Regularized or Constrained Linear Regression

- goal: restrict the possible coefficients β to avoid overfitting
- overfitting occurs when features are (almost) redundant, i.e. (almost) linearly dependent:

$$X_j \approx \sum_{j' \neq j} w_{j'} X_{j'} \qquad (*)$$

happens in two cases: ① if system of equations is underdetermined, $D > N$ (more features than instances!)

theorem of linear algebra: $(*)$ is always fulfilled [$\text{rank}(X) \leq \min(N, D) = N < D$]

② if condition $\kappa(X)$ is bad (large), $(*)$ is approximately true

why does this cause overfitting: for simplicity, suppose $X_j = X_{j'}$ are exactly equal, $\beta_j X_j + \beta_{j'} X_{j'} = 0$ if $\beta_j = -\beta_{j'}$

now, if X are slightly noisy, $\beta_j X_j \approx \beta_{j'} X_{j'}$ with ~~very~~ to coefficients of very big magnitude $|\beta_j|, |\beta_{j'}| \gg 0$

if ~~that~~ the test data have different noise, the balance between $\beta_j X_j$ and $\beta_{j'} X_{j'}$ no longer works $\Rightarrow$ if $|\beta_j|, |\beta_{j'}| \gg 0$, we get huge error

- conclusion: we want to prevent $|\beta_j| \gg 0$ "constraint" or "regularization"

$\Rightarrow$ **Ridge Regression**

$$\hat{\beta} = \arg\min_{\beta} (y - X\beta)^2 \quad \text{s.t.} \quad \underbrace{\beta^T \beta = \|\beta\|_2^2 \leq t}_{\text{constraint}}$$

(t: hyperparameter)

if $\hat{\beta}_{OLS}$ of OLS solution already has small norm $\|\hat{\beta}_{OLS}\|_2^2 \leq t$, then the constraint is "inactive" and nothing changes. Otherwise, the optimal solutions will differ. Rewrite objective with Lagrange multiplier τ

$$\hat{\beta} = \arg\min_{\beta} (Y - X\beta)^2 + \tau \beta^T\beta \qquad \tau \geq 0$$

($\tau\beta^T\beta$: Lagrange multipl.)

Theorem: the two variants of the objective give the same solution $\hat{\beta}$ if τ is chosen appropriately for each given t.

- $\frac{\partial Loss}{\partial \beta} = 2\left[X^T(Y - X\beta) + \tau\beta \right] \stackrel{!}{=} 0$

$$\boxed{\hat{\beta}_\tau = (X^TX + \tau \mathbb{I})^{-1} X^T Y}$$

regularized pseudo-inverse

- for $\tau = 0$, this reduces to ordinary pseudo-inverse $(X^TX)^{-1}X^T$
- for $\tau > 0$, the diagonal of $X^TX + \tau\mathbb{I}$ is bigger than the diagonal of X^TX (scatter matrix $= S$) $\Rightarrow \kappa(\underbrace{X^TX + \tau\mathbb{I}}_{\text{regularized scatter matrix } S_\tau}) < \kappa(X^TX)$, i.e. better condition

[reminder: X and Y must be centered, $\bar{x} = 0$, $\bar{Y} = 0$]

The effect of adding τ to diagonal element $(X^TX)_{jj}$ depends on the magnitude of $(X^TX)_{jj}$ $\Rightarrow$ scale the features beforehand to have unit variance $\Rightarrow$ standardization $\tilde{x}_j = \frac{x_j - \bar{x}_j}{std(x_j)}$ $\Rightarrow$ $(X^TX)_{jj} = 1$ for all j

solutions of ridge regression:

(1) compute and apply regularized pseudo-inverse: easy, but numerically less accurate

(2) reduce it to ordinary least squares: augmented data matrix

$$X' = \begin{bmatrix} X \\ \begin{matrix} \sqrt{\tau} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\tau} \end{matrix} \end{bmatrix} \begin{matrix} \} N \text{ rows} \\ \} D \text{ rows} \end{matrix} \qquad Y' = \begin{bmatrix} Y \\ 0 \end{bmatrix} \begin{matrix} N \text{ elements} \\ D \text{ zeros} \end{matrix}$$

$$\underbrace{(Y' - X'\beta)^T (Y' - X'\beta)}_{\text{OLS objective for } X', Y'} = \underbrace{(Y - X\beta)^T (Y - X\beta) + \tau \beta^T \beta}_{\text{ridge regression objective}}$$

(3) use singular value decomposition (SVD):

$$X = U \Lambda V^T \qquad X^T = V \Lambda U^T \qquad X^T X = V \Lambda^2 V^T$$

$$X^T X + \tau I\!I = V \Lambda^2 V^T + \tau \underbrace{I\!I}_{V V^T = I\!I} = V \Lambda^2 V^T + \tau V I\!I V^T$$

$$= V (\Lambda^2 + \tau I\!I) V^T$$

$$(X^T X + \tau I\!I)^{-1} X^T = (V (\Lambda^2 + \tau I\!I) V^T)^{-1} V \Lambda U^T = V \underbrace{(\Lambda^2 + \tau I\!I)^{-1} \Lambda}_{\Lambda'} U^T$$

