Mean Field Asymptotics in Statistical Learning.

Mar 17.

Lecture 16. Double descent and generalized linear models.

① Double descent in linear models.

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

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

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

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

Linear ridgeless regression:

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

$$\hat{\beta}_0 = \lim_{\lambda \to 0+} \hat{\beta}_\lambda = X^{+} y = \begin{cases} \text{OLS solution} \\ \text{Min norm interpolating} \end{cases} \quad \text{sol.}$$

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

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

Bias-Variance decomposition:

$$\mathbb{E}\|\hat{\beta}_\lambda - \beta\|_2^2 = \mathbb{E}\|((X^T X + d\lambda I)^{-1} X^T X - I_d)\beta + (X^T X + d\lambda I)^{-1} X w\|_2^2$$

$$= \underbrace{\mathbb{E}[\|((X^T X + d\lambda I)^{-1} X^T X - I_d)\beta\|_2^2]}_{B(\lambda)} + \underbrace{\mathbb{E}[\|(X^T X + d\lambda I)^{-1} X^T w\|_2^2]}_{V(\lambda)}.$$

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

[Dobriban, Wager, 2015].

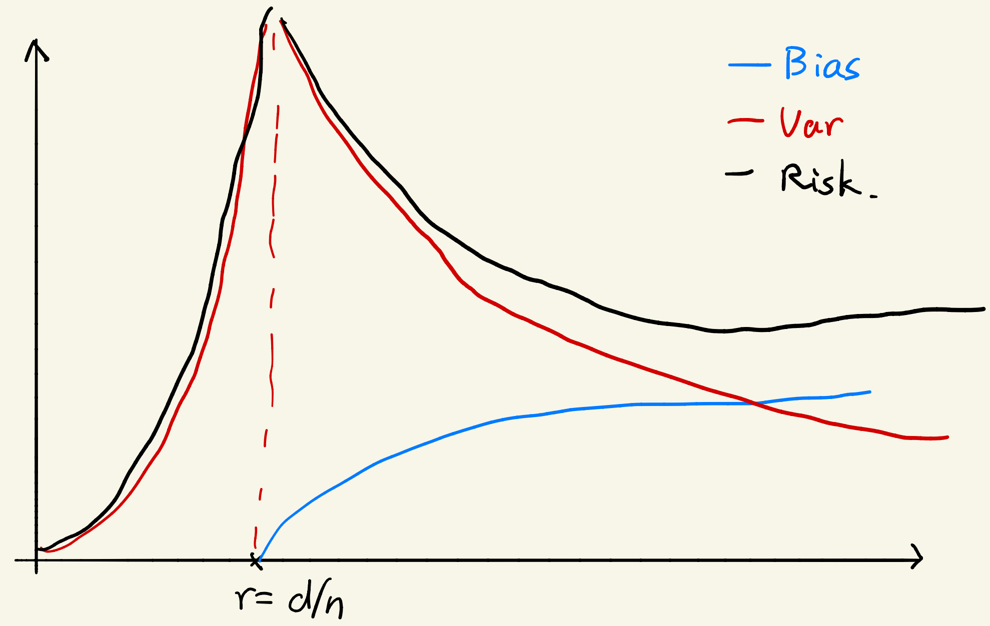
Under the assumptions above, as $n, d \to \infty$ and $d/n \to r$, we have

$$B(0) \longrightarrow (1 - \tfrac{1}{r})\, \mathbb{1}\{r > 1\}$$

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

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

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



Double descent curve.

Appoach 1: Spectrum of the Wishart matrix.

$$B(\lambda) = \mathbb{E}_{\beta, X} \| ((X^TX + d\lambda I)^{-1} X^TX - I)\beta\|_2^2 = \lambda^2 d\, \mathbb{E}_X \mathrm{tr}\left[(X^TX + \lambda I)^{-2} \right].$$

$$= \lambda^2 \times \mathbb{E}\left[\int_{(0,\infty)} \frac{1}{(s+\lambda)^2} \hat{\mu}(ds) \right]$$

$\hat{\mu}$: Empirical distribution of eigenvalues.

where

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

$$V(\lambda) = \mathbb{E}_{w, X}\left[\| (X^TX + d\lambda I)^{-1} X^T w \|_2^2 \right] = \sigma^2 \cdot \mathbb{E}_X\left[\mathrm{tr}((X^TX + d\lambda I_d)^{-2} X^TX) \right]$$

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

Thm: (Marchenko – Pastur law)

Assume $d, n \to \infty$, $d/n \to r \in (0,1)$.

