A Measure-Theoretic Definition of Memory in Neural Networks via Wasserstein Gradient Flows

Nujen Yuksel

Abstract

This paper introduces a rigorous, measure-theoretic formulation of memory in neural networks, proposing that memory corresponds to a stationary distribution in the space of probability measures over network parameters. Rather than treating memory as a static value or architectural feature, we define it as a fixed point under the gradient flow in Wasserstein space. We develop the mathematical underpinnings for this theory by drawing from optimal transport, measure theory, and partial differential equations. Our results establish the existence and uniqueness of such stationary measures and offer a new framework to understand learning dynamics in infinite-width networks.

Introduction and Motivation

The concept of memory in neural networks is traditionally understood through heuristic or architectural terms—weights, activations, and hidden states are all regarded as vessels through which networks "remember" data. However, these interpretations depend on finite-dimensional representations and specific architectural constraints, making them fragile across changing model scales or designs. In this paper, we seek a more fundamental definition: a mathematically intrinsic characterization of memory that is independent of implementation details.

We posit that memory is best understood not as a stored value or a functional state, but as a stationary structure within a dynamical system—specifically, as a stationary probability measure in the space of distributions over network parameters, evolving under the learning dynamics of gradient descent. This reframing is motivated by the growing body of literature on the mean-field limit of neural networks, where the empirical distribution of parameters, rather than the parameters themselves, becomes the object of study. In the infinite-width limit, the learning process of a neural network can be described by a partial differential equation (PDE) in the space of probability measures, typically endowed with the Wasserstein metric from optimal transport theory.

Let us illustrate this with a simple example. Consider a one-hidden-layer neural network with n neurons, where each neuron's parameters are denoted $\theta_i \in \mathbb{R}^d$. The output of the network for input $x \in \mathbb{R}^p$ is:

$$f_n(x; \theta_1, \dots, \theta_n) = \frac{1}{n}\sum_{i=1}^{n} \sigma(x; \theta_i)$$

where $\sigma(x;\theta_i)$ is the activation function applied to input x with parameters θ_i. We define the empirical measure $\mu^{(n)}$ associated with the network as:

$$\mu^{(n)} := \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$$

where δ_θ is the Dirac delta measure centered at θ. As $n \to \infty$, under mild regularity conditions, this empirical measure converges weakly to a deterministic measure μ on $\mathbb{R}^d$. Thus, the neural network is no longer described by a finite tuple of parameters, but by a distribution μ, which becomes the fundamental object in the mean-field framework.

This convergence allows us to reinterpret training as a dynamical flow of probability measures**.** The key insight is that gradient descent on the loss function induces a Wasserstein gradient flow in the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second moments. The dynamics of training can then be formulated as a continuity equation:

$$\partial_t \mu_t + \nabla \cdot (v_t \mu_t) = 0$$

where μ_t is the time-dependent measure of parameters and v_t is a velocity field derived from the functional gradient of the loss.

To proceed rigorously, we begin by defining the empirical measure over network parameters:

Definition (Empirical Measure of Network Parameters)

Let $\{\theta_i\}_{i=1}^n \subset \mathbb{R}^d$ be the parameters of a neural network with n units. The empirical measure associated with these parameters is defined as:

$$\mu^{(n)} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$$

where δ_θ denotes the Dirac measure centered at $\theta_i \in \mathbb{R}^d$.

As $n \to \infty$, under appropriate regularity conditions, $\mu^{(n)} \rightharpoonup \mu$ (weak convergence), where $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is a probability measure with finite second moment.

Definition (Training Dynamics in the Mean-Field Limit).

Let $F[\mu]$ be the expected population loss associated with a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. The gradient descent dynamics in the mean-field limit induce a continuity equation on μ_t, given by:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{F}}{\delta \mu}\right) = 0$$

where $\frac{\delta F}{\delta \mu}$ denotes the first variation (functional derivative) of F with respect to μ, and the flow occurs in the Wasserstein-2 space $\mathcal{P}_2(\mathbb{R}^d)$.

Definition (Memory Measure).