Λ' is a diagonal matrix with elements $\lambda_j' = \dfrac{\lambda_j}{\lambda_j^2 + \tau}$

for $\tau = 0$, $\lambda_j' = \dfrac{1}{\lambda_j^2} \Rightarrow \Lambda' = \Lambda^{-1}$, reduces to ordinary pseudo-inverse

- if $\tau > 0$, rewrite $\lambda_j' = \frac{\lambda_j}{\lambda_j^2 + \tau} = \frac{1}{\lambda_j + \frac{\tau}{\lambda_j}}$

two cases: (1) $\lambda_j >> 0$ (no redundancy) $\Rightarrow \frac{\tau}{\lambda_j} \approx 0 \Rightarrow \lambda_j' \approx \frac{1}{\lambda_j}$

$\Rightarrow$ if $\lambda_j >> \tau$, regularization has no effect

(2) $\lambda_j \approx 0$ (redundant features) $\Rightarrow \lambda_j' = \frac{0}{0^2 + \tau} = 0$

$\Rightarrow$ in ordinary least squares $\lambda_j' = \frac{1}{\lambda_j} >> 0$ pseudo-inverse "explodes"

ridge regression $\lambda_j' = \frac{\lambda_j}{\tau} \approx 0 \Rightarrow$ no "explosion"

(4) solve via dual optimization problem $\Rightarrow$ later

- What's the drawback of regularization? Does improved conditioning come at a price?

Bias - Variance - Trade-off

Suppose, we have M training sets of size N. What happens when $M \to \infty$?

Repeat experiment M times and check how much β varies: if it a lot varies $\Rightarrow$ overfitting, otherwise little/no overfitting.

Mathematically: $\hat{\beta}_m$ = OLS-solution for TS m $(m = 1 \dots M)$

expectation of all TS: $\mathbb{E}_m[\hat{\beta}_m]$, β^*: unknown true solution

bias (systematic error) $= \beta^* - \mathbb{E}_m[\hat{\beta}_m]$, variance $= \mathbb{E}_m\left[(\hat{\beta}_m - \mathbb{E}_m[\hat{\beta}_m])^2\right]$

- calculate the expected mean-square error of $\hat{\beta}_m$

$$MSE = \mathbb{E}_m\left[(\hat{\beta}_m - \beta^*)^2\right] = \mathbb{E}_m\left[\left((\hat{\beta}_m - \mathbb{E}_m[\hat{\beta}_m]) + (\mathbb{E}_m[\hat{\beta}_m] - \beta^*)\right)^2\right]$$

$$= \underbrace{\mathbb{E}_m\left[(\hat{\beta}_m - \mathbb{E}_m[\hat{\beta}_m])^2\right]}_{\text{expected variance}} + \underbrace{\mathbb{E}_m\left[(\mathbb{E}_m[\hat{\beta}_m] - \beta^*)^2\right]}_{\text{expected squared bias}}$$

[cross terms vanish:

$$-2\,\mathbb{E}_m\left[(\hat{\beta}_m - \mathbb{E}_m[\hat{\beta}_m])(\mathbb{E}_m[\hat{\beta}_m] - \beta^*)\right]$$

$$= -2\left[\mathbb{E}_m[\hat{\beta}_m](\mathbb{E}_m[\hat{\beta}_m] - \beta^*) - \underbrace{\mathbb{E}_m[\mathbb{E}_m[\hat{\beta}_m]]}_{\mathbb{E}_m[\hat{\beta}_m]}(\mathbb{E}_m[\hat{\beta}_m] - \beta^*)\right) = 0\Big]$$

- apply bias-variance decomposition to OLS: $\hat{\beta}_m = (X_m^T X_m)^{-1} X_m^T \underbrace{y_m}_{= X_m\beta^* + \varepsilon_m}$

$$\mathbb{E}_m[\hat{\beta}_m] = \mathbb{E}_m\left[(X_m^T X_m)^{-1} X_m^T (X_m\beta^* + \varepsilon_m)\right]$$

$$= \mathbb{E}_m\Big[\underbrace{(X_m^T X_m)^{-1}(X_m^T X_m)}_{= \mathbb{I}}\beta^* + \underbrace{(X_m^T X)^{-1} X_m^T \varepsilon_m}_{= \mathbb{E}_m[(X_m^T X_m)^{-1} X_m^T]\,\underbrace{\mathbb{E}_m[\varepsilon_m]}_{=0}}\Big]$$

