Deep Learning theory - Lecture 10.

Recap: Single layer NNs

$$f(\vartheta) = \frac{1}{\sqrt{m}} \sum^{m} a_j \sigma(w_j x + b_j) \qquad \vartheta = \{(a_j, w_j, b_j)\}$$

behave like linear models: tangent to $\mathcal{F}_\vartheta$ at ϑ_0:

$$\bar{f}(\bar{\vartheta}) = f(\vartheta_0) + D_\vartheta f(\vartheta_0)(\bar{\vartheta} - \vartheta_0)$$

Thm: As $m \to \infty$, $\mathrm{Lip}\, D_\vartheta f \sim O(\frac{1}{\sqrt{m}})$ so that

- $\hat{R}(\vartheta_t) = e^{-\lambda_* t} \hat{R}(\vartheta_0)$
- $\| f(\vartheta_t) - \bar{f}(\bar{\vartheta}_t) \|^2_{L^2(\mathbb{P})} = O(\frac{1}{\sqrt{m}})$ unif. in t.

Program: - study ridge (linear) regression models in overpar. regime
- study kernel methods ($p \to \infty$) in overpar. regime.

Reminder: Ridge regression: $\mathcal{F}_\Theta = \{z \mapsto z \cdot \beta, \ \beta \in \mathbb{R}^d\}$

$$\hat{R}_n(\beta) = \frac{1}{n} \sum_{j=1}^{n} (z_j \beta - y_j)^2 + \lambda \|\beta\|_2^2 = \frac{1}{n} \|Z\beta - y\|_2^2 + \lambda \|\beta\|_2^2$$

$$Z = \begin{pmatrix} — & z_1 & — \\ & \vdots & \\ — & z_n & — \end{pmatrix} \in \mathbb{R}^{n \times d}$$

λ-convex $\Rightarrow$ unique minimizer $\hat{\beta}_\lambda = \frac{1}{n}\left(\frac{1}{n} Z^T Z + \lambda \mathbb{1}\right)^{-1} Z^T y$

Depending on the rank of $Z^T Z$ we have $= \frac{1}{n} Z^T \left(\frac{1}{n} Z Z^T + \lambda \mathbb{1}\right)^{-1} y$

a) $Z^T Z$ full rank $\Rightarrow \hat{\beta}_{0+} = (Z^T Z)^{-1} Z^T y$

b) rank $Z^T Z \neq d \Rightarrow \hat{\beta}_{0+} = \operatorname{argmin}\{\|\beta\|_2^2 : Z\beta = y\}$.

Consider a simple model for distribution of z_j, y_j:

$$y_j = \beta^* \cdot z_j + \varepsilon_j$$

$\mathbb{E}(\varepsilon_j | w_j) = 0$ $\quad$ $\mathbb{E}(z_j) = 0$

$\mathbb{E}(\varepsilon_j^2 | w_j) = \tau^2$ $\quad$ $\mathbb{E}(z_j z_j^T) = \Sigma$ $\longrightarrow$ $z_j \overset{iid}{\sim} N(0, \Sigma)$ (Assume to simplify)

In general (bias-variance decomposition)

$$R(Z, \lambda) = B(Z, \lambda) + V(Z, \lambda)$$

$$\mathbb{E}_\varepsilon R_{exc}(Z, \lambda) = \mathbb{E}_\varepsilon \mathbb{E}_{z_{new}}\left(\left(\hat{\beta}(Z, \varepsilon) \cdot z_{new} - \beta^* z_{new}\right)^2\right)$$

$$= \mathbb{E}_\varepsilon \mathbb{E}_{z_{new}}\left(\left(\hat{\beta}(Z, \varepsilon) - \beta^*\right)^T z_{new} z_{new}^T \left(\hat{\beta}(Z, \varepsilon) - \beta^*\right)\right)$$

$$= \mathbb{E}_\varepsilon \|\hat{\beta}(Z, \varepsilon) - \beta^*\|_\Sigma^2 = \underbrace{\|\hat{\beta}(Z, \varepsilon) - \mathbb{E}_\varepsilon(\hat{\beta}(Z, \varepsilon))\|_\Sigma^2}_{B(Z, \lambda)}$$

$$+ \underbrace{\mathbb{E}_\varepsilon \| \mathbb{E}_\varepsilon (\hat\beta(z,\varepsilon)) - \beta^* \|_\Sigma^2}_{V(z,\lambda)}$$