A memory measure $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$. is defined as a stationary solution of the continuity equation above. That is:

$$\nabla \cdot \left(\mu^* \nabla \frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) \right) = 0$$

Equivalently, μ^* satisfies the Euler–Lagrange condition:

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) = \text{constant on } supp(\mu^*)$$

Intuitively, a memory measure is a fixed point of the learning dynamics in measure space. It corresponds to a state of dynamical equilibrium, where the distribution of parameters no longer evolves under gradient descent. Unlike weight vectors or activations, this structure is independent of the neural architecture and admits analysis via tools from optimal transport and PDE theory.

Proposition (Well-posedness of Gradient Flow).

Under suitable assumptions on F (e.g., displacement convexity, lower semi continuity, coercivity), the gradient flow

$$\frac{\partial \mu_t}{\partial t} = -\nabla_{W_2} \mathcal{F}(\mu_t)$$

in $\mathcal{P}_2(\mathbb{R}^d)$ is well-posed. That is, there exists a unique curve μ_t satisfying the gradient flow equation, and for $t \to \infty$, $\mu_t \to \mu^*$ in Wasserstein-2 distance, where μ^* is a memory measure.

Proof: See Ambrosio, Gigli, and Savaré (2008), Chapter 11.

Preliminaries

This section introduces the mathematical tools necessary for the formulation of our theoretical framework. We begin by reviewing foundational concepts in probability measure theory, followed by a detailed definition of the Wasserstein-2 space. Finally, we describe how gradient flows can

be rigorously defined in this geometric setting, preparing the groundwork for analyzing neural network dynamics through distributional evolution.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of Borel probability measures on $\mathbb{R}^d$. A measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfies $\mu(\mathbb{R}^d) = 1$, ensuring unit total mass. Given a sequence $(\mu_n) \subset \mathcal{P}(\mathbb{R}^d)$ we say that μ_n converges weakly to $\mu \in \mathcal{P}(\mathbb{R}^d)$, denoted $\mu_n \rightharpoonup \mu$, if for every bounded and continuous function $\phi: \mathbb{R}^d \to \mathbb{R}$, the following condition holds:

$$\lim_{n\to\infty} \int_{R^d} \varphi(x)\ d\mu_n(x) = \int_{R^d} \varphi(x)\ d\mu(x)$$

This notion of convergence is particularly suitable when dealing with distributions over parameter spaces, especially in the infinite-width limit of neural networks, where pointwise convergence of parameters loses meaning.

The Wasserstein-2 Space

We define the Wasserstein-2 space $\mathcal{P}_2(\mathbb{R}^d)$ as the set of probability measures in $\mathcal{P}(\mathbb{R}^d)$ with finite second moment:

$$\mathcal{P}_2(\mathbb{R}^d) \coloneqq \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) \;\middle|\; \int_{\mathbb{R}^d} |x|^2\ d\mu(x) < \infty \right\}.$$

The Wasserstein-2 distance between two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$W_2^2(\mu,\nu) \coloneqq \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{R^d\times R^d} |x-y|^2\, d\gamma(x,y),$$

where $\Gamma(\mu,\nu)$ denotes the set of all couplings of μ and ν, i.e., probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. This distance metrizes weak convergence along with convergence of second moments and endows $\mathcal{P}_2(\mathbb{R}^d)$ with a rich geometric structure that enables differential calculus on the space of distributions.

Gradient Flows in Wasserstein Space

Let $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$ be a functional defined on the space of probability measures. This function may represent, for instance, the expected loss in a learning problem. The gradient flow of F in Wasserstein space is defined as a curve $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ that satisfies:

$$\frac{d\mu_t}{dt} = -\nabla_{W_2}\mathcal{F}(\mu_t),$$

where $\nabla_{W_2}\mathcal{F}$ denotes the Wasserstein gradient of F with respect to the metric structure induced by W_2. This evolution equation admits a PDE representation known as the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{F}}{\delta \mu}(\mu_t) \right) = 0,$$

where $\frac{\delta F}{\delta \mu}$ denotes the first variation (functional derivative) of F. The quantity $\nabla \frac{\delta F}{\delta \mu}$ serves as the velocity field guiding the mass transport of the measure μ_t . This formulation, pioneered by Otto and further developed by Ambrosio, Gigli, and Savaré, provides a rigorous analytic framework for studying evolution equations on spaces of measures and underpins the theoretical understanding of training as an optimization process in infinite-dimensional space.

