A Mean Field View of Two Layer Neural Networks

Maxwell Lovig

June 26, 2025

Outline

Background:

Two layer neural networks, stochastic gradient descent, previous infinite width limits.

Heuristics:

Three atom example, small step sizes “cancel out” randomness, distributional dynamics.

Results:

Dynamics of noiseless and noisy SGD, fixed point solutions, applications to separating isotropic Gaussians.

What’s Next?:

More complex architectures, applications, other infinite width limits?

Background

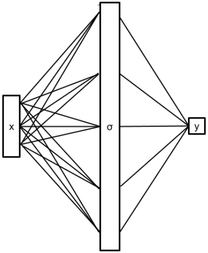
What is a two layer neural network?

Given a function $\sigma : \mathbb{R} \to \mathbb{R}$, weights $\theta = (\theta_i)_{i \in [N]}$ $(\theta_i \in \mathbb{R}^D)$, we define a two layer neural network, $\hat{y}(\theta; x)$, as

$$\hat{y}(\theta; x) = \frac{1}{N} \sum_{i=1}^{N} a_i \cdot \sigma(\langle w_i, x \rangle + b_i)$$

Or more generally,

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



A Two Layer Neural Network

Two common choices of σ are $\text{ReLu}(x) = x\mathbb{1}_{x>0}$ or $\tanh(x)$.

The 2 Layer MLP strikes a balance between richness and tractability. In general, how can such a complex model work so well for many problems?

Problem Setup

Model:

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

Training: Stochastic gradient descent with batch size 1 and step size s_k

$$\theta_i^{k+1} = \theta_i^k + 2s_k(y_k - \hat{y}(x_k;\theta^k))\nabla_{\theta_i}\sigma_*(x_k;\theta_i^k)$$

Here we make a "one-pass assumption" meaning that each $\{(x_k, y_k)\}_{k\geq 1}$ are *iid* draws from $\mathbb{P}$

Major Goal: Characterize the population risk

$$R_N(\theta) = \mathbb{E}^{(x,y)}[(y - \hat{y}(x;\theta)^2]$$

as the number of neurons, N, goes to infinity.

The Large Width Limit

Most neural network analysis is of the asymptotic variety, in the sense that we let the number of neurons go to infinity. Why do we do this?

- It simplifies the analysis (concentration of measure, LLN)
- Even in the limit, these models are still rich enough to be worth studying.
- Modern applications concern massive networks, which in a sense should approach some large width limit. So by understanding what happens for an infinite number of neurons can be relevant.

Lazy Training, [Misiakiewicz and Montanari, 2023]

Previous work has been done on a modification of our model

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

This model enters a "lazy regime" where each θ_i has little action and thus one can analyze the dynamics of a linear approximation called the neural tangent kernel [Jacot et al., 2020].

Lazy Training, [Misiakiewicz and Montanari, 2023]

Previous work has been done on a modification of our model

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

This model enters a "lazy regime" where each θ_i has little action and thus one can analyze the dynamics of a linear approximation called the neural tangent kernel [Jacot et al., 2020].

Say we fix an activation $\sigma_* = \sigma(\langle \theta_i, x \rangle)$ and $w^* \in S_{d-1}(1)$, and generate n data-points $(x_i, y_i) \in \mathbb{R}^{d+1}$ with

$$x_i \sim U(S_{d-1}(\sqrt{d})), \quad y_i = \sigma(\langle w^*, x_i \rangle) + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1)$$

[Ghorbani et al., 2020] showed that if $d^\ell << n << d^{\ell+1}$, then as the number of neurons $N \to \infty$ we have

$$R(\theta_t) \approx ||P_{>\ell}\sigma||^2_{L^2(S_{d-1}(\sqrt{d})}$$

Where $P_{>\ell}$ is the projection onto ℓ degree polynomials.

Lazy Training, [Misiakiewicz and Montanari, 2023]

Previous work has been done on a modification of our model

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

This model enters a "lazy regime" where each θ_i has little action and thus one can analyze the dynamics of a linear approximation called the neural tangent kernel [Jacot et al., 2020].