For our model we can compute

$$\hat\beta(z,\lambda) = \frac{1}{n} S_\lambda z^\top y \quad \text{for} \quad S_\lambda = \left(\frac{1}{n} z^\top z + \lambda \mathrm{Id}_d\right)^{-1}$$

so that

$$B(\lambda, z) = \lambda^2 \langle \beta_*, S_\lambda \Sigma S_\lambda \beta_* \rangle$$

$$V(\lambda, z) = \frac{\tau^2}{n} \mathrm{Tr}\left(\frac{1}{n} S_\lambda^2 \hat\Sigma \Sigma\right). \qquad \hat\Sigma = \frac{1}{n} z^\top z$$

These quantities are Random in z, but they concentrate around deterministic values

Write formulas for $\lambda = 0_+$ for simplicity:

Thm: Assume that $c^{-1} \le \frac{n}{p} \le c$, $\lambda_{\max}(\Sigma) \le c$, $\frac{1}{p} \sum \lambda_j^{-1}(\Sigma) \le c$, $\|\Sigma\| = 1$

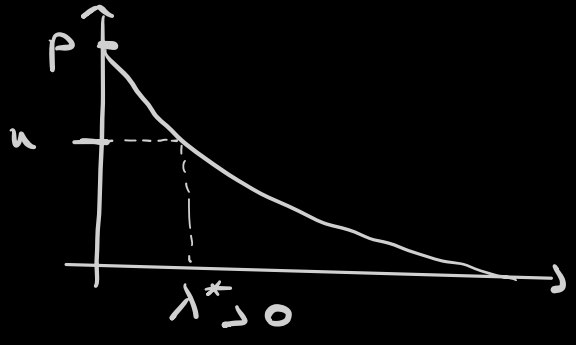
whp.
$$|B_n(z,\lambda) - \mathcal{B}_n(\lambda_*)| \le \bar c\, n^{-\frac{1}{6}}$$
$$|V_n(z,\lambda) - \mathcal{V}_n(\lambda_*)| \le \bar c\, n^{-\frac{1}{6}}$$

where

$$\mathcal{B}_n(\lambda_*) = \frac{\lambda_*^2 \beta_*^T (\Sigma + \lambda_* I)^{-2} \Sigma \beta_*}{1 - \frac{1}{n} \mathrm{Tr}(\Sigma^2 (\Sigma + \lambda_* I)^{-2})}$$

$$\mathcal{V}_n(\lambda_*) = \frac{\tau^2}{n} \frac{\mathrm{Tr}(\Sigma + \lambda_* I)^{-1} \Sigma^2}{1 - \frac{1}{n} \mathrm{Tr}(\Sigma^2 (\Sigma + \lambda_* I)^{-2})}$$

and λ^* solves $n = \mathrm{Tr}(\Sigma (\Sigma + \lambda_* I)^{-1})$ (*)



Note (implicit regularization) The model

deterministic $y_s = \Sigma^{1/2} \beta^* + \frac{\omega}{\sqrt{n}} \varepsilon'$ $\quad \varepsilon' \sim iid\ \mathcal{N}(0, I)$

fitted through

$$\hat{\beta}(\lambda_*) = \operatorname{argmin}_\beta \underbrace{\frac{1}{n} \| \Sigma^{1/2} \beta - y \|^2 + \lambda_* \| \beta \|_2^2}_{\hat{R}(\lambda_*, \varepsilon')}$$

has the risk decomposition

$$R_n(\lambda_*) = \mathbb{E}_{\varepsilon'} \hat{R}(\lambda_*, \varepsilon') = \mathcal{B}_n(\lambda_*) + \mathcal{V}_n(\lambda_*)$$

$\Rightarrow$ asymptotically, it is like fitting a simpler deterministic model with positive regularization! in $\mathbb{Z}$.

Bounding terms $\mathcal{B}_n, \mathcal{V}_n$.

