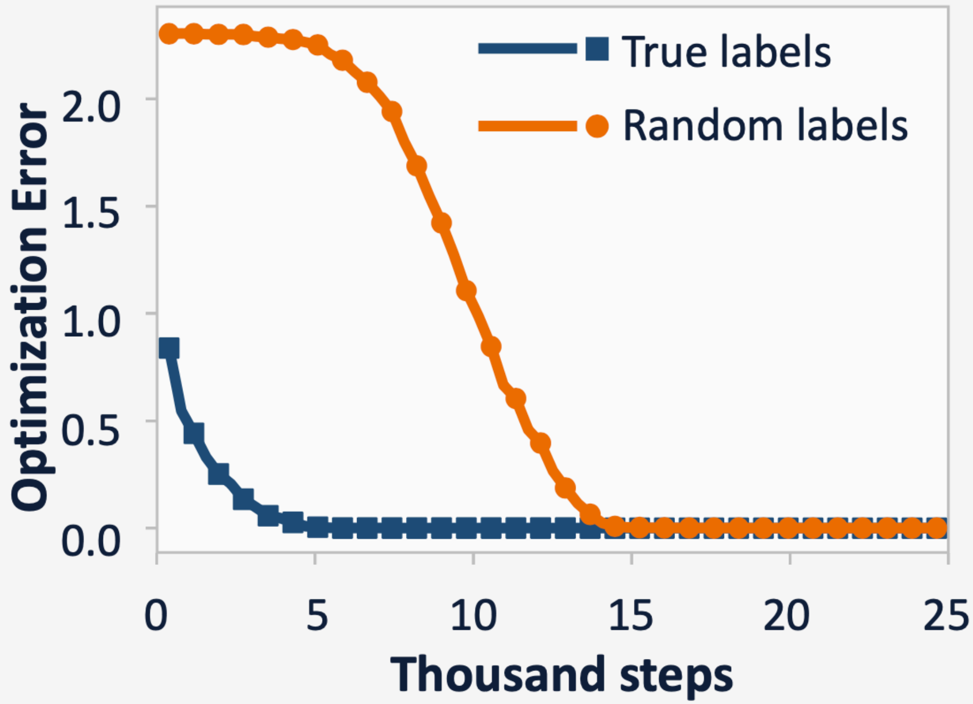
Global convergence of gradient descent

Gradient descent finds global minima

over-parameterized

Practice: gradient descent

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



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

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 - optimal

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

Gradient Flow: a Kernel Point of View

- GF: $\frac{d\theta(t)}{dt} = -\frac{\partial L(\theta(t))}{\partial \theta(t)}$
- $L(\theta) = \frac{1}{n}\sum_{i=1}^{n} l(f(\theta, x_i), y_i)$

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

- If L is strongly convex
  $\exists$ unique θ^*, $\theta(t) \to \theta^*$
- if over-parameterized
  multiple θ^*

Gradient Flow: a Kernel Point of View

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

ℓ-loss: $u(t) \to y$

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

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

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

$[H(t)]_{ij}$

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

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

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

$= \ell'(u_i(t), y_i)$

Gradient Flow: a Kernel Point of View

If ℓ is quadratic, $\ell(u_i(t), y_i) = \frac{1}{2}(u_i(t) - y_i)^2$

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

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

Claim: If $H(t)$ is always positive definite $\forall t$, $\exists \lambda_0 > 0$, $\lambda_{min}(H(t)) > 0$

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

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

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

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

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

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

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

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

$\Rightarrow$ $u(t)\rightarrow y$ as $t\rightarrow\infty$

$\square$

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

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

$w_r \sim N(0, I)$

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

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

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

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

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

H^*: $n \times n$

Kernel: NTK

Idea:

$H(t)$ almost constant for $\forall t$

$H(t) \approx H^*$

$$(H^*)_{ij} = \lim_{m \to \infty} \mathbb{E}_{\text{init}} \left\langle \frac{\partial u_i(0)}{\partial \theta}, \frac{\partial u_j(0)}{\partial \theta} \right\rangle$$

