CS 2243 Fall 2024 Scribes: <names>
10/16 Based on notes by: Alan Chung

Lecture 12: Mean Field Limit

1 Finishing up the NTK Analysis

1.1 Recalling the Setting

We are given a dataset $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ and a student network $f_\theta : \mathbb{R}^d \to \mathbb{R}$. We are trying to learn parameters $\theta \in \mathbb{R}^p$, where θ is initialized to some θ_0. We define the loss of some function g to be $\hat{L}(g) \triangleq \frac{1}{2}||g(X) - y||_2^2$ and $\hat{L}_0 \triangleq \hat{L}(\gamma f_{\theta_0})$, where $\gamma > 0$ is a scaling parameter.

We can define a **gradient flow** on the parameter space Θ through the equation

$$d\theta_t = -\nabla_\theta \hat{L}(\gamma f_{\theta_t})dt = -\gamma J_t^T \nabla \hat{L}(\gamma f_{\theta_t})dt,$$

where the **Jacobian** matrix is given by

$$J_t = J_{\theta_t} = \begin{pmatrix} \nabla_\theta f_{\theta_t}(x_1) \\ \vdots \\ \nabla_\theta f_{\theta_t}(x_1). \end{pmatrix} \in \mathbb{R}^{n \times p}$$

The main idea is to compare the gradient flow to **linearized dynamics**, which is given by the equations

$$f_\theta^{lin}(x) = f_{\theta_0}(x) + J_0(\theta - \theta_0)$$

$$d\tilde{\theta}_t = -\nabla_\theta \hat{L}(\gamma f_{\tilde{\theta}_t}^{lin}) = -\gamma J_0^T \nabla \hat{L}(\gamma f_{\tilde{\theta}_t}^{lin}),$$

where we note here that the Jacobian does not change in time for the linearized network. Two assumptions from last time were

1. J_θ is Lipschitz in θ with some constant β.
2. $J_0 = J_{\theta_0}$ is full rank.

Before we complete the analysis, we recall the lemmas we proved last time.

Lemma 1. *Suppose that $Q(t) \geq \lambda \cdot Id$ for all t. Then for (g_t) given by*

$$dg_t = -Q(t)\nabla \hat{L}(g_t)dt,$$

we have

$$\hat{L}(g_t) \leq \hat{L}(g_0) \cdot \exp(-2\lambda t).$$

In lecture, the way this was written was

$$\hat{L}(\tilde{\theta}_t) \leq \exp\left(-\Omega(\sigma_{\min}^2(J_\theta)\gamma^2 t)\right)\hat{L}_0$$

Proof. The proof idea was through a strong convexity style argument. □

Lemma 2 (Lemma 2)**.** *If θ is close to θ_0, then J_θ is close to J_{θ_0}.*

Proof. The idea of the proof was the triangle inequality and the assumption that J is Lipschitz, that is,

$$||J_\theta - J_{\theta_0}||_{op} \leq \beta||\theta - \theta_0||_2$$

for some $\beta > 0$. □

Today, we will do the final step, proving

Lemma 3. *Suppose that $(\hat{\theta}_t)$ satisfies*

$$d\hat{\theta}_t = -S(t)^T \nabla \hat{L}(g_{\hat{\theta}_t})$$

for some network G, and that $\underline{\lambda} \cdot Id \preceq S(t)S(t)^T \preceq \bar{\lambda} \cdot Id$. Then

$$||\hat{\theta}_t - \hat{\theta}_0||_2 \leq \frac{\sqrt{\bar{\lambda}}}{\underline{\lambda}} ||g_{\hat{\theta}_0}(x) - y||.$$

The key feature here is that the numerator features a square root.

Proof.

$$\hat{\theta}_t = \hat{\theta}_0 - \int_0^t S(s)^T \nabla \hat{L}(g_{\hat{\theta}_s}) ds$$

$$\begin{aligned}
\implies ||\hat{\theta}_t = \hat{\theta}_0|| &- \int_0^t ||S(s)^T \nabla \hat{L}(g_{\hat{\theta}_s})|| ds \\
&\leq \int_0^t \underbrace{||S(s)||_{op}}_{\leq \sqrt{\bar{\lambda}}} \cdot || \underbrace{\nabla \hat{L}(g_{\hat{\theta}_s})}_{g_{\hat{\theta}_s}(x) - y} || ds \\
&\leq \sqrt{\bar{\lambda}} \cdot \int_0^t \underbrace{||g_{\hat{\theta}_s}(x) - y||}_{\leq \exp(-\lambda s) \cdot ||g_{\hat{\theta}_0} - y|| \text{ (by previous lemma)}} ds \\
&\leq \sqrt{\bar{\lambda}} \cdot ||g_{\hat{\theta}_0} - y|| \cdot \int_0^t \exp(-\lambda s) ds \\
&\leq \sqrt{\bar{\lambda}} \cdot ||g_{\hat{\theta}_0} - y|| \cdot \frac{1}{\lambda} ds
\end{aligned}$$

