STAT260 Mean Field Asymptotics in Statistical Learning Lecture 16 - 03/17/2021

Lecture 16: Double Descent and Generalized Linear Models

Lecturer: Song Mei *Scriber: Kumar Krishna Agrawal* *Proof Reader: Zitong Yang*

1 Double Descent in Linear Models

We study the phenomenon of double descent in linear models. Consider the following setup

$$\begin{aligned}\textbf{Training Dataset}&: (x_i, y_i)_{i\in[n]} \subseteq \mathbb{R}^d \times R\\ \textbf{Linear model}&: y_i = \langle x_i, \beta_0\rangle + w_i\\ &\text{where,}\quad x_i \sim_{iid} N(0, I_d) \quad w_i \sim_{iid} N(0, \sigma^2 \quad \beta_0 \sim N\Big(0, \frac{1}{d} I_d\Big)\end{aligned}$$

Under the linear ridgeless regression formulation we recover

$$\begin{aligned}\hat{\beta}_\lambda &= \arg\min_\beta \frac{1}{2n}\|Y - X\beta\|_2^2 + \frac{\lambda d}{2n}\|\beta\|_2^2 = (X^T X + d\lambda I)X^T Y\\ \therefore \hat{\beta}_0 &= \lim_{\lambda\to 0_+} \hat{\beta}_\lambda = X^\dagger Y\end{aligned}$$

This is the ordinary least squares, minimum norm interpolating solution. Given the above solution, we can compute the test error as

$$\begin{aligned}\textbf{Test error: } \mathbb{E}_X\Big[(\langle X, \hat{\beta}_\lambda\rangle - \langle X, \beta_0\rangle)^2\Big] &= \|\hat{\beta}_\lambda - \beta_0\|_2^2\\ R(\lambda, \gamma, \sigma^2) &\equiv \lim_{\substack{d\to\infty\\ d/n\to\gamma}} \mathbb{E}_{\beta_0, X, w}[\|\hat{\beta_\lambda} - \beta_)\|_2^2]\end{aligned}$$

Consider the Bias-Variance decomposition of the above error term

$$\begin{aligned}\mathbb{E}[\|\hat{\beta_\lambda} - \beta_0\|_2^2] &= \mathbb{E}\Big[\|((X^T X + d\lambda I)^{-1}X^T X - I_d)\beta + (X^T X + d\lambda I)^{-1} X w\|_2^2\Big]\\ &= \underbrace{\mathbb{E}\Big[\|((X^T X + d\lambda I)^{-1}X^T X - I_d)\beta\Big]}_{Bias(\lambda)\|_2^2} + \underbrace{\mathbb{E}\Big[\|(X^T X + d\lambda I)^{-1} X^T w\|_2^2\Big]}_{Variance(\lambda)}\end{aligned}$$

Theorem 1 (Hastie, Montanari, Rosset, Tibshirani, 2020. Dobriban, Wager, 2015)**.** *Under the assumptions above, as $n, d \to \infty$ and $d/n \to \gamma$, we have*

$$\begin{aligned}B(0) &\to \Big(1 - \frac{1}{\gamma}\Big)1\{\gamma > 1\}\\ V(0) &\to \sigma^2\Big[\frac{\gamma}{1-\gamma}1\{\gamma < 1\} + \frac{1}{1-\gamma}1\{\gamma > 1\}\Big]\\ \therefore\quad R(\gamma, \sigma^2) &= \begin{cases}\sigma^2\frac{\gamma}{1-\gamma} & \gamma < 1\\ \Big(1 - \frac{1}{\gamma}) + \sigma^2\frac{1}{\gamma - 1} & \gamma > 1\end{cases}\end{aligned}$$

In the rest of this lecture, we explore different approaches to recover the above solution.

