• midterm course evaluation due this Friday (4/29)

Global convergence of gradient descent

Global convergence of gradient descent

Theorem (Du et al. '18, Allen-Zhu et al. '18, Zou et al 19') If the width of each layer is poly(n) where n is the number of data. Using random initialization with a particular scaling, gradient descent finds an approximate global minimum in polynomial time.

Neural Tangent Kernel

$$|f(w) - f(w^*)| \le \varepsilon$$

$$\text{poly}\left(\frac{1}{\varepsilon}, \text{width}, \text{depth}\right)$$

Gradient Flow: a Kernel Point of View

- $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(f(\theta, x_i), y_i)$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{1}{n} \sum_{i=1}^{n} l'(f(\theta, x_i), y_i) \cdot \frac{\partial f(\theta, x_i)}{\partial \theta}$$

GF: $\frac{d\theta(t)}{dt} = -\frac{\partial L(\theta)}{\partial \theta}$

if strong convex, $\Rightarrow \theta^*$, $\theta(t) \to \theta^*$

in NN, # of parameters $> n$

we know $t \to \infty$, $f(\theta(t), x_i) \to y_i$

Gradient Flow: a Kernel Point of View

$$u_i(t) = f(\theta(t), x_i), \quad u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}$$

$$\frac{d\, u_i(t)}{dt} = \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{d\theta(t)}{dt} \right\rangle$$

$$= \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, -\frac{1}{n}\sum_{j=1}^{n} \ell'(u_j(t), y_j) \cdot \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle$$

$$= -\frac{1}{n}\left[\ell'(u_1(t), y_1), \dots, \ell'(u_n(t), y_n)\right] \cdot \left(\left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_1(t)}{\partial \theta(t)} \right\rangle, \dots, \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_n(t)}{\partial \theta(t)} \right\rangle \right)$$

$$\ell'(u(t), y) \in \mathbb{R}^n$$

$$[\ell'(u(t), y)]_i = \ell'(u_i(t), y_i)$$

$$H(t) \in \mathbb{R}^{n \times n}$$

$$[H(t)]_{ij} = \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle,$$

$$\frac{d\, u(t)}{dt} = -\frac{1}{n} H(t) \cdot \ell'(u(t), y),$$

Gradient Flow: a Kernel Point of View

If ℓ : quadratic loss, $\ell(u(t), y) = \frac{1}{2}(u(t)-y)^2$

$$\ell'(u(t), y) = u(t) - y$$

$$(*) \quad \frac{du(t)}{dt} = -\frac{1}{n} \underbrace{H(t)}_{\text{does not depend on loss}} (u(t) - y)$$

- If $\exists \lambda_0 > 0$, $\forall t$, $\lambda_{\min}(H(t)) \geq \lambda_0$

$$\frac{d\left(\underbrace{\frac{1}{2}\|u(t)-y\|_2^2}_{\text{loss}}\right)}{dt} = -\frac{1}{n}(u(t)-y)^T H(t)(u(t)-y)$$

$$\leq -\frac{\lambda_0}{n}\|u(t)-y\|_2^2$$

Gradient Flow: a Kernel Point of View

Consider $\frac{d}{dt}\left(\exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2\right)$

$$= \frac{\lambda_0}{2n}\exp\left(\frac{\lambda_0 t}{n}\right)\|u(t)-y\|_2^2 + \frac{d\left(\frac{1}{2}\|u(t)-y\|_2^2\right)}{dt}\cdot\exp\left(\frac{\lambda_0 t}{n}\right)$$

$$\leq \exp\left(\frac{\lambda_0 t}{n}\right)\|u(t)-y\|_2^2\left(\frac{\lambda_0 t}{2n}-\frac{\lambda_0 t}{n}\right) < 0$$

$\Rightarrow \exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2$ is decreasing

$t=0$, $\frac{1}{2}\|u(0)-y\|_2^2$ $\bigcirc\bigcirc\bigcirc$

$\forall t$ $\exp\left(\frac{\lambda_0 t}{n}\right)\cdot\frac{1}{2}\|u(t)-y\|_2^2 \leq C$

loss $\left(\frac{1}{2}\right)$ rate $\frac{1}{2}\|u(t)-y\|_2 \leq C\cdot\exp\left(-\frac{\lambda_0 t}{n}\right)$

$t\to\infty$, loss $\to 0$, $u(t)\to y$

Gradient Flow: a Kernel Point of View

kernel: $[u(t)]_i = \phi(x_i)^T \theta(t)$ ϕ: feature

$$[H(t)]_{ij} = \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle$$

$$= \langle \phi(x_i), \phi(x_j) \rangle$$

$$= K(x_i, x_j)$$ does not depend on t

$$H(t) = \begin{pmatrix} & K(x_i, x_j) & \end{pmatrix}$$ if kernel is full rank

$\lambda_{min}(H(t)) > 0$ if kernel is universal

(Gaussian) $\lambda_{min}(H(t)) > 0$

Gradient Flow: a Kernel Point of View

NTK scaling

- $f(\theta, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \, \sigma(w_r^T x)$
- $x \in \mathbb{R}^d$, $w_r \in \mathbb{R}^d$, $a_r \in \mathbb{R}$, $\sigma(\cdot)$: ReLU
- Initialization: $a_r \sim \text{unif}\{1, -1\}$ only for simplicity
  $w_r \sim \mathcal{N}(0, I)$
- Training: only train $w_1, \dots, w_m$
  $$\min_{w_1, \dots, w_m} \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i; a, w) - y_i\right)^2$$

