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:

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

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

- Risk function:

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

- SGD/gradient flow:

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

$$\frac{\mathrm{d}}{\mathrm{d}t} \Theta^t = -\nabla R_N(\Theta^t).$$

Gradient dynamics of two-layers neural network

- Two layers neural network:

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

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

- Risk function:

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

- SGD/gradient flow:

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

$$\frac{\mathrm{d}}{\mathrm{d}t} \Theta^t = - \nabla R_N(\Theta^t).$$

Gradient dynamics of two-layers neural network

- Two layers neural network:

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

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

- Risk function:

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

- SGD/gradient flow:

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

$$\frac{\mathrm{d}}{\mathrm{d}t} \Theta^t = - \nabla R_N(\Theta^t).$$

Two-layers neural networks

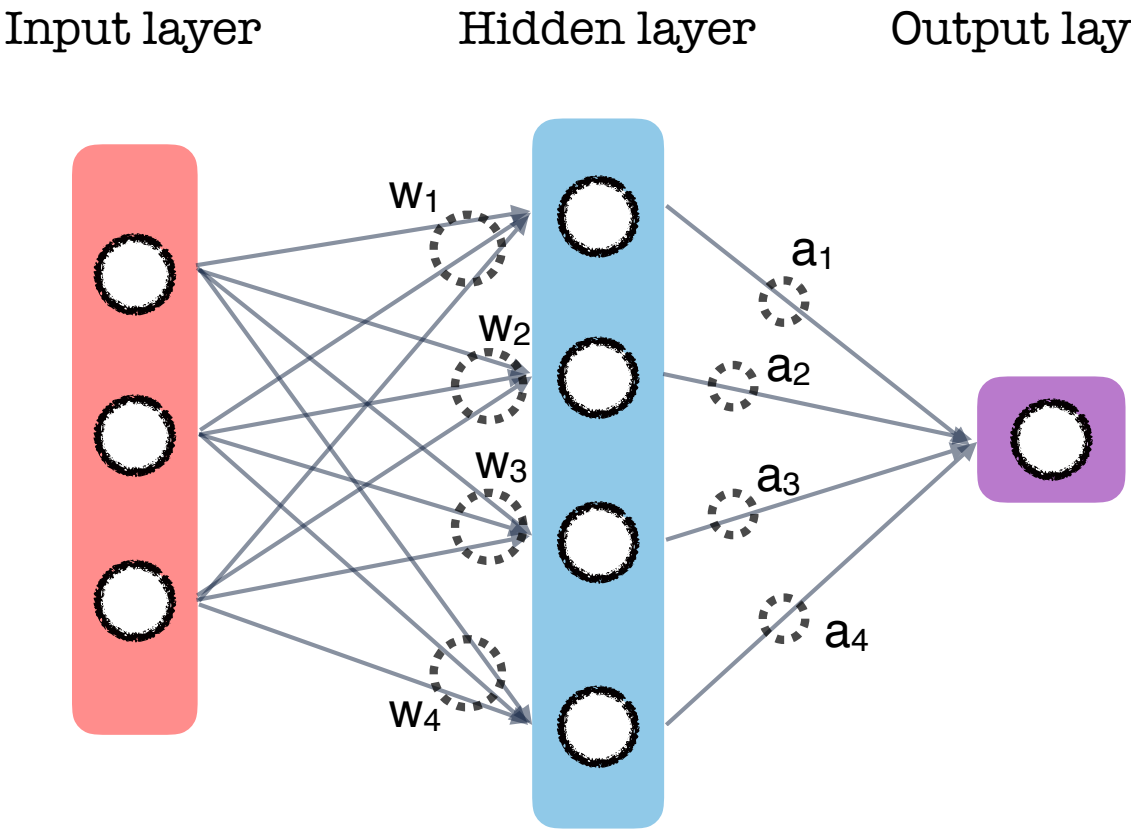

Figure: Architecture for $N = 4$. $\boldsymbol{\theta}_i = (a_i, w_i)$

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:

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

- 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:

$$\partial_t \rho_t(\boldsymbol{\theta}) = \nabla \cdot (\nabla \Psi(\boldsymbol{\theta}; \rho_t) \rho_t).$$

- 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:

$$\partial_t \|\boldsymbol{u}_t\|_2^2 = -\langle \boldsymbol{u}_t, \mathcal{H} \boldsymbol{u}_t \rangle.$$

- 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:

$$\partial_t \rho_t(\boldsymbol{\theta}) = \nabla \cdot (\nabla \Psi(\boldsymbol{\theta}; \rho_t) \rho_t).$$

- 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:

$$\partial_t \|\boldsymbol{u}_t\|_2^2 = -\langle \boldsymbol{u}_t, \mathcal{H} \boldsymbol{u}_t \rangle.$$

- 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 Θ^k, with $(x_k, y_k) \sim \mathbb{P}_{x,y}$, $i \in [N]$,

$$\theta_i^{k+1} = \theta_i^k - 2s_k N \nabla_{\theta_i} \ell_N(\Theta^k; x_k, y_k). \tag{SGD}$$

