A Mean Field Theory of Two-Layers Neural Networks

Song Mei

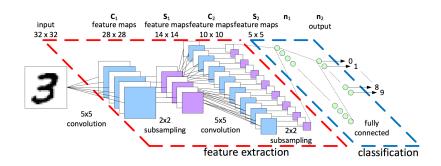
Stanford University

January 16, 2019

Joint work with Andrea Montanari and Phan-Minh Nguyen



Applications of neuralnets



- ► Computer vision (video surveillance).
- Generative modeling (generating arts).
- ▶ Reinforcement learning (robotics).

Usually # parameter much larger than # data (overparameterization).

Usually # parameter much larger than # data (overparameterization).

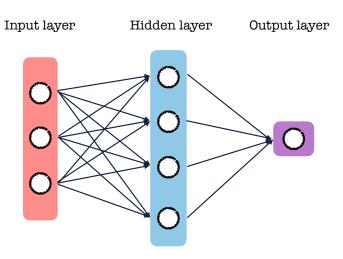
The mystery

- Optimized efficiently.
- ► Generalize well.

Usually # parameter much larger than # data (overparameterization).

The mystery

- Optimized efficiently. Why not trapped at bad local min?
- ▶ Generalize well. Why not overfitting?



- ▶ Parameter: $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times D}$.
- ▶ Prediction:

$$\hat{y}(x;oldsymbol{ heta}) = rac{1}{N} \sum_{i=1}^N \sigma_{\star}(x;oldsymbol{ heta}_i)$$

- lacksquare An example: $oldsymbol{ heta}_i = (oldsymbol{a}_i, oldsymbol{w}_i), \ \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta}_i) = oldsymbol{a}_i \sigma(\langle oldsymbol{x}, oldsymbol{w}_i
 angle).$
- ▶ Data distribution: $(x, y) \sim \mathbb{P}_{x,y}$.
- ▶ Risk function:

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y}ig[ig(y-rac{1}{N}\sum_{i=1}^N\sigma_{\star}(oldsymbol{x};oldsymbol{ heta}_i)ig)^2ig]$$

- ▶ Parameter: $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times D}$.
- ▶ Prediction:

$$\hat{y}(oldsymbol{x};oldsymbol{ heta}) = rac{1}{N} \sum_{i=1}^N \sigma_{\star}(oldsymbol{x};oldsymbol{ heta_i}).$$

- lacktriangle An example: $m{\theta}_i = (a_i, w_i), \ \sigma_{\star}(x; m{\theta}_i) = a_i \sigma(\langle x, w_i \rangle).$
- ▶ Data distribution: $(x, y) \sim \mathbb{P}_{x,y}$.
- ▶ Risk function

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} ig[ig(y - rac{1}{N} \sum_{i=1}^N \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta}_i) ig)^2 ig]$$

- ▶ Parameter: $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times D}$.
- ▶ Prediction:

$$\hat{y}(x; oldsymbol{ heta}) = rac{1}{N} \sum_{i=1}^N \sigma_{\star}(x; oldsymbol{ heta}_i).$$

- ▶ An example: $\theta_i = (a_i, w_i), \ \sigma_{\star}(x; \theta_i) = a_i \sigma(\langle x, w_i \rangle).$
- ▶ Data distribution: $(x, y) \sim \mathbb{P}_{x,y}$.
- ▶ Risk function:

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} ig[ig(y - rac{1}{N} \sum_{i=1}^N \sigma_\star(oldsymbol{x}; oldsymbol{ heta}_i) ig)^2 ig]$$

- ▶ Parameter: $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times D}$.
- ▶ Prediction:

$$\hat{y}(x; oldsymbol{ heta}) = rac{1}{N} \sum_{i=1}^N \sigma_{\star}(x; oldsymbol{ heta}_i).$$

- An example: $\theta_i = (a_i, w_i), \ \sigma_{\star}(x; \theta_i) = a_i \sigma(\langle x, w_i \rangle).$
- ▶ Data distribution: $(x, y) \sim \mathbb{P}_{x,y}$.
- Risk function

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} \Big[\Big(y - rac{1}{N} \sum_{i=1}^N \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta}_i) \Big)^2 \Big]$$



- ▶ Parameter: $\theta = (\theta_1, \dots, \theta_N) \in \mathbb{R}^{N \times D}$.
- ▶ Prediction:

$$\hat{y}(x; oldsymbol{ heta}) = rac{1}{N} \sum_{i=1}^N \sigma_{\star}(x; oldsymbol{ heta}_i).$$

