Memory as a Stationary Measure in the Mean-Field Limit of Neural Networks

A Measure-Theoretic Formulation of Memory

Structure of the Talk

- ⟶ Motivation: Why redefine memory
- ⟶ Neural networks as distributions
- ⟶ Training as a PDE in measure space
- ⟶ Memory as a stationary measure

Motivation

⟶ Traditional memory definitions: weights, activations, hidden states

⟶ These are heuristic and architecture-dependent

⟶ Core question: Can we define memory purely mathematically?

⟶ “Memory is not a stored value, but a stationary structure in distribution space.”

Neural Networks as Measures

$$f^{(n)}(x) = \frac{1}{n} \sum_{i=1}^{n} a_i \sigma(w_i \cdot x + b_i)$$

Empirical measure over neurons:

$$\mu^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \delta_{(a_i, w_i, b_i)} \in \mathcal{P}(\mathbb{R}^{2d+1})$$

As $n \to \infty$, $\mu^{(n)} \xrightarrow{w} \mu$ (weak convergence)

Mean-Field Limit

Limit behavior:

$$f_\mu(x) = \int a\,\sigma(w \cdot x + b)\, d\mu(a, w, b)$$

⟶ The network is now described by μ, not finite parameters

Gradient Descent Becomes a PDE

Let the expected loss be:

$$\longrightarrow \quad \mathcal{L}(\mu) = \mathbb{E}_{(x,y)}\left[\ell(f_\mu(x), y)\right]$$

Gradient descent in parameter space becomes a distributional PDE:

$$\longrightarrow \quad \partial_t \mu_t + \nabla \cdot (\mu_t V[\mu_t]) = 0$$

Where the velocity field is:

$$\longrightarrow \quad V[\mu_t](a, w, b) = -\nabla_{(a,w,b)} \left(\frac{\delta \mathcal{L}}{\delta \mu_t} \right)$$

Wasserstein Space

Wasserstein-2 distance between two distributions:

$$W_2^2(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^2 \, d\gamma(x, y)$$

Γ(μ,ν): couplings with marginals μ,ν

Training becomes a gradient flow of $\mathcal{L}$ in $\mathcal{P}_2(\mathbb{R}^{d+2})$

Training as Gradient Flow

The continuity equation describes the flow of μ_t:

$$\partial_t \mu_t = -\nabla_{\mathcal{W}} \mathcal{L}(\mu_t)$$

Wasserstein gradient flows minimize $\mathcal{L}(\mu)$ over time:

$$\frac{d}{dt}\mathcal{L}(\mu_t) = -\|\nabla_{\mathcal{W}} \mathcal{L}(\mu_t)\|^2 \leq 0$$

Defining Memory

Memory = stationary distribution:

$$\partial_t \mu_t = 0 \Rightarrow \mu_t = \mu^*$$

Equivalently, memory measure minimizes loss:

$$\mu^* = \arg\min_{\mu \in \mathcal{P}_2} \mathcal{L}(\mu)$$

Why the Definition Works

Memory is a fixed point in Wasserstein space

⟶ The definition is:
⟶ Independent of network architecture
⟶ Dynamically stable under training
⟶ Compatible with variational and PDE analysis

THANK YOU/ ANY QUESTIONS?