$u_i(t) = f(\theta(t), x_i)$

$u(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix}$

$$\frac{d\, u(t)}{dt} = -\frac{1}{n} H(t) \cdot (u(t) - y)$$

Idea: show $H(t) \approx H^*$ (NTK), $H^*_{ij} = \lim_{m \to \infty} \mathbb{E}_{\text{init dist}} \left\langle \frac{\partial f(x_i; a, w)}{\partial w}, \frac{\partial f(x_j; a, w)}{\partial w} \right\rangle$

$$H^*_{ij} = \mathbb{E}_{\theta} \left\langle \frac{\partial f(\theta, x_i)}{\partial \theta}, \frac{\partial f(\theta, x_j)}{\partial \theta} \right\rangle$$

~ init. dist

$$H_{ij}(t) = \left\langle \frac{\partial f(\theta(t), x_i)}{\partial \theta}, \frac{\partial f(\theta(t), x_j)}{\partial \theta} \right\rangle$$

$H^*, H(t) \in \mathbb{R}^{n \times n}$

Gradient Flow: a Kernel Point of View

$$W \in \mathbb{R}^{m \times d}$$

$$H_{ij}(t) = \left\langle \frac{\partial u_i(t)}{\partial W(t)}, \frac{\partial u_j(t)}{\partial W(t)} \right\rangle,$$

$$= \sum_{r=1}^{m} \left\langle \frac{\partial u_i(t)}{\partial w_r(t)}, \frac{\partial u_j(t)}{\partial w_r(t)} \right\rangle$$

$$\frac{\partial u_i(t)}{\partial w_r(t)} = \frac{1}{\sqrt{m}} a_r \cdot x_i \cdot \mathbb{1}\{w_r^T x_i \geq 0\}$$

$$H_{ij}(t) = \sum_{r=1}^{m} \frac{1}{m} \left\langle a_r \cdot x_i \cdot \mathbb{1}\{w_r^T x_i \geq 0\}, a_r \cdot x_j \mathbb{1}\{w_r^T x_j \geq 0\} \right\rangle$$

$$= \frac{1}{m} x_i^T x_j \sum_{r=1}^{m} \mathbb{1}\{w_r^T x_i \geq 0, w_r^T x_j \geq 0\}$$

To show $H(t) \approx H^*$: ① $H(0) \approx H^*$

② $H(t) \approx H(0)$

Gradient Flow: a Kernel Point of View

Initialization:

Hoeffding Inequality:

R.V. $z_1, \dots, z_n \overset{i.i.d.}{\sim}$, $|z_i| \le 1$

if $n = \Omega\left(\frac{\log(\frac{1}{\delta})}{\varepsilon^2}\right)$, $0 < \delta < 1$, $0 < \varepsilon < 1$

w.p. $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}[z_i]\right| \le \varepsilon$

$$H_{ij}(0) = x_i^T x_j \cdot \underbrace{\frac{1}{m}\sum_{r=1}^{m}}_{\text{average}} \cdot \underbrace{\mathbb{1}\{w_r(0)^T x_i \ge 0,\ w_r(0)^T x_j \ge 0\}}_{\text{random variable}}$$

$$H_{ij}^* = \mathbb{E}_{w \sim N(0, I)}\, x_i^T x_j \cdot \mathbb{1}\{w^T x_i \ge 0,\ w^T x_j \ge 0\}$$

when m is large enough, $|H_{ij}(0) - H_{ij}^*| \le \varepsilon$

Gradient Flow: a Kernel Point of View

$m = \mathrm{poly}(n)$

$$\|H^* - H(0)\|_F$$

$$\leq \sum_{i,j} |H^*_{ij} - H_{ij}(0)|$$

$$\leq n^2 \varepsilon$$

choose ε small

Figure : power of over-parameterization

$\rightarrow$ convergence

Gradient Flow: a Kernel Point of View

ε - training error

want $H(t) \approx H(0)$

for simplicity: 1) just train till time t

2) $y_i = O(1)$

3) $\|x_i\|_2 = 1$

$$H_{ij}^{t} = x_i^T x_j \cdot \frac{\pi - \arccos(x_i^T x_j)}{2\pi}$$

key idea: every weight vector only moves a little $\left(\frac{1}{\sqrt{m}}\right)$, lazy training

$$\|w_v(t) - w_v(0)\|_2 = \left\| \int_0^t \frac{dw_v(\tau)}{d\tau} d\tau \right\|_2$$

$$\leq \int_0^t \left\| \frac{dw_v(\tau)}{d\tau} \right\|_2 d\tau$$

$$= \int_0^t \left\| \frac{1}{\sqrt{m}} \frac{1}{n} \sum_{i=1}^{n} (u_i(\tau) - y_i) a_v x_i \cdot 1\{w_v^T x_i > 0\} \right\|_2 d\tau$$