- $\qquad \text{An example: } \boldsymbol{\theta}_i = (\boldsymbol{a}_i, \boldsymbol{w}_i), \ \boldsymbol{\sigma}_{\star}(\boldsymbol{x}; \boldsymbol{\theta}_i) = \boldsymbol{a}_i \boldsymbol{\sigma}(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle).$
- ▶ Data distribution: $(x, y) \sim \mathbb{P}_{x,y}$.
- ▶ Risk function:

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} ig[ig(y - rac{1}{N} \sum_{i=1}^N \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta_i}) ig)^2 ig].$$



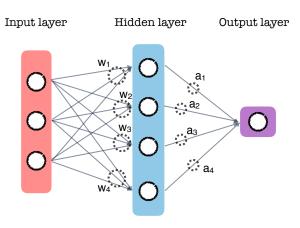


Figure: $\boldsymbol{\theta}_i = (a_i, \boldsymbol{w}_i)$.

Related literatures (before 2018)

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} \Big[\Big(y - rac{1}{N} \sum_{j=1}^N \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta_j}) \Big)^2 \Big].$$

- ▶ Optimization based on landscape analysis: [Soudry, Carmon, 2016], [Freeman, Bruna, 2016], [Ge, Lee, Ma, 2017], [Soltanolkotabi, Javanmard, Lee, 2017], [Zhong, Song, Jain, Bartlett, Dhillon, 2017], [Tian, 2017], [Soltanolkotabi, 2017], [Li, Yuan, 2017]...
- ► Generalization based on margin theory: [Bartlett, Foster, Telgarsky, 2017], [Neyshabur, Bhojanapalli, McAllester, Srebro, 2017]...

Related literatures (before 2018)

$$R_N(oldsymbol{ heta}) = \mathbb{E}_{oldsymbol{x},y} \Big[\Big(y - rac{1}{N} \sum_{j=1}^N \sigma_{\star}(oldsymbol{x}; oldsymbol{ heta}_j) \Big)^2 \Big].$$

- ▶ Optimization based on landscape analysis: [Soudry, Carmon, 2016], [Freeman, Bruna, 2016], [Ge, Lee, Ma, 2017], [Soltanolkotabi, Javanmard, Lee, 2017], [Zhong, Song, Jain, Bartlett, Dhillon, 2017], [Tian, 2017], [Soltanolkotabi, 2017], [Li, Yuan, 2017]...
- ► Generalization based on margin theory: [Bartlett, Foster, Telgarsky, 2017], [Neyshabur, Bhojanapalli, McAllester, Srebro, 2017]...

▶ [Bengio, et. al, 2006]. Expand the square

$$R_N(oldsymbol{ heta}) = \mathbb{E}[y^2] + rac{2}{N} \sum_{i=1}^N V(oldsymbol{ heta}_i) + rac{1}{N^2} \sum_{i,j=1}^N U(oldsymbol{ heta}_i, oldsymbol{ heta}_j),$$

$$egin{aligned} V(oldsymbol{ heta}_i) &= - \mathbb{E}[y \sigma_{\star}(x; oldsymbol{ heta}_i)], \ U(oldsymbol{ heta}_i, oldsymbol{ heta}_j) &= & \mathbb{E}[\sigma_{\star}(x; oldsymbol{ heta}_i) \sigma_{\star}(x; oldsymbol{ heta}_j)]. \end{aligned}$$

- ▶ R_N depends on $(\theta_i)_{i \leq N}$ through $\rho_N = (1/N) \sum_{i=1}^N \delta_{\theta_i}$.
- ▶ Motivate us to define $R(\rho)$, $\rho \in \mathcal{P}(\mathbb{R}^D)$,

$$R(
ho) = \mathbb{E}[y^2] + 2 \int V(oldsymbol{ heta})
ho(\mathrm{d}oldsymbol{ heta}) + \int U(oldsymbol{ heta}_1, oldsymbol{ heta}_2)
ho(\mathrm{d}oldsymbol{ heta}_1)
ho(\mathrm{d}oldsymbol{ heta}_2).$$



▶ [Bengio, et. al, 2006]. Expand the square

$$R_N(oldsymbol{ heta}) = \mathbb{E}[y^2] + rac{2}{N} \sum_{i=1}^N V(oldsymbol{ heta}_i) + rac{1}{N^2} \sum_{i,j=1}^N U(oldsymbol{ heta}_i, oldsymbol{ heta}_j),$$

$$egin{aligned} V(oldsymbol{ heta}_i) &= - \mathbb{E}[y\sigma_{\star}(x;oldsymbol{ heta}_i)], \ U(oldsymbol{ heta}_i,oldsymbol{ heta}_j) = & \mathbb{E}[\sigma_{\star}(x;oldsymbol{ heta}_i)\sigma_{\star}(x;oldsymbol{ heta}_j)]. \end{aligned}$$