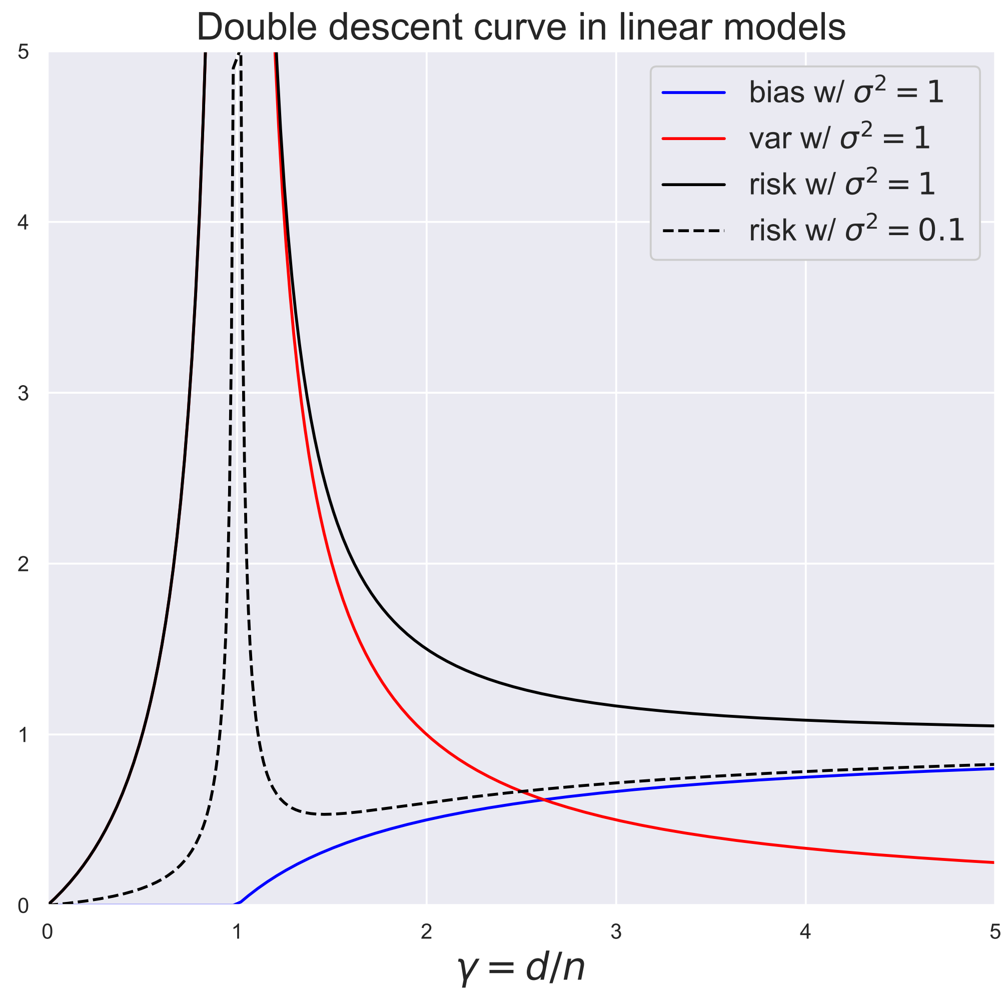

Figure 1: Double descent curve in linear models with bias-variance decomposition

1.1 Approach 1 : Spectrum of the Wishart matrix

The bias-variance decomposition terms are expressible in terms of the eigenspectrum of certain matrices.

$$\begin{aligned} Bias(\lambda) &= \mathbb{E}_{\beta,X} \| \Big((X^T X + d\lambda I)^{-1} X^T X - I \Big) \beta \|_2^2 = \lambda^2 d \mathbb{E}_X tr \Big[(X^T X + \lambda I)^{-2} \Big] \\ &= \lambda^2 \times \mathbb{E} \Big[\int_{[0,\infty)} \frac{1}{(s+\lambda)^2} \hat{\mu}(ds) \Big] \end{aligned}$$

where $\hat{\mu}$ is the empirical distribution of eigenvalues, such that $\hat{\mu}(ds) = \frac{1}{d} \sum_{j=1}^{d} \delta_{\lambda_j (X^T X/d)}$

Similarly, the variance can be simplified as

$$\begin{aligned} Var(\lambda) &= \mathbb{E}_{w,X} \Big[\| X^T X + d\lambda I)^{-1} X^T w \|_2^2 \Big] = \sigma^2 \mathbb{E}_X \Big[tr((X^T X + d\lambda I_d)^{-2} X^T X) \Big] \\ &= \sigma^2 \mathbb{E} \int_{(0,\infty)} \frac{s}{(s+\lambda)^2} \hat{\mu}(ds) \end{aligned}$$

