Model/estimator selection

Sylvain Arlot[1,2,3]

[1]Université Paris-Saclay

[2]Inria Saclay, Celeste project-team

[3]Institut Universitaire de France

CEA/EDF/INRIA Summer School 2021
June 18, 2021

Main references: arXiv:1901.07277 (parts 1 & 2)
arXiv:0907.4728 or hal-01485508 (part 3)

Outline

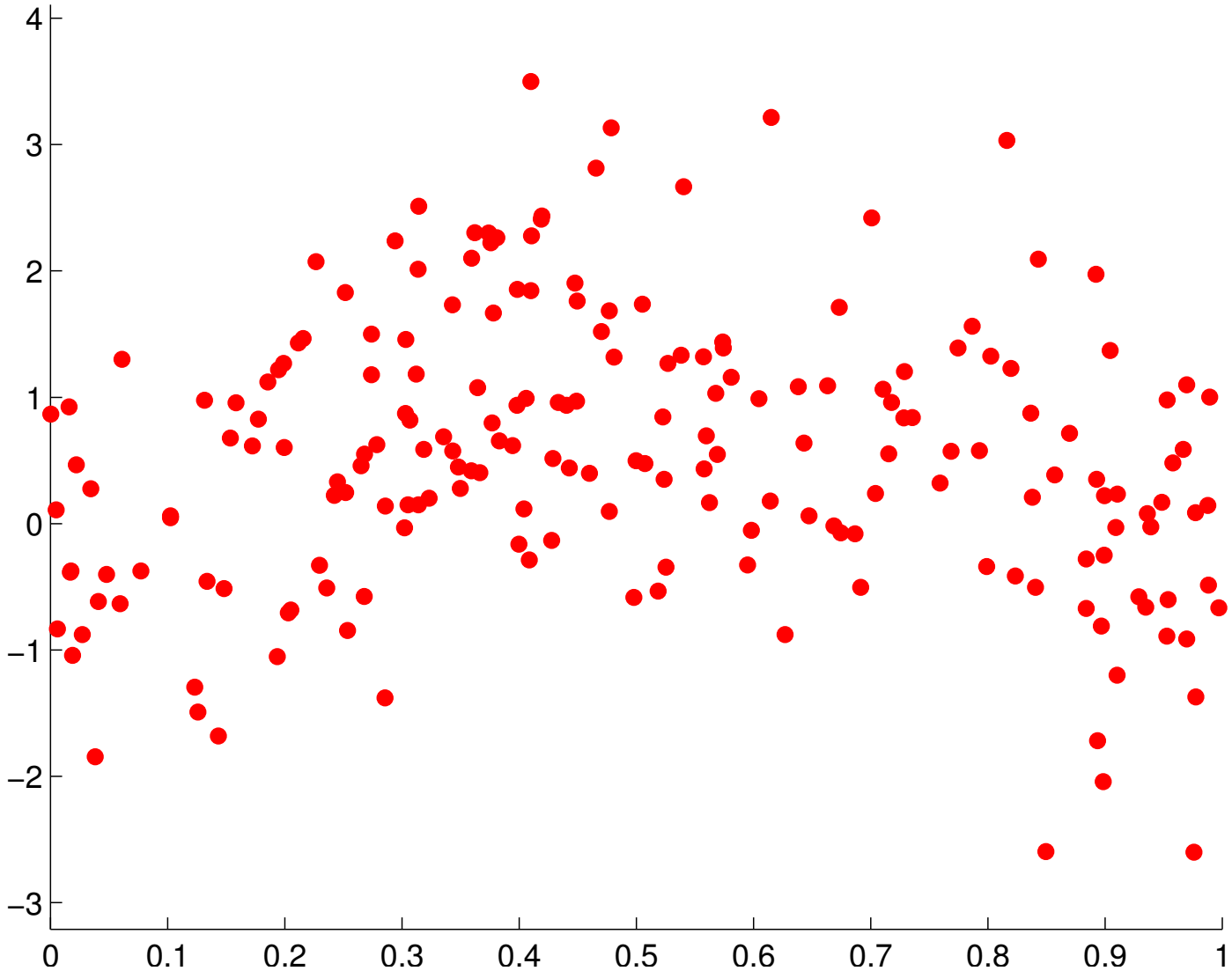
- Part I: Model selection for fixed-design regression
- Part II: Model/estimator selection in the prediction setting
- Part III: Cross-validation for estimator selection/aggregation

Part I

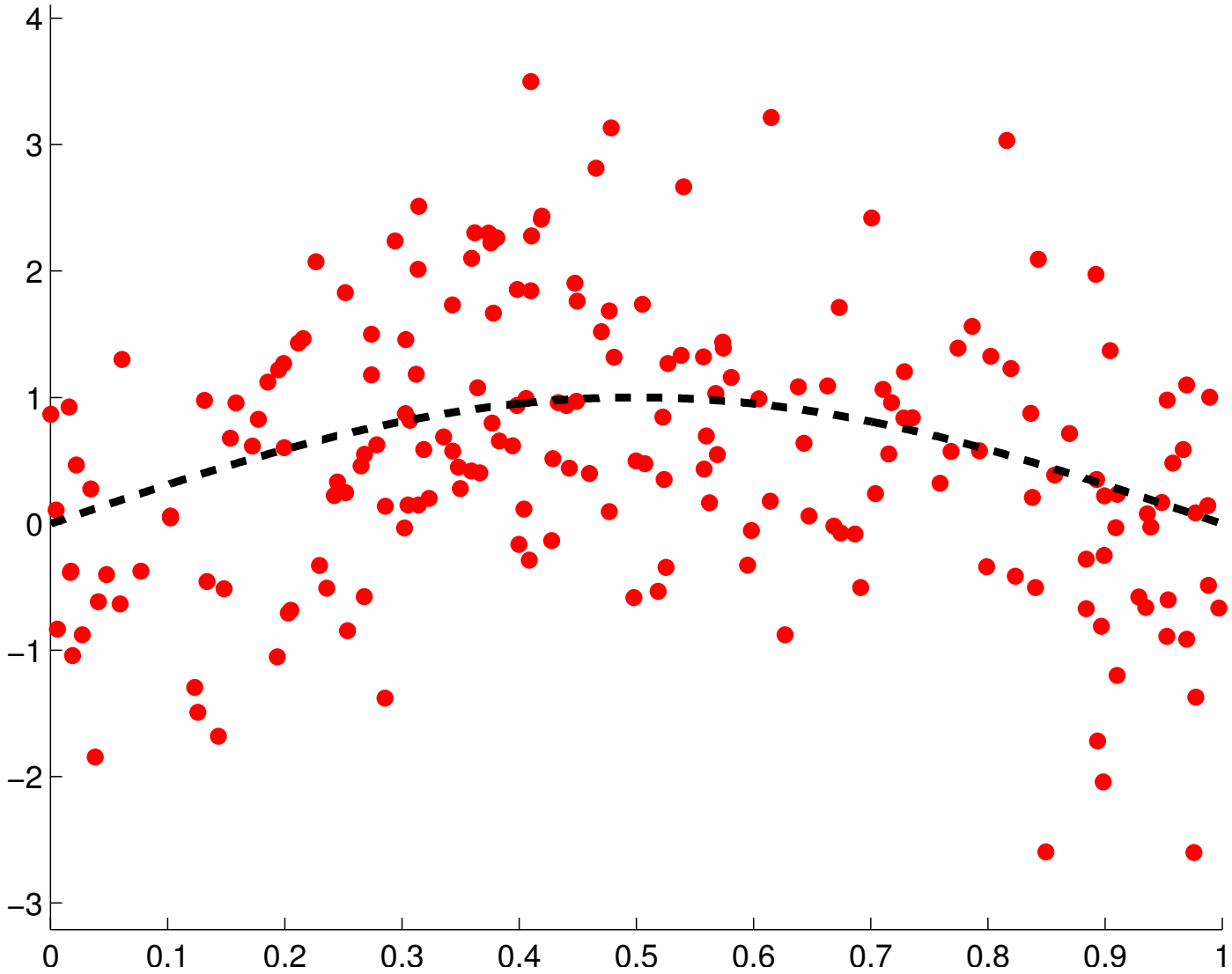
Model selection for fixed-design regression

Outline

Regression: data $(X_1, Y_1), \ldots, (X_n, Y_n)$



Goal: find the signal (denoising)



Statistical framework: regression, least-squares risk

- Observations:

$$Y_i = f^\star(x_i) + \varepsilon_i \in \mathbb{R}$$

with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$, $f^\star$ and σ^2 unknown

- Fixed design: $x_i \in \mathcal{X}$ deterministic
- Notation: $Y = F + \varepsilon$ with

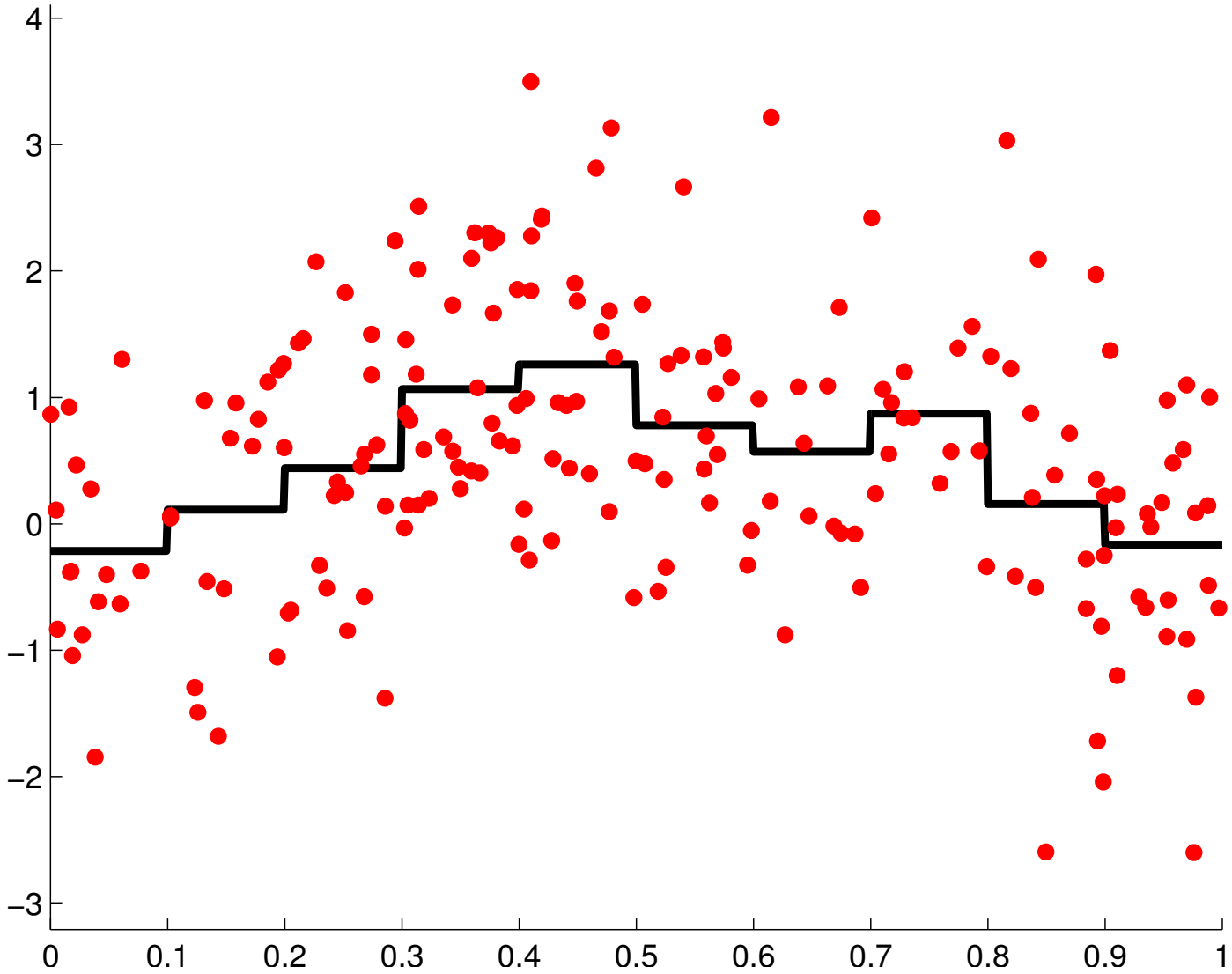
$$Y = (Y_i)_{1 \leqslant i \leqslant n}, \quad F = \left(f^\star(x_i)\right)_{1 \leqslant i \leqslant n}, \quad \varepsilon = (\varepsilon_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$$

- Least-squares risk of a predictor $t \in \mathbb{R}^n$ ("$t_i = t(x_i)$"):

$$\frac{1}{n} \|t - F\|^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - F_i)^2$$

$\Rightarrow$ Estimator $\widehat{F}(Y) \in \mathbb{R}^n$?