Say we fix an activation $\sigma_* = \sigma(\langle \theta_i, x \rangle)$ and $w^* \in S_{d-1}(1)$, and generate n data-points $(x_i, y_i) \in \mathbb{R}^{d+1}$ with

$$x_i \sim U(S_{d-1}(\sqrt{d})), \quad y_i = \sigma(\langle w^*, x_i \rangle) + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1)$$

[Ghorbani et al., 2020] showed that if $d^\ell << n << d^{\ell+1}$, then as the number of neurons $N \to \infty$ we have

$$R(\theta_t) \approx ||P_{>\ell}\sigma||^2_{L^2(S_{d-1}(\sqrt{d})}$$

Where $P_{>\ell}$ is the projection onto ℓ degree polynomials. **Wait...**

Animation

The Mean Field Limit: Heuristics

The Mean Field Approach

Recall we are now going to use the "mean-field scaling" model

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

From [Chizat et al., 2020], we know that the model Jacobian, will evolve during training unlike in the NTK model. This will hopefully represent the "feature learning" phenomena we say in the previous animation.

Again, our goal is to characterize

$$R_N(\theta) = \mathbb{E}^{(x,y)}[(y - \hat{y}(x; \theta)^2]$$

as the number of neurons N goes to infinity.

Let's decompose the population risk into the following 3 parts,

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

Where:

$R_\# := E[y^2]$

$V(\theta_i) := -E[y\sigma_*(x;\theta_i)]$

$U(\theta_i,\theta_j) := E[\sigma_*(x;\theta_i)\sigma_*(x;\theta_j)]$

$$R_N(\theta) = R_\# + \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i, \theta_j)$$

This formulation is invariant to the permutation of the neurons, so we can equivalently write

$$R_N(\theta) = R_\# + 2E^\theta[V(\theta)] + E^{\theta,\theta'}[U(\theta_i, \theta_j)]$$

Where θ, θ' are independently drawn from

$$\frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i}$$

This further motivates a risk on any probability measure ρ

$$R(\rho) = R_\# + \int V(\theta)\rho(d\theta) + \int U(\theta, \theta')\rho(d\theta)\rho(d\theta')$$

Let $\hat{\rho}$ be a uniform distribution on N atoms and $\hat{\rho} \implies \rho$.

We then see that,

$$\begin{aligned}
\nabla_{\theta_i} R(\hat{\rho}) &= \frac{2}{N}\nabla_{\theta_i} V(\theta_i) + \frac{1}{N^2}\nabla_{\theta_i} U(\theta_i,\theta_i) + \frac{2}{N^2}\sum_{j\neq i}\nabla_{\theta_i} U(\theta_i,\theta_j) \\
&= 2\hat{\rho}(\theta_i)\left(\nabla_{\theta_i} V(\theta_i) + \frac{1}{2N}\nabla_{\theta_i} U(\theta_i,\theta_i) + \sum_{j\neq i}\hat{\rho}(\theta_j)\nabla_{\theta_i} U(\theta_i,\theta_j)\right) \\
&\overset{N\to\infty}{\to} 2\rho(\theta_i)\left(\nabla_{\theta_i} V(\theta_i) + \int \nabla_{\theta_i} U(\theta_i,\theta')\rho(\theta')\right)
\end{aligned}$$

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

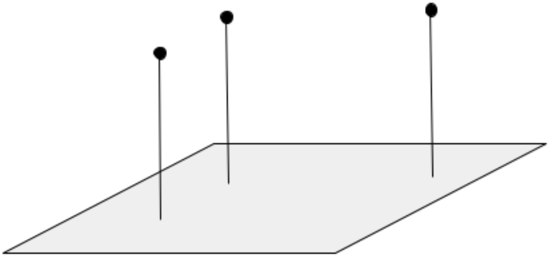


Figure: Here we have 3 atoms in $\mathbb{R}^2 : \theta_1, \theta_2, \theta_3$

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

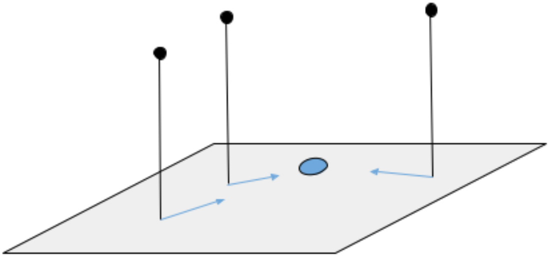


Figure: Say that our random draw of (x, y) has each of these atoms to move towards the blue patch. The blue vectors represent the influence of the external field, which the average case should be something like $\nabla_\theta V(\theta)$.

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

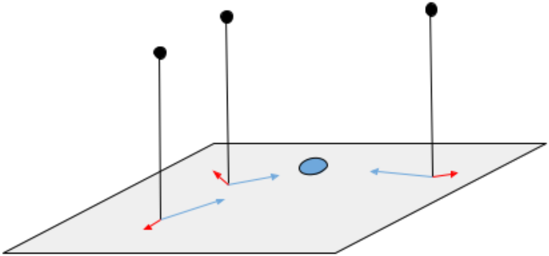


Figure: We then have the mean-field repulsive effect between the atoms represented by the red vectors. In the average case, this should look like $\nabla_\theta \int U(\theta, \theta')\rho(\theta')$.