Theorem 2 (Marchenko-Pastur law)**.** *Assume $d,\ n \to \infty, d/n \to \gamma \in (0,1)$. For sufficient nice funciton ψ*

$$\frac{1}{d}\sum_{j=1}^{d} \psi(\lambda_j(X^T X/d)) \underset{d\to\infty}{\overset{a.s}{\to}} \int \psi(s)\mu(ds)$$

$$\mu(ds) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+ - s)(x-\lambda_-)}}{\gamma x} 1(s \in [\lambda_-, \lambda_+])ds$$
$$\textit{where}\ \ \lambda_\pm = \sigma^2(1 \pm \sqrt{\lambda})^2$$

In our context, examples of nice functions include $\psi_{Bias}(s) = \frac{s}{(s+\lambda)^2} \quad \psi_{Var}(s) = \frac{1}{(s+\lambda)^2}$

1.2 Approach 2: Stiejles transform of Wishart matrix

Lemma 3. *Define $S(t,\lambda) = tr\Big[(tX^T X + d\lambda I_d)^{-1}\Big]$, then*

$$\partial_t S(t,\lambda) = -tr\Big[(X^T X + d\lambda I_d)^{-2}(X^T X)\Big]$$
$$\partial_\lambda S(t,\lambda) = -tr\Big[(tX^T X + d\lambda I_d)^{-2}\Big] \times d$$
$$\textit{So that}\ \ Bias(\lambda) = -\lambda^2 \times \partial_\lambda S(1,\lambda)$$
$$Var(\lambda) = -\sigma^2 \times \partial_t S(1,\lambda)$$

This gives us the following two steps to get asymptotics of bias-variance

(A) Calculate the asymptotics of $S(t,\lambda)$
(B) Show that $\partial_\lambda S(t,\lambda)$ and $\partial_t S(t,\lambda)$ converges to $\partial_\lambda, \partial_t$ of asymptotics

- **(A)** Apply CGMT

$$\mathbb{E}_X S(t,\lambda) = \mathbb{E}_X tr[(tX^T X + d\lambda I_d)^{-1}] \qquad \text{(note } X_{ij} \sim N(0,1))$$
$$= \mathbb{E}_{\bar{g}X}\Big[\langle \bar{g}, (tX^T X + d\lambda I_d)^{-1}\bar{g}\rangle\Big] \qquad \text{(where } \bar{g} = N(0, I_d))$$

Using the variational form $\langle \bar{g} A^{-1}\bar{g}\rangle = \sup_{u\in\mathbb{R}^d}(2\langle g, u\rangle - \langle u, Au\rangle)$

$$\mathbb{E}_X S(t,\lambda) = \mathbb{E}_{\bar{g}X}\Big[\sup_{u\in\mathbb{R}^d}\Big(2\langle \bar{g}, u\rangle - \langle u, (tX^T X + d\lambda I_d)u\rangle\Big)\Big]$$
$$= \mathbb{E}_{\bar{g}X}\Big[\sup_{u\in\mathbb{R}^d}\Big(2\langle \bar{g}, u\rangle - t\|Xu\|_2^2 - d\lambda\|u\|_2^2\Big)\Big]$$

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

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

Introducing random variables $\bar{g} \sim N(0, I_d), g \sim N(0, I_d), h \sim N(0, I_n)$ we apply CGMT

$$\mathbb{E}_X S(t,\lambda) = \mathbb{E}_{\bar{g},g,h}\Big[\sup_{u\in\mathbb{R}^d}\inf_{v\in\mathbb{R}^n} 2\langle \bar{g}, u\rangle - 2t\|v\|\langle u, g\rangle - 2t\|u\|_2\langle v, h\rangle + s\|v\|_2^2 - d\lambda\|u\|_2^2\Big]$$

Using the simplification $\inf_{v\in\mathbb{R}^n} f(v) = \inf_{\beta\geq 0}\inf_{\|v\|_2=\beta} f(v)$