Estimators: example: regressogram



Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n} \|t - F\|^2$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n}\|t-F\|^2$
- Least-squares criterion:

$$\frac{1}{n}\|t-Y\|^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i-Y_i)^2$$

$$\forall t \in \mathbb{R}^n \ , \qquad \mathbb{E}\left[\frac{1}{n}\|t-Y\|^2\right] = \frac{1}{n}\|t-F\|^2 + \frac{1}{n}\mathbb{E}\left[\|\varepsilon\|^2\right]$$

Least-squares estimators

- Natural idea: minimize an estimator of the risk $\frac{1}{n}\|t-F\|^2$
- Least-squares criterion:

$$\frac{1}{n}\|t-Y\|^2 = \frac{1}{n}\sum_{i=1}^{n}(t_i-Y_i)^2$$

$$\forall t \in \mathbb{R}^n \ , \qquad \mathbb{E}\left[\frac{1}{n}\|t-Y\|^2\right] = \frac{1}{n}\|t-F\|^2 + \frac{1}{n}\mathbb{E}\left[\|\varepsilon\|^2\right]$$

- Model: $S \subset \mathbb{R}^n \Rightarrow$ Least-squares estimator on S:

$$\widehat{F}_S \in \operatorname*{argmin}_{t\in S}\left\{\frac{1}{n}\|t-Y\|^2\right\} = \operatorname*{argmin}_{t\in S}\left\{\frac{1}{n}\sum_{i=1}^{n}(t_i-Y_i)^2\right\}$$

so that

$$\widehat{F}_S = \Pi_S(Y) \qquad \text{(orthogonal projection)}$$

Model examples

- histograms on some partition m of $\mathcal{X}$
  $\Rightarrow S_m = \text{vect}\left\{(\mathbb{1}_{x_i \in \lambda})_{1 \leqslant i \leqslant n} \,/\, \lambda \in m\right\}$
  $\Rightarrow$ the least-squares estimator (regressogram) can be written

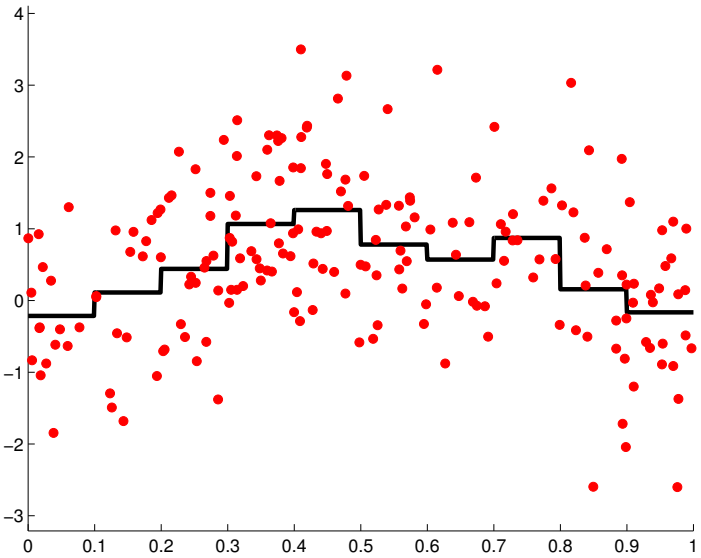
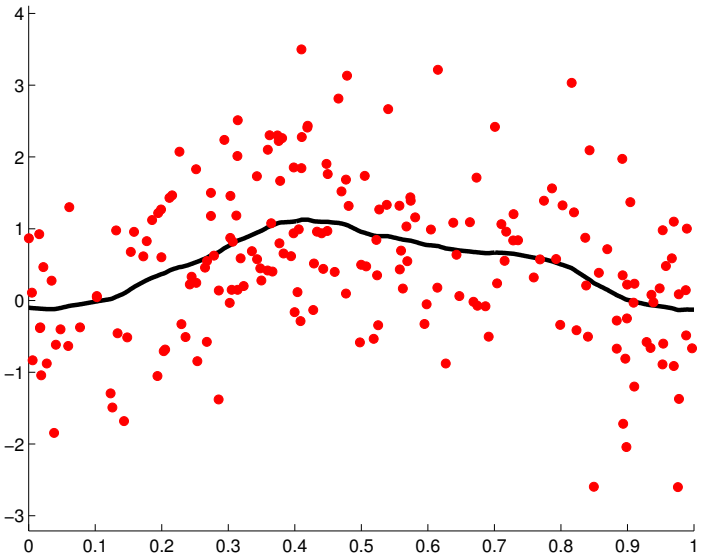
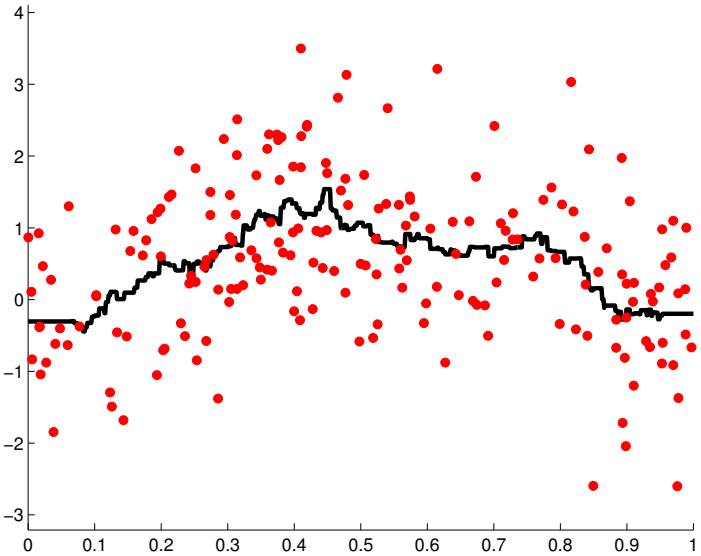
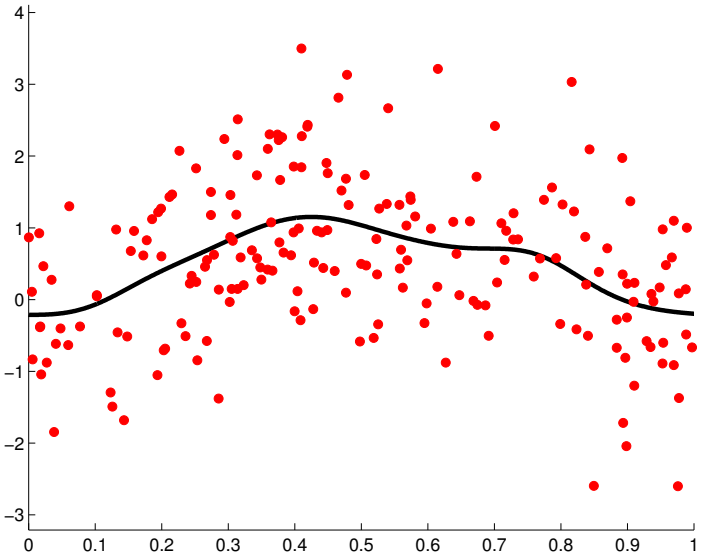
$$\widehat{F}_{S_m}(x_i) = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_{x_i \in \lambda} \qquad \widehat{\beta}_\lambda = \frac{1}{\text{Card}\left\{x_i \in \lambda\right\}} \sum_{x_i \in \lambda} Y_i$$

- variable selection: $x_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

$$\forall m \subset \{1, \ldots, p\} \quad , \quad S_m = \text{vect}\left\{x^{(j)} \text{ s.t. } j \in m\right\}$$

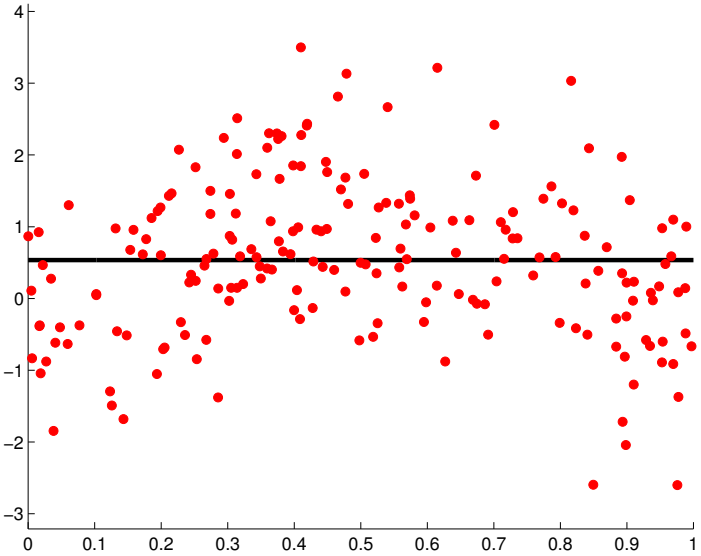
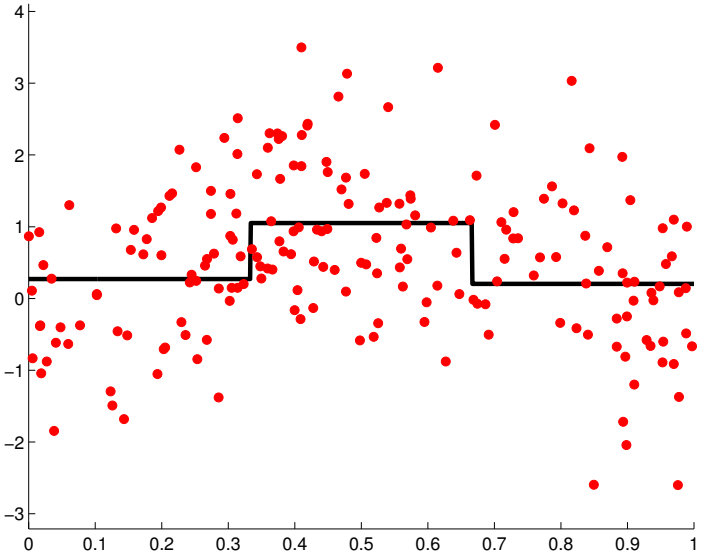
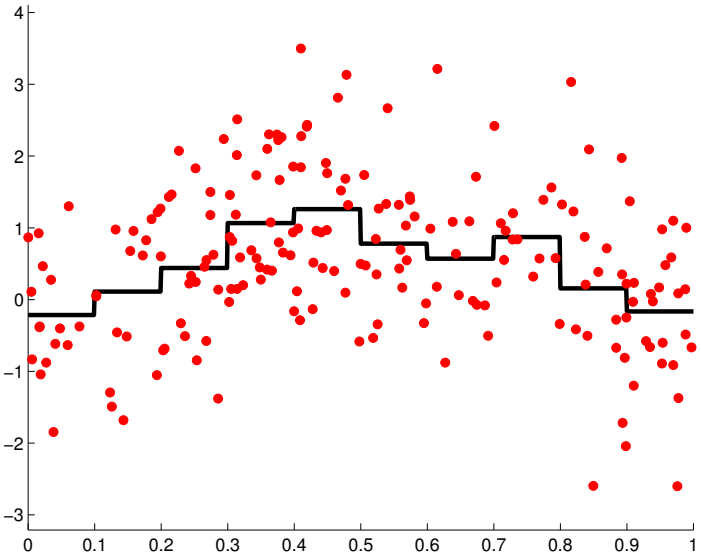
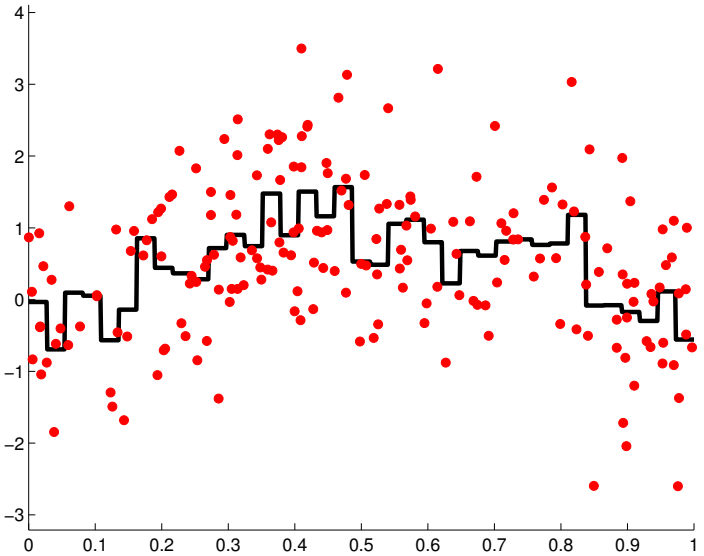
- S_m subspace generated by a subset of an orthogonal basis of $\mathbb{R}^n$ (Fourier, wavelets, ...)

Estimators: regressogram, ridge, k-NN, Nadaraya-Watson

Outline

Model selection: regular regressograms, choose D?

Model selection

- Model collection $(S_m)_{m\in\mathcal{M}} \Rightarrow (\widehat{F}_m)_{m\in\mathcal{M}} \Rightarrow \widehat{m}(Y)$?

$$\widehat{F}_m = \Pi_m Y = \Pi_{S_m} Y$$

- Goal: minimize the risk, i.e.,
  Oracle inequality (in expectation or with a large probability):

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant C \inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + R_n$$

Approximation and estimation error

Fixed model S_m, linear subspace of $\mathbb{R}^n$, dimension $D_m = \dim(S_m)$.

- Approximation error of S_m:

$$\inf_{t \in S_m} \frac{1}{n} \|t - F\|^2 = \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|(I_n - \Pi_m) F\|^2$$

where $F_m = \Pi_m F$ orthogonal projection of F onto S_m

- $\widehat{F}_m \in S_m \Rightarrow \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \geqslant \frac{1}{n} \|F_m - F\|^2$

Approximation and estimation error

Fixed model S_m, linear subspace of $\mathbb{R}^n$, dimension $D_m = \dim(S_m)$.

