Statistical Learning II

Lecture 8 - Ridge regression (continued)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
13.11.2024

Ridge regression

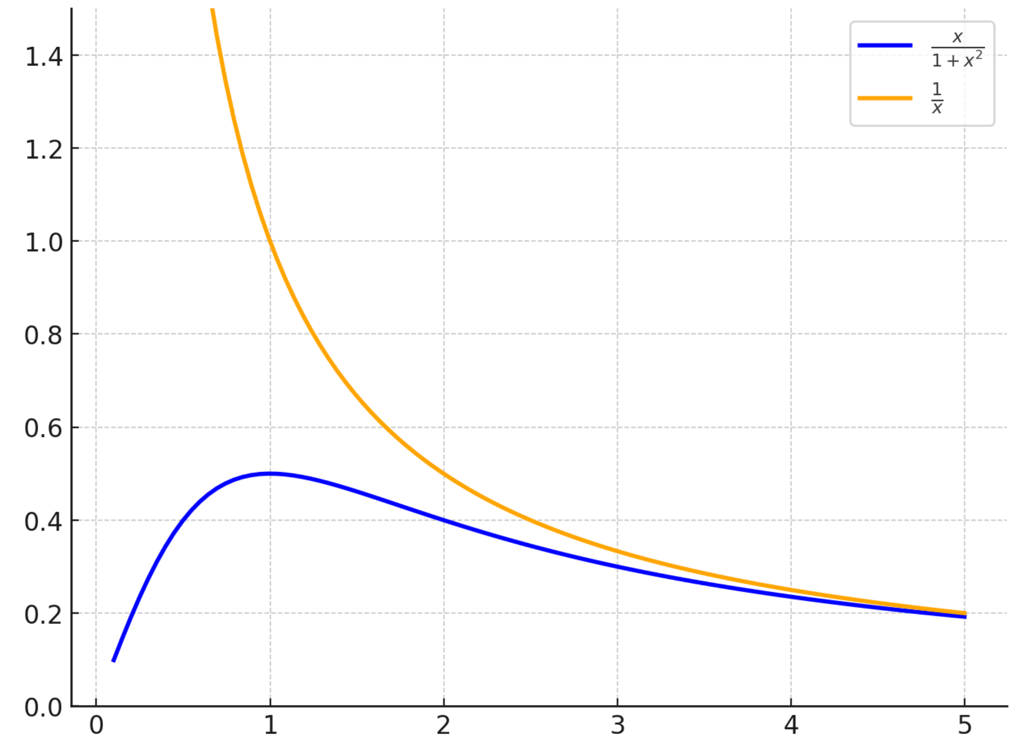
$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \lambda \boldsymbol{I}_d\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

<u>Remarks:</u>

- As before, consider s.v.d. of $\boldsymbol{X} = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j$

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \sum_{j=1}^{\mathrm{rank}(\boldsymbol{X})} \frac{\sigma_j}{\sigma_j^2 + n\lambda}\langle \boldsymbol{u}_j, \boldsymbol{y}\rangle \boldsymbol{v}_j$$

Ridge performs shrinkage:
small s.v.s are suppressed!



