A Mean Field View of the Landscape of Two-layer Neural Network

Song Mei, Andrea Montanari, Phan-Minh Nguyen

I - Learning with a shallow NN

$Y_i = f(X_i) + \varepsilon_i$

depth 2 width N neural net $\hat f(x, \vec\theta) = \frac{1}{N}\sum_{i=1}^N \sigma_*(x, \theta_i)$ $\quad \vec\theta \in (\mathbb{R}^d)^N = \Theta^N$, $\Theta \subset \mathbb{R}^D$

typically $\theta_i = (a_i, b_i, w_i)$ $\quad \sigma_*(x,\theta_i) = a_i\,\sigma(\langle w_i, x\rangle + b_i)$ $\quad \sigma$ = relu or sigmoid

loss at (x,y): $R_N(\vec\theta, x, y) = (y - \hat f(x,\vec\theta))^2$ $\quad$ avg risk: $R_N(\theta) = \mathbb{E}[R_N(\theta, X, Y)]$

minimized by SGD: (X_k, Y_k) iid, Robbins Monroe

$$\vec\theta^{k+1} = \vec\theta^k - s_k \nabla_{\vec\theta} R_N(\vec\theta, X_k, Y_k) \qquad \mathbb{E}[\nabla_{\vec\theta} R_N(\vec\theta, X_k, Y_k) \mid \mathcal{F}_k] = \nabla R_N(\vec\theta)$$

here: $\vec\theta_i^{k+1} = \vec\theta_i^k + 2 s_k \nabla_{\theta_i}\sigma_*(X_k, \theta_i^k)\left(Y_k - \frac{1}{N}\sum_{j=1}^N \sigma_*(X_k, \theta_j^k)\right)$

PC: does it converge to $\min R_N$? $\to$ predict dynamics

ex: $Y_i \sim \mathcal{U}\{\pm 1\}$ $\quad X_i \sim (1 + Y_i \Delta) Z_i$ $\quad Z_i \overset{iid}{\sim} \mathcal{N}(0,1)$ Montanari

or $Y = \hat f(X, \vec\theta_0)$ and try to fit $\hat f$ (Chizat & Bach)

II Mean Field View

2 "idealizations": • $\widetilde{\text{(gradient)}}$ $\to$ expectation

• mean-field: ∞ of neurons (particles)

$$R_N(\vec\theta, x, y) = y^2 + \frac{2}{N}\sum_{i=1}^N -y\,\sigma_*(x, \theta_i) + \frac{1}{N^2}\sum_{i,j}\sigma_*(x,\theta_i)\sigma_*(x,\theta_j)$$

$\rho_N = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i}$ $\quad \rho_{N,2}^{\otimes} = \frac{1}{N^2}\sum_{i,j=1}^N \delta_{\theta_i,\theta_j}$

$$= y^2 + 2\int \underbrace{-y\,\sigma_*(x,\theta)}_{V(\theta, x, y)} d\rho_N(\theta) + \iint \underbrace{\sigma_*(x,\theta)\,\sigma_*(x,\theta')}_{U(\theta,\theta', x, y)} d\rho_{N,2}^{\otimes}(\theta,\theta')$$

$V(\theta) = \mathbb{E}[V(\theta, x, y)]$ = external potential $\qquad U(\theta,\theta') = \mathbb{E}[U(\theta,\theta', X, Y)]$ = pairwise potential

≥ 0, s-definite: $\to$ repulsive $\quad \iint U(\theta,\theta') h(\theta) h(\theta') d\theta d\theta' = \mathbb{E}[\int \sigma_*(x,\theta) h(\theta)]^2$

$$R_N(\vec\theta) = \underset{\mathbb{E}(Y^2)}{R_\#} + 2\int V(\theta)\, d\rho_N(\theta) + \iint U(\theta,\theta')\, d\rho_{N,2}^{\otimes}(\theta,\theta')$$

$\downarrow$ idealize

$$R(\rho) = R_\# + 2\int V(\theta)\, d\rho(\theta) + \iint U(\theta,\theta')\, d\rho(\theta)\, d\rho(\theta') \quad \to R_N(\theta) = R(\rho_N)$$

idea: $\left|\inf_{\vec\theta \in \Theta^N} R_N(\vec\theta) - \inf_\rho R(\rho)\right| \leq K/N$

if $\exists K > 0, \varepsilon_0 > 0$ st $\forall \rho : R(\rho) \leq \inf_\rho R(\rho) + \varepsilon_0$, $\int U(\theta,\theta)\, d\rho(\theta) \leq K$

Proof: let ρ^* st $R(\rho^*) \leq R(\rho) + \varepsilon$ and $\vec\theta = \theta_1, \dots, \theta_N \overset{iid}{\sim} \rho^*$

then $\mathbb{E}[R_N(\vec\theta)] = \int \mathbb{E} R_N(\vec\theta, X, Y)\, d\rho^*(\vec\theta) = R_\# + \int V(\theta)\, d\rho^*(\theta) + \int \frac{1}{N^2}\sum_{i,j} U(\theta_i,\theta_j)\, d\rho^*(\theta)\, d\rho^*(\theta)$