- $ightharpoonup R_N$ depends on $(\theta_i)_{i\leq N}$ through $ho_N=(1/N)\sum_{i=1}^N \delta_{\theta_i}$.
- ▶ Motivate us to define $R(\rho)$, $\rho \in \mathcal{P}(\mathbb{R}^D)$,

$$R(\rho) = \mathbb{E}[y^2] + 2 \int V(\boldsymbol{\theta}) \rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \rho(\mathrm{d}\boldsymbol{\theta}_1) \rho(\mathrm{d}\boldsymbol{\theta}_2).$$



[Bengio, et. al, 2006]. Expand the square

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

$$egin{aligned} V(oldsymbol{ heta}_i) &= - \mathbb{E}[y\sigma_\star(x;oldsymbol{ heta}_i)], \ U(oldsymbol{ heta}_i,oldsymbol{ heta}_j) = & \mathbb{E}[\sigma_\star(x;oldsymbol{ heta}_i)\sigma_\star(x;oldsymbol{ heta}_j)]. \end{aligned}$$

- $ightharpoonsight R_N$ depends on $(\theta_i)_{i \leq N}$ through $\rho_N = (1/N) \sum_{i=1}^N \delta_{\theta_i}$.
- ▶ Motivate us to define $R(\rho)$, $\rho \in \mathcal{P}(\mathbb{R}^D)$,

$$R(
ho) = \mathbb{E}[y^2] + 2 \int V(heta)
ho(\mathrm{d} heta) + \int U(heta_1, heta_2)
ho(\mathrm{d} heta_1)
ho(\mathrm{d} heta_2).$$



Correspondence $R_N(\theta) = R((1/N) \sum_{i=1}^N \delta_{\theta_i})$, where

$$egin{aligned} R_N(oldsymbol{ heta}) =& \mathbb{E}[y^2] + rac{2}{N} \sum_{i=1}^N V(oldsymbol{ heta}_i) + rac{1}{N^2} \sum_{i,j=1}^N U(oldsymbol{ heta}_i, oldsymbol{ heta}_j), \ R(oldsymbol{
ho}) =& \mathbb{E}[y^2] + 2 \int V(oldsymbol{ heta})
ho(\mathrm{d}oldsymbol{ heta}) + \int U(oldsymbol{ heta}_1, oldsymbol{ heta}_2)
ho(\mathrm{d}oldsymbol{ heta}_1)
ho(\mathrm{d}oldsymbol{ heta}_2). \end{aligned}$$

Correspondence $R_N(\theta) = R((1/N) \sum_{i=1}^N \delta_{\theta_i})$, where

$$\begin{split} R_N(\boldsymbol{\theta}) = & \mathbb{E}[y^2] + \frac{2}{N} \sum_{i=1}^N V(\boldsymbol{\theta_i}) + \frac{1}{N^2} \sum_{i,j=1}^N U(\boldsymbol{\theta_i}, \boldsymbol{\theta_j}), \\ R(\boldsymbol{\rho}) = & \mathbb{E}[y^2] + 2 \int V(\boldsymbol{\theta}) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta_1}) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta_2}). \end{split}$$

What is the relationship of minimum value of R_N and R?

Correspondence $R_N(\theta) = R((1/N) \sum_{i=1}^N \delta_{\theta_i})$, where

$$\begin{split} R_N(\boldsymbol{\theta}) = & \mathbb{E}[y^2] + \frac{2}{N} \sum_{i=1}^N V(\boldsymbol{\theta}_i) + \frac{1}{N^2} \sum_{i,j=1}^N U(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j), \\ R(\boldsymbol{\rho}) = & \mathbb{E}[y^2] + 2 \int V(\boldsymbol{\theta}) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta}_1) \boldsymbol{\rho}(\mathrm{d}\boldsymbol{\theta}_2). \end{split}$$

What is the relationship of minimum value of R_N and R?

Lemma

If U bounded, then

$$\inf_{\rho} R(\rho) \leq \inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) \leq \inf_{\rho} R(\rho) + O(1/N).$$

How to optimize $R(\rho)$?

[Bengio, et. al, 2006] proposed to optimize over ρ

$$R(
ho) = \mathbb{E}[y^2] + 2 \int V(heta)
ho(\mathrm{d} heta) + \int U(heta_1, heta_2)
ho(\mathrm{d} heta_1)
ho(\mathrm{d} heta_2).$$

Exponential bases functions to discretize ρ !

How to optimize $R(\rho)$?

