10 - Linearization

- a.k.a. Neural Tangent Kernel
- Last lecture, we computed risk curves associated with overparametrized linear models under implicit bias (eg. min-norm solution).

* Gradient Flows on $\mathbb{R}^d$

- We want to train the model using a gradient-based algorithm.
- Gradient flows are easier to analyze, but not practical.
- Gradient descent is practical:

$\theta_* \triangleq \arg\min_{\theta \in \mathbb{R}^d} R(\theta)$ via $\theta_{t+\eta} = \theta_t - \eta \nabla R(\theta_t)$ (GD)

(η: step size.)

(re-arrange) $\Leftrightarrow \frac{\theta_{t+\eta} - \theta_t}{\eta} = \nabla R(\theta_t)$

$\left(\lim_{\eta \to 0}\right) \Rightarrow \boxed{\frac{d\theta_t}{dt} = \dot{\theta}_t = -\nabla R(\theta_t)}$ (GF)

- GD is the Euler discretization of GF. For small η, GD behaves like GF.

Proposition: If $\boxed{\frac{d}{dt}\Delta_t \le -\mu \Delta_t}$ (C), then $\Delta_t \le e^{-\mu t}\Delta_0$.

proof:

$$\frac{d}{dt}(e^{\mu t}\Delta_t) = \mu e^{\mu t}\Delta_t + e^{\mu t}\frac{d}{dt}\Delta_t$$
$$\le \mu e^{\mu t}\Delta_t - \mu e^{\mu t}\Delta_t \quad \text{(by (C))}$$
$$= 0.$$

Integrating both sides yields the result. ∎

Assumption (Polyak-Łojasiewicz (PL) inequality): For $\mu > 0$,

$$\forall \theta, \quad R(\theta) - R(\theta_*) \leq \frac{1}{\mu} \|\nabla R(\theta)\|^2 .$$

Proposition: PL $\Rightarrow$ C for $\Delta_t = R(\theta_t) - R(\theta_*)$.

proof:

$$\frac{d}{dt}\underbrace{\{R(\theta_t) - R(\theta_*)\}}_{\Delta_t} = \langle \nabla R(\theta_t), \dot{\theta}_t \rangle$$

$$= \langle \nabla R(\theta_t), -\nabla R(\theta_t) \rangle$$

$$= -\|\nabla R(\theta_t)\|^2$$

$$\leq -\mu \underbrace{\{R(\theta_t) - R(\theta_t)\}}_{\Delta_t} \quad \text{(by PL)}$$

□

Assumption (Strong Convexity (SC)): At θ_*, R is SC:

$$\forall \theta, \quad \langle \theta - \theta_*, \nabla R(\theta) \rangle \geq \frac{\mu}{2} \|\theta - \theta_*\|^2 .$$

Proposition: SC $\Rightarrow$ C for $\Delta_t = \frac{1}{2}\|\theta_t - \theta_*\|^2$.

proof:

$$\frac{d}{dt}\underbrace{\frac{1}{2}\|\theta_t - \theta_*\|^2}_{\Delta_t} = \langle \theta_t - \theta_*, \dot{\theta}_t \rangle$$

$$= \langle \theta_t - \theta_*, -\nabla R(\theta_t) \rangle$$

$$\leq -\mu \cdot \underbrace{\frac{1}{2}\|\theta_t - \theta_*\|^2}_{\Delta_t} \quad \text{(by SC)}$$

□.

Remark: - In fact, SC $\Rightarrow$ PL so it is a stronger condition.

- In both cases, GF converges exponentially fast!

* Linearization

- Recall: We want to train a function s.t.

$$Y \approx f(x;\theta) \qquad \text{e.g. } f \text{ is NN.}$$

(Y: output, x: input, θ: parameter)

- Data: (Y_i, x_i) for $i = 1, \dots n$, let $x_i \in \mathbb{R}^d$ and $\theta \in \mathbb{R}^p$.

$$Y_n = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n \qquad f_n(\theta) = \begin{bmatrix} f(x_1,\theta) \\ \vdots \\ f(x_n;\theta) \end{bmatrix} \in \mathbb{R}^n$$

Empirical risk: $\hat{R}(\theta) = \frac{1}{2n} \| Y_n - f_n(\theta) \|^2$.