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

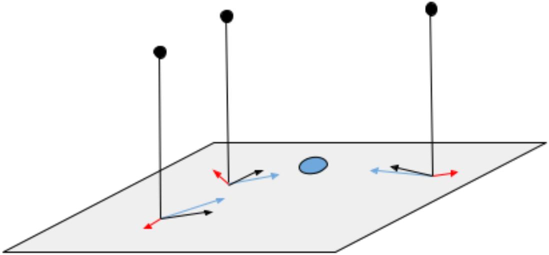


Figure: Adding the two vectors, we get our net direction of each particle.

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

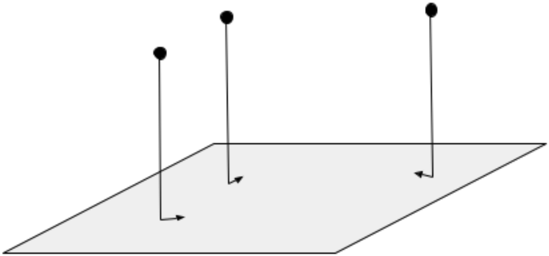


Figure: We then weight this step according to our step size s_k, let's say $s_k \approx 1/2$.

3 Atom Example

Now let's implement our stochastic gradient descent on a simple example with 3 neurons in $\mathbb{R}^2$.

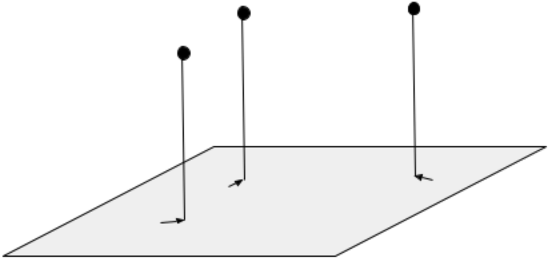


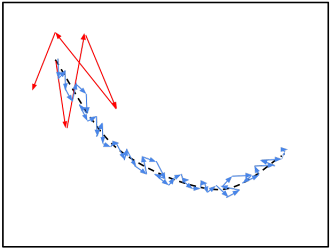
Figure: We then finally take the step in the direction of this resized vector.

Passing to the large width, small step-size limit

Assume that the step-size given by SGD is $s_k = \varepsilon\xi(k\varepsilon)$ for some $\xi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.

As $\varepsilon \to 0$ and $N \to \infty$ we would hope the randomness in SGD is "cancelled out" by a small step-size. Meaning that over the course of many small steps, we take the path the minimizes the average case risk for the limiting model. This step is in the direction of,

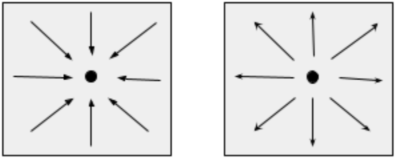
$$-\nabla_\theta \Psi(\theta; \rho) = -\nabla_\theta \left(V(\theta) + \int U(\theta, \theta')\rho(d\theta') \right)$$



Now let's consider a single point $\mathbb{R}^d, \theta_*$. We have that the instantaneous change in the mass on θ_* at time t, $\partial_t \theta_*$, is

$$\begin{aligned}
\partial_t \theta_* &= \text{Mass Entering at } t - \text{Mass Leaving at } t \\
&= (\text{Mass at } t) \cdot (\text{Rate Entering at } t - \text{Rate Leaving at } t) \\
&= (\text{Mass at } t) \cdot (\text{Step Size at } t) \cdot (\text{Net in/out flow for } \theta_* \text{ at } t) \\
&= 2\xi(t) \nabla_{\theta, \theta_*} \cdot (\rho_t \nabla_\theta \Psi(\theta; \rho_t))
\end{aligned}$$

Where $\nabla_{\theta,\theta_*} \cdot (v(\theta))$ represents the divergence of the vector field $v(\theta)$ at θ_*.



Div < 0

Div > 0

Distributional Dynamics

To recap, under the assumptions $s_k = \varepsilon\xi(k\varepsilon)$ we heuristically expect the empirical distribution $\hat{\rho}_k^{(N)} = N^{-1}\sum_{i=1}^N \delta_{\theta_i^k}$ of the N neurons after $k = t/\varepsilon$ steps of SGD to have

$$\hat{\rho}_{t/\varepsilon}^{(N)} \implies \rho_t$$

as $N \to \infty$ and $\varepsilon \to 0$. Here ρ_t evolves according to the PDE

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t \nabla_\theta \Psi(\theta; \rho_t))$$
$$\Psi(\theta; \rho) = V(\theta) + \int U(\theta, \theta')\rho(\theta')$$

We call the solution to this PDE the *Distributional Dynamics*.