[Bengio, et. al, 2006] proposed to optimize over ρ

$$R(
ho) = \mathbb{E}[y^2] + 2 \int V(heta)
ho(\mathrm{d} heta) + \int U(heta_1, heta_2)
ho(\mathrm{d} heta_1)
ho(\mathrm{d} heta_2).$$

Exponential bases functions to discretize ρ !

[This work]: run SGD on θ , and give a scaling limit dynamics for ρ .

SGD and distributional dynamics (DD)

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

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

SGD and distributional dynamics (DD)

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

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

• Claim: $s_k = \varepsilon \xi(k\varepsilon), \ k = t/\varepsilon, \ N \to \infty, \ \varepsilon \to 0$:

$$\hat{
ho}_k^{(N)} \equiv rac{1}{N} \sum_{i=1}^N \delta_{oldsymbol{ heta}_i^k} \Rightarrow oldsymbol{
ho_t}.$$

SGD and distributional dynamics (DD)

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

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

lacksquare Claim: $s_k = \varepsilon \xi(k\varepsilon), \ k = t/\varepsilon, \ N \to \infty, \ \varepsilon \to 0$:

$$\hat{
ho}_k^{(N)} \equiv rac{1}{N} \sum_{i=1}^N \delta_{oldsymbol{ heta}_i^k} \Rightarrow oldsymbol{
ho}_t.$$

▶ Distributional dynamics (DD) for ρ_t ,

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

$$\Psi(oldsymbol{ heta}; oldsymbol{
ho}) = rac{\delta R(oldsymbol{
ho})}{\delta \,
ho(oldsymbol{ heta})} = V(oldsymbol{ heta}) + \int U(oldsymbol{ heta}, oldsymbol{ heta}') oldsymbol{
ho}(\mathrm{d}oldsymbol{ heta}').$$

More precisely

Assumption

(i) σ_{\star} bounded; (ii) $\nabla_{\theta}\sigma_{\star}(x;\theta)$ sub-Gaussian; (iii) ∇V , ∇U bdd. Lipschitz.

Theorem (M., Montanari, Nguyen, 2018)

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

$$\sup_{t \leq T} \Big| rac{1}{N} \sum_{i=1}^N f(oldsymbol{ heta}_i^{\lfloor t/arepsilon
floor}) - \int f(oldsymbol{ heta}) oldsymbol{
ho}_t(oldsymbol{ heta}) \Big| \leq K e^{KT} \mathrm{err}_{N,D}(z),$$

where

$$\operatorname{err}_{N,D}(z) \equiv \sqrt{rac{1}{N}ee arepsilon} \cdot \left\lceil \sqrt{Dee \lograc{N}{arepsilon}} + z
ight
ceil,$$

with probability at least $1-4e^{-z^2/2}$.

4 D > 4 D > 4 D > 4 D > 4 D > 9

More precisely

Assumption

(i) σ_{\star} bounded; (ii) $\nabla_{\theta}\sigma_{\star}(x;\theta)$ sub-Gaussian; (iii) $\nabla V, \nabla U$ bdd. Lipschitz.

Theorem (M., Montanari, Nguyen, 2018)

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

$$\sup_{t \leq T} \Big| \frac{1}{N} \sum_{i=1}^N f(\boldsymbol{\theta}_i^{\lfloor t/\varepsilon \rfloor}) - \int f(\boldsymbol{\theta}) \rho_t(\boldsymbol{\theta}) \Big| \leq K e^{KT} \mathrm{err}_{N,D}(z),$$

where

$$\operatorname{err}_{N,D}(z) \equiv \sqrt{rac{1}{N} ee arepsilon} \cdot \left\lceil \sqrt{D ee \log rac{N}{arepsilon}} + z
ight
ceil$$
 ,

with probability at least $1 - 4e^{-z^2/2}$.

一《中》《御》《唐》《唐》 唐 《

Number of neurons, sample size, and dimensions

$$\operatorname{err}_{N,D} symp \sqrt{rac{D}{N}} ee (Darepsilon).$$

N: number of neurons; D: feature dimension; ε : stepsize.

▶ Small if $N \gg D$, $\varepsilon \ll 1/D$. This is very practical!

Number of neurons, sample size, and dimensions

$$\operatorname{err}_{N,D} symp \sqrt{rac{D}{N} ee (Darepsilon)}.$$

N: number of neurons; D: feature dimension; ε : stepsize.

▶ Small if $N \gg D$, $\varepsilon \ll 1/D$. This is very practical!

