ON THE ANALYSIS OF STOCHASTIC GRADIENT DESCENT IN NEURAL NETWORKS VIA GRADIENT FLOWS

VIET NGUYEN

Abstract

Research in neural network theory is steadily gaining traction, as there is a growing interest in the thorough understanding of the functionalities and the mechanisms through which these models achieve strong performances in decision problems. Several methods have been proposed to quantitatively assess the optimization process of the neural network's high dimensional non-convex objective, employing various tools such as kernel methods, global optimization, optimal transport, and functional analysis. In this work, we focus on Mei et al.'s analysis of the mean risk field of two-layer neural networks in [12], which associates stochastic gradient descent's (SGD) training dynamics to a partial differential equation (PDE) in the space of probability measures with the topology of weak convergence. Precisely, we dissect the proof of the convergence of SGD's dynamics to the solution of the PDE, showcase several results regarding the analysis of the latter and their implications on the training process of neural networks via SGD, and discuss related work as well as potential further explorations stemming from various fields.

Contents

1. Neural networks, stochastic gradient descent, gradient flow

The widespread of use and research related to neural networks as a supervised (and more recently, a semi-supervised and even unsupervised) learning model is largely due to their capabilities of learning complex correspondences between data and labels. They found use in diverse fields: natural language processing and understanding [16], computer vision [8], and lately in reinforcement learning [9], to name a few. Their theoretical properties have also been subject to a lot of research, most notably the universal approximation theorem [6] for continuous functions over a compact domain, and extensions of these theorems to cases where neural networks enjoy a countably infinite or even uncountable number of hidden nodes. We restrict ourselves to the study of two layer neural networks in the supervised study setting.

Given a finite dataset $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\} \subset (\mathcal{X} \times \mathcal{Y})^M$ from an unknown distribution $\mathbb{P}$ of features and their corresponding labels, neural networks leverage the descriptive potential of linear combinations of perceptrons arranged in a multi-layered structure to infer complex and general correspondences $f : \mathcal{X} \longrightarrow \mathcal{Y}$ from samples in $\mathcal{D}$. We remark that there are little to no assumptions on $\mathcal{Y}$, our analysis applies to both categorical and continuous labels, i.e. machine learning classification and regression respectively.

Definition 1.1. (Neural network) A two-layer neural network of N hidden units with a parametrization $\boldsymbol{\theta}$ is a **function** $\hat{y} : \mathbf{R}^d \longrightarrow \mathbf{R}$ such that on input $\boldsymbol{x} \in \mathbf{R}^d$, computes:

$$\hat{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i)$$

where $\sigma_* : \mathbf{R}^d \longrightarrow \mathbf{R}$ is an activation function, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$ is a collection of parameters of each hidden unit, and $\boldsymbol{\theta}_i = (a_i, b_i, \boldsymbol{w}_i)$ with $\dim \boldsymbol{\theta}_i = D$ where:

$$\sigma_*(\boldsymbol{x}; \boldsymbol{\theta}_i) = a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle + b_i)$$

where $\sigma : \mathbf{R} \longrightarrow \mathbf{R}$ is a squashing function. Historically, σ's range is restricted to a bounded set, however more recent studies and applications use the rectified linear unit (ReLU) and its variants as a baseline, a function of infinite range.

One can think of a neural network as a weighted average of affine units. Here, $\boldsymbol{w} = \boldsymbol{w}_i$ is the input-to-hidden weight matrix, b_i a layer bias, and a_i the hidden-to-output weights. It is easier to think of $\boldsymbol{\theta}$ as a collection of random variables $\boldsymbol{\theta}_i$ rather than a vector in $\mathbf{R}^{D \times N}$. We want to choose parameters $\boldsymbol{\theta}_i, i \leq N$, such that the neural network parametrized by $\boldsymbol{\theta}$ minimizes an objective function, here chosen to be the risk $R_N(\boldsymbol{\theta}) = \mathbb{E}\left[\ell(y, \hat{y}(\boldsymbol{x}; \boldsymbol{\theta}))\right]$ for some loss function ℓ. We focus on the squared loss, i.e. $\ell(y, \hat{y}) = (y - \hat{y})^2$, however we note that the same analysis can be done with different choices of loss functions.

Stochastic gradient descent is an iterative optimization algorithm that can be used to learn the parameters of the neural network that minimizes our objective. In practice, we apply the weak law of large numbers and work directly on the empirical risk, as they converge in probability to the population risk $\mathbf{R}_N$.

Definition 1.2. (Stochastic gradient descent, SGD) Given a loss function ℓ, stochastic gradient descent amounts to the following iteration for discrete timesteps $k \geq 0$:

$$\boldsymbol{\theta}^{k+1} \leftarrow \boldsymbol{\theta}^k - s_k \nabla_{\boldsymbol{\theta}} \ell(y_k, \hat{y}(\boldsymbol{x}_k; \boldsymbol{\theta}^k))$$

for some variable learning rate $s_k \in \mathbf{R}^+$.

For neural networks, this amounts to the iteration:

$$\boldsymbol{\theta}_i^{k+1} \leftarrow \boldsymbol{\theta}_i^k + 2 s_k (y_k - \hat{y}(\boldsymbol{x}_k; \boldsymbol{\theta}^k)) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\boldsymbol{x}_k; \boldsymbol{\theta}_i^k)$$

In a sense, one is taking a calculated step in the direction of steepest descent in the loss landscape over the parameter space $\mathbf{R}^D$, for each hidden unit $i \leq N$. We assume the algorithm never visits the same data sample $(\boldsymbol{x}_k, y_k)$. This is reasonable, given the large datasets at disposal in many real world applications.

Often, it is the case that the aforementioned loss landscape is highly non-convex. However, empirical results do indeed demonstrate that neural networks trained via SGD (or variants of it) not only achieves near optimal losses but also very strong generalization properties, which motivates the study of 1. local minima to which SGD weights converge to and the corresponding network's' generalization properties and 2. SGD dynamics during training itself. To this end, one can think of essentially shrinking the learning rate to 0, obtaining a gradient flow in $\mathbf{R}^D \times \mathbf{R}$. More precisely, we consider the smooth curve $\gamma : \mathbf{R} \longrightarrow \mathbf{R}^D$ such that:

$$\gamma'(t) = -\nabla_{\boldsymbol{\theta}_i} \ell(y, \hat{y}(\boldsymbol{x}; \boldsymbol{\theta}_i^t))$$

as a trajectory in the loss landscape of the two-layer neural network (with the appropriate Euclidean metric). Remarkably, authors of [12] proved that in a suitable scaling limit, the SGD dynamics admits an asymptotic description in terms of a PDE, which corresponds exactly to a gradient flow in the space of probability measures over $\mathbf{R}^D$, with the Kantorovich metric[1], which minimizes a generalized risk function defined for $\rho \in (\mathscr{P}(\mathbf{R}^D), W_2)$, denoted $R(\rho)$. The authors suggest that this association simplifies the analysis of the loss landscape of two-layer neural networks, and comes with various theoretical guarantees from PDE theory and propagation of chaos.

2. Setup: generalized risks, distributional dynamics

One begins by considering the risk function $R_N(\boldsymbol{\theta})$:

$$\begin{aligned}\mathbb{E}\left[(y-\hat{y}(\boldsymbol{x};\boldsymbol{\theta}))^2\right] &= \mathbb{E}\left[y^2\right] - \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left[y\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)\right] + \frac{1}{N^2}\sum_{i,j}\mathbb{E}\left[\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_j)\right] \\ &:= R_{\#} + \frac{2}{N}V(\boldsymbol{\theta}_i) + \frac{1}{N^2}\sum_{i,j}U(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\end{aligned}$$

where we define $R_{\#} = \mathbb{E}\left[y^2\right]$, the risk of the trivial predictor $\hat{y} = 0$, potentials $V(\boldsymbol{\theta}_i) = \mathbb{E}\left[y\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)\right]$, and $U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_J) = \mathbb{E}\left[\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_i)\sigma_*(\boldsymbol{x};\boldsymbol{\theta}_j)\right]$. We shall always assume that the expectations defining V and U exist for all parameters $\boldsymbol{\theta} \in \mathbf{R}^D$. Since R_N depends on $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N)$ only through their empirical distribution $\hat{\rho}^{(N)} = N^{-1}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_i}$, one can consider the following generalization for $\rho \in \mathscr{P}(\mathbf{R}^D)$:

$$R(\rho) = R_{\#} + 2\int V\, d\rho + \iint U\,(d\rho)^2$$

we readily see that $R(\hat{\rho}^{(N)}) = R_N(\boldsymbol{\theta})$, as the point masses discretize the integrals. Intuitively, considering all probability measures over $\mathbf{R}^D$ amounts to taking into consideration neural networks of any hidden layer width, which corresponds to the support of the measure in question. With finite support, we recover the usual risk R_N and finite hidden units. Similarly, with countably infinite and uncountable support, we recover the infinite width neural network with the generalized risk. Under some mild assumptions, we can prove that $\inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) = \inf_\rho R(\rho) + O(1/N)$.

