

Mean Field Asymptotics in Statistical Learning.

Mar 17.

Lecture 16. Double descent and generalized linear models.

① Double descent in linear models.

Training dataset: $(x_i, y_i)_{i \in [n]} \in \mathbb{R}^d \times \mathbb{R}$.

Linear model: $y_i = \langle x_i, \beta_0 \rangle + w_i$.

where $x_i \sim \text{iid } N(0, I_d)$ $w_i \sim \text{iid } N(0, \sigma^2)$.

Further assume $\beta_0 \sim N(0, \frac{1}{d} I_d)$

Linear ridgeless regression:

$$\hat{\beta}_\lambda = \arg \min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 + \frac{\lambda d}{2n} \|\beta\|_2^2 = (X^T X + d\lambda I)^{-1} X^T y.$$

$$\hat{\beta}_0 = \lim_{\lambda \rightarrow 0^+} \hat{\beta}_\lambda = X^T y = \begin{cases} \text{OLS solution} & d < n \\ \text{Min norm interpolating} & d > n \end{cases} \text{ sol.}$$

Test error: $\mathbb{E}_x[\langle x, \hat{\beta}_\lambda \rangle - \langle x, \beta_0 \rangle]^2 = \|\hat{\beta}_\lambda - \beta_0\|_2^2$.

$$R(\lambda, r, \sigma^2) \equiv \lim_{\substack{d \rightarrow \infty \\ d/n \rightarrow r}} \mathbb{E}_{\beta_0, X, w} [\|\hat{\beta}_\lambda - \beta_0\|_2^2]$$

$n/d \rightarrow r$

Bias-Variance decomposition:

$$\begin{aligned} \mathbb{E} \|\hat{\beta}_\lambda - \beta_0\|_2^2 &= \mathbb{E} \left\| ((X^T X + d\lambda I)^{-1} X^T X - I_d) \beta_0 + (X^T X + d\lambda I)^{-1} X^T w \right\|_2^2 \\ &= \underbrace{\mathbb{E} [\|((X^T X + d\lambda I)^{-1} X^T X - I_d) \beta_0\|_2^2]}_{B(\lambda)} + \underbrace{\mathbb{E} [\|(X^T X + d\lambda I)^{-1} X^T w\|_2^2]}_{V(\lambda)}. \end{aligned}$$

Theorem: [Hastie, Montanari, Rosset, Tibshirani, 2020]

[Dobriban, Wager, 2015].

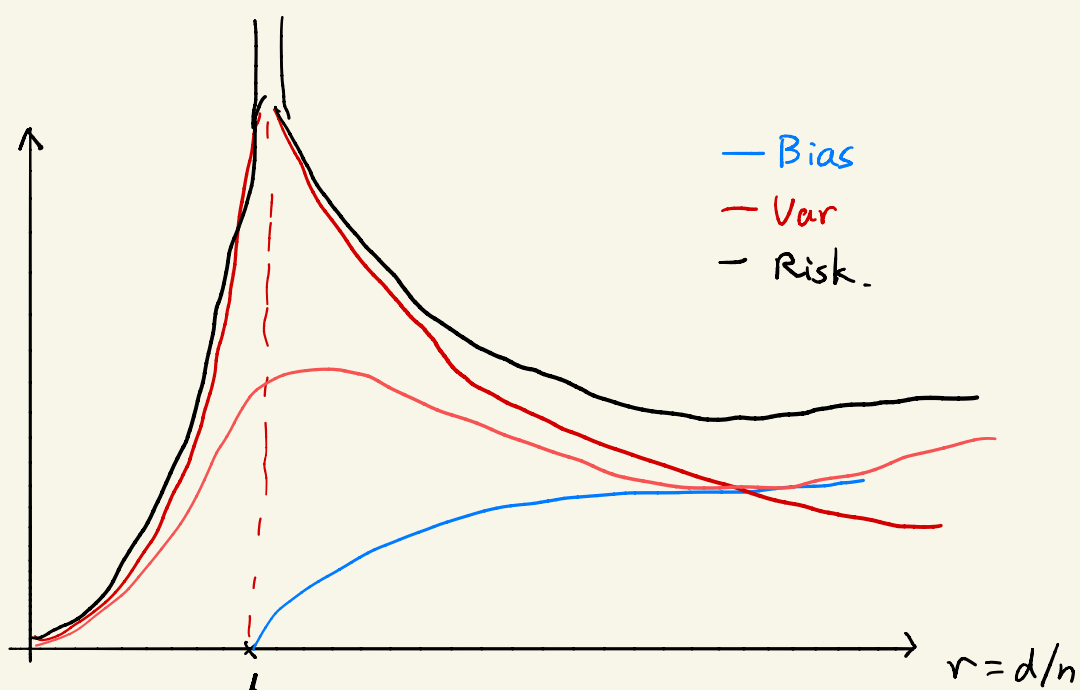
Under the assumptions above, as $n, d \rightarrow \infty$
and $d/n \rightarrow r$, we have