$= R_\# + \int V(\theta)\, d\rho^*(\theta) + \int U(\theta,\theta')\, d\rho^*(\theta)\, d\rho^*(\theta') + \frac{1}{N}\left(\int U(\theta,\theta)\, d\rho^*(\theta) - \int U(\theta,\theta')\, d\rho^*(\theta)\, d\rho^*(\theta')\right)$

$\leq R(\rho) + K/N - \frac{1}{N}\int (\sigma_*(x,\theta)\, d\rho(\theta))^2$

III Mean Field Dynamics

What does SGD correspond to in the continuous world?

answer: Gradient flow for $R(\rho)$ in Wasserstein metric

SGD: $\theta_i^{k+1} = \theta_i^k + s_k \underbrace{\left(-\nabla_{\theta_i} R_N(\vec{\theta}^k, x_k, y_k)\right)}$ $\qquad \rho_N^k = \frac{1}{N}\sum_{i=1}^N \delta_{\theta_i^k}$

$$v(\theta_i^k, \rho_N^k, x_k, y_k) = -2\nabla V(\theta_i^k, x_k, y_k) - 2\int \nabla U(\theta_i^k, \theta', x_k, y_k)\, d\rho_N^k(\theta')$$

$$= -\nabla \psi(\theta_i^k, \rho_N^k, x_k, y_k) = -\nabla\psi(\theta_i^k, \rho_N^k) + \text{noise}$$

where $\psi(\theta, \rho, x, y) = 2V(\theta, x, y) + 2\int U(\theta, \theta', x, y)\, d\rho(\theta')$

$$\psi(\theta, \rho) = \mathbb{E}\left[\psi(\theta, \rho, x_k, y_k)\right] = 2V(\theta) + 2\int U(\theta, \theta')\, d\rho(\theta)$$

$$= \frac{\delta R(\rho)}{\delta \rho(\theta)} \quad : \text{"additional energy when adding one particle at } \theta\text{"}$$

→ idealized particle speed: $v(\theta, \rho) = -\nabla\psi(\theta, \rho)$

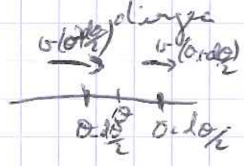
such that $\mathbb{E}[v(\theta_i^k, \rho_N, x_k, y_k)] = v(\theta_i^k, \rho_N)$

Continuous dynamic: continuity equation

$$(*) \qquad \frac{\partial \rho_t(\theta)}{\partial t} = -\nabla_\theta \cdot \left[\rho_t(\theta)\, v(\theta, \rho_t)\right] = -\nabla_\theta \cdot \left[\rho_t(\theta) \nabla\psi(\theta, \rho_t)\right]$$

divergence

intuition in 1D:



$$\rho_{t+dt}(\theta) - \rho_t(\theta) = \left(-\rho_t(\theta + \tfrac{d\theta}{2})\, v(\theta + \tfrac{d\theta}{2}) + v(\theta - \tfrac{d\theta}{2})\, \rho(\theta - \tfrac{d\theta}{2})\right) dt$$

- Gradient flow: $\dot{x}(t) = -\nabla F(x(t))$ $\qquad x(t+\varepsilon) = \arg\min_z \left\{F(z) + \frac{1}{2\varepsilon}\|z - x(t)\|^2\right\}$

- Wasserstein metric: $\rho_{t+\varepsilon} = \arg\min_{\rho \in \mathcal{P}(\Theta)} \left\{R(\rho) + \frac{1}{2\varepsilon} W_2(\rho, \rho_t)^2\right\}$

where $W_2(\rho, \rho')^2 = \inf_{\gamma \in \mathcal{C}(\rho, \rho')} \int_\Theta \|\theta - \theta'\|^2\, \gamma(d\theta, d\theta')$

IV Approximation with Particles

Thm 5.1: Assumptions: • $(x,\theta) \mapsto \sigma_*(x,\theta)$ bounded with sub-Gaussian gradient
$\|\sigma_*\|_\infty \le K_2$ $\|\nabla_\theta \sigma_*(X,\theta)\|_{\psi_2} \le K_2$
labels $|y_k| \le K_2$

• gradient of V and U are bounded, Lipschitz continuous:
$\|\nabla_\theta V(\theta)\|_2$ $\|\nabla_{\theta_1} U(\theta_1,\theta_2)\| \le K_3$
$\|\nabla V(\theta) - \nabla V(\theta')\|_2 \le K_3 \|\theta - \theta'\|_2$
$\|\nabla_{\theta_1} U(\theta_1,\theta_2) - \nabla U(\theta_1',\theta_2')\| \le K_3 \left\| \begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} - \begin{pmatrix}\theta_1'\\ \theta_2'\end{pmatrix} \right\|_2$