Applying this inequality to $\hat{\theta}_t = \tilde{\theta}_t$, $g_\theta = f_\theta^{lin}, S(t) = \gamma J_0$, we have that

$$\begin{aligned}
||\tilde{\theta}_t - \theta_0|| &\leq \frac{\sqrt{\sigma_{\max}^2(J_0)}}{\sigma_{\min}^2(J_0)} \cdot ||f_{\theta_0}(x) - y|| \\
&\leq \frac{\sqrt{2\gamma^2 \sigma_{\max}^2(J_0)}}{\gamma^2 \sigma_{\min}^2(J_0)} \cdot \sqrt{\hat{L}_0} \\
&\lesssim \frac{\sigma_{\max}(J_0)}{\gamma \sigma_{\min}^2} \cdot \sqrt{\hat{L}_0}
\end{aligned}$$

Then, substituting our bounds on the eigenvalues of J_0 suffices for the proof. □

2 Mean-Field Limit

To recall, in the NTK regime, we have as the trainer network

$$f_\theta(x) = \gamma \sum_{i=1}^N a_i \sigma(\langle w_i, x \rangle),$$

and the NTK regime is characterized as $\gamma \gg 1/N$. As showed previously, the Jacobian does not change much over this scaling, and so the network is well-approximated by its Taylor expansion around the parameters at

initialization. However, gradient descent in the NTK regime is bottlenecked by what kernel methods can do. Then, the natural question is what happens in the regime $\gamma \asymp \frac{1}{N}$, which is known as the *mean field regime*.

We first introduce notation, then explain the defining features of the mean field limit. We first introduce the student network

$$f_\theta(x) = \frac{1}{N}\sum_{i=1}^{N} a_i \sigma\left(\langle w_i, x\rangle\right)$$

It will be convenient to denote this by $f_\theta(x) = \frac{1}{N}\sum_{i=1}^{N} \sigma\left(x;\theta_i\right)$ We will consider population gradient descent:

$$\theta^{(k+1)} \leftarrow \theta^{(k)} - \eta_k \nabla L\left(\theta^{(k)}\right)$$

where $L(\theta) = \mathbb{E}_{x,y}\left[(y - f_\theta(x))^2\right]$ is the test loss (can be approximated using online gradient descent). This loss function $L(\theta)$ can be decomposed as

$$L(\theta) = \mathbb{E}\left[y^2\right] + \frac{2}{N}\sum_{i=1}^{N} V\left(\theta_i\right) + \frac{1}{N^2}\sum_{i,j=1}^{N} U\left(\theta_i, \theta_j\right),$$

where

$$V\left(\theta_i\right) \triangleq -\mathbb{E}\left[y \cdot \sigma\left(x;\theta_i\right)\right]$$

is known as the external field and

$$U\left(\theta_i, \theta_j\right) \triangleq \mathbb{E}\left[\sigma\left(x;\theta_i\right) \cdot \sigma\left(x;\theta_j\right)\right]$$

are "pairwise interactions" between particles/neurons. Some convenient notation will be to set

$$\nabla_{\theta_i} L(\theta) = \frac{2}{N}\nabla \Psi_\theta\left(\theta_i\right) \quad \text{for} \quad \Psi_\theta\left(\theta_i\right) \triangleq V\left(\theta_i\right) + \frac{1}{N}\sum_{j=1}^{N} U\left(\theta_i, \theta_j\right).$$

2.1 Physics Intuition for the Mean Field Limit

The physics intuition is to regard each θ_i is an interacting particle. The idea is that as $N \to \infty$, the fluctuations in the "environment" around any given particle average out, so every particle experiences same "average environment," hence the name mean field. In the mean field limit, at any point in time, all particles are i.i.d. draws from same distribution. The natural question, hence, is how does this distribution evolve over the course of training?

Let's consider the loss

$$L(\theta) = \mathbb{E}\left[y^2\right] + \frac{2}{N}\sum_{i=1}^{N} V\left(\theta_i\right) + \frac{1}{N^2}\sum_{i,j=1}^{N} U\left(\theta_i, \theta_j\right)$$

We can think of the term $\frac{2}{N}\sum_{i=1}^{N} V\left(\theta_i\right)$ as similar-valued to $2\mathbb{E}\left[V\left(\theta_i\right)\right]$ for θ_i drawn from the "empirical distribution"$\hat{\rho}_\theta \triangleq \frac{1}{N}\sum_i \delta_{\theta_i}$, and the second term $\frac{1}{N^2}\sum_{i,j=1}^{N} U\left(\theta_i, \theta_j\right)$ can similarly be thought of as some U-Statistic.