Neural Networks in the Mean-Field Limit

In this section, we reformulate neural networks as systems of interacting particles and study their behavior as the number of neurons tends to infinity. This asymptotic perspective, often referred to as the **mean-field limit**, provides a rigorous mathematical framework for analyzing the collective behavior of parameters as probability distributions. This perspective allows learning dynamics to be captured not by trajectories in finite-dimensional parameter space, but by the evolution of distributions governed by partial differential equations.

Neural Networks as Particle Systems

We begin with a standard one-hidden-layer neural network defined as

$$f_n(x; \theta_1, \dots, \theta_n) = \frac{1}{n} \sum_{i=1}^{n} \sigma(x; \theta_i)$$

where $\theta_i \in \mathbb{R}^d$

encodes the parameters (e.g., weights and biases) of the iii-th neuron, and σ: $\mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ is a fixed activation-output map. The subscript n reflects the network width. In this formulation, the output is the empirical average of individual neuron responses.

To capture the statistical behavior of the parameter ensemble, we define the empirical measure associated with the neural network as

$$\mu^{(n)} := \frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i},$$

where δ_θ denotes the Dirac delta measure at $\theta \in \mathbb{R}^d$. This measure encodes the entire parameter configuration and permits analysis through measure-theoretic and probabilistic tools.

As $n \to \infty$, we are interested in the convergence of $\mu^{(n)}$ to a limit $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. This convergence is understood in the weak sense and forms the basis of the mean-field approximation.

Limit Network Representation

Assuming weak convergence $\mu^{(n)} \rightharpoonup \mu$, we define the mean-field network as the limit function

$$f(x;\mu) \coloneqq \int_{R^d} \sigma(x;\theta)\ d\mu(\theta).$$

This integral representation replaces the finite sum in the original network with an expectation under the limiting measure μ, effectively transitioning from a discrete model to a continuous ensemble of neurons. The network is now entirely characterized by the probability measure μ, which serves as its infinite-width representation.

This transformation is not merely formal. Under mild assumptions on the smoothness and boundedness of σ, one can show that $f_n(x;\theta_1,\ldots,\theta_n) \to f(x;\mu)$ uniformly over compact domains with high probability. The result connects statistical approximation theory with empirical process theory and underlies the validity of the mean-field model.

Dynamics of Training in the Mean-Field Limit

Let $L(\mu)$ denote the expected loss of the network induced by measure μ, defined as

$$\mathcal{L}(\mu) \coloneqq E_{(x,y)\sim D}[\ell\ (f(x;\mu), y)],$$

where ℓ is a standard loss function (e.g., squared loss or cross-entropy), and D is the data distribution. The evolution of μ_t under gradient descent on L gives rise to a time-dependent family of distributions governed by a Wasserstein gradient flow:

$$\frac{d\mu_t}{dt} = -\nabla_{W_2}\mathcal{L}(\mu_t).$$

This gradient flow can be equivalently formulated as the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{L}}{\delta \mu}(\mu_t)\right) = 0,$$

where the velocity field is derived from the first variation of the loss functional. The training process thus corresponds to a transport of mass in parameter space, where each particle (or neuron) evolves under the influence of the loss-induced potential field.

Interpretation of the Mean-Field Limit

The mean-field limit provides several critical advantages. First, it removes dependence on specific parameterizations or architectures, treating the network as a dynamical system in probability space. Second, it enables a variational formulation of learning, allowing one to study convergence and

generalization through convex analysis and PDE techniques. Third, the framework naturally supports the definition of **stationary distributions**, which will serve as our formal definition of memory in the subsequent section.

The measure μ_t captures the entire state of the network at time *t*, and its evolution is fully described by the gradient flow of the loss functional in $\mathcal{P}_2(\mathbb{R}^d)$. As training progresses, we expect μ_t to approach a fixed point μ^*, whose properties reflect the system's learned representations. In this formulation, learning is not a movement through parameter space, but through the space of distributions — an inherently geometric and collective process.