The remainder of the talk will be devoted to why such a representation is helpful and a proof of this convergence.

Results

Assumptions

- *$t \mapsto \xi(t)$ is the instantaneous step-size at time t, we assume it is absolutely bounded by K and is K-Lipschitz.*
- *$(x, \theta) \mapsto \sigma_*(x; \theta)$ is absolutely bounded by K and whose gradient has a sub-Gaussian norm $||\nabla_\theta \sigma_*(X, \theta)||_{\psi_2} \leq K$. We also have the labels y_k are absolutely bounded.*
- *We have that $\nabla_\theta V(\theta)$ and $\nabla_{\theta_1} U(\theta_1, \theta_2)$ are also absolutely bounded by K and K-Lipschitz.*
- *We also consider sequences (N, ε_N) such that $N \to \infty, \varepsilon_N \to 0, N/\log(N/\varepsilon_N) \to \infty$ and $\varepsilon_N \log(N/\varepsilon_N) \to 0$*

Theorem (Noiseless Dynamics)

Under the assumptions of the previous slide, consider SGD with initialization $\theta_i \overset{iid}{\sim} \rho_0$ with step-size $s_k = \varepsilon\xi(k\varepsilon)$. For $t \geq 0$, let ρ_t be the distributional dynamics. Then for fixed t we have $\hat{\rho}_t^{(N)} \implies \rho_t$ almost surely along the sequence (N, ε_N).

Moreover, for all test functions $f : \mathbb{R}^D \to \mathbb{R}$ which are absolutely bounded by K and K-Lipschitz, there exists a K dependent constant C such that

$$\sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} |\frac{1}{N} \sum_{i=1}^{N} f(\theta_i^k) - \int f(\theta) \rho_{k\varepsilon}(d\theta)| \leq Ce^{CT} \mathbf{err}_{N,D}(z)$$

$$\sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} |R_n(\theta^k) - R(\rho_{k\varepsilon})| \leq Ce^{CT} \mathbf{err}_{N,D}(z)$$

with probability $1 - e^{-z^2}$ and

$$\mathbf{err}_{N,D}(z) := \sqrt{(1/N) \vee \varepsilon_N} \cdot [\sqrt{D + \log(N/\varepsilon_N)} + z]$$

Theorem ((Informal) Noisy Dynamics)

Consider a noisy variant of SGD:

$$\theta_i^{k+1} = (1 - 2\lambda s_k)\theta_i^k + 2s_k(y_k - \hat{y}(x_k; \theta^k))\nabla_{\theta_i}\sigma_*(x_k; \theta_i^k) + \sqrt{2s_k/\beta}g_i^k$$

We can get a result similar to the noiseless theorem with noisy distributional dynamics, specifically we can get the same error bound with the following PDE

$$\partial_t \rho_t = 2\xi(t)\nabla_\theta \cdot (\rho_t \nabla_\theta \Psi_\lambda(\theta; \rho_t)) + 2\xi(t)\beta^{-1}\Delta_\theta \rho_t$$
$$\Psi_\lambda(\theta; \rho) = V(\theta) + \int U(\theta, \theta')\rho(\theta') + (\lambda/2)||\theta||_2^2$$

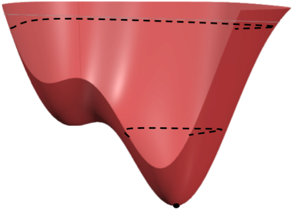
Where Δ_θ is the Laplacian.

Theorem ((Informal) Noiseless Convergence)

We know that any fixed point ρ has the following property

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

Further under assumptions that the initialization ρ_0 is not too far from a point θ^* where the Hessian $H(\rho) = \nabla^2_\theta \Psi(\theta; \rho)$ has $\lambda_{min}(H(\delta_{\theta*})) \geq 0$, then $\rho_t \implies \delta_{\theta*}$ exponentially fast.