- Approximation error of S_m:

$$\inf_{t \in S_m} \frac{1}{n} \|t - F\|^2 = \frac{1}{n} \|F_m - F\|^2 = \frac{1}{n} \|(I_n - \Pi_m) F\|^2$$

where $F_m = \Pi_m F$ orthogonal projection of F onto S_m

- $\widehat{F}_m \in S_m \Rightarrow \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \geqslant \frac{1}{n} \|F_m - F\|^2$
- Estimation error of S_m:

$$\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \|F_m - F\|^2 \geqslant 0$$

Expectation of the estimation error

$$\left\|\widehat{F}_m - F\right\|^2 = \|\Pi_m(F+\varepsilon) - F\|^2$$
$$= \|\Pi_m F - F\|^2 + 2\underbrace{\langle \Pi_m F - F, \Pi_m \varepsilon \rangle}_{=0} + \|\Pi_m \varepsilon\|^2$$

Expectation of the estimation error

$$\left\|\widehat{F}_m - F\right\|^2 = \|\Pi_m(F+\varepsilon) - F\|^2$$
$$= \|\Pi_m F - F\|^2 + 2\underbrace{\langle \Pi_m F - F,\, \Pi_m \varepsilon\rangle}_{=0} + \|\Pi_m \varepsilon\|^2$$

$\Rightarrow$ Estimation error (of $\widehat{F}_m$):

$$\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\|F_m - F\|^2 = \frac{1}{n}\|\Pi_m \varepsilon\|^2 = \frac{1}{n}\langle \Pi_m \varepsilon,\, \varepsilon\rangle$$

Expectation of the estimation error

$$\left\|\widehat{F}_m - F\right\|^2 = \|\Pi_m(F+\varepsilon) - F\|^2$$
$$= \|\Pi_m F - F\|^2 + 2\underbrace{\langle \Pi_m F - F,\, \Pi_m \varepsilon\rangle}_{=0} + \|\Pi_m \varepsilon\|^2$$

$\Rightarrow$ Estimation error (of $\widehat{F}_m$):

$$\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\|F_m - F\|^2 = \frac{1}{n}\|\Pi_m \varepsilon\|^2 = \frac{1}{n}\langle \Pi_m \varepsilon,\, \varepsilon\rangle$$

$\Rightarrow$ Expectation of the estimation error (of $\widehat{F}_m$):

$$\frac{1}{n}\mathbb{E}\left[\langle \Pi_m \varepsilon,\, \varepsilon\rangle\right] = \frac{\sigma^2 \operatorname{tr}(\Pi_m)}{n} = \frac{\sigma^2 D_m}{n}$$

Bias-variance trade-off

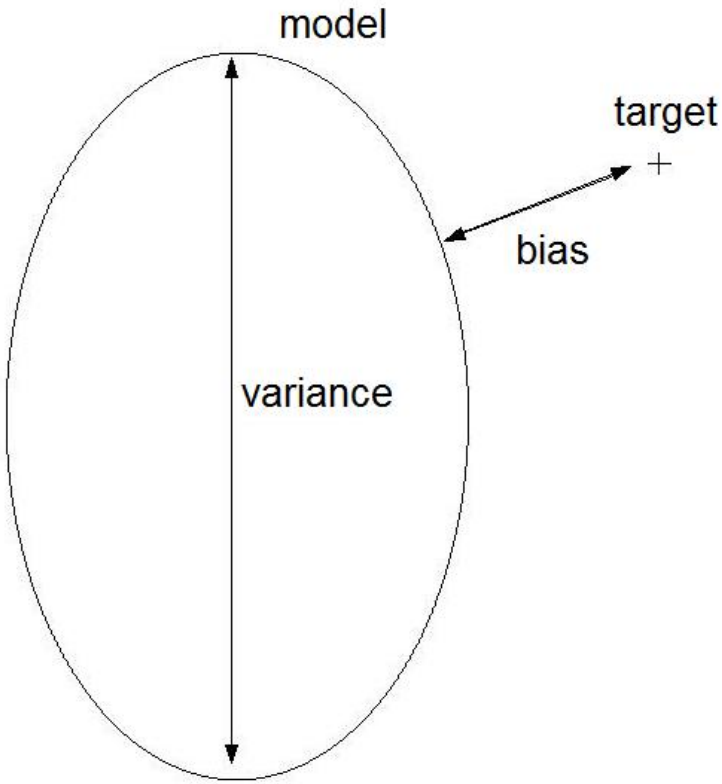
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|F_m - F\right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error or Bias:

$$\frac{1}{n}\left\|F_m - F\right\|^2 = \frac{1}{n}\left\|\Pi_m F - F\right\|^2$$

$\mathbb{E}$[Estimation error] or Variance:

$$\frac{\sigma^2 D_m}{n}$$



Bias-variance trade-off

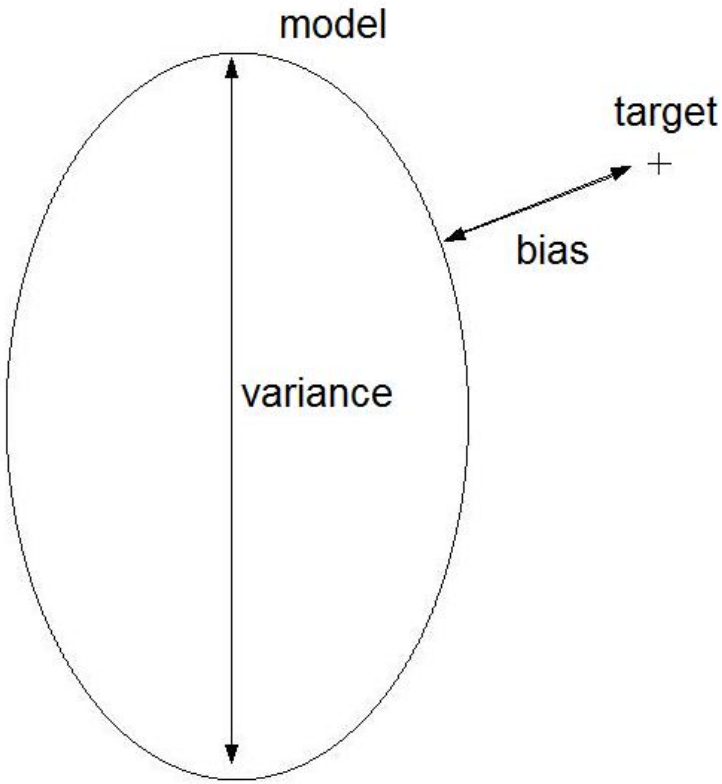
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|F_m - F\right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error or Bias:

$$\frac{1}{n}\left\|F_m - F\right\|^2 = \frac{1}{n}\left\|\Pi_m F - F\right\|^2$$

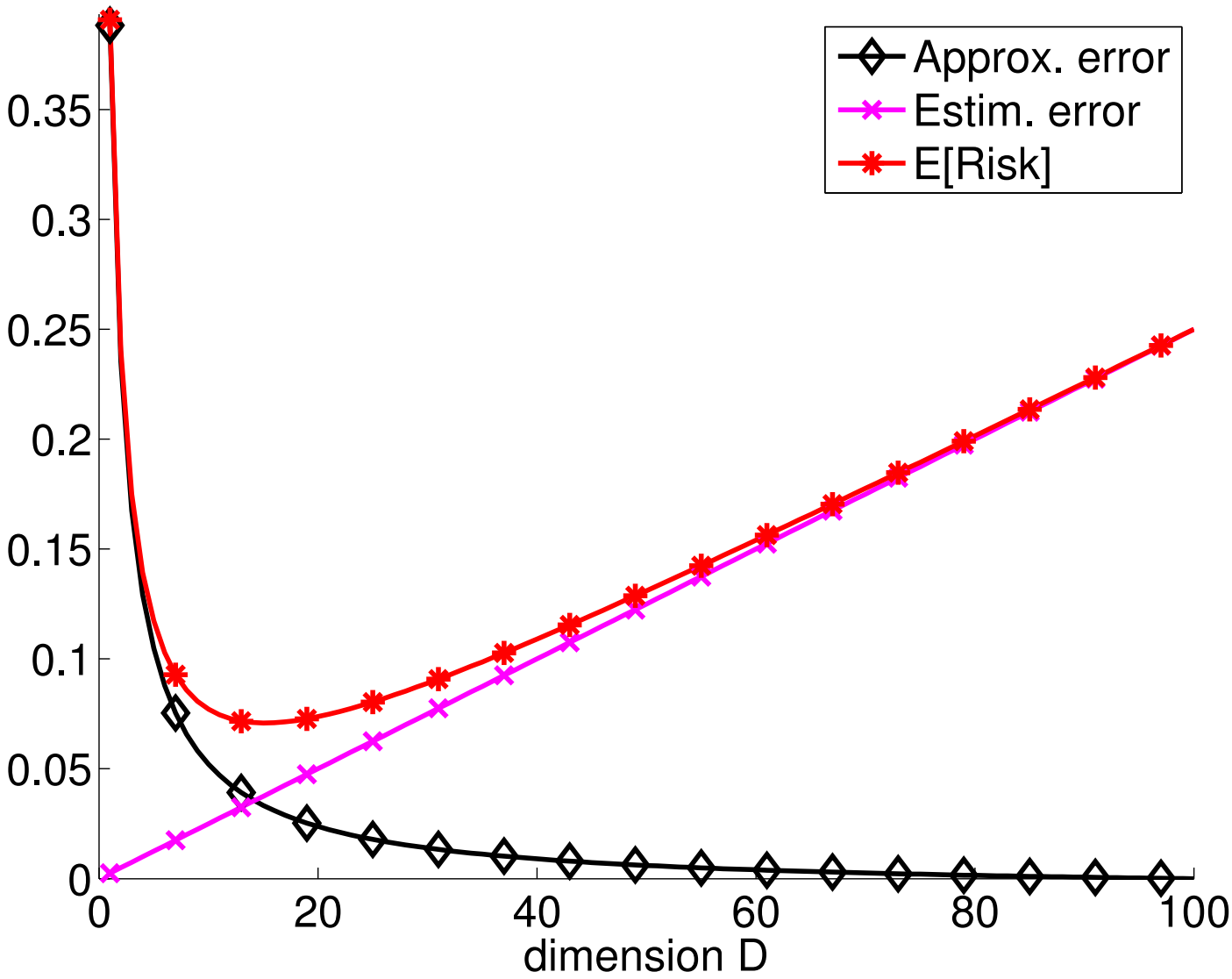
$\mathbb{E}$[Estimation error] or Variance:

$$\frac{\sigma^2 D_m}{n}$$

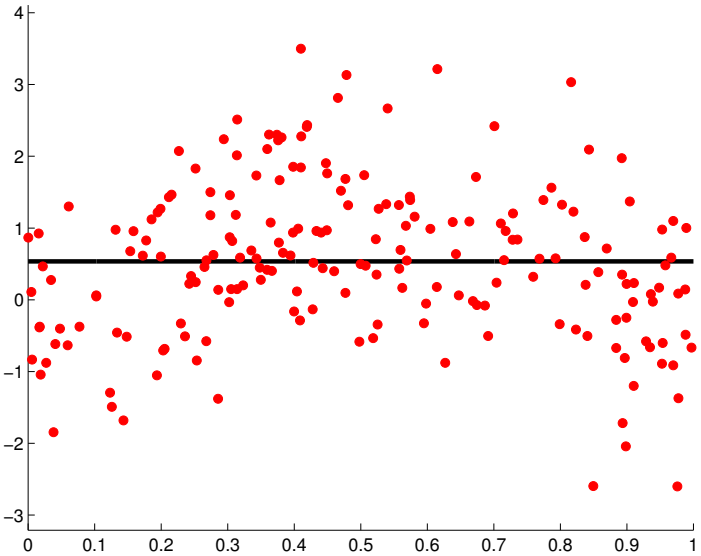


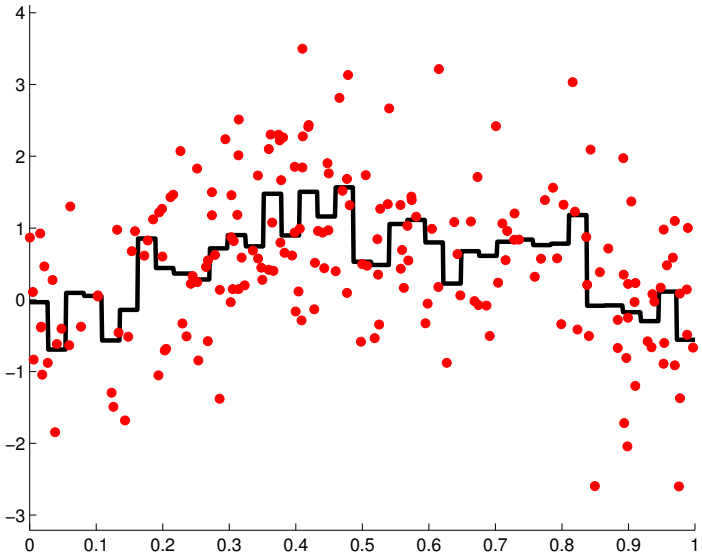
Bias-variance trade-off
$\Leftrightarrow$ avoid overfitting and underfitting

