

Mean-Field Theory of Two-Layers Neural Networks: dimension free bounds and examples

Theodor Misiakiewicz

Stanford University

October 28th, 2021

Dynamics and Discretization: PDEs, Sampling, and Optimization workshop

(*Geometric Methods in Optimization and Sampling*, Simons Institute)

Joint work with Song Mei (UC Berkeley) and Andrea Montanari (Stanford)

- A. *Approximation theory for two-layers NNs.*
- B. *Mean-Field description of SGD on two-layers NNs.*
- C. *Example: classifying centered anisotropic Gaussians.*
- D. *Dimension-free bounds between SGD dynamics and mean-field PDE.*
- E. *Outline of the proof of the dimension-free bounds.*

Classical supervised learning setting:

- ▶ Given n i.i.d. samples $\{(y_i, \mathbf{x}_i)\}_{i \in [n]}$:
 - ▶ $\mathbf{x}_i \in \mathbb{R}^d$ vector of covariates.
 - ▶ $y_i \in \mathbb{R}$ response variable.
 - ▶ Common probability distribution $(y_i, \mathbf{x}_i) \sim_{i.i.d.} \mathbb{P} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d)$.
- ▶ Learn model $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ s.t. given a new data point \mathbf{x}_{new} predicts y_{new} via $\hat{f}(\mathbf{x}_{\text{new}})$. Measure the quality of the prediction via the squared error loss:

$$R(\mathbb{P}, \hat{f}) := \mathbb{E}_{(y, \mathbf{x}) \sim \mathbb{P}} \{ (y - \hat{f}(\mathbf{x}))^2 \}.$$

- ▶ Take \hat{f} parametrized by a vector of parameters $\theta \in \mathbb{R}^p$, i.e., $\hat{f} : (\mathbf{x}, \theta) \rightarrow \hat{f}(\mathbf{x}; \theta)$.
- ▶ E.g., fit $\hat{\theta}$ by minimizing the empirical risk

$$\hat{R}^{(n)}(\theta) := \frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}(\mathbf{x}_i; \theta)]^2.$$

Classical supervised learning setting:

- ▶ Given n i.i.d. samples $\{(y_i, \mathbf{x}_i)\}_{i \in [n]}$:
 - ▶ $\mathbf{x}_i \in \mathbb{R}^d$ vector of covariates.
 - ▶ $y_i \in \mathbb{R}$ response variable.
 - ▶ Common probability distribution $(y_i, \mathbf{x}_i) \sim_{i.i.d.} \mathbb{P} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d)$.
- ▶ Learn model $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ s.t. given a new data point \mathbf{x}_{new} predicts y_{new} via $\hat{f}(\mathbf{x}_{\text{new}})$. Measure the quality of the prediction via the squared error loss:

$$R(\mathbb{P}, \hat{f}) := \mathbb{E}_{(y, \mathbf{x}) \sim \mathbb{P}} \{ (y - \hat{f}(\mathbf{x}))^2 \}.$$

- ▶ Take \hat{f} parametrized by a vector of parameters $\theta \in \mathbb{R}^p$, i.e., $\hat{f} : (\mathbf{x}, \theta) \rightarrow \hat{f}(\mathbf{x}; \theta)$.
- ▶ E.g., fit $\hat{\theta}$ by minimizing the empirical risk

$$\hat{R}^{(n)}(\theta) := \frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}(\mathbf{x}_i; \theta)]^2.$$

Classical supervised learning setting:

- ▶ Given n i.i.d. samples $\{(y_i, \mathbf{x}_i)\}_{i \in [n]}$:
 - ▶ $\mathbf{x}_i \in \mathbb{R}^d$ vector of covariates.
 - ▶ $y_i \in \mathbb{R}$ response variable.
 - ▶ Common probability distribution $(y_i, \mathbf{x}_i) \sim_{i.i.d.} \mathbb{P} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d)$.
- ▶ Learn model $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ s.t. given a new data point \mathbf{x}_{new} predicts y_{new} via $\hat{f}(\mathbf{x}_{\text{new}})$. Measure the quality of the prediction via the squared error loss:

$$R(\mathbb{P}, \hat{f}) := \mathbb{E}_{(y, \mathbf{x}) \sim \mathbb{P}} \{ (y - \hat{f}(\mathbf{x}))^2 \}.$$

- ▶ Take \hat{f} parametrized by a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^p$, i.e., $\hat{f} : (\mathbf{x}, \boldsymbol{\theta}) \rightarrow \hat{f}(\mathbf{x}; \boldsymbol{\theta})$.
- ▶ E.g., fit $\hat{\theta}$ by minimizing the empirical risk

$$\hat{R}^{(n)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}(\mathbf{x}_i; \boldsymbol{\theta})]^2.$$

Classical supervised learning setting:

- ▶ Given n i.i.d. samples $\{(y_i, \mathbf{x}_i)\}_{i \in [n]}$:
 - ▶ $\mathbf{x}_i \in \mathbb{R}^d$ vector of covariates.
 - ▶ $y_i \in \mathbb{R}$ response variable.
 - ▶ Common probability distribution $(y_i, \mathbf{x}_i) \sim_{i.i.d.} \mathbb{P} \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d)$.

- ▶ Learn model $\hat{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ s.t. given a new data point \mathbf{x}_{new} predicts y_{new} via $\hat{f}(\mathbf{x}_{\text{new}})$. Measure the quality of the prediction via the squared error loss:

$$R(\mathbb{P}, \hat{f}) := \mathbb{E}_{(y, \mathbf{x}) \sim \mathbb{P}} \{ (y - \hat{f}(\mathbf{x}))^2 \}.$$

- ▶ Take \hat{f} parametrized by a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^p$, i.e., $\hat{f} : (\mathbf{x}, \boldsymbol{\theta}) \rightarrow \hat{f}(\mathbf{x}; \boldsymbol{\theta})$.

- ▶ E.g., fit $\hat{\boldsymbol{\theta}}$ by minimizing the empirical risk

$$\hat{R}^{(n)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}(\mathbf{x}_i; \boldsymbol{\theta})]^2.$$

Two-layers neural networks

- ▶ Need a rich enough class of functions to fit complex data.
- ▶ Consider two-layers neural networks:

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

- ▶ N : number of hidden units (neurons).
 - ▶ $\sigma_* : \mathbb{R}^d \times \mathbb{R}^D$ is an activation function.
 - ▶ $\theta_i \in \mathbb{R}^D$ parameters which we denote collectively $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{ND}$.
- ▶ Standard choice: $\theta_i = (a_i, b_i, \mathbf{w}_i)$ with $a_i \in \mathbb{R}, b_i \in \mathbb{R}, \mathbf{w}_i \in \mathbb{R}^d, D = d + 2,$

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

where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, e.g.: $\sigma(x) = \max(x, 0)$ (ReLU) or $\sigma(x) = \frac{1}{1+e^{-2x}}$ (sigmoid).

Two-layers neural networks

- ▶ Need a rich enough class of functions to fit complex data.
- ▶ Consider **two-layers neural networks**:

$$\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^N \sigma_*(\mathbf{x}; \boldsymbol{\theta}_i).$$

- ▶ N : number of hidden units (neurons).
 - ▶ $\sigma_* : \mathbb{R}^d \times \mathbb{R}^D$ is an activation function.
 - ▶ $\boldsymbol{\theta}_i \in \mathbb{R}^D$ parameters which we denote collectively $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \in \mathbb{R}^{ND}$.
- ▶ Standard choice: $\boldsymbol{\theta}_i = (a_i, b_i, \mathbf{w}_i)$ with $a_i \in \mathbb{R}, b_i \in \mathbb{R}, \mathbf{w}_i \in \mathbb{R}^d, D = d + 2$,

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

where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, e.g.: $\sigma(x) = \max(x, 0)$ (ReLU) or $\sigma(x) = \frac{1}{1+e^{-2x}}$ (sigmoid).

Two-layers neural networks

- ▶ Need a rich enough class of functions to fit complex data.
- ▶ Consider **two-layers neural networks**:

$$\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) := \frac{1}{N} \sum_{i=1}^N \sigma_*(\mathbf{x}; \boldsymbol{\theta}_i).$$

- ▶ N : number of hidden units (neurons).
 - ▶ $\sigma_* : \mathbb{R}^d \times \mathbb{R}^D$ is an activation function.
 - ▶ $\boldsymbol{\theta}_i \in \mathbb{R}^D$ parameters which we denote collectively $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \in \mathbb{R}^{ND}$.
- ▶ Standard choice: $\boldsymbol{\theta}_i = (a_i, b_i, \mathbf{w}_i)$ with $a_i \in \mathbb{R}, b_i \in \mathbb{R}, \mathbf{w}_i \in \mathbb{R}^d, D = d + 2$,

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

where $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, e.g.: $\sigma(x) = \max(x, 0)$ (ReLU) or $\sigma(x) = \frac{1}{1+e^{-2x}}$ (sigmoid).