$$\begin{aligned}
\mathbb{E}_X S(t,\lambda) &= \mathbb{E}_{g,h}\Big[\sup_{u\in\mathbb{R}^d}\inf_{\beta\geq 0} 2\sqrt{1+s^2\beta^2}\langle u,g\rangle - 2t\|u\|_2\beta\|h\|_2 + t\beta^2 - d\lambda\|u\|_2^2\Big] \\
&= \mathbb{E}_{g,h}\Big[\sup_{\alpha\geq 0}\inf_{\beta\geq 0} 2\sqrt{1+t^2\beta^2}\times\alpha\frac{\|g\|_2}{d} - 2t\alpha\beta\frac{\|h\|_2}{\sqrt{d}} + t\beta^2 - \lambda\alpha^2\Big] \\
&\approx \sup_{\alpha\geq 0}\inf_{\beta\geq 0}(2\sqrt{1+t^2\beta^2} - 2t\beta/\sqrt{\gamma})\alpha + t\beta^2 - \lambda\alpha^2 \\
&= \inf_{\beta\geq 0}\Big[\frac{(\sqrt{1+t^2\beta^2} - t\beta/\sqrt{\gamma})^2}{\lambda} + t\beta^2\Big] \equiv S(t,\lambda)
\end{aligned}$$

Calculus problem:

$$\lim_{\lambda\to 0^+}\Big[-\lambda^2\partial_\lambda s(1,\lambda)\Big] = \Big(1-\frac{1}{\gamma}\Big)1\{\gamma>0\}$$

$$\lim_{\lambda\to 0^+}\sigma^2\partial_t S(1,\lambda) = \begin{cases}\sigma^2\frac{\gamma}{1-\gamma} & \gamma<1\\ \sigma^2\frac{1}{\gamma-1} & \gamma>1\end{cases}$$

- **(B)** Here we want to consider why $\partial_\lambda\mathbb{E}[S(t,\lambda)]\to\partial_\lambda S(t,\lambda)$

  Lemma 4. *If* $\lim_{d\to\infty} f_d(\lambda) = f(\lambda)$ *and* $\lim_{d\to\infty}\sup_{\lambda\in\Lambda}|f_d''(\lambda)| < \infty$, *then*

$$\lim_{d\to\infty} f_d'(\lambda) = f(\lambda)$$

1.3 Approach 3: The free energy approach

Recall the setting of LASSO example

$$Y = Ax_0 + w \qquad W_i \sim_{iid} N(0,\sigma^2)$$
$$A_{ij}\sim N\Big(0,\frac{1}{n}\Big) \qquad X_{0,i}\sim_{iid}\mathbb{P}_0$$
$$\hat{x} = \arg\min_x \frac{1}{2d}\|y-Ax\|_2^2 + \frac{\lambda}{d}\sum_{i=1}^d\Gamma(x_i)$$

In this setup, we are interested in $\frac{1}{d}\sum_{i=1}^d\psi(\hat{x}_i, X_{0,i})$ as $d\to\infty, n/d\to\delta$

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

$$\implies f'(h) = \frac{1}{d}\sum_{i=1}^d\psi(\hat{x}_i,x_{o,i}) = \mathbb{E}[\psi(\hat{x},x_0)]$$

$$\hat{X} = \min_u\Big[\frac{\beta}{2\tau}u^2 - \beta G u + \lambda\Gamma(u+x_0)\Big] + X_0$$

where (τ_*,β_*) solves
$$\tau^2 = \sigma^2 + \delta^{-1}\mathbb{E}\Big[(\eta(x_0+\tau G;\frac{\lambda\tau}{\beta}) - x_0)^2\Big]$$
$$\beta = \tau\Big(1-\delta^{-1}\mathbb{E}\Big[\eta'(x_0+\tau G;\frac{\lambda\tau}{\beta})\Big]\Big)$$

Remark 5. *Consider taking* $\Gamma = \frac{1}{2}x^2$

$$
\begin{aligned}
\implies \eta(x;t) &= \min_u \frac{1}{2}(u-x)^2 + \frac{t}{2}u^2 \\
&= \frac{x}{1+t} \qquad\qquad \textit{also} \quad \partial_x \eta(x;t) = \frac{1}{1+t} \\
\therefore \hat{x} &= \frac{x_0 + \tau G}{1 + \frac{\lambda\tau}{\beta}} \qquad \tau^2 = \frac{\sigma^2}{1-\delta^{-1}} \quad \textit{for}(\delta > 1)
\end{aligned}
$$

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

$$
\begin{aligned}
&\textit{For} \qquad \delta > 1 \lim_{\lambda\to 0^+} \beta(\lambda) = \tau_*(1-\delta^{-1}) \quad \tau_* = \frac{\sigma^2}{1-\delta^{-1}} \\
&\textit{For} \qquad \delta < 1 \lim_{\lambda\to 0^+} \frac{\beta(\lambda)}{\lambda} = \frac{\delta\tau_*}{1-\delta}
\end{aligned}
$$