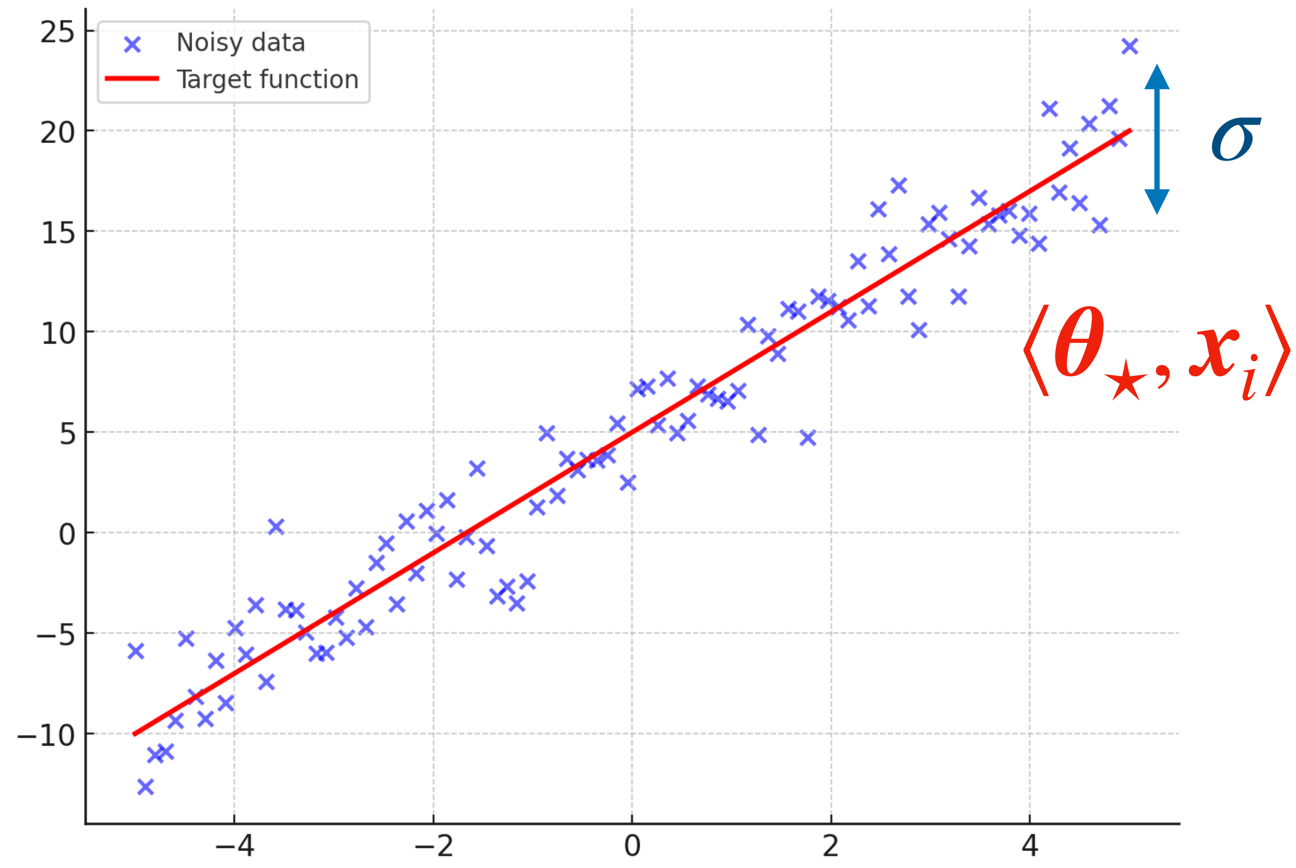
Statistical analysis of ridge regression

Fixed design assumption

As we did for the OLS, now let's assume:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:
- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$



Decomposition of ridge

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

The ridge estimator is given by:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Decomposition of ridge

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_{\star} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

The ridge estimator is given by:

$$\hat{\boldsymbol{\theta}}_{\lambda}(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^{\top}\boldsymbol{y}$$

Decomposition of ridge

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

The ridge estimator is given by:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon}\right)$$

Decomposition of ridge

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

The ridge estimator is given by:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) &= \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon}\right) \\
&= \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\theta}_\star + \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}
\end{aligned}$$

Decomposition of ridge

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

The ridge estimator is given by:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) &= \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \left(\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon}\right) \\
&= \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\theta}_\star + \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \\
&= \boldsymbol{\theta}_\star - \lambda\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star + \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}
\end{aligned}$$

Decomposition of ridge

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{\theta}_\star + \frac{1}{n} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

"signal" "noise"

Decomposition of ridge

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star + \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

"signal" "noise"

In particular:

- Bias: $$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y})\right] = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star$$

Decomposition of ridge

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{\theta}_\star + \frac{1}{n} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

"signal" "noise"

In particular:

- Bias: $$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) \right] = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-1} \boldsymbol{\theta}_\star$$
- Variance: $$\mathrm{Var}_{\boldsymbol{\varepsilon}} \left[\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) \right] = \frac{\sigma^2}{n} \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d \right)^{-2} \hat{\boldsymbol{\Sigma}}_n$$

Decomposition of ridge

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star + \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

"signal" "noise"

In particular:

- Bias: $$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y})\right] = \boldsymbol{\theta}_\star - \lambda \left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{\theta}_\star$$
- Variance: $$\mathrm{Var}_{\boldsymbol{\varepsilon}}\left[\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y})\right] = \frac{\sigma^2}{n}\left(\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d\right)^{-2} \hat{\boldsymbol{\Sigma}}_n$$

- Ridge is a biased estimator.
- Regularisation shrinks both signal and noise

Risk of ridge

Recall that in Lecture 5 we have shown that for any $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) - \sigma^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}_\star)^\top \hat{\boldsymbol{\Sigma}}_n (\boldsymbol{\theta} - \boldsymbol{\theta}_\star)$$

Risk of ridge

Recall that in Lecture 5 we have shown that for any $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) - \sigma^2 = (\boldsymbol{\theta} - \boldsymbol{\theta}_\star)^\top \hat{\boldsymbol{\Sigma}}_n (\boldsymbol{\theta} - \boldsymbol{\theta}_\star)$$

Therefore, inserting the solution $\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{X}, \boldsymbol{y})$:

$$\begin{aligned}\mathscr{R}(\hat{\boldsymbol{\theta}}_\lambda) - \sigma^2 = \lambda^2 \boldsymbol{\theta}_\star^\top (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2} \boldsymbol{\theta}_\star \\ + \frac{1}{n^2} \boldsymbol{\varepsilon}^\top \boldsymbol{X} (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{X} (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \\ - \frac{\lambda}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{X} (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2} \boldsymbol{\theta}_\star\end{aligned}$$

Risk of ridge

Taking the expectation with respect to the noise:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathscr{R}(\hat{\boldsymbol{\theta}}_\lambda)] - \sigma^2 = \lambda^2 \boldsymbol{\theta}_\star^\top (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\theta}_\star + \frac{\sigma^2}{n} \mathrm{Tr}\ \hat{\boldsymbol{\Sigma}}_n^2 (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2}$$

Risk of ridge

Taking the expectation with respect to the noise:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathscr{R}(\hat{\boldsymbol{\theta}}_\lambda)] - \sigma^2 = \lambda^2 \boldsymbol{\theta}_\star^\top (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\theta}_\star + \frac{\sigma^2}{n} \mathrm{Tr}\ \hat{\boldsymbol{\Sigma}}_n^2 (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2}$$

Alternatively, we can also write in terms of a bias-variance decomposition of the risk:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathscr{R}(\hat{\boldsymbol{\theta}}_\lambda)] - \sigma^2 = \mathscr{B} + \mathscr{V}$$

Where:

$$\mathscr{B} = \lambda^2 \boldsymbol{\theta}_\star^\top (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2} \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\theta}_\star \qquad \mathscr{V} = \frac{\sigma^2}{n} \mathrm{Tr}\ \hat{\boldsymbol{\Sigma}}_n^2 (\hat{\boldsymbol{\Sigma}}_n + \lambda \boldsymbol{I}_d)^{-2}$$

Risk of ridge

Considering the SVD of $\boldsymbol{X} = \sum_{k=1}^{\mathrm{rank}(X)} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$, we can also write:

$$\mathscr{B} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{(n\lambda)^2 \lambda_k \langle \boldsymbol{v}_k, \boldsymbol{\theta}_\star \rangle^2}{(\lambda_k + n\lambda)^2} \quad \mathscr{V} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{\sigma^2 \lambda_k^2}{(\lambda_k + n\lambda)^2}$$