Approximation properties

- ▶ *Is the function class of two-layers NNs rich enough?*
- ▶ **Universal Approximation [Cybenko, 1989]:** Take $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ continuous with $\lim_{x \rightarrow \infty} \sigma(x) = 1$ and $\lim_{x \rightarrow -\infty} \sigma(x) = 0$. For any $\mathbb{E}\{f(x)^2\} < \infty$ and $\varepsilon > 0$, there exists $N = N(\varepsilon, f)$ such that

$$R_{\text{approx}}(f; N) := \inf_{\{a_i, b_i, \mathbf{w}_i\}} \mathbb{E} \left\{ \left[f(\mathbf{x}) - \frac{1}{N} \sum_{i=1}^N a_i \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i) \right]^2 \right\} \leq \varepsilon.$$

- ▶ *How big should $N(\varepsilon, f)$ be for reasonable functions?*
- ▶ **Barron's Theorem [Barron, 1993]:** $\|\mathbf{x}\|_2 \leq r$ on the support of \mathbb{P} and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has Fourier transform F such that $f(\mathbf{x}) = \int e^{i\langle \mathbf{x}, \mathbf{w} \rangle} F(\mathbf{w}) d\mathbf{w}$. Then

$$R_{\text{approx}}(f; N) \leq \frac{\Delta(f)^2}{N}, \quad \Delta(f) := 2r \int \|\mathbf{w}\|_2 |F(\mathbf{w})| d\mathbf{w}.$$

Hence, $N(\varepsilon, f) \leq \Delta(f)^2 / \varepsilon$.

Approximation properties

- ▶ *Is the function class of two-layers NNs rich enough?*
- ▶ **Universal Approximation [Cybenko, 1989]:** Take $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ continuous with $\lim_{x \rightarrow \infty} \sigma(x) = 1$ and $\lim_{x \rightarrow -\infty} \sigma(x) = 0$. For any $\mathbb{E}\{f(\mathbf{x})^2\} < \infty$ and $\varepsilon > 0$, there exists $N = N(\varepsilon, f)$ such that

$$R_{\text{approx}}(f; N) := \inf_{\{a_i, b_i, \mathbf{w}_i\}} \mathbb{E} \left\{ \left[f(\mathbf{x}) - \frac{1}{N} \sum_{i=1}^N a_i \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i) \right]^2 \right\} \leq \varepsilon.$$

- ▶ *How big should $N(\varepsilon, f)$ be for reasonable functions?*
- ▶ **Barron's Theorem [Barron, 1993]:** $\|\mathbf{x}\|_2 \leq r$ on the support of \mathbb{P} and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has Fourier transform F such that $f(\mathbf{x}) = \int e^{i\langle \mathbf{x}, \mathbf{w} \rangle} F(\mathbf{w}) d\mathbf{w}$. Then

$$R_{\text{approx}}(f; N) \leq \frac{\Delta(f)^2}{N}, \quad \Delta(f) := 2r \int \|\mathbf{w}\|_2 |F(\mathbf{w})| d\mathbf{w}.$$

Hence, $N(\varepsilon, f) \leq \Delta(f)^2 / \varepsilon$.

Approximation properties

- ▶ *Is the function class of two-layers NNs rich enough?*
- ▶ **Universal Approximation [Cybenko, 1989]:** Take $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ continuous with $\lim_{x \rightarrow \infty} \sigma(x) = 1$ and $\lim_{x \rightarrow -\infty} \sigma(x) = 0$. For any $\mathbb{E}\{f(\mathbf{x})^2\} < \infty$ and $\varepsilon > 0$, there exists $N = N(\varepsilon, f)$ such that

$$R_{\text{approx}}(f; N) := \inf_{\{(a_i, b_i, \mathbf{w}_i)\}} \mathbb{E} \left\{ \left[f(\mathbf{x}) - \frac{1}{N} \sum_{i=1}^N a_i \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i) \right]^2 \right\} \leq \varepsilon.$$

- ▶ *How big should $N(\varepsilon, f)$ be for reasonable functions?*
- ▶ **Barron's Theorem [Barron, 1993]:** $\|\mathbf{x}\|_2 \leq r$ on the support of \mathbb{P} and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has Fourier transform F such that $f(\mathbf{x}) = \int e^{i\langle \mathbf{x}, \mathbf{w} \rangle} F(\mathbf{w}) d\mathbf{w}$. Then

$$R_{\text{approx}}(f; N) \leq \frac{\Delta(f)^2}{N}, \quad \Delta(f) := 2r \int \|\mathbf{w}\|_2 |F(\mathbf{w})| d\mathbf{w}.$$

Hence, $N(\varepsilon, f) \leq \Delta(f)^2 / \varepsilon$.

Approximation properties

- ▶ *Is the function class of two-layers NNs rich enough?*
- ▶ **Universal Approximation [Cybenko, 1989]:** Take $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ continuous with $\lim_{x \rightarrow \infty} \sigma(x) = 1$ and $\lim_{x \rightarrow -\infty} \sigma(x) = 0$. For any $\mathbb{E}\{f(\mathbf{x})^2\} < \infty$ and $\varepsilon > 0$, there exists $N = N(\varepsilon, f)$ such that

$$R_{\text{approx}}(f; N) := \inf_{\{a_i, b_i, \mathbf{w}_i\}} \mathbb{E} \left\{ \left[f(\mathbf{x}) - \frac{1}{N} \sum_{i=1}^N a_i \sigma(\langle \mathbf{w}_i, \mathbf{x} \rangle + b_i) \right]^2 \right\} \leq \varepsilon.$$

- ▶ *How big should $N(\varepsilon, f)$ be for reasonable functions?*
- ▶ **Barron's Theorem [Barron, 1993]:** $\|\mathbf{x}\|_2 \leq r$ on the support of \mathbb{P} and $f : \mathbb{R}^d \rightarrow \mathbb{R}$ has Fourier transform F such that $f(\mathbf{x}) = \int e^{i\langle \mathbf{x}, \mathbf{w} \rangle} F(\mathbf{w}) d\mathbf{w}$. Then

$$R_{\text{approx}}(f; N) \leq \frac{\Delta(f)^2}{N}, \quad \Delta(f) := 2r \int \|\mathbf{w}\|_2 |F(\mathbf{w})| d\mathbf{w}.$$

Hence, $N(\varepsilon, f) \leq \Delta(f)^2 / \varepsilon$.

Insights from approximation theory

- ▶ Suggest that we should *represent two-layers NN with distribution* $\rho \in \mathcal{P}(\mathbb{R}^D)$:

$$\hat{f}(\mathbf{x}; \rho) = \int \sigma_*(\mathbf{x}; \boldsymbol{\theta}) \rho(d\boldsymbol{\theta}).$$

E.g., take $\hat{\rho}^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\boldsymbol{\theta}_i}$ for finite networks: $\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) = \hat{f}(\mathbf{x}; \hat{\rho}^{(N)})$.

- ▶ *Small population risk achieved by many NNs*: what matters is ρ , not $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$. Behavior is insensitive to the number of neurons N , as long as it is large enough for $\hat{\rho}^{(N)}$ to approximate ρ .
- ▶ Minimum number of neurons to achieve certain accuracy depends on the *intrinsic regularity* of f (e.g., $\Delta(f)$) and not on the dimension d .

These insights concern *ideal representations*.

Insights from approximation theory

- ▶ Suggest that we should *represent two-layers NN with distribution* $\rho \in \mathcal{P}(\mathbb{R}^D)$:

$$\hat{f}(\mathbf{x}; \rho) = \int \sigma_*(\mathbf{x}; \boldsymbol{\theta}) \rho(d\boldsymbol{\theta}).$$

E.g., take $\hat{\rho}^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\boldsymbol{\theta}_i}$ for finite networks: $\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) = \hat{f}(\mathbf{x}; \hat{\rho}^{(N)})$.

- ▶ *Small population risk achieved by many NNs*: what matters is ρ , not $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$. Behavior is insensitive to the number of neurons N , as long as it is large enough for $\hat{\rho}^{(N)}$ to approximate ρ .
- ▶ Minimum number of neurons to achieve certain accuracy depends on the *intrinsic regularity* of f (e.g., $\Delta(f)$) and not on the dimension d .

These insights concern *ideal representations*.

