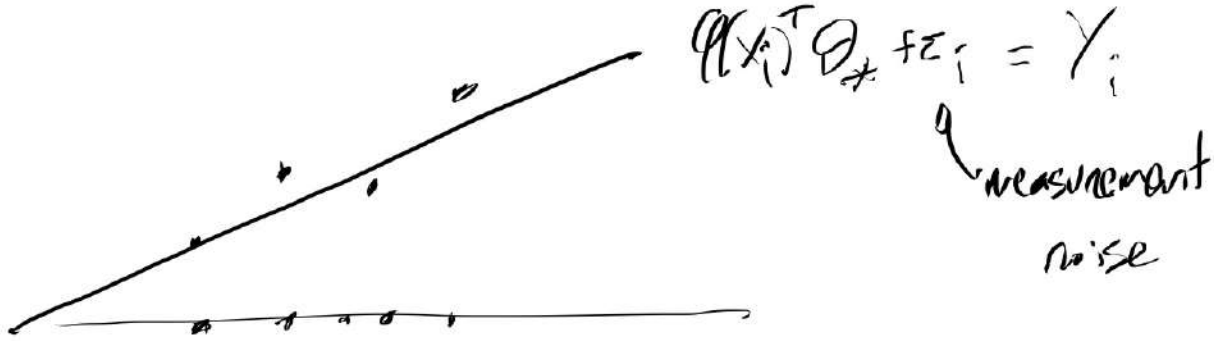


H: | ε_i are independant
$\mathbb{E}\varepsilon_i = 0$, $\mathbb{E}\varepsilon_i^2 = \sigma^2 < +\infty$

excess risk of prediction f_θ $\left(x \longmapsto \varphi(x)^\top \theta\right)$:

$R(f_\theta) - R^* =$

Prop. 1: $R^* = \sigma^2$; $\overline{R(\theta) - R^* = \|\theta - \theta^*\|^2_\Sigma}$ to get

expected excess risk = $\mathbb{E}[R(\tilde{\theta})] - R^*$
of the ERM : Lif $\tilde{\theta}$ is random

Prop 2: $\mathbb{E}[R(\tilde{\theta})] - R^*$ = bias + variance

Prop 3: bias $= 0$; var $= \frac{\sigma^2}{n} \hat{\Sigma}^{-1}$

Say V random vector.

$(Cov(V))_{jk} = \mathbb{E}[(V_j - \mathbb{E}V_j)(V_k - \mathbb{E}V_k)]$

$=:$ Var.

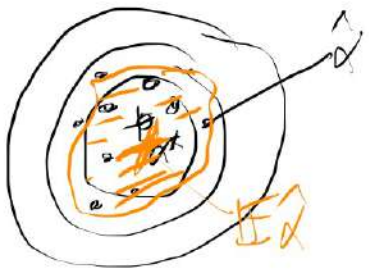
Prop. 4: $\mathbb{E}R(\hat{\theta}) - R^* = \frac{\sigma^2 d}{n}$ → dimension

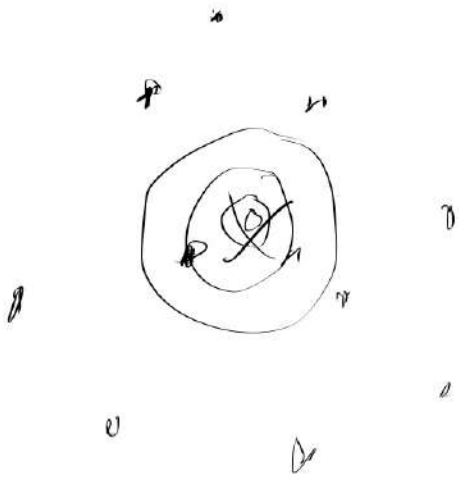
- $\mathbb{E}R(\hat{\theta})$: expected risk of $\hat{\theta}$
- R^*: Bayes risk
- σ^2: noise level
- n: sample size

Q re bias: in statistics, we estimate true quantities. An estimator is a random variable which depends only on $X_1, \dots, X_n$

