A Mean Field Theory of Two-Layers Neural Networks

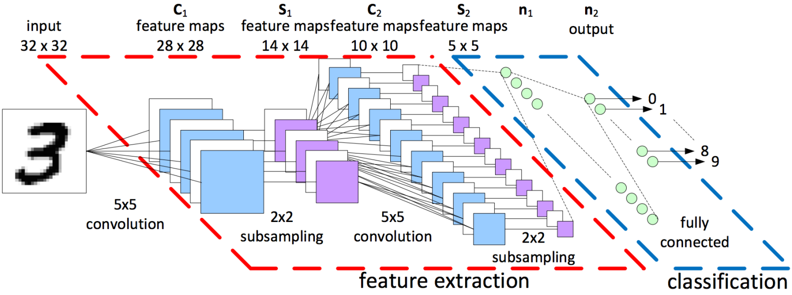
Song Mei

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).

The mystery of neuralnets

The mystery of neuralnets

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

The mystery of neuralnets

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

The mystery

- Optimized efficiently.
- Generalize well.

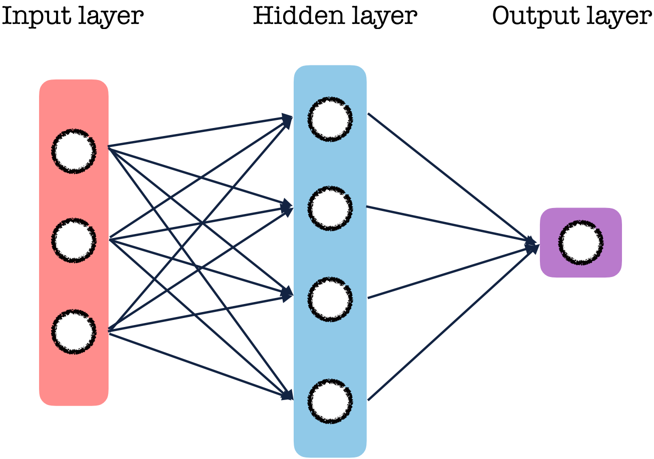
The mystery of neuralnets

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

The mystery

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

Two-layers neural networks



Two-layers neural networks

- Parameter: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \in \mathbb{R}^{N \times D}$.
- Prediction:

$$\hat{y}(x; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_\star(x; \boldsymbol{\theta}_i).$$

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

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

Two-layers neural networks

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

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

- 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:

$$R_N(\theta) = \mathbb{E}_{x,y}\Big[\Big(y - \frac{1}{N}\sum_{i=1}^{N}\sigma_\star(x;\theta_i)\Big)^2\Big].$$

Two-layers neural networks

- Parameter: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N) \in \mathbb{R}^{N \times D}$.
- Prediction:

$$\hat{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_i).$$

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

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

Two-layers neural networks

- Parameter: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \in \mathbb{R}^{N \times D}$.
- Prediction:

$$\hat{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_i).$$

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

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

Two-layers neural networks

- Parameter: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \in \mathbb{R}^{N \times D}$.
- Prediction:

$$\hat{y}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \sigma_\star(\boldsymbol{x}; \boldsymbol{\theta}_i).$$

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

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

Two-layers neural networks

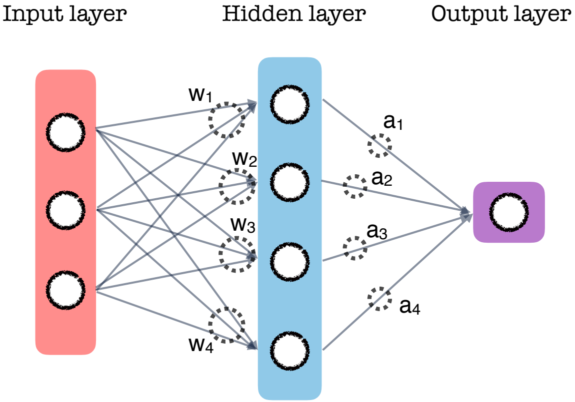


Figure: $\boldsymbol{\theta}_i = (a_i, \boldsymbol{w}_i)$.

Related literatures (before 2018)

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

- 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)

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

- 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]...