Insights from approximation theory

- ▶ Suggest that we should *represent two-layers NN with distribution* $\rho \in \mathcal{P}(\mathbb{R}^D)$:

$$\hat{f}(\mathbf{x}; \rho) = \int \sigma_*(\mathbf{x}; \boldsymbol{\theta}) \rho(d\boldsymbol{\theta}).$$

E.g., take $\hat{\rho}^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\boldsymbol{\theta}_i}$ for finite networks: $\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) = \hat{f}(\mathbf{x}; \hat{\rho}^{(N)})$.

- ▶ *Small population risk achieved by many NNs*: what matters is ρ , not $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$. Behavior is insensitive to the number of neurons N , as long as it is large enough for $\hat{\rho}^{(N)}$ to approximate ρ .
- ▶ Minimum number of neurons to achieve certain accuracy depends on the *intrinsic regularity* of f (e.g., $\Delta(f)$) and not on the dimension d .

These insights concern *ideal representations*.

Insights from approximation theory

- ▶ Suggest that we should *represent two-layers NN with distribution* $\rho \in \mathcal{P}(\mathbb{R}^D)$:

$$\hat{f}(\mathbf{x}; \rho) = \int \sigma_*(\mathbf{x}; \boldsymbol{\theta}) \rho(d\boldsymbol{\theta}).$$

E.g., take $\hat{\rho}^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\boldsymbol{\theta}_i}$ for finite networks: $\hat{f}_N(\mathbf{x}; \boldsymbol{\theta}) = \hat{f}(\mathbf{x}; \hat{\rho}^{(N)})$.

- ▶ *Small population risk achieved by many NNs*: what matters is ρ , not $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$. Behavior is insensitive to the number of neurons N , as long as it is large enough for $\hat{\rho}^{(N)}$ to approximate ρ .
- ▶ Minimum number of neurons to achieve certain accuracy depends on the *intrinsic regularity* of f (e.g., $\Delta(f)$) and not on the dimension d .

These insights concern *ideal representations*.

- ▶ In practice, the parameters of NNs are learned by SGD or its variants.
- ▶ **SGD**: initialize weights $\theta_i \sim_{iid} \rho_0$. At each step k , sample (\mathbf{x}_k, y_k) and update

$$\theta_i^{k+1} = \theta_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \theta^k)) \nabla_{\theta_i} \sigma_*(\mathbf{x}_k; \theta_i^k).$$

ε : step size; $\theta^k = (\theta_i^k)_{i \leq N}$: parameters after k iterations.

What are the properties of NNs reached by SGD?

- ▶ Do they have small test error? Are they fairly insensitive to the number N of neurons (as long as N is large enough) and the dimension d , as in approximation theory?
- ▶ Recent analysis connects naturally SGD dynamics and approximation theory.
[Mei, Montanari, Nguyen, '18], [Chizat, Bach, '18], [Sirignano, Spiliopoulos, '18], [Rotskoff, Vanden-Eijnden, '18]

Mean-field theory: SGD dynamics admits an asymptotic description as $N \rightarrow \infty, \varepsilon \rightarrow 0$ in terms of a PDE in the space of probability distributions on \mathbb{R}^D .

- ▶ In practice, the parameters of NNs are learned by SGD or its variants.
- ▶ **SGD**: initialize weights $\theta_i \sim_{iid} \rho_0$. At each step k , sample (\mathbf{x}_k, y_k) and update

$$\theta_i^{k+1} = \theta_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \theta^k)) \nabla_{\theta_i} \sigma_*(\mathbf{x}_k; \theta_i^k).$$

ε : step size; $\theta^k = (\theta_i^k)_{i \leq N}$: parameters after k iterations.

What are the properties of NNs reached by SGD?

- ▶ Do they have small test error? Are they fairly insensitive to the number N of neurons (as long as N is large enough) and the dimension d , as in approximation theory?
- ▶ Recent analysis connects naturally SGD dynamics and approximation theory.

[Mei, Montanari, Nguyen, '18], [Chizat, Bach, '18], [Sirignano, Spiliopoulos, '18], [Rotskoff, Vanden-Eijnden, '18]

Mean-field theory: SGD dynamics admits an asymptotic description as $N \rightarrow \infty, \varepsilon \rightarrow 0$ in terms of a PDE in the space of probability distributions on \mathbb{R}^D .

- ▶ In practice, the parameters of NNs are learned by SGD or its variants.
- ▶ **SGD:** initialize weights $\theta_i \sim_{iid} \rho_0$. At each step k , sample (\mathbf{x}_k, y_k) and update

$$\theta_i^{k+1} = \theta_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \theta^k)) \nabla_{\theta_i} \sigma_*(\mathbf{x}_k; \theta_i^k).$$

ε : step size; $\theta^k = (\theta_i^k)_{i \leq N}$: parameters after k iterations.

What are the properties of NNs reached by SGD?

- ▶ Do they have small test error? Are they fairly insensitive to the number N of neurons (as long as N is large enough) and the dimension d , as in approximation theory?
- ▶ Recent analysis connects naturally SGD dynamics and approximation theory.

[Mei, Montanari, Nguyen, '18], [Chizat, Bach, '18], [Sirignano, Spiliopoulos, '18], [Rotskoff, Vanden-Eijnden, '18]

Mean-field theory: SGD dynamics admits an asymptotic description as $N \rightarrow \infty, \varepsilon \rightarrow 0$ in terms of a PDE in the space of probability distributions on \mathbb{R}^D .

Mean-field limit

- ▶ **One-pass SGD:** training examples are never revisited, i.e., $\{(\mathbf{x}_k, y_k)\}_{k \geq 1}$ are iid.
- ▶ Denote $\hat{\rho}_k^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\theta_i^k}$ after k SGD steps with step size ε and $\theta_i^0 \sim_{iid} \rho_0$:

$$\hat{\rho}_{t/\varepsilon}^{(N)} \Rightarrow \rho_t, \quad \text{as } N \rightarrow \infty, \varepsilon \rightarrow 0.$$

- ▶ Evolution of ρ_t given by the following PDE (of McKean-Vlasov type):

$$\begin{aligned} \partial_t \rho_t &= \nabla_{\theta} \cdot \left(\rho_t \nabla_{\theta} \Psi(\theta; \rho_t) \right), \\ \Psi(\theta; \rho_t) &\equiv V(\theta) + \int U(\theta, \theta') \rho(d\theta'), \end{aligned}$$

where $V(\theta) := -\mathbb{E}\{y\sigma_*(\mathbf{x}; \theta)\}$ and $U(\theta_1, \theta_2) := \mathbb{E}_{\mathbf{x}}\{\sigma_*(\mathbf{x}; \theta_1)\sigma_*(\mathbf{x}; \theta_2)\}$.

This is referred to as the *mean-field* description, or *distributional dynamics* (DD).

- ▶ Wasserstein gradient flow on risk $R(\rho) := \mathbb{E}\{(y - \hat{f}(\mathbf{x}; \rho))^2\}$, with $\rho \in \mathcal{P}(\mathbb{R}^D)$

Mean-field limit

- ▶ **One-pass SGD:** training examples are never revisited, i.e., $\{(\mathbf{x}_k, y_k)\}_{k \geq 1}$ are iid.
- ▶ Denote $\hat{\rho}_k^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\theta_i^k}$ after k SGD steps with step size ε and $\theta_i^0 \sim_{iid} \rho_0$:

$$\hat{\rho}_{t/\varepsilon}^{(N)} \Rightarrow \rho_t, \quad \text{as } N \rightarrow \infty, \varepsilon \rightarrow 0.$$

- ▶ Evolution of ρ_t given by the following PDE (of *McKean-Vlasov* type):

$$\begin{aligned} \partial_t \rho_t &= \nabla_{\theta} \cdot \left(\rho_t \nabla_{\theta} \Psi(\theta; \rho_t) \right), \\ \Psi(\theta; \rho_t) &\equiv V(\theta) + \int U(\theta, \theta') \rho(d\theta'), \end{aligned}$$

where $V(\theta) := -\mathbb{E}\{y\sigma_*(\mathbf{x}; \theta)\}$ and $U(\theta_1, \theta_2) := \mathbb{E}_{\mathbf{x}}\{\sigma_*(\mathbf{x}; \theta_1)\sigma_*(\mathbf{x}; \theta_2)\}$.

This is referred to as the *mean-field* description, or *distributional dynamics* (DD).