Bias-variance trade-off

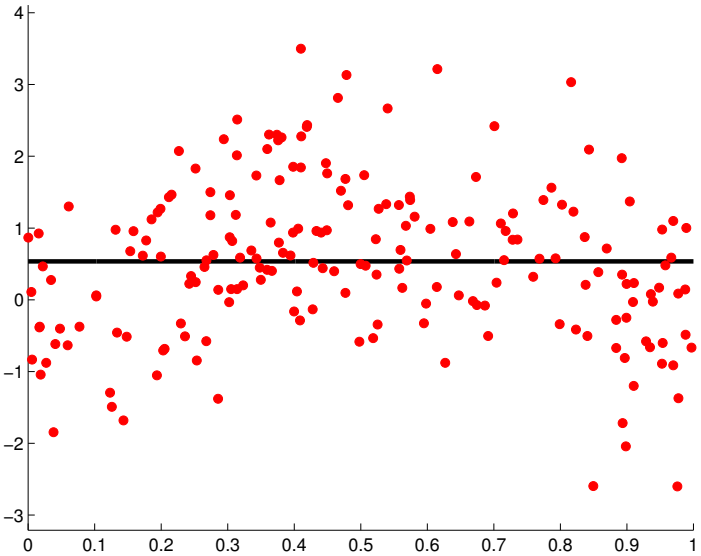


Overfitting/underfitting: regressograms

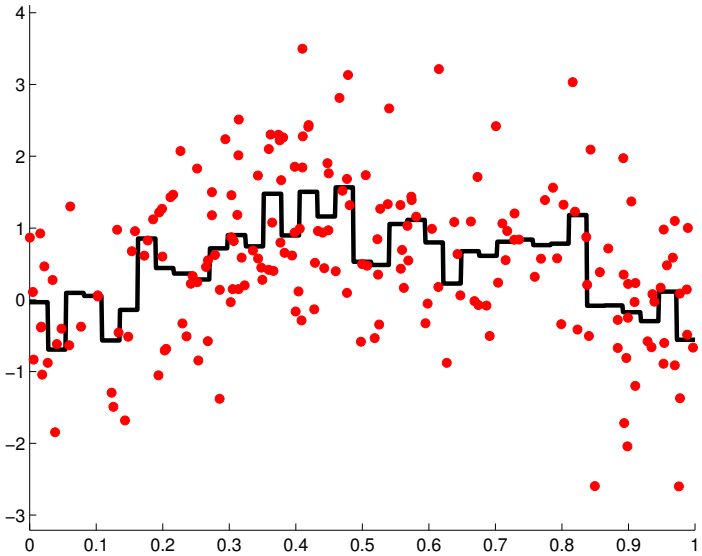




Overfitting/underfitting: regressograms

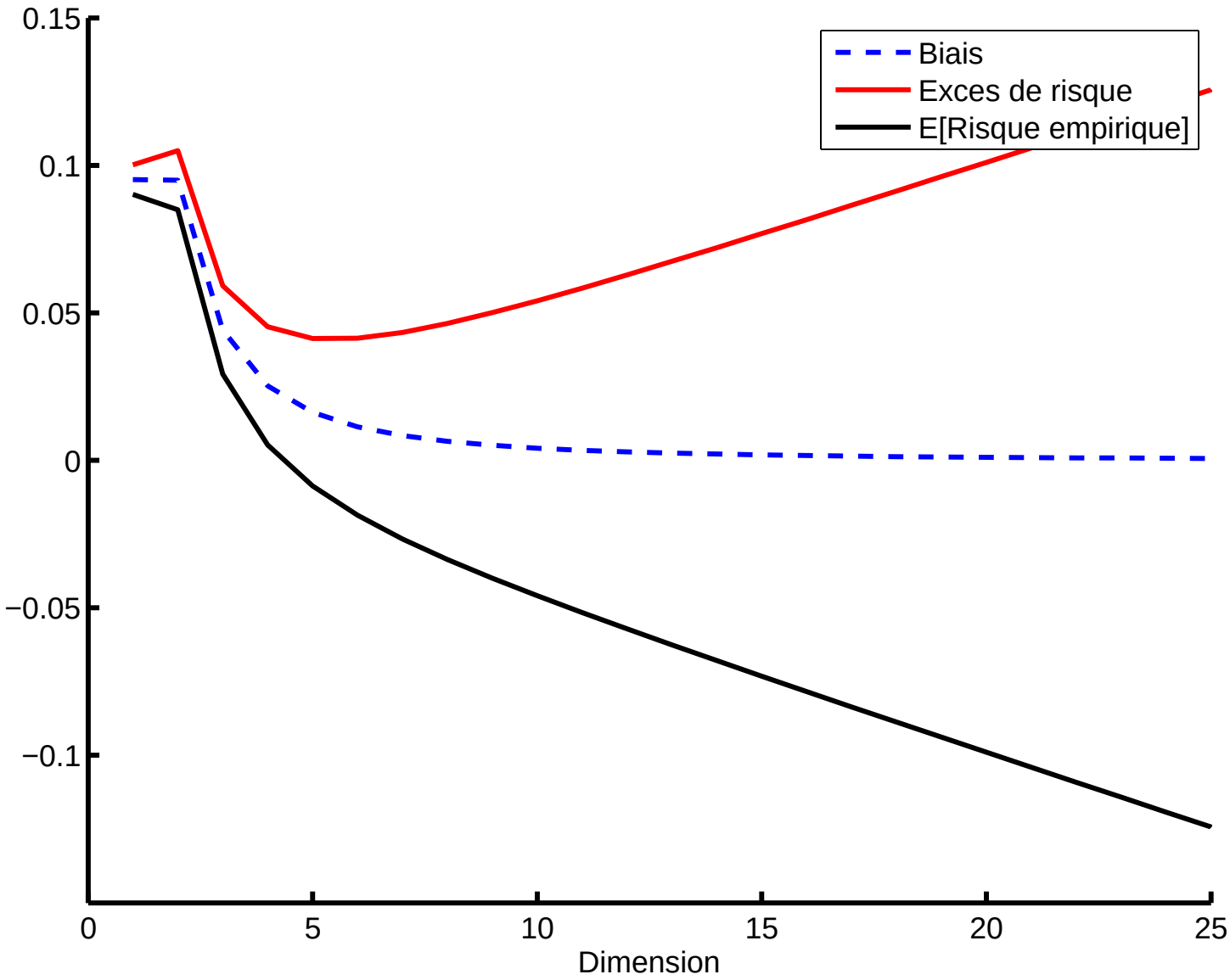


Underfitting
$D = 1$ (too small)



Overfitting
$D = 37$ (too large)

Why should the empirical risk be penalized?



Penalization

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$$

Penalization

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$$

- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m) := \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 - \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 = \text{Risk} - \text{Empirical risk}$$

- Mallows' heuristic: $\mathrm{pen}(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$
  $\Rightarrow$ oracle inequality if $\mathrm{Card}(\mathcal{M})$ not too large
  $(+$ concentration inequalities$)$

Ideal penalty and its expectation

$$\begin{aligned}
\left\|\widehat{F}_m - Y\right\|^2 &= \left\|\widehat{F}_m - F - \varepsilon\right\|^2 \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \widehat{F}_m - F, \varepsilon \right\rangle \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F + \Pi_m \varepsilon - F, \varepsilon \right\rangle \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F - F, \varepsilon \right\rangle - 2\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle
\end{aligned}$$

Ideal penalty and its expectation

$$\begin{aligned}
\left\|\widehat{F}_m - Y\right\|^2 &= \left\|\widehat{F}_m - F - \varepsilon\right\|^2 \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \widehat{F}_m - F, \varepsilon \right\rangle \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F + \Pi_m \varepsilon - F, \varepsilon \right\rangle \\
&= \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F - F, \varepsilon \right\rangle - 2\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle
\end{aligned}$$

$\Rightarrow$ Ideal penalty

$$\begin{aligned}
\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\
&= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle - \frac{1}{n}\|\varepsilon\|^2
\end{aligned}$$

Ideal penalty and its expectation

$\Rightarrow$ Ideal penalty

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\ &= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon\right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon\right\rangle - \frac{1}{n}\left\|\varepsilon\right\|^2\end{aligned}$$

$\Rightarrow$ Expectation of the ideal penalty

$$\begin{aligned}\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] &= \frac{2}{n}\mathbb{E}\left[\left\langle \Pi_m F - F, \varepsilon\right\rangle\right] + \frac{2}{n}\mathbb{E}\left[\left\langle \Pi_m \varepsilon, \varepsilon\right\rangle\right] - \frac{1}{n}\mathbb{E}\left[\left\|\varepsilon\right\|^2\right] \\ &= \underbrace{\frac{2\sigma^2 D_m}{n}}_{\text{Mallows' } C_p} - \sigma^2\end{aligned}$$

Towards theoretical guarantees: a key lemma

Lemma

Let crit : $\mathcal{M} \to \mathbb{R}$ *be any function* (*possibly data-dependent*).
On the event Ω *on which,* $\forall m, m' \in \mathcal{M}$

$$\left[\mathrm{crit}(m) - \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] - \left[\mathrm{crit}(m') - \frac{1}{n}\left\|\widehat{F}_{m'} - F\right\|^2\right] \leqslant A(m) + B(m')\,,$$

$$\textit{we have} \quad \forall \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{\mathrm{crit}(m)\right\},$$

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m)\right\}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2$$

$$= \mathrm{crit}(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m})$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2$$

$$= \mathrm{crit}(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m})$$

$$\leqslant \mathrm{crit}(m) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m})$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$
\begin{aligned}
&\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \\
&= \mathrm{crit}(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m}) \\
&\leqslant \mathrm{crit}(m) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m}) \\
&= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \mathrm{crit}(m) - \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \mathrm{crit}(\widehat{m})
\end{aligned}
$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$
\begin{aligned}
&\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \\
&= \operatorname{crit}(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&\leqslant \operatorname{crit}(m) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \operatorname{crit}(m) - \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&\leqslant \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m) + B(\widehat{m})
\end{aligned}
$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$
\begin{aligned}
&\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \\
&= \operatorname{crit}(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&\leqslant \operatorname{crit}(m) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \operatorname{crit}(m) - \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - \operatorname{crit}(\widehat{m}) \\
&\leqslant \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m) + B(\widehat{m})
\end{aligned}
$$

hence

$$
\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m) \right\}. \quad \square
$$

Key lemma (reformulated)

Lemma

Let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}$ *be any penalty* (*possibly data-dependent*).
On the event Ω *on which,* $\forall m, m' \in \mathcal{M}$

$$\left[\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m)\right] - \left[\mathrm{pen}(m') - \mathrm{pen}_{\mathrm{id}}(m')\right] \leqslant A(m) + B(m') \,,$$

$$\textit{we have} \quad \forall \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

Proof: take $\mathrm{crit}(m) = \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m)$. □

How to use the key lemma here?

1. Take $\mathrm{pen}(m) = \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$ (up to a translation).
2. Prove that $\forall m \in \mathcal{M}$, $\mathrm{pen}_{\mathrm{id}}(m)$ concentrates around its expectation.
3. Union bound over $m \in \mathcal{M}$ (if $\mathrm{Card}(\mathcal{M})$ "small").
4. Apply the key lemma with $A(m) \propto$ deviations of $\mathrm{pen}_{\mathrm{id}}(m)$.

Ideal penalty (reminder)

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\ &= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon\right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon\right\rangle - \frac{1}{n}\|\varepsilon\|^2\end{aligned}$$

Ideal penalty (reminder)

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\ &= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle - \frac{1}{n}\left\|\varepsilon\right\|^2\end{aligned}$$

- Linear term:

  $\frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle \quad \Rightarrow \quad \text{mean } = 0\,, \quad \text{Gaussian distribution}$

Ideal penalty (reminder)

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\ &= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle - \frac{1}{n}\|\varepsilon\|^2\end{aligned}$$

- Linear term:

  $$\frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle \quad \Rightarrow \quad \text{mean } = 0\,, \quad \text{Gaussian distribution}$$

- Quadratic term:

  $$\frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle \quad \Rightarrow \quad \text{mean } = \frac{2\sigma^2 D_m}{n}\,, \quad \chi^2 \text{ distribution}$$

Ideal penalty (reminder)

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 \\ &= \frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle + \frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle - \frac{1}{n}\left\|\varepsilon\right\|^2\end{aligned}$$

- Linear term:

  $$\frac{2}{n}\left\langle \Pi_m F - F, \varepsilon \right\rangle \quad \Rightarrow \quad \text{mean } = 0\,, \quad \text{Gaussian distribution}$$

- Quadratic term:

  $$\frac{2}{n}\left\langle \Pi_m \varepsilon, \varepsilon \right\rangle \quad \Rightarrow \quad \text{mean } = \frac{2\sigma^2 D_m}{n}\,, \quad \chi^2 \text{ distribution}$$

- Constant term:

  $$\frac{1}{n}\left\|\varepsilon\right\|^2 \quad \Rightarrow \quad \text{can be discarded}$$

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $\alpha \in \mathbb{R}^n$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(|\langle \varepsilon, \alpha \rangle| \leqslant \sigma \sqrt{2x}\, \|\alpha\| \right) \geqslant 1 - 2e^{-x} .$$

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $\alpha \in \mathbb{R}^n$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(|\langle \varepsilon, \alpha \rangle| \leqslant \sigma \sqrt{2x}\, \|\alpha\| \right) \geqslant 1 - 2e^{-x} .$$

$\Rightarrow$ with probability $\geqslant 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} |\langle \Pi_m F - F, \varepsilon \rangle| \leqslant \frac{2\sigma\sqrt{2x}}{n} \|\Pi_m F - F\| \leqslant \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n}$$

since $2ab \leqslant \theta a^2 + \theta^{-1} b^2 \ \forall a, b \geqslant 0, \theta > 0$.