Then for sufficient nice function ψ,

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

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

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

Nice functions $\psi_1(s) = \frac{s}{(s+\lambda)^2}$, $\psi_2(s) = \frac{1}{(s+\lambda)^2}$.

Approach 2: Stieltjes transform of Wishart matrix.

Lemma: Define $S(t,\lambda) = \mathrm{tr}\left[(tX^TX + d\cdot\lambda I_d)^{-1}\right]$,

then $\partial_t S(t,\lambda) = -\mathrm{tr}\left[(X^TX + d\lambda I_d)^{-2}(X^TX)\right]$.

$\partial_\lambda S(t,\lambda) = -\mathrm{tr}\left[(tX^TX + d\lambda I_d)^{-2}\right]\times d$.

So that $B(\lambda) = -\lambda^2 \times \partial_\lambda S(1,\lambda)$

$V(\lambda) = -\sigma^2 \times \partial_t S(1,\lambda)$

(A) Calculate the asymptotics of $S(t,\lambda)$

(B) Show that $\partial_\lambda S(t,\lambda)$ and $\partial_t S(t,\lambda)$ converges to ∂_λ, ∂_t of asymptotics. (Only $\lambda > 0$).

(A) Apply CGMT.

$$\mathbb{E}_X S(t,\lambda) = \mathbb{E}_X \mathrm{tr}\left[(tX^TX + d\cdot\lambda\cdot I_d)^{-1}\right] \qquad X_{ij} \sim N(0,1).$$

$$\boxed{① : \quad \bar g \sim N(0, I_d)}$$

$$\overset{①}{=} \mathbb{E}_{\bar g X}\left[\langle \bar g, (tX^TX + d\cdot\lambda\cdot I_d)^{-1}\bar g\rangle\right]$$

$$\boxed{② : \quad \langle \bar g, A^{-1}\bar g\rangle = \sup_{u\in\mathbb{R}^d}\left(2\langle g, u\rangle - \langle u, Au\rangle\right).}$$

$$\overset{②}{=} \mathbb{E}_{\bar g X}\left[\sup_{u\in\mathbb{R}^d}\left(2\langle \bar g, u\rangle - \langle u, (tX^TX + d\lambda I_d)u\rangle\right)\right]$$

$$= \mathbb{E}_{\bar g X}\left[\sup_{u\in\mathbb{R}^d}\left(2\langle \bar g, u\rangle - t\|Xu\|_2^2 - d\lambda\|u\|_2^2\right)\right]. \qquad \|u\|_2 = O\left(\tfrac{1}{\sqrt d}\right)$$

$$\boxed{③ : \quad \|a\|_2^2 = \sup_{b\in\mathbb{R}^n} 2\langle a, b\rangle - \|b\|_2^2}$$

$$\overset{③}{=} \mathbb{E}_{\bar g X}\left[\sup_{u\in\mathbb{R}^d}\inf_{v\in\mathbb{R}^n} \underbrace{2\langle \bar g, u\rangle - 2t\langle v, Xu\rangle + t\|v\|_2^2 - d\lambda\|u\|_2^2}_{\text{Concave - Convex}}\right]$$

$$\bar g \sim N(0, I_d), \quad g \sim N(0, I_d), \quad h \sim N(0, I_n)$$

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

$$\boxed{④ : \quad \inf_{v\in\mathbb{R}^n} f(v) = \inf_{\beta\ge 0}\inf_{\|v\|_2 = \beta} f(v)}$$

$$\overset{④}{=} \mathbb{E}_{g,h}\left[\sup_{u\in\mathbb{R}^d}\inf_{\beta\ge 0} 2\sqrt{1+s^2\beta^2}\langle u, g\rangle - 2t\|u\|_2\beta\|h\|_2 + t\beta^2 - d\lambda\|u\|_2^2\right]$$