Message

Approximately $(1/N) \sum_{i=1}^{N} \delta_{\theta_{i}^{k}} \approx \rho_{t}$, where

$$oldsymbol{ heta}_i^{k+1} = oldsymbol{ heta}_i^k - 2s_k N
abla_{oldsymbol{ heta}_i} \ell(x_k, y_k; oldsymbol{ heta}^k), \quad i \in [N],$$
 (SGD)

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_\theta \cdot (\rho_t(\theta) \nabla_\theta \Psi(\theta; \rho_t)).$$
 (DD)

Overparameterization $N \to \infty$ does not affect the limiting dynamics, and therefore

- Overparameterization does not affect optimization!
- Overparameterization does not affect generalization!

Message

Approximately $(1/N) \sum_{i=1}^{N} \delta_{\theta_{i}^{k}} \approx \rho_{t}$, where

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k - 2s_k N \nabla_{\boldsymbol{\theta}_i} \ell(\boldsymbol{x}_k, y_k; \boldsymbol{\theta}^k), \quad i \in [N],$$
(SGD)

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

Overparameterization $N \to \infty$ does not affect the limiting dynamics, and therefore

- Overparameterization does not affect optimization!
- Overparameterization does not affect generalization!

Message

Approximately $(1/N)\sum_{i=1}^{N} \delta_{\theta_{i}^{k}} \approx \rho_{t}$, where

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k - 2s_k N \nabla_{\boldsymbol{\theta}_i} \ell(\boldsymbol{x}_k, y_k; \boldsymbol{\theta}^k), \quad i \in [N],$$
(SGD)

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_\theta \cdot (\rho_t(\theta) \nabla_\theta \Psi(\theta; \rho_t)).$$
 (DD)

Overparameterization $N \to \infty$ does not affect the limiting dynamics, and therefore

- Overparameterization does not affect optimization!
- Overparameterization does not affect generalization!

What is this?

$$\partial_t \rho_t = \nabla_{m{ heta}} \cdot \left(\rho_t \nabla_{m{ heta}} \Psi(m{ heta}; \rho_t) \right).$$

Existence and uniqueness: [Sznitman, 1991].

▶ Physics: nonlinear transport equation describing motions of particles with pairwise interaction (mean field approach).

- ▶ Math: Gradient flow of $R(\rho)$
- ightharpoonup ... in the metric space $(\mathcal{P}(\mathbb{R}^D),W_2)$
- ▶ [Jordan, Kinderlehrer, Otto, 1998; Ambrosio, Gigli, Savaré, 2006; Carrillo, McCann, Villani, 2013;...]



What is this?

$$\partial_t
ho_t = \nabla_{m{ heta}} \cdot \left(
ho_t
abla_{m{ heta}} \Psi(m{ heta};
ho_t)
ight).$$

Existence and uniqueness: [Sznitman, 1991].

▶ Physics: nonlinear transport equation describing motions of particles with pairwise interaction (mean field approach).

- ▶ Math: Gradient flow of $R(\rho)$...
- ... in the metric space $(\mathcal{P}(\mathbb{R}^D), W_2)$.
- ► [Jordan, Kinderlehrer, Otto, 1998; Ambrosio, Gigli, Savaré, 2006; Carrillo, McCann, Villani, 2013;...]



What is this?

$$\partial_t
ho_t = \nabla_{m{ heta}} \cdot \left(
ho_t
abla_{m{ heta}} \Psi(m{ heta};
ho_t)
ight).$$

Existence and uniqueness: [Sznitman, 1991].

▶ Physics: nonlinear transport equation describing motions of particles with pairwise interaction (mean field approach).

- ▶ Math: Gradient flow of $R(\rho)$...
- ... in the metric space $(\mathcal{P}(\mathbb{R}^D), W_2)$
- ► [Jordan, Kinderlehrer, Otto, 1998; Ambrosio, Gigli, Savaré, 2006; Carrillo, McCann, Villani, 2013;...]



What is this?

$$\partial_t
ho_t = \nabla_{m{ heta}} \cdot \left(
ho_t
abla_{m{ heta}} \Psi(m{ heta};
ho_t)
ight).$$

Existence and uniqueness: [Sznitman, 1991].

▶ Physics: nonlinear transport equation describing motions of particles with pairwise interaction (mean field approach).

- ▶ Math: Gradient flow of $R(\rho)$...
- ... in the metric space $(\mathcal{P}(\mathbb{R}^D), W_2)$.
- ► [Jordan, Kinderlehrer, Otto, 1998; Ambrosio, Gigli, Savaré, 2006; Carrillo, McCann, Villani, 2013;...]



What is this?

$$\partial_t
ho_t = \nabla_{m{ heta}} \cdot \left(
ho_t
abla_{m{ heta}} \Psi(m{ heta};
ho_t)
ight).$$

Existence and uniqueness: [Sznitman, 1991].

▶ Physics: nonlinear transport equation describing motions of particles with pairwise interaction (mean field approach).