Given this setting, we will see how SGD dynamics in $\mathbf{R}^D \times \mathbf{R}$ can be approximated by a gradient flow in $\mathscr{P}(\mathbf{R}^D)$, namely for step sizes $s_k = \epsilon\xi(k\epsilon)$ for some $\xi : \mathbf{R}^+ \to \mathbf{R}^+$ with nice properties, and $\hat{\rho}_k^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{\theta}_i^k}$ the point process of parameters $\boldsymbol{\theta}$ at the k-th iteration of SGD, we take a closer look at Mei et al.'s proof in [12] of the weak convergence:

$$\hat{\rho}_{t/\epsilon}^{(N)} \Rightarrow \rho_t$$

as $N \to \infty, \epsilon \to 0$, i.e. convergence in the weak topology of $\mathscr{P}(\mathbf{R}^D)$. Here, the dynamics of ρ_t is described by the following PDE:

$$\begin{aligned}\partial_t\rho_t &= 2\xi(t)\nabla_{\boldsymbol{\theta}}\left(\rho_t\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta};\rho_t)\right) \\ \Psi(\boldsymbol{\theta};\rho) &= \frac{1}{2}\frac{\delta R(\rho)}{\delta\rho(\boldsymbol{\theta})} = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta},\boldsymbol{\theta}')\,d\rho(\theta')\end{aligned}$$

therefore, establishing that the analysis of SGD' training dynamics amounts to analyzing the above PDE. Ψ is a functional derivative, and can be interpreted as the additional energy of adding a single particle at $\boldsymbol{\theta} \in \mathbf{R}^D$. We restrict ourselves to mathematical interpretations of results, however we remark that authors of [12] provided an elegant interpretation from a particle dynamics perspective, in which the above PDE enforces a principle of local conservation of mass due to particles not being able to move discontinuously. In addition, the authors pointed out several benefits of this method, most notably being factoring out the invariance of the risk $R_{\#}$ and of the SGD dynamics, as well as exploiting underlying symmetries, such as the fact that U is a symmetric positive semidefinite kernel. Furthermore, powerful tools from the literature of gradient flows on the space of probability measures can be used in the analysis of the above PDE, we

[1] The Kantorovich metric is more often known as the Wasserstein metric, it is unfortunate that it wasn't named after its inventor.

refer the reader to [18]. In what's to come, we briefly outline several important mathematical tools, provide detailed proofs of several major results mentioned above, as well as discuss their implications and various future directions.

3. Background: space of probability measures over (X, d), metrics, transport

Consider the space of probability measures over a metric space (X, d). We define a base of open neighborhoods of $\mu \in \mathscr{P}(X)$ as follows:
Fix $f_1, \dots, f_k$ bounded continuous functions on X and $\epsilon > 0$. Then,

$$V_\mu(f_1, \dots, f_k, \epsilon) = \left\{ \nu \in \mathscr{P}(X) : \left| \int f_i \, d\nu - \int f_i \, d\mu \right| < \epsilon, i = 1, \dots, k \right\}$$

for any such collection of k functions form a base for the **weak topology** on $\mathscr{P}(X)$. Convergence in this topology is the weak convergence, with a very important characterization via the Portmanteau theorem:

Theorem 3.1. (Portmanteau theorem) The following are equivalent:

a) $\mu_n \Rightarrow \mu$ (μ_n converges weakly to μ)

b) $\lim_{n \to \infty} \displaystyle\int g \, d\mu_n = \int g \, d\mu$, for all g uniformly continuous and bounded.

c) $\limsup_n \mu_n(C) \leq \mu(C)$, for all C closed.

d) $\liminf_n \mu_N(U) \geq \mu(U)$, for all U open.

e) $\lim_{n \to \infty} \mu_n(A) = \mu(A)$, for all A Borel such that $\mu(\partial A) = 0$.

Crucial for our purposes, statement *b)* in particular applies for all g bounded Lipschitz continuous.

Theorem 3.2. $\mathscr{P}(X)$ can be metrized as a **separable** metric space if and only if X can. Furthermore, $\mathscr{P}(X)$ is compact if and only if X is compact.

Stemming from this, we observe the rich structure of $\mathscr{P}(X)$ depending on X's structure. If the underlying space X is Polish, Prokhorov's theorem (omitted) gives rise to the Lévy–Prokhorov metric:

Definition 3.3. (Lévy–Prokhorov metric) Take A a Borel set, consider $A^\epsilon := \{x : d(x, A) < \epsilon\}$ is open, as $d(\cdot, A)$ is continuous in its first parameter. We define π as:

$$\pi(\mu, \nu) = \inf \{\epsilon > 0 : \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ and } \nu(A) \leq \mu(A^\epsilon) + \epsilon\}$$

One can easily show that $\pi(\mu_n, \mu) \to 0 \implies \mu_n \Rightarrow \mu$ by taking limit suprema and using the Portmanteau theorem. Thus, the Lévy–Prokhorov metric metrizes the weak topology, and therefore weak convergence.

There are many other metrics that metrize weak convergence but for our purposes, we consider the bounded Lipschitz distance and the Kantorovich metric.

Definition 3.4. (Bounded Lipschitz distance) For any two probabiility measures μ, ν, the bounded Lipschitz distance between μ and ν is defined by:

$$d_{BL}(\mu, \nu) = \sup \left\{ \left| \int f \, d\mu - \int f \, d\nu \right| : \|f\|_\infty + \|f\|_{Lip} \leq 1 \right\}$$

where the (homogeneous) Lipschitz norm is defined as $\|f\|_{Lip} = \sup\limits_x \sup\limits_{h \neq 0} \dfrac{|f(x+h) - f(x)|}{|h|}$, a common tool in Sobolev space theory. Intuitively, one can think of the Lipschitz norm as the Lipschitz semi-norm (which returns the Lipschitz constant of f) with additional boundary conditions.

An important metric on $\mathscr{P}(X)$ arising from the theory of optimal transport is the Kantorovich metric.

Definition 3.5. (p-Kantorovich metric) Let (X, d) Polish, $p \in [1, \infty)$. For any two probability measures μ, ν, the Kantorovich metric of order p is defined by:

$$\begin{aligned} W_p(\mu, \nu) &= \left(\inf_{\pi \in \Pi(\mu,\nu)} \int d(x,y)^p \, d\pi(x,y) \right)^{1/p} \\ &= \inf \left\{ \left[\mathbb{E} d(X,Y)^p \right]^{1/p}, \operatorname{law}(X) = \mu, \operatorname{law}(Y) = \nu \right\} \end{aligned}$$

The metric induces the Kantorovich space defined as:

$$\mathscr{P}_p(X) \coloneqq \left\{ \mu \in \mathscr{P}(X) : \int d(x_0, x)^p \, d\mu(x) < \infty \right\}$$

when $x_0 \in X$ is arbitrary. The space $P_p(X)$ does not depend on the choice of x_0.

Remarkably, the p-Kantorovich metric metrizes $\mathscr{P}_p(X)$, i.e. W_p fully characterizes weak convergence. Furthermore, they have a rich duality theory due to the Kantorovich duality, and in particular, for $p = 1$, the Kantorovich-Rubinstein duality gives:

$$W_1(\mu, \nu) = \sup \left\{ \left| \int f \, d\mu - \int f \, d\nu \right| : \|f\|_{Lip} \leq 1 \right\}$$

For more related to optimal transport theory, we refer the reader to [18]. When (X, d) corresponds to $(\mathbf{R}^D, \|\cdot\|_2)$, we recover the original setup of neural networks.