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

$$\approx \sup_{\alpha \geq 0} \inf_{\beta \geq 0} \Big(\big(2\sqrt{1+t^2\beta^2} - 2t\beta/\sqrt{r}\big)\alpha + t\beta^2 - \lambda\alpha^2 \Big)$$

$$= \inf_{\beta \geq 0} \left[\frac{\big(\sqrt{1+t^2\beta^2} - t\beta/\sqrt{r}\big)^2}{\lambda} + t\beta^2 \right] \equiv s(t, \lambda).$$

Calculus problem: $\lim_{\lambda \to 0_+} \big[-\lambda^2 \cdot \partial_\lambda s(1, \lambda) \big] = \big(1 - \frac{1}{r}\big) \mathbb{1}\{r > 0\}.$

$$\lim_{\lambda \to 0+} \big[\sigma^2 \partial_t s(1, \lambda) \big] = \begin{cases} \sigma^2 \dfrac{r}{1-r} & r < 1, \\ \sigma^2 \dfrac{1}{r-1} & r > 1. \end{cases}$$

(B) If $\lim_{d \to \infty} f_d(\lambda) = f(\lambda)$, how to show $\lim_{n \to \infty} f_d'(\lambda) = f(\lambda)$.

Need: $\lim_{d \to \infty} \sup_{\lambda \in \Lambda} |f_d''(\lambda)| < \infty$.

Approch 3: The free energy approach.

Recall the setting of LASSO example.

$y = A x_0 + w.$ $\quad w_i \sim_{iid} N(0, \sigma^2)$

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

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

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

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

$$\Rightarrow \quad f'(h) \doteq \frac{1}{d} \sum_{i=1}^d \psi(\hat{x}_i, x_{0,i}) = \mathbb{E}[\psi(\hat{X}, X_0)]$$

$$\hat{X} = \min_u \Big[\frac{\beta}{2\tau} u^2 - \beta G u + \lambda I(u + X_0) \Big] + X_0$$

τ_*, β_* solves

$$\begin{cases} \tau^2 = \sigma^2 + \delta^{-1} \mathbb{E}\big[\big(\eta(X_0 + \tau G; \frac{\lambda\tau}{\beta}) - X_0 \big)^2 \big]. \\ \beta = \tau \big(1 - \delta^{-1} \mathbb{E}\big[\eta'(X_0 + \tau G; \frac{\lambda\tau}{\beta}) \big] \big). \end{cases}$$

$$I = \frac{1}{2}x^2, \quad \Rightarrow \quad \eta(x;t) = \min_u \frac{1}{2}(u-x)^2 + \frac{t}{2}u^2$$
$$= \frac{x}{1+t} \qquad \partial_x \eta(x;t) = \frac{1}{1+t}$$

$$\hat{X} = \frac{X_0 + \tau G}{1 + \frac{\lambda\tau}{\beta}}$$

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

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

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

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

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

② How to apply CGMT to generalized linear models?

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

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

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

$\boxed{\ell^*(y,v) = \max_{t\in\mathbb{R}}\left[vt - \ell(y,t)\right]}$ ℓ^* convex in v

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

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

$$\boxed{\begin{array}{l} y_i = f(\langle a_i, x_0\rangle; z_i) \qquad z_i \sim_{\text{i.i.d.}} \mathbb{P}_z. \\ y_i \text{ depend on } A\,! \end{array}}$$

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

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

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

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

$A = AP_0 + AP_0^\perp = B + C$ $\quad C$ is independent of (B, y).

$$= \min_x \max_v \Big\{ \langle v, AP_0^\perp x \rangle + \langle v, AP_0 x \rangle - \frac{1}{n}\sum_{i=1}^n \ell^*(f(\zeta_i \|x_0\|_2 ; z_i), v_i) + \frac{1}{d}\sum_{j=1}^d \Gamma(x_j) \Big\}.$$

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

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

$$\approx \min_t \min_{x : \langle x_0, x \rangle = t} \max_v \Big\{ \|P_0^\perp x\|_2 \langle g, v \rangle / \sqrt{n} + \|v\|_2 \langle h, P_0^\perp x \rangle / \sqrt{n} + \langle v, \zeta \rangle \langle x, x_0 \rangle / \|x_0\|_2 - \frac{1}{n}\sum_{i=1}^n \ell^*(f(\zeta_i \|x_0\|_2 ; z_i), v_i) + \frac{1}{d}\sum_{j=1}^d \Gamma(x_j) \Big\}.$$