For $\rho \in \mathcal{M}_1(\mathbb{R})$, consider SGD with initialization $(\theta_i^0)_{i \le N} \overset{iid}{\sim} \rho_0$
and step size $s_k = \varepsilon/2$. For $t \geq 0$, let ρ_t be the solution of (*)
then there exists $C = C(K_i)$ st $\forall f \in \mathcal{C}(\Theta) \times \mathbb{R} \to \mathbb{R}^2$ $\|f\|_\infty \le 1$ $\|f\|_{Lip} \le 1$

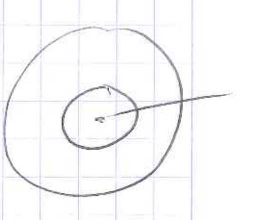
$\varepsilon \le 1$: $\sup_{k \in [0,T/\varepsilon]} \left| \frac{1}{N} \sum_i f(\theta_i^k) - \int f \, d\rho_{k\varepsilon} \right|$

and in particular: $\sup_{k \in [0,T/\varepsilon] \cap \mathbb{N}} \left| R_N(\theta^k) - R(\rho_{k\varepsilon}) \right| \le C e^{CT} \sqrt{\frac{1}{N} \vee \varepsilon} \left(\sqrt{D + \log \frac{N}{\varepsilon}} + z \right)$

with proba $\geq 1 - e^{-z^2}$

$\Rightarrow$ PDE approx accurate as soon as $N \gg D$, $\varepsilon \ll 1/D$

Speed is indep of N!



example: $y_k \sim \mathcal{U}(\{\pm 1\})$ $X_k = (1 + y_k \Delta) Z_k$ where $Z_k \sim \mathcal{N}(0,1)$
(= radial optimal classifier)

$\theta_i = w_i$ (no a_i, no b_i) then ρ_t is radial, $\rho_t(|w|)$ predicting well $\frac{1}{N}\sum_{i=1}^N \delta_{|w_i^k|}$

can see learning fail in some examples

4

Analysis "Propagation of chaos argument of Sznitman 91" "topics in propagation of chaos".

$\theta_i^0 \overset{iid}{\sim} \rho_0$ $\qquad$ ρ_0

$\theta_i^{k+1} = \theta_i^k + s_k v(\theta_i^k, \rho_N^k, x_k, y_k)$ $\qquad$ $t = k\varepsilon$, $s_k = \varepsilon/2$ $\qquad$ $\frac{\partial \rho_t(\theta)}{\partial t} = -\nabla_\theta \cdot \left(\rho_t(\theta) v(\theta, \rho_t)\right)$

squeeze $\quad$ hybrid $\quad$ called "non-linear dynamics"

$\bar{\theta}_i^0 \overset{iid}{\sim} \rho_0$

$\frac{d}{dt} \bar{\theta}_i^t = v(\theta_i^t, \rho_t) = -\psi(\theta_i^t, \rho_t)$

$P_{i \neq j}: (\bar{\theta}_i^t)_i \overset{iid}{\sim} \rho_t$

5

$\| \theta_i^{\lfloor T/\varepsilon \rfloor} - \bar{\theta}_i^T \| = \int_0^T \| \psi(\bar{\theta}_i^t, \rho_t) - \psi(\theta_i^{\lfloor t/\varepsilon \rfloor}, \rho_N^{\lfloor t/\varepsilon \rfloor}, x_{\lfloor t/\varepsilon \rfloor}, y_{\lfloor t/\varepsilon \rfloor}) \| \, dt$

$\leq \int_0^T \| \psi(\bar{\theta}_i^t, \rho_t) - \psi(\theta_i^{\lfloor t/\varepsilon \rfloor}, \rho_N^{\lfloor t/\varepsilon \rfloor}) \| \, dt$

$+ \| \varepsilon \sum_{k=0}^{T/\varepsilon - 1} \psi(\theta_i^{\lfloor k\varepsilon \rfloor}, \rho_N^{\lfloor k\varepsilon \rfloor}) - \psi(\theta_i^{\lfloor k\varepsilon \rfloor}, \rho_N^{\lfloor k\varepsilon \rfloor}, x_{\lfloor k\varepsilon \rfloor}, y_{\lfloor k\varepsilon \rfloor}) \|$ (3)

$\leq \int_0^T \| \psi(\bar{\theta}_i^t, \rho_t) - \psi(\theta_i^{\lfloor t/\varepsilon \rfloor}, \rho_t) \| \, dt$ + ~~...~~ $/ \leq k_3 \int \| \bar{\theta}_i^t - \theta_i^{\lfloor t/\varepsilon \rfloor} \| \, dt$ + Gronwall

$+ \int_0^T \| \psi(\theta_i^{\lfloor t/\varepsilon \rfloor}, \rho_t) - \psi(\theta_i^{\lfloor t/\varepsilon \rfloor}, \rho_N^{\lfloor t/\varepsilon \rfloor}) \| \, dt$ $/ \leq k_3 / \sup_j \| \bar{\theta}_j^t - \theta_j^{\lfloor t/\varepsilon \rfloor} \| \, dt$ → Gronwall

$+$ (3) / sum of martingale increments → by Azuma-Hoeffding

$\leq$ ~~...~~