Start with denominator:

$$\frac{1}{n}\mathrm{Tr}\left(\Sigma^2(\Sigma+\lambda_* I)^{-2}\right) \le \frac{1}{n}\mathrm{Tr}\left(\Sigma(\Sigma+\lambda_* I)^{-1}\right) = 1$$

cond $*$

assume that $\frac{1}{n} \mathrm{Tr}\left(\Sigma^2 \left(\Sigma + \lambda_* \mathrm{Id}\right)^{-2}\right) < 1 - \frac{1}{c^*}$ (**)

$$\Longrightarrow \begin{cases} \mathcal{B}_n(\lambda_*) \le \lambda_*^2 \, \beta_*^T \left(\Sigma + \lambda_* \mathrm{Id}\right)^{-2} \Sigma \, \beta_* \\ V_n(\lambda_*) \le \frac{T^2}{n} \mathrm{Tr}\left(\Sigma + \lambda_* \mathrm{Id}\right)^{-1} \Sigma^2 \end{cases}$$

Consider λ_* small, $\{\sigma_k\} = \sigma(\Sigma)$ $\Sigma = \sum \sigma_j v_j v_j^T$

$\sigma_1 \ge \sigma_2 \ge \dots \sigma_k \ge \dots$

Define k_* such that $\sigma_k \ge \lambda_* \ge \sigma_{k+1}$

Prop: If (**) holds

$$V_n \le c_* T^2 \frac{k_*}{n}$$

$$\mathcal{B}_n(\lambda_*) \le c_* \left(\sigma_{k_*}^2 \| \beta_{*, \le k_*} \|_{\Sigma^{-1}}^2 + \| \beta_{*, \ge k} \|_{\Sigma}^2 \right)$$

where $\beta_{*, \le k_*} = \sum_j^{k_*} (\beta_* \cdot v_j) \cdot v_j$

$\Longrightarrow$ even if $p > n$ we can have.

$V_n \xrightarrow{n \to \infty} 0$ $\left(\frac{k_*}{n} \to 0 \right)$

$\mathcal{B}_n \xrightarrow{n \to \infty} 0$ (β_* concentrated on top eigenvalues of Σ)

"Example": $\beta_* = \begin{pmatrix} \beta_1 \\ \beta_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & \ddots \end{pmatrix}$

Benign overfitting

$$\Rightarrow \quad n = \sum \frac{\sigma_k}{\sigma_k + \lambda_*} \geq \sum_{1}^{k_*} \frac{\sigma_k}{2\sigma_k} + \sum_{k_*}^{p} \frac{\sigma_k}{2\lambda_*}$$

$$\geq k + \frac{1}{\sigma_k} \sum_{k_*}^{p} \sigma_k \geq k + r_k^{(1)}$$

Example: $\sigma_k \simeq k^{-\alpha} \quad \alpha > 1 \Rightarrow r_k^{(1)} = \frac{1}{k^{-\alpha}} \int_k^{\infty} x^{-\alpha} dx \simeq k$

$$\Rightarrow n \simeq k$$

$$\Rightarrow V_n \simeq c_* \frac{T^2}{n} \operatorname{Tr}\left(\Sigma^2 (\Sigma + \lambda_* I)^{-2}\right)$$

$$\simeq c_* \frac{T^2}{n} \left(k_* + \frac{1}{\sigma_k} \sum_{k_*} \sigma_j^2\right) \simeq c_* \frac{T^2}{n} k_*$$

$$\Rightarrow B_n(\beta_*) \leq c_* \sum \frac{\lambda_*^2 \sigma_k}{(\sigma_k + \lambda_*)^2} (v_k^T \beta_*)^2 \qquad \bar{\beta}_k$$

$$= c_* \sigma_{k_*}^2 \left(\sum_{1}^{k_*} \bar{\beta}_k^2 \frac{1}{\sigma_k} + \sum_{k_*} \bar{\beta}_k^2 \frac{\sigma_k^2}{\sigma_{k_*}^2} \right)$$

Example (latent space model)

$$\begin{cases} y_i = \theta_*^T x_i + \xi_i & \xi_i \sim \mathcal{N}(0, \tau^2) \\ z_i = W x_i + v_i & v_i \sim \mathcal{N}(0, \mathbb{1}). \end{cases}$$

$x_i \in \mathcal{N}(0, \mathbb{1}_d)$ $\quad W^T W = \frac{p}{d}$

assume $\frac{p}{n} \to \gamma \in (0, \infty)$, $\frac{d}{p} \to \psi \in (0, 1)$