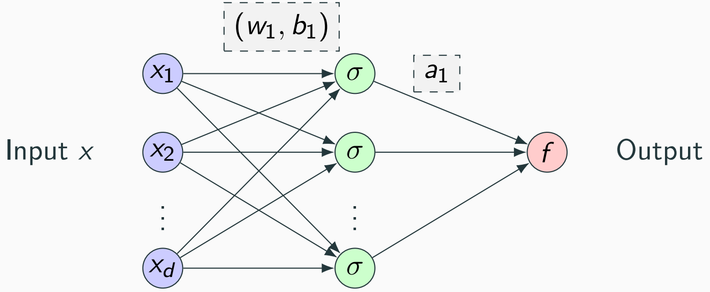
Mean-Field Neural Networks and Transformers

September 17, 2025

Mean-field Neural Networks

Two-Layer Neural Network: Definition

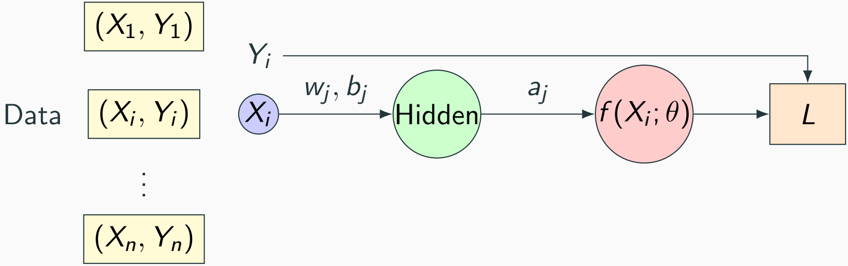


The output $f(x;\theta)$ is given by:

$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle + b_j)$$

- $\theta = \{(a_j, w_j, b_j)\}_{j=1}^{m}$ are the parameters.
- m is the number of neurons.

The Optimization Problem: Non-Convexity

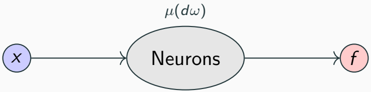


Loss function (e.g., squared loss): $L(\theta) = \sum_{i=1}^{n} (Y_i - f(X_i; \theta))^2$

$$L(\theta) = \sum_{i=1}^{n} \left(Y_i - \frac{1}{m} \sum_{j=1}^{m} a_j \sigma(\langle w_j, X_i \rangle + b_j) \right)^2$$

This loss function $L(\theta)$ is highly **non-convex**.

Lifting to Measure Space



Let $\mu = \frac{1}{m}\sum_{j=1}^{m} \delta_{(a_j, w_j, b_j)}$.

Network output becomes an integral:

$$f(x;\mu) = \int_\Omega a\sigma(\langle w, x\rangle + b)\,\mu(\mathrm{d}\omega) = \int_\Omega \rho(x;\omega)\,\mu(\mathrm{d}\omega)$$

The loss becomes a functional on the space of measures:
$L : \mathcal{P}_2(\Omega) \to \mathbb{R}$

$$L(\mu) = \sum_{i=1}^{n} \left(Y_i - \int_\Omega \rho(X_i;\omega)\mu(\mathrm{d}\omega) \right)^2$$

Gradient Flows: Euclidean vs. Wasserstein

Correspondence Proposition

The Wasserstein gradient flow $(\mu_t)_{t\geq 0}$ of $L(\mu)$, when initialised at an empirical measure $\mu_0 = \mu_\theta$ (where $\theta = \{(a_j, w_j, b_j)\}_{j=1}^m$ and $\mu_\theta = \frac{1}{m}\sum_{j=1}^m \delta_{(a_j, w_j, b_j)}$), satisfies $\mu_t = \mu_{\theta_t}$ for all $t \geq 0$. Here $(\theta_t)_{t\geq 0}$ is the (time-rescaled) Euclidean gradient flow of $L(\theta)$ initialised at θ.