Hence, the idea is that instead of parametrizing in terms of $\theta = (\theta_1, \ldots, \theta_N)$, we can parametrize in terms of a probability distribution ρ over $\mathbb{R}^d$. Then L becomes

$$L(\rho) \triangleq \mathbb{E}\left[y^2\right] + 2\int V(\theta)\mathrm{d}\rho(\theta) + \int U\left(\theta, \theta'\right)\mathrm{d}\rho(\theta)\mathrm{d}\rho\left(\theta'\right)$$

To give some more intuition, the basic idea in statistical physics is that when the number of particles in a system is very large, instead of considering the dynamics of all of the individual particles, we can instead consider some probability distribution induced by the particles. This assumption is justified in the $N \to \infty$ limit due to a law of large numbers argument.

Let's do a comparison of what the dynamics look like in the finite N case and in the mean-field limit. In the finite N case, the loss function, Ψ (parameterized by θ), and the dynamics are given by

$$\begin{aligned}
L(\theta) &= \mathbb{E}\left[y^2\right] + \frac{2}{N}\sum_{i=1}^{N} V\left(\theta_i\right) + \frac{1}{N^2}\sum_{i,j=1}^{N} U\left(\theta_i, \theta_j\right) \\
\Psi_\theta\left(\theta_i\right) &\triangleq V\left(\theta_i\right) + \frac{1}{N}\sum_{j=1}^{N} U\left(\theta_i, \theta_j\right) \\
\mathrm{d}\theta_i^t &= -\nabla\Psi_\theta\left(\theta_i^t\right)\mathrm{d}t, \quad \theta_i^0 \sim \rho_0.
\end{aligned}$$

In the mean-field limit, these equations are given by

$$\begin{aligned}
L(\rho) &\triangleq \mathbb{E}\left[y^2\right] + 2\int V(\theta)\mathrm{d}\rho(\theta) + \int U\left(\theta, \theta'\right)\mathrm{d}\rho(\theta)\mathrm{d}\rho\left(\theta'\right) \\
\Psi_\rho(\theta) &\triangleq V(\theta) + \int U\left(\theta, \theta'\right)\mathrm{d}\rho\left(\theta'\right) \\
\mathrm{d}\bar{\theta}_i^t &= -\nabla\Psi_{\rho_t}\left(\bar{\theta}_i^t\right)\mathrm{d}t, \quad \rho_t = \mathrm{law}\left(\bar{\theta}_i^t\right)
\end{aligned}$$

One difference here is that the parameters in question are not individual θ's, but rather probability distributions. Secondly, the function Ψ in the dynamics is now parameterized by ρ_t. The idea here is that the dynamics of the mean-field are now parameterized by the mean-field variables themselves. In particular, one can show that the dynamics are given by a continuity PDE:

$$\partial_t\rho_t = \mathrm{Div}(\rho_t \cdot \nabla\Psi_{\rho_t}),$$

which holds in the weak sense. In the above, $-\nabla\Psi_{\rho_t}$ is called the velocity field of ρ_t, and the idea is that the process ρ_t is performing gradient descent in the space of probability distributions (equipped with the Wasserstein metric) with respect to the function $L(\rho)$. There is a rich theory about this material; one good source is [AGS06].

2.2 Derivation of the Continuity Equation

Recall again the continuity equation given above:

$$\partial_t\rho_t = \mathrm{Div}\left(\rho_t \cdot \nabla\Psi_{\rho_t}\right).$$

In this statement, we mean that the PDE holds in the weak sense, in that for any "nice" (e.g. bounded, differentiable, with bounded gradient) test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we have

$$\int \varphi(\theta)\partial_t\rho_t(\theta)d\theta = \int \varphi(\theta)\cdot \mathrm{div}\left(\rho_t \cdot \nabla\Psi_{\rho_t}\right)(\theta)d\theta \quad (\star)$$

We formulate this as being satisfied in the weak sense because differentiable solutions to $(\star)$ might not exist. We now show the derivation for this equation. Note that for $\bar{\theta}_t \backsim \rho_t$,

$$\begin{aligned}
\text{LHS of } (\star) &= \frac{\partial}{\partial t}\mathbb{E}\left[\varphi\left(\bar{\theta}_t\right)\right] \\
&= \mathbb{E}\left[\left\langle \nabla\varphi\left(\bar{\theta}_t\right), \frac{d}{dt}\bar{\theta}_t\right\rangle\right] \quad \text{(differentiate under integral)} \\
&= \int \left\langle \nabla\varphi(\theta), -\nabla\Psi_{\rho_t}(\theta)\right\rangle d\rho_t(\theta) \quad \text{(gradient flow for } \bar{\theta}_t\text{)} \\
&= -\int \left\langle \nabla\varphi(\theta), \nabla\Psi_{\rho_t}(\theta)\right\rangle d\rho_t(\theta) \\
&= \text{RHS of } (\star) \quad \text{(integration by parts)}
\end{aligned}$$

