Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit

Song Mei, Theodor Misiakiewicz, and Andrea Montanari

Stanford University

June 26, 2019

COLT 2019
Gradient dynamics of two-layers neural network

- Two layers neural network:

\[ \Theta = (\theta_1, \ldots, \theta_N), \quad \theta_i = (a_i, w_i) \in \mathbb{R}^D. \]

\[ \hat{y}(x; \Theta) = \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle). \]

- Risk function:

\[ R_N(\Theta) = \mathbb{E}_{x,y} \left[ \left( y - \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle) \right)^2 \right]. \]

- SGD/gradient flow:

\[ \Theta^{k+1} = \Theta^k - \eta_k \nabla \ell_N(\Theta^k; x_k, y_k), \]

\[ \frac{d}{dt} \Theta^t = - \nabla R_N(\Theta^t). \]
Gradient dynamics of two-layers neural network

- Two layers neural network:

\[ \Theta = (\theta_1, \ldots, \theta_N), \quad \theta_i = (a_i, w_i) \in \mathbb{R}^D. \]

\[ \hat{y}(x; \Theta) = \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle). \]

- Risk function:

\[ R_N(\Theta) = \mathbb{E}_{x,y} \left[ \left( y - \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle) \right)^2 \right]. \]

- SGD/gradient flow:

\[ \Theta^{k+1} = \Theta^k - \eta_k \nabla l_N(\Theta^k; x_k, y_k), \]

\[ \frac{d}{dt} \Theta^t = -\nabla R_N(\Theta^t). \]
Gradient dynamics of two-layers neural network

- Two layers neural network:

\[ \Theta = (\theta_1, \ldots, \theta_N), \quad \theta_i = (a_i, w_i) \in \mathbb{R}^D. \]

\[ \hat{y}(x; \Theta) = \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle). \]

- Risk function:

\[ R_N(\Theta) = \mathbb{E}_{x,y} \left[ (y - \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle w_i, x \rangle))^2 \right]. \]

- SGD/gradient flow:

\[ \Theta^{k+1} = \Theta^k - \eta_k \nabla \ell_N(\Theta^k; x_k, y_k), \]

\[ \frac{d}{dt} \Theta^t = - \nabla R_N(\Theta^t). \]
Two-layers neural networks

Figure: Architecture for $N = 4$. $\theta_i = (a_i, w_i)$
Related literatures

- **Mean field distributional dynamics:**

\[ \partial_t \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t). \]

- Non-linear dynamics. Converges in some cases.
- [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018], [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- **Neural tangent kernel (NTK) dynamics:**

\[ \partial_t \|u_t\|^2_2 = -\langle u_t, \mathcal{H}u_t \rangle. \]

- Linear dynamics. Always converges to 0 empirical risk.
- [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018], [Du, Zhai, Poczos, Singh, 2018].
Related literatures

- **Mean field distributional dynamics:**
  \[ \partial_t \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t). \]

- **Non-linear dynamics. Converges in some cases.**
  - [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018], [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- **Neural tangent kernel (NTK) dynamics:**
  \[ \partial_t \|u_t\|_2^2 = -\langle u_t, H u_t \rangle. \]

- **Linear dynamics. Always converges to 0 empirical risk.**
  - [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018], [Du, Zhai, Poczos, Singh, 2018].
Related literatures

- Mean field distributional dynamics:
  \[
  \frac{\partial}{\partial t} \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t).
  \]

- Non-linear dynamics. Converges in some cases.
  - [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018],
    [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- Neural tangent kernel (NTK) dynamics:
  \[
  \frac{\partial}{\partial t} \|u_t\|^2_2 = -\langle u_t, \mathcal{H} u_t \rangle.
  \]

- Linear dynamics. Always converges to 0 empirical risk.
  - [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018],
    [Du, Zhai, Poczos, Singh, 2018].
Related literatures

- **Mean field distributional dynamics:**
  \[
  \partial_t \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t).
  \]
  Non-linear dynamics. Converges in some cases.

  - [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018], [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- **Neural tangent kernel (NTK) dynamics:**
  \[
  \partial_t \|u_t\|_2^2 = -\langle u_t, \mathcal{H}u_t \rangle.
  \]
  Linear dynamics. Always converges to 0 empirical risk.