- ▶ Wasserstein gradient flow on risk $R(\rho) := \mathbb{E}\{(y - \hat{f}(\mathbf{x}; \rho))^2\}$, with $\rho \in \mathcal{P}(\mathbb{R}^D)$

Mean-field limit

- ▶ **One-pass SGD:** training examples are never revisited, i.e., $\{(\mathbf{x}_k, y_k)\}_{k \geq 1}$ are iid.
- ▶ Denote $\hat{\rho}_k^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\theta_i^k}$ after k SGD steps with step size ε and $\theta_i^0 \sim_{iid} \rho_0$:

$$\hat{\rho}_{t/\varepsilon}^{(N)} \Rightarrow \rho_t, \quad \text{as } N \rightarrow \infty, \varepsilon \rightarrow 0.$$

- ▶ Evolution of ρ_t given by the following PDE (of McKean-Vlasov type):

$$\begin{aligned} \partial_t \rho_t &= \nabla_{\theta} \cdot \left(\rho_t \nabla_{\theta} \Psi(\theta; \rho_t) \right), \\ \Psi(\theta; \rho_t) &\equiv V(\theta) + \int U(\theta, \theta') \rho(d\theta'), \end{aligned}$$

where $V(\theta) := -\mathbb{E}\{y \sigma_*(\mathbf{x}; \theta)\}$ and $U(\theta_1, \theta_2) := \mathbb{E}_{\mathbf{x}}\{\sigma_*(\mathbf{x}; \theta_1) \sigma_*(\mathbf{x}; \theta_2)\}$.

This is referred to as the *mean-field* description, or *distributional dynamics* (DD).

- ▶ Wasserstein gradient flow on risk $R(\rho) := \mathbb{E}\{(y - \hat{f}(\mathbf{x}; \rho))^2\}$, with $\rho \in \mathcal{P}(\mathbb{R}^D)$

Mean-field limit

- ▶ **One-pass SGD:** training examples are never revisited, i.e., $\{(\mathbf{x}_k, y_k)\}_{k \geq 1}$ are iid.
- ▶ Denote $\hat{\rho}_k^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\theta_i^k}$ after k SGD steps with step size ε and $\theta_i^0 \sim_{iid} \rho_0$:

$$\hat{\rho}_{t/\varepsilon}^{(N)} \Rightarrow \rho_t, \quad \text{as } N \rightarrow \infty, \varepsilon \rightarrow 0.$$

- ▶ Evolution of ρ_t given by the following PDE (of McKean-Vlasov type):

$$\begin{aligned} \partial_t \rho_t &= \nabla_{\theta} \cdot \left(\rho_t \nabla_{\theta} \Psi(\theta; \rho_t) \right), \\ \Psi(\theta; \rho_t) &\equiv V(\theta) + \int U(\theta, \theta') \rho(d\theta'), \end{aligned}$$

where $V(\theta) := -\mathbb{E}\{y \sigma_*(\mathbf{x}; \theta)\}$ and $U(\theta_1, \theta_2) := \mathbb{E}_{\mathbf{x}}\{\sigma_*(\mathbf{x}; \theta_1) \sigma_*(\mathbf{x}; \theta_2)\}$.

This is referred to as the *mean-field* description, or *distributional dynamics* (DD).

- ▶ Wasserstein gradient flow on risk $R(\rho) := \mathbb{E}\{(y - \hat{f}(\mathbf{x}; \rho))^2\}$, with $\rho \in \mathcal{P}(\mathbb{R}^D)$

Mean-field limit

- ▶ **One-pass SGD:** training examples are never revisited, i.e., $\{(\mathbf{x}_k, y_k)\}_{k \geq 1}$ are iid.
- ▶ Denote $\hat{\rho}_k^{(N)} = N^{-1} \sum_{i \leq N} \delta_{\theta_i^k}$ after k SGD steps with step size ε and $\theta_i^0 \sim_{iid} \rho_0$:

$$\hat{\rho}_{t/\varepsilon}^{(N)} \Rightarrow \rho_t, \quad \text{as } N \rightarrow \infty, \varepsilon \rightarrow 0.$$

- ▶ Evolution of ρ_t given by the following PDE (of McKean-Vlasov type):

$$\begin{aligned} \partial_t \rho_t &= \nabla_{\theta} \cdot \left(\rho_t \nabla_{\theta} \Psi(\theta; \rho_t) \right), \\ \Psi(\theta; \rho_t) &\equiv V(\theta) + \int U(\theta, \theta') \rho(d\theta'), \end{aligned}$$

where $V(\theta) := -\mathbb{E}\{y \sigma_*(\mathbf{x}; \theta)\}$ and $U(\theta_1, \theta_2) := \mathbb{E}_{\mathbf{x}}\{\sigma_*(\mathbf{x}; \theta_1) \sigma_*(\mathbf{x}; \theta_2)\}$.

This is referred to as the *mean-field* description, or *distributional dynamics* (DD).

- ▶ Wasserstein gradient flow on risk $R(\rho) := \mathbb{E}\{(y - \hat{f}(\mathbf{x}; \rho))^2\}$, with $\rho \in \mathcal{P}(\mathbb{R}^D)$

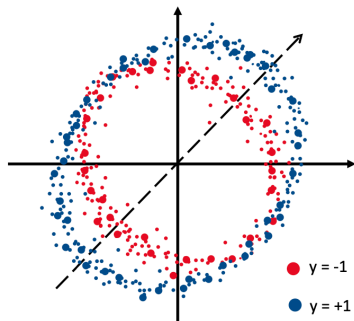
Example: classifying centered anisotropic Gaussians (I)

Data distribution (y, \mathbf{x}) :

With proba 1/2: $y = +1$, $\mathbf{x} \sim N(0, \Sigma_+)$,

With proba 1/2: $y = -1$, $\mathbf{x} \sim N(0, \Sigma_-)$,

- $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$.
- $\mathbf{U} \in \mathbb{R}^{d \times d}$ orthogonal matrix.
- $\mathbf{P}_{\mathcal{V}}$: projection on the subspace $\mathcal{V} = \text{span}(\mathbf{U}_{1:s_0})$.



Consider activation function $\sigma_*(\mathbf{x}; \theta) = \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$ (i.e., $\theta = \mathbf{w} \in \mathbb{R}^d$).

Goal: study SGD on

$$R_N(\theta) = \mathbb{E} \left\{ \left(y - \frac{1}{N} \sum_{i=1}^N \sigma(\langle \mathbf{x}, \mathbf{w}_i \rangle) \right)^2 \right\}.$$

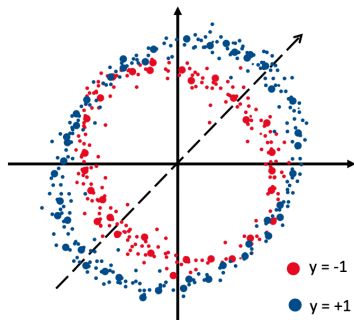
Example: classifying centered anisotropic Gaussians (I)

Data distribution (y, \mathbf{x}) :

With proba 1/2: $y = +1$, $\mathbf{x} \sim N(0, \Sigma_+)$,

With proba 1/2: $y = -1$, $\mathbf{x} \sim N(0, \Sigma_-)$,

- $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$.
- $\mathbf{U} \in \mathbb{R}^{d \times d}$ orthogonal matrix.
- $\mathbf{P}_{\mathcal{V}}$: projection on the subspace $\mathcal{V} = \text{span}(\mathbf{U}_{1:s_0})$.



Consider activation function $\sigma_*(\mathbf{x}; \boldsymbol{\theta}) = \sigma(\langle \mathbf{x}, \mathbf{w} \rangle)$ (i.e., $\boldsymbol{\theta} = \mathbf{w} \in \mathbb{R}^d$).

Goal: study SGD on

$$R_N(\boldsymbol{\theta}) = \mathbb{E} \left\{ \left(y - \frac{1}{N} \sum_{i=1}^N \sigma(\langle \mathbf{x}_i, \mathbf{w}_i \rangle) \right)^2 \right\}.$$

Example: classifying centered anisotropic Gaussians (II)

Mean-field description of SGD in this problem:

- ▶ $(\mathbf{x}, y) \sim \mathbb{P}$ is invariant under $\mathcal{O}(\mathcal{V}) \times \mathcal{O}(\mathcal{V}^\perp)$.
- ▶ Denote $r_1 := \|\mathbf{P}_\mathcal{V} \mathbf{w}\|_2$ and $r_2 := \|(\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}\|_2$. If ρ_0 is spherically symmetric, solution ρ_t of DD remains uniform conditional on r_1, r_2 :

$$\rho_t(\mathbf{w}) = \bar{\rho}_t(r_1, r_2) \times \mu_{s_0}(\mathbf{P}_\mathcal{V} \mathbf{w}/r_1) \times \mu_{d-s_0}((\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}/r_2), \quad \mu_p \equiv \text{Unif}(\mathbb{S}^{p-1}).$$

(global optimum must be of this form by Jensen's inequality).