4. Results: statics and dynamics

As pointed out previously, there is an intricate connection between the risk $R_N(\boldsymbol{\theta})$ and the risk $R(\rho)$, as we will state and outline the proof below:

Proposition 4.1. Assume either:

a) $\inf_\rho R(\rho)$ is attained by ρ_* such that $\displaystyle\int U(\boldsymbol{\theta}, \boldsymbol{\theta}) \, d\rho_*(\boldsymbol{\theta}) \leq K$

b) $\exists \epsilon_0 > 0$ such that $\forall \rho \in \mathscr{P}(\mathbf{R}^D)$ such that $R(\rho) \leq \inf_\rho R(\rho) + \epsilon_0$, we have that $\displaystyle\int U(\boldsymbol{\theta}, \boldsymbol{\theta}) \, d\rho_*(\boldsymbol{\theta}) \leq K$

Then,

$$\left| \inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) - \inf_\rho R(\rho) \right| \leq \frac{K}{N}$$

Furthermore, if V and U are jointly continuous with U bounded below, $\rho_* \in \mathscr{P}(\mathbf{R}^D)$ is a global minimum of R if $\inf_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_*) > -\infty$ and

$$\operatorname{supp}(\rho_*) \subseteq \arg\min_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_*)$$

Proof. For any $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i \leq N}$, we have $R_N(\boldsymbol{\theta}) \geq \inf_\rho R(\rho)$ since $R_N(\boldsymbol{\theta}) = R(\rho)$ for $\rho = N^{-1} \sum \delta_{\boldsymbol{\theta}_i}$. Now, let $\rho_* \in \mathscr{P}(\mathbf{R}^D)$ be such that $R(\rho_*) = R_*$ under assumption a), or $R(\rho_*) \leq R_* + \epsilon$ under assumption b). We thus obtain:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) - R(\rho_*) &= \frac{1}{N} \left(\int U(\boldsymbol{\theta}, \boldsymbol{\theta}) \, d\rho_*(\boldsymbol{\theta}) - \iint U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \, d\rho_*(\boldsymbol{\theta}_1) d\rho_*(\boldsymbol{\theta}_2) \right) \\ &\leq \frac{1}{N} \int U(\boldsymbol{\theta}, \boldsymbol{\theta}) \, d\rho_*(\boldsymbol{\theta}) \\ &\leq K/N \end{aligned}$$

where we've applied Fubini-Tonelli twice to interchange expectations w.r.t. $\mathbb{P}_{\boldsymbol{x}}$, the underlying data's distribution $\boldsymbol{x}$-marginal, and w.r.t. $\boldsymbol{\theta}$ in the term $\mathbb{E}_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta})$. The inequality follows from:

$$\iint U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \, d\rho_*(\boldsymbol{\theta}_1) d\rho_*(\boldsymbol{\theta}_2) = \mathbb{E} \left[\left(\int \sigma_*(\boldsymbol{x}; \boldsymbol{\theta}) \, d\rho_*(\boldsymbol{\theta}) \right)^2 \right] \geq 0$$

We omit the proof of the conditions of global minimum. ∎

We now present the mild assumptions and main theorem that establishes the connection between SGD dynamics and the distributional dynamics PDE described previously.

Assumption 4.2. The map $t \mapsto \xi(t)$ is bounded Lipschitz, i.e. $\|\xi\|_{\infty}, \|\xi\|_{Lip} \le K_1, \displaystyle\int_{\mathbf{R}^+} \xi(t)\, dt = \infty$.

Assumption 4.3. The neural network activation $(\boldsymbol{x}, \boldsymbol{\theta}) \mapsto \sigma_*(\boldsymbol{x}, \boldsymbol{\theta})$ is bounded, and has sub-Gaussian gradient: $\|\sigma_*\|_{\infty}, \|\nabla_{\boldsymbol{\theta}} \sigma_*(\boldsymbol{X}, \boldsymbol{\theta})\|_{\psi_2} \le K_2, |y_k| \le K_2$, where $\|\cdot\|_{\psi_2}$ is the Orlicz norm on the Orlicz space generated by the N-function $z \mapsto \exp\left\{z^2\right\} - 1$. We refer the reader to[14] and [10] for the characterization of Orlicz spaces of exponential type and the sub-Gaussian property.

Assumption 4.4. Gradients $\nabla V, \nabla_1 U$ are bounded, Lipschitz continuous, that is to say that $\|\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta})\|_2$, and $\|\nabla_1 U(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})\| \le K_3$. Furthermore,

$$\|\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta'})\|_2 \le K_3 \|\boldsymbol{\theta} - \boldsymbol{\theta'}\|_2$$
$$\left\|\nabla_1 U(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) - \nabla_1 U(\boldsymbol{\theta'_1}, \boldsymbol{\theta'_2})\right\|_2 \le K_3 \left\|(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) - (\boldsymbol{\theta'_1}, \boldsymbol{\theta'_2})\right\|_2$$

The following theorem quantifies the extent to which the SGD dynamics converges to the PDE dynamics previously foreshadowed.

Theorem 4.5. (Convergence of SGD to PDE) Assume that **Assumptions 4.2, 4.3, 4.4** hold. For an initial distribution $\rho_0 \in \mathscr{P}(\mathbf{R}^D)$, let the random sample $\left(\boldsymbol{\theta}_{\boldsymbol{i}}^{\boldsymbol{0}}\right)_{i \le N} \sim_{iid} \rho_0$ be the weight initialization to our neural network, and let our step size be $s_k = \epsilon \xi(k\epsilon)$. For $t \ge 0$, let ρ_t solve the **distributional dynamics (DD)**:

$$\partial_t \rho_t = 2\xi(t) \nabla_{\boldsymbol{\theta}} \left(\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)\right)$$
$$\Psi(\boldsymbol{\theta}; \rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \boldsymbol{\theta'})\, d\rho(\theta')$$

Then, for a fixed $t \ge 0$, we have that $\hat{\rho}^{(N)}_{\lfloor t/\epsilon \rfloor} \Rightarrow \rho_t$ a.s. along any sequence $(N, \epsilon = \epsilon(N))$ such that $N \to \infty, \epsilon_N \to 0, \frac{N}{\log(N/\epsilon_N)} \to \infty, \epsilon_N \log(N/\epsilon_N) \to 0$. Furthermore, $\exists C \in \mathbf{R} = C(K_1, K_2, K_3)$ depending on the constants in the assumptions such that for any $f : \mathbf{R}^D \to \mathbf{R}$ with $\|f\|_{\infty}, \|f\|_{Lip} \le 1, \epsilon \le 1$, we have

$$\sup_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left| N^{-1} \sum_{i=1}^{N} f(\boldsymbol{\theta}_i^k) - \int f(\boldsymbol{\theta})\, d\rho_{k\epsilon}(\boldsymbol{\theta}) \right| \le C e^{CT} \operatorname{err}_{N,D}(z)$$
$$\sup_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left| R_N(\boldsymbol{\theta}^k) - R(\rho_{k\epsilon}) \right| \le C e^{CT} \operatorname{err}_{N,D}(z)$$

where

$$\operatorname{err}_{N,D}(z) = \sqrt{1/N \vee \epsilon} \left(\sqrt{D + \log(N/\epsilon)} + z\right)$$

is an error term that quantifies the accuracy of the DD.