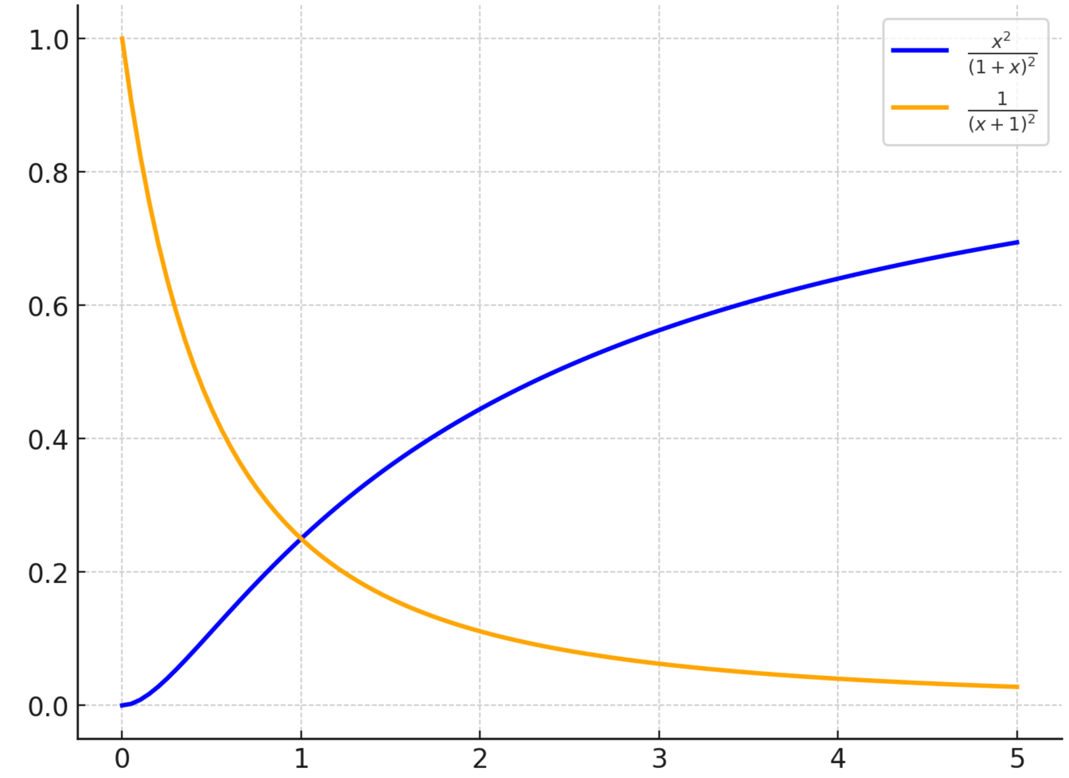
Risk of ridge

Considering the SVD of $\boldsymbol{X} = \sum_{k=1}^{\mathrm{rank}(X)} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$, we can also write:

$$\mathscr{B} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{(n\lambda)^2 \lambda_k \langle \boldsymbol{v}_k, \boldsymbol{\theta}_\star \rangle^2}{(\lambda_k + n\lambda)^2} \qquad \mathscr{V} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{\sigma^2 \lambda_k^2}{(\lambda_k + n\lambda)^2}$$

<u>Remarks:</u>

- For $\lambda \to 0^+$, we get the OLS excess risk
- $\mathscr{B}(\lambda)$ is an increasing function of λ
- $\mathscr{V}(\lambda)$ is a decreasing function of λ