  - [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018], [Du, Zhai, Poczos, Singh, 2018].
Related literatures

- Mean field distributional dynamics:
  \[ \partial_t \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t). \]

- Non-linear dynamics. Converges in some cases.
- [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018], [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- Neural tangent kernel (NTK) dynamics:
  \[ \partial_t \|u_t\|_2^2 = -\langle u_t, \mathcal{H}u_t \rangle. \]

- Linear dynamics. Always converges to 0 empirical risk.
- [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018], [Du, Zhai, Poczos, Singh, 2018].
Related literatures

- Mean field distributional dynamics:
  \[ \partial_t \rho_t(\theta) = \nabla \cdot (\nabla \Psi(\theta; \rho_t) \rho_t). \]

- Non-linear dynamics. Converges in some cases.
  - [Mei, Montanari, Nguyen, 2018], [Rotskoff and Vanden-Eijnden, 2018], [Chizat and Bach, 2018a], [Sirignano and Spiliopoulos, 2018].

- Neural tangent kernel (NTK) dynamics:
  \[ \partial_t \|u_t\|_2^2 = -\langle u_t, \mathcal{H}u_t \rangle. \]

- Linear dynamics. Always converges to 0 empirical risk.
  - [Jacot, Gabriel, and Clement, 2018], [Li and Liang, 2018], [Du, Zhai, Poczos, Singh, 2018].
This work

(a) Improved bound for SGD - PDE interpolation.

(b) Relationship of the mean field limit and the kernel limit.
SGD and distributional dynamics (DD)

SGD for $\Theta^k$, with $(x_k, y_k) \sim P_{x,y}$, $i \in [N]$,

$$\theta_i^{k+1} = \theta_i^k - 2s_kN \nabla \theta_i \ell_N(\Theta^k; x_k, y_k). \quad \text{(SGD)}$$

[MMN18]: $s_k = \epsilon \xi(k\epsilon), k = t/\epsilon, N \to \infty, \epsilon \to 0$:

$$\hat{\rho}_k^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i^k} \Rightarrow \rho_t \in \mathcal{P}(\mathbb{R}^D) \times [0, \infty).$$

Distributional dynamics (DD) for $\rho_t$,

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla \theta \cdot (\rho_t(\theta) \nabla \Psi(\theta; \rho_t)),$$ \quad \text{(DD)}

where

$$\Psi(\theta; \rho) = \frac{\delta R(\rho)}{\delta \rho(\theta)} = V(\theta) + \int U(\theta, \theta') \rho(d\theta').$$
SGD and distributional dynamics (DD)

- **SGD for** $\Theta^k$, with $(x_k, y_k) \sim P_{x, y}$, $i \in [N],
  \[ \theta_{i}^{k+1} = \theta_{i}^{k} - 2s_k N \nabla_{\theta_i} \ell_N(\Theta^k; x_k, y_k). \] \hspace{1cm} (SGD)

- **[MMN18]**: $s_k = \epsilon \xi(k\epsilon)$, $k = t/\epsilon$, $N \to \infty$, $\epsilon \to 0$: 
  \[ \hat{\rho}_k^{(N)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_{i}^{k}} \Rightarrow \rho_t \in \mathcal{P}(\mathbb{R}^D) \times [0, \infty). \]

- **Distributional dynamics (DD) for** $\rho_t$,
  \[ \partial_t \rho_t(\theta) = 2\xi(t) \nabla_{\theta} \cdot (\rho_t(\theta) \nabla_{\theta} \Psi(\theta; \rho_t)), \] \hspace{1cm} (DD)

where
  \[ \Psi(\theta; \rho) = \frac{\delta R(\rho)}{\delta \rho(\theta)} = \nabla V(\theta) + \int U(\theta, \theta') \rho(d\theta'). \]
SGD and distributional dynamics (DD)

- **SGD** for $\Theta^k$, with $(x_k, y_k) \sim P_{x,y}, i \in [N]$,

$$
\theta_{i}^{k+1} = \theta_{i}^{k} - 2s_k N \nabla \theta_i \ell_N (\Theta^k; x_k, y_k).
$$

(SGD)