Theorem ((Informal) Noisy Convergence)

The (Noisy) distributional dynamics has a fixed point of the form,

$$\rho_*(\theta) = \frac{1}{Z(\beta)} e^{-\beta \Psi_\lambda(\theta;\rho_*)}$$

This fixed point is the global minimizer of the free energy. We have that

$$\rho_t \implies \rho_*$$

as $t \to \infty$. Moreover, this solution has

$$R(\rho_*) \leq \inf_{\theta \in \mathbb{R}^{N \times D}} R_N(\theta) + \frac{CD}{\beta}$$

Where C is some constant dependent on the absolute/gradient bounds of V and U. The convergence to this fix point is N (the number of neurons) independent.

Applications (Informal)

For application type results, they choose a distribution on (x, y) and analyze the distributional dynamics. For example

$$\text{With probability } 1/2 : y = 1, x \sim N(0, (1+\Delta)^2 I_d)$$
$$\text{With probability } 1/2 : y = -1, x \sim N(0, (1-\Delta)^2 I_d)$$

If ρ_0 is spherically symmetric then by the rotational invariance of isotropic Gaussian we can reduce the infinite dimensional distributional dynamics down to a PDE of the distribution $\bar{\rho}_t$ with one parameter, $r = ||w||_2$, the norm of the weights. We then have the evolution

$$\partial_t \bar{\rho}_t = 2\xi(t)\, \partial_r(\bar{\rho}_t \psi(r; \bar{\rho}_t))$$

Where ψ is derived from Ψ under our rotational invariance. They then go on to show that SGD preforms well for this problem by analyzing this PDE.

Theorem (Noiseless Dynamics)

Consider SGD with initialization $\theta_i \overset{iid}{\sim} \rho_0$ with step-size $s_k = \varepsilon\xi(k\varepsilon)$. For $t \geq 0$, let ρ_t be the distributional dynamics. Then for fixed t we have $\hat{\rho}_t^{(N)} \implies \rho_t$ almost surely along the sequence (N, ε_N).

Moreover, for all test functions $f : \mathbb{R}^D \to \mathbb{R}$ which are absolutely bounded by K and K-Lipschitz, there exists a K dependent constant C such that

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |\frac{1}{N}\sum_{i=1}^{N} f(\theta_i^k) - \int f(\theta)\rho_{k\varepsilon}(d\theta)| \leq Ce^{CT}\mathbf{err}_{N,D}(z)$$

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |R_n(\theta^k) - R(\rho_{k\varepsilon})| \leq Ce^{CT}\mathbf{err}_{N,D}(z)$$

with probability $1 - e^{-z^2}$ and

$$\mathbf{err}_{N,D}(z) \coloneqq \sqrt{(1/N) \vee \varepsilon_N} \cdot [\sqrt{D + \log(N/\varepsilon_N)} + z]$$

Proof of The Theorem

Definition

Consider trajectories $(\bar{\theta}_i)_{i\in[N],t\in\mathbb{R}_{\geq 0}}$ generated according to the following non-linear dynamics (where P_X denotes the law of random variable X),

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

with initialization $\bar{\theta}_i^0 = \theta_i^0 \overset{iid}{\sim} \rho_0$. This is interpreted as an evolution of the law of the trajectory of $\bar{\theta}_i^t$. Under the assumptions of our theorem, we know these non-linear dynamics have a unique solution ρ_t [Sznitman, 1991].

Also, for the remainder of the talk, we define a constant K that is dependent on our assumptions and may change value for position to position (for example $K^2 = K$). We also drop the N subscript from ε_N.

High Level Sketch

I will present the proof for test function f, the proof for R_N is similar but slightly more technical.

$$
\begin{aligned}
&\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |\frac{1}{N}\sum_{i=1}^{N} f(\theta_i^k) - \int f(\theta)\rho_{k\varepsilon}(d\theta)| \\
&\quad \leq \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |\frac{1}{N}\sum_{i=1}^{N} f(\theta_i^k) - \frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k)| + |\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k) - \int f(\theta)\rho_{k\varepsilon}(d\theta)| \\
&\quad \leq \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \max_{i\leq N} |f(\theta_i^k) - f(\bar{\theta}_i^k)| + |\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k) - \mathbb{E}_{\rho_0}\left[\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k)\right]| \\
&\quad \leq K \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \max_{i\leq N} ||\theta_i^k - \bar{\theta}_i^{k\varepsilon}||_2 + |\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k) - \mathbb{E}_{\rho_0}\left[\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k)\right]|
\end{aligned}
$$

$$\sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} |\frac{1}{N}\sum_{i=1}^{N} f(\theta_i^k) - \int f(\theta)\rho_{k\varepsilon}(d\theta)|$$

$$\leq K \sup_{k\in[0,T/\varepsilon]\cap\mathbb{N}} \underbrace{\max_{i\leq N} ||\theta_i^k - \bar{\theta}_i^{k\varepsilon}||_2}_{I} + \underbrace{|\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k) - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} f(\bar{\theta}_i^k)\right]|}_{II}$$

If we are able to show that both I and II are bounded by

$$Ke^{KT}\mathbf{err}_{N,D}(z)$$

With probability $1 - e^{-z^2}$ then we are done.