Concentration of the ideal penalty (1): linear term

Proposition (Gaussian concentration)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $\alpha \in \mathbb{R}^n$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(|\langle \varepsilon, \alpha \rangle| \leqslant \sigma \sqrt{2x}\, \|\alpha\| \right) \geqslant 1 - 2e^{-x}.$$

$\Rightarrow$ with probability $\geqslant 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} |\langle \Pi_m F - F, \varepsilon \rangle| \leqslant \frac{2\sigma\sqrt{2x}}{n} \|\Pi_m F - F\| \leqslant \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n}$$

since $2ab \leqslant \theta a^2 + \theta^{-1} b^2 \ \forall a, b \geqslant 0, \theta > 0$.

- Can be generalized to sub-Gaussian noise, see arXiv:1901.07277 (Remark 1).

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $M \in \mathcal{M}_n(\mathbb{R})$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(\left| \langle \varepsilon, M\varepsilon \rangle - \sigma^2 \operatorname{tr}(M) \right| \leqslant 2\sigma^2 \sqrt{x \operatorname{tr}\left(M^\top M\right)} + 2\sigma^2 \left\| M \right\| x \right) \geqslant 1 - 2e^{-x} .$$

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $M \in \mathcal{M}_n(\mathbb{R})$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(\left| \langle \varepsilon, M\varepsilon \rangle - \sigma^2 \operatorname{tr}(M) \right| \leqslant 2\sigma^2 \sqrt{x \operatorname{tr}\left(M^\top M\right) + 2\sigma^2 \left\| M \right\| x} \right) \geqslant 1 - 2e^{-x}.$$

$\Rightarrow$ with probability $\geqslant 1 - 2e^{-x}$, for every $\theta > 0$,

$$\frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| \leqslant \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n}$$

.

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $M \in \mathcal{M}_n(\mathbb{R})$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(\left| \langle \varepsilon, M\varepsilon \rangle - \sigma^2 \operatorname{tr}(M) \right| \leqslant 2\sigma^2 \sqrt{x \operatorname{tr}\left(M^\top M\right)} + 2\sigma^2 \left\| M \right\| x \right) \geqslant 1 - 2e^{-x} .$$

$\Rightarrow$ with probability $\geqslant 1 - 2e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} \frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| &\leqslant \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n} \\ &\leqslant \frac{\theta \sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta} \right) \frac{\sigma^2 x}{n} . \end{aligned}$$

Concentration of the ideal penalty (2): quadratic term

Proposition (see A. & Bach 2011 (arXiv:0909.1884), Proposition 6)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ *and* $M \in \mathcal{M}_n(\mathbb{R})$, *for every* $x \geqslant 0$,

$$\mathbb{P}\left(\left| \langle \varepsilon, M\varepsilon \rangle - \sigma^2 \operatorname{tr}(M) \right| \leqslant 2\sigma^2 \sqrt{x \operatorname{tr}\left(M^\top M\right)} + 2\sigma^2 \left\| M \right\| x \right) \geqslant 1 - 2e^{-x} .$$

$\Rightarrow$ with probability $\geqslant 1 - 2e^{-x}$, for every $\theta > 0$,

$$\begin{aligned} \frac{2}{n} \left| \langle \Pi_m \varepsilon, \varepsilon \rangle - \sigma^2 D_m \right| &\leqslant \frac{4\sigma^2}{n} \sqrt{x D_m} + \frac{4\sigma^2 x}{n} \\ &\leqslant \frac{\theta \sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n} . \end{aligned}$$

- Can be generalized to sub-Gaussian noise, see arXiv:1901.07277 (Remark 1).

Concentration of the ideal penalty: summary

- With probability $\geqslant 1 - 4e^{-x}$, for every $\theta > 0$,

$$\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] \right|$$
$$\leqslant \frac{\theta}{n} \|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta \sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right) \frac{\sigma^2 x}{n}$$

Concentration of the ideal penalty: summary

- With probability $\geqslant 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned}
&\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2\right] \right| \\
&\leqslant \frac{\theta}{n}\|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right)\frac{\sigma^2 x}{n} \\
&= \theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \left(6 + \frac{4}{\theta}\right)\frac{x\sigma^2}{\theta n} \\
&=: A(m) = B(m)\,.
\end{aligned}$$

Concentration of the ideal penalty: summary

- With probability $\geqslant 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned}
&\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2\right] \right| \\
&\leqslant \frac{\theta}{n}\|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right)\frac{\sigma^2 x}{n} \\
&= \theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \left(6 + \frac{4}{\theta}\right)\frac{x\sigma^2}{\theta n} \\
&=: A(m) = B(m)\,.
\end{aligned}$$

$\Rightarrow$ end of step 2 (concentration of $\mathrm{pen}_{\mathrm{id}}(m)$)

Concentration of the ideal penalty: summary

- With probability $\geqslant 1 - 4e^{-x}$, for every $\theta > 0$,

$$\begin{aligned}
&\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2\right] \right| \\
&\leqslant \frac{\theta}{n}\|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right)\frac{\sigma^2 x}{n} \\
&= \theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \left(6 + \frac{4}{\theta}\right)\frac{x\sigma^2}{\theta n} \\
&=: A(m) = B(m)\,.
\end{aligned}$$

- Step 3: union bound $\Rightarrow$ with probability $\geqslant 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, $\forall\theta > 0$, $\forall m, m' \in \mathcal{M}$,

$$\begin{aligned}
&\left[\frac{2\sigma^2 D_m}{n} - \mathrm{pen}_{\mathrm{id}}(m)\right] - \left[\frac{2\sigma^2 D_{m'}}{n} - \mathrm{pen}_{\mathrm{id}}(m')\right] \\
&\qquad \leqslant A(m) + B(m')\,.
\end{aligned}$$

Application of the key lemma (step 4)

- With probability $\geqslant 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, $\forall \theta > 0$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$(1-\theta)\mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \right] - \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{\theta n}$$
$$\leqslant (1+\theta) \inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] + \left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{\theta n},$$

Application of the key lemma (step 4)

- With probability $\geqslant 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, $\forall \theta > 0$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$(1-\theta)\mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \right]$$
$$\leqslant (1+\theta) \inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] + 2\left(6 + \frac{4}{\theta}\right) \frac{x\sigma^2}{\theta n},$$

Application of the key lemma (step 4)

- With probability $\geqslant 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, $\forall \theta \in (0,1)$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \frac{2\sigma^2 D_m}{n} \right\},$$

$$\mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \right]$$
$$\leqslant \frac{1+\theta}{1-\theta} \inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right] + \frac{2}{1-\theta} \left(6 + \frac{4}{\theta} \right) \frac{x\sigma^2}{\theta n}.$$

Conclusion: oracle inequality for C_p

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n}$, *i.i.d. Gaussian noise.*
Then, for every $x \geqslant 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, *for every* $\theta \in (0, 1/3)$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant (1+3\theta) \underbrace{\inf_{m\in\mathcal{M}} \left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\}}_{\textit{oracle risk}} + \frac{L(\theta)\sigma^2 x}{n}\,.$$

Conclusion: oracle inequality for C_p

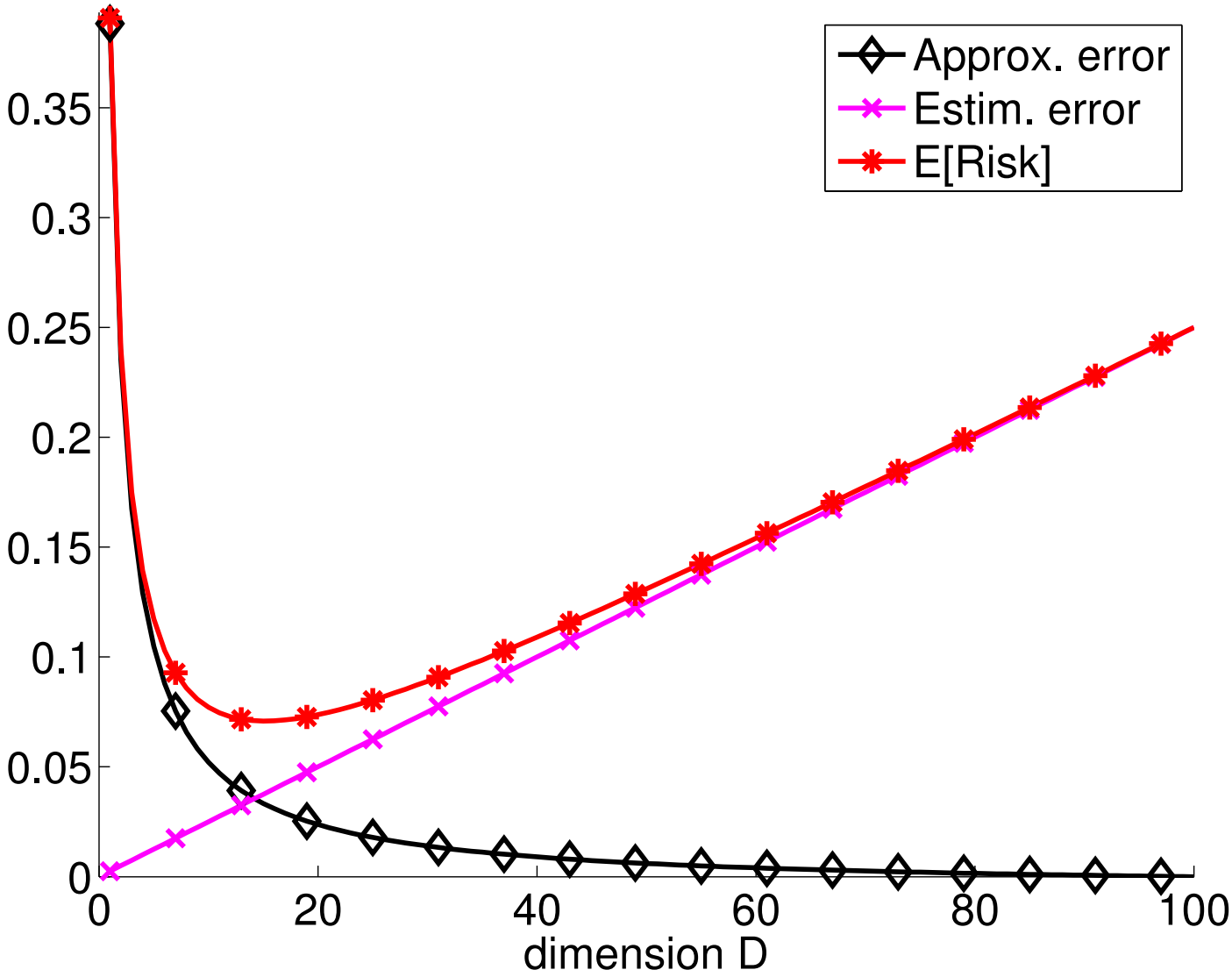
Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\mathrm{pen}(m) = \frac{2\sigma^2 D_m}{n}$, *i.i.d. Gaussian noise.*
Then, for every $x \geqslant 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, *for every* $\theta \in (0, 1/3)$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant (1+3\theta) \underbrace{\inf_{m\in\mathcal{M}} \left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\}}_{\textit{oracle risk}} + \frac{L(\theta)\sigma^2 x}{n}\,.$$

Generalization (arXiv:1901.07277): sub-Gaussian noise.

Meaning of an oracle inequality



Outline

Practical and theoretical questions about C_p

Mallows' C_p penalty:

$$\text{pen}(m) = \frac{2\sigma^2 D_m}{n}$$

- σ^2 unknown in general

$$\Rightarrow \qquad \text{pen}(m) = \frac{C D_m}{n}$$

with C (hopefully) close to $2\sigma^2$.

Practical and theoretical questions about C_p

Mallows' C_p penalty:

$$\text{pen}(m) = \frac{2\sigma^2 D_m}{n}$$

- σ^2 unknown in general

$$\Rightarrow \qquad \text{pen}(m) = \frac{C D_m}{n}$$

with C (hopefully) close to $2\sigma^2$.

- Oracle inequality for any $C > 0$? Minimal value for C?