2.3 Non-asymptotic Convergence to the Mean-Field Limit

We use a method known as the "Propagation of chaos." Some original references are by Kac in 1956, McKean in 1969, and Sznitman in 1991; a comprehensive reference is in [CD22]. We want to compare $(\theta_i^{(k)})_{k=0,1,2\ldots}$ and $\left(\bar{\theta}_i^t\right)_{t\geq 0}$, where the first are the gradient descent iterates given by $\theta_i^{(k+1)} \leftarrow \theta_i^{(k)} - h\nabla L(\theta^{(k)})$, and the second are the mean-field iterates given by $d\bar{\theta}_i^t = -\nabla L_{\rho_t}(\bar{\theta}_i^t)dt$, where $\rho_t = \text{law}(\bar{\theta}_i^t)$. Note that

$$\theta_i^{(k)} = \theta_i^{(0)} + 2h\sum_{l=0}^{k-1} F_i\left(\theta^{(l)}; (x_{l+1}, y_{l+1})\right)$$
$$\bar{\theta}_i^t = \bar{\theta}_i^{(0)} + 2\int_0^t G\left(\bar{\theta}_i^s; \rho_s\right) ds$$

for $F_i\left(\theta; (x,y)\right) \triangleq (y - f_\theta(x))\cdot\nabla_{\theta_i}\sigma\left(x;\theta_i\right), G(\theta;\rho) \triangleq -\nabla\Psi_\rho(\theta)$. The ultimate goal is to upper bound $\left\|\bar{\Theta}_i^{kh} - \theta_i^{(k)}\right\|$. To do so, will bound by a self-similar expression of the form

$$\text{(small terms)} + \int_0^{kh} \left\|\bar{\theta}_i^s - \theta_i^{(\lfloor s/h\rfloor)}\right\| ds.$$

This will imply (by Grönwall's inequality), the desired bound.

Now, let $[S] = h\cdot\lfloor s/h\rfloor$ and consider

$$\begin{aligned}
&\left\|\bar{\theta}_i^{kh} - \theta_i^k\right\| \\
&= 2\left\|\int_0^{kh} G\left(\bar{\theta}_i^s;\rho_s\right) ds - h\sum_{l=0}^{k-1} F_i\left(\theta^{(l)};(x_{l+1},y_{l+1})\right)\right\| \\
&\leq 2\left\|\int_0^{kh}\left[G\left(\bar{\theta}_i^s;\rho_s\right) - G\left(\bar{\theta}_i^{[s]};\rho_{[s]}\right)\right] ds\right\| \quad (1) \\
&+ 2\left\|\int_0^{kh}\left[G\left(\bar{\theta}_i^{[s]};\rho_{[s]}\right) - G\left(\theta_i^{(\lfloor s/h\rfloor)};\rho_{[s]}\right)\right] ds\right\| \quad (2) \\
&+ 2\left\|h\sum_{l=0}^{k-1}\left[G\left(\theta_i^{(l)};\rho_{lh}\right) - F_i\left(\theta^{(l)};(x_{l+1},y_{l+1})\right)\right]\right\| \quad (3)
\end{aligned}$$

We first bound term (1) (easy). This term is small because because G is Lipschitz by assumption, and we can show f varies smoothly over time so that ρ_s and $\rho_{[s]}$ are close. Term (2) is bounded by the Lipschitzness of G:

$$\left\|G\left(\bar{\theta}_i^s;\rho_{[s]}\right) - G\left(\theta_i^{(\lfloor s/h\rfloor)};\rho_{[s]}\right)\right\| \lesssim \left\|\bar{\theta}_i^s - \theta_i^{(\lfloor s/h\rfloor)}\right\|,$$

so (2) is bounded by

$$\int_0^{kh} \underbrace{\left\|\bar{\theta}_i^s - \theta_i^{(\lfloor s/h\rfloor)}\right\|}_{\substack{\text{looks analogous}\\ \text{to what we want}\\ \text{to bound on the LHS}}} ds$$

Now to bound (3), consider

$$\sum_{l=0}^{k-1}\left[G\left(\theta_i^{(l)};\rho_{lh}\right) - F_i\left(\theta^{(l)};(x_{l+1},y_{l+1})\right)\right].$$