State Represents: θ_t	State Represents: $\mu_t = \mu_{\theta_t}$
Loss Function: $L(\theta_t)$	Loss Functional: $L(\mu_t) = L(\mu_{\theta_t})$
Optimization Dynamics: Euclidean Gradient Flow $\dot{\theta}_t = -\nabla_\theta L(\theta_t)$	Optimization Dynamics: Wasserstein Gradient Flow (PDE) $\partial_t \mu_t = \nabla \cdot \left(\mu_t \nabla \frac{\delta L}{\delta \mu_t}\right)$

Geodesic Non-Convexity Remains!

The lifted loss functional $L(\mu)$:

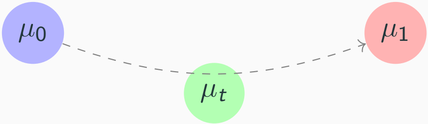
$$L(\mu) = \sum_{i=1}^{n} \left(Y_i - \int_\Omega \rho(x;\omega)\mu(\mathrm{d}\omega) \right)^2$$

- $L(\mu)$ *is* "convex" for linear combinations $\lambda\mu_1 + (1-\lambda)\mu_2$.
- But convex combinations (geodesics) in Wasserstein space $\mathcal{P}_2(\Omega)$ are defined via optimal transport, not linear mixing of measures.
- $L(\mu)$ is generally **not** geodesically convex.

Geodesic Non-Convexity Remains!

Wasserstein Geodesics vs. Linear Interpolation:

Particle Path $X_t = (1-t)X_0 + tX_1$



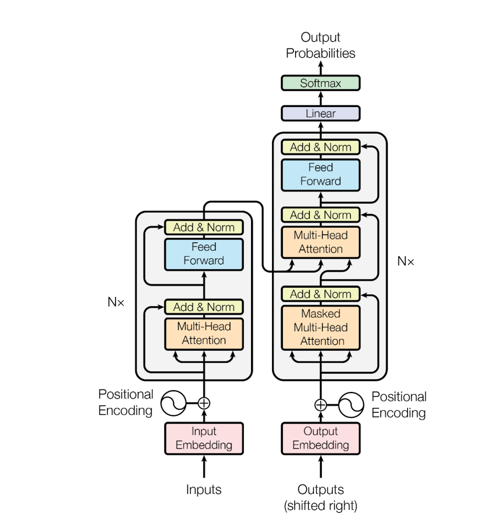
Advantages of the Measure Perspective

Allows considering measures other than the empirical $\frac{1}{m}\sum_{i=1}^{m}\delta_{\omega_j}$ and provides analytical benefits:

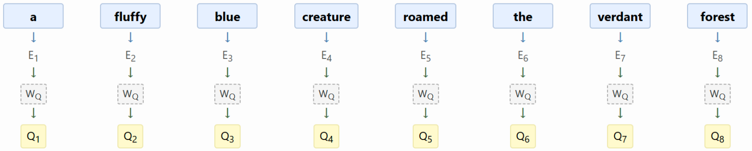
- **Chizat & Bach '18, "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport"**
- **Mei et al. '18, "A Mean Field View of the Landscape of Two-Layers Neural Networks"**