Practical and theoretical questions about C_p

Mallows' C_p penalty:

$$\text{pen}(m) = \frac{2\sigma^2 D_m}{n}$$

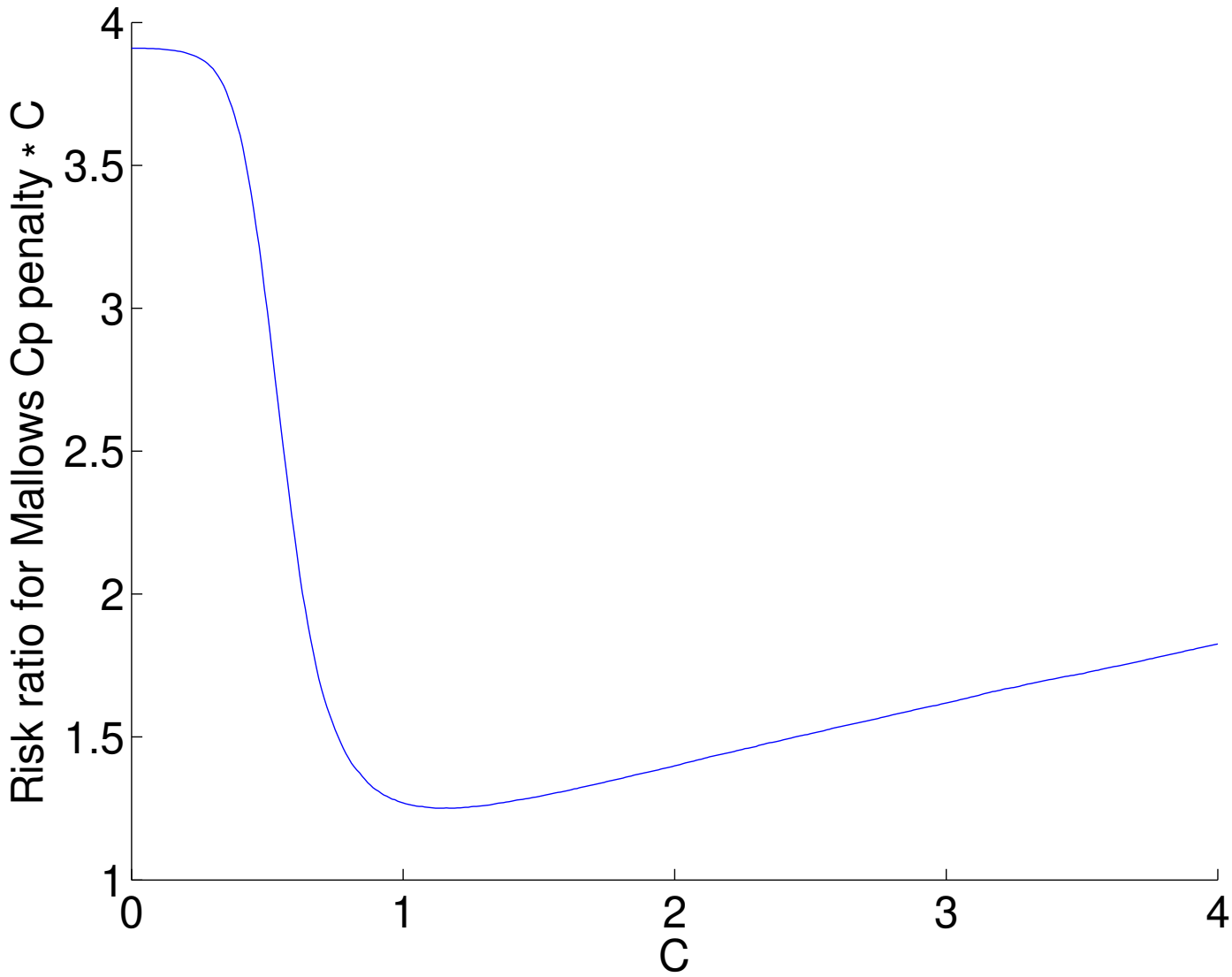
- σ^2 unknown in general

$$\Rightarrow \qquad \text{pen}(m) = \frac{C D_m}{n}$$

  with C (hopefully) close to $2\sigma^2$.

- Oracle inequality for any $C > 0$? Minimal value for C?
- How to estimate σ^2?

Performance of the penalty $K \times \sigma^2 D_m/n$, $K > 0$



Oracle inequality for the penalty CD_m/n

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\mathrm{pen}(m) = \frac{CD_m}{n}$, $C > \sigma^2$, *i.i.d. Gaussian noise.*
Then, for every $x \geqslant 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant L_1\left(\frac{C}{\sigma^2}\right) \inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{20\sigma^2 x}{n}\,.$$

Proof: same arguments as when $C = 2\sigma^2$.
Generalization to sub-Gaussian noise.

Oracle inequality for the penalty CD_m/n

Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

Assumptions: $\mathrm{pen}(m) = \frac{CD_m}{n}$, $C > \sigma^2$, *i.i.d. Gaussian noise.*
Then, for every $x \geqslant 0$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant L_1\left(\frac{C}{\sigma^2}\right) \inf_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{20\sigma^2 x}{n}\,.$$

Proof: same arguments as when $C = 2\sigma^2$.
Generalization to sub-Gaussian noise.

Remark: $L_1(h) \xrightarrow[h\to 1^+]{} +\infty$, $L_1(h) \xrightarrow[h\to+\infty]{} +\infty$.
$\Rightarrow$ overfitting when $C < \sigma^2$?

Expectation of the penalized criterion

$$\left\|\widehat{F}_m - Y\right\|^2 = \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F - F, \varepsilon\right\rangle - 2\left\langle \Pi_m \varepsilon, \varepsilon\right\rangle$$

implies that $\forall C \in \mathbb{R}$,

$$\mathbb{E}\left[\left\|\widehat{F}_m - Y\right\|^2 + \frac{C D_m}{n}\right] = \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] + \sigma^2 + \frac{(C - 2\sigma^2) D_m}{n}$$

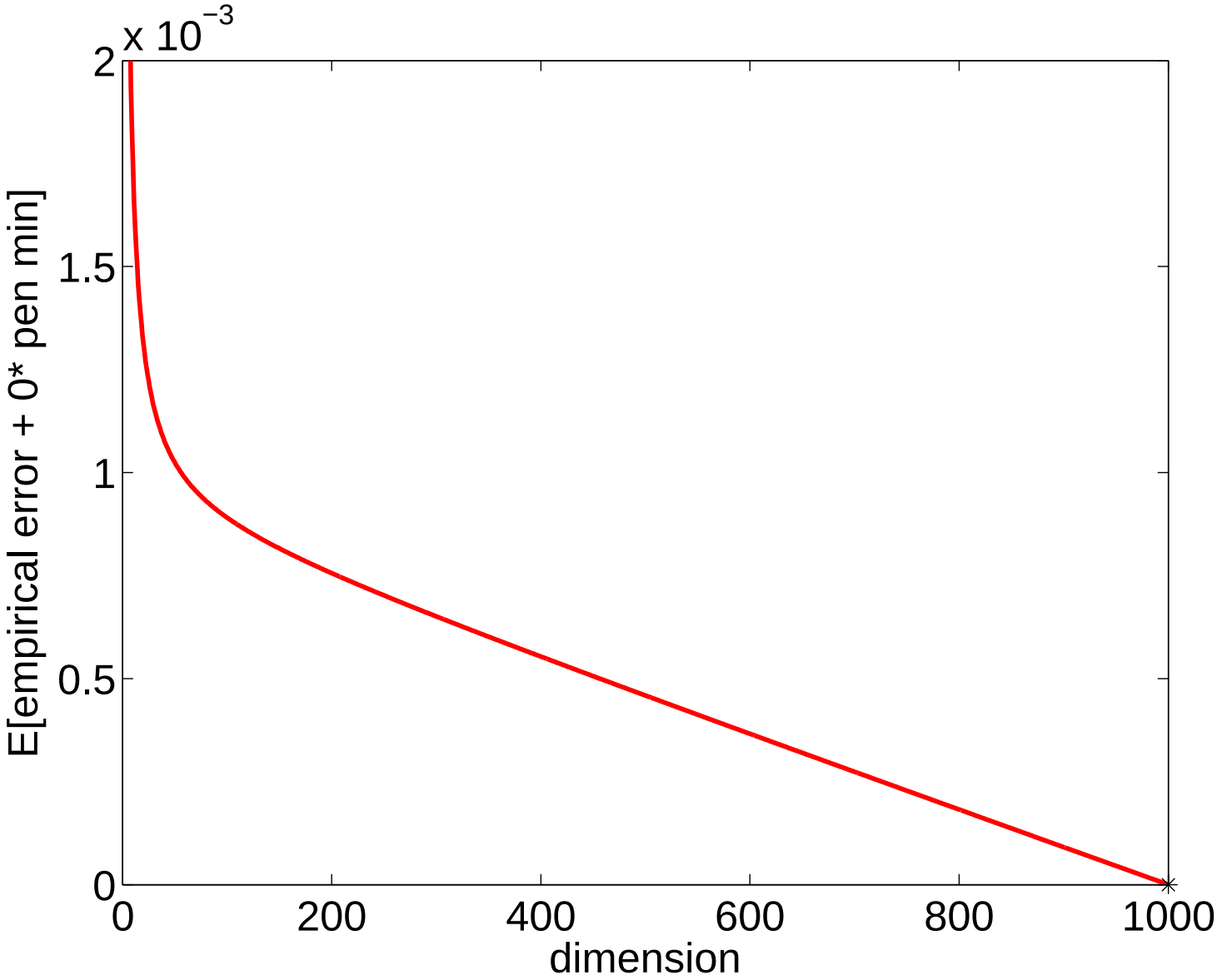
Expectation of the penalized criterion

$$\left\|\widehat{F}_m - Y\right\|^2 = \left\|\widehat{F}_m - F\right\|^2 + \|\varepsilon\|^2 - 2\left\langle \Pi_m F - F, \varepsilon\right\rangle - 2\left\langle \Pi_m \varepsilon, \varepsilon\right\rangle$$

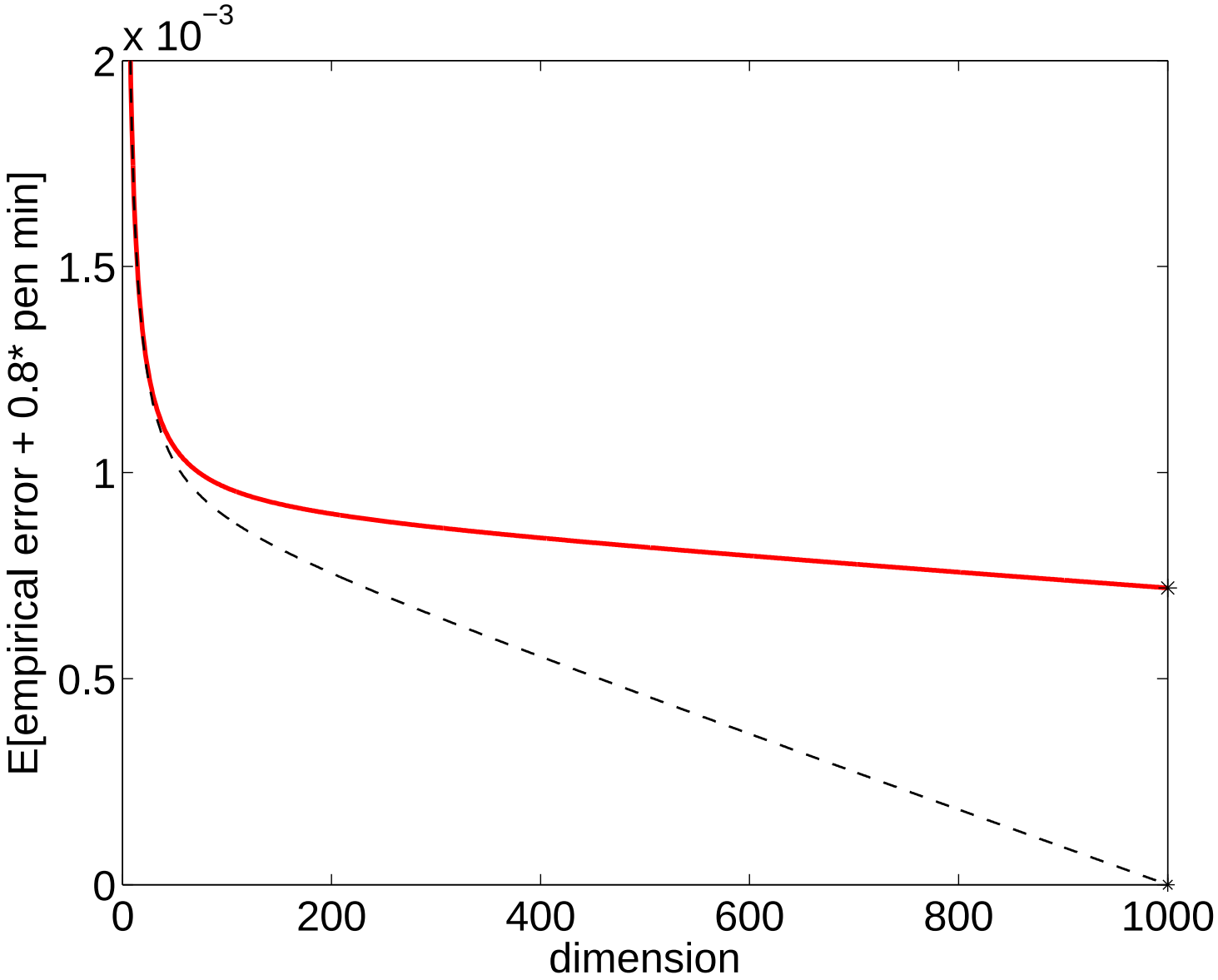
implies that $\forall C \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}\left[\left\|\widehat{F}_m - Y\right\|^2 + \frac{CD_m}{n}\right] &= \mathbb{E}\left[\left\|\widehat{F}_m - F\right\|^2\right] + \sigma^2 + \frac{(C-2\sigma^2)D_m}{n} \\ &= \mathbb{E}\left[\|F_m - F\|^2\right] + \frac{(C-\sigma^2)D_m}{n} + \sigma^2 .\end{aligned}$$