Training as a Distributional Partial Differential Equation

In the mean-field formulation of neural networks, the training process is no longer viewed as discrete updates to finite-dimensional parameters, but rather as the evolution of a probability distribution over parameter space. This distributional viewpoint allows us to describe learning dynamics as a continuum flow — a process that unfolds not in parameter vectors, but in the space of measures. In this section, we formalize that intuition and demonstrate how the gradient descent training of infinitely wide neural networks converges, in the large *n* limit, to a partial differential equation (PDE) in the space of probability measures. This PDE governs the temporal evolution of the network's parameter distribution and constitutes the core dynamical object in our analysis of memory.

Loss Functional and Functional Derivative

Let $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ denote the time-dependent probability measure representing the network's parameter distribution at training time *t*. Suppose we are given a data distribution *D* over input-output pairs $(x,y) \in \mathbb{R}^p \times \mathbb{R}$, and an activation function or neuron output mapping $\sigma: \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$. Then we can define a functional $F: \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$, representing the expected loss under the parameter distribution μ, as follows:

$$\mathcal{F}(\mu) \coloneqq E_{(x,y)\sim D}\left[\ell\left(\int_{R^d} \sigma(x;\theta)\ d\mu(\theta), y\right)\right],$$

where $\ell: \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ is a convex loss function, such as the squared error $\ell(a,b) = \frac{1}{2}(a-b)^2$. The integral $\int \sigma(x;\theta)\ d\mu(\theta)$ represents the prediction made by the infinite-width neural network, and the expectation over *D* computes the average prediction error across the data distribution.

To compute the gradient flow of this loss functional in Wasserstein space, we require its **first variation**, also referred to as the functional derivative. Informally, this derivative quantifies how a small perturbation in the measure μ influences the value of the functional *F*. Formally, it is defined via the Gâteaux derivative:

$$\frac{d}{d\epsilon}\mathcal{F}\big(\mu + \epsilon(\nu-\mu)\big)\big|_{\epsilon=0} = \int_{R^d} \frac{\delta\mathcal{F}}{\delta\mu}(\theta)\ d(\nu-\mu)(\theta),$$

for any perturbation measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$. When the functional F is sufficiently regular — for instance, when σ is smooth and bounded and the loss function ℓ is differentiable — the functional derivative exists and admits an explicit representation:

$$\frac{\delta \mathcal{F}}{\delta \mu}(\theta) = E_{(x,y)\sim D}\left[\ell'(f(x;\mu), y) \cdot \frac{\partial \sigma(x;\theta)}{\partial \theta}\right].$$

This derivative defines a scalar field over the parameter space $\mathbb{R}^d$. Its gradient gives rise to the **velocity field** that governs the evolution of the distribution μ_t under training.

Continuity Equation Formulation

With the functional derivative in hand, we can describe the temporal evolution of the parameter distribution as a PDE. Specifically, the measure μ_t evolves according to a continuity equation, which expresses conservation of probability mass under a velocity field. The equation takes the form:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v_t \mu_t) = 0,$$

where the velocity field $v_t(\theta)$ is given by the negative gradient of the functional derivative:

$$v_t(\theta) := -\nabla\left(\frac{\delta \mathcal{F}}{\delta \mu}(\theta)\right).$$

This formulation is the infinite-dimensional analogue of gradient descent in finite-dimensional Euclidean space. Rather than evolving a single point θ_t, the continuity equation describes how the entire distribution of parameters flows through the landscape of the loss functional F. Each infinitesimal mass of probability moves in the direction of steepest descent, and the full evolution reflects the aggregate behavior of infinitely many such particles.

Properties of the Gradient Flow

The gradient flow described above is well-posed under standard assumptions. If F is lower semicontinuous and **displacement convex** — a condition that generalizes convexity to geodesic paths in Wasserstein space — then the evolution equation admits a unique solution μ_t that depends continuously on the initial distribution μ_0. Moreover, the flow converges to a limiting distribution $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ as $t \to \infty$, which satisfies the stationary condition

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) = \text{constant on } supp(\mu^*)$$

This Euler–Lagrange condition characterizes equilibrium points of the flow and will form the mathematical basis for our definition of memory in the next section. These stationary measures are not transient or noisy but instead reflect fixed structures in distribution space — structures that retain the result of learning even after training has ceased.

