A mean field view of the landscape of two-layers neural network

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. Stanford University

Challenge

Neural network is highly non-convex so that it is hard to analyze its landscape. Though, empirically SGD "works well" for neural networks. Any explanations?

Model: Two layers neural network

Let $(\boldsymbol{X}, Y) \sim \mathbb{P}$, $\boldsymbol{X} \in \mathbb{R}^d$, $Y \in \mathbb{R}$. Consider the two-layers neural network with decision variable $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{i=1}^N \in \mathbb{R}^{N \times D}$,

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\, R_N(\boldsymbol{\theta}) = \mathbb{E}\Big[Y - \frac{1}{N}\sum_{i=1}^N \sigma_*(\boldsymbol{X}; \boldsymbol{\theta}_i)\Big]^2 \Big(+ \frac{\lambda}{N}\|\boldsymbol{\theta}\|_2^2 \Big). \quad (1)$$

An example of σ_* gives

$$\sigma_*(\boldsymbol{X}; \boldsymbol{\theta}_i) = a_i \sigma(\langle \boldsymbol{X}, \boldsymbol{w}_i \rangle + b_i) \quad (2)$$

where $\boldsymbol{\theta}_i = (a_i, b_i, \boldsymbol{w}_i)$ and $\sigma(\,\cdot\,)$ is ReLU.

Algorithm: SGD and noisy SGD

Consider the SGD / noisy SGD algorithm minimizing risk R_N,

$$\boldsymbol{\theta}_i^{k+1} = \boldsymbol{\theta}_i^k - 2Ns_k \nabla_{\boldsymbol{\theta}_i} \widehat{R}_N(\boldsymbol{\theta}^k; (\boldsymbol{x}_k, y_k)) + \sqrt{\frac{2s_k}{\beta}}\boldsymbol{g}_i^k.$$

In each iteration we get a fresh sample $(\boldsymbol{x}_k, y_k)$, and N independent Gaussian random variable $(\boldsymbol{g}_i^k)_{i \in [N]}$. Parameter s_k gives the step size, β the inverse temperature (could be infinity).

Idea: Mean field reformulation

Let $\rho = (1/N)\sum_{i=1}^N \delta(\boldsymbol{\theta}_i)$ be the empirical distribution of the neuron parameters. Then $R_N(\boldsymbol{\theta}) = R(\rho)$, with

$$R(\rho) = \mathbb{E}\Big[\Big(Y - \int \sigma_*(\boldsymbol{X}; \boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta})\Big)^2\Big] + \lambda \int \|\boldsymbol{\theta}\|_2^2 \rho(\mathrm{d}\boldsymbol{\theta})$$
$$= 1 + 2\int V(\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta})\rho(\mathrm{d}\boldsymbol{\theta}'),$$

where

$$V(\boldsymbol{\theta}) = -\mathbb{E}[Y\sigma_*(\boldsymbol{X}; \boldsymbol{\theta})] + \lambda\|\boldsymbol{\theta}\|_2^2,$$
$$U(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}[\sigma_*(\boldsymbol{X}; \boldsymbol{\theta})\sigma_*(\boldsymbol{X}; \boldsymbol{\theta}')].$$

R is convex in ρ. But since ρ is infinite dimensional, the convex problem is still hard to solve.

Idea: Distributional dynamics (DD)

The noisy SGD dynamics can be well approximated by the PDE, in which we call distributional dynamics (DD)

$$\partial_t \rho(\boldsymbol{\theta}, t) = \nabla_{\boldsymbol{\theta}} \cdot (\nabla_{\boldsymbol{\theta}} \Psi(\boldsymbol{\theta}; \rho)\rho) + \frac{1}{\beta}\Delta_{\boldsymbol{\theta}}\rho, \quad (3)$$

where $\Psi(\boldsymbol{\theta}; \rho) = V(\boldsymbol{\theta}) + \int U(\boldsymbol{\theta}, \boldsymbol{\theta}')\rho(\mathrm{d}\boldsymbol{\theta}')$. The PDE can be interpreted as the gradient flow of free energy $F_\beta(\rho)$ (PDE is minimizing F_β)

$$F_\beta(\rho) = \frac{1}{2}R(\rho) + \frac{1}{\beta}\int \rho \log \rho \,\mathrm{d}\boldsymbol{\theta}. \quad (4)$$

