- HW 1 Due Tomorrow 11:59PM
- 2 Late Days
- Course Evaluation 2/3

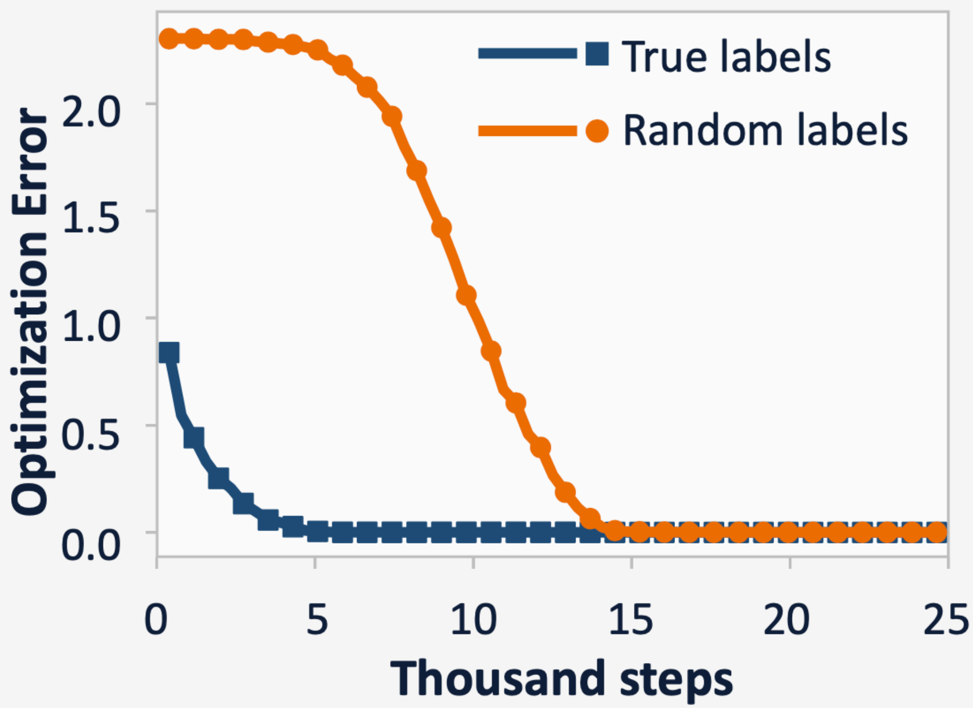
Global convergence of gradient descent

Gradient descent finds global minima

Practice: gradient descent

$$\theta(t+1) \leftarrow \theta(t) - \eta \frac{\partial L(\theta(t))}{\partial \theta(t)}$$

parameter $\gg n$



Optimization error **-> 0** for both ***true labels*** and ***random labels*** !

Zhang Bengio Hardt Recht Vinyals 2017

Understanding DL Requires Rethinking Generalization

Global convergence of gradient descent

convex loss

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 - global min

poly(n) $\log(\frac{1}{\varepsilon})$

for quadratic

neural tangent kernel

proof for a two-layer NN

Gradient Flow: a Kernel Point of View

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

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

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

if $L(\theta)$ strongly convex, $\exists$ unique θ^*, $\theta(t) \to \theta^*$

for NN, # of parameter, $\dim(\theta) > n$

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

Gradient Flow: a Kernel Point of View

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

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

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

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

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

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

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

$$[H(t)]_{ij} = \left\langle \frac{\partial u_i(t)}{\partial \theta(t)}, \frac{\partial u_j(t)}{\partial \theta(t)} \right\rangle, \qquad \boxed{\frac{du(t)}{dt} = -\frac{1}{n} H(t) \cdot \ell'(u(t), y)}$$

Gradient Flow: a Kernel Point of View

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

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

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

If $H(t)$ is always positive definite

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

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

H p.d.

$\lambda_{min}(H) \geq \lambda_0$

any vector u

$u^T H u \geq \lambda_0 \|u\|_2^2$

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

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

Gradient Flow: a Kernel Point of View

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

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

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

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

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

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

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

$\log\left(\frac{1}{\varepsilon}\right) \qquad t\to\infty$, loss $\to 0$, $u(t)-y$

kernel: $u_i(t) = \underbrace{\phi(x_i)}_{\text{feature map for some kernel}}{}^T \theta(t)$

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

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

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

$$H(t) = \begin{pmatrix} \ddots & & \\ & K(x_i, x_j) & \\ & & \end{pmatrix} \triangleq K \in \mathbb{R}^{n \times n}$$