Def: $\alpha^* \in \mathbb{R}$, $\hat{\alpha}$ estimator of α^*,

$\text{bias}(\hat{\alpha}) := \mathbb{E}[\hat{\alpha}] - \alpha^*$.





gone bins "good," but not the end of the story

RIDGE REGRESSION

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2}_{\text{empirical risk}} + \underbrace{\lambda \|\theta\|^2}_{\text{regularization / penalization / penalty for}} \right\}$$

your loss

$\lambda \geq 0$, fixed

in our notation:

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \|Y - X\theta\|^2 + \lambda \|\theta\|^2 \right\}$$

We know this: CVX + Smooth

CVX + Smooth (Ψ)

$F(\theta) :=$ red + blue is convex, smooth $\rightarrow \nabla F$?

$$\nabla F(\theta) = (\text{already done it}) + \nabla(\lambda\|\theta\|^2)$$

$$\|\theta\|^2 = \theta^T\theta = \theta^T I\theta$$

$$\nabla(\theta^T I\theta) = (I + I^T)\theta = 2I\theta = 2\theta$$

$$\nabla F(\theta) = \frac{2}{n}(X^TX\theta - X^TY) + 2\lambda\theta$$

$$= 2\left(\left(\hat{\Sigma} + \lambda I\right)\theta - X^TY\right)$$

Claim 1: $\nabla F = 0 \iff (\hat{\Sigma} + \lambda I)\theta = X^TY$

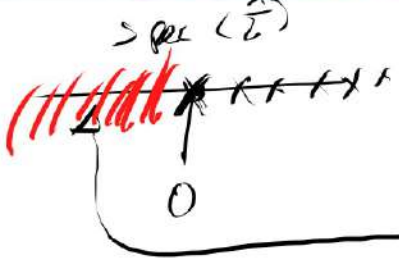
equation in θ. how to solve it?

claim 2: $\hat{\Sigma} + \lambda I$ is always invertible ($\lambda > 0$)

$$\Rightarrow \hat{\theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} X^TY$$

(original matrix.)

$\hat{\Sigma} \in S_d^+$



pendant dio: $\hat{\Sigma} + \lambda I$ is not invertible.
$\det(\hat{\Sigma} + \lambda I) = 0$
$(-\lambda)$ is eigenvalue of $\hat{\Sigma}$.

Goal: expected excess risk of ridge

Q: bias-variance decomposition still holds?

→ not true for any estimator, in particular for $\hat{\theta}_\lambda$

For $\lambda = 1$: $\mathbb{E}[(\alpha - \alpha^*)^2] = \mathbb{E}[(\alpha - \mathbb{E}\hat{\alpha} + \mathbb{E}\hat{\alpha} - \alpha^*)^2]$

$= \mathbb{E}[(\alpha - \mathbb{E}\hat{\alpha})^2 + 2(\alpha - \mathbb{E}\hat{\alpha})(\mathbb{E}\hat{\alpha} - \alpha^*) + (\mathbb{E}\hat{\alpha} - \alpha^*)^2]$

$= \mathbb{E}[(\alpha - \mathbb{E}\hat{\alpha})^2] + 2\,\mathbb{E}[(\alpha - \mathbb{E}\hat{\alpha})(\mathbb{E}\hat{\alpha} - \alpha^*)] + \mathbb{E}[(\mathbb{E}\hat{\alpha} - \alpha^*)^2]$

($\mathbb{E} = 0$)

Todo: bias and variance of $\hat{\theta}_\lambda$

Prop 5: $\text{bias}(\hat{\theta}_\lambda) = -\lambda(\hat{\Sigma} + \lambda I)^{-1}\theta^*$

Rk: $\lambda = 0$: $\hat{\theta}_\lambda = \hat{\theta}$, bias $= 0$ as expected

Proof: $\mathbb{E}\hat{\theta}_\lambda = \mathbb{E}[\frac{1}{n}(\hat{\Sigma} + \lambda I)^{-1} X^T Y]$

$= \mathbb{E}[\frac{1}{n}(\hat{\Sigma} + \lambda I)^{-1} X^T (X\theta^* + \varepsilon)]$

$= \mathbb{E}[\frac{1}{n}(\hat{\Sigma} + \lambda I)^{-1} X^T X \theta^*] + \mathbb{E}[\dots \varepsilon]$