The mathematical theory underlying this flow was first developed in the context of thermodynamics and fluid dynamics by Jordan, Kinderlehrer, and Otto, and was later rigorously formalized by Ambrosio, Gigli, and Savaré. These results connect the geometry of $\mathcal{P}_2(\mathbb{R}^d)$ to variational problems in infinite-dimensional spaces, enabling a rich theory of dynamics and equilibria.

Conceptual Implications

The reformulation of training as a PDE over measures introduces a profound shift in perspective. In traditional optimization, training is a path through a finite-dimensional parameter space, governed by local updates. In the mean-field setting, training becomes a trajectory through the space of distributions, governed by global flows. This abstraction strips away architectural dependencies and instead focuses on the geometry of learning itself.

The continuity equation enables us to characterize learning trajectories, convergence rates, and stability properties at a macroscopic level. It permits analysis using tools from functional analysis, calculus of variations, and optimal transport. More importantly, it sets the stage for defining and analyzing the notion of **memory** in purely mathematical terms — as a fixed point of this dynamical system, independent of network instantiation.

In what follows, we will define memory as a stationary distribution under this PDE and explore its theoretical and practical consequences.

Defining Memory as a Stationary Distribution

In the preceding sections, we have described the evolution of neural network parameters in the infinite-width limit as a gradient flow in the Wasserstein-2 space $\mathcal{P}_2(\mathbb{R}^d)$. Within this continuous framework, the behavior of the parameter distribution as training progresses is governed by a distributional partial differential equation (PDE). In this section, we formalize the notion of "memory" in neural networks rigorously, defining it explicitly as a stationary distribution of the aforementioned PDE. We will explore the mathematical foundations underpinning this definition, discuss conditions ensuring existence and uniqueness, and interpret these stationary solutions as capturing the learned representations within the network.

Stationary Measures and Equilibrium Conditions

Recall from the previous section that the temporal evolution of the parameter distribution μ_t is governed by the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot \left(\mu_t \nabla \frac{\delta \mathcal{F}}{\delta \mu}(\mu_t) \right) = 0,$$

where F denotes the expected loss functional and $\frac{\delta \mathcal{F}}{\delta \mu}$ its functional derivative. We say a measure $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ is a stationary solution or equilibrium measure if it satisfies:

$$\nabla \cdot \left(\mu^* \nabla \frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) \right) = 0.$$

Equivalently, the equilibrium condition can be expressed via the Euler–Lagrange condition, requiring the first variation of the loss functional to be constant on the support of the equilibrium measure:

$$\frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) = \text{constant on } supp(\mu^*)$$

for some constant $c \in \mathbb{R}$. Intuitively, this condition means that at equilibrium, no infinitesimal re-distribution of mass within the support of μ^* can yield a decrease in the loss.

We now adopt this equilibrium notion formally as our definition of memory:

Definition (Memory Distribution): A probability measure $\mu^* \in \mathcal{P}_2(\mathbb{R}^d)$ satisfying the stationary condition described above is called a memory distribution of the infinitely wide neural network.

In other words, memory is defined not as a specific choice of weights or activations, but rather as a global statistical configuration of parameters that is invariant under training dynamics.

Existence and Uniqueness of Memory Measures

The question of existence and uniqueness of stationary measures hinges on the properties of the loss functional F. Standard assumptions ensuring existence and uniqueness typically include lower semicontinuity and displacement convexity of the functional. More explicitly, if F is proper, lower semicontinuous, and displacement convex, then a minimizer of F exists and is unique. Such minimizers necessarily satisfy the equilibrium conditions described above and thus constitute memory distributions.

Ambrosio, Gigli, and Savaré (2008) established general existence results for gradient flows in Wasserstein space, providing the mathematical basis for our formulation. Under standard regularity assumptions—such as Lipschitz continuity and smoothness conditions on the activation functions and convexity conditions on the loss—the existence of at least one stationary measure μ^* is guaranteed. Strict convexity further ensures uniqueness, resulting in a well-defined notion of memory.