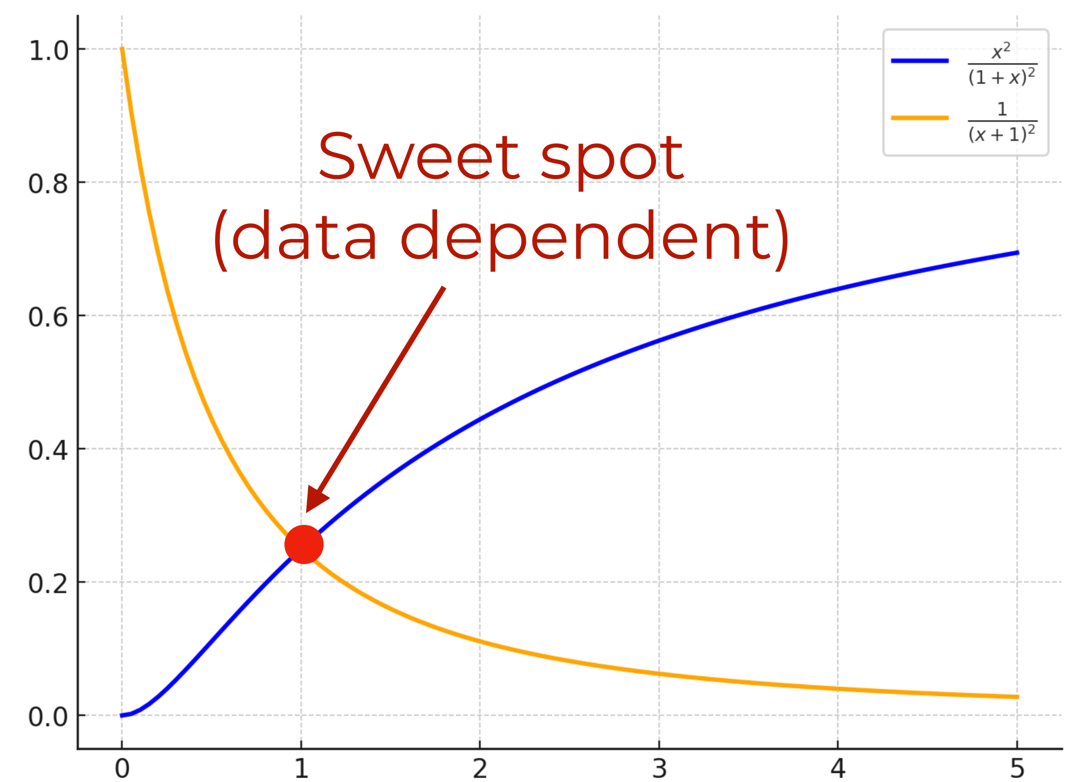
Risk of ridge

Considering the SVD of $\boldsymbol{X} = \sum_{k=1}^{\mathrm{rank}(X)} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$, we can also write:

$$\mathscr{B} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{(n\lambda)^2 \lambda_k \langle \boldsymbol{v}_k, \boldsymbol{\theta}_\star \rangle^2}{(\lambda_k + n\lambda)^2} \qquad \mathscr{V} = \sum_{k=1}^{\mathrm{rank}(X)} \frac{\sigma^2 \lambda_k^2}{(\lambda_k + n\lambda)^2}$$

Remarks:

- For $\lambda \to 0^+$, we get the OLS excess risk
- $\mathscr{B}(\lambda)$ is an increasing function of λ
- $\mathscr{V}(\lambda)$ is a decreasing function of λ



Interpretation of variance

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a positive definite matrix with decreasing eigenvalues $\mathrm{spec}(\boldsymbol{A}) = \{\lambda_k : k = 1, \cdots, d\}$. Define the cumulative:

$$\phi(\lambda) = \#\{k : \lambda_k > \lambda\}$$

"Count eigenvalues bigger than λ"