Transformers

Transformer Architecture



Self-Attention Example: QKV Calculation



	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
a → E_1 → W_k → K_1	$K_1 \cdot Q_1$	$K_1 \cdot Q_2$	$K_1 \cdot Q_3$	$K_1 \cdot Q_4$	$K_1 \cdot Q_5$	$K_1 \cdot Q_6$	$K_1 \cdot Q_7$	$K_1 \cdot Q_8$
fluffy → E_2 → W_k → K_2	$K_2 \cdot Q_1$	$K_2 \cdot Q_2$	$K_2 \cdot Q_3$	$K_2 \cdot Q_4$	$K_2 \cdot Q_5$	$K_2 \cdot Q_6$	$K_2 \cdot Q_7$	$K_2 \cdot Q_8$
blue → E_3 → W_k → K_3	$K_3 \cdot Q_1$	$K_3 \cdot Q_2$	$K_3 \cdot Q_3$	$K_3 \cdot Q_4$	$K_3 \cdot Q_5$	$K_3 \cdot Q_6$	$K_3 \cdot Q_7$	$K_3 \cdot Q_8$
creature → E_4 → W_k → K_4	$K_4 \cdot Q_1$	$K_4 \cdot Q_2$	$K_4 \cdot Q_3$	$K_4 \cdot Q_4$	$K_4 \cdot Q_5$	$K_4 \cdot Q_6$	$K_4 \cdot Q_7$	$K_4 \cdot Q_8$
roamed → E_5 → W_k → K_5	$K_5 \cdot Q_1$	$K_5 \cdot Q_2$	$K_5 \cdot Q_3$	$K_5 \cdot Q_4$	$K_5 \cdot Q_5$	$K_5 \cdot Q_6$	$K_5 \cdot Q_7$	$K_5 \cdot Q_8$
the → E_6 → W_k → K_6	$K_6 \cdot Q_1$	$K_6 \cdot Q_2$	$K_6 \cdot Q_3$	$K_6 \cdot Q_4$	$K_6 \cdot Q_5$	$K_6 \cdot Q_6$	$K_6 \cdot Q_7$	$K_6 \cdot Q_8$
verdant → E_7 → W_k → K_7	$K_7 \cdot Q_1$	$K_7 \cdot Q_2$	$K_7 \cdot Q_3$	$K_7 \cdot Q_4$	$K_7 \cdot Q_5$	$K_7 \cdot Q_6$	$K_7 \cdot Q_7$	$K_7 \cdot Q_8$
forest → E_8 → W_k → K_8	$K_8 \cdot Q_1$	$K_8 \cdot Q_2$	$K_8 \cdot Q_3$	$K_8 \cdot Q_4$	$K_8 \cdot Q_5$	$K_8 \cdot Q_6$	$K_8 \cdot Q_7$	$K_8 \cdot Q_8$

Self-Attention Example: Scores and Normalization

Unnormalized Attention Pattern

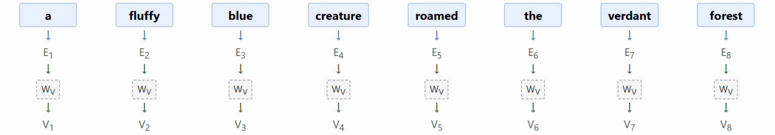
+3.53	+0.80	+1.96	+4.48	+3.74	-1.95
-∞	-0.30	-0.21	+0.82	+0.29	+2.91
-∞	+0.89	+0.67	+2.99	-0.41	
-∞	-∞	+1.31	+1.73	-1.48	
-∞	-∞	-∞	+3.07	+2.94	
-∞	-∞	-∞	-∞	+0.31	

→ softmax

Normalized Attention Pattern

1.00	0.75	0.69	0.92	0.46	0.00
0.00	0.25	0.08	0.02	0.01	0.46
0.00	0.00	0.24	0.02	0.22	0.02
0.00	0.00	0.00	0.04	0.06	0.01
0.00	0.00	0.00	0.00	0.24	0.48
0.00	0.00	0.00	0.00	0.00	0.03