Remark 4.6. We note here that from Assumptions 4.2 and 4.4, the existence and uniqueness of weak solutions ρ_t of the DD given some time $t \ge 0$ is guaranteed by [13]. The evolution in $\mathscr{P}(\mathbf{R}^D)$ is to be interpreted in weak sense, i.e. ρ_t is a weak solution of DD if for any bounded differentiable function $\phi : \mathbf{R}^D \to \mathbf{R}$ with bounded gradient, we have:

$$d_t \langle \rho_t, \phi \rangle = -2\xi(t) \int \langle \nabla \phi(\boldsymbol{\theta}), \nabla \Psi(\boldsymbol{\theta}; \rho_t) \rangle\, d\rho_t(\boldsymbol{\theta})$$

We now introduce a nonlinear dynamical system (which from here onwards will be referred to as ND for simplicity's sake) that will be essential to proving SGD convergence to DD. Let $(\bar{\boldsymbol{\theta}}_i^t)_{i \le N, t \in \mathbf{R}^+}$ be trajectories initialized by $\bar{\boldsymbol{\theta}}_i^0 = \boldsymbol{\theta}_i^0$, the same initialization from ρ_0, and for $t \ge 0$, we have:

$$\bar{\boldsymbol{\theta}}_i^t = \boldsymbol{\theta}_i^0 - 2 \int_0^t \xi(s) \nabla \Psi(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\, ds$$

where $\rho_s = \bar{\boldsymbol{\theta}}_{i\#}^s P_s$, the pushforward measure of the distribution of parameters at time s onto $\mathbf{R}$. With boundary conditions $\bar{\boldsymbol{\theta}}_i^0 \sim_{iid} \rho_0$, this is indeed a PDE with existence and uniqueness of solution guaranteed in a similar fashion to the DD, and with pushforwards of the solutions $\rho_t = \bar{\boldsymbol{\theta}}_{i\#}^t P_t$ satisfying the DD. In the ND, the individual trajectories $(\bar{\boldsymbol{\theta}}_1^t), \dots, \bar{\boldsymbol{\theta}}_n^t$ are iid and thus, we have that a.s.,

$$N^{-1} \sum_{i=1}^{N} \delta_{\bar{\boldsymbol{\theta}}_i^t} \Rightarrow \rho_t$$

Remark 4.7. We previously used ρ_t to denote a measure in parameter space $\mathbf{R}^D$, however in the proof of convergence that follows it will denote the pushforward, described above, i.e. a measure on $\mathbf{R}$.

Before proceeding, we establish several useful results in the following lemma, which we state without proof.

Lemma 4.8. Assume 4.2, 4.3, 4.4 hold, and let (ρ_t) be the solution to the DD, and $(\bar{\boldsymbol{\theta}}_i^t)$ the solution to the ND. Then, $t \mapsto \bar{\boldsymbol{\theta}}_i^t$ is K_1K_3-Lipschitz continuous in Euclidean 2-norm metric, i.e. in the Euclidean topology, $t \mapsto \rho_t$ is K_1K_3-Lipschitz continuous in W_2 Kantorovich metric, i.e. in the weak topology on $\mathscr{P}(\mathbf{R}^D)$.

We are now able to prove Theorem 4.5. The proof roadmap goes as follows: we first define a continuous trajectory in $\mathbf{R}^D$ similar to SGD dynamics and control the difference between our trajectory and the discrete weights obtained from SGD. Then, we control the risk difference between SGD weights and ρ_t, the solution to the DD via our trajectory, as well as the expectation difference of bounded Lipschitz functions of SGD weights and ρ_t, again via our trajectory. Weak convergence follows naturally from the Portmanteau theorem.

5. Proof of convergence to the Distributional Dynamics

Proof. (of 4.5) We use the term K to denote the maximum of constants in terms of K_1, K_2, K_3 which arise in many parts of the proof, for ease of factoring. Letting $\boldsymbol{z}_k = (\boldsymbol{x}_k, y_k)$ be the k-th data point, we define the following:

$$F_i(\boldsymbol{\theta}; \boldsymbol{z}_k) = (y_k - \hat{y}(\boldsymbol{x}_k; \boldsymbol{\theta}))\nabla_{\boldsymbol{\theta}_i}\sigma_*(\boldsymbol{x}_k; \boldsymbol{\theta}_i), \quad \boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i \leq N} \in \mathbf{R}^{D \times N}$$

$$G(\boldsymbol{\theta}; \rho) = -\nabla\Psi(\boldsymbol{\theta}, \rho) = -\nabla V(\boldsymbol{\theta}) - \int \nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\, d\rho(\boldsymbol{\theta}') \quad \boldsymbol{\theta} \in \mathbf{R}^D$$

Since gradients ∇V and $\nabla_1 U$ are bounded, we have that $\|G(\boldsymbol{\theta}; \rho)\|_2 \leq K$, $\|G(\boldsymbol{\theta}_1; \rho) - G(\boldsymbol{\theta}_2; \rho)\|_2 \leq K\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$. We also have Lipschitz continuity in the measure parameter of G:

$$\|G(\boldsymbol{\theta}; \rho_1) - G(\boldsymbol{\theta}; \rho_2)\|_2 = \left\| \int \nabla_1 U(\boldsymbol{\theta}, \boldsymbol{\theta}')\, d(\rho_1 - \rho_2)(\boldsymbol{\theta}') \right\|_2 \leq K d_{BL}(\rho_1, \rho_2) \tag{1}$$

where d_{BL} is the bounded Lipschitz metric on $\mathscr{P}(\mathbf{R}^D)$.
Recalling SGD dynamics, we rewrite with our newly defined expressions to yield:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + 2\epsilon\xi(k\epsilon)F_i(\boldsymbol{\theta}_i^k; \boldsymbol{z}_{k+1}) \tag{2}$$

thus,

$$\boldsymbol{\theta}_i^k = \boldsymbol{\theta}_i^0 + 2\epsilon \sum_{l=0}^{k-1} \xi(l\epsilon)F_i(\boldsymbol{\theta}_i^l; \boldsymbol{z}_{l+1}) \tag{3}$$

Let $[t] = \epsilon\lfloor \frac{t}{\epsilon} \rfloor$. Our ND trajectory defined previously is then:

$$\bar{\boldsymbol{\theta}}_i^t = \boldsymbol{\theta}_i^0 + 2\int_0^t \xi(s)G(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\, ds \tag{4}$$

We can thus control the difference between the above SGD dynamics and the ND.

Lemma 5.1. Assume Assumptions 4.2, 4.3, and 4.4. There exists $K = K(K_1, K_2, K_3)$ such that $\forall T \geq 0$,

$$\max_{i \leq N} \sup_{k \in [T/\epsilon] \cap \mathbf{N}} \left\| \boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\epsilon} \right\|_2 \leq K e^{KT} \sqrt{1/N \vee \epsilon} \left(\sqrt{D + \log(N(T/\epsilon \vee 1))} + z \right) \tag{5}$$

with probability at least $1 - e^{-z^2}$.

Proof. Consider $t \in \mathbf{N}\epsilon \cap [0, T]$. We have by subtracting (4) from (3) that:

$$\left\| \boldsymbol{\theta}_i^{t/\epsilon} - \bar{\boldsymbol{\theta}}_i^t \right\|_2 = \left\| \boldsymbol{\theta}_i^0 + 2\epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) F_i(\boldsymbol{\theta}_i^k; \boldsymbol{z}_{l+1}) - \left(\boldsymbol{\theta}_i^0 + 2 \int_0^t \xi(s) G(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\, ds \right) \right\|_2 \tag{6}$$

$$= 2 \left\| \int_0^t \xi(s) G(\bar{\boldsymbol{\theta}}_i^s; \rho_s)\, ds - \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) \right\|_2 \tag{7}$$

$$\begin{aligned} &\leq 2 \int_0^t \left\| \xi(s) G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) \right\|_2 ds \\ &\quad + 2 \int_0^t \left\| \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) - \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{\lfloor s/\epsilon \rfloor}; \rho_{[s]}) \right\|_2 ds \\ &\quad + 2 \left\| \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) \left(F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \right) \right\|_2 \end{aligned} \tag{8}$$

$$:= 2E_1^i(t) + 2E_2^i(t) + 2E_3^i(t) \tag{9}$$

where in (8), we introduced and subtracted terms $\xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]})$ and $\xi([s]) G(\bar{\boldsymbol{\theta}}_i^{\lfloor s/\epsilon \rfloor}; \rho_{[s]})$ within integrals, and applied Minkowski's inequality. The last term in (8) follows from:

$$\int_0^t \xi([s]) G(\boldsymbol{\theta}_i^{\lfloor s/\epsilon \rfloor}; \rho_{[s]})\, ds \leq \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \epsilon \tag{10}$$

as operations $\lfloor \cdot \rfloor$ and $[\cdot]$ discretizes the integral on the left-hand side in s-lengths of at most ϵ, and recombining the sum in (10) with the sum in (7) yields the last term in (8). We now control terms E_i through various techniques. For E_1^i, consider:

$$\begin{aligned} \left\| \xi(s) G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) \right\|_2 \leq\ & \left\| G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) \left(\xi(s) - \xi([s]) \right) \right\|_2 \\ & + \left\| \xi([s]) \left(G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) - G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_s) \right) \right\|_2 \\ & + \left\| \xi([s]) \left(G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_s) - G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) \right) \right\|_2 \end{aligned} \tag{11}$$

$$\leq K\epsilon \left\| G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) \right\|_2 + \xi([s]) K \left\| \bar{\boldsymbol{\theta}}_i^s - \bar{\boldsymbol{\theta}}_i^{[s]} \right\|_2 + \xi([s]) K d_{BL}(\rho_s, \rho_{[s]}) \tag{12}$$

$$\leq K\epsilon \tag{13}$$

where we applied Minkowski's inequality in (11), used Lipschitz continuity w.r.t. $\boldsymbol{\theta}$ and ρ of $G(\boldsymbol{\theta}, \rho)$ from (1) in (12), and (13) follows from bounds on G and Lemma 4.8, and

$$d_{BL}(\rho_s, \rho_{[s]}) \leq W_2(\rho_s, \rho_{[s]}) \leq \left(\int \left\| \bar{\boldsymbol{\theta}}_i^s - \bar{\boldsymbol{\theta}}_i^{[s]} \right\|_2^2 d\gamma(\bar{\boldsymbol{\theta}}_i^s, \bar{\boldsymbol{\theta}}_i^{[s]}) \right)^{1/2} \leq K_1 K_3 |s - [s]| \leq K_1 K_3 \epsilon$$

where γ is some arbitrary coupling of the laws of $\bar{\boldsymbol{\theta}}_i^s$ and $\bar{\boldsymbol{\theta}}_i^{[s]}$. Therefore, we have that

$$E_1^i(t) \leq t \sup_{s \in [0,t]} \left\{ \left\| \xi(s) G(\bar{\boldsymbol{\theta}}_i^s; \rho_s) - \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) \right\|_2 \right\} \tag{14}$$

$$\leq Kt\epsilon \tag{15}$$

To control E_2^i, consider:

$$E_2^i(t) = \int_0^t \left\| \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) - \xi([s]) G(\bar{\boldsymbol{\theta}}_i^{\lfloor s/\epsilon \rfloor}; \rho_{[s]}) \right\|_2 ds \tag{16}$$

$$\leq K \int_0^t \left\| G(\bar{\boldsymbol{\theta}}_i^{[s]}; \rho_{[s]}) - G(\bar{\boldsymbol{\theta}}_i^{\lfloor s/\epsilon \rfloor}; \rho_{[s]}) \right\|_2 ds \tag{17}$$

$$\leq K^2 \int_0^t \left\| \bar{\boldsymbol{\theta}}_i^{[s]} - \boldsymbol{\theta}_i^{\lfloor s/\epsilon \rfloor} \right\|_2 ds \tag{18}$$

where we controlled ξ by its upper bound from Assumption 4.2 in (17) and Lipschitz continuity in the first variable of G in (18). We control E_3^i via a variant of Azuma-Hoeffding's inequality for martingales (henceforth will be referred to as Azuma-Hoeffding, information regarding this is displayed in the Appendix section). Let $\mathcal{F}_k$ be the sub-sigma algebra generated by $(\boldsymbol{\theta}_i^0)_{i \leq N}$ and $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_k$, i.e. the knowledge of the first k data points. We have that:

$$\mathbb{E}\left[F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) | \mathcal{F}_k\right] = \mathbb{E}\left[y_{k+1} \nabla_{\boldsymbol{\theta}_i} \sigma_*(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}_i^k) | \mathcal{F}_k\right] - \mathbb{E}\left[\hat{y}(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}^k) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\boldsymbol{x}_{k+1}; \boldsymbol{\theta}_i^k) | \mathcal{F}_k\right] \tag{19}$$

$$= -\nabla_{\boldsymbol{\theta}_i} V(\boldsymbol{\theta}_i^k) - N^{-1} \sum_{j=1}^N \nabla_1 U(\boldsymbol{\theta}_i^k; \boldsymbol{\theta}_j^k) \tag{20}$$

$$= G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) \tag{21}$$

where we recall $\hat{\rho}_k^{(N)} = N^{-1} \sum_i \delta_{\boldsymbol{\theta}_i^k}$ is the empirical distribution of the realizations of $\boldsymbol{\theta}_i^k$. Therefore, we observe:

$$E_3^i(t) = \left\| \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) \left\{ F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) + G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - \mathbb{E}\left[F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) | \mathcal{F}_k\right] \right\} F \right\|_2 \tag{22}$$

$$\leq \left\| \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) \left[G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \right] \right\|_2 + \left\| \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) \boldsymbol{Z}_k^i \right\|_2 \tag{23}$$

where $\boldsymbol{Z}_k^i := F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) - \mathbb{E}\left[F_i(\boldsymbol{\theta}^k; \boldsymbol{z}_{k+1}) | \mathcal{F}_k\right]$. Defining $Q_1^i(t) := \left\| \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) \boldsymbol{Z}_k^i \right\|_2$, we apply Azuma-Hoeffding to yield:

$$\mathbb{P}\left(\max_{k \in [0, t/\epsilon] \cap \mathbf{N}} Q_1^i(k\epsilon) \geq K\sqrt{t\epsilon}(\sqrt{D} + u)) \right) \leq e^{-u^2} \tag{24}$$

$$\mathbb{P}\left(\max_{i \leq N} \max_{k \in [0, t/\epsilon] \cap \mathbf{N}} Q_1^i(k\epsilon) \leq K\sqrt{t\epsilon}(\sqrt{D + \log N} + z)) \right) \geq 1 - e^{-z^2} \tag{25}$$

where the second inequality follows from taking union bounds over all naturals $i \leq N$. The conditions for Azuma-Hoeffding follows from Assumption 4.3, since $\xi(k\epsilon)\boldsymbol{Z}_k^i$ is sub-Gaussian. Defining $E_{3,0}^i(t) :=$

$\left\| \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) \left[G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \right] \right\|_2$, we can consider the difference within the sum:

$$\left\| G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \right\|_2 = \left\| -\frac{1}{N} \sum_{j=1}^{N} \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k)) + \int \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}') \, d\rho_{k\epsilon}(\boldsymbol{\theta}')) \right\|_2 \tag{26}$$

$$= \left\| \frac{1}{N} \sum_{j=1}^{N} \left\{ \nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k) - \mathbb{E}_{\bar{\boldsymbol{\theta}}} \left[\nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) \right] \right\} \right\|_2 \tag{27}$$

$$\begin{aligned} &\leq \left\| \frac{1}{N} \sum_{j=1}^{N} \left[\nabla_1 U(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k) - \nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) \right] \right\|_2 \\ &\quad + \left\| \frac{1}{N} \sum_{j=1}^{N} \left\{ \nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) - \mathbb{E}_{\bar{\boldsymbol{\theta}}} \left[\nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) \right] \right\} \right\|_2 \end{aligned} \tag{28}$$

$$\leq \frac{K}{N} \sum_{j=1}^{N} \left\| \boldsymbol{\theta}_j^k - \bar{\boldsymbol{\theta}}_j^{k\epsilon} \right\|_2 + Q_2^i(k\epsilon) + \frac{K}{N} \tag{29}$$

where we used the Lipschitz continuity of $\nabla_1 U$ and Minkowski's inequality in (28). We define $Q_2^i(k\epsilon) := \left\| \frac{1}{N} \sum_{j \neq i} \left\{ \nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) - \mathbb{E}_{\bar{\boldsymbol{\theta}}} \left[\nabla_1 U(\boldsymbol{\theta}_i^k, \bar{\boldsymbol{\theta}}_j^{k\epsilon}) \right] \right\} \right\|_2$, since for $j = i$, we invoke the boundedness of $\nabla_1 U$ to yield constant K, obtaining the term K/N in (29). Since fixing k, $(\bar{\boldsymbol{\theta}}_j^{k\epsilon})_{j \leq N, j \neq i}$ are independent of $\boldsymbol{\theta}_i^k$, with $\nabla_1 U$ bounded, we meet the conditions of Azuma-Hoeffding, applying the inequality and union bounds yields:

$$\mathbb{P} \left(\max_{k \in [0, t/\epsilon] \cap \mathbf{N}} Q_2^i(k\epsilon) \geq K \sqrt{1/N} (\sqrt{D} + u)) \right) \leq 1 - e^{-u^2} \tag{30}$$

$$\mathbb{P} \left(\max_{i \leq N} \max_{k \in [0, t/\epsilon] \cap \mathbf{N}} Q_2^i(k\epsilon) \leq K \sqrt{1/N} (\sqrt{D + \log(N(t/\epsilon \vee 1))} + z)) \right) \geq 1 - e^{-z^2} \tag{31}$$