Convergence to Memory Distributions

A central feature of the gradient flow structure in Wasserstein space is that the functional F decreases monotonically along the trajectory of the distribution μ_t . Formally, this monotonic decrease is expressed by the energy dissipation identity:

$$\frac{d}{dt}\mathcal{F}(\mu_t) = -\int_{R^d} ||\nabla \frac{\delta \mathcal{F}}{\delta \mu}(\theta)||^2 \, d\mu_t(\theta) \leq 0.$$

This inequality demonstrates that the expected loss functional acts as a Lyapunov function for the system. Due to the monotone decrease and boundedness from below (assuming F is proper and lower bounded), the parameter distribution μ_t converges, in the Wasserstein-2 distance, towards the set of stationary measures as time goes to infinity. Under strict displacement convexity conditions, this convergence is stronger: the parameter distribution μ_t converges uniquely to the stationary measure μ^*.

Interpretation and Implications

The definition of memory as a stationary distribution offers a natural and rigorous interpretation: memory encapsulates those parameter configurations that remain invariant under training dynamics. Each equilibrium measure represents a global, self-consistent geometric configuration of parameters for which no local improvement in predictive performance can be achieved via infinitesimal perturbations.

This perspective has several important implications for understanding learning and generalization. Concentrated stationary measures with compact or tightly localized support correspond to configurations with implicit regularization properties, akin to finite-width networks with strong inductive biases. Conversely, diffuse or multimodal stationary measures reflect networks that can memorize multiple patterns, indicative of highly over-parameterized settings.

Relation to Finite-Width Networks

To bridge the gap between finite-width and infinite-width networks, we consider the empirical measure $\mu_t^{(n)}$ defined by the finite-width neural network parameters at time t:

$$\mu_t^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i^{(n)}(t)}.$$

As the width $n \to \infty$, standard propagation-of-chaos arguments ensure that $\mu_t^{(n)}$ converges in Wasserstein-2 sense towards the mean-field distribution μ_t. Consequently, in the limit $t \to \infty$, the finite-width empirical distributions converge to the stationary memory distribution:

$$\mu_t^{(n)} \xrightarrow[n\to\infty]{W_2} \mu_t, \quad \mu_\infty^{(n)} \xrightarrow[n\to\infty]{W_2} \mu^*.$$

Thus, finite-width networks can be interpreted as finite-sample approximations to the infinite-width memory distribution. Memory in finite-width networks emerges as a statistical approximation of the stationary solutions characterizing the infinite-width limit.

In summary, we have formally defined memory in neural networks as a stationary measure of the Wasserstein gradient flow associated with the infinite-width network training dynamics. Existence, uniqueness, and convergence results have been discussed, along with interpretative insights connecting geometric properties of memory distributions to learning and generalization.

The subsequent sections of the paper discuss the analytical characterizations of these memory distributions, investigate their structural properties in concrete scenarios, and explore how varying network architectures and training algorithms influence the formation and geometry of stationary solutions.

Wasserstein Gradient Flows: Mathematical and Analytical Insights

Wasserstein gradient flows can be naturally viewed as curves of steepest descent in the metric space $\mathcal{P}_2(\mathbb{R}^d)$, endowed with the Wasserstein-2 distance. From a geometric standpoint, the Wasserstein metric endows the space of probability measures with a Riemannian-like structure, enabling a rich geometrical interpretation of distributional dynamics. This structure facilitates a deeper understanding of concepts like geodesics, curvature, and convexity in the space of probability measures, which in turn impact convergence rates, stability, and generalization.

A major practical advantage of interpreting training dynamics as Wasserstein gradient flows is the availability of advanced analytical and numerical methods. These include:

Variational schemes: Minimizing movements and JKO (Jordan–Kinderlehrer–Otto) schemes provide powerful numerical approximations for solutions of gradient flow equations.

Energy-dissipation inequalities: They quantify convergence rates and enable precise stability analyses.

Functional inequalities: Tools such as Talagrand inequalities, log-Sobolev inequalities, and displacement convexity conditions yield direct control on the convergence behavior of the parameter distributions.

While much of the theory of Wasserstein gradient flows rests on convexity assumptions, neural network loss landscapes are often non-convex. Extending classical gradient-flow theory beyond convexity remains a significant and largely open mathematical challenge. Recent research has begun exploring generalized notions of convexity, such as Polyak-Łojasiewicz conditions and gradient dominance, to characterize convergence behavior in these more general settings.