- ▶ Math: Gradient flow of $R(\rho)$...
- ... in the metric space $(\mathcal{P}(\mathbb{R}^D), W_2)$.
- ► [Jordan, Kinderlehrer, Otto, 1998; Ambrosio, Gigli, Savaré, 2006; Carrillo, McCann, Villani, 2013;...]



Does distributional dynamics converge?

Gradient flow minimizing $R(\rho)$,

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

- ▶ Does distributional dynamics converge to minimizers?
- ▶ In general, no; sometimes, yes.

In the following

- ► Concrete examples with convergence
- ▶ A general convergence result for noisy SGD.

Does distributional dynamics converge?

Gradient flow minimizing $R(\rho)$,

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

- ▶ Does distributional dynamics converge to minimizers?
- ▶ In general, no; sometimes, yes.

In the following

- ► Concrete examples with convergence
- ▶ A general convergence result for noisy SGD.

Does distributional dynamics converge?

Gradient flow minimizing $R(\rho)$,

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_\theta \cdot (\rho_t(\theta) \nabla_\theta \Psi(\theta; \rho_t)).$$
 (DD)

- Does distributional dynamics converge to minimizers?
- ▶ In general, no; sometimes, yes.

In the following

- ► Concrete examples with convergence.
- ▶ A general convergence result for noisy SGD.

Concrete examples

Simplest example requiring more than one neuron

With probability 1/2: y = +1, $x \sim \mathcal{N}(0, \Sigma_+)$, With probability 1/2: y = -1, $x \sim \mathcal{N}(0, \Sigma_-)$.

$$\Sigma_{\pm} = egin{bmatrix} au_{\pm}^2 \operatorname{I}_{s_0} & \mathbf{0} \ \mathbf{0} & \operatorname{I}_{d-s_0} \end{bmatrix}.$$

Invariant under $\mathcal{O}(s_0) \times \mathcal{O}(d-s_0) \Rightarrow \text{Reduced PDE}$.

Classifying anisotropic Gaussians: analysis

Assumption

(i) $\sigma: \mathbb{R} \to \mathbb{R}$ truncated ReLU; (ii) $s_0 = \gamma d$, $\gamma \in (0, 1)$ fixed; (iii) $\bar{\rho}_0 \in \mathcal{P}(\mathbb{R}_+)$ has bounded density and $R(\rho_0) < 1$.

Theorem (M., Montanari, Nguyen, 2018)

For $T \geq T_0$, $d \geq d_0$, $N \geq C_0 d \log d$ $(T_0, d_0, C_0$ depend on $(\eta, \bar{\rho}_0, \tau_{\pm}))$, consider SGD initialized with $(\boldsymbol{\theta}_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \leq 1/(C_0 d)$. Then, for any $k \in [T/\varepsilon, 10T/\varepsilon]$, whp

$$R_N(oldsymbol{ heta}^k) \leq \inf_{oldsymbol{ heta} \in \mathbb{R}^{d imes N}} R_N(oldsymbol{ heta}) + \eta.$$

- Learning from $k = O(1/\varepsilon) = O(d)$ samples.
- ▶ Independent of number of neurons $N \ge O(d \log d)$.

4 D > 4 B > 4 B > 3 C

Classifying anisotropic Gaussians: analysis

Assumption

(i) $\sigma: \mathbb{R} \to \mathbb{R}$ truncated ReLU; (ii) $s_0 = \gamma d$, $\gamma \in (0, 1)$ fixed; (iii) $\bar{\rho}_0 \in \mathcal{P}(\mathbb{R}_+)$ has bounded density and $R(\rho_0) < 1$.

Theorem (M., Montanari, Nguyen, 2018)

For $T \geq T_0$, $d \geq d_0$, $N \geq C_0 d \log d$ $(T_0, d_0, C_0$ depend on $(\eta, \bar{\rho}_0, \tau_{\pm})$), consider SGD initialized with $(\theta_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \leq 1/(C_0 d)$. Then, for any $k \in [T/\varepsilon, 10T/\varepsilon]$, whp

$$R_N({m{ heta}}^k) \leq \inf_{{m{ heta}} \in \mathbb{R}^{d imes N}} R_N({m{ heta}}) + \eta.$$

- Learning from $k = O(1/\varepsilon) = O(d)$ samples.
- ▶ Independent of number of neurons $N \ge O(d \log d)$.

- 4日 > 4日 > 4目 > 4目 > 目 り900

Classifying anisotropic Gaussians: analysis

Assumption

(i) $\sigma: \mathbb{R} \to \mathbb{R}$ truncated ReLU; (ii) $s_0 = \gamma d$, $\gamma \in (0, 1)$ fixed; (iii) $\bar{\rho}_0 \in \mathcal{P}(\mathbb{R}_+)$ has bounded density and $R(\rho_0) < 1$.

