- BEHAVIOR IN HIGH-DIMENSIONS: $d = d(n)$ s.t. $\frac{d}{n} \to \gamma$

  The fixed dimensional case: $\gamma = 0$. Also if $d = o(n)$ we
  are in the $\gamma = 0$ scenario. Today we will study the
  case of $\gamma > 0$ and even $\gamma \geq 1$.

- Let's assume a linear model with random covariates:

$$Y_i = \underline{\Phi}_i^\top \beta^* + \varepsilon_i \qquad i = 1, \dots, n$$

$$Y_i, \underline{\Phi}_i \quad \text{are} \quad \text{iid from some } \mathbb{P}_{Y, \underline{\Phi}}$$

$$\beta^* \in \mathbb{R}^d \qquad\qquad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- We are interested in the excess risk of $\hat{\beta}$ (the OLS)

$$R(\hat{\beta}) = \|\hat{\beta} - \beta^*\|_{\Sigma}^2 \qquad \Sigma = \mathbb{E}\left[\underline{\Phi}_i \underline{\Phi}_i^\top\right] \qquad \text{\textcircled{1}}$$

excess
risk

- We saw that an upper bound for this risk is
$$\frac{\sigma^2}{n} \, \mathrm{tr}\left(\hat{\Sigma}^{-1} \Sigma \right) \qquad \hat{\Sigma} = \frac{1}{d} \sum_{i=1}^{n} \Phi_i \Phi_i^{\top}$$

- Mourtada's paper. Theorem 1: this is also the value of the minimax risk.

  Corollary 2: if $n > d+1$ and the distribution of the Φ_i's is not degenerate the the minimax risk is lower bounded by
  $$\sigma^2 \, \frac{d}{n-d-1} \;\approx\; \boxed{\sigma^2 \, \frac{\delta}{1-\delta}}$$

  for large n and d.

- Also Mourtade shows that if $d > n$ the minimax value for the excess risk is infinity!

- Belkin, Hsu and Xu (2020) SIAM Journal of Mathematics & Data Science

  An exact analysis of the excess risk when the Φ_i's are iid $N(0, I_d)$. In this case the excess risk is
  $$\frac{\sigma^2}{n} \, \mathbb{E}\left[\mathrm{tr}\left(\hat{\Sigma}^{-1} \underbrace{\Sigma}_{I_d} \right) \right] = \sigma^2 \, \mathbb{E}\left[\left(\sum_{i=1}^{n} \Phi_i \Phi_i^{\top} \right)^{-1} \right]$$

  ②

$\downarrow$

Wishart distribution

with parameters $I_d$ol

aund n degrees of freedom

$$= \begin{cases} \dfrac{\sigma^2 \, tr(I_d)}{n-d-1} = \sigma^2 \dfrac{d}{n-d-1} & n > d+1 \\ \\ \infty & n \le d+1 \end{cases}$$

$\llcorner$ at interpolation, i.e. $d \le n$, this is infinity!

- Assume $d > n$, we cannot use OLS $\hat{\beta}$. We use the min-norm estimator

$$\hat{\beta}_{MN} = \Phi^{+} Y = \Phi^{T} (\underbrace{\Phi \Phi^{T}}_{\substack{n \times n \\ \text{invertible with prob. 1}}})^{-1} Y$$

$\underset{\substack{n \times d \text{ matrix} \\ \text{whose } i^{th} \text{ row} \\ \text{is } \Phi_n^{T}}}{\nearrow}$ $\underset{\substack{n\text{-dim} \\ \text{vector of} \\ y_i's}}{\searrow}$

- Let's compute $\mathbb{E}\left[\| \hat{\beta}_{MN} - \beta^* \|^2\right]$.

The first thing to notice is that there exists a bias.

$$\beta^* - \hat{\beta}_{MN} = \beta^* - \Phi^{T}(\Phi \Phi^{T})^{-1} \Phi \beta^* - \Phi^{T}(\Phi \Phi^{T})^{-1} \underset{\llcorner}{\varepsilon} \quad \varepsilon = \begin{bmatrix} c_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \underbrace{\left(I_d - \Phi^{T}(\Phi \Phi^{T})^{-1}\Phi\right)}_{\pi} \beta^* - \Phi^{T}(\Phi \Phi^{T})^{-1}\varepsilon$$