Wasserstein gradient flows offer a rigorous yet flexible mathematical foundation for analyzing neural network training dynamics. They provide a geometric and analytical framework that not only clarifies the interpretation of memory as stationary measures but also suggests powerful

computational tools and opens pathways to solving contemporary mathematical problems arising from practical non-convex optimization scenarios in deep learning.

Stationary Distributions and the Concept of Memory

Stationary distributions serve as a foundational component of the framework developed throughout this paper. By characterizing neural network training as a gradient flow in the Wasserstein-2 space, we arrive at a mathematically rigorous and intrinsic definition of memory. A stationary distribution, formally introduced as an equilibrium solution to the gradient-flow PDE, is a probability measure μ* that remains invariant under the training dynamics. More explicitly, a stationary measure satisfies the condition:

$$\nabla \cdot \left(\mu^* \nabla \frac{\delta \mathcal{F}}{\delta \mu}(\mu^*) \right) = 0.$$

From a variational viewpoint, this condition implies that the first variation (functional derivative) of the loss functional F is constant across the support of the stationary measure. Thus, no local perturbation of the distribution can produce a further decrease in the loss. Geometrically, stationary distributions represent points in the infinite-dimensional distribution space where the gradient of the loss functional vanishes or is uniform, indicating stable equilibria.

Conceptually, the notion of memory as stationary distributions differs significantly from traditional definitions reliant upon specific neural architectures or parameter vectors. Instead of representing memory by particular weights, neuron activations, or hidden states, memory is now viewed as a robust global structure in distribution space. This abstraction is both powerful and natural, allowing memory to be studied as an intrinsic property of the training dynamics, independent of any finite-dimensional parameterization or specific neural architecture. Hence, memory emerges as a stable, self-consistent statistical configuration of parameters that encapsulates the network's learned representation of the data.

Furthermore, this distributional notion of memory sheds light on critical aspects of neural learning and generalization. Concentrated stationary distributions, characterized by low entropy and compact support, are indicative of strong implicit regularization effects, mirroring classical notions of good generalization. Conversely, more diffuse or multimodal stationary distributions correspond to over-parameterized networks capable of memorizing diverse patterns. Thus, the geometric properties of stationary distributions directly translate into meaningful implications about the network's generalization capabilities and the nature of its learned representations.

Existence and Uniqueness of Stationary Measures

Given the foundational significance of stationary distributions as the formal definition of memory, it becomes essential to address questions of their existence and uniqueness rigorously. Classical results in the theory of Wasserstein gradient flows ensure existence under general assumptions:

First, consider the loss functional F: $\mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R} \cup \{+\infty\}$. If F is proper, lower-semicontinuous, and displacement convex, then it admits at least one stationary distribution μ^*. Such conditions are typically satisfied by common neural network loss functions—for example, mean squared error losses combined with sufficiently regular activation functions.

Moreover, the uniqueness of stationary measures is determined primarily by the strictness of displacement convexity. Strict displacement convexity ensures that the stationary measure is unique, providing an unambiguous notion of memory. Conversely, if displacement convexity is non-strict or degenerate, multiple stationary distributions might coexist. Such multiplicity corresponds to networks capable of memorizing distinct modes or patterns, each represented by a separate equilibrium measure.

These theoretical guarantees have important practical implications: existence ensures a meaningful notion of memory is always achievable, while uniqueness governs the interpretability and predictability of learning outcomes. In practice, most standard neural architectures and losses produce landscapes where at least local uniqueness is achievable, although global uniqueness may not always hold.

Applications and Implications for Learning Theory

The formal definition of memory as stationary distributions in Wasserstein space has profound theoretical implications and broad applications in the learning theory of neural networks. One immediate consequence is an enhanced understanding of generalization and capacity control. Stationary distributions inherently embody regularization properties encoded through their geometric structure. For example, concentrated stationary distributions reflect implicit regularization akin to norm-based methods, thus leading to strong generalization. Conversely, diffuse stationary distributions characterize models with excessive memorization capacity, typically arising in highly overparameterized scenarios.