Moreover, since we would show this bound for any $1-$Lipschitz f bounded by 1 absolutely, we would then have that $\hat{\rho}^{(N)}_{\lfloor k/\varepsilon\rfloor} \implies \rho_t$ along any sequence (N, ε) such that the above bound goes to 0 in the limit.

Analysis of II

Since we have that f is bounded absolutely by K then we have that $\phi_k(\bar{\theta}_1, \cdots, \bar{\theta}_N) = \frac{1}{N}\sum_{i=1}^N f(\bar{\theta}_i^k)$ changes by at most K/N when perturbing a coordinate $\bar{\theta}_i$. Using McDiarmind's Inequality we have

$$\begin{aligned}
&\mathbb{P}\left(\phi_k(\bar{\theta}_1, \cdots, \bar{\theta}_N) - \mathbb{E}[\phi_k(\bar{\theta}_1, \cdots, \bar{\theta}_N)] \geq \frac{K}{\sqrt{2}}(1/\sqrt{N})(1+z)\right) \\
&\leq 2\exp\left(-2\frac{N}{K^2}(\frac{K}{\sqrt{2}}(1/\sqrt{N})(1+z))^2\right) \\
&\leq 2\exp(-1-z^2) \\
&\leq e^{-z^2}
\end{aligned}$$

Thus, by a union bound, we have that

$$\begin{aligned}
(II) &\leq \frac{K}{\sqrt{2}}(1/\sqrt{N})(\sqrt{1+\log(T/\varepsilon \vee 1)}+z) \\
&\leq Ke^{KT}\mathbf{err}_{N,D}(z)
\end{aligned}$$

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

Main Technical Lemma

Now we need just show the same upper bound for $\max_{i \leq N} \sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} ||\theta_i^k - \bar{\theta}^{k\varepsilon}||_2$, unfortunately this is much more involved.

Lemma

Under the assumptions of the Theorem, there exists a constant K such that for any $T \geq 0$, we have:

$$\max_{i \leq N} \sup_{k \in [0, T/\varepsilon] \cap \mathbb{N}} ||\theta_i^k - \bar{\theta}^{k\varepsilon}||_2 \\ \leq K e^{KT} \sqrt{(1/N) \vee \varepsilon} \left[\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z \right]$$

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

Proof of the Technical Lemma

Before we go into proving this Lemma, we define the following short-hands:

$$
\begin{aligned}
F_i(\theta; z_k) &= (y_k - \hat{y}(x_k; \theta))\nabla_{\theta_i}\sigma_*(x_k; \theta) \\
G(\theta; \rho) &= -\nabla\Psi(\theta; \rho) = -\nabla V(\theta) - \int \nabla_\theta U(\theta, \theta')\rho(d\theta') \\
[t] &= \varepsilon\lfloor t/\varepsilon \rfloor
\end{aligned}
$$

One can see that G is both bounded and Lipschitz in θ and ρ.

We now can rewrite SGD as

$$\theta_i^{k+1} = \theta_i^k + 2\varepsilon\xi(k\varepsilon)F_i(\theta_i^k; z_{k+1}),$$

unrolling this definition in θ_i^k inductively gives

$$\theta_i^{k+1} = \theta_i^0 + 2\varepsilon \sum_{\ell=0}^{k-1} \xi(\ell\varepsilon)F_i(\theta_i^l; z_{l+1}).$$

Compare this to the non-linear dynamics

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

The proof follows a propagation of chaos argument [Sznitman, 1991]:

$$
\begin{aligned}
||\theta_i^t - \bar{\theta}_i^t||_2 &= 2\left|\left| \int_0^t \xi(s) G(\bar{\theta}_i^s; \rho_s)\, ds - \varepsilon \sum_{k=1}^{t/\varepsilon - 1} \xi(k\varepsilon) F_i(\theta^k; z_{k+1}) \right|\right|_2 \\
&\leq 2 \underbrace{\int_0^t \left|\left| \xi(s) G(\bar{\theta}_i^s; \rho_s) - \xi([s]) G(\bar{\theta}_i^s; \rho_s) \right|\right|_2 ds}_{E_1^i(t)} \\
&\quad + 2 \underbrace{\int_0^t \left|\left| \xi([s]) G(\bar{\theta}_i^s; \rho_s) - \xi([s]) G(\theta_i^{\lfloor s/\varepsilon \rfloor}; \rho_{[s]}) \right|\right|_2 ds}_{E_2^i(t)} \\
&\quad + 2 \underbrace{\left|\left| \varepsilon \sum_{k=1}^{t/\varepsilon - 1} \xi(k\varepsilon) \left[F_i(\theta^k; z_{k+1}) - G(\theta_i^k; \rho_{k\varepsilon}) \right] \right|\right|_2}_{E_3^i(t)}
\end{aligned}
$$