DD is the mean field version of Fokker-Planck Eq. for Langevin dynamics

$$\partial_t \rho_N(\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N, t) = \nabla_{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N} \cdot (N \nabla R_N(\boldsymbol{\theta})\rho_N) + \frac{1}{\beta}\Delta_{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N}\rho_N. \quad (5)$$

DD reduced the dimension of Fokker-Planck Eq. using symmetry.

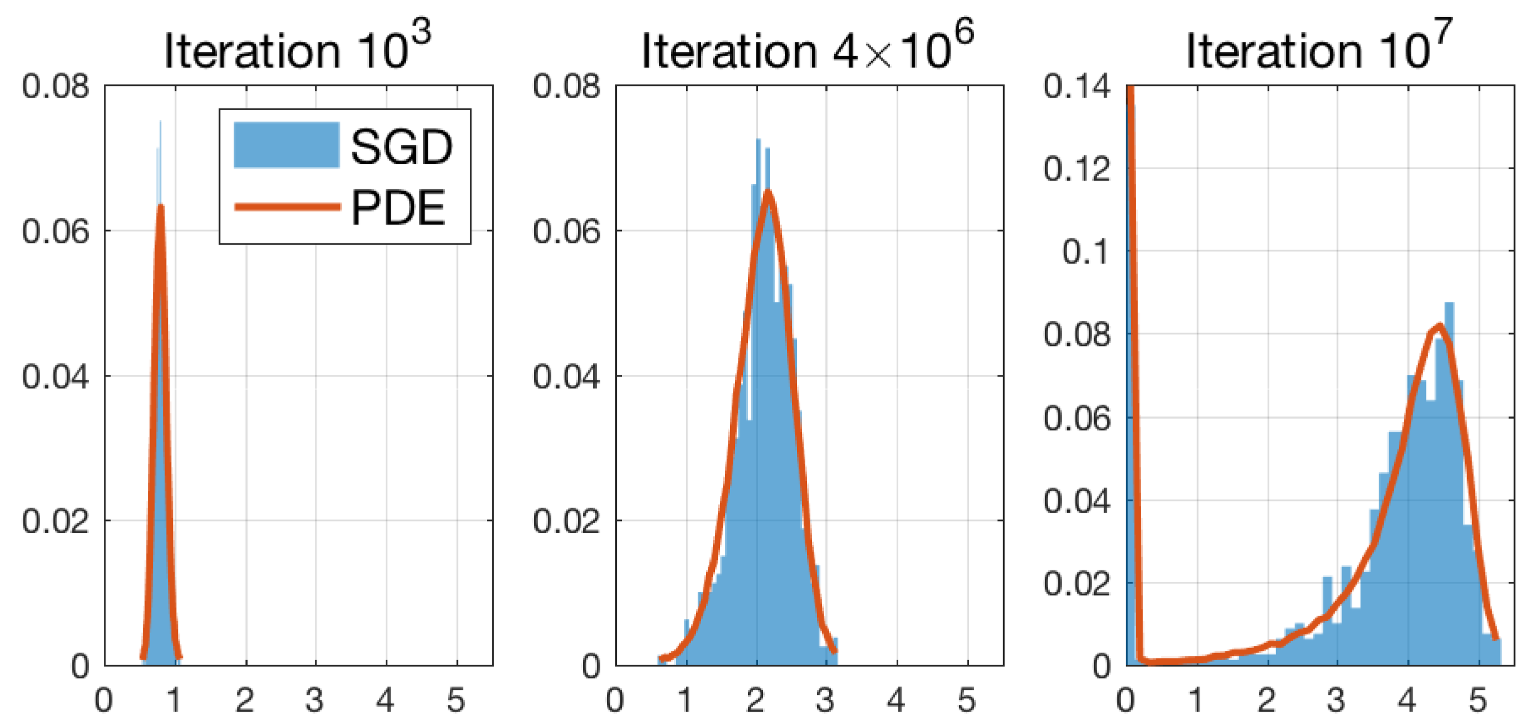
Key Theorem (informal)