- Minimize $\hat{R}(\theta)$ with GF!

$$\dot{\theta}_t \triangleq \frac{d\theta_t}{dt} = \frac{1}{n} \nabla f_n(\theta_t) \left(Y_n - f_n(\theta_t) \right)$$

$\nabla f_n(\theta_t) \in \mathbb{R}^{p \times n}$ (Jacobian^T of $f_n : \mathbb{R}^p \to \mathbb{R}^n$)

- JGH18 argues that, in highly overparametrized regime, θ_t changes only slightly w.r.t. initialization θ_0.

- Therefore we compare θ_t with its linearization at θ_0:

$$\dot{\bar{\theta}}_t \triangleq \frac{d\bar{\theta}_t}{dt} = \frac{1}{n} \nabla f_n(\theta_0) \left(Y_n - \underbrace{f_n(\theta_0) - \nabla f_n(\theta_0)^T (\bar{\theta}_t - \theta_0)}_{= f_n^{lin}(\bar{\theta}_t)} \right)$$

$f_n^{lin}(\bar{\theta}_t)$: first-order Taylor approx around θ_0.

which minimizes $\hat{R}^{lin}(\bar{\theta}) = \frac{1}{2n} \| Y - f_n(\theta_0) - \nabla f_n(\theta_0)^T (\bar{\theta} - \theta_0) \|^2$.

* Why expect generalization?

- If $\nabla f_n(\theta_0)$ is full-rank, $\mathcal{E}_0 = \{ \bar{\theta} : \hat{R}^{lin}(\bar{\theta}) = 0 \}$ is an affine subspace of dim $p-n$.

- $\bar{\theta}_\infty = \underset{\bar{\theta}}{\operatorname{argmin}} \left\{ \|\bar{\theta} - \theta_0\| : \nabla f_n(\theta_0)^T (\bar{\theta} - \theta_0) = y - f_n(\theta_0) \right\}$

  $\in \mathcal{E}_0 \Rightarrow$ min-norm solution $\Rightarrow$ implicit regularization!

  (see previous lecture) (A3)

- But how realistic is this approximation:

  1. $\theta_t \approx \bar{\theta}_t \quad \forall t$? 2. $f_n(\theta_t) \approx f_n^{lin}(\bar{\theta}_t)$?

  This regime is called the linear regime.

- $\bar{\theta}_\infty$ generalizes + $\theta_t \approx \bar{\theta}_t \quad \forall t$.

  $\Rightarrow$ GF trained NN generalizes!

Theorem: Let $\sigma_{max/min} = \sigma_{max/min}(\nabla f_n(\theta_0))$. Assume ∇f_n is L-Lipschitz and

$$\|y - f_n(\theta_0)\| < \frac{\sigma_{min}^2(\nabla f_n(\theta_0))}{4L}.$$

1- Then, the empirical risk decreases exponentially fast to 0, i.e.

for $\mu = \sigma_{min}^2 / 2n$ $\quad \hat{R}(\theta_t) \leq \hat{R}(\theta_0) \cdot e^{-\lambda_0 t}$.

2- Parameters stay close to initialization,

$$\|\theta_t - \theta_0\| \leq \frac{2}{\sigma_{min}} \|y - f_n(\theta_0)\|$$

Remarks:

1- Theorem is about the original GF. No linearization yet.

2- Initialize close to the global min

$\Rightarrow$ converge to it exponentially fast (local convergence).

Proof: 1. Let $f_t \triangleq f_n(\theta_t)$ so $\hat{R}(\theta_t) = \frac{1}{2}\|f_t - Y_n\|^2$.

$$\dot{f}_t = \nabla f_n(\theta_t)\,\dot{\theta}_t = -\frac{1}{n}\underbrace{\nabla f_n(\theta_t)\nabla f_n(\theta_t)^T}_{:= K_t \text{ empirical kernel}}(f_t - Y_n)$$

$$= -\frac{1}{n}K_t(f_t - Y_n) \qquad \simeq K_0 : \text{Neural Tangent Kernel}$$

Also, $\frac{d}{dt}\hat{R}(\theta_t) = \frac{d}{dt}\frac{1}{2}\|f_t - Y_n\|^2 = \langle f_t - Y_n, \dot{f}_t \rangle$ (*)

$$= \frac{-1}{n}\langle f_t - Y_n, K_t(f_t - Y_n)\rangle \;\because$$

Technical detail

- Let $r_* = \frac{\sigma_{\min}}{2L}$ and $t_* = \inf\{t : \|\theta_t - \theta_0\| > r_*\}$.

Below calculations are for $t \le t_*$.