- [MMN18]: $s_k = \epsilon \xi(k\epsilon), k = t/\epsilon, N \to \infty, \epsilon \to 0$:

$$
\hat{\rho}_k^{(N)} \equiv \frac{1}{N} \sum_{i=1}^{N} \delta_{\Theta_{i}^{k}} \Rightarrow \rho_t \in \mathcal{P}(\mathbb{R}^D) \times [0, \infty).
$$

- **Distributional dynamics (DD)** for $\rho_t$,

$$
\partial_t \rho_t (\theta) = 2\xi(t) \nabla \theta \cdot (\rho_t (\theta) \nabla \theta \Psi(\theta; \rho_t)),
$$

(DD)

where

$$
\Psi(\theta; \rho) = \frac{\delta R(\rho)}{\delta \rho(\theta)} = V(\theta) + \int \mathcal{U}(\theta, \theta') \rho(d\theta').
$$
An improved bound

Assumption

(i) \( \sigma \) bounded; (ii) \( \nabla_w \sigma(\langle x, w \rangle) \) sub-Gaussian; (iii) \( \nabla \Psi \) bdd. Lipschitz.

Theorem (M., Misiakiewicz, Montanari, 2019)

Let \((\theta_i^0)_{i \leq N} \sim_{iid} \rho_0\). Then, \( \forall f \) bounded Lipschitz, w.h.p,

\[
\sup_{t \leq T} \left| \frac{1}{N} \sum_{i=1}^{N} f(\theta_i^{[t/\varepsilon]}) - \int f(\theta) \rho_t(\theta) \right| \leq \text{Func}(T) \cdot \sqrt{\frac{1}{N} \lor D\varepsilon}.
\]

An example: learning a spherically symmetric Lipschitz function using \( N = O_d(1) \) neurons and \( n = O_d(d) \) samples.

Caveat: this improved bound is not strong. In other cases the factor \( \text{Func}(T) \) could potentially be huge.
An improved bound

Assumption

(i) \( \sigma \) bounded; (ii) \( \nabla_w \sigma(\langle x, w \rangle) \) sub-Gaussian; (iii) \( \nabla \Psi \) bdd. Lipschitz.

Theorem (M., Misiakiewicz, Montanari, 2019)

Let \( (\theta_i^0)_{i \leq N} \sim_{\text{iid}} \rho_0 \). Then, \( \forall f \) bounded Lipschitz, w.h.p,

\[
\sup_{t \leq T} \left| \frac{1}{N} \sum_{i=1}^{N} f(\theta_i^t) - \int f(\theta) \rho_t(\theta) \right| \leq \text{Func}(T) \cdot \sqrt{\frac{1}{N} \vee D\varepsilon}.
\]

An example: learning a spherically symmetric Lipschitz function using \( N = O_d(1) \) neurons and \( n = O_d(d) \) samples.

Caveat: this improved bound is not strong. In other cases the factor \( \text{Func}(T) \) could potentially be huge.
An improved bound

Assumption

(i) \( \sigma \) bounded; (ii) \( \nabla_w \sigma(\langle x, w \rangle) \) sub-Gaussian; (iii) \( \nabla \Psi \) bdd. Lipschitz.

Theorem (M., Misiakiewicz, Montanari, 2019)

Let \((\theta^0_i)_{i \leq N} \sim_{iid} \rho_0\). Then, \( \forall f \) bounded Lipschitz, w.h.p,

\[
\sup_{t \leq T} \left| \frac{1}{N} \sum_{i=1}^{N} f(\theta^{|t/\epsilon|}_i) - \int f(\theta) \rho_t(\theta) \right| \leq \text{Func}(T) \cdot \sqrt{\frac{1}{N}} \vee D\epsilon.
\]

An example: learning a spherically symmetric Lipschitz function using \( N = O_d(1) \) neurons and \( n = O_d(d) \) samples.

Caveat: this improved bound is not strong. In other cases the factor \( \text{Func}(T) \) could potentially be huge.
An improved bound

**Assumption**

(i) $\sigma$ bounded; (ii) $\nabla_w \sigma(\langle x, w \rangle)$ sub-Gaussian; (iii) $\nabla \Psi$ bdd. Lipschitz.