($\hat{\Sigma}$; since $\mathbb{E}\varepsilon_i = 0$)

$$\mathbb{E}[\hat{\theta}_\lambda] = (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} \theta^*$$

$$I - (\hat{\Sigma} + \lambda I)^{-1}(\hat{\Sigma} + \lambda I) = (\hat{\Sigma} + \lambda I)^{-1}\hat{\Sigma} + \lambda(\hat{\Sigma} + \lambda I)^{-1}$$

$$(\hat{\Sigma} + \lambda I)^{-1} = I - \lambda(\hat{\Sigma} + \lambda I)^{-1}$$

$$\Rightarrow \mathbb{E}\theta_\lambda - \theta^* = -\lambda(\hat{\Sigma} + \lambda I)^{-1}\theta^*$$

$$\| \cdot \|_{\hat{\Sigma}}^2 \longrightarrow \cdots$$

Prop 6: $\mathrm{Var}(\hat{\theta}_\lambda) = \frac{\sigma^2}{n} \mathrm{trace}(\hat{\Sigma}^2(\hat{\Sigma} + \lambda I)^{-2})$

Proof: $\hat{\theta}_\lambda - \mathbb{E}\hat{\theta}_\lambda = (\hat{\Sigma} + \lambda I)^{-1} X^T y - (\hat{\Sigma} + \lambda I)^{-1} X^T X \theta^*$

$y = X\theta^* + \varepsilon$

$= (\hat{\Sigma} + \lambda I)^{-1} X^T \varepsilon$.

$z \in \mathbb{R}$, $z = \mathrm{trace}(z)$

$$\mathbb{E}\left[\|\hat{\theta}_\lambda - \mathbb{E}\theta_\lambda\|_{\hat{\Sigma}}^2\right] = \mathbb{E}\left[\frac{1}{n}\|(\hat{\Sigma} + \lambda I)^{-1} X^T \varepsilon\|_{\hat{\Sigma}}^2\right]$$

$$= \frac{1}{n^2}\mathbb{E}\left[\varepsilon^T X (\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda I)^{-1} X^T \varepsilon\right] \in \mathbb{R}$$

$$= \frac{1}{n^2}\mathbb{E}\left[\mathrm{trace}(\cdots)\right]$$

$$\text{Var} = \frac{1}{n^2}\mathbb{E}\left[\text{trace}\left(\varepsilon^T \bar{X}\left(\hat{\Sigma}+\lambda I\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1}\bar{X}^T\varepsilon\right)\right]$$

[cyclic property : $\text{trace}(AB) = \text{trace}(BA)$]

$$= \frac{1}{n^2}\mathbb{E}\left[\text{trace}\left(\varepsilon\varepsilon^T \underbrace{\qquad\qquad}_{A}\right)\right]$$

$$= \frac{1}{n^2}\text{trace}\left(\mathbb{E}[\varepsilon\varepsilon^T]\qquad\qquad\right)$$

$$= \frac{\sigma^2}{n}\text{trace}\left(\mathbb{E}\left(\hat{\Sigma}+\lambda I\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1}\bar{X}^T\right)$$

$$= \frac{\sigma^2}{n}\text{trace}\left(\underbrace{\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1}\hat{\Sigma}\left(\hat{\Sigma}+\lambda I\right)^{-1}}_{A}\right)$$

$$= \frac{\sigma^2}{n}\text{trace}\left(\hat{\Sigma}^2\left(\hat{\Sigma}+\lambda I\right)^{-2}\right) \quad \square$$

Prop 7: $\mathbb{E}R(\theta_\lambda) - R^* = \lambda^2\left(\theta^{*T}\left(\hat{\Sigma}+\lambda I\right)^{-2}\hat{\Sigma}\,\theta^*\right)$

$$+ \frac{\sigma^2}{n}\text{trace}\left(\hat{\Sigma}^2\left(\hat{\Sigma}+\lambda I\right)^{-2}\right)$$

$\frac{d^2}{\sigma^2+3}$ $\qquad\qquad$ $\approx \frac{1}{d^2}$

Rk: $d=0$ has NO reason to be the best.