This distributional viewpoint also suggests novel analytical tools. Energy-dissipation inequalities, integral curvature bounds, and other functional inequalities become accessible through the Wasserstein geometry, enabling rigorous quantitative analyses of convergence rates, stability properties, and generalization errors.

Additionally, computational schemes arising naturally from the gradient-flow formulation, such as Jordan–Kinderlehrer–Otto (JKO) approximations, can offer new numerical methods to simulate training dynamics, analyze convergence behavior, and identify memory distributions explicitly in practice.

Furthermore, this rigorous notion of memory may inspire new algorithms explicitly designed to find equilibrium measures efficiently, improving practical training convergence and robustness.

Comparison with Traditional Memory Models

The concept of memory presented herein contrasts significantly with traditional notions used in neural network theory. Conventionally, memory is represented explicitly by weights, activation

vectors, or discrete hidden states, dependent strongly on the architecture and parameterization of specific networks. Such definitions are inherently finite-dimensional, localized, and tied to particular implementations, limiting general theoretical insights and robustness under perturbations.

In sharp contrast, defining memory as a stationary distribution provides a universal, intrinsic characterization independent of network parameterization or architecture specifics. This approach removes the inherent limitations associated with traditional, finite-dimensional memory models, allowing deeper theoretical analysis of learned representations, generalization behavior, and robustness.

Moreover, the geometric nature of stationary distributions facilitates direct mathematical insights into their structural properties, stability under perturbations, and generalization capacity. Traditional finite-dimensional models typically require heuristic or empirical studies, while our distributional framework admits rigorous mathematical analysis through tools from optimal transport, PDE theory, and geometric measure theory.

In summary, viewing memory through the lens of stationary distributions represents a substantial theoretical advancement, bridging a fundamental conceptual gap by providing a rigorous, unified, and robust mathematical notion of memory. This approach offers both conceptual clarity and powerful analytical tools, enriching theoretical understanding and practical methodology in neural network learning.

Conclusion

In this paper, we have presented a mathematically rigorous formulation of memory in neural networks as stationary distributions arising naturally from the Wasserstein gradient-flow dynamics of infinitely wide networks. Unlike traditional finite-dimensional approaches that depend explicitly on particular network architectures, our measure-theoretic definition provides an intrinsic and universal characterization, bridging the gap between mathematical theory and practical learning dynamics.

We began by formulating neural network training as gradient flows in the Wasserstein-2 space, providing a distributional partial differential equation (PDE) governing parameter evolution. Within this context, we defined memory formally as equilibrium points or stationary solutions of this PDE. These stationary distributions represent robust, stable parameter configurations invariant under training dynamics, capturing the essence of learned information within the network.

Subsequently, we established conditions ensuring existence and uniqueness of stationary distributions, highlighting displacement convexity and related analytical conditions from optimal transport theory. We showed that stationary measures inherently encode generalization and regularization properties, allowing new analytical insights into network training dynamics, convergence behavior, and robustness.

A comparison with traditional neural memory models revealed significant theoretical and practical advantages of our distributional formulation, particularly its architecture-independence, robustness

to perturbations, and amenability to rigorous mathematical analysis through advanced tools from measure theory and PDEs.

This rigorous viewpoint opens numerous directions for future research. Extending these theoretical tools beyond displacement convexity to non-convex landscapes represents an exciting mathematical challenge. Moreover, translating our theoretical insights into computationally efficient algorithms for identifying and analyzing memory distributions can significantly impact practical deep learning methodologies. Ultimately, our measure-theoretic framework enriches the theoretical understanding of memory in neural networks, providing a foundation for deeper analysis and more robust applications.

References

Ambrosio, L., Gigli, N., & Savaré, G. (2008). *Gradient flows in metric spaces and in the space of probability measures.* Birkhäuser.

Chizat, L., & Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 3040–3050.

Jordan, R., Kinderlehrer, D., & Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1–17.

Mei, S., Montanari, A., & Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), E7665–E7671.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1–2), 101–174.

Rotskoff, G. M., & Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv preprint arXiv:1805.00915.*

Santambrogio, F. (2015). *Optimal transport for applied mathematicians.* Birkhäuser.

Villani, C. (2008). *Optimal transport: Old and new.* Springer-Verlag.