**Theorem (M., Misiakiwicz, Montanari, 2019)**

Let $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$. Then, $\forall f$ bounded Lipschitz, w.h.p,

$$
\sup_{t \leq T} \left| \frac{1}{N} \sum_{i=1}^{N} f(\theta_i^t) - \int f(\theta) \rho_t(\theta) \right| \leq \text{Func}(T) \cdot \sqrt{\frac{1}{N} \vee D\epsilon}.
$$

An example: learning a spherically symmetric Lipschitz function using $N = O_d(1)$ neurons and $n = O_d(d)$ samples.

Caveat: this improved bound is not strong. In other cases the factor $\text{Func}(T)$ could potentially be huge.
This work

(a) Improved bound for SGD - PDE interpolation.

(b) Relationship of the mean field limit and the kernel limit.
Recovering the kernel limit

Same idea appeared in [Chizat and Bach, 2018b], where the kernel limit was called “lazy training”.

Setup:

Prediction function: \( \hat{f}_{\alpha,N}(x; \theta) = \frac{\alpha}{N} \sum_{j=1}^{N} \sigma_j(x; \theta_j) \),

Risk function: \( R_{\alpha,N}(\theta) = \mathbb{E}_x \left[ \left( f(x) - \hat{f}_{\alpha,N}(x; \theta) \right)^2 \right] \),

Gradient flow: \( \frac{d\theta_j^t}{dt} = -\frac{N}{2\alpha^2} \nabla_{\theta_j} R_{\alpha,N}(\theta^t) \).
The coupled dynamics

Denote \( \rho_t^{\alpha,N} = (1/N) \sum_{j=1}^{N} \delta_{\theta_j^t} \). Distributional dynamics:

\[
\partial_t \rho_t^{\alpha,N} = (1/\alpha) \nabla \theta \cdot (\rho_t^{\alpha,N} \nabla \theta \Psi(\theta; \rho_t^{\alpha,N})).
\]

Denote \( u_t^{\alpha,N}(z) = f(z) - \hat{f}_{\alpha,N}(z; \theta^t) \). Residual dynamics:

\[
\partial_t ||u_t^{\alpha,N}||^2_{L^2} = -\langle u_t^{\alpha,N}, \mathcal{H}_{\rho_t^{\alpha,N}} u_t^{\alpha,N} \rangle.
\]

Here

\[
\mathcal{H}_\rho(x, z) \equiv \int \langle \nabla_\theta \sigma_\star(x; \theta), \nabla_\theta \sigma_\star(z; \theta) \rangle \rho(d\theta),
\]

\[
\Psi_\alpha(\theta; \rho^{\alpha,N}) = -\mathbb{E}_x [u_t^{\alpha,N}(x) \sigma_\star(x; \theta)].
\]
The coupled dynamics

Denote $\rho_{t}^{\alpha,N} = (1/N) \sum_{j=1}^{N} \delta_{\theta_{j}^{t}}$. Distributional dynamics:

$$\partial_{t} \rho_{t}^{\alpha,N} = (1/\alpha) \nabla_{\theta} \cdot (\rho_{t}^{\alpha,N} \nabla_{\theta} \Psi(\theta; \rho_{t}^{\alpha,N})).$$

Denote $u_{t}^{\alpha,N}(z) = f(z) - \hat{f}_{\alpha,N}(z; \theta^{t})$. Residual dynamics:

$$\partial_{t} \|u_{t}^{\alpha,N}\|_{L^{2}}^{2} = -\langle u_{t}^{\alpha,N}, H_{\rho_{t}^{\alpha,N}} u_{t}^{\alpha,N} \rangle.$$

Here

$$H_{\rho}(x, z) \equiv \int \langle \nabla_{\theta} \sigma_{*}(x; \theta), \nabla_{\theta} \sigma_{*}(z; \theta) \rangle \rho(d\theta),$$

$$\Psi_{\alpha}(\theta; \rho_{\alpha,N}^{\alpha,N}) = -\mathbb{E}_{x}[u_{t}^{\alpha,N}(x) \sigma_{*}(x; \theta)].$$
The coupled dynamics