- ▶ PDE on $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ simplifies to PDE on $\bar{\rho}_t \in \mathcal{P}([0, \infty)^2)$:

$$\partial_t \bar{\rho}_t(r_1, r_2) = \nabla \cdot (\bar{\rho}_t(r_1, r_2) \nabla \bar{\Psi}(r_1, r_2; \bar{\rho}_t)).$$

- ▶ PDE in 2 dimensions: efficient to solve numerically, can classify stationary points...
[Mei, Montanari, Nguyen, 2018] for some good initialization $\bar{\rho}_0$ and activation σ , and denoting $s_0 = \gamma d$, the PDE converges to global optimum in time $T(\Delta, \gamma, \bar{\rho}_0)$.

Example: classifying centered anisotropic Gaussians (II)

Mean-field description of SGD in this problem:

- ▶ $(\mathbf{x}, y) \sim \mathbb{P}$ is invariant under $\mathcal{O}(\mathcal{V}) \times \mathcal{O}(\mathcal{V}^\perp)$.
- ▶ Denote $r_1 := \|\mathbf{P}_\mathcal{V} \mathbf{w}\|_2$ and $r_2 := \|(\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}\|_2$. If ρ_0 is spherically symmetric, solution ρ_t of DD remains uniform conditional on r_1, r_2 :

$$\rho_t(\mathbf{w}) = \bar{\rho}_t(r_1, r_2) \times \mu_{s_0}(\mathbf{P}_\mathcal{V} \mathbf{w}/r_1) \times \mu_{d-s_0}((\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}/r_2), \quad \mu_p \equiv \text{Unif}(\mathbb{S}^{p-1}).$$

(global optimum must be of this form by Jensen's inequality).

- ▶ PDE on $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ simplifies to PDE on $\bar{\rho}_t \in \mathcal{P}([0, \infty)^2)$:

$$\partial_t \bar{\rho}_t(r_1, r_2) = \nabla \cdot (\bar{\rho}_t(r_1, r_2) \nabla \bar{\Psi}(r_1, r_2; \bar{\rho}_t)).$$

- ▶ PDE in 2 dimensions: efficient to solve numerically, can classify stationary points...

[Mei, Montanari, Nguyen, 2018] for some good initialization $\bar{\rho}_0$ and activation σ , and denoting $s_0 = \gamma d$, the PDE converges to global optimum in time $T(\Delta, \gamma, \bar{\rho}_0)$.

Example: classifying centered anisotropic Gaussians (II)

Mean-field description of SGD in this problem:

- ▶ $(\mathbf{x}, y) \sim \mathbb{P}$ is invariant under $\mathcal{O}(\mathcal{V}) \times \mathcal{O}(\mathcal{V}^\perp)$.
- ▶ Denote $r_1 := \|\mathbf{P}_\mathcal{V} \mathbf{w}\|_2$ and $r_2 := \|(\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}\|_2$. If ρ_0 is spherically symmetric, solution ρ_t of DD remains uniform conditional on r_1, r_2 :

$$\rho_t(\mathbf{w}) = \bar{\rho}_t(r_1, r_2) \times \mu_{s_0}(\mathbf{P}_\mathcal{V} \mathbf{w}/r_1) \times \mu_{d-s_0}((\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}/r_2), \quad \mu_p \equiv \text{Unif}(\mathbb{S}^{p-1}).$$

(global optimum must be of this form by Jensen's inequality).

- ▶ PDE on $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ simplifies to PDE on $\bar{\rho}_t \in \mathcal{P}([0, \infty)^2)$:

$$\partial_t \bar{\rho}_t(r_1, r_2) = \nabla \cdot (\bar{\rho}_t(r_1, r_2) \nabla \bar{\Psi}(r_1, r_2; \bar{\rho}_t)).$$

- ▶ PDE in 2 dimensions: efficient to solve numerically, can classify stationary points...

[Mei, Montanari, Nguyen, 2018] for some good initialization $\bar{\rho}_0$ and activation σ , and denoting $s_0 = \gamma d$, the PDE converges to global optimum in time $T(\Delta, \gamma, \bar{\rho}_0)$.

Example: classifying centered anisotropic Gaussians (II)

Mean-field description of SGD in this problem:

- ▶ $(\mathbf{x}, y) \sim \mathbb{P}$ is invariant under $\mathcal{O}(\mathcal{V}) \times \mathcal{O}(\mathcal{V}^\perp)$.
- ▶ Denote $r_1 := \|\mathbf{P}_\mathcal{V} \mathbf{w}\|_2$ and $r_2 := \|(\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}\|_2$. If ρ_0 is spherically symmetric, solution ρ_t of DD remains uniform conditional on r_1, r_2 :

$$\rho_t(\mathbf{w}) = \bar{\rho}_t(r_1, r_2) \times \mu_{s_0}(\mathbf{P}_\mathcal{V} \mathbf{w}/r_1) \times \mu_{d-s_0}((\text{Id} - \mathbf{P}_\mathcal{V}) \mathbf{w}/r_2), \quad \mu_p \equiv \text{Unif}(\mathbb{S}^{p-1}).$$

(global optimum must be of this form by Jensen's inequality).

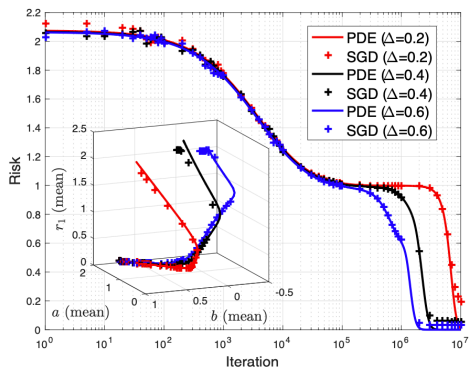
- ▶ PDE on $\rho_t \in \mathcal{P}(\mathbb{R}^d)$ simplifies to PDE on $\bar{\rho}_t \in \mathcal{P}([0, \infty)^2)$:

$$\partial_t \bar{\rho}_t(r_1, r_2) = \nabla \cdot (\bar{\rho}_t(r_1, r_2) \nabla \bar{\Psi}(r_1, r_2; \bar{\rho}_t)).$$

- ▶ PDE in 2 dimensions: efficient to solve numerically, can classify stationary points...

[Mei, Montanari, Nguyen, 2018] for some good initialization $\bar{\rho}_0$ and activation σ , and denoting $s_0 = \gamma d$, the PDE converges to global optimum in time $T(\Delta, \gamma, \bar{\rho}_0)$.

Example: classifying centered anisotropic Gaussians (III)

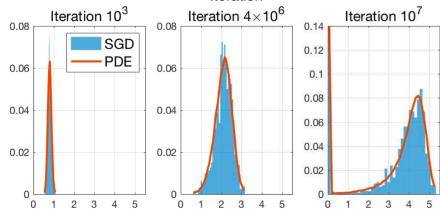


ReLU activation:

$$\sigma_*(\mathbf{x}; \theta_i) = a_i(\langle \mathbf{x}, \mathbf{w}_i \rangle + b_i)_+.$$

Evolution of some statistics:

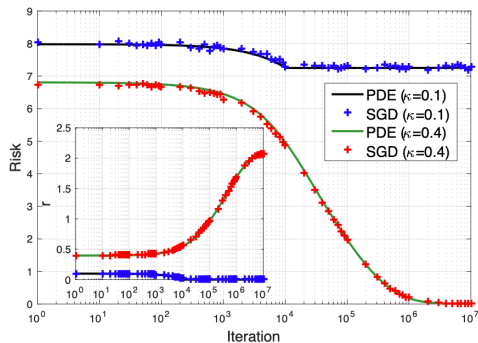
$$d = 320, s_0 = 60, N = 800, \\ \varepsilon = 2 \times 10^{-4}.$$



Evolution of $\bar{\rho}(r_1)$ for $d = s_0 = 40$,
 $N = 800$, $\Delta = 0.8$, $\varepsilon = 10^{-6}$,
 $\rho_0 = N(0, 0.8^2 \text{Id}_d/d)$.

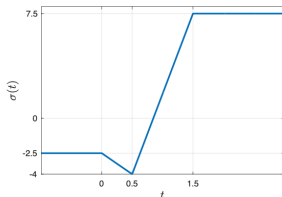
[Mei, Montanari, Nguyen, 2018]

Example: classifying centered anisotropic Gaussians (IV)



Evolution of the risk for
 $d = s_0 = 320$, $\Delta = 0.5$, $N = 800$.