The key idea is to note that $F_i\left(\theta^{(l)};(x_{l+1},y_{l+1})\right)$ has expectation $G\left(\theta_i^{(l)};\hat{\rho}_l\right)$, where $\hat{\rho}_l$ is the empirical distribution of $\frac{1}{N}\sum_{i=1}^N \delta_{\theta_i^{(l)}}$. Then, over many steps l, the total deviation between the $F_i\left(\theta^{(l)};(x_{l+1},y_{l+1})\right)$ and the $G\left(\theta_i^{(l)};\hat{\rho}_l\right)$ is of order $h\sqrt{kp}$ by Martingale concentration. Then, replacing F_i with its expectation, it remains to bound

$$\sum_{l=0}^{k-1}\left[G\left(\theta_i^{(l)};\rho_{lh}\right)-G\left(\theta_i^{(l)};\hat{\rho}_l\right)\right]$$
$$=\frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\left[\mathbb{E}_{\bar{\theta}}U\left(\theta_i^{(l)},\bar{\theta}_j^{lh}\right)-U\left(\theta_i^{(l)},\theta_j^{(l)}\right)\right]$$

Again, by Martingale concentration we can essentially replace $\mathbb{E}_{\bar{\theta}}U(\theta_i^{(l)},\bar{\theta}_j^{lh})$ (deterministic) with $U(\theta_i^{(l)},\bar{\theta}_j^{lh})$ (random). Then, we use Lipschitzness of U to get

$$\frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\left\|U\left(\theta_i^{(l)},\bar{\theta}_j^{lh}\right)-U\left(\theta_i^{(l)},\theta_j^{(l)}\right)\right\|$$
$$\leq\frac{1}{N}\sum_{l=0}^{k-1}\sum_{j=1}^{N}\left\|\bar{\theta}_j^{lh}-\theta_j^{(l)}\right\|$$

Once again, this last term looks similar to what we want to bound. This yields the self-similar equation we sought after.

2.4 Note on the PDE when the Data Distribution has Symmetries

When data distribution has symmetries, PDE simplifies considerably. Suppose that the training data $\{(x_i,y_i)\}$ satisfies $x_i\sim N(0,I)$ and $y_i=\varphi(\Pi_x)$ where Π is a projection to a low-dimensional subspace V^*.

Then the joint dist over (x,y) is invariant under rotations of x that preserve V^*, ie. $Rv\in V^*\ \forall v\in V^*$. Consider this observation: Let R be such a rotation. If ρ_0 and ρ_0' are two different initializations of the weights related by $\rho_0'=R_{\#}\rho_0$ (i.e., to sample (a',w') from ρ_0', one can sample (a,w) from ρ_0 and then take $a'=a,w'=Rw)$), then $\rho_t'=R_{\#}\rho_t$.

Hence, if ρ_0 is rotation-invariant, then ρ_t is invariant to rotations preserving V^* for any $t\geq 0$! ρ_t is thus completely specified by the distribution on

$$\left(a,\underbrace{\Pi w}_{\vec{s}},\underbrace{\left\|\Pi^1 w\right\|_2}_{r}\right),$$

i.e., we get a $\dim(V^*)+2$ dimensional PDE! Then denoting the distribution, or $(a,\vec{s},r)$ by $\bar{\rho}_t$, we have

$$\partial_t\bar{\rho}_t=\operatorname{div}\left(\bar{\rho}_t\cdot\nabla_{\vec{s}}\Psi_{\bar{\rho}_t}\right)+\partial_a\left(\bar{\rho}_t\cdot\partial_a\Psi_{\bar{\rho}_t}\right)+\frac{1}{r}\partial_r\left(r\cdot\bar{\rho}_t\cdot\partial_r\bar{\Psi}_{\bar{\rho}_t}\right).$$

3 Analysis of the Mean-Field Theory

Recall that the game plan for understanding training dynamics of overparameterized neural networks was

1. Define the limiting object.
2. Show that we quickly converge to the limiting object as $N\to\infty$.
3. Prove optimization/generalization guarantees for limiting object.

3.1 Theorems Regarding Speed of Convergence

Works such as [MMN18] and [MMM19] address this second point. One theorem that they proved is:

Theorem 1 (Mei-Montanari-Nguyen '18). *Assumptions: (1) $\nabla V, \nabla U$ bounded Lipschitz, (2) σ bounded, (3) $\nabla_\theta \sigma(x;\theta)$ has sub-Gaussian tails.*

Then let $(\theta^{(k)})_{k=0,1,2,\dots}$ denote iterates of gradient descent with step size h and let $(\bar{\theta}^t)_{t\geq 0}$ denote mean-field gradient flow. Then with probability $1-\delta$ in the randomness of the initialization and the training examples,

$$\max_{i\in[N]} \sup_{k=0,\dots,T/h} \left\| \theta_i^{(k)} - \bar{\theta}_i^{kh} \right\|_2$$

$$\lesssim e^{O(T)} \cdot \sqrt{\max(1/N, h)} \cdot [\sqrt{p + \log(N \max(1, T/h))} + \sqrt{\log 1/\delta}]$$

3.2 Theorems Regarding Asymptotics

Regarding the third point above, it is in general quite difficult to prove things about the PDE. The current understanding is largely limited to asymptotics, toy examples (e.g., generalized linear models), and experiments. One example is from [CB18], from which there is the following information theorem about the evolution of $L(\rho_t)$ in noiseless gradient descent.