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

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

- Distributional dynamics (DD) for ρ_t,

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_\theta \cdot (\rho_t(\theta) \nabla_\theta \Psi(\theta; \rho_t)), \tag{DD}$$

where

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

SGD and distributional dynamics (DD)

- SGD for Θ^k, with $(x_k, y_k) \sim \mathbb{P}_{x,y}$, $i \in [N]$,

$$\theta_i^{k+1} = \theta_i^k - 2s_k N \nabla_{\theta_i} \ell_N(\Theta^k; x_k, y_k). \tag{SGD}$$

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

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

- Distributional dynamics (DD) for ρ_t,

$$\partial_t \rho_t(\theta) = 2\xi(t)\nabla_\theta \cdot (\rho_t(\theta)\nabla_\theta \Psi(\theta; \rho_t)), \tag{DD}$$

where

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

SGD and distributional dynamics (DD)

- SGD for Θ^k, with $(x_k, y_k) \sim \mathbb{P}_{x,y}$, $i \in [N]$,

$$\theta_i^{k+1} = \theta_i^k - 2s_k N \nabla_{\theta_i} \ell_N(\Theta^k; x_k, y_k). \tag{SGD}$$

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

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

- Distributional dynamics (DD) for ρ_t,

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_\theta \cdot (\rho_t(\theta) \nabla_\theta \Psi(\theta; \rho_t)), \tag{DD}$$

where

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

An improved bound

Assumption

(i) σ bounded; (ii) $\nabla_{\boldsymbol{w}}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ sub-Gaussian; (iii) $\nabla\Psi$ bdd. Lipschitz.

Theorem (M., Misiakiwicz, Montanari, 2019)

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

$$\sup_{t\le T}\Big|\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^{\lfloor t/\varepsilon\rfloor}) - \int f(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})\Big| \le \mathrm{Func}(T)\cdot\sqrt{\frac{1}{N}\vee D\varepsilon}.$$

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 $\mathrm{Func}(T)$ could potentially be huge.

An improved bound

Assumption

(i) σ bounded; (ii) $\nabla_{\boldsymbol{w}}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ sub-Gaussian; (iii) $\nabla\Psi$ bdd. Lipschitz.

Theorem (M., Misiakiwicz, Montanari, 2019)

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

$$\sup_{t\le T}\left|\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^{\lfloor t/\varepsilon\rfloor}) - \int f(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})\right| \le \mathrm{Func}(T)\cdot\sqrt{\frac{1}{N}\vee D\varepsilon}.$$

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 $\mathrm{Func}(T)$ could potentially be huge.

An improved bound

Assumption

(i) σ bounded; (ii) $\nabla_{\boldsymbol{w}}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ sub-Gaussian; (iii) $\nabla\Psi$ bdd. Lipschitz.

Theorem (M., Misiakiwicz, Montanari, 2019)

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

$$\sup_{t\le T}\left|\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^{\lfloor t/\varepsilon\rfloor}) - \int f(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})\right| \le \mathrm{Func}(T)\cdot\sqrt{\frac{1}{N}\vee D\varepsilon}.$$

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 $\mathrm{Func}(T)$ could potentially be huge.

An improved bound

Assumption

(i) σ bounded; (ii) $\nabla_{\boldsymbol{w}}\sigma(\langle \boldsymbol{x}, \boldsymbol{w}\rangle)$ sub-Gaussian; (iii) $\nabla\Psi$ bdd. Lipschitz.

Theorem (M., Misiakiwicz, Montanari, 2019)

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

$$\sup_{t\le T}\left|\frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}_i^{\lfloor t/\varepsilon\rfloor}) - \int f(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})\right| \le \mathrm{Func}(T)\cdot\sqrt{\frac{1}{N}\vee D\varepsilon}.$$

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 $\mathrm{Func}(T)$ 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:

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

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

$$\text{Gradient flow:} \quad \frac{\mathrm{d}\theta_j^t}{\mathrm{d}t} = -\frac{N}{2\alpha^2}\nabla_{\theta_j} R_{\alpha,N}(\theta^t).$$

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:

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

Here

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

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

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:

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

Here

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

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

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:

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

Here

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

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

The mean field limit and kernel limit

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

- 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

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

- 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

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

- 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].

Summary

- Gave an interpretation of neural tangent kernel in the mean field point of view. (Also in [Chizat and Bach, 2018b])
- Whether NTK can explain the success of neural network is still an open problem. [Arora, Du, Hu, Li, Salakhutdinov, Wang, 2019], [Ghorbani, Mei, Misiakiwics, and Montanari, 2019a, 2019b], [Allen-Zhu and Li, 2019].

Summary

- Gave an interpretation of neural tangent kernel in the mean field point of view. (Also in [Chizat and Bach, 2018b])
- Whether NTK can explain the success of neural network is still an open problem. [Arora, Du, Hu, Li, Salakhutdinov, Wang, 2019], [Ghorbani, Mei, Misiakiwics, and Montanari, 2019a, 2019b], [Allen-Zhu and Li, 2019].