$$B(0) \rightarrow \left(1 - \frac{1}{r}\right) \mathbb{1}_{\{r > 1\}}$$

$$V(0) \rightarrow \sigma^2 \left[\frac{r}{1-r} \mathbb{1}_{\{r < 1\}} + \frac{1}{r-1} \mathbb{1}_{\{r > 1\}} \right].$$

$$\text{So } R(r, \sigma^2) = \begin{cases} \sigma^2 \left(\frac{r}{1-r} \right), & r < 1. \\ \left(1 - \frac{1}{r}\right) + \sigma^2 \frac{1}{r-1}, & r > 1. \end{cases} \quad r = \frac{d}{n}$$

* Easier to calculate $\lim_{\lambda \rightarrow 0} \lim_{d \rightarrow \infty} B(\lambda)$, harder to prove $\lim_{d \rightarrow \infty} B(0)$.



Double descent curve.

Approach 1: Spectrum of Wishart matrix.

$$B(\lambda) = \mathbb{E} \left\| \left((X^T X + d\lambda I_d)^{-1} X^T X - I_d \right) \beta_0 \right\|_2^2$$

$$= \mathbb{E} \operatorname{tr} \left[(X^T X + d\lambda I_d)^{-2} \right] \times \lambda^2 d.$$

$$= \lambda^2 \cdot \mathbb{E} \left[\int_{(0, \infty)} \frac{1}{(s + \lambda)^2} \hat{\mu}(ds) \right] \quad \hat{\mu} \text{ empirical dist. of eigenvalue of } X^T X / d.$$

$$\hat{\mu}(ds) = \frac{1}{d} \sum_{j=1}^d \delta_{\lambda_j(X^T X / d)}$$

$$V(\lambda) = \sigma^2 \cdot \mathbb{E}_X \left[\operatorname{tr} \left((X^T X + d\lambda I_d)^{-2} X^T X \right) \right]$$

$$= \sigma^2 \cdot \mathbb{E} \int_{(0, \infty)} \frac{s}{(s + \lambda)^2} \hat{\mu}(ds)$$

Thm: (Marchenko - Pastur law).

$$X \in \mathbb{R}^{n \times d}$$

As $d, n \rightarrow \infty$, $d/n \rightarrow r \in (0, 1)$,

$$X_{ij} \sim N(0, 1).$$

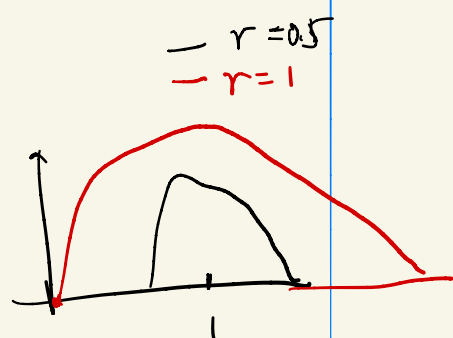
For sufficient nice function ψ ,

$$X^T X \quad X X^T$$

$$\frac{1}{d} \sum_{j=1}^d \psi(\lambda_j(X^T X / d)) \xrightarrow[d \rightarrow \infty]{a.s.} \int \psi(s) \mu(ds).$$

$$\mu(ds) = \frac{1}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - s)(s - \lambda_-)}}{rx} \mathbb{1}(s \in [\lambda_-, \lambda_+]) ds.$$

$$\lambda_{\pm} = \sigma^2(1 \pm \sqrt{r}).$$



in $B(\lambda)$ $\psi_B = \frac{1}{(s + \lambda)^2}$

in $V(\lambda)$ $\psi_V(s) = \frac{s}{(s + \lambda)^2}$

Approach 2: Stieltjes transform of Wishart matrix.