At iteration k of noisy SGD giving the weights $(\boldsymbol{\theta}_i^k)_{i \in [N]}$, and time t_k of distributional dynamics PDE giving distribution $\rho(\boldsymbol{\theta}, t_k)$, we have

$$\rho(\boldsymbol{\theta}, t_k) \approx (1/N)\textstyle\sum_{i \in [N]} \delta(\boldsymbol{\theta}_i^k) \quad (6)$$

The approximation is consistent as long as $N \geq \Omega(D)$ and $t_k = O(1)$.

Illustration: SGD vs DD



Evolution of the empirical distribution of $\|\boldsymbol{w}_i\|_2$ for "classifying two Gaussians" example. Dimension $d = 40$, number of neurons $N = 800$. Histograms are obtained from SGD experiments. Continuous lines correspond to a numerical solution of the distributional dynamics PDE.

Key message

We can in turn to study the geometry of the free energy $F_\beta(\rho)$ to analyze neural networks!

Theorem: Landscape of two layers NN

For two layers neural network (1), $F_\beta(\rho)$ is strongly convex in ρ. As $t \to \infty$, $\rho(\boldsymbol{\theta}, t)$ following PDE (3) converges to the unique minimizer of $F_\beta(\rho)$.

Theorem: Convergence of noisy SGD

Take $\beta = O(d)$. Suppose PDE (3) takes $T = T(d, \beta)$ time to converge to a distribution ρ_T with $F_\beta(\rho_T) \leq \inf_\rho F_\beta(\rho) + \eta/2$. Then as we take $N \geq Ce^{CT}d$ and run noisy SGD with stepsize $s = 1/(Ce^{CT}d)$, we have $R_N(\boldsymbol{\theta}^{T/s}) \leq \inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) + \eta$.

Remarks

- Amazingly, the convergence time for SGD on two-layers neural network does not depend on the number of neurons N. Overparameterization does not harm generalization!
- The time $T = T(d, \beta)$ for PDE converging to global minimizer of $F_\beta(\rho)$ requires a case by case study. Sometimes the time is independent of d and β. Here is an example in the following.

Example: Classifying two Gaussians

Let the joint law of $(\boldsymbol{X}, Y) \in \mathbb{R}^d \times \mathbb{R}$ to be:
With probability 1/2: $Y = +1$, $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I}_d + (\tau_+^2 - 1)\mathsf{P}_V)$.
With probability 1/2: $Y = -1$, $\boldsymbol{X} \sim \mathcal{N}(0, \mathbf{I}_d + (\tau_-^2 - 1)\mathsf{P}_V)$.
Here $\tau_\pm = (1 \pm \Delta)$ and P_V is the projector onto $V \subseteq \mathbb{R}^d$.
Activation $\sigma_*(\boldsymbol{X}; \boldsymbol{\theta}_i) = \sigma(\langle \boldsymbol{X}, \boldsymbol{\theta}_i \rangle)$ for σ to be truncated ReLU.

Then as long as $N = \Omega(d)$, we run $k = \Omega(d)$ number of SGD iterations, we get $\boldsymbol{\theta}^k$ such that $R_N(\boldsymbol{\theta}^k) \leq \inf_{\boldsymbol{\theta}} R_N(\boldsymbol{\theta}) + \eta$.