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

$$
\therefore \mathbb{E}\|\hat{x} - x_0\|_2^2 \to \delta(\tau_*^2 0 \delta^2) = \begin{cases} \frac{\sigma^2}{1-\delta^{-1}} & \delta > 1 \\ (1-\delta) + \frac{\delta^2\sigma^2}{1-\delta} & \delta < 1 \end{cases}
$$

Question: How to apply CGMT to generalized linear models?
Recall the problem setup, where

$$
\begin{aligned}
A_{ij} \sim N\Big(0, \frac{1}{n}\Big) \quad X_{0,i} \sim \mathbb{P}_0 \\
\mathbb{P}(Y_i = 1 | a_i) = \sigma(\langle a_i, x\rangle)
\end{aligned}
$$

Defining the loss function,

$$
\min_X \mathcal{L}(x) = \min_x \Big\{ \frac{1}{n}\sum_{i=1}^n l(y_i, \langle a_i, x\rangle) + \frac{1}{d}\sum_{j=1}^d \Gamma(x_j) \Big\}
$$

$$
\text{Define} \quad L^*(y,v) = \max_{t\in\mathbb{R}} vt - l(y,t) \qquad \text{(which is convex in v)}
$$

$$
\begin{aligned}
\therefore \min_x \mathcal{L}(x) &= \min_x \max_v \Big\{ \frac{1}{n}\sum_{i=1}^n \Big[\langle a_i, x\rangle v_i - L^*(y_i, v_i)\Big] + \frac{1}{d}\sum_{j=1}^d \Gamma(x_j) \Big\} \\
&= \min_x \max_v \Big\{ \frac{\langle v, Ax\rangle}{n} - \frac{1}{n}\sum_{i=1}^n L^*(y_i, v_i) + \frac{1}{d}\sum_{j=1}^d \Gamma(x_j) \Big\}
\end{aligned}
$$

$$
\text{Define} \quad y_i = f(\langle a_i, x_0\rangle; z_i) \quad \text{where } z_i \sim_{iid} \mathbb{P}_z
$$

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

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

$$b_i = P_0 a_i = \xi_i \frac{x_0}{\|x_0\|_2} \qquad x_i = P_0^{\perp} a_i = a_i - \xi_i x_0$$

$$\textbf{note} : b_i \text{ are indpendent of } c_i \qquad B = \begin{bmatrix} b_1^T \\ \vdots \\ b_n^T \end{bmatrix}, C = \begin{bmatrix} c_1^T \\ \vdots \\ c_n^T \end{bmatrix}, \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}$$

$$A = AP_0 + AP_0^{\perp} = B + C \qquad \text{C is independent of (B, Y)}$$

Using these, we can simplify the equation as

$$\therefore \min_x \mathcal{L}(x) = \min_x \max_v \Big\{ \langle v, AP_0^{\perp} x \rangle + \langle v, AP_o x \rangle - \frac{1}{n} \sum_{i=1}^{n} + \frac{1}{d} \sum_{j=1}^{d} \Gamma(x_j) \Big\}$$

$$= \min_t \min_{x:\langle x_0, x \rangle = t} \max_v \Big\{ \langle v, CP_0^{\perp} x \rangle + \langle v, \xi \rangle \langle x, x_0 \rangle / \|x_0\|_2$$

$$- \frac{1}{n} \sum_{i=1}^{n} l^*(f(\xi_i \|x_0\|_i; z_i), v_i) + \frac{1}{d} \sum_{j=1}^{d} \Gamma(x_j) \Big\}$$

For any fixed $\|x_0\|_2, \xi, z_i$, this is a Gaussian process where C is the source of randomness.

$$\therefore \min_x \mathcal{L}(x) \approx \min_t \min_{x:\langle x_0, x \rangle = t} \max_v \Big\{ \|P_0^{\perp} x|_2 \langle g, v \rangle / \sqrt{n} + \|v\|_2 \langle h, P_0^{\perp} x \rangle / \sqrt{n}$$

$$+ \langle v, \xi \rangle \langle x, x_0 \rangle / \|x_0\|_2 - \frac{1}{n} \sum_{i=1}^{n} l^*(f(\xi_i \|x_0\|_i; z_i), v_i) + \frac{1}{d} \sum_{j=1}^{d} \Gamma(x_j) \Big\}$$