Theorem (M., Montanari, Nguyen, 2018)

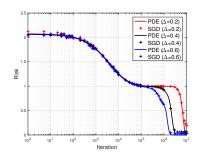
For $T \geq T_0$, $d \geq d_0$, $N \geq C_0 d \log d$ $(T_0, d_0, C_0$ depend on $(\eta, \bar{\rho}_0, \tau_{\pm})$), consider SGD initialized with $(\theta_i^0)_{i \leq N} \sim_{iid} \bar{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \leq 1/(C_0 d)$. Then, for any $k \in [T/\varepsilon, 10T/\varepsilon]$, whp

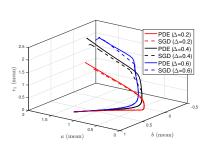
$$R_N({m{ heta}}^k) \leq \inf_{{m{ heta}} \in \mathbb{R}^{d imes N}} R_N({m{ heta}}) + \eta.$$

- Learning from $k = O(1/\epsilon) = O(d)$ samples.
- ▶ Independent of number of neurons $N \ge O(d \log d)$.

◆ロト ◆部ト ◆注ト ◆注ト 注 りなべ

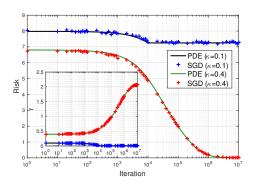
ReLU activation





- $ightharpoonup d = 320, s_0 = 60, N = 800, \tau_+^2 = 1 \pm \Delta.$
- ▶ ReLU activation.

Predicting failure



- $ightharpoonup s_0 = d = 320, N = 800, \tau_+^2 = 1.5, \tau_-^2 = 0.5.$
- ▶ Non-monotone activation.
- ▶ Two different initialization (κ = initialization variance).



Predicting failure

▶ SGD does not necessarily converge to global min.

► Can we fix it?

Noisy stochastic gradient descent

Regularized noisy SGD

SGD

$$oldsymbol{ heta}_i^{k+1} = oldsymbol{ heta}_i^k - 2 s_k N
abla_{ heta_i} oldsymbol{\ell}(x_k, y_k; oldsymbol{ heta}^k)$$

Distributional dynamics

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

Regularized noisy SGD

SGD with $(g_i^k)_{i \leq N, k \geq 0} \sim_{iid} \mathcal{N}(\mathbf{0}, I)$,

$$oldsymbol{ heta}_i^{k+1} = (1-2\lambda s_k)oldsymbol{ heta}_i^k - 2s_k N
abla_{ heta_i} oldsymbol{\ell}(x_k,y_k;oldsymbol{ heta}^k) + \sqrt{s_k/oldsymbol{eta}} g_i^k.$$

Distributional dynamics with diffusion term

$$\partial_t \rho_t(\boldsymbol{\theta}) = 2\xi(t) \nabla_{\boldsymbol{\theta}} \cdot (\rho_t(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \Psi_{\lambda}(\boldsymbol{\theta}; \rho_t)) + \boldsymbol{\beta}^{-1} \Delta_{\boldsymbol{\theta}} \rho_t(\boldsymbol{\theta}).$$

Regularized noisy SGD

SGD with $(g_i^k)_{i \leq N, k \geq 0} \sim_{iid} \mathcal{N}(\mathbf{0}, I)$,

$$oldsymbol{ heta}_i^{k+1} = (1-2\lambda s_k)oldsymbol{ heta}_i^k - 2s_k N
abla_{ heta_i} oldsymbol{\ell}(x_k,y_k;oldsymbol{ heta}^k) + \sqrt{s_k/oldsymbol{eta}}g_i^k.$$

Distributional dynamics with diffusion term

$$\partial_t \rho_t(\theta) = 2\xi(t) \nabla_{\theta} \cdot (\rho_t(\theta) \nabla_{\theta} \Psi_{\lambda}(\theta; \rho_t)) + \beta^{-1} \Delta_{\theta} \rho_t(\theta).$$

Theorem

Same approximation theorem: noisy $SGD \leftrightarrow PDE$.

Gradient flow interpretation

$$egin{aligned} F_{oldsymbol{eta},\lambda}(oldsymbol{
ho}) = &rac{1}{2}R(oldsymbol{
ho}) + rac{\lambda}{2}\int \|oldsymbol{ heta}\|_2^2
ho(\mathrm{d}oldsymbol{ heta}) - oldsymbol{eta}^{-1}\mathrm{Ent}(oldsymbol{
ho}), \ \mathrm{Ent}(oldsymbol{
ho}) = &-\int oldsymbol{
ho}(oldsymbol{ heta})\logoldsymbol{
ho}(oldsymbol{ heta})\mathrm{d}oldsymbol{ heta}. \end{aligned}$$

- ▶ Distributional dynamics is the gradient flow of $F_{\beta,\lambda}(\rho)$...
- ... in Wasserstein metric space.