$E_1^i(t) = \int_0^t \left|\left|\xi(s)G(\bar{\theta}_i^s;\rho_s) - \xi([s])G(\bar{\theta}_i^s;\rho_s)\right|\right|_2 ds$

$$\begin{aligned}
E_1^i(t) \le t \sup_{s\in[0,t]} & \left|\left|\xi(s)G(\bar{\theta}_i^s;\rho_s) - \xi([s])G(\bar{\theta}_i^s;\rho_s)\right|\right|_2 && (1)\\
& + \left|\left|\xi([s])G(\bar{\theta}_i^s;\rho_s) - \xi([s])G(\bar{\theta}_i^{[s]};\rho_s)\right|\right|_2 && (2)\\
& + \left|\left|\xi([s])G(\bar{\theta}_i^{[s]};\rho_s) - \xi([s])G(\bar{\theta}_i^{[s]};\rho_{[s]})\right|\right|_2 && (3)\\
\le Kt\varepsilon &
\end{aligned}$$

Where: (1) comes from ξ Lipschitz, $|s-[s]| \le \varepsilon$ and $||G||_\infty \le C_1$. (2) and (3) follow from similar reasoning depending on what variable is different in their respective lines.

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

$$\begin{aligned}
E_2^i(t) &\leq \int_0^t ||\xi([s])G(\bar{\theta}_i^{[s]};\rho_{[s]}) - \xi([s])G(\theta_i^{\lfloor s/\varepsilon \rfloor};\rho_{[s]})||_2 \, ds \\
&\leq \int_0^t |\xi([s])| \cdot ||G(\bar{\theta}_i^{[s]};\rho_{[s]}) - G(\theta_i^{\lfloor s/\varepsilon \rfloor};\rho_{[s]})||_2 \, ds \\
&\leq K \int_0^t ||G(\bar{\theta}_i^{[s]};\rho_{[s]}) - G(\theta_i^{\lfloor s/\varepsilon \rfloor};\rho_{[s]})||_2 \, ds \\
&\leq K \int_0^t ||\bar{\theta}_i^{[s]} - \theta_i^{\lfloor s/\varepsilon \rfloor}||_2 \, ds
\end{aligned}$$

Here we have used the Lipschitz properties of G and the boundedness of ξ.

$E_3^i(t) = \left|\left|\varepsilon \sum_{k=1}^{t/\varepsilon-1} \xi(k\varepsilon)\left[F_i(\theta^k; z_{k+1}) - G(\theta_i^k; \rho_{k\varepsilon})\right]\right|\right|_2$

We bound $E_3^i(t)$ in the following way:

$$E_3^i(t) \leq \frac{K}{N}\sum_{j=1}^{N}\int_0^t ||\theta_j^{\lfloor s/\varepsilon\rfloor} - \bar{\theta}_j^{[s]}||_2 + K(\sqrt{t} \vee t)\sqrt{(1/N) \vee \varepsilon}(\sqrt{D + \log(N(t/\varepsilon \vee 1))} + z) + \frac{Kt}{N}$$

with probability $1 - e^{-z^2}$. This bound comes from conditioning on the filtration $\mathcal{F}_k$ created by $(x_1, y_1), \cdots, (x_k, y_k)$. This works since

$$E[F_i(\theta^k; z_{k+1})|\mathcal{F}_k] = G(\theta_i^k; \hat{\rho}_k^{(N)})$$

As we know that $||\nabla_\theta \sigma_*(X; \theta_i)||_{\psi_2}$ and $|y_k|$ is bounded then we have that the martingale difference $F_i(\theta^k; z_{k+1}) - E[F_i(\theta^k; z_{k+1})|\mathcal{F}_k]$ is sub-Gaussian, a mixture of Azuma-Hoeffding and using the Lipschitz and boundedness of G and F gives the stated bound.