We are now able to control E_3^i as follows:

$$E_3^i(t) \leq E_{3,0}^i(t) + Q_1^i(t) \tag{32}$$

$$\leq \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) \left\| G(\boldsymbol{\theta}_i^k; \hat{\rho}_k^{(N)}) - G(\boldsymbol{\theta}_i^k; \rho_{k\epsilon}) \right\|_2 + Q_1^i(t) \tag{33}$$

$$\leq \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) \left[\frac{K}{N} \sum_{j=1}^{N} \left\| \boldsymbol{\theta}_j^k - \bar{\boldsymbol{\theta}}_j^{k\epsilon} \right\|_2 + \frac{K}{N} + Q_2^i(k\epsilon) \right] + Q_1^i(t) \tag{34}$$

$$= \epsilon \frac{K}{N} \sum_{k=0}^{t/\epsilon-1} \sum_{j=1}^{N} \left\| \boldsymbol{\theta}_j^k - \bar{\boldsymbol{\theta}}_j^{k\epsilon} \right\|_2 + \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) \frac{K}{N} + \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) Q_2^i(k\epsilon) + Q_1^i(t) \tag{35}$$

$$\leq \frac{K}{N} \sum_{j=1}^{N} \int_0^t \left\| \boldsymbol{\theta}_j^{\lfloor s/\epsilon \rfloor} - \bar{\boldsymbol{\theta}}_j^{[s]} \right\|_2 ds + \frac{Kt}{N} + \epsilon \sum_{k=0}^{t/\epsilon-1} \xi(k\epsilon) Q_2^i(k\epsilon) + Q_1^i(t) \tag{36}$$

$$\tag{37}$$

where (33) follows from Minkowski, we substituted our control for (26) in (34), we absorbed $\xi(k\epsilon)$ into K in (35), we upper-bounded the discrete sum over k by an integral and modified the appropriate arguments in (36), finally passing it inside the finite sum over j by linearity. To control the last two terms, we define

$Q(t) = \epsilon \sum_{k=0}^{t/\epsilon - 1} \xi(k\epsilon) Q_2^i(k\epsilon) + Q_1^i(t)$ and use the bounds from Azuma-Hoeffding:

$$Q(t) \leq Kt \cdot \max_{i \leq N} \max_{k \in [0, t/\epsilon] \cap \mathbf{N}} Q_2^i(k\epsilon) + \max_{i \leq N} Q_1^i(t) \tag{38}$$

$$\leq K\sqrt{t\epsilon}\left(\sqrt{D + \log N} + z\right) + Kt\sqrt{1/N}\left(\sqrt{D + \log\left(N(t/\epsilon \vee 1)\right)} + z\right) \tag{39}$$

$$\leq K\left(\sqrt{t} \vee t\right)\sqrt{1/N \vee \epsilon}\left[\sqrt{D + \log\left(N(t/\epsilon \vee 1)\right)} + z\right] \tag{40}$$

with probability at least $1 - e^{-z^2}$. Back to (5), we define:

$$\Delta(t; N, \epsilon) = \max_{i \leq N} \sup_{k \in [0, t/\epsilon] \cap \mathbf{N}} \left\| \boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\epsilon} \right\|_2 \tag{41}$$

Using the bounds we found for $E_i, i = 1, 2, 3$, we have that:

$$\Delta(t; N, \epsilon) \leq 2(E_1^i(t) + E_2^i(t) + E_3^i(t)) \tag{42}$$

$$\leq Kt\epsilon + K\int_0^t \Delta(s; N, \epsilon)\, ds + \frac{Kt}{N} + Q(t) \tag{43}$$

Gronwall's inequality thus yields:

$$\Delta(t; N, \epsilon) \leq \exp\left\{\int_0^t K\, ds\right\}\left(K\epsilon + \frac{K}{N} + KQ(t)\right) \tag{44}$$

Applying the bounds from (40) and absorbing constants into the term K yields the desired result. ∎

Lemma 5.2. Assuming Assumptions 4.2, 4.3, and 4.4, we have:

$$\max_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left| R_N(\bar{\boldsymbol{\theta}}^{k\epsilon}) - R_N(\boldsymbol{\theta}^k) \right| \leq K \cdot \max_{i \leq N} \max_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left\| \boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\epsilon} \right\|_2 \tag{45}$$

Furthermore, for $f : \mathbf{R}^D \to \mathbf{R}$ bounded Lipschitz, we have:

$$\max_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left| \mathbb{E}\left[f(\bar{\boldsymbol{\theta}}_i^{k\epsilon})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}_i^k)\right] \right| \leq K \cdot \max_{k \in [0, T/\epsilon] \cap \mathbf{N}} \left\| \boldsymbol{\theta}_i^k - \bar{\boldsymbol{\theta}}_i^{k\epsilon} \right\|_2 \tag{46}$$

Proof. Let $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i', \ldots, \boldsymbol{\theta}_N)$ be identical to $\boldsymbol{\theta}$, except with the i-th component resampled. We observe:

$$\left| R_N(\boldsymbol{\theta}) - R_N(\boldsymbol{\theta}') \right| \leq \frac{1}{N}\left| V(\boldsymbol{\theta}_i) - V(\boldsymbol{\theta}_i') \right| + \frac{1}{N^2}\left| U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_i') \right| + \frac{2}{N^2}\sum_{j \neq i}\left| U_(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) - U(\boldsymbol{\theta}_i', \boldsymbol{\theta}_j) \right| \tag{47}$$

$$\leq \frac{K}{N}\left(\left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}_i' \right\|_2 \wedge 1 \right) \tag{48}$$

The conclusion is immediate since differences in all N dimensions of $\boldsymbol{\theta} \in \mathbf{R}^{D \times N}$ reduces the denominator of (47), and maximizing over all $i \leq N$ ensures the inequality stays valid, thus we recover (45). To see (46), we observe:

$$\left| \mathbb{E}\left[f(\bar{\boldsymbol{\theta}}_i^{k\epsilon})\right] - \mathbb{E}\left[f(\boldsymbol{\theta}_i^k)\right] \right| \leq \left| \int f d\left(\bar{\boldsymbol{\theta}}_i^{k\epsilon}{}_{\#} P_{k\epsilon}\right) - \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{\theta}_i^k) \right| \tag{49}$$

$$= \left| \frac{1}{N}\sum_{i=1}^N \left[f(\boldsymbol{\theta}_i^k) - \int f d\left(\bar{\boldsymbol{\theta}}_i^{k\epsilon}{}_{\#} P_{k\epsilon}\right) \right] \right| \tag{50}$$

$$\leq \frac{1}{N}\int \sum_{i=1}^N \left| f(\boldsymbol{\theta}_i^k) - f(\bar{\boldsymbol{\theta}}_i^{k\epsilon}) \right| d\left(\bar{\boldsymbol{\theta}}_i^{k\epsilon}{}_{\#} P_{k\epsilon}\right) \tag{51}$$

$$\leq K\left\| \bar{\boldsymbol{\theta}}_i^{k\epsilon} - \boldsymbol{\theta}_i^k \right\|_2 \tag{52}$$

where we applied Jensen's inequality in (51) and used the Lipschitz condition of f in (52). Maximizing over k gives the desired statement. ∎

Lemma 5.3. Assuming Assumptions 4.2, 4.3, and 4.4, we have:

$$\max_{k\in[0,T/\epsilon]\cap\mathbf{N}} \left|R_N(\bar{\boldsymbol{\theta}}^{k\epsilon}) - R(\rho_{k\epsilon})\right| \leq K\sqrt{1/N}\left(\sqrt{D+\log\left(N(T/\epsilon\vee 1)\right)}+z\right) \tag{53}$$

with probability at least $1-e^{-z^2}$.