Lemma: Define $S(t, \lambda) = \text{tr}[(tX^T X + d\lambda I_d)^{-1}]$.

$$\text{then } \partial_t S(t, \lambda) = -\text{tr}[(X^T X + d\lambda I_d)^{-2} X^T X]$$

$$\partial_\lambda S(t, \lambda) = -\text{tr}[(X^T X + d\lambda I_d)^{-2}] \times d.$$

$$B(\lambda) = -\lambda^2 \times \mathbb{E} \partial_\lambda S(1, \lambda)$$

$$V(\lambda) = -\sigma^2 \times \mathbb{E} \partial_t S(1, \lambda).$$

Linear pencil approach.

(A) Calculate ^{asymptotic} $S(t, \lambda)$.

(B) Show $\partial_t, \partial_\lambda$ can be exchanged with asymptotic.

(A). Apply CGMT.

$$\mathbb{E}_X S(t, \lambda) = \mathbb{E}_X \text{tr}[(tX^T X + d\lambda I_d)^{-1}] \quad X_{ij} \sim N(0, 1).$$

$$= \mathbb{E}_{X, \bar{g}} [\langle \bar{g}, (tX^T X + d\lambda I_d)^{-1} \bar{g} \rangle] \quad \bar{g} \sim N(0, I_d).$$

$$\langle g, A^{-1}g \rangle = \sup_u [2\langle g, u \rangle - \langle u, Au \rangle]$$

$$= \mathbb{E}_{X, \bar{g}} \left[\sup_{u \in \mathbb{R}^d} (2\langle \bar{g}, u \rangle - t \|Xu\|_2^2 - d\lambda \|u\|_2^2) \right]$$

$$\|z\|_2^2 = \sup_v 2\langle v, z \rangle - \|z\|_2^2$$

$$= \mathbb{E}_{X, \bar{g}} \left[\sup_{u \in \mathbb{R}^d} \inf_{v \in \mathbb{R}^n} -2t \langle v, Xu \rangle + 2\langle \bar{g}, u \rangle + t \|v\|_2^2 - d\lambda \|u\|_2^2 \right]$$

$$\stackrel{\text{CGMT}}{\approx} \mathbb{E}_{\bar{g}, g, h} \left[\sup_{u \in \mathbb{R}^d} \inf_{v \in \mathbb{R}^n} -2t \|v\|_2 \langle u, g \rangle - 2t \frac{\|u\|_2}{\sqrt{d}} \langle v, h \rangle + 2\langle \bar{g}, u \rangle + t \frac{\|v\|_2^2}{\sqrt{d}} - d\lambda \|u\|_2^2 \right].$$

$$\|u\|_2 = O(1).$$

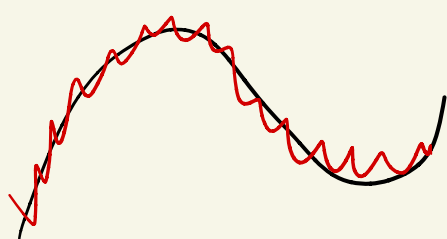
$$d \|u\|_2^2 = O(1)$$

$$= \mathbb{E}_{\bar{g}, g, h} \left[\sup_{\alpha \geq 0} \inf_{\beta \geq 0} \inf_{\|u\|_2 = \beta^2} \frac{\sup_{\|v\|_2 = \alpha^2} \langle v, Xu \rangle}{\beta^2} \right]$$

$$= \mathbb{E}_{\bar{g}, h} \left[\sup_{\alpha \geq 0} \inf_{\beta \geq 0} 2 \sqrt{1+t\beta^2} \times \alpha \frac{\|g\|_2}{\sqrt{d}} - 2t \alpha \beta \left(\frac{\|h\|_2}{\sqrt{d}} \right) + t \beta^2 - \lambda \alpha^2 \right]$$

$$\approx \sup_{\alpha \geq 0} \inf_{\beta \geq 0} \left[2 \sqrt{1+t\beta^2} \alpha - 2t \alpha \beta / \sqrt{r} + t \beta^2 - \lambda \alpha^2 \right] = S(t, \lambda).$$

$$\lim_{\lambda \rightarrow 0^+} \left[-\lambda^2 \partial_\lambda S(1, \lambda) \right] = \left(1 - \frac{1}{r}\right) \mathbb{1}_{\{r > 0\}}.$$



wavy is smooth

(B) why $\partial_\lambda \mathbb{E}[S(t, \lambda)] \rightarrow \partial_\lambda s(t, \lambda)$

Lemma: If $\lim_{d \rightarrow \infty} f_d(\lambda) = f(\lambda)$. $\lim_{d \rightarrow \infty} \sup_{\lambda \in \Lambda} |f_d''(\lambda)| < \infty$.

$$\Rightarrow \lim_{d \rightarrow \infty} f_d'(\lambda) = f'(\lambda).$$

Approach 3: The free energy approach.