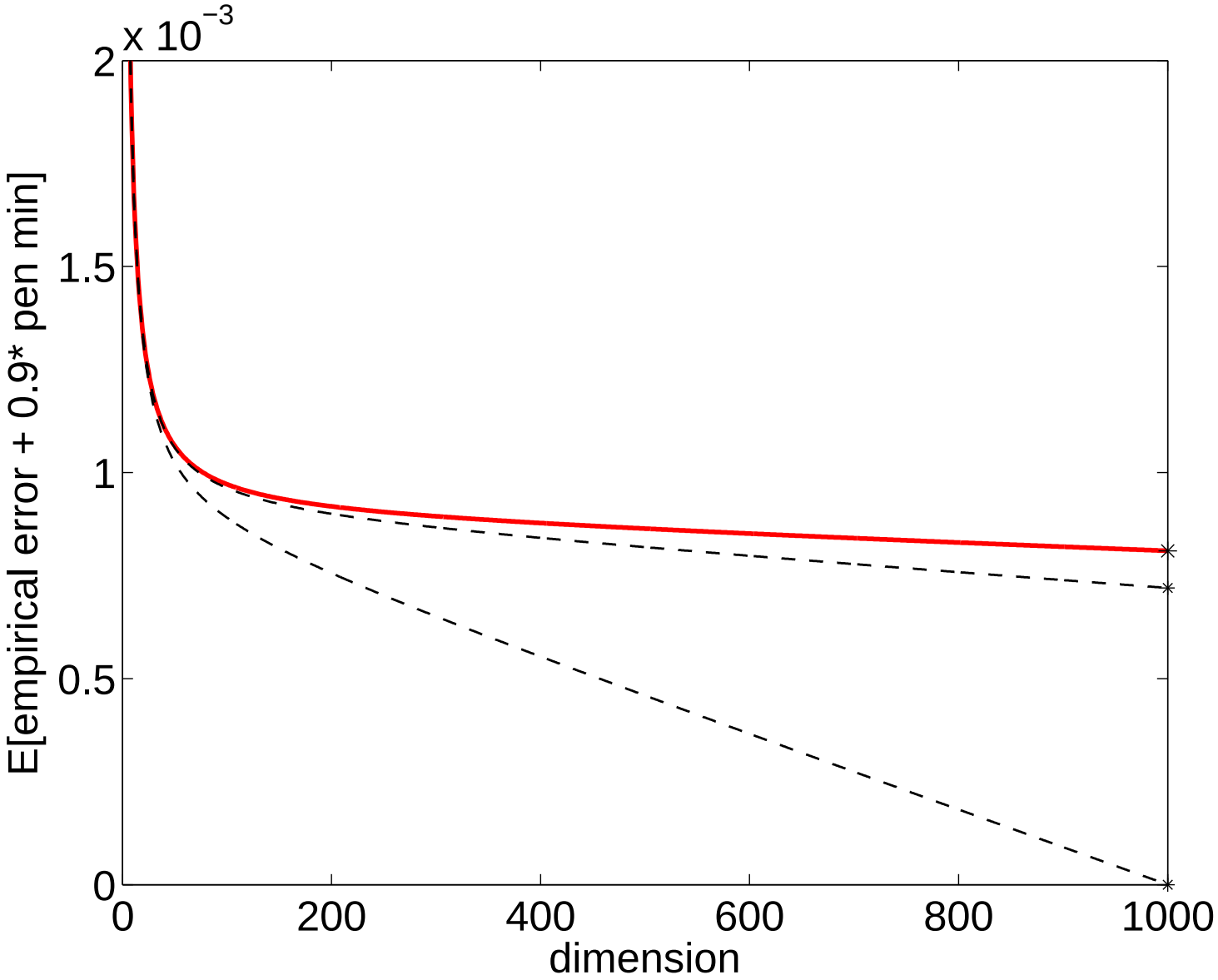
$\mathbb{E}[$ Empirical risk $] + 0 \times \sigma^2 D_m n^{-1}$ (OLS)



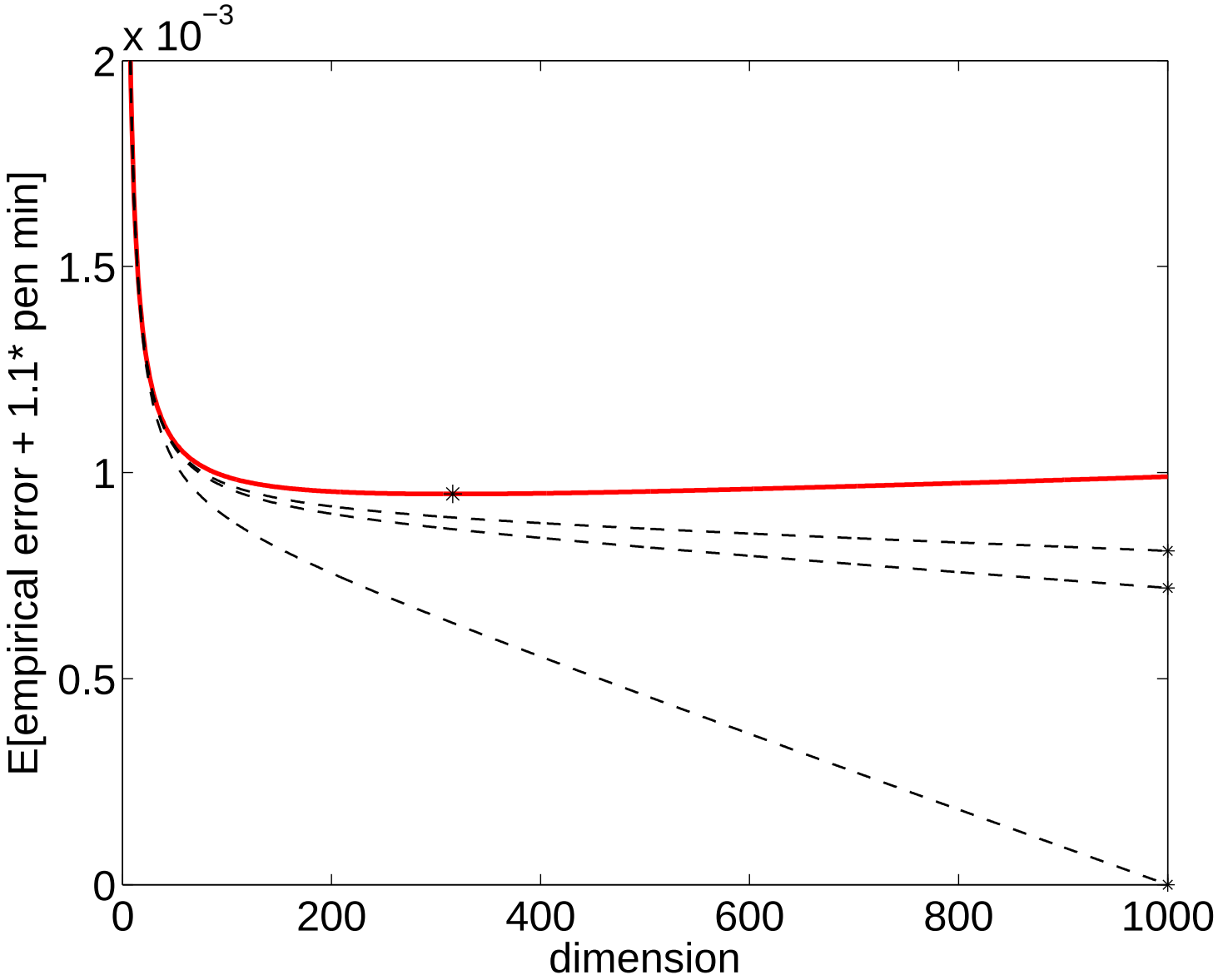
$\mathbb{E}[$ Empirical risk $] + 0.8 \times \sigma^2 D_m n^{-1}$ (OLS)



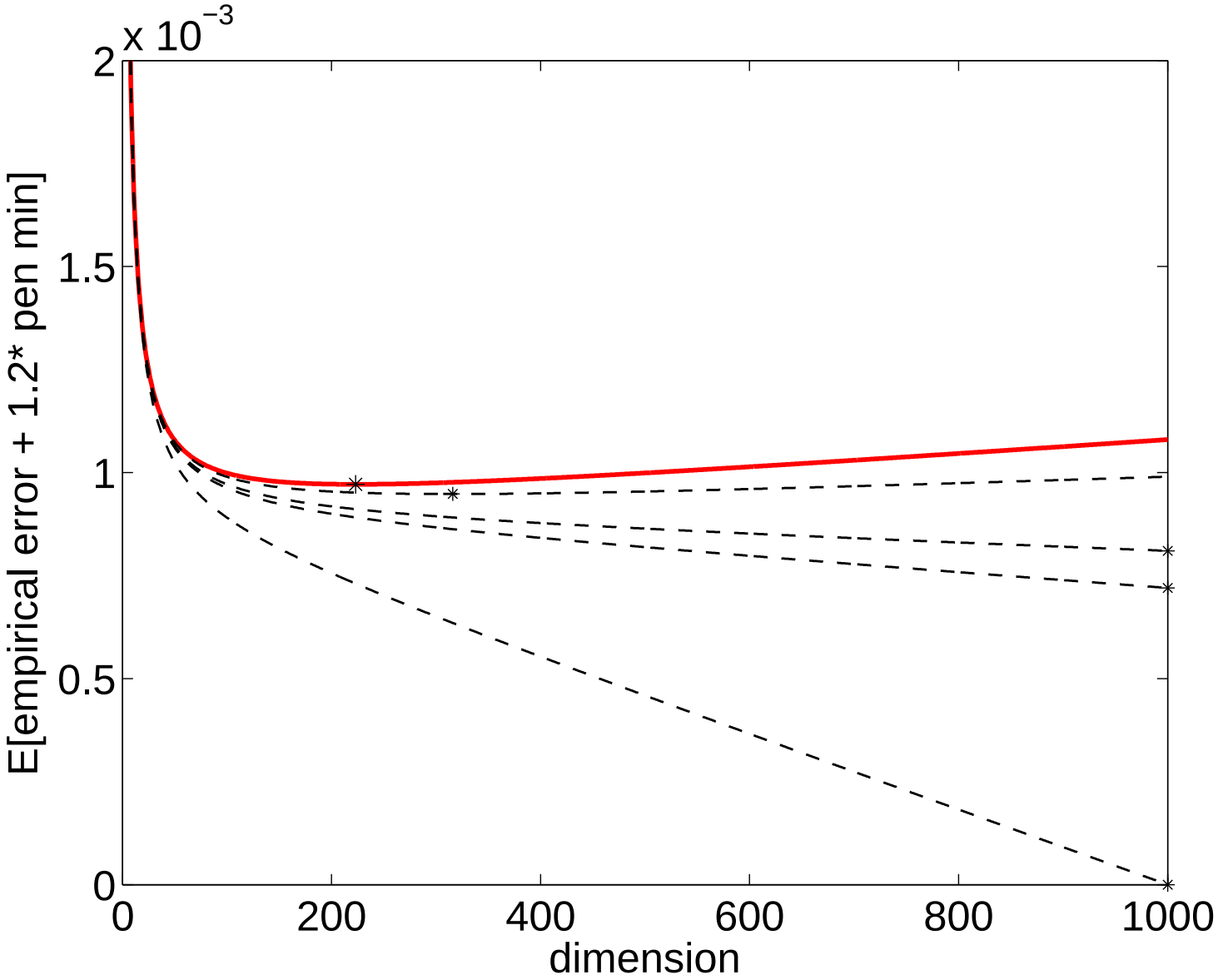
$\mathbb{E}[$ Empirical risk $] + 0.9 \times \sigma^2 D_m n^{-1}$ (OLS)