Self-Attention Example: Weighted Value Sum



a → E_1 → W_v → V_1	1.00 V_1	0.00 V_1	0.00 V_1	0.00 V_1	0.00 V_1	0.00 V_1	0.00 V_1	0.00 V_1
fluffy → E_2 → W_v → V_2	0.00 V_2	1.00 V_2	0.00 V_2	0.42 V_2	0.00 V_2	0.00 V_2	0.00 V_2	0.00 V_2
blue → E_3 → W_v → V_3	0.00 V_3	0.00 V_3	1.00 V_3	0.58 V_3	0.00 V_3	0.00 V_3	0.00 V_3	0.00 V_3
creature → E_4 → W_v → V_4	0.00 V_4	0.00 V_4	0.00 V_4	0.00 V_4	0.00 V_4	0.00 V_4	0.00 V_4	0.00 V_4
roamed → E_5 → W_v → V_5	0.00 V_5	0.00 V_5	0.00 V_5	0.00 V_5	0.00 V_5	0.01 V_5	0.00 V_5	0.00 V_5
the → E_6 → W_v → V_6	0.00 V_6	0.00 V_6	0.00 V_6	0.00 V_6	0.99 V_6	1.00 V_6	0.00 V_6	0.00 V_6
verdant → E_7 → W_v → V_7	0.00 V_7	0.00 V_7	0.00 V_7	0.00 V_7	0.00 V_7	0.00 V_7	1.00 V_7	0.00 V_7
forest → E_8 → W_v → V_8	0.00 V_8	0.00 V_8	0.00 V_8	0.00 V_8	0.00 V_8	0.00 V_8	0.00 V_8	1.00 V_8
	Σ	Σ	Σ	Σ	Σ	Σ	Σ	Σ
	↓	↓	↓	↓	↓	↓	↓	↓
	ΔE_1	**ΔE_2**	**ΔE_3**	**ΔE_4**	**ΔE_5**	**ΔE_6**	**ΔE_7**	**ΔE_8**

Self-Attention as Dynamics (1/4): Iterative Scheme

Consider applying the attention update iteratively, like residual connections in deep networks:

$$x_{t+1}^i = x_t^i + \left(V \frac{\sum_{j=1}^{N} x_t^j e^{\langle Qx_t^i, Kx_t^j \rangle}}{\sum_{l=1}^{N} e^{\langle Qx_t^i, Kx_t^l \rangle}} \right)$$

Self-Attention as Dynamics (2/4): Continuous ODE

Consider applying the attention update iteratively, like residual connections in deep networks:

$$\dot{x}_t^i = V \frac{\sum_{j=1}^{N} x_t^j e^{\langle Q x_t^i, K x_t^j \rangle}}{\sum_{l=1}^{N} e^{\langle Q x_t^i, K x_t^l \rangle}}$$

Self-Attention as Dynamics (3/4): Mean-Field Limit

Consider applying the attention update iteratively, like residual connections in deep networks:

$$\dot{x}_t^i = V \frac{\int y e^{\langle Q x_t^i, K y \rangle} \mu_t(\mathrm{d}y)}{\int e^{\langle Q x_t^i, K y \rangle} \mu_t(\mathrm{d}y)}$$

Assume $Q = K = V = I$ for simplicity:

$$\dot{x}_t^i = \frac{\int y e^{\langle x_t^i, y \rangle} \mu_t(\mathrm{d}y)}{\int e^{\langle x_t^i, y \rangle} \mu_t(\mathrm{d}y)} = \nabla_x \left[\log \int e^{\langle x, y \rangle} \mu_t(\mathrm{d}y) \right]_{x = x_t^i}$$

Self-Attention as Dynamics (4/4): WGF Connection?

Recall: Particle Interpretation of WGF:

$$\dot{X}_t = -\nabla \frac{\delta \mathcal{F}}{\delta \mu}(X_t)$$

From the previous slide (with $Q = K = V = I$):

$$\dot{x}_t^i = \nabla \underbrace{[\log \int e^{\langle x,y \rangle} \mu_t(\mathrm{d}y)]}_{\Psi(x;\mu_t)}(x_t^i)$$

Question

$\Psi(x; \mu_t) \stackrel{?}{=} -\frac{\delta \mathcal{F}}{\delta \mu_t}(x)$ for some energy functional $\mathcal{F}$ describing the self-attention dynamics?

Self-Attention as Dynamics: WGF Connection? (Cont.)

No!

- **Reference:** Sander et al. '22, "Sinkformers: Transformers with Doubly Stochastic Attention".
- **Issue:** Asymmetry.

Unnormalised Self-Attention (1/2)

Remedy: What if we remove the denominator (the Softmax normalization)?