Empirical distribution of weights

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

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

where

$$\begin{aligned} V(\theta_i) =& -\mathbb{E}[y\sigma_\star(x;\theta_i)], \\ U(\theta_i,\theta_j) =& \mathbb{E}[\sigma_\star(x;\theta_i)\sigma_\star(x;\theta_j)]. \end{aligned}$$

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

$$R(\rho) = \mathbb{E}[y^2] + 2\int V(\theta)\rho(\mathrm{d}\theta) + \int U(\theta_1,\theta_2)\rho(\mathrm{d}\theta_1)\rho(\mathrm{d}\theta_2).$$

Empirical distribution of weights

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

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

where

$$V(\theta_i) = -\mathbb{E}[y\sigma_\star(x;\theta_i)],$$
$$U(\theta_i,\theta_j) = \mathbb{E}[\sigma_\star(x;\theta_i)\sigma_\star(x;\theta_j)].$$

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

$$R(\rho) = \mathbb{E}[y^2] + 2\int V(\theta)\rho(\mathrm{d}\theta) + \int U(\theta_1,\theta_2)\rho(\mathrm{d}\theta_1)\rho(\mathrm{d}\theta_2).$$

Empirical distribution of weights

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

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

where

$$\begin{aligned} V(\theta_i) =& -\mathbb{E}[y\sigma_\star(x; \theta_i)], \\ U(\theta_i, \theta_j) =& \mathbb{E}[\sigma_\star(x; \theta_i)\sigma_\star(x; \theta_j)]. \end{aligned}$$

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

$$R(\rho) = \mathbb{E}[y^2] + 2\int V(\theta)\rho(\mathrm{d}\theta) + \int U(\theta_1, \theta_2)\rho(\mathrm{d}\theta_1)\rho(\mathrm{d}\theta_2).$$

Empirical distribution of weights

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

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

Empirical distribution of weights

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

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

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

Empirical distribution of weights

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

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

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

Lemma

If U bounded, then

$$\inf_\rho R(\rho) \le \inf_\theta R_N(\theta) \le \inf_\rho R(\rho) + O(1/N).$$

How to optimize $R(\rho)$?

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

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

Exponential bases functions to discretize ρ!

How to optimize $R(\rho)$?

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

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

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)

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

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

SGD and distributional dynamics (DD)

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

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

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

$$\hat{\rho}_k^{(N)} \equiv \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i^k} \Rightarrow \rho_t.$$

SGD and distributional dynamics (DD)

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

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k - 2 s_k N \nabla_{\boldsymbol{\theta}_i} \ell(\boldsymbol{x}_k, y_k; \boldsymbol{\theta}^k). \tag{SGD}$$

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

$$\hat{\rho}_k^{(N)} \equiv \frac{1}{N} \sum_{i=1}^N \delta_{\boldsymbol{\theta}_i^k} \Rightarrow \rho_t.$$

- Distributional dynamics (DD) for ρ_t,

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

where

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

More precisely

Assumption

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

Theorem (M., Montanari, Nguyen, 2018)

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

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

where

$$\mathrm{err}_{N,D}(z) \equiv \sqrt{\frac{1}{N} \vee \varepsilon} \cdot \left[\sqrt{D \vee \log \frac{N}{\varepsilon}} + z \right],$$

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

More precisely

Assumption

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

Theorem (M., Montanari, Nguyen, 2018)

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

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

where

$$\mathrm{err}_{N,D}(z) \equiv \sqrt{\frac{1}{N}\vee\varepsilon}\cdot\left[\sqrt{D\vee\log\frac{N}{\varepsilon}} + z\right],$$

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

Number of neurons, sample size, and dimensions

$$\mathrm{err}_{N,D} \asymp \sqrt{\frac{D}{N} \vee (D\varepsilon)}.$$

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

$$\mathrm{err}_{N,D} \asymp \sqrt{\frac{D}{N} \vee (D\varepsilon)}.$$

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

$$\theta_i^{k+1} = \theta_i^k - 2s_k N\nabla_{\theta_i}\ell(x_k, y_k; \theta^k), \quad i \in [N], \tag{SGD}$$

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

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

$$\theta_i^{k+1} = \theta_i^k - 2s_k N \nabla_{\theta_i} \ell(x_k, y_k; \theta^k), \quad i \in [N], \tag{SGD}$$

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

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

$$\theta_i^{k+1} = \theta_i^k - 2s_k N \nabla_{\theta_i} \ell(x_k, y_k; \theta^k), \quad i \in [N], \tag{SGD}$$

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

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

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

What is this?

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

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?

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

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?

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

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?

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

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?

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

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 $R(\rho)$,

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

- 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 $R(\rho)$,

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

- 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 $R(\rho)$,

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

- 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.

Concrete examples

Simplest example requiring more than one neuron

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

$$\Sigma_\pm = \begin{bmatrix} \tau_\pm^2 \mathrm{I}_{s_0} & \mathbf{0} \\ \mathbf{0} & \mathrm{I}_{d-s_0} \end{bmatrix}.$$

Invariant under $\mathcal{O}(s_0) \times \mathcal{O}(d - s_0) \Rightarrow$ 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

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

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

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

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

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

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 $R(\rho_0) < 1$.