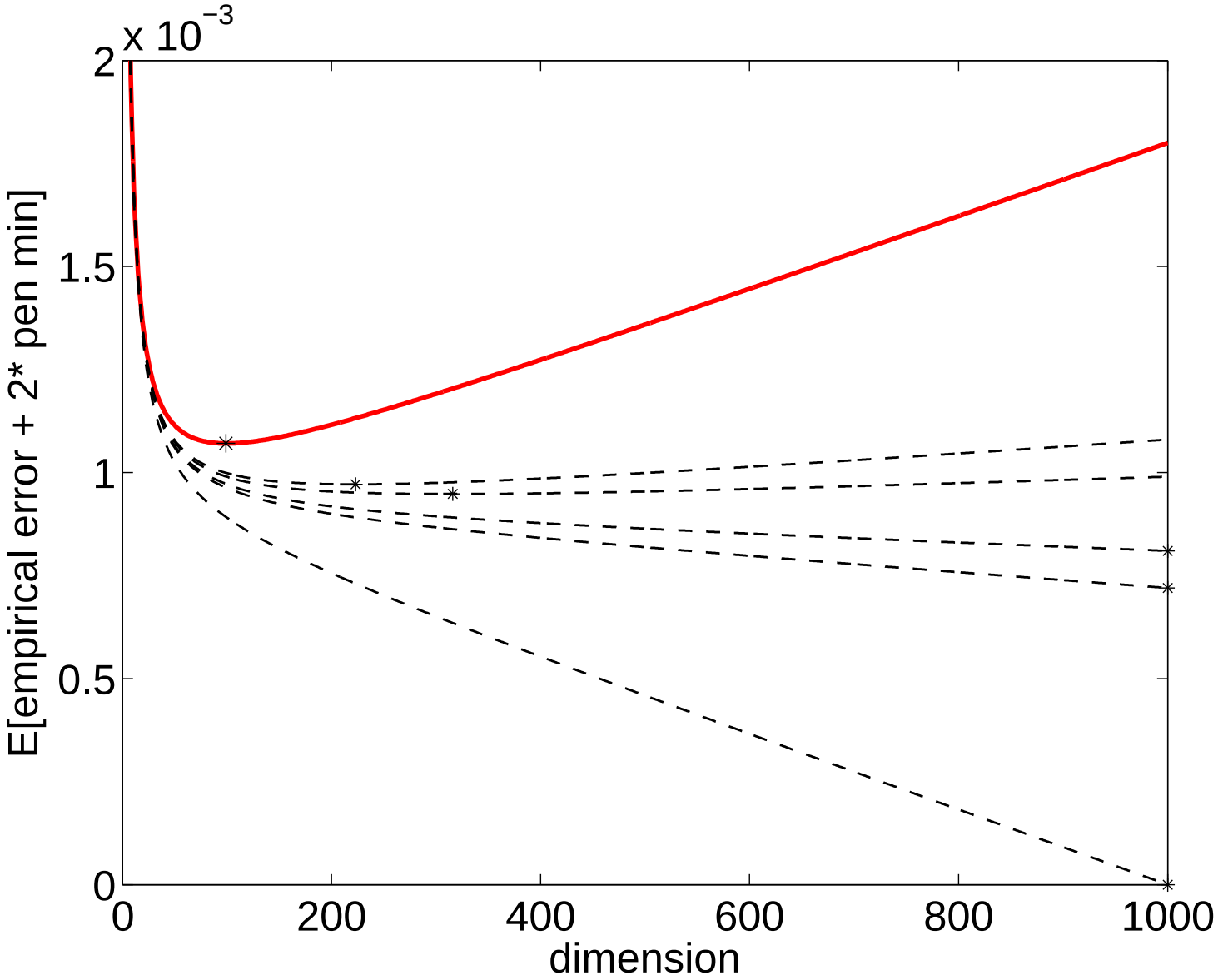
$\mathbb{E}[$ Empirical risk $] + 1.1 \times \sigma^2 D_m n^{-1}$ (OLS)



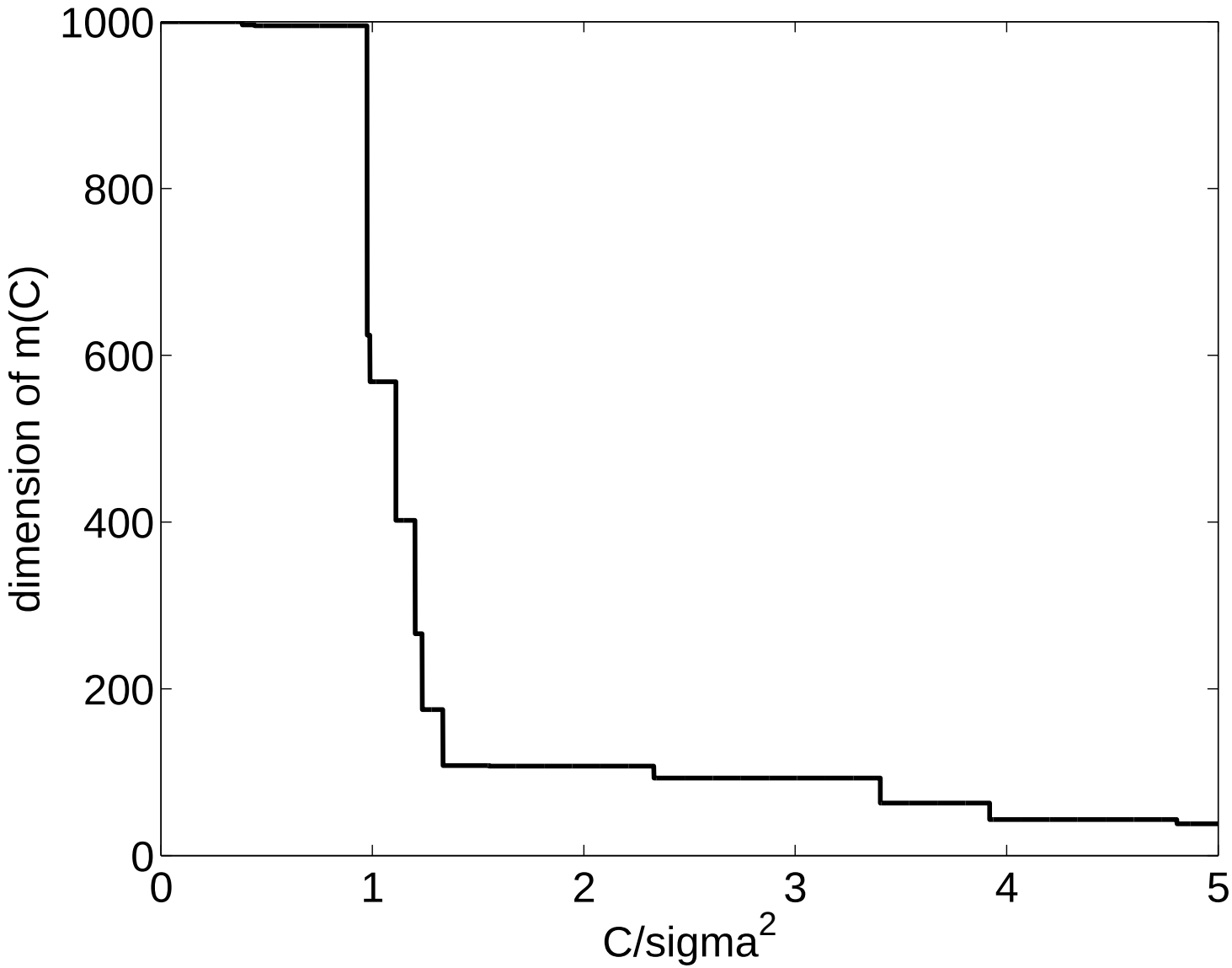
$\mathbb{E}[$ Empirical risk $] + 1.2 \times \sigma^2 D_m n^{-1}$ (OLS)



$\mathbb{E}[$ Empirical risk $] + 2 \times \sigma^2 D_m n^{-1}$ (OLS)



Dimension jump around $C = \sigma^2$



Slope heuristics algorithm (Birgé & Massart 2007)

1. For every $C > 0$, compute

$$\widehat{m}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + C \frac{D_m}{n} \right\} .$$

2. Find $\widehat{C}_{\text{jump}}$ such that $D_{\widehat{m}(C)}$ is "very large" when $C < \widehat{C}_{\text{jump}}$ and "reasonably small" when $C > \widehat{C}_{\text{jump}}$.

3. Select $\widehat{m} = \widehat{m}\left(2\widehat{C}_{\text{jump}}\right)$.

Practical use: CAPUSHE package (Baudry, Maugis & Michel, 2011)

`http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

Dimension jump / Minimal penalty: theory

Theorem (A. & Bach 2011, reformulated as Theorem 1 of arXiv:1901.07277)

Assumptions: i.i.d. Gaussian noise and $\exists m_1 \in \mathcal{M}$, $S_{m_1} = \mathbb{R}^n$, *i.e.,* $\widehat{F}_{m_1} = Y$. *Then,* $\forall x \in [0, n]$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$,

$$\forall C < \left(1 - 100\sqrt{\frac{x}{n}}\right)\sigma^2\,, \quad D_{\widehat{m}(C)} \geqslant \frac{9n}{10}$$

$$\forall C > \left(1 + 240\sqrt{\tfrac{x}{n}}\right)\sigma^2 + 40\inf_{m\,/\,D_m \leqslant n/20}\left\{\tfrac{1}{n}\|F_m - F\|^2\right\}, \quad D_{\widehat{m}(C)} \leqslant \frac{n}{10}$$

In the first case, $\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2 \geqslant \frac{7\sigma^2}{8}$ on the same event.

Remark: generalization to sub-Gaussian noise (arXiv:1901.07277).

Dimension jump / Minimal penalty: theory

Theorem (A. & Bach 2011, reformulated as Theorem 1 of arXiv:1901.07277)

Assumptions: i.i.d. Gaussian noise and $\exists m_1 \in \mathcal{M}$, $S_{m_1} = \mathbb{R}^n$, *i.e.,* $\widehat{F}_{m_1} = Y$. *Then,* $\forall x \in [0, n]$, *with probability at least* $1 - 4\,\mathrm{Card}(\mathcal{M}) e^{-x}$,

$$\forall C < \left(1 - 100\sqrt{\frac{x}{n}}\right)\sigma^2\,, \quad D_{\widehat{m}(C)} \geqslant \frac{9n}{10}$$

$$\forall C > \left(1 + 240\sqrt{\tfrac{x}{n}}\right)\sigma^2 + 40\inf_{m\,/\,D_m \leqslant n/20}\left\{\tfrac{1}{n}\|F_m - F\|^2\right\}, \quad D_{\widehat{m}(C)} \leqslant \frac{n}{10}$$

Consequences: $\widehat{C}_{\mathrm{jump}}$ estimates well σ^2 & oracle inequality for $\widehat{m}(2\widehat{C}_{\mathrm{jump}})$ (see arXiv:1901.07277).

Practical qualities of the algorithm

- visual checking of existence of a jump
- calibration independent from the choice of some m_0
- too strong overfitting almost impossible
- one remaining parameter: how to localize the jump

How to localize the jump in practice?

- Dimension jump: largest jump? jump on a geometrical window? complexity threshold?

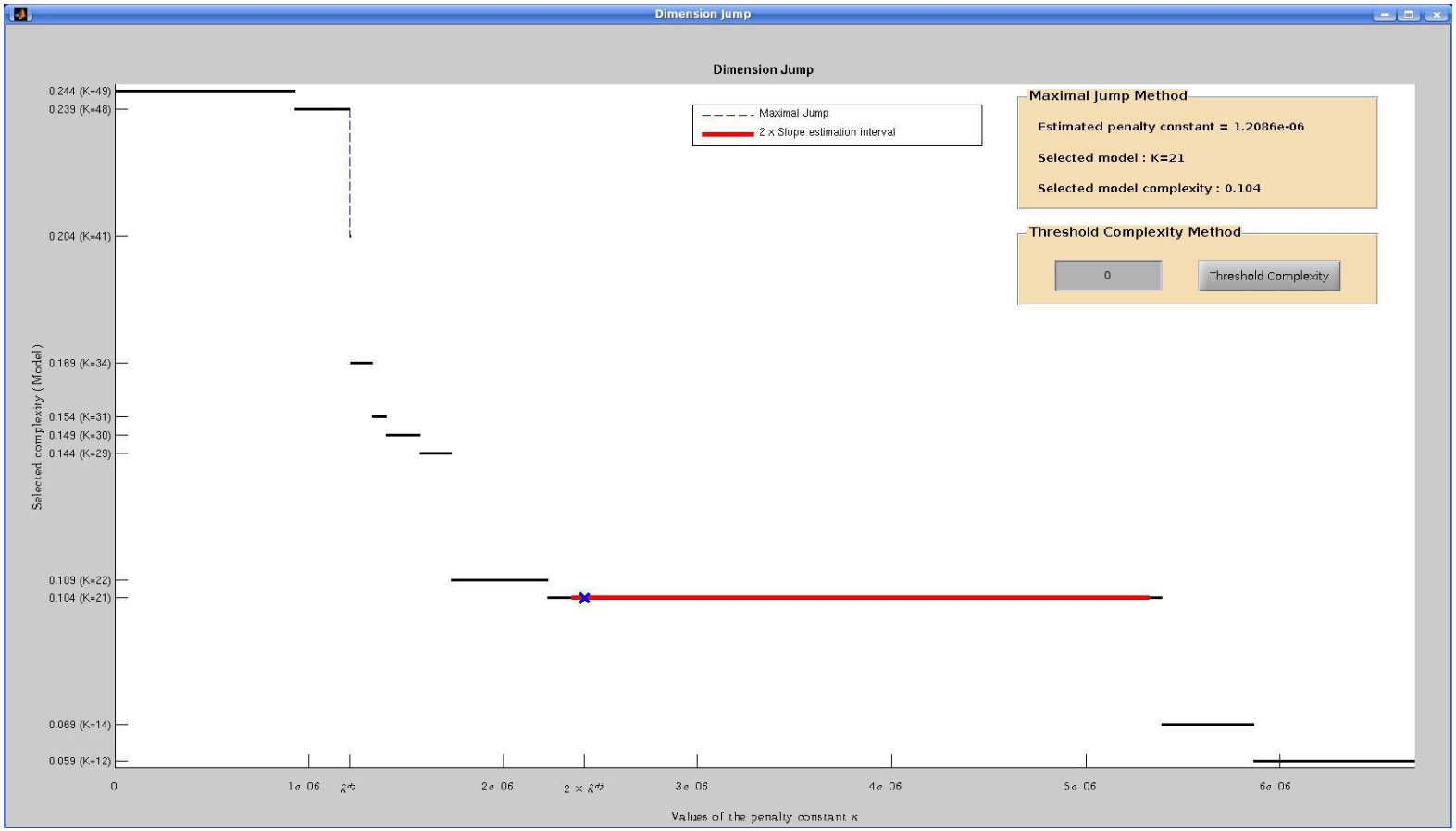
How to localize the jump in practice?

- Dimension jump: largest jump? jump on a geometrical window? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the dimension:
  computed with which models? robust regression?