Recall the setting of LASSO example.

$$y = Ax_0 + w, \quad w_i \sim \text{iid } N(0, \sigma^2)$$

$$A_{ij} \sim N(0, \frac{1}{n}), \quad x_{0,i} \sim \text{iid } P_0$$

$$\hat{x} = \arg \min_x \frac{1}{2d} \|y - Ax\|_2^2 + \frac{\lambda}{d} \sum_{i=1}^d I(x_i)$$

Interested in $\frac{1}{d} \sum_{i=1}^d \psi(\hat{x}_i, x_{0,i})$ as $d \rightarrow \infty$ and $n/d \rightarrow \delta$

$$f(h) \equiv \lim_{d \rightarrow \infty} \mathbb{E} \left[\min_x \left\{ \frac{1}{2d} \|y - Ax\|_2^2 + \frac{\lambda}{2d} \sum_{i=1}^d I(x_i) + h \frac{1}{d} \sum_{i=1}^d \psi(x_i, x_{0,i}) \right\} \right].$$

$$\Rightarrow f'(h) \doteq \frac{1}{d} \sum_{i=1}^d \psi(\hat{x}_i, x_{0,i}) = \mathbb{E}[\psi(\hat{X}, x_0)] \quad x_0 \sim P_0.$$

$$\hat{X} = \min_u \left[\frac{\beta}{2\tau} u^2 - \beta G u + \lambda I(u + x_0) \right] + x_0$$

τ_* , β_* solves

$$\begin{cases} \tau^2 = \sigma^2 + \delta^{-1} \mathbb{E} \left[\left(\eta(x_0 + \tau G; \frac{\lambda \tau}{\beta}) - x_0 \right)^2 \right] \\ \beta = \tau \left(1 - \delta^{-1} \mathbb{E} \left[\eta'(x_0 + \tau G; \frac{\lambda \tau}{\beta}) \right] \right) \end{cases}$$

$$\begin{aligned} \Gamma &= \frac{1}{2} x^2, \quad \Rightarrow \quad \eta(x; t) = \min_u \frac{1}{2} (u-x)^2 + \frac{t}{2} u^2 \\ &= \frac{x}{1+t} \quad \partial_x \eta(x; t) = \frac{1}{1+t} \end{aligned}$$

$$\hat{x} = \frac{x_0 + \tau G}{1 + \frac{\lambda \tau}{\beta}}$$

$$\begin{aligned} \tau^2 &= \sigma^2 + \delta^{-1} \tau^2 \\ \tau^2 &= \frac{\sigma^2}{1 - \delta^{-1}} \quad \delta > 1 \end{aligned}$$

$$\begin{cases} \tau^2 = \sigma^2 + \delta^{-1} \left[\overset{\lambda \rightarrow 0}{\lambda \tau / (\beta + \lambda \tau)} \right]^2 + \delta^{-1} \tau^2 \beta^2 / (\beta + \lambda \tau)^2 \\ \beta = \tau \left[1 - \beta \delta^{-1} / (\beta + \lambda \tau) \right] \end{cases}$$

$$\text{For } \delta > 1, \quad \lim_{\lambda \rightarrow 0^+} \beta(\lambda) = \tau_* (1 - \delta^{-1}), \quad \tau_* = \frac{\sigma^2}{1 - \delta^{-1}}$$

$$\text{For } \delta < 1, \quad \lim_{\lambda \rightarrow 0^+} \frac{\beta(\lambda)}{\lambda} = \frac{\delta \tau_*}{1 - \delta}$$

$$\text{where} \quad \tau_*^2 = \frac{\sigma^2}{1 - \delta} + \frac{1 - \delta}{\delta}$$

$$\Rightarrow \mathbb{E} \|\hat{x} - x_0\|_2^2 \rightarrow \delta (\tau_*^2 - \sigma^2) = \begin{cases} \frac{\sigma^2}{1 - \delta^{-1}}, & \delta > 1 \\ (1 - \delta) + \frac{\delta^2 \sigma^2}{1 - \delta}, & \delta < 1. \end{cases}$$