Theorem 2. *Suppose that σ is sigmoid or ReLU, and the distribution over x has finite fourth-order moments. If the support of the random initialization is chosen appropriately and the distribution over x satisfies certain a certain "Sard-type regularity" condition, then if $\rho_t \to \rho_\infty$, then ρ_∞ is a global minimizer of $L(\rho)$.*

The proof of this theorem exploits facts from the Wasserstein gradient flow, as well as homogeneity/partial homogeneity of the activation function. Follow-up works include [NP23] and [EW21].

Another theorem regarding asymptotics is the evolution of $L(\rho_t)$.

Theorem 3 ([MMN18] (informal)). *The theorem extends the mean-field picture to "noisy gradient descent", i.e. $\theta \leftarrow (1-2\lambda\eta)\cdot\theta - 2\eta\cdot\nabla L(\theta) + N(0, \eta/\beta)$. As $t\to\infty$, the resulting continuity PDE converges to minimizer of the regularized loss (free energy)*

$$F_{\beta,\lambda}(\rho) \overset{def}{=} L(\rho)/2 + (\lambda/2)\cdot \mathbb{E}_\rho\left[\|\theta\|^2\right] - \beta^{-1}\operatorname{Ent}(\rho)$$

In fact, limit distribution satisfies

$$\rho_\infty(\theta) \propto \exp\left(-\beta \Psi_{\rho_\infty}^{(\lambda)}(\theta)\right)$$

3.3 Toy Models

One setting in which the continuity PDE greatly simplifies is when the underlying data distribution has symmetries. For instance, suppose that the x's are Gaussian and that $y = \phi(\langle w^*, x\rangle)$ (the single index model), so that the function secretly only depends on a 1D subspace. Recall this is a setting that kernel methods perform terribly at because they fail to efficiently learn the relevant feature $\langle w^*, \cdot\rangle$. Then, because data distribution is invariant to any rotation that preserves w^*, the PDE simplifies dramatically! In such settings where symmetries drastically reduce the dimension of the PDE, we can numerically solve it and obtain sharp predictions. The following is an example from [ABAM22]. Another example is from [MMN18]. In this paper, the context is classifying isotropic Gaussians. The data is a mixture of

$$\begin{cases} x \sim N\left(0, (1+\Delta)^2 \cdot \mathrm{Id}_d\right) & y = 1 \\ x \sim N\left(0, (1-\Delta)^2 \cdot \mathrm{Id}_d\right) & y = -1 \end{cases}$$

In this case, the reduced PDE is 1-dimensional (only need to track distribution of $\|\theta\|_2$), and they obtain rigorous end-to-end guarantees for this problem.

Another example is [BMZ23]. In this paper, the continuity PDE is for learning $f(x) = \phi(\langle w^*, x\rangle)$ for $\phi(z) = \mathrm{He}_0(z) - \mathrm{He}_1(z) + \frac{2}{3}\mathrm{He}_2(z)$ using a one-hidden-layer ReLU network over Gaussian examples. They demonstrate two empirically observed phenomena: (1) plateaus in the loss curve, interspersed with sharp drops and (2) longer and longer time scales. A key (partially rigorous) finding is that the mean field gradient flow incrementally learns low-degree Hermite components of single index models.

