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.

ε - close to global min

0 error

$\text{poly}(n, \frac{1}{\varepsilon})$

Gradient Flow: a Kernel Point of View

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

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

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

If L is strongly convex $\Rightarrow$ unique θ^*, $\theta(t) \to \theta^*$

if over-parameterized multiple θ^*

Gradient Flow: a Kernel Point of View

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

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

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

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

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

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

$y \in \mathbb{R}^n$, labels

$u(t) \to y$

Gradient Flow: a Kernel Point of View

If l is quadratic, $l(u(t), y) = \frac{1}{2}\|u(t) - y\|_2^2$

$$l'(u(t), y) = u(t) - y$$

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

Claim:

If $H(t)$ is always positive definite:

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

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

Pf:

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

$$\leq -\frac{1}{n}\lambda_0 \|u(t) - y\|_2^2$$

Gradient Flow: a Kernel Point of View

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

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

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

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

when $t=0$, assume $\frac{1}{2}\|u(0)-y\|_2^2 = C$, $O(1)$

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

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

$t = O\left(\log\left(\frac{1}{\varepsilon}\right)\right)$ rate

Gradient Flow: a Kernel Point of View

$$f(\theta, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^T x)$$

m: width, $x \in \mathbb{R}^d$, $a_r \in \mathbb{R}$, $w_r \in \mathbb{R}^d$, $\sigma(\cdot): \mathbb{R} \to \mathbb{R}$

- Initialization: $a_r \sim \text{unif}\{-1, 1\}$ only for simplicity of proof

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

$$\Rightarrow f(\theta(0), x) = O(1)$$ NTK

- Training: only train $w_1, \dots w_m$

$$\min_{w_1, \dots, w_m} \frac{1}{n} \sum_{i=1}^{n} (f(x_i, a, w) - y_i)^2$$

$$u_i(t) = f(x_i, a, w(t))$$

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

Idea: $H(t) \approx H^*$ for all t

$[H^*]_{ij} = \lim_{m \to \infty} E\left[\frac{\partial u_i(0)}{\partial w} \cdot \frac{\partial u_j(0)}{\partial w}\right]$

Gradient Flow: a Kernel Point of View

$\lambda_{\min}(H^*) > 0$ $\Leftarrow$ kernel is universal

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

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

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

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

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

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

1) $H(0) \approx H^*$

2) $H(t) \approx H(0), \forall t$

Hoeffding inequality:

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

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

w.p. $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}[z_1]\right| \le \varepsilon$

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

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

when m is sufficiently large

$$H_{ij}(\theta) \longrightarrow H_{ij}^{*}$$

$$\Rightarrow \quad H(\theta) \longrightarrow H^{*}$$

Want to show $H(t) \approx H(0)$

for simplicity: 1) just train till time t

2) $y_i = O(1)$

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

$1 - h_{ij}^{\infty} = x_i^T x_j \cdot \dfrac{\pi - \arccos(x_i^T x_j)}{2\pi}$

☆ every weight w_r only moves a little: $O\left(\frac{1}{\sqrt{m}}\right)$

lazy training

$$
\begin{aligned}
&\| w_r(t) - w_r(0) \|_2 \\
&= \left\| \int_0^t \frac{d w_r(\tau)}{d\tau} d\tau \right\|_2 \\
&\le \int_0^t \left\| \frac{d w_r(\tau)}{d\tau} \right\|_2 d\tau \\
&= \int_0^t \left\| -\frac{1}{\sqrt{m}} \frac{1}{n} \sum_{i=1}^{n} (u_i(\tau) - y_i)\, a_r x_i \mathbb{1}\{w_r(\tau)^\top x_i \ge 0\} \right\|_2 d\tau
\end{aligned}
$$

$O(1)$

$$
\begin{aligned}
&\le C \cdot \int_0^t \frac{1}{\sqrt{m}} d\tau \\
&\le \frac{C \cdot t}{\sqrt{m}}
\end{aligned}
$$

— ReLU Smoothness

smooth : small deviation in para

⇒ small deviation in function derivative

$$\begin{cases} H_{ij}(t) = X_i^T X_j \frac{1}{m} \sum_{r=1}^{m} \mathbb{1}\{w_r(t)^T X_i \geq 0, w_r(t)^T X_j \geq 0\} \\ H_{ij}(0) = X_i^T X_j \quad \cdots \cdots w_r(0) \cdots w_r(0) \cdots \end{cases}$$

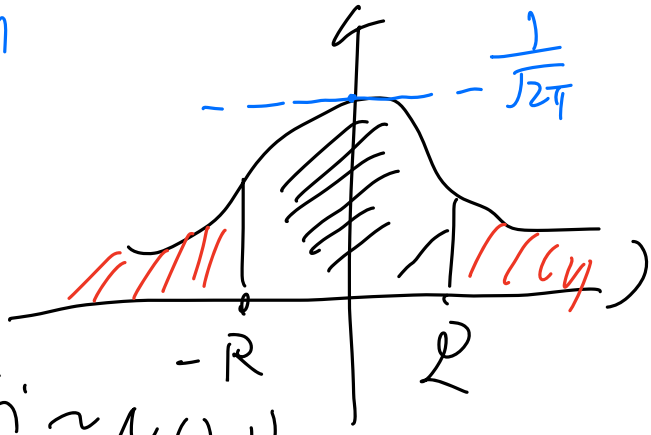
$$|H_{ij}(t) - H_{ij}(0)| \leq \frac{X_i^T X_j}{m} \left[\sum_{r=1}^{m} \mathbb{1}\{ sgn(w_r(t)^T X_i) \neq sgn(w_r(0)^T X_i) \} + \sum_{r=1}^{m} \mathbb{1}\{ sgn(w_r(t)^T X_j) \neq sgn(w_r(0)^T X_j) \} \right]$$

$$\rightarrow O\left(\frac{1}{\sqrt{m}}\right)$$

⇒ want to show # of pattern changes

Gaussian Anti-concentration

$$P_{z\sim N(0,1)}(|z| \le R) \le \frac{2R}{\sqrt{2\pi}}$$



$w_r(0) \sim \mathcal{N}(0, I) \Rightarrow w_r(0)^T x_i \sim \mathcal{N}(0,1)$

If $\forall r, \|w_r(t) - w_r(0)\| \le \delta_W$

$(\delta_W \to 0, \text{ as } m \to \infty)$

Let's choose $R > \delta_W$

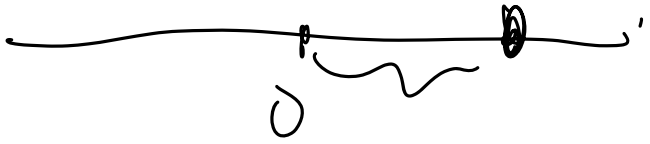
Suppose $|w_r(0)^T x_i| \ge R$, w.p. $1 - \frac{2R}{\sqrt{2\pi}}$

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

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

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

$< R \le |w_r(0)^T x_i|$



$$\Pr\left(|w_r(0)^T x_i| \le \sigma_w\right) \le \frac{2\sigma_w}{\sqrt{2\pi}}$$

we know $\sigma_w \to 0$, as $m \to \infty$

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

$$\|H(t) - H(0)\|_F \xrightarrow{m\to\infty} 0$$

$$\|H(0) - H^*\|_F \xrightarrow{m\to\infty} 0$$

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

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

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

$$\Rightarrow u(t) \to y$$

if $m = \text{poly}\left(n, \frac{1}{\varepsilon}\right)$

$$\|H(t) - H^*\| \le \varepsilon$$