$\underbrace{\qquad\qquad}_{}$

$I_d - \pi$ orthogonal projection
onto the null space of Φ

(3)

$$- \quad \mathbb{E}\left[\| \beta^* - \hat{\beta}_{MN} \|^2\right] = \mathbb{E}_{\Phi}\left[\|(\mathbb{I} - \pi)\beta^*\|^2\right] +$$

$$\mathbb{E}_{\Phi}\left[\mathrm{tr}\left((\Phi\Phi^T)^{-1}\Phi\cancel{\Phi^T(\Phi\Phi^2)^{-1}}\mathbb{E}[\varepsilon\varepsilon^T]\right)\right]$$

$$\overline{}\;\sigma^2\;\mathbb{E}\left[(\Phi\Phi^2)^{-1}\right]$$

$$= \quad T_1 + T_2$$

Next $\quad T_1 = \|\overset{*}{\beta}\|^2 - \mathbb{E}\left[\|\pi\beta^*\|^2\right]$

To compute $\mathbb{E}\left[\|\pi\beta^*\|^2\right]$ we will use the fact that

if $\quad Z \sim N(0, I_d)$ then $\quad UZ \sim N(0, I_d)$ for any

U orthogonal. Let $\quad U_1, \dots, U_d$ be d orthogonal
$d \times d$

matrices s.t. $\quad U_j \beta^* = \|\beta\|^* e_j$

$\qquad\qquad\qquad j = 1, \dots, d \qquad \hookrightarrow \; j^{th}$ standard
$\qquad\qquad\qquad\qquad\qquad\qquad$ unit vector

$\downarrow$

So $\quad \forall j = 1, \dots, d$

$$\|\pi\beta^*\|^2 = \beta^{*T}\Phi^T(\Phi\Phi^T)^{-1}\Phi\beta^* \overset{d}{=} \beta^{*T} U_j^T \Phi^T \left(\Phi U_j U_j^T \Phi^T\right)^{-1} \Phi U_j \beta^* \quad \overset{I_d}{\overbrace{}}$$

$$= \|\overset{*}{\beta}\|^2 e_j^T \Phi^T(\Phi\Phi^T)^{-1}\Phi e_j$$

$$= \|\overset{*}{\beta}\|^2 \mathrm{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi e_j e_j^T\right)$$

$\downarrow$

$$\mathbb{E}\left[\|\pi\beta^*\|^2\right] = \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\left[\|\pi\beta^*\|^2\right] = \frac{1}{d}\sum_{j=1}^{d}\mathbb{E}\left[\|\beta^*\|^2 \mathrm{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi e_j e_j^T\right)\right]$$

$$= \frac{\|\overset{*}{\beta}\|^2}{d}\mathbb{E}\left[\mathrm{tr}\left(\Phi^T(\Phi\Phi^T)^{-1}\Phi \underbrace{\sum_{j=1}^{d} e_j e_j^T}_{I_d}\right)\right]$$

$\textcircled{4}$

$$= \frac{\|\hat{\beta}\|^2}{d} \, \mathbb{E}\left[\mathrm{tr}\left(\underbrace{(\Phi\Phi^T)^{-1}\Phi\Phi^T}_{I_n} \right) \right]$$

$$= \|\hat{\beta}\|^2 \, \frac{n}{d}$$

Next $\quad T_2 = \sigma^2 \, \mathbb{E}\left[\mathrm{tr}\,(\Phi\Phi^T)^{-1} \right] = \begin{cases} \sigma^2 \dfrac{n}{d-n-1} & d > n+1 \\ \infty & d = n \text{ or } \\ & d = n+1 \end{cases}$

$\downarrow$

The excess risk for $d > n+1$ is

$$\|\hat{\beta}\|^2 \left(1 - \frac{n}{d} \right) + \sigma^2 \, \frac{n}{d-n-1}$$

and it is infinity if $d = n$ or $n+1$

$\hookrightarrow$ To summarize, the excess risk is

$$\begin{cases} \sigma^2 \dfrac{d}{n-d-1} & d \le n-2 \\[3mm] \|\hat{\beta}\| \left(1 - \frac{n}{d} \right) + \sigma^2 \dfrac{n}{d-n-1} & d \ge n+2 \\[3mm] \infty & \text{otherwise} \end{cases}$$

$$= \begin{cases} \sigma^2 \dfrac{\gamma}{1-\gamma} & \gamma < 1 \\[3mm] \|\hat{\beta}\|^2 \dfrac{\gamma-1}{\gamma} + \sigma^2 \dfrac{1}{\gamma-1} & \gamma > 1 \end{cases}$$

- Optimally tuned ridge regression does not suffer from these issues - the risk is monotonic in γ and is uniformly (in γ) smaller than the risk of ridgeless estimator, avoiding the double descent around $\gamma = 1$

The optimal value of the ridge parameter is

$$\lambda_{\text{optimal}} = \frac{\sigma^2}{\|\hat{\beta}\|^2} \gamma$$

↳ CV is a good procedure to find this!