IMPLICIT BIAS of GRADIENT DESCENT

$\min_{\theta \in \mathbb{R}^p} g(\theta)$ unique minimizer if g is strongly convex.

When there are multiple global minimizers,

$$g\left(\frac{1}{T}\sum_{t<T} \theta_t\right) - \inf_{\theta \in \mathbb{R}^p} g(\theta) \le O\left(\frac{1}{T^\beta}\right), \quad \beta > 0.$$

$\longrightarrow$ convergence in function value.

An important question : Which $\theta_* \in \arg\min_{\theta \in \mathbb{R}^p} g(\theta)$ does $(\theta_t)_{t \ge 0}$ under a given optimization algorithm?

ML perspective : $g(\theta) = \frac{1}{n}\sum_{j=1}^{n} \hat{R}_s(f_\theta)$

$p \gg n$

no regularization used.

$\Big\}$ multiple ERM

an arbitrary ERM does not generalize well.

In a nutshell, GD $\longrightarrow$ minimum ℓ_2-norm solutions

$\Rightarrow$ good generalization.

Least - Squares

$$\mathcal{X} = \mathbb{R}^d,\ \mathcal{Y} = \mathbb{R},\quad f_\theta(x) = \theta^T x,$$

$$g(\theta) = \frac{1}{2n} \sum_j \left(y_j - f_\theta(x_j) \right)^2$$

Let $\Phi \triangleq \begin{bmatrix} — x_1 — \\ \vdots \\ — x_n — \end{bmatrix} \in \mathbb{R}^{n\times d}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$. Then,

$$g(\theta) = \frac{1}{2n} \| y - \Phi\theta \|_2^2$$

<u>Overparameterization</u>: $d > n$ (more parameters than data points)

$\Phi\Phi^T$ is non-singular. $\longrightarrow$ Full-rank assump.

$z^T \Phi \Phi^T z = |\Phi^T z|^2 = 0$ iff $z = 0$.

$\Rightarrow (x_1, \dots, x_n)$ are linearly independent.

Since $\mathrm{Col}(\Phi) = \mathbb{R}^n$ and $d > n$, there are infinitely many solutions s.t. $\Phi\theta = y$.

Gradient descent: $\eta \leq \dfrac{1}{\lambda_{max}(\frac{1}{n}\Phi\Phi^T)} = \dfrac{1}{\lambda_{max}(\frac{1}{n}\Phi^T\Phi)}$,

$\theta_0 = 0$,

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{n} \Phi^T (\Phi\theta_t - y)$$

Thus,

$$\Phi\theta_{t+1} - y = \Phi\theta_t - y - \eta \cdot \frac{1}{n} \Phi\Phi^T (\Phi\theta_t - y)$$

$$= \left[I - \eta \cdot \frac{1}{n} \Phi\Phi^T \right] (\Phi\theta_t - y)$$

$$= \left[I - \frac{\eta}{n} \Phi\Phi^T \right]^{t+1} \cdot (-y)$$

$$\Rightarrow \quad \| \Phi\theta_{t+1} - y \|_2^2 \leq \left(1 - \eta \cdot \frac{1}{n} \lambda_{max}(\Phi\Phi^T) \right)^{2(t+1)} \cdot \| y \|_2^2 .$$

$\Rightarrow \Phi\theta_t \to y$ as $t \to \infty$ at an exponential rate.

Starting from $\theta_0 = 0$,
$$\theta_t = \Phi^T \alpha_t \quad \text{for some } \alpha_t \in \mathbb{R}^n.$$

$\Phi\theta_t$ converges to y $\Rightarrow$ $\Phi\theta_t = \Phi\Phi^T\alpha_t$ converges to y
since $\theta \mapsto \Phi\theta$ is a continuous mapping

By the full-rank assumption, $\Phi\Phi^T =: K$ is non-singular.

Thus,
$$\Phi\Phi^T\alpha_t \xrightarrow[t\to\infty]{} y \quad \Rightarrow \quad \alpha_t \xrightarrow[t\to\infty]{} K^{-1}y.$$

$$\Rightarrow \Phi^T\alpha_t = \boxed{\theta_t \xrightarrow[t\to\infty]{} \Phi^T K^{-1} y.}$$

What is special about $\Phi^T(\Phi\Phi^T)^{-1}y$?

Let $\theta_{LN} := \Phi^T(\Phi\Phi^T)^{-1}y$, and θ be any solution of $\Phi\theta = y$. Then,

$$(\theta-\theta_{LN})^T\theta_{LN} = (\theta-\theta_{LN})^T\Phi^T(\Phi\Phi^T)^{-1}y$$
$$= [\Phi\theta - \Phi\theta_{LN}]^T(\Phi\Phi^T)^{-1}y = 0$$

Thus,
$$\|\theta\|_2^2 = \|\theta_{LN} + \theta - \theta_{LN}\|_2^2$$
$$= \|\theta_{LN}\|_2^2 + \cancelto{0}{2.(\theta-\theta_{LN})^T\theta_{LN}} + \underbrace{\|\theta-\theta_{LN}\|_2^2}_{\geq 0,\ \forall\theta.}$$
$$\geq \|\theta_{LN}\|_2^2.$$

Hence, θ_{LN} is the solution of $\Phi\theta = y$ with the minimum l_2-norm.

An alternative solution : Lagrange duality,

$$\inf_{\theta \in \mathbb{R}^d} \frac{1}{2}\|\theta\|_2^2 \quad \text{s.t.} \quad \Phi\theta = y \quad = \inf_{\theta \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}^n} \left\{ \frac{1}{2}\|\theta\|_2^2 + \lambda^T (y - \Phi\theta) \right\}$$

$$= \sup_{\lambda \in \mathbb{R}^n} \left\{ \lambda^T y - \frac{1}{2} \|\Phi^T \lambda\|_2^2 \right\} \quad \text{with} \quad \theta = \Phi^T \lambda \quad \text{at opt.}$$

$$= \sup_{\lambda \in \mathbb{R}^n} \left\{ \lambda^T y - \frac{1}{2} \lambda^T K \lambda \right\} \quad \text{where} \quad K = \Phi\Phi^T$$

Solution of the above : $\lambda^* = (\Phi\Phi^T)^{-1} y$ with optimum at

$$\theta_{LN} = \Phi^T \lambda^* = \Phi^T (\Phi\Phi^T)^{-1} y.$$

Generalization Performance under Implicit Bias

$x \sim \mathcal{N}(0, I_d)$, $y = \theta_*^T x + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$(x_j, y_j)_{j \leq n}$ given. The excess risk for $f_\theta(x) = \theta^T x$ is

$$\mathcal{R}(f_\theta) = (\theta - \theta_*)^T \mathbb{E}[x x^T] (\theta - \theta_*) = \|\theta - \theta_*\|_2^2.$$

If $\hat{\theta} = \theta_{LN}$, $d \geq n+2$, then,

$$\mathbb{E}\mathcal{R}(f_{\hat{\theta}}) = \frac{\sigma^2 n}{d-n-1} + \|\theta_*\|_2^2 \frac{d-n}{d}.$$