$\varepsilon_m \sim N(0, \sigma^2)$

$$= \mathbb{E}_m[\beta^*] = \beta^*$$

$$\Rightarrow \mathbb{E}_m\left[(\mathbb{E}_m[\hat{\beta}_m] - \beta^*)^2\right] = \mathbb{E}_m\left[(\beta^* - \beta^*)^2\right] = 0$$

squared bias of OLS is zero: "OLS is unbiased"

expected co-variance matrix of $\hat{\beta}_m$ across infinitely many TS

$$\mathbb{E}_m\left[\left(\hat{\beta}_m - \mathbb{E}_m[\hat{\beta}_m]\right)^2\right] = \mathbb{E}_m\left[\left(\hat{\beta}_m - \beta^*\right)^2\right]$$

$$\left(\hat{\beta}_m - \beta^*\right)^2 = \left(\hat{\beta}_m - \beta^*\right)\left(\hat{\beta}_m - \beta^*\right)^T$$

$$\hat{\beta} = \left(X_m^T X_m\right)^{-1} X_m^T \underbrace{y_m}_{X_m\beta^* + \varepsilon_m}$$

$$= \beta^* + \left(X_m^T X_m\right)^{-1} X_m^T \varepsilon_m$$

$$\hat{\beta}_m - \beta^* = \left(X_m^T X_m\right)^{-1} X_m^T \varepsilon_m$$

$$\left(\hat{\beta}_m - \beta^*\right)\left(\hat{\beta}_m - \beta^*\right)^T = \left(X_m^T X_m\right)^{-1} X_m^T \varepsilon_m \varepsilon_m^T X_m \left(X_m^T X_m\right)^{-1}$$

$$\mathbb{E}_m\left[\left(\hat{\beta}_m - \beta^*\right)\left(\hat{\beta}_m - \beta^*\right)^T\right] = \mathbb{E}_m\left[\left(X_m^T X_m\right)^{-1} X_m^T\right] \underbrace{\mathbb{E}_m\left[\varepsilon_m \varepsilon_m^T\right]}_{=\sigma^2 I \text{ because}} \mathbb{E}_m\left[X_m \left(X_m^T X_m\right)^{-1}\right]$$

noise of all instances is independent $\varepsilon_m \sim N(0, \sigma^2)$

$$= \sigma^2 \mathbb{E}_m\left[\underbrace{\left(X_m^T X_m\right)^{-1} X_m^T X_m}_{=I} \left(X_m^T X_m\right)^{-1}\right]$$

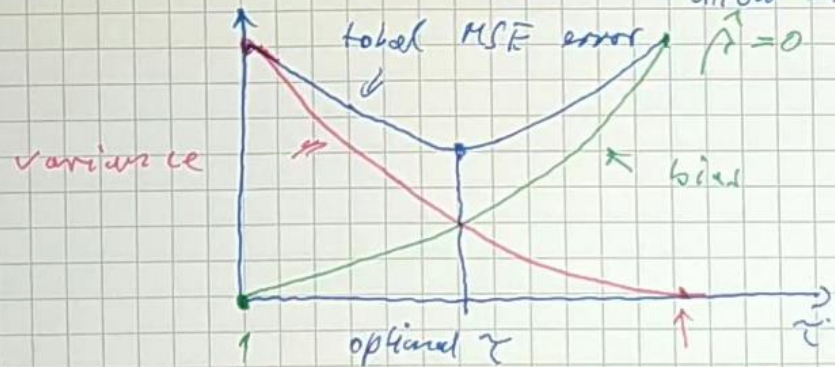
$$\mathbb{E}_m\left[\left(\hat{\beta}_m - \beta^*\right)^2\right] = \mathbb{E}_m\left(X_m^T X_m\right)^{-1} \sigma^2, \text{ for single TS: } \left(X^T X\right)^{-1} \sigma^2$$

variance of solution $\hat{\beta}$ is proportional to noise variance σ^2 and to inverse scatter matrix $\left(X_m^T X_m\right)^{-1}$

$\Rightarrow$ if features are redundant, $X_m^T X_m$ has bad condition, $\left(X_m^T X_m\right)^{-1}$ explodes

$\Rightarrow$ high variance of $\hat{\beta}_m$ over TS $\Rightarrow$ we cannot profit from unbiasedness

⇒ bias-variance trade-off: sacrifice some bias, if it reduces the variance a lot
"allow some bias"



$\hat{\beta}=0 \Rightarrow \text{bias}^2 = |\beta^*|^2$

MSE error has U-shape ⇒ select τ at the minimal MSE (cross-validation)

τ so big that $\hat{\beta}=0 \Rightarrow$ variance $=0$

bias and variance of ridge regression:

$$\mathbb{E}_n[\hat{\beta}_n(\tau)] = (X^TX+\tau\mathbb{I})^{-1}X^TX\cdot\beta^*, \quad \text{for } \tau=0: \mathbb{E}_n[\hat{\beta}_n(0)]=\beta^*$$

$$= V \operatorname{diag}\left(\frac{\lambda_j^2}{\lambda_j^2+\tau}\right) V^T\beta^*$$

variance $\text{Cov}[\hat{\beta}_\tau] = (X^TX+\tau\mathbb{I})^{-1}X^TX(X^TX+\tau\mathbb{I})^{-1}\sigma^2$

$$= V \operatorname{diag}\left(\frac{\lambda_j^2}{(\lambda_j^2+\tau)^2}\right) V^T\sigma^2$$