Fact (Weyl's inequality): Eigen/Singular values are 1-Lipschitz.

$$t \le t_* \Rightarrow \left|\sigma_{\min}(\nabla f_n(\theta_t)) - \sigma_{\min}(\nabla f_n(\theta_0))\right| \le \|\nabla f_n(\theta_t) - \nabla f_n(\theta_0)\|_{op}$$

$$\le L\|\theta_t - \theta_0\| \le \frac{\sigma_{\min}}{2}$$

$$\Rightarrow \lambda_{\min}(K_t) = \sigma_{\min}^2(\nabla f_n(\theta_t)) \ge \frac{\sigma_{\min}^2}{4} \quad \text{for } t \le t_*.$$

$\Rightarrow$ (*) reads $\frac{d}{dt}\frac{1}{2}\|f_t - Y_n\|^2 = \frac{-1}{n}\langle f_t - Y_n, K_t(f_t - Y_n)\rangle$ ($)

$$\le -\underbrace{\frac{\sigma_{\min}^2}{2n}}_{:=\mu}\frac{1}{2}\|f_t - Y_n\|^2 \qquad (c) \text{ is satisfied.}$$

$$\Rightarrow \underbrace{\|f_t - Y_n\|^2}_{\hat{R}(\theta_t)} \le e^{-\mu t}\underbrace{\|f_0 - Y_n\|^2}_{\hat{R}(\theta_0)}$$

2.

$$\|\dot\theta_t\| = \frac{1}{n}\|\nabla f_n(\theta_t)^\top (f_t - y_n)\|$$

$$\frac{d}{dt}\|f_t - y_n\| = \frac{\langle f_t - y_n, \dot f_t\rangle}{\|f_t - y_n\|} = -\frac{1}{n}\frac{\|\nabla f_n(\theta_t)^\top (f_t - y_n)\|^2}{\|f_t - y_n\|}$$

$$\le -\frac{\sigma_{\min}}{2n}\|\nabla f_n(\theta_t)^\top (f_t - y_n)\|$$

Thus, $\frac{d}{dt}\left\{\|f_t - y_n\| + \frac{\sigma_{\min}}{2}\|\theta_t - \theta_0\|\right\}$

$$= \frac{d}{dt}\|f_t - y_n\| + \frac{\sigma_{\min}}{2}\frac{d}{dt}\|\theta_t - \theta_0\|$$

$$\frac{d}{dt}\|\theta_t - \theta_0\| = \frac{\langle \theta_t - \theta_0, \dot\theta_t\rangle}{\|\theta_t - \theta_0\|} \le \|\dot\theta_t\| \quad (a)$$

(unit norm)

$$\le \frac{d}{dt}\|f_t - y_n\| + \frac{\sigma_{\min}}{2}\|\dot\theta_t\| \le 0. \quad (b)$$

$\Rightarrow$ $\frac{d}{dt}\|\theta_t - \theta_0\| \le \|\dot\theta_t\|$ by (a)

$$\le -\frac{2}{\sigma_{\min}}\frac{d}{dt}\|f_t - y_n\| \quad \text{by (b)}$$

Integrating $\|\theta_t - \theta_0\| \le \frac{2}{\sigma_{\min}}\|f_0 - y_n\|.$

- Finally, if $t_* < \infty$, $\|\theta_{t_*} - \theta_0\| \le \frac{2}{\sigma_{\min}}\|f_0 - y_n\|$

$$< \frac{2}{\sigma_{\min}}\frac{\sigma_{\min}^2}{4L} = \frac{\sigma_{\min}}{2L} = r_*.$$

(contradiction!) ∎

$t_* = \infty$.

Technical details

* So far, we analyzed the actual GF. How about the linearization?
* Need $\theta_t \approx \bar\theta_t$ so that we get generalization via implicit bias!

Theorem: Original and linearized flows closely track each other

$$\|\theta_t - \bar{\theta}_t\| \le \frac{4}{\sigma_{\min}^2}\|Y_n - f_n(\theta_0)\|\left\{\frac{L}{\sigma_{\min}}\|Y_n - f_n(\theta_0)\| + \sigma_{\max}\right\}.$$

Remark: $\Rightarrow$ Implicit bias.

! Sharper analysis is possible (See BMR21).

proof: Let $\bar{f}_t = f_n(\theta_0) + \nabla f_n(\theta_0)(\bar{\theta}_t - \theta_0)$. As before, we get

$$\frac{d}{dt}\bar{f}_t = -\frac{1}{n}K_0(\bar{f}_t - Y_n)$$

Neural Tangent Kernel

↳ This is now constant due to linearization.