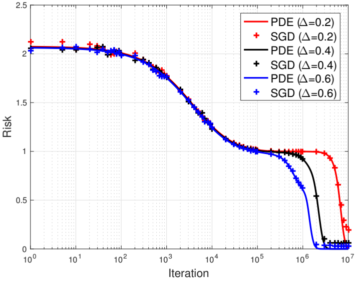
Theorem (M., Montanari, Nguyen, 2018)

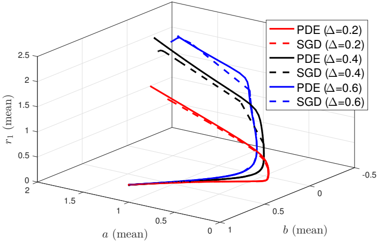
For $T \ge T_0$, $d \ge d_0$, $N \ge C_0 d \log d$ (T_0, d_0, C_0 depend on $(\eta, \bar{\rho}_0, \tau_\pm)$), consider SGD initialized with $(\boldsymbol{\theta}_i^0)_{i \le N} \sim_{iid} \bar{\rho}_0 \times \mathrm{Unif}(\mathbb{S}^{d-1})$ and step size $\varepsilon \le 1/(C_0 d)$. Then, for any $k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

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

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

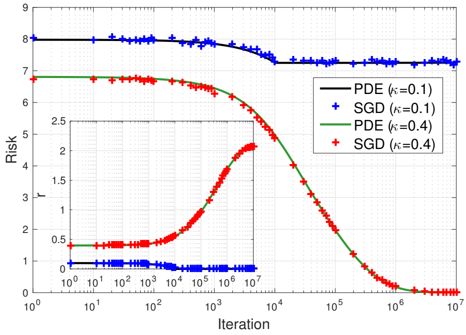
ReLU activation





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

Predicting failure



- $s_0 = d = 320$, $N = 800$, $\tau_+^2 = 1.5$, $\tau_-^2 = 0.5$.
- 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

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

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 $(g_i^k)_{i \le N, k \ge 0} \sim_{iid} \mathcal{N}(\mathbf{0}, \mathrm{I})$,

$$\boldsymbol{\theta}_i^{k+1} = (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k - 2 s_k N \nabla_{\boldsymbol{\theta}_i} \ell(\boldsymbol{x}_k, y_k; \boldsymbol{\theta}^k) + \sqrt{s_k/\beta}\, \boldsymbol{g}_i^k.$$

Distributional dynamics with diffusion term

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

Regularized noisy SGD

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

$$\boldsymbol{\theta}_i^{k+1} = (1 - 2\lambda s_k)\boldsymbol{\theta}_i^k - 2 s_k N \nabla_{\boldsymbol{\theta}_i} \ell(\boldsymbol{x}_k, y_k; \boldsymbol{\theta}^k) + \sqrt{s_k/\beta}\, \boldsymbol{g}_i^k.$$

Distributional dynamics with diffusion term

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

Theorem

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

Gradient flow interpretation

$$
\begin{aligned}
F_{\beta,\lambda}(\rho) =& \frac{1}{2} R(\rho) + \frac{\lambda}{2} \int \|\boldsymbol{\theta}\|_2^2 \rho(\mathrm{d}\boldsymbol{\theta}) - \beta^{-1} \mathrm{Ent}(\rho), \\
\mathrm{Ent}(\rho) =& - \int \rho(\boldsymbol{\theta}) \log \rho(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.
\end{aligned}
$$

- 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:

$$\partial_t F_{\beta,\lambda}(\rho_t) = -\int \Big\|\nabla\Big(\Psi_\lambda(\boldsymbol{\theta};\rho_t) - \frac{1}{\beta}\log\rho_t(\boldsymbol{\theta})\Big)\Big\|_2^2 \rho_t(\mathrm{d}\boldsymbol{\theta}) \le 0.$$

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

$$\rho_\star(\boldsymbol{\theta}) = \frac{1}{Z_\star(\beta,\lambda)}\exp\{-\beta\Psi_\lambda(\boldsymbol{\theta};\rho_\star)\}.$$

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

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

Key remark

- $$\rho_\star(\boldsymbol{\theta}) = \frac{1}{Z_\star(\beta,\lambda)} \exp\{-\beta\Psi(\boldsymbol{\theta}; \rho_\star)\}.$$

  is the stationery equation for

  $$F_{\beta,\lambda}(\rho) = \frac{1}{2}R(\rho) + \frac{\lambda}{2}\int \|\boldsymbol{\theta}\|_2^2 \rho(\mathrm{d}\boldsymbol{\theta}) - \beta^{-1}\mathrm{Ent}(\rho).$$
- $F_{\beta,\lambda}(\cdot)$ is strongly convex.
- The fixed point is unique!

General convergence for noisy SGD

Theorem (M., Montanari, Nguyen, 2018)

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

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

General convergence for noisy SGD

Theorem (M., Montanari, Nguyen, 2018)

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

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

- For general distribution $(x, y) \sim \mathbb{P}_{x,y}$!
- 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.