Proof. By Azuma-Hoeffding bound, we have that:

$$\max_{k\in[0,T/\epsilon]\cap\mathbf{N}} \left|R_N(\bar{\boldsymbol{\theta}}^{k\epsilon}) - \mathbb{E}\left[R_N(\bar{\boldsymbol{\theta}}^{k\epsilon})\right]\right| \leq K\sqrt{1/N}\left(\sqrt{D+\log\left(N(T/\epsilon\vee 1)\right)}+z\right) \tag{54}$$

with probability at least $1-e^{-z^2}$. Also,

$$\left|\mathbb{E}\left[R_N(\bar{\boldsymbol{\theta}}^t)\right] - R(\rho_t)\right| = \frac{1}{N}\left|\int U(\boldsymbol{\theta},\boldsymbol{\theta})\,d\rho_t(\boldsymbol{\theta}) - \int U(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\,d\rho_t(\boldsymbol{\theta}_1)d\rho_t(\boldsymbol{\theta}_2)\right| \leq \frac{K}{N} \tag{55}$$

giving us the desired conclusion. ∎

Remark 5.4. A similar statement for the difference $\left|\mathbb{E}\left[f(\bar{\boldsymbol{\theta}}_i^{k\epsilon})\right] - \mathbb{E}_{\rho_{k\epsilon}}[f]\right|$ holds true with the same upper bounds with probability at least $1-e^{-z^2}$, following from the same Azuma-Hoeffding approach. We leave the proof to the reader.

Controlling the upper bound in Lemma 5.2 with Lemma 5.1, using this new estimate with Lemma 5.3, we are able to control the risk difference between our SGD weights and the DD's generalized risk, yielding us:

$$\sup_{k\in[0,T/\epsilon]\cap\mathbf{N}} \left|R_N(\boldsymbol{\theta}^k) - R(\rho_{k\epsilon})\right| \leq Ke^{KT}\sqrt{1/N\vee\epsilon}\left[\sqrt{D+\log(N/\epsilon)}+z\right] \tag{56}$$

with probability $1-e^{-z^2}$. Applying the formulation for bounded Lipschitz functions f in Lemma 5.2 with Remark 5.4, we yield the statement:

$$\sup_{k\in[0,T/\epsilon]\cap\mathbf{N}} \left|\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^k) - \int f\,d\rho_{k\epsilon}\right| \leq Ke^{KT}\sqrt{1/N\vee\epsilon}\left[\sqrt{D+\log(N/\epsilon)}+z\right] \tag{57}$$

with probability $1-e^{-z^2}$. Therefore, fixing some $t\geq 0$, letting $N\to\infty, \epsilon\to 0$, we have that the empirical distribution of the weights obtained from SGD iterations converges weakly to the solution to the DD, via the Portmanteau theorem:

$$\hat{\rho}^{(N)}_{\lfloor t/\epsilon\rfloor} \Rightarrow \rho_t \tag{58}$$

∎

6. Properties, discussions

The generalized risk is non-increasing. This is used as a sanity check to confirm that the DD is tending towards a local minimum in a continuous way. This is contrasted to SGD training dynamics where local steps could be non-optimal (depending on the input data point being used), i.e. local steps may temporarily increase the risk.

Proposition 6.1. Assume V, U differentiable with bounded gradient. If ρ_t is a solution to the distributional dynamics (DD), then $R(\rho_t)$ is non-increasing. Further, ρ is a fixed point of DD if and only if

$$\text{supp}(\rho)\subset\{\boldsymbol{\theta}:\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta};\rho)=0\}$$

Proof. We have that:

$$R(\rho_{t+h}) - R(\rho_t) = 2\int \Psi(\boldsymbol{\theta};\rho_t)\,d(\rho_{t+h}-\rho_t)(\boldsymbol{\theta}) + \iint U\ \left(d(\rho_{t+h}-\rho_t)\right)^{\otimes 2}$$

Since from Lemma 4.8, $t \mapsto \rho_t$ is Lipschitz continuous in the Kantorovich metric, thus $W_(\rho_{t+h}, \rho_t) \leq K\,|h|$. Using Remark 4.6, and reducing the term in U we obtain:

$$\begin{aligned} R(\rho_{t+h}) - R(\rho_t) &= 2\int \Psi(\boldsymbol{\theta};\rho_t)\, d(\rho_{t+h}-\rho_t)(\boldsymbol{\theta}) + O(h^2) \\ &= -4\xi(t)h \int \left\|\nabla_{\boldsymbol{\theta}}\Psi(\boldsymbol{\theta};\rho_t)\right\|_2^2\, d\rho_t(\boldsymbol{\theta}) + o(h) \end{aligned}$$

as the left-hand side is negative for all choices of $h > 0$, it follows that $R(\rho_t)$ is non-increasing in t. We omit the proof for fixed points. ∎

The term N does not appear in the formulation of the distributional dynamics. Relating back to SGD where we established weak convergence of the empirical distribution of its iterated weights to the solution of the DD, this implies that the asymptotic loss landscape as N grows to infinity remains essentially unchanged. In particular, assume that the DD converges close to an optimum in some time $t_* = t_*(D)$. Then, this does not depend on the number of hidden units N, as soon as $N \gg D$, thus SGD can achieve a population risk independent of N. Although this result has been found in several other works in the literature of neural networks, it has been shown here via an interpretation of the DD.

We observe that the error term in the DD grows exponentially with the time horizon T, limiting the applications of this SGD approximation scheme via PDE to cases where the latter converges quickly to a good solution. Authors of [12] stated that it is believed to still be able to prove good convergence and approximation results despite this exponential term, however did not expand further on this idea. They are also not able to obtain a general convergence result of the distributional dynamics, but they do have a stability condition for fixed points containing one point mass, which they use to characterize possible limiting points. The essence of their ideas are illustrated in the two theorems below, which display the stability and instability conditions of the DD respectively.

Theorem 6.2. (Stability conditions of the DD) Assume V and U to be twice differentiable with bounded gradient and bounded continuous Hessian, and let $\boldsymbol{\theta}_* \in \mathbf{R}^D$ be given. Then, $\rho_* = \delta_{\boldsymbol{\theta}_*}$ is a fixed point of the DD if and only if $\nabla V(\boldsymbol{\theta}_*) + \nabla_1 U(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) = 0$. Defining

$$\boldsymbol{H}_0(\rho_*) = \nabla^2 V(\boldsymbol{\theta}_*) + \int \nabla^2_{1,1} U(\boldsymbol{\theta}_*, \boldsymbol{\theta})\, d\rho_*(\boldsymbol{\theta})$$

if $\lambda_{\min}(\boldsymbol{H}_0(\delta_{\boldsymbol{\theta}_*})) > 0$, then there exists $r_0 > 0$ such that if $\operatorname{supp}(\rho_{t_0}) \subseteq B_{r_0}(\boldsymbol{\theta}_*) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_*\|_2 \leq r_0\}$, then $\rho_t \Rightarrow \rho_*$ as $t \to \infty$.

Theorem 6.3. (Instability conditions of the DD) Under the same assumptions as the previous theorem, let $\rho_* = p_*\delta_{\boldsymbol{\theta}_8} + (1-p_*)\tilde{\rho}_* \in \mathscr{P}(\mathbf{R}^D)$ be a fixed point of DD with $p_* \in (0,1]$ and $\nabla\Psi(\boldsymbol{\theta}_*;\rho_*) = 0$ (implied by Proposition 6.1). Define level sets $\mathcal{L}(\eta) = \{\boldsymbol{\theta} : \Psi(\boldsymbol{\theta};\rho_*) \leq \Psi(\boldsymbol{\theta}_*;\rho_*) - \eta\}$ and assume

(1) The eigenvalues of $\boldsymbol{H}_0 = \boldsymbol{H}_0(\rho_*)$ are non-zero, with $\lambda_{\min}(\boldsymbol{H}_0) < 0$

(2) $\tilde{\rho}_* \uparrow 1$ as $\eta \downarrow 0$

(3) $\exists \eta_0 > 0$ such that the sets $\partial\mathcal{L}(\eta)$ are compact for all $\eta \in (0, \eta_0)$

If ρ_0 has bounded density w.r.t. the Lebesgue measure, then it **cannot** be that ρ_t converges weakly to ρ_* as $t \to \infty$.