Consider the unnormalised dynamics (assuming $Q = K = V = I$ for now):

$$\dot{x}_t^i = \int y e^{\langle x_t^i, y \rangle} \mu_t(\mathrm{d}y) = \nabla_x \left[\int e^{\langle x, y \rangle} \mu_t(\mathrm{d}y) \right]_{x = x_t^i}$$

Unnormalised Self-Attention (2/2)

Remedy: What if we remove the denominator (the Softmax normalization)?

Consider the unnormalised dynamics (assuming $Q = K = V = I$ for now):

$$\dot{x}_t^i = \int y e^{\langle x_t^i, y\rangle} \mu_t(dy) = \nabla_x \left[\underbrace{\int e^{\langle x, y\rangle} \mu_t(dy)}_{\delta \mathcal{F}(\mu_t)} \right]_{x = x_t^i}$$

where

$$\mathcal{F}(\mu) = - \int \int e^{\langle x, y\rangle} \mu(dx) \mu(dy)$$

Unnormalised Attention as WGF

Proposition
The unnormalised self-attention dynamics (with $Q = K = V = I$, after time rescaling):

$$\dot{x}_t^i = \int y e^{\langle x_t^i, y \rangle} \mu_t(\mathrm{d}y)$$

is the Wasserstein gradient flow of the interaction energy:

$$\mathcal{F}(\mu) = -\int \int e^{\langle x, y \rangle} \mu(\mathrm{d}x) \mu(\mathrm{d}y)$$

defined on $\mathcal{P}_2(\mathbb{R}^d)$.

Implication: Divergence

$\mathcal{F}(\delta_z) = -e^{\langle z, z \rangle} = -e^{\|z\|^2}$. As $\|z\| \to \infty$, $\mathcal{F}(\delta_z) \to -\infty$.

Adding Layer Normalisation

$$\begin{tikzcd} x_t^i \arrow[d] \\ \text{Unnorm.Attention} \arrow[d] \\ \boxed{\text{Layer Norm}} \end{tikzcd}$$

Corresponds to the dynamics

$$\dot{x}_t^i = P_{x_t^i}\left(\int y e^{\langle x_t^i, y\rangle} \mu_t(\mathrm{d}y)\right)$$

where P_x is the projection onto $T_x(S^{d-1})$

Layer-Normalised Unnormalised Attention as WGF

Proposition

The layer-normalised unnormalised self-attention dynamics:

$$\dot{x}_t^i = P_{x_t^i}\left(\int y e^{\langle x_t^i, y\rangle} \mu_t(\mathrm{d}y)\right)$$

is the Wasserstein gradient flow of the same interaction energy:

$$\mathcal{F}(\mu) = -\int\int e^{\langle x, y\rangle} \mu(\mathrm{d}x)\mu(\mathrm{d}y)$$

but now considered on the space of probability measures on the sphere, $\mathcal{P}_2(S^{d-1})$.

Implication: Concentration

Minimizers are Dirac masses δ_z for $z \in S^{d-1}$

The dynamics tend to form a single cluster on the sphere.

Clustering and Convergence

The interaction energy on the sphere:

$$\mathcal{F}(\mu) = -\int_{S^{d-1}}\int_{S^{d-1}} e^{\langle x,y\rangle}\mu(\mathrm{d}x)\mu(\mathrm{d}y)$$

- This functional $\mathcal{F}(\mu)$ is **not** geodesically convex on $\mathcal{P}_2(S^{d-1})$.
- It admits many stationary points where the Wasserstein gradient vanishes

Clustering and Convergence

The interaction energy on the sphere:

$$\mathcal{F}(\mu) = -\int_{S^{d-1}} \int_{S^{d-1}} e^{\langle x,y \rangle} \mu(\mathrm{d}x)\mu(\mathrm{d}y)$$

- **Geshkovski et al. '24, "A Mathematical Perspective on Transformers":** These points are in fact saddle points, guaranteeing asymptotic convergence to a single cluster when dynamics are initialized in a generic position.