To prove the Lemma, we now define the random variable

$$\Delta(t; N, \varepsilon) = \max_{i \leq N} \max_{k \in [0, t/\varepsilon] \cap \mathbb{N}} ||\theta_i^k - \bar{\theta}_i^{k\varepsilon}||_2$$

We can then get the following bound with probability $1 - e^{-z^2}$.

$$\begin{aligned} \Delta(t; N, \varepsilon) \leq K \int_0^t \Delta(s; N, \varepsilon)\, ds + Kt\varepsilon + \frac{Kt}{N} \\ + K(\sqrt{t} \vee t)\sqrt{(1/N) \vee \varepsilon}(\sqrt{D + \log(N(t/\varepsilon \vee 1))} + z) \end{aligned} \tag{4}$$

We now invoke Gronwall's Inequality, which states that when

$$u(t) \leq \alpha(t) + \int_0^t \beta(s) u(s)\, ds$$

then one has

$$u(t) \leq \alpha(t) e^{\int_0^t \beta(s)\, ds}$$

on line (4) with $\beta(s) = K$ (and thus $\int_0^t 1\, ds = Kt$) we have that

$$\Delta(t; N, \varepsilon) \leq Ke^{Kt}(t\varepsilon + \frac{1}{N} + (\sqrt{t} \vee t)\sqrt{(1/N) \vee \varepsilon}(\sqrt{D + \log(N(t/\varepsilon \vee 1))} + z))$$

The claim follows by absorbing t pre-factors into K and recognizing that $\sqrt{\varepsilon}$ and $N^{-1/2}$ vanish slower than ε and N^{-1}. Meaning that,

$$\begin{aligned} \Delta(t; N, \varepsilon) &= \max_{i \leq N} \max_{k \in [0, t/\varepsilon] \cap \mathbb{N}} ||\theta_i^k - \bar{\theta}_i^{k\varepsilon}||_2 \\ &\leq Ke^{KT} \sqrt{(1/N) \vee \varepsilon} \left[\sqrt{D + \log(N(T/\varepsilon \vee 1))} + z \right] \\ &\leq Ke^{KT} \mathbf{err}_{N,D}(z) \end{aligned}$$

What's Next?

Future Directions 1

Other Possible Infinite Width Limits: There have been two approaches to taking on this problem,

- Trying to Fix NTK to give it some type of feature learning, an example of which is the Neural Tangent Hierarchy. [Huang and Yau, 2019]
- Classifying all possible infinite width limits, even for general optimization procedures. This was done in [Yang and Littwin, 2023] albeit with a non-vanishing s_k.

Future Directions 2

Applications to Specific Problems

- A lot of work has been done on the k-index (or k-ridge) problem. An example of such an analysis was done in [Abbe et al., 2023]
- Expanding to more general function classes, is there an exact characterization of what SGD can and can't learn? An analysis of sparse functions mapping to the hypercube was done by [Abbe et al., 2022] and could be further extended.
- There is also the possibility of statements related to optimal hyperparameters like drop-out or batch-normalization, which would require extensions on the mean field framework to analyze.

Future Directions 3

Going Beyond 2 Layer MLP's

- There have been efforts to establish a mean field approach to attention, specifically to study the clustering behavior of the embedding from repeated single-head attention [Geshkovski et al., 2024].
- In general, there have been steps to try to extend the mean field model to multi-layer MLP's and other architectures such as convolutional layers, RNN's, etc. An example of such analysis can be found in [Yang and Littwin, 2023] and [Nguyen and Pham, 2023].

References I

Abbe, E., Boix-Adsera, E., and Misiakiewicz, T. (2022).
The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks.

Abbe, E., Boix-Adsera, E., and Misiakiewicz, T. (2023).
Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics.

Chizat, L., Oyallon, E., and Bach, F. (2020).
On lazy training in differentiable programming.

Geshkovski, B., Letrouit, C., Polyanskiy, Y., and Rigollet, P. (2024).
The emergence of clusters in self-attention dynamics.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. (2020).
Linearized two-layers neural networks in high dimension.

Huang, J. and Yau, H.-T. (2019).
Dynamics of deep neural networks and neural tangent hierarchy.

Jacot, A., Gabriel, F., and Hongler, C. (2020).
Neural tangent kernel: Convergence and generalization in neural networks.

Misiakiewicz, T. and Montanari, A. (2023).
Six lectures on linearized neural networks.

Nguyen, P.-M. and Pham, H. T. (2023).
A rigorous framework for the mean field limit of multilayer neural networks.

References II

Sznitman, A. S. (1991).
Topics in propagation of chaos.

Yang, G. and Littwin, E. (2023).
Tensor programs ivb: Adaptive optimization in the infinite-width limit.