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

for kernel $\lambda_{min}(K) > 0 \iff K$ is full-rank

if kernel is universal: Gaussian, NTK

$\Rightarrow K$ is full-rank

Gradient Flow: a Kernel Point of View

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

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

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

$w_r \sim N(0, I)$

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

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

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

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

H^*: NTK

Idea: $H(t)$ stays the same for $\forall t$

$H(t) \approx H^*$

$$H^*_{ij} = \lim_{m \to \infty} \mathbb{E}_{\text{init dist}} \left\langle \frac{\partial f(x_i)}{\partial \theta}, \frac{\partial f(x_j)}{\partial \theta} \right\rangle$$

Gradient Flow: a Kernel Point of View

$$H_{ij}(t) = \left\langle \frac{\partial u_i(t)}{\partial w(t)}, \frac{\partial u_j(t)}{\partial w(t)} \right\rangle, \quad w \in \mathbb{R}^{m \times d}$$

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

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

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

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

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

② $H(t) \approx H(0), \forall t$

Gradient Flow: a Kernel Point of View

Initialization:

Hoeffding inequality:

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

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

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

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

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

when m is large enough

$$|H_{ij}(0) - H_{ij}^*| \le \epsilon$$

Gradient Flow: a Kernel Point of View

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

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

$$\leq n^2 \cdot \varepsilon$$

a) $m \to \infty$

$$H(0) \to H^*$$

Gradient Flow: a Kernel Point of View

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

for simplicity 1) just train till time t

2) $y_i = O(1)$

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

$$H^*_{ij} = x_i^T x_j \cdot \frac{\pi - \arccos(x_i^T x_j)}{2\pi}$$

Key idea every weight vector only moves

a little $O(\frac{1}{\sqrt{m}})$: lazy training

Gradient Flow: a Kernel Point of View

$$\|w_r(t) - w_r(0)\|_2$$

$$= \left\| \int_0^t \frac{d\, w_r(\tau)}{d\tau} \, d\tau \right\|_2$$

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

$$= \int_0^t \Big\| \frac{1}{\sqrt{m}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \left(u_i(\tau) - y_i\right) \cdot a_r x_i \cdot \mathbb{1}\{w_r(\tau)^\top x_i \geq 0\}}_{O(1)} \Big\| d\tau$$

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

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

Gradient Flow: a Kernel Point of View

— ReLU Smoothness

smoothness : small movement in paramete
=) small deviation in function's derivative

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

$$|H_{ij}(t) - H_{ij}(0)| \leq \frac{x_i^T x_j}{m} \Big[\sum_{r=1}^{m} \mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_i) \neq \mathrm{sgn}(w_r(0)^T x_i)\} + \mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_j) \neq \mathrm{sgn}(w_r(0)^T x_j)\} \Big]$$

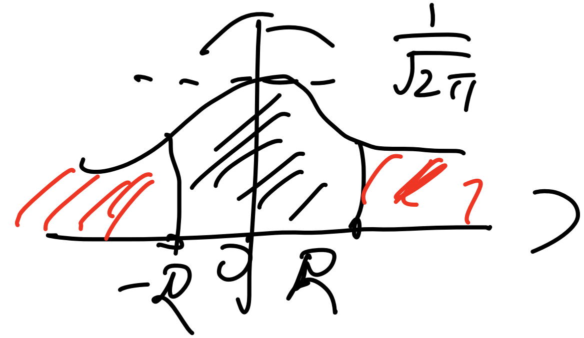
=) want to bound # of pattern changes

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

Gradient Flow: a Kernel Point of View

$N(0,1)$

Gaussian Anti-Concentration



$P(|Z| \le R) \le \frac{2R}{\sqrt{2\pi}}$

$Z \sim N(0,1)$

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

$w_r(0) \sim N(0, I) \Rightarrow w_r(0)^T x_i \sim N(0,1)$ (when $\|x_i\|_2 = 1$)

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

($\Delta w \to 0$ as $m \to \infty$, $\Delta w = O(\frac{1}{\sqrt{m}})$)

Let's choose $R > \Delta w$

Suppose $|w_r(0)^T x_i| \geq R$

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

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

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

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

Gradient Flow: a Kernel Point of View

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

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

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

$$\|H(t) - H(0)\|_F \le \varepsilon, \quad m = \frac{n^6}{\varepsilon^2}$$

$$H(t) \to H^*$$

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

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

universal $\to$ full rank

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

For predicting

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

$f(\theta, x_{te}) \longrightarrow$ kernel predicator

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

H^*: NTK

$$K_{te} \in \mathbb{R}^n, \quad \begin{pmatrix} K(x_1, x_{te}) \\ \vdots \\ K(x_n, x_{te}) \end{pmatrix}$$