[Jordan, Kinderlehrer, Otto, 1998]



Convergence of DD

Theorem (M., Montanari, Nguyen, 2018)

Assume V, U, ρ_0 "sufficiently" regular. If ρ_t is a solution of DD, then $F_{\beta,\lambda}(\rho_t)$ is non-increasing:

$$\partial_t F_{oldsymbol{eta},\lambda}(oldsymbol{
ho_t}) = -\int \left\|
abla \Big(\Psi_{\lambda}(oldsymbol{ heta}; oldsymbol{
ho_t}) - rac{1}{oldsymbol{eta}} \log oldsymbol{
ho_t}(oldsymbol{ heta}) \Big)
ight\|_2^2 oldsymbol{
ho_t}(\mathrm{d}oldsymbol{ heta}) \leq 0.$$

In particular, there exists a unique fixed point ρ_{\star} of $F_{\beta,\lambda}$ satisfies

$$m{
ho}_{\star}(m{ heta}) = rac{1}{Z_{\star}(m{eta}, \lambda)} \exp\{-m{eta}\Psi_{\lambda}(m{ heta}; m{
ho}_{\star})\}.$$

Moreover, as $t \to \infty$, $\rho_t \to \rho_{\star}$.

Generalized the analysis of [Carrillo, McCann, Villani, 2013].

4 D > 4 A > 4 E > 4 E > E = 90 P

Key remark

$$\label{eq:rho_psi} \frac{\rho_{\star}(\boldsymbol{\theta}) = \frac{1}{Z_{\star}(\boldsymbol{\beta}, \boldsymbol{\lambda})} \exp\{-\beta \Psi(\boldsymbol{\theta}; \textcolor{red}{\rho_{\star}})\}.$$

is the stationery equation for

$$F_{oldsymbol{eta},\lambda}(oldsymbol{
ho}) = rac{1}{2}R(oldsymbol{
ho}) + rac{\lambda}{2}\int \|oldsymbol{ heta}\|_2^2
ho(\mathrm{d}oldsymbol{ heta}) - oldsymbol{eta}^{-1}\mathrm{Ent}(oldsymbol{
ho}).$$

- $F_{\beta,\lambda}(\cdot)$ is strongly convex.
- ▶ The fixed point is unique!

General convergence for noisy SGD

Theorem (M., Montanari, Nguyen, 2018)

Assumptions of previous theorem. Initialization $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$. Then there exists $\beta_0 = \beta_0(D, U, V, \eta)$, such that, for $\beta \geq \beta_0$, there exists $T = T(D, U, V, \beta, \eta)$ such that for any $k \in [T/\varepsilon, 10T/\varepsilon]$, $N \geq C_0 D \log D$, $\varepsilon \leq 1/(C_0 D)$, we have, whp

$$R_{\lambda,N}(\boldsymbol{\theta}^k) \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^{D \times N}} R_{\lambda,N}(\boldsymbol{\theta}) + \eta.$$

General convergence for noisy SGD

Theorem (M., Montanari, Nguyen, 2018)

Assumptions of previous theorem. Initialization $(\theta_i^0)_{i \leq N} \sim_{iid} \rho_0$. Then there exists $\beta_0 = \beta_0(D, U, V, \eta)$, such that, for $\beta \geq \beta_0$, there exists $T = T(D, U, V, \beta, \eta)$ such that for any $k \in [T/\varepsilon, 10T/\varepsilon]$, $N \geq C_0 D \log D$, $\varepsilon \leq 1/(C_0 D)$, we have, whp

$$R_{\lambda,N}(\boldsymbol{\theta}^k) \leq \inf_{\boldsymbol{\theta} \in \mathbb{R}^{D \times N}} R_{\lambda,N}(\boldsymbol{\theta}) + \eta.$$

- ▶ For general distribution $(x, y) \sim \mathbb{P}_{x,y}$!
- \triangleright Convergence time depends on D, but not on N!

Conclusion

Conclusion

Correspondence

- ► Two layer neural networks.
- ▶ Dynamics of particles with pairwise interactions.
- ▶ Gradient flow in measure spaces.

Conclusion

Correspondence

- ► Two layer neural networks.
- ▶ Dynamics of particles with pairwise interactions.
- ▶ Gradient flow in measure spaces.

Partially explained the optimization/generalization mystery.