Denote \( \rho_{t}^{\alpha,N} = (1/N) \sum_{j=1}^{N} \delta_{\theta_{j}}^{t} \). Distributional dynamics:

\[
\partial_{t} \rho_{t}^{\alpha,N} = (1/\alpha) \nabla_{\theta} \cdot (\rho_{t}^{\alpha,N} \nabla_{\theta} \Psi(\theta; \rho_{t}^{\alpha,N})).
\]

Denote \( u_{t}^{\alpha,N}(z) = f(z) - \hat{f}_{\alpha,N}(z; \theta^{t}) \). Residual dynamics:

\[
\partial_{t} \| u_{t}^{\alpha,N} \|_{L^{2}}^{2} = - \langle u_{t}^{\alpha,N}, \mathcal{H}_{\rho_{t}^{\alpha,N}} u_{t}^{\alpha,N} \rangle.
\]

Here

\[
\mathcal{H}_{\rho}(x, z) \equiv \int \langle \nabla_{\theta} \sigma_{\star}(x; \theta), \nabla_{\theta} \sigma_{\star}(z; \theta) \rangle \rho(d\theta),
\]

\[
\Psi_{\alpha}(\theta; \rho^{\alpha,N}) = - \mathbb{E}_{x}[u_{t}^{\alpha,N}(x) \sigma_{\star}(x; \theta)].
\]
The mean field limit and kernel limit

\[ \partial_t \rho_t^{\alpha,N} = \frac{1}{\alpha} \nabla_\theta \cdot (\rho_t^{\alpha,N} [\nabla_\theta \Psi(\theta; \rho_t^{\alpha,N})]), \]

\[ \partial_t \|u_t^{\alpha,N}\|_{L^2}^2 = - \langle u_t^{\alpha,N}, \mathcal{H}_{\rho_t^{\alpha,N}} u_t^{\alpha,N} \rangle. \]

- The mean field limit: fix \( \alpha = O(1) \) and let \( N \to \infty \).
- The kernel limit: let \( \alpha \to \infty \) after \( N \to \infty \).
- The benefit of kernel limit: the kernel will not change, and the residual dynamics becomes self contained. The empirical risk will converge to 0. Full derivation see appendix H of [Mei, Misiakiewics, Montanari, 2019].
The mean field limit and kernel limit

\[
\begin{align*}
\partial_t \rho_t^{\alpha,N} &= \left(1/\alpha\right) \nabla_{\theta} \cdot \left( \rho_t^{\alpha,N} \nabla_{\theta} \Psi(\theta; \rho_t^{\alpha,N}) \right), \\
\partial_t \|u_t^{\alpha,N}\|_{L^2}^2 &= - \langle u_t^{\alpha,N}, H_{\rho_t^{\alpha,N}} u_t^{\alpha,N} \rangle.
\end{align*}
\]

- The mean field limit: fix \( \alpha = O(1) \) and let \( N \to \infty \).
- The kernel limit: let \( \alpha \to \infty \) after \( N \to \infty \).
- The benefit of kernel limit: the kernel will not change, and the residual dynamics becomes self contained. The empirical risk will converge to 0. Full derivation see appendix H of [Mei, Misiakiewics, Montanari, 2019].
The mean field limit and kernel limit

\[ \partial_t \rho_t^{\alpha,N} = \frac{1}{\alpha} \nabla \theta \cdot (\rho_t^{\alpha,N} [\nabla \theta \Psi(\theta; \rho_t^{\alpha,N})]), \]
\[ \partial_t \|u_t^{\alpha,N}\|_{L^2}^2 = - \langle u_t^{\alpha,N}, \mathcal{H}^{\alpha,N}_{\rho_t} u_t^{\alpha,N} \rangle. \]

- The mean field limit: fix \( \alpha = O(1) \) and let \( N \to \infty \).
- The kernel limit: let \( \alpha \to \infty \) after \( N \to \infty \).
- The benefit of kernel limit: the kernel will not change, and the residual dynamics becomes self contained. The empirical risk will converge to 0. Full derivation see appendix H of [Mei, Misiakiewics, Montanari, 2019].
Gave an interpretation of neural tangent kernel in the mean field point of view. (Also in [Chizat and Bach, 2018b])

Summary

- Gave an interpretation of neural tangent kernel in the mean field point of view. (Also in [Chizat and Bach, 2018b])