② How to apply CGMT to generalized linear models?

$$A_{ij} \sim N(0, \frac{1}{n}), \quad x_{0,i} \sim \mathbb{P}_0.$$

$$\mathbb{P}(y_i = 1 | a_i) = \sigma(\langle a_i, x \rangle).$$

$$\min_x \mathcal{L}(x) = \min_x \left\{ \frac{1}{n} \sum_{i=1}^n \ell(y_i, \langle a_i, x \rangle) + \frac{1}{d} \sum_{j=1}^d \Gamma(x_j) \right\}.$$

$$\ell^*(y, v) = \max_{t \in \mathbb{R}} [vt - \ell(y, t)] \quad \ell^* \text{ convex in } v$$

$$= \min_x \max_v \left\{ \frac{1}{n} \sum_{i=1}^n [\langle a_i, x \rangle v_i - \ell^*(y_i, v_i)] + \frac{1}{d} \sum_{j=1}^d \Gamma(x_j) \right\}$$

$$= \min_x \max_v \left\{ \langle v, Ax \rangle / n - \frac{1}{n} \sum_{i=1}^n \ell^*(y_i, v_i) + \frac{1}{d} \sum_{j=1}^d \Gamma(x_j) \right\}.$$

$$y_i = f(\langle a_i, x_0 \rangle; Z_i) \quad Z_i \sim \text{i.i.d. } \mathbb{P}_Z.$$

y_i depend on A !

Denote $P_0 = \frac{x_0 x_0^T}{\|x_0\|_2^2}$ $P_0^\perp = I_d - P_0$.

$$\zeta_i = \frac{\langle x_0, a_i \rangle}{\|x_0\|_2} \sim N(0, \frac{1}{n}). \quad y_i = f(\langle a_i, x_0 \rangle; z_i) = f(\zeta_i \|x_0\|_2; z_i)$$

$$b_i = P_0 a_i = \zeta_i \frac{x_0}{\|x_0\|_2} \quad c_i = P_0^\perp a_i = a_i - \zeta_i x_0$$

$$b_i \text{ are independent of } c_i, \quad B = \begin{bmatrix} b_1^T \\ \vdots \\ b_n^T \end{bmatrix}, \quad C = \begin{bmatrix} c_1^T \\ \vdots \\ c_n^T \end{bmatrix}, \quad \zeta = \begin{bmatrix} \zeta_1 \\ \vdots \\ \zeta_n \end{bmatrix}$$

$$A = A P_0 + A P_0^\perp = B + C \quad C \text{ is independent of } (B, y).$$

$$= \min_x \max_v \left\{ \langle v, A P_0^\perp x \rangle + \langle v, A P_0 x \rangle - \frac{1}{n} \sum_{i=1}^n \ell^*(f(\zeta_i \|x_0\|_2; z_i), v_i) + \frac{1}{2} \sum_{j=1}^d \mathbb{I}(x_j) \right\}.$$

$$= \min_t \min_{x: \langle x_0, x \rangle = t} \max_v \left\{ \langle v, C P_0^\perp x \rangle + \langle v, \zeta \rangle \langle x, x_0 \rangle / \|x_0\|_2 - \frac{1}{n} \sum_{i=1}^n \ell^*(f(\zeta_i \|x_0\|_2; z_i), v_i) + \frac{1}{2} \sum_{j=1}^d \mathbb{I}(x_j) \right\}.$$

For any fixed $\|x_0\|_2, \zeta, z_i$, Gaussian process where randomness from C .

$$\begin{aligned} \approx \min_t \min_{x: \langle x_0, x \rangle = t} \max_v \left\{ \|P_0^\perp x\|_2 < g, v > / \sqrt{n} + \|v\|_2 < h, P_0^\perp x > / \sqrt{n} \right. \\ \left. + \langle v, \zeta \rangle \langle x, x_0 \rangle / \|x_0\|_2 - \frac{1}{n} \sum_{i=1}^n \ell^*(f(\zeta_i \|x_0\|_2; z_i), v_i) + \frac{1}{2} \sum_{j=1}^d \mathbb{I}(x_j) \right\}. \end{aligned}$$