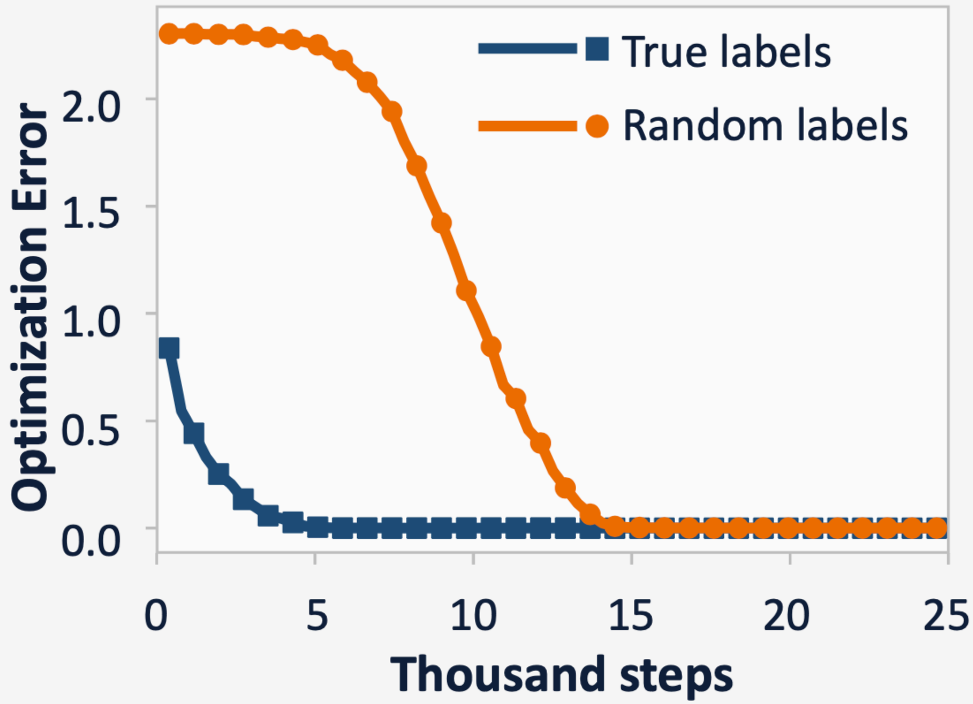
Global convergence of gradient descent

Gradient descent finds global minima

Practice: gradient descent

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



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.

Gradient Flow: a Kernel Point of View

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

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

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

- classical optimization
  if L is strongly convex
  $\Rightarrow$ $\exists$ unique θ^*, $\theta(t) \to \theta^*$

- if over-parameterized: # of params $> n$
  $\Rightarrow$ more than 1 global min

Gradient Flow: a Kernel Point of View

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

loss: $u(t) \to y$ as $t \to \infty$

(quadratic loss: $\frac{1}{n}\|u(t) - y\|_2^2$)

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

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

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

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

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

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

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

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

Gradient Flow: a Kernel Point of View

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

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

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

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

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

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

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

Gradient Flow: a Kernel Point of View

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

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

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

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

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

$$\Rightarrow \quad \frac{1}{2}\|u(t)-y\|_2^2 \leq \frac{1}{2}\|u(0)-y\|_2^2\cdot\exp\left(-\frac{\lambda_0 t}{n}\right)$$

as $t\to\infty$, $\frac{1}{2}\|u(t)-y\|_2^2 \to 0$

$\square$

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, $x \in \mathbb{R}^d$, $a_r \in \mathbb{R}$, $w_r \in \mathbb{R}^d$, σ: ReLU
- Initialization: $a_r \sim \text{unif}\{1, -1\}$, just for simplicity

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

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

assume $y = O(1)$

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

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

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

Gradient Flow: a Kernel Point of View

ReLU

$\sigma'(z) = \mathbb{1}\{z \geq 0\}$

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

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

$$\frac{\partial u_i(t)}{\partial w_r(t)} = \frac{1}{\sqrt{m}} a_r \cdot x_i \cdot \sigma'(w_r^T x_i)$$

$$H_{ij}(t) = \sum_{r=1}^{m} \left\langle \frac{1}{\sqrt{m}} a_r x_i \sigma'(w_r^T x_i), \frac{1}{\sqrt{m}} a_r x_j \sigma'(w_r^T x_j) \right\rangle$$

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

Key idea:

H^* matrix (fixed)

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

1) $H(0) \approx H^*$ if m is large

2) $H(t) \approx H(0)$ if m is large

3) H^* is positive definite

problem-dependent, data-dependent

use functional analysis $\Rightarrow$ under general conditions $\Rightarrow$ H^* p.d.

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

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

when m is large, $H_{ij}(0) \to H_{ij}^*$

$$\Rightarrow \quad \|H(0) - H^*\|_2 \to 0$$

as $m \to \infty$

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

for simplicity: 1) just train till some time t

2) $y_i = O(1)$

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

$\Rightarrow$ to show W_r only moves $O\left(\frac{1}{\sqrt{m}}\right)$

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

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

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

$$= \int_0^t \left\| -\frac{1}{\sqrt{m}} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^{n} (u_i(\tau) - y_i) a_r x_i \mathbb{I}\{ w_r(\tau)^T x_i > 0 \}}_{O(1)} \right\|_2 d\tau$$

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

- ReLU smoothness

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

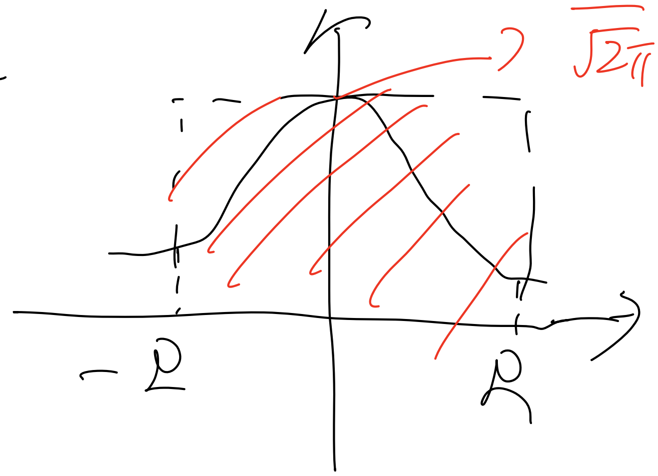
$$H_{ij}(0) = \_\ \_\ \_\ \_\ \_\ w_r(0)\ \_\ \_\ \_\ w_r(0)\ \_\ \_\ \_$$

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

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

Gaussian Anti-concentration

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



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

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

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

$|(w_r(t)-w_r(0))^T x_i|\le\Delta w<R$

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

if can control Δw be small

we can choose small R

$\Rightarrow$ can show most sgn stay the same

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

We know $\Delta w = O\left(\frac{1}{\sqrt{m}}\right)$

as $m \to \infty$

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

$$H_{ij}(t) \longrightarrow H_{ij}(0)$$

$$\|H(t) - H(0)\|_F \to 0$$

as $m \to \infty$

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

$$\xrightarrow{m\to\infty} \frac{du(t)}{dt} = -H^{*}(u(t) - y)$$

$$\Rightarrow u(t) \to y \text{ as } t \to \infty$$

$\square$