Starting at two initializations:
 $N(0, \kappa^2 \text{Id}_d/d)$ with $\kappa \in \{0.1, 0.4\}$.



Non-monotonic activation function.

[Mei, Montanari, Nguyen, 2018]

Mean-field: good theory for SGD?

Mean-field description:

- ▶ Independent of N (as long as N is large enough).

Simplify the analysis of SGD:

- ▶ Factors-out some landscape complexities of NNs (e.g., permutation invariance).
- ▶ Allows to exploit symmetries in the data distribution \mathbb{P} .
- ▶ Can focus on studying the PDE (global convergence, stationary points, etc.).

For this approach to be meaningful:

In what regime is the distributional dynamics a good approximation to SGD?

Mean-field: good theory for SGD?

Mean-field description:

- ▶ Independent of N (as long as N is large enough).

Simplify the analysis of SGD:

- ▶ Factors-out some landscape complexities of NNs (e.g., permutation invariance).
- ▶ Allows to exploit symmetries in the data distribution \mathbb{P} .
- ▶ Can focus on studying the PDE (global convergence, stationary points, etc.).

For this approach to be meaningful:

In what regime is the distributional dynamics a good approximation to SGD?

Concentration of SGD process on PDE

More precisely:

- ▶ $\theta^k = (\theta_i^k)_{i \leq N}$: weights after k steps of one-pass SGD with step-size ε and $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$.
- ▶ $(\rho_t)_{t \geq 0}$: solution of the *distribution dynamics* with initialization ρ_0 .

Goal: compare population risks $R_N(\theta^k)$ and $R(\rho_t)$.

Show that for $T \geq 0$ and with probability at least $1 - \delta$,

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq \text{Error}(T, \varepsilon, N, \delta, \dots).$$

Assumptions

Take $\theta = (a, \mathbf{w})$ with $a \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^{D-1}$ and activation $\sigma_*(\mathbf{x}; \theta) = a\sigma(\mathbf{x}; \mathbf{w})$.

Denote $V(\theta) = av(\mathbf{w})$ and $U(\theta_1, \theta_2) = a_1a_2u(\mathbf{w}_1, \mathbf{w}_2)$ where

$$v(\mathbf{w}) = -\mathbb{E}\{y\sigma(\mathbf{x}; \mathbf{w})\}, \quad u(\mathbf{w}_1, \mathbf{w}_2) = \mathbb{E}_{\mathbf{x}}\{\sigma(\mathbf{x}; \mathbf{w}_1)\sigma(\mathbf{x}; \mathbf{w}_2)\}.$$

Assumptions:

A1. $\sigma : \mathbb{R}^d \times \mathbb{R}^{D-1}$ and y are bounded, i.e., $\|\sigma\|_\infty, |y| \leq K_1$.

For any \mathbf{w} , $\nabla_{\mathbf{w}}\sigma(\mathbf{x}; \mathbf{w})$ is K_1 -sub-Gaussian with respect to $\mathbf{x} \sim \mathbb{P}$.

A2. Functions $\mathbf{w} \mapsto v(\mathbf{w})$ and $(\mathbf{w}_1, \mathbf{w}_2) \mapsto u(\mathbf{w}_1, \mathbf{w}_2)$ are differentiable with bounded and Lipschitz gradients: $\|\nabla v(\mathbf{w})\|_2 \leq K_2$, $\|\nabla u(\mathbf{w}_1, \mathbf{w}_2)\|_2 \leq K_2$,

$$\|\nabla v(\mathbf{w}) - \nabla v(\mathbf{w}')\|_2 \leq K_2\|\mathbf{w} - \mathbf{w}'\|_2,$$

$$\|\nabla u(\mathbf{w}_1, \mathbf{w}_2) - \nabla u(\mathbf{w}'_1, \mathbf{w}'_2)\|_2 \leq K_2\|(\mathbf{w}_1, \mathbf{w}_2) - (\mathbf{w}'_1, \mathbf{w}'_2)\|_2.$$

A3. Initialization $\rho_0 \in \mathcal{P}(\mathbb{R}^D)$ is supported on $|a_i| \leq K_3$.

Dimension-free bound (I)

Consider two cases:

General coefficients: initialize parameters $\theta_i^0 = (a_i^0, \mathbf{w}_i^0)$ as $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$. Update both a_i and \mathbf{w}_i during the dynamics.

Fixed coefficients: same initialization but only update \mathbf{w}_i during the dynamics.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

Let σ_* verifies assumptions A1-A3. Take $T \geq 1$.

Fixed coefficients: $\exists K$ depending only on K_1-K_3 such that with proba at least $1 - e^{-z^2}$,

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq K e^{KT} \frac{1}{\sqrt{N}} [\sqrt{\log N} + z] + K e^{KT} [\sqrt{D + \log(N)} + z] \sqrt{\varepsilon}.$$

General coefficients: same result with $e^{KT} \rightarrow e^{KT^3}$.

Dimension-free bound (I)

Consider two cases:

General coefficients: initialize parameters $\theta_i^0 = (a_i^0, \mathbf{w}_i^0)$ as $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$. Update both a_i and \mathbf{w}_i during the dynamics.

Fixed coefficients: same initialization but only update \mathbf{w}_i during the dynamics.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

Let σ_* verifies assumptions A1-A3. Take $T \geq 1$.

Fixed coefficients: $\exists K$ depending only on K_1-K_3 such that with proba at least $1 - e^{-z^2}$,

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq Ke^{KT} \frac{1}{\sqrt{N}} [\sqrt{\log N} + z] + Ke^{KT} [\sqrt{D + \log(N)} + z] \sqrt{\varepsilon}.$$

General coefficients: same result with $e^{KT} \rightarrow e^{KT^3}$.

Dimension-free bound (II)

With probability at least $1 - 1/N$:

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq \underbrace{Ke^{KT} \sqrt{\frac{\log N}{N}}}_{\text{error due to finite } N} + \underbrace{Ke^{KT} \sqrt{D + \log(N)} \sqrt{\varepsilon}}_{\text{error due to discretization } \varepsilon > 0}.$$

Provided $T, K = O(1)$, the mean-field approximation is accurate for

- ▶ *Number of neurons:* $N \gg 1$ independent of D , and only depends on intrinsic properties of the activation and data distribution.
- ▶ *Step-size:* $\varepsilon \ll 1/D$.

'Dimension-free bound': N does not depend directly on D .

The K_i 's in the assumption can potentially depend on D . However in a number of settings of interests, the K_i 's are independent of D .

Dimension-free bound (II)

With probability at least $1 - 1/N$:

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq \underbrace{Ke^{KT} \sqrt{\frac{\log N}{N}}}_{\text{error due to finite } N} + \underbrace{Ke^{KT} \sqrt{D + \log(N)} \sqrt{\varepsilon}}_{\text{error due to discretization } \varepsilon > 0}.$$

Provided $T, K = O(1)$, the mean-field approximation is accurate for

- ▶ *Number of neurons*: $N \gg 1$ independent of D , and only depends on intrinsic properties of the activation and data distribution.
- ▶ *Step-size*: $\varepsilon \ll 1/D$.

'Dimension-free bound': N does not depend directly on D .

The K_i 's in the assumption can potentially depend on D . However in a number of settings of interests, the K_i 's are independent of D .

Dimension-free bound (II)

With probability at least $1 - 1/N$:

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq \underbrace{Ke^{KT} \sqrt{\frac{\log N}{N}}}_{\text{error due to finite } N} + \underbrace{Ke^{KT} \sqrt{D + \log(N)} \sqrt{\varepsilon}}_{\text{error due to discretization } \varepsilon > 0}.$$

Provided $T, K = O(1)$, the mean-field approximation is accurate for

- ▶ *Number of neurons*: $N \gg 1$ independent of D , and only depends on intrinsic properties of the activation and data distribution.
- ▶ *Step-size*: $\varepsilon \ll 1/D$.

'Dimension-free bound': N does not depend directly on D .

The K_i 's in the assumption can potentially depend on D . However in a number of settings of interests, the K_i 's are independent of D .