Interpretation of variance

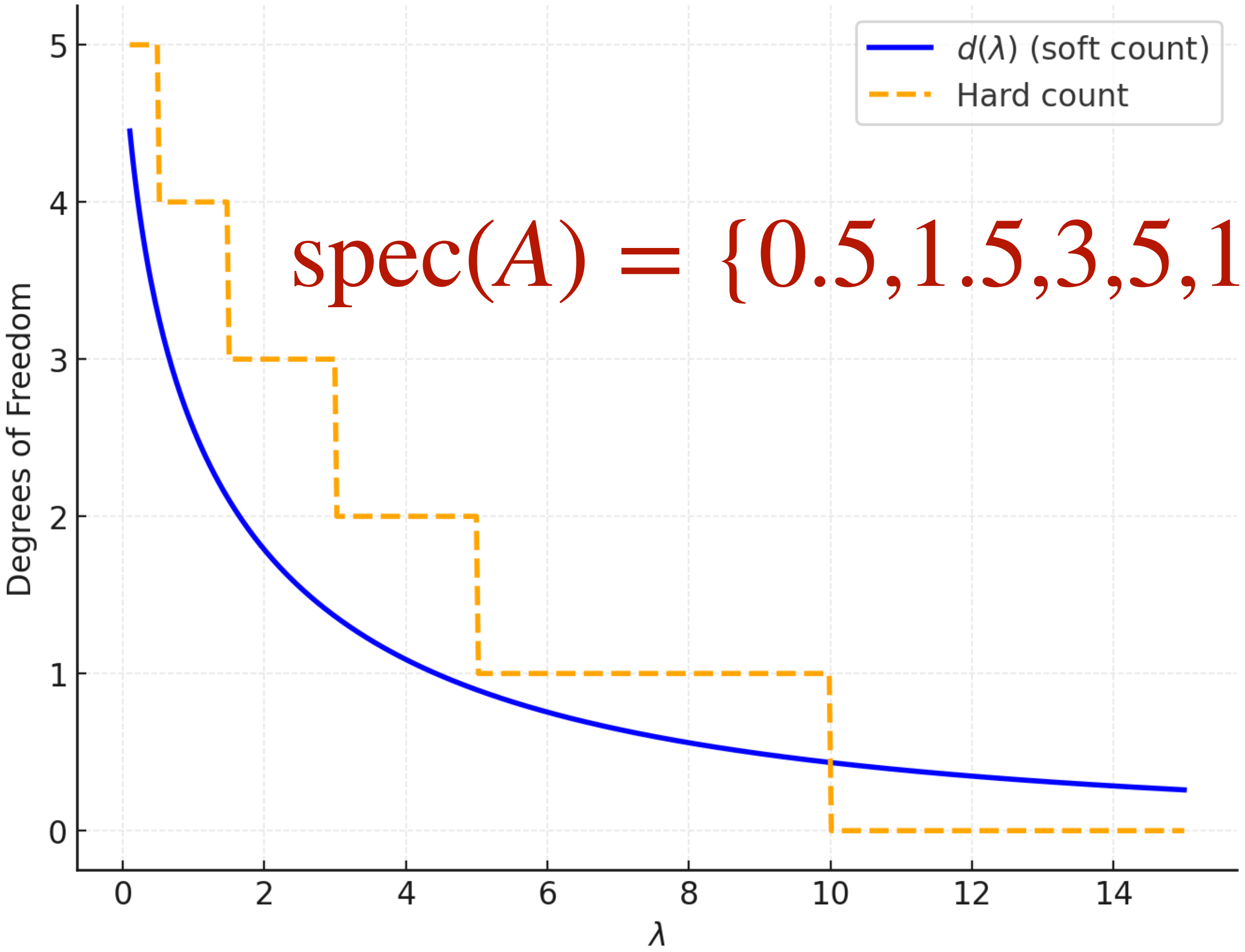
Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a positive definite matrix with decreasing eigenvalues $\mathrm{spec}(\boldsymbol{A}) = \{\lambda_k : k = 1, \cdots, d\}$. Define the cumulative:

$$\phi(\lambda) = \#\{k : \lambda_k > \lambda\}$$

"Count eigenvalues bigger than λ"

The variance of the ridge risk can be seen as a soft version:

$$\mathrm{d}f_2(\lambda) = \sum_{k=1}^{d} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2}$$



$\mathrm{spec}(A) = \{0.5, 1.5, 3, 5, 10\}$

Interpretation of variance

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a positive definite matrix with decreasing eigenvalues $\mathrm{spec}(\boldsymbol{A}) = \{\lambda_k : k = 1, \cdots, d\}$. Define the cumulative:

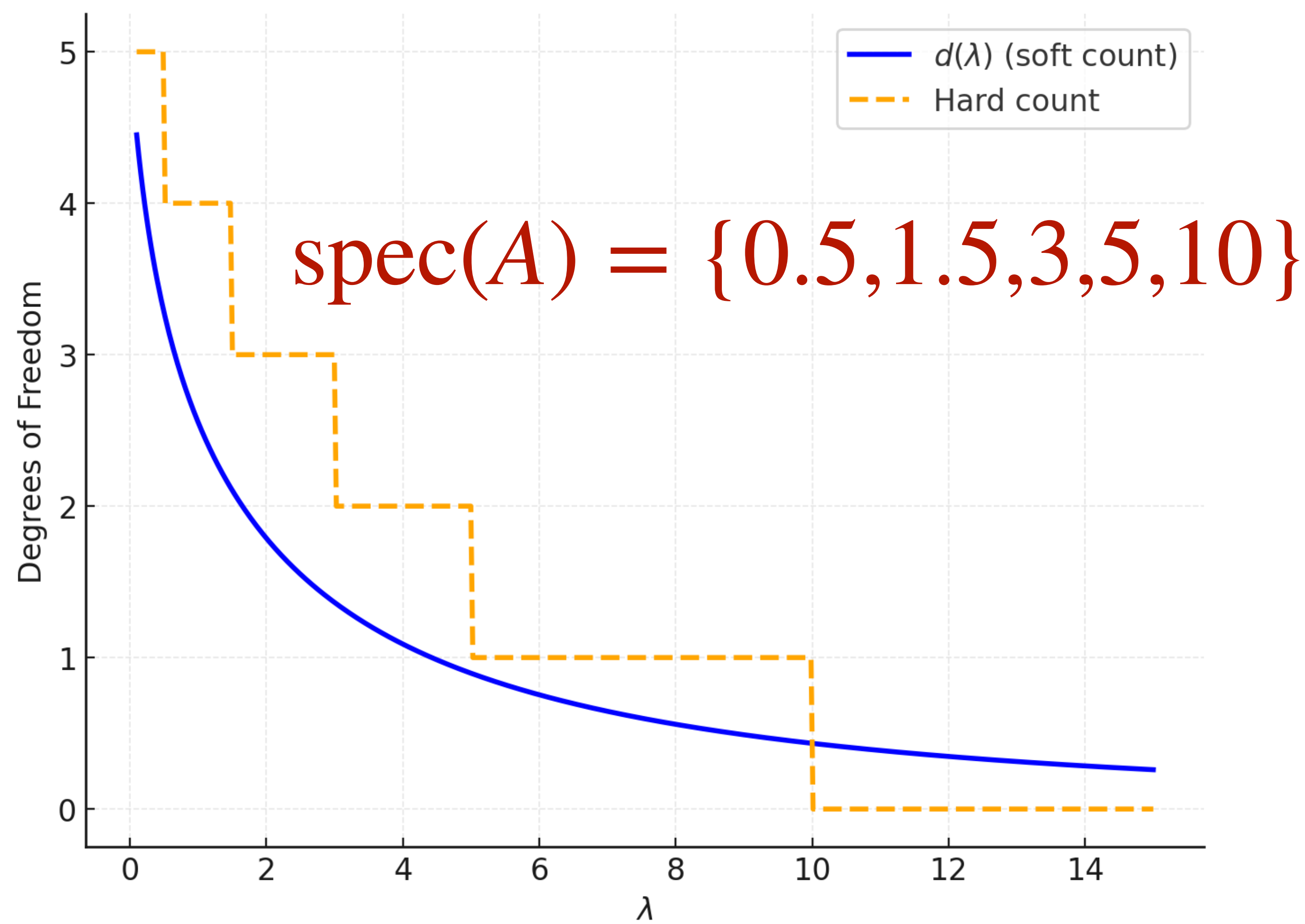
$$\phi(\lambda) = \#\{k : \lambda_k > \lambda\}$$

"Count eigenvalues bigger than λ"

The variance of the ridge risk can be seen as a soft version:

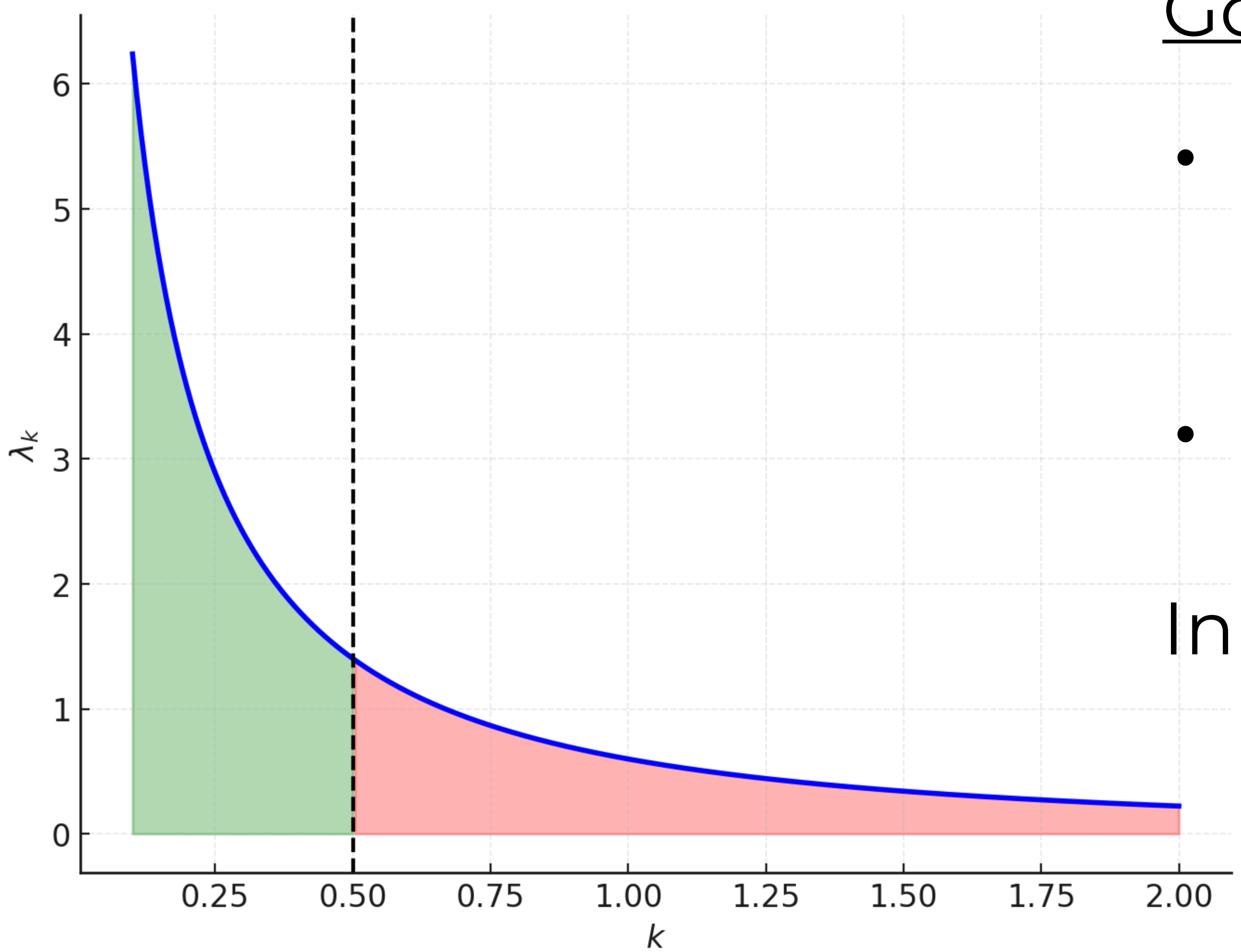
$$\mathrm{d}f_2(\lambda) = \sum_{k=1}^{d} \frac{\lambda_k^2}{(\lambda_k + \lambda)^2}$$

- Fast decay: small λ
- Slow decay: large λ



Choosing regularisation

$\mathrm{d}f_2(\lambda)$



Low-frequency High-frequency

Goal: pick λ such that:

- directions in $\boldsymbol{X}$ that better correlate with $\boldsymbol{\theta}_\star$ are retained
- Shrink remaining directions

In practice, cross-validation...