If $u_i(\tau) = 0$ ( )

$$\leq C \cdot \int_0^t \frac{1}{\sqrt{m}} 1 \cdot d\tau$$

$$\leq \frac{C \cdot t}{\sqrt{m}}$$

- ReLU smoothness

smoothness: small deviation in parameter

$\longrightarrow$ small deviation in quantity of interest

$$\begin{cases} H_{ij}(t) = x_i^T x_j \cdot \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{w_r(t)^T x_i \geq 0,\ w_r(t)^T x_j \geq 0\} \\ H_{ij}(0) = x_i^T x_j \cdot \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{w_r(0)^T x_i \geq 0,\ w_r(0)^T x_j \geq 0\} \end{cases}$$

$$H_{ij}(t) - H_{ij}(0)$$

$$= x_i^T x_j \frac{1}{m} \sum_{r=1}^{m} \left(\mathbb{1}\{w_r(t)^T x_i \geq 0,\ w_r(t)^T x_j \geq 0\} - \mathbb{1}\{w_r(0)^T x_i \geq 0,\ w_r(0)^T x_j \geq 0\} \right)$$

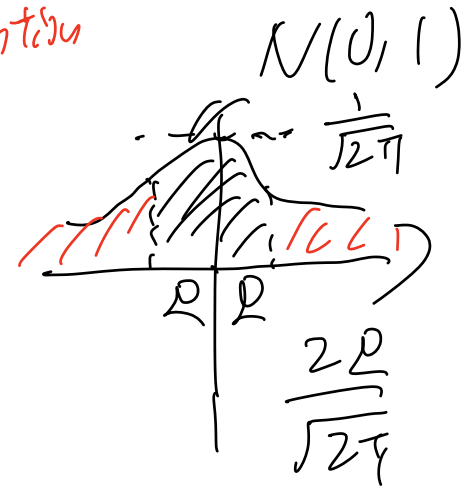
$$\leq \frac{x_i^T x_j}{m} \left[\sum_{r=1}^{m} \left(\mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_i) \neq w_r(0)^T x_i\} + \mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_j) \neq w_r(0)^T x_j\} \right) \right]$$

$\Rightarrow$ Want to bound # of pattern changes

$$\frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_i) \neq \mathrm{sgn}(w_r(0)^T x_i)\}$$

Gaussian Anti-Concentration

$$1 - \frac{2\rho}{\sqrt{2\pi}}$$



$$\Pr_{Z\sim N(0,1)}(|Z| \le \rho) \le \frac{2\rho}{\sqrt{2\pi}}$$

$$w_r(0) \sim N(0, I)$$

$$\Rightarrow w_r(0)^T x_i \sim N(0,1) \quad (\text{if } \|x_i\|_2 = 1)$$

If $\forall r$, $\|w_r(t) - w_r(0)\|_2 \le \Delta w$

$$\left(\Delta w \to 0 \text{ as } m \to \infty \quad \Delta w = O\left(\frac{1}{\sqrt{m}}\right) \right)$$

Let's choose $\rho > \Delta w$

Suppose $|w_r(0)^T x_i| \ge \rho$

$$\Rightarrow \operatorname{sgn}(w_r(t)^T x_i) = \operatorname{sgn}(w_r(0)^T x_i)$$

because $|w_r(t)^T x_i - w_r(0)^T x_i|$

$$\le \|x_i\|_2 \cdot \|w_r(t) - w_r(0)\|_2$$

$$\le \Delta w < \rho \le |w_r(0)^T x_i|$$

$$\Pr(|w_r^T(0) x_i| \le \Delta w) \le \frac{2\Delta w}{\sqrt{2\pi}}$$

we know $\delta w \to 0$ as $m \to \infty$

$$\frac{1}{m} \sum_{r=1}^{m} 1\{\mathrm{sgn}(w_r^T(t) x_i) \neq \mathrm{sgn}(w_r(0)^T x_i)\}$$

$\to 0$ as $m \to \infty$

$$|H_{ij}(t) - H_{ij}(0)| \to 0 \text{ as } m \to \infty$$

$$m = \frac{n^6}{\varepsilon^2}$$

$$\| H(t) - H(0) \|_F \leq \varepsilon$$

$$H(t) \to H^* \text{ as } m \to \infty$$

$$\frac{du(t)}{dt} = -\frac{1}{n} H(t) (u(t) - y)$$

$$\simeq -\frac{1}{n} H^* (u(t) - y)$$

universal $\to \lambda_{min}(H^*) > 0$

$\Rightarrow u(t) \to y$

For prediction

$m \to \infty$, X_{te} test point

$f(\Theta, X_{te}) \Rightarrow$ Kernel

$$K_{te} (H^*)^{-1} y \quad , y \in \mathbb{R}^n$$

H^*: NTK

$$K_{te} \in \mathbb{R}^n \quad , \begin{pmatrix} K(X_1, X_{te}) \\ \vdots \\ K(X_n, X_{te}) \end{pmatrix}$$