Application: classifying centered anisotropic Gaussians (V)

- ▶ $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$ with $s_0 = \gamma d$ and $\gamma \in (0, 1)$.
- ▶ $\sigma(t) = 0$ for $t \leq 0$, $\sigma(t) = 1$ for $t \geq 1$, and $\sigma(t) = t$ for $0 \leq t \leq 1$ (truncated ReLU).
- ▶ $(\mathbf{w}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \text{Unif}(\mathbb{S}^{d-1})$ with $\bar{\rho}_0$ with bounded density and $R_{d=\infty}(\bar{\rho}_0) < 1$.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

For any $\eta, \Delta, \delta > 0$, there exists

$$d_0 \equiv d_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad C_0 \equiv C_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad T \equiv T(\eta, \bar{\rho}_0, \Delta, \gamma),$$

such that for $d \geq d_0$, $N \geq C_0$ and $\varepsilon \leq 1/(C_0 d)$, we have at $k = T/\varepsilon$ with probability at least $1 - \delta$,

$$R_N(\theta^k) \leq \inf_{\rho} R(\rho) + \eta/2 \leq \inf_{\theta \in \mathbb{R}^{N \times d}} R_N(\theta) + \eta.$$

For $N = O(1)$ and $n = O(d)$, one-pass SGD finds a near global minimizer of the population risk (near global minimizer over all two-layers neural networks).

Application: classifying centered anisotropic Gaussians (V)

- ▶ $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$ with $s_0 = \gamma d$ and $\gamma \in (0, 1)$.
- ▶ $\sigma(t) = 0$ for $t \leq 0$, $\sigma(t) = 1$ for $t \geq 1$, and $\sigma(t) = t$ for $0 \leq t \leq 1$ (truncated ReLU).
- ▶ $(\mathbf{w}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \text{Unif}(\mathbb{S}^{d-1})$ with $\bar{\rho}_0$ with bounded density and $R_{d=\infty}(\bar{\rho}_0) < 1$.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

For any $\eta, \Delta, \delta > 0$, there exists

$$d_0 \equiv d_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad C_0 \equiv C_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad T \equiv T(\eta, \bar{\rho}_0, \Delta, \gamma),$$

such that for $d \geq d_0$, $N \geq C_0$ and $\varepsilon \leq 1/(C_0 d)$, we have at $k = T/\varepsilon$ with probability at least $1 - \delta$,

$$R_N(\theta^k) \leq \inf_{\rho} R(\rho) + \eta/2 \leq \inf_{\theta \in \mathbb{R}^{N \times d}} R_N(\theta) + \eta.$$

For $N = O(1)$ and $n = O(d)$, one-pass SGD finds a near global minimizer of the population risk (near global minimizer over all two-layers neural networks).

Application: classifying centered anisotropic Gaussians (V)

- ▶ $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$ with $s_0 = \gamma d$ and $\gamma \in (0, 1)$.
- ▶ $\sigma(t) = 0$ for $t \leq 0$, $\sigma(t) = 1$ for $t \geq 1$, and $\sigma(t) = t$ for $0 \leq t \leq 1$ (truncated ReLU).
- ▶ $(\mathbf{w}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \text{Unif}(\mathbb{S}^{d-1})$ with $\bar{\rho}_0$ with bounded density and $R_{d=\infty}(\bar{\rho}_0) < 1$.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

For any $\eta, \Delta, \delta > 0$, there exists

$$d_0 \equiv d_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad C_0 \equiv C_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad T \equiv T(\eta, \bar{\rho}_0, \Delta, \gamma),$$

such that for $d \geq d_0$, $N \geq C_0$ and $\varepsilon \leq 1/(C_0 d)$, we have at $k = T/\varepsilon$ with probability at least $1 - \delta$,

$$R_N(\theta^k) \leq \inf_{\rho} R(\rho) + \eta/2 \leq \inf_{\theta \in \mathbb{R}^{N \times d}} R_N(\theta) + \eta.$$

For $N = O(1)$ and $n = O(d)$, one-pass SGD finds a near global minimizer of the population risk (near global minimizer over all two-layers neural networks).

Application: classifying centered anisotropic Gaussians (V)

- ▶ $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$ with $s_0 = \gamma d$ and $\gamma \in (0, 1)$.
- ▶ $\sigma(t) = 0$ for $t \leq 0$, $\sigma(t) = 1$ for $t \geq 1$, and $\sigma(t) = t$ for $0 \leq t \leq 1$ (truncated ReLU).
- ▶ $(\mathbf{w}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \text{Unif}(\mathbb{S}^{d-1})$ with $\bar{\rho}_0$ with bounded density and $R_{d=\infty}(\bar{\rho}_0) < 1$.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

For any $\eta, \Delta, \delta > 0$, there exists

$$d_0 \equiv d_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad C_0 \equiv C_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad T \equiv T(\eta, \bar{\rho}_0, \Delta, \gamma),$$

such that for $d \geq d_0$, $N \geq C_0$ and $\varepsilon \leq 1/(C_0 d)$, we have at $k = T/\varepsilon$ with probability at least $1 - \delta$,

$$R_N(\boldsymbol{\theta}^k) \leq \inf_{\rho} R(\rho) + \eta/2 \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times d}} R_N(\boldsymbol{\theta}) + \eta.$$

For $N = O(1)$ and $n = O(d)$, one-pass SGD finds a near global minimizer of the population risk (near global minimizer over all two-layers neural networks).

Application: classifying centered anisotropic Gaussians (V)

- ▶ $\Sigma_{\pm} = \mathbf{U} \text{diag}((1 \pm \Delta)^2 \cdot \text{Id}_{s_0}, \text{Id}_{d-s_0}) \mathbf{U}^T$ with $s_0 = \gamma d$ and $\gamma \in (0, 1)$.
- ▶ $\sigma(t) = 0$ for $t \leq 0$, $\sigma(t) = 1$ for $t \geq 1$, and $\sigma(t) = t$ for $0 \leq t \leq 1$ (truncated ReLU).
- ▶ $(\mathbf{w}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \text{Unif}(\mathbb{S}^{d-1})$ with $\bar{\rho}_0$ with bounded density and $R_{d=\infty}(\bar{\rho}_0) < 1$.

Theorem (Mei, Misiakiewicz, Montanari, 2019)

For any $\eta, \Delta, \delta > 0$, there exists

$$d_0 \equiv d_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad C_0 \equiv C_0(\eta, \bar{\rho}_0, \Delta, \gamma), \quad T \equiv T(\eta, \bar{\rho}_0, \Delta, \gamma),$$

such that for $d \geq d_0$, $N \geq C_0$ and $\varepsilon \leq 1/(C_0 d)$, we have at $k = T/\varepsilon$ with probability at least $1 - \delta$,

$$R_N(\boldsymbol{\theta}^k) \leq \inf_{\rho} R(\rho) + \eta/2 \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^{N \times d}} R_N(\boldsymbol{\theta}) + \eta.$$

For $N = O(1)$ and $n = O(d)$, one-pass SGD finds a near global minimizer of the **population risk** (near global minimizer over all two-layers neural networks).

Strategy to prove (quantitative) global convergence of SGD:

- ▶ *Global convergence of the PDE* (e.g., [Chizat, Bach, 2018]).
- ▶ *Bound the time to convergence T_c* (e.g., [Javanmard, Mondelli, Montanari, 2019]).
- ▶ *Bound between SGD and PDE*: unfortunately, current bound is e^{KT}/\sqrt{N} , hence it is non-vacuous only if $T_c \ll \log(N)$.

Strategy already applies to some non-trivial examples (e.g., anisotropic Gaussians).

Noisy SGD and entropic regularization

- ▶ Noisy SGD:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \boldsymbol{\theta}^k)) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\mathbf{x}_k; \boldsymbol{\theta}_i^k) + \sqrt{\varepsilon/\beta} \cdot \mathbf{g}_i^k,$$

where $\mathbf{g}_i^k \sim_{iid} N(0, \text{Id}_D)$.

- ▶ Mean-field description:

$$\partial_t \rho_t = \nabla_{\boldsymbol{\theta}} \cdot (\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)) + \frac{1}{\beta} \Delta_{\boldsymbol{\theta}} \rho_t.$$

- ▶ Wasserstein gradient flow for the free energy: $F_{\beta}(\rho) = R(\rho) + \frac{1}{\beta} \int \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) d\boldsymbol{\theta}$.
- ▶ [Mei, Montanari, Nguyen, 2018] convergence $\rho_t \Rightarrow \rho_*^{\beta}$ global minimizer $F_{\beta}(\rho)$.
Under some regularity conditions,

$$R(\rho_*^{\beta}) \leq \inf_{\rho} R(\rho) + O\left(\frac{D}{\beta}\right).$$

- ▶ Noisy SGD:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \boldsymbol{\theta}^k)) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\mathbf{x}_k; \boldsymbol{\theta}_i^k) + \sqrt{\varepsilon/\beta} \cdot \mathbf{g}_i^k,$$

where $\mathbf{g}_i^k \sim_{iid} N(0, \text{Id}_D)$.