7. Beyond: theories, analyses, and hypotheses

Authors of [12] studied extensively the corresponding diffusion PDE to noisy SGD, the latter iterates with:

$$\boldsymbol{\theta}_i^{k+1} \leftarrow (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k + 2s_k(y_k - \hat{y}(\boldsymbol{x}_k;\boldsymbol{\theta}^k))\nabla_{\boldsymbol{\theta}_i}\sigma_*(\boldsymbol{x}_k;\boldsymbol{\theta}_i^k) + \sqrt{2s_k/\beta} + \boldsymbol{g}_i^k$$

where $\boldsymbol{g}_i^k \sim \mathcal{N}(0, \boldsymbol{I}_D)$ a standard noise. The term $-2s_k\lambda\boldsymbol{\theta}_i^k$ corresponds to a ℓ^2 regularization. The resulting scaling limit hence becomes:

$$\partial_t \rho_t = 2\xi(t)\nabla_{\boldsymbol{\theta}}\left(\rho_t \nabla_{\boldsymbol{\theta}}\Psi_\lambda(\boldsymbol{\theta};\rho_t)\right) + 2\xi(t)\beta^{-1}\Delta_{\boldsymbol{\theta}}\rho_t$$

called the diffusion dynamics, where $\Psi_\lambda(\boldsymbol{\theta};\rho) = \Psi(\boldsymbol{\theta};\rho) + (\lambda/2)\,\|\boldsymbol{\theta}\|_2^2$. They regarded this diffusion process as a gradient flow for the free energy:

$$F_{\beta,\lambda}(\rho) = \frac{1}{2}R(\rho) + \frac{\lambda}{2}\int \|\boldsymbol{\theta}\|_2^2 \, d\rho(\boldsymbol{\theta}) - \beta^{-1}\,\mathrm{Ent}(\rho)$$

where $\mathrm{Ent}(\rho)$ is the entropy of ρ. Essentially, $F_{\beta,\lambda}$ can be thought of as an entropy-regularized risk. Unlike the case for noiseless SGD, they are able to prove that for $\beta < \infty$, the diffusion process admits a unique fixed point, which is the global minimum of $F_{\beta,\lambda}(\rho)$ and converges to it, if initialized so that $F_{\beta,\lambda} < \infty$. Therefore, noisy SGD generically converges to a global optimum. We refer the reader to [19] which provided an analysis of global optima for noisy SGD.

Various other methods have been able to provide a theory of neural network, including detailed analyses of asymptotic behavior and convergence. Notably, [7] provided a framework of analysis of general neural networks via Neural Tangent Kernels (NTKs), arguing that during SGD training, the network function $f_{\boldsymbol{\theta}} : \mathcal{X} \longrightarrow \mathcal{Y}$ follows the kernel gradient of the functional cost w.r.t. the limiting NTK of the neural network in function space, and the NTK only depends on the depth of the neural network. Furthermore, they are able to establish convergence properties of neural networks via the positive definiteness of the infinite width limiting NTK. Remarkably, the functional cost, i.e. the cost incurred on the function space (in which neural networks are a subset of) is convex, in contrast to the parameter cost, e.g. in [12] where authors used the empirical risk to train SGD. NTKs allow a correspondence between a particular neural network structure and a unique NTK. Several follow up work from various authors provided NTK analyses of convolutional neural networks, recurrent neural networks and residual networks.

Several other analyses, notably [2], formulated the problem of learning a function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ as a search for a function in a Hilbert space $\mathcal{H}$ that minimizes a functional cost $R : \mathcal{H} \longrightarrow \mathbf{R}$, such that the function f is a linear combination of a few elements from a large given parametrized set $\{\phi(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \subset \mathcal{H}$. This corresponds in general to describing the linear combination through an unknown signed measure μ on the parameter space and solving the objective

$$J^* = \min_{\mu \in \mathscr{P}(\boldsymbol{\Theta})} J(\mu) \qquad J(\mu) = R\left(\int \phi \, d\mu\right) + G(\mu)$$

where G is a convex regularizer. The authors aimed to explain when and why the non-convex stochastic particle gradient descent finds a global minima for the discretized version of the above objective, J_m where $J_m = J(m^{-1}\sum_{i=1^m} w_i \delta_{\boldsymbol{\theta}_i})$, where m is the number of particles and w_i are the weights on each particle. The many-particle limit as $m \to \infty$, is characterized as a Kantorovich gradient flow in the space of probability measures over the parameter space. Finally, they proved that if this flow converges, then its limit is the global minimizer of J, using tools from optimal transport theory and mathematical physics.

More recent work aimed to generalize neural networks to functions between topological vector spaces (TVS), finite or infinite dimensional. Authors of [4] proved the universal approximation theorem for neural networks between TVSs, in particular, operator networks, that could have strong implications in operator theory. In essence, they showed that a generalized version of neural networks, namely function machines [11], are able to uniformly approximate to arbitrary precision operators and functionals, and [17] outlined similar approximation results for other configurations of networks, mixing up finite and infinite layers in between. A potential promising area of research would be to apply the theory of function machines to analyze SGD dynamics as well as convergence properties of two-layer finite neural networks. Beyond the above, it would be interesting to apply function machine theory to studying gradient flows in function spaces as well as Kantorovich spaces.

Appendix A. Concentration inequalities

Lemma A.1. (Azuma-Hoeffding variant) Let $(\boldsymbol{X}_k)_{k\geq 0}$ be a martingale with values in $\mathbf{R}^d$ w.r.t. the filtration $(\mathcal{F}_k)_{k\geq 0}$, with $\boldsymbol{X}_0 = 0$. Assume that a.s., the following hold $\forall k \geq 1$:

$$\mathbb{E}\left[\exp\left\{\langle \lambda, \boldsymbol{X}_k - \boldsymbol{X}_{k-1}\rangle\right\}|\mathcal{F}_{k-1}\right] \leq \exp\left\{\frac{L^2\left\|\lambda\right\|^2}{2}\right\}$$

Then, we have

$$\mathbb{P}\left(\max_{k\leq n}\left\|\boldsymbol{X}_k\right\|_2 \geq 2L\sqrt{n}\left(\sqrt{d}+t\right)\right) \leq e^{-t^2}$$

Appendix B. Notation

- Given f measurable and μ a measure, we denote $\langle f, \mu\rangle = \langle \mu, f\rangle = \int f\, d\mu$.
- Given a probability space $(\Omega, \mathcal{F}, \mu)$ and a random variable $X : \Omega \longrightarrow \mathbf{R}$, we denote the pushforward measure of X onto $\mathbf{R}$ by $X_{\#}\mu$, i.e. for all $B \subset X$ Borel, we have that $X_{\#}\mu = \mu\left(X^{-1}(B)\right)$.
- Given a matrix A, we denote the smallest eigenvalue of A by $\lambda_{\min}(A)$.
- We write $\langle U, \rho\rangle^{\otimes 2} = \iint U(\boldsymbol{\theta}, \boldsymbol{\theta})\, d\rho(\boldsymbol{\theta})d\rho(\boldsymbol{\theta})$.

Acknowledgments

I would like to thank Prof. Khalili for his guidance on this research project, as well as his productive discussions and intuition.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.

[2] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.

[3] Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

[4] William H Guss and Ruslan Salakhutdinov. On universal approximation by neural networks with uniform guarantees on approximation of infinite dimensional maps. *arXiv preprint arXiv:1910.01545*, 2019.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[7] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[10] Viet Nguyen. On the concentration of measure in orlicz spaces of exponential type. 2020.

[11] nLab authors. function machine, May 2020. https://ncatlab.org/nlab/show/function+machine.

[12] Mei Song, Andrea Montanari, and P Nguyen. A mean field view of the landscape of two-layers neural networks. *Proceedings of the National Academy of Sciences*, 115:E7665–E7671, 2018.

[13] Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

[14] Yu. V. Kozachenko V. V. Buldygin. *Metric Characterization of Random Variables and Random Processes (Translations of Mathematical Monographs).* American Mathematical Society, 2000.

[15] Aad W Van Der Vaart and Jon A Wellner. Weak convergence and empirical processes. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[17] Eric Hu Viet Nguyen, Johnny Huang. Neural networks: A continuum of potential. 2019.

[18] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[19] Lei Wu, Chao Ma, and E Weinan. How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pages 8279–8288, 2018.