Applying ($) for K_t replaced with K_0, we get

$$\|\bar{f}_t - Y_n\| \le e^{-\mu t}\|\bar{f}_0 - Y_n\|.$$

↳ $= f_0$ since $\theta_0 = \bar{\theta}_0$.

$$\Rightarrow \|f_t - \bar{f}_t\| \le \|f_t - Y_n\| + \|\bar{f}_t - Y_n\|$$
$$\le 2e^{-\mu t}\|f_0 - Y_n\|$$

Recall $\dot{\theta}_t = \frac{1}{n}\nabla f_n(\theta_t)^T(Y_n - f_t)$

$\dot{\bar{\theta}}_t = \frac{1}{n}\nabla f_n(\theta_0)^T(Y_n - \bar{f}_t)$

Thus,

$$\frac{d}{dt}\|\theta_t - \bar{\theta}_t\| \le \|\dot{\theta}_t - \dot{\bar{\theta}}_t\|$$
$$\le \frac{1}{n}\|\nabla f_n(\theta_t) - \nabla f_n(\theta_0)\| \cdot \|Y_n - f_t\| + \frac{1}{n}\|\nabla f_n(\theta_0)\| \cdot \|f_t - \bar{f}_t\|$$
$$\le \frac{L}{n}\underbrace{\|\theta_t - \theta_0\|}_{\le \frac{2}{\sigma_{\min}}\|Y - f_0\|} \cdot \underbrace{\|Y_n - f_t\|}_{\le e^{-\mu t}\|f_0 - Y_n\|} + \frac{\sigma_{\max}}{n}\underbrace{\|f_t - \bar{f}_t\|}_{\le 2e^{-\mu t}\|f_0 - Y_n\|}$$

$$\leq 2e^{-\mu t}\|y_n - f_0\|\left\{\frac{L}{n\cdot\sigma_{min}}\|y_n - f_0\| + \frac{\sigma_{max}}{n}\right\}.$$

Integrating this, we get (note $\theta_0 = \bar{\theta}_0$)

$$\|\theta_t - \bar{\theta}_t\| \leq \frac{2}{\mu}\|y_n - f_0\|\left\{\frac{L}{n\cdot\sigma_{min}}\|y_n - f_0\| + \frac{\sigma_{max}}{n}\right\}.$$

✳ Two-layer Neural Networks

- Consider $f(x;\theta) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_j\,\sigma(\langle w_j, x\rangle)$, $\theta = (w_1, \dots, w_m)$. $p = md$.
  - m: width
  - a_j: 1st layer weights
  - w_j: 2nd layer weights
- For simplicity choose half of a_j's $+1$ and other half -1.
- Initialize θ_0: $w_i \sim Unif(S^{d-1})$
- Train via GF.
- We can compute: $\left[\nabla f(x;\theta)\right]_{i,(j,k)} = \frac{1}{\sqrt{m}}\,a_j\,\sigma'(\langle w_j, x_i\rangle)\,x_{ik}$

  $i \in [n]$, $(j,k) \in \underbrace{[m]\times[d]}_{p = md}$.

Lemma (BMR21): Under certain conditions, w.h.p.

1. $\|y_n - f_n(\theta_0)\| \lesssim \sqrt{n}$
2. $\sigma_{min} \gtrsim \sqrt{d}$
3. $\sigma_{max} \lesssim \sqrt{n} + \sqrt{d}$
4. $L \lesssim \sqrt{\frac{d}{m}}\left(\sqrt{n} + \sqrt{d}\right)$

Corollary: If $m > n^2$ and $d > n$,

$$\|\theta_t - \bar{\theta}_t\| \longrightarrow o_n(1).$$

Remarks: 1- Since $\bar{\theta}_0$ is the min-norm solution $\Rightarrow$ implicit bias $\Rightarrow$ generalization.

2- This result can be sharpened. See [BMR21].

proof:

By the theorem, we get

$$\|\theta_t - \bar{\theta}_t\| \le \frac{4}{\sigma_{\min}^2}\|Y_n - f_n(\theta_0)\|\left\{\frac{L}{\sigma_{\min}}\|Y_n - f_n(\theta_0)\| + \sigma_{\max}\right\}$$

(using the estimates in the Lemma)

$$\lesssim \frac{1}{d}\sqrt{n}\left\{\sqrt{\frac{d}{\mu}}(\sqrt{d}+\sqrt{n})\sqrt{n} + n + \sqrt{nd}\right\}$$

$$\lesssim \frac{n}{\sqrt{\mu}} + \sqrt{\frac{n}{d}}.$$

□