Gradient Flow: a Kernel Point of View

use functional analysis to show $\lambda_{min}(H^*) = \lambda_0 > 0$

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

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

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

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

$$(a_r^2 = 1) \qquad = \frac{1}{m} \sum_{r=1}^{m} x_i^T x_j \mathbb{1}\{w_r^T x_i \geq 0, w_r^T x_j \geq 0\}$$

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

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

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

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

$$H_{ij}(0) = X_i^T X_j \frac{1}{m}\sum_{r=1}^{m} \underbrace{\mathbb{1}\{w_r(0)^T X_i \geq 0, w_r(0)^T X_j \geq 0\}}_{z_r}$$

m R.V. average

when m is sufficiently large

$$H_{ij}(0) \longrightarrow H_{ij}^*$$

$$\Rightarrow \quad H(0) \rightarrow H^*$$

|

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

for simplicity 1) just train till some time t

2) $y_i = O(1)$

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

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

will show: every w_r only moves

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

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

— ReLU smoothness

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

$$H_{ij}(0) = \text{- - - - - - } w_r(0) \text{ - - -, } w_r(0) \cdots$$

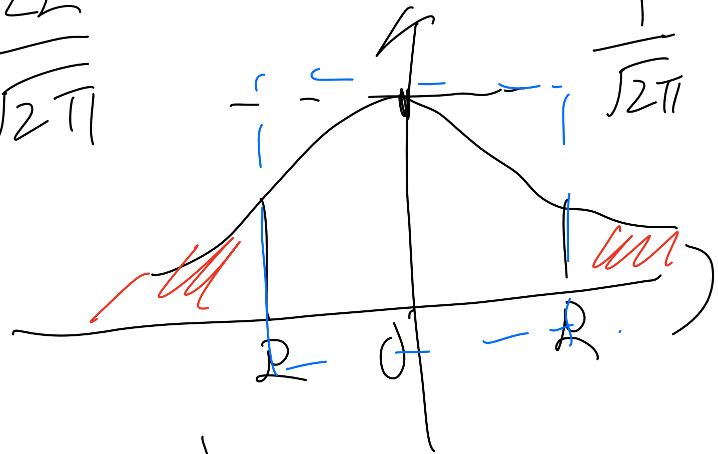
$$|H_{ij}(t) - H_{ij}(0)|$$

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

$$+ \mathbb{1}\{\mathrm{sgn}(w_r(t)^T x_j) \neq \mathrm{sgn}(w_r(0)^T x_j)\}\Big]$$

Gaussian Anti-Concentration

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



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

$\Rightarrow w_r(0)^T x_i \sim N(0,1)$

Let's choose $R > \Delta W$

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

if $\|w_r(t) - w_r(0)\| \le \Delta W < R$

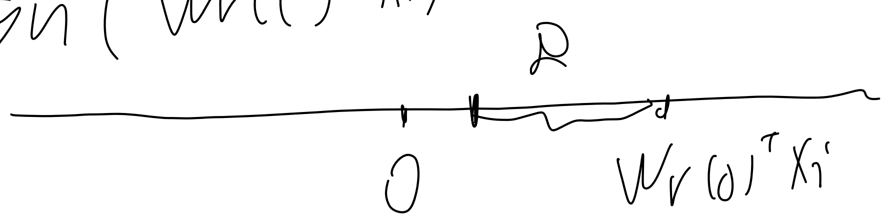
claim: $sgn(w_r(t)^T x_i) = sgn(w_r(0)^T x_i)$

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

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

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

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



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

We know $\delta w \to 0$, as $m \to \infty$

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

$H_{ij}(t) \to H(0)$ as $m \to \infty$

$\Downarrow$

$\|H(t) - H(0)\|_F \to 0$ as $m \to \infty$

$\|H(0) - H^*\|_F \longrightarrow 0$ as $m \to \infty$

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

$$(m \to \infty) \longrightarrow -\frac{1}{n} H^*(u(t) - y)$$

$\Rightarrow u(t) \to y$

$m = \text{poly}(n)$, $u(t) \to y$ $\square$