- ▶ Mean-field description:

$$\partial_t \rho_t = \nabla_{\boldsymbol{\theta}} \cdot (\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)) + \frac{1}{\beta} \Delta_{\boldsymbol{\theta}} \rho_t.$$

- ▶ Wasserstein gradient flow for the free energy: $F_{\beta}(\rho) = R(\rho) + \frac{1}{\beta} \int \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) d\boldsymbol{\theta}$.
- ▶ [Mei, Montanari, Nguyen, 2018] convergence $\rho_t \Rightarrow \rho_*^{\beta}$ global minimizer $F_{\beta}(\rho)$.
Under some regularity conditions,

$$R(\rho_*^{\beta}) \leq \inf_{\rho} R(\rho) + O\left(\frac{D}{\beta}\right).$$

Noisy SGD and entropic regularization

- ▶ Noisy SGD:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \boldsymbol{\theta}^k)) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\mathbf{x}_k; \boldsymbol{\theta}_i^k) + \sqrt{\varepsilon/\beta} \cdot \mathbf{g}_i^k,$$

where $\mathbf{g}_i^k \sim_{iid} N(0, \text{Id}_D)$.

- ▶ Mean-field description:

$$\partial_t \rho_t = \nabla_{\boldsymbol{\theta}} \cdot (\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)) + \frac{1}{\beta} \Delta_{\boldsymbol{\theta}} \rho_t.$$

- ▶ Wasserstein gradient flow for the free energy: $F_{\beta}(\rho) = R(\rho) + \frac{1}{\beta} \int \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) d\boldsymbol{\theta}$.

- ▶ [Mei, Montanari, Nguyen, 2018] convergence $\rho_t \Rightarrow \rho_*^{\beta}$ global minimizer $F_{\beta}(\rho)$.
Under some regularity conditions,

$$R(\rho_*^{\beta}) \leq \inf_{\rho} R(\rho) + O\left(\frac{D}{\beta}\right).$$

- ▶ Noisy SGD:

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k + \varepsilon(y_k - \hat{f}_N(\mathbf{x}_k; \boldsymbol{\theta}^k)) \nabla_{\boldsymbol{\theta}_i} \sigma_*(\mathbf{x}_k; \boldsymbol{\theta}_i^k) + \sqrt{\varepsilon/\beta} \cdot \mathbf{g}_i^k,$$

where $\mathbf{g}_i^k \sim_{iid} N(0, \text{Id}_D)$.

- ▶ Mean-field description:

$$\partial_t \rho_t = \nabla_{\boldsymbol{\theta}} \cdot (\rho_t \nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho_t)) + \frac{1}{\beta} \Delta_{\boldsymbol{\theta}} \rho_t.$$

- ▶ Wasserstein gradient flow for the free energy: $F_{\beta}(\rho) = R(\rho) + \frac{1}{\beta} \int \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) d\boldsymbol{\theta}$.
- ▶ [Mei, Montanari, Nguyen, 2018] convergence $\rho_t \Rightarrow \rho_*^{\beta}$ global minimizer $F_{\beta}(\rho)$.
Under some regularity conditions,

$$R(\rho_*^{\beta}) \leq \inf_{\rho} R(\rho) + O\left(\frac{D}{\beta}\right).$$

Theorem (Mei, Misiakiewicz, Montanari, 2019)

Let σ_* verifies assumptions A1-A3, $\tau \leq K_4$ and $T \geq 1$.

Fixed coefficients: $\exists K$ depending only on K_1 - K_4 such that with proba at least $1 - e^{-z^2}$

$$\sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \leq Ke^{KT} \frac{1}{\sqrt{N}} [\sqrt{\log N} + z] + Ke^{KT} [\sqrt{D + \log(N)} + z] \sqrt{\varepsilon}.$$

General coefficients: $\exists K$ depending only on K_1 - K_4 such that with proba at least $1 - e^{-z^2}$

$$\begin{aligned} \sup_{k \in [0, T/\varepsilon] \cup \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| &\leq Ke^{e^{KT} [\sqrt{\log N} + z^2]} [\sqrt{D \log N} + \log^{3/2}(NT) + z^5] / \sqrt{N} \\ &\quad + Ke^{e^{KT} [\sqrt{\log N} + z^2]} [\sqrt{D} \log(NT/\varepsilon) + \log^{3/2}(N) + z^6] \sqrt{\varepsilon}. \end{aligned}$$

General coefficients: harder to control. The bound is not dimension-free and only allows us to control the approximation error up to $T = o(\log \log N)$ instead of $T = o(\log N)$.

Outline of the proof of the non-asymptotic bound (I)

Ingredients: isolating different error terms + coupling + concentration-of-measure.

Consider four coupled dynamics:

- ▶ *Nonlinear dynamics (ND):* $\bar{\theta}_i^0 \sim_{iid} \rho_0$,

$$\frac{d}{dt} \bar{\theta}_i^t = - \left[\nabla V(\bar{\theta}_i^t) + \int \nabla_1 U(\bar{\theta}_i^t, \theta) \rho_t(d\theta) \right].$$

- ▶ *Particle dynamics (PD):* $\underline{\theta}_i^0 = \bar{\theta}_i^0$,

$$\frac{d}{dt} \underline{\theta}_i^t = - \left[\nabla V(\underline{\theta}_i^t) + \frac{1}{N} \sum_{j=1}^N \nabla_1 U(\underline{\theta}_i^t, \underline{\theta}_j^t) \right].$$

- ▶ *Gradient descent (GD):* $\tilde{\theta}_i^0 = \bar{\theta}_i^0$,

$$\tilde{\theta}_i^{k+1} = \tilde{\theta}_i^k - \varepsilon \left[\nabla V(\tilde{\theta}_i^k) + \frac{1}{N} \sum_{j=1}^N \nabla_1 U(\tilde{\theta}_i^k, \tilde{\theta}_j^k) \right].$$

- ▶ *Stochastic gradient descent (SGD):* $\theta_i^0 = \bar{\theta}_i^0$,

$$\theta_i^{k+1} = \theta_i^k - \varepsilon \left[-y_k \nabla_{\theta} \sigma_*(\mathbf{x}_k; \theta_i^k) + \frac{1}{N} \sum_{j=1}^N \sigma_*(\mathbf{x}_k; \theta_j^k) \nabla_{\theta} \sigma_*(\mathbf{x}_k; \theta_i^k) \right].$$

Outline of the proof of the non-asymptotic bound (II)

$$\begin{aligned} \left| R(\rho_{k\varepsilon}) - R_N(\boldsymbol{\theta}^k) \right| &\leq \underbrace{\left| R(\rho_{k\varepsilon}) - R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon}) \right|}_{\text{PDE-ND}} + \underbrace{\left| R_N(\bar{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) \right|}_{\text{ND-PD}} \\ &\quad + \underbrace{\left| R_N(\underline{\boldsymbol{\theta}}^{k\varepsilon}) - R_N(\tilde{\boldsymbol{\theta}}^k) \right|}_{\text{PD-GD}} + \underbrace{\left| R_N(\tilde{\boldsymbol{\theta}}^k) - R_N(\boldsymbol{\theta}^k) \right|}_{\text{GD-SGD}}. \end{aligned}$$

- ▶ **PDE-ND:** $\bar{\boldsymbol{\theta}}^{k\varepsilon} \sim_{iid} \rho_{k\varepsilon}$ + McDiarmid's inequality.
- ▶ **ND-PD:** McDiarmid's inequality + Gronwall's inequality.
- ▶ **PD-GD:** Lipschitzness + Gronwall's lemma.
- ▶ **GD-SGD:** Azuma-Hoeffding inequality + Gronwall's lemma.

Details in:

Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit.
Mei, Misiakiewicz, Montanari, COLT 2019.

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to focus on key elements of the dynamics (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to **focus on key elements of the dynamics** (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to **focus on key elements of the dynamics** (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to **focus on key elements of the dynamics** (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to **focus on key elements of the dynamics** (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!

Final remarks

- ▶ **Mean-field theory:** describe SGD for $N \rightarrow \infty, \varepsilon \rightarrow 0$, in terms of a PDE in the space of probability distributions.
- ▶ It allows to **focus on key elements of the dynamics** (global convergence, stationary points), and in some cases vastly simplifies the analysis of SGD.
- ▶ **Dimension-free bounds:** to approximate the SGD dynamics by the distributional dynamics, we only need $N = O(1)$ that depends on intrinsic properties of the activation and data distribution, and $\varepsilon = O(1/D)$.
- ▶ Capturing the **correct dimension-dependence** is crucial in order to compare neural networks to other learning techniques.
- ▶ Lots of open problems: global convergence guarantees, bounding convergence time, improving the exponential dependency in T , multi-layer etc.

Thank you!