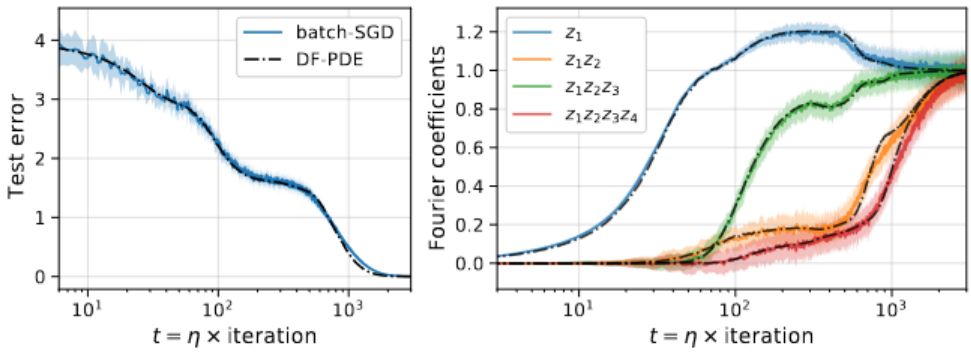


Figure 1: Comparison between (bSGD) and (DF-PDE) dynamics for $h_*(\boldsymbol{z}) = z_1 + z_1z_2 + z_1z_2z_3 + z_1z_2z_3z_4$. Left: Test error. Right: Fourier coefficients of $\hat{f}_{\mathsf{NN}}(\boldsymbol{x}; \boldsymbol{\Theta}^{t/\eta})$ and $\hat{f}_{\mathsf{NN}}(\boldsymbol{z}; \overline{\rho}_t)$. The dashed-dotted black lines correspond to (DF-PDE) and the continuous colored line to (bSGD). The test errors and Fourier coefficients are evaluated with $m = 300$ test samples and for (bSGD), we report the average and 95% confidence interval over 10 experiments.

Currently, the only "end-to-end" non-asymptotic, global convergence result for learning single-index models with "standard" gradient descent (no layerwise training, no weird learning rate schedules, no sharp gradient clipping, etc.) is

Theorem 4 ([MHD+23] (informal)). *Projected gradient descent on a one-hidden-layer student network with quartic polynomial activations and polynomial width learns certain functions of the form $y = p(\langle w^*, x\rangle)$, where p is a degree-4 polynomial, over Gaussian inputs using $O(d^{3.1})$ samples.*

3.4 CSQ Revisited

It's helpful to keep in mind what is generally possible (by any algorithm). Consider single index models ($y = \phi(\langle w^*, x\rangle)$), where the complexity of CSQ algorithms is dictated by the "information exponent," which is the smallest s for which the s-th Hermite coefficient of ϕ is nonzero.

As examples, [AGJ21] shows that online SGD on single neuron learns in time d^S. [ABAM23] generalizes ("leap complexity") to multi-index models ($y = \phi(\Pi_W x)$), layerwise training of overparametrized model learns in time d^{leap} . These are optimal for CSQ algorithms (next unit: lower bound), but there are more efficient non-CSQ algorithms (filtered PCA) that achieve $O(d)$ sample complexity and (fixed) polynomial runtime [CM20].

References

[ABAM22] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks, 2022.

[ABAM23] Emmanuel Abbe, Enric Boix-Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics, 2023.

[AGJ21] Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference, 2021.

[AGS06] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2006.

[BMZ23] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks, 2023.

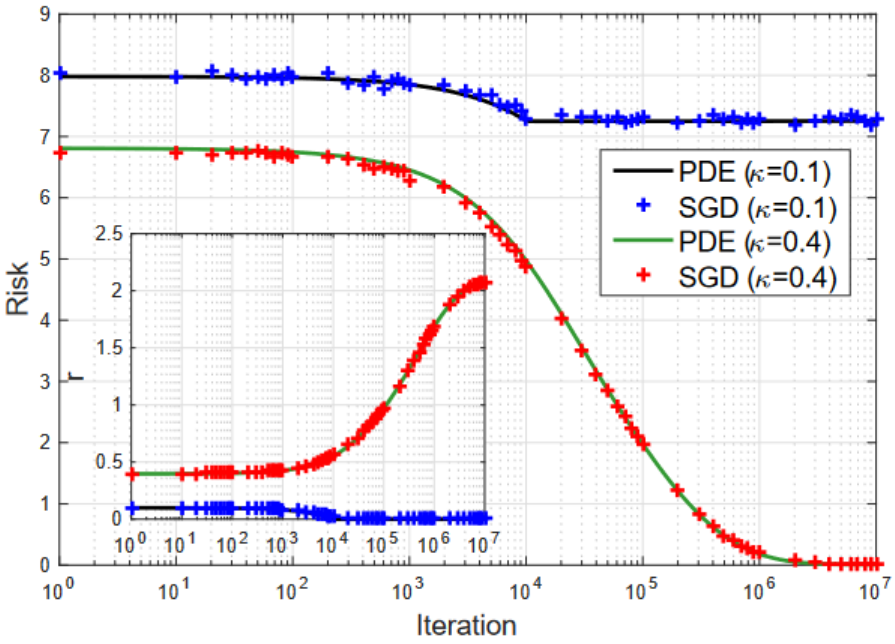


Figure 4: Separating two isotropic Gaussians, with a non-monotone activation function (see text for details). Here $N = 800$, $d = 320$, $\Delta = 0.5$. The main frame presents the evolution of the population risk along the SGD trajectory, starting from two different initializations of $(\boldsymbol{w}_i^0)_{i\le N} \sim_{iid} \mathsf{N}(0, \kappa^2/d \cdot \mathbf{I}_d)$ for either $\kappa = 0.1$ or $\kappa = 0.4$. In the inset, we plot the evolution of the average of $\|\boldsymbol{w}\|_2$ for the same conditions. Symbols are empirical results. Continuous lines are prediction obtained with the reduced PDE (13).

[CB18] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport, 2018.

[CD22] Louis-Pierre Chaintron and Antoine Diez. Propagation of chaos: A review of models, methods and applications. *Kinetic and Related Models*, 15(6):1017, 2022.

[CM20] Sitan Chen and Raghu Meka. Learning polynomials of few relevant dimensions, 2020.

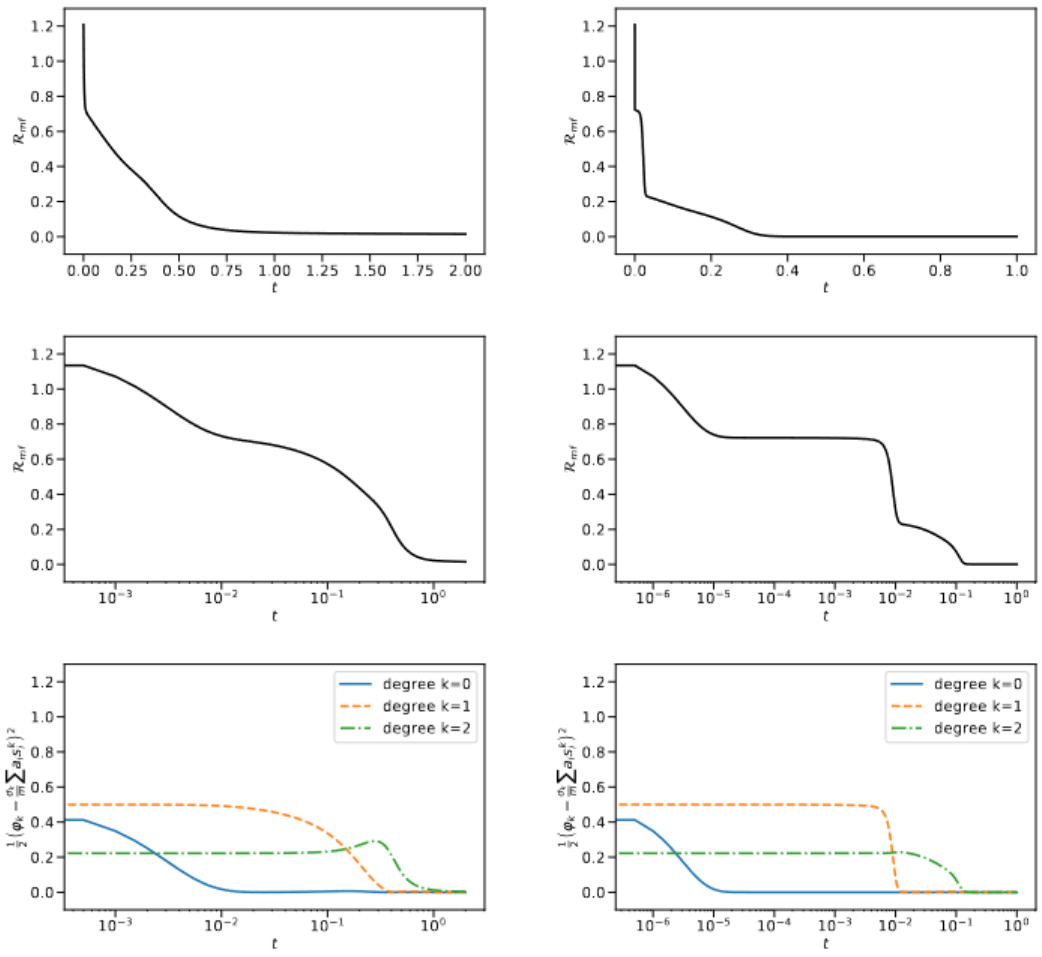
[EW21] Weinan E and Stephan Wojtowytsch. Representation formulas and pointwise properties for barron functions, 2021.

[MHD+23] Arvind Mahankali, Jeff Z. Haochen, Kefan Dong, Margalit Glasgow, and Tengyu Ma. Beyond ntk with vanilla gradient descent: A mean-field analysis of neural networks with polynomial width, samples, and time, 2023.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2388–2464. PMLR, 25–28 Jun 2019.

[MMN18] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), jul 2018.

[NP23] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks, 2023.



(a) $\varepsilon = 10^{-3}$ (b) $\varepsilon = 10^{-6}$

Figure 2: Simulation of the mean field neuron dynamics of Eqs. (23), with the target function of Eq. (37) and ReLU activations. We use learning rate ratios $\varepsilon = 10^{-3}$ (left) and $\varepsilon = 10^{-6}$ (right) and we use $m = 10$ neurons. First two rows: evolution of the risk $\mathscr{R}_{\text{mf}}$ of Eq. (22), in linear and log-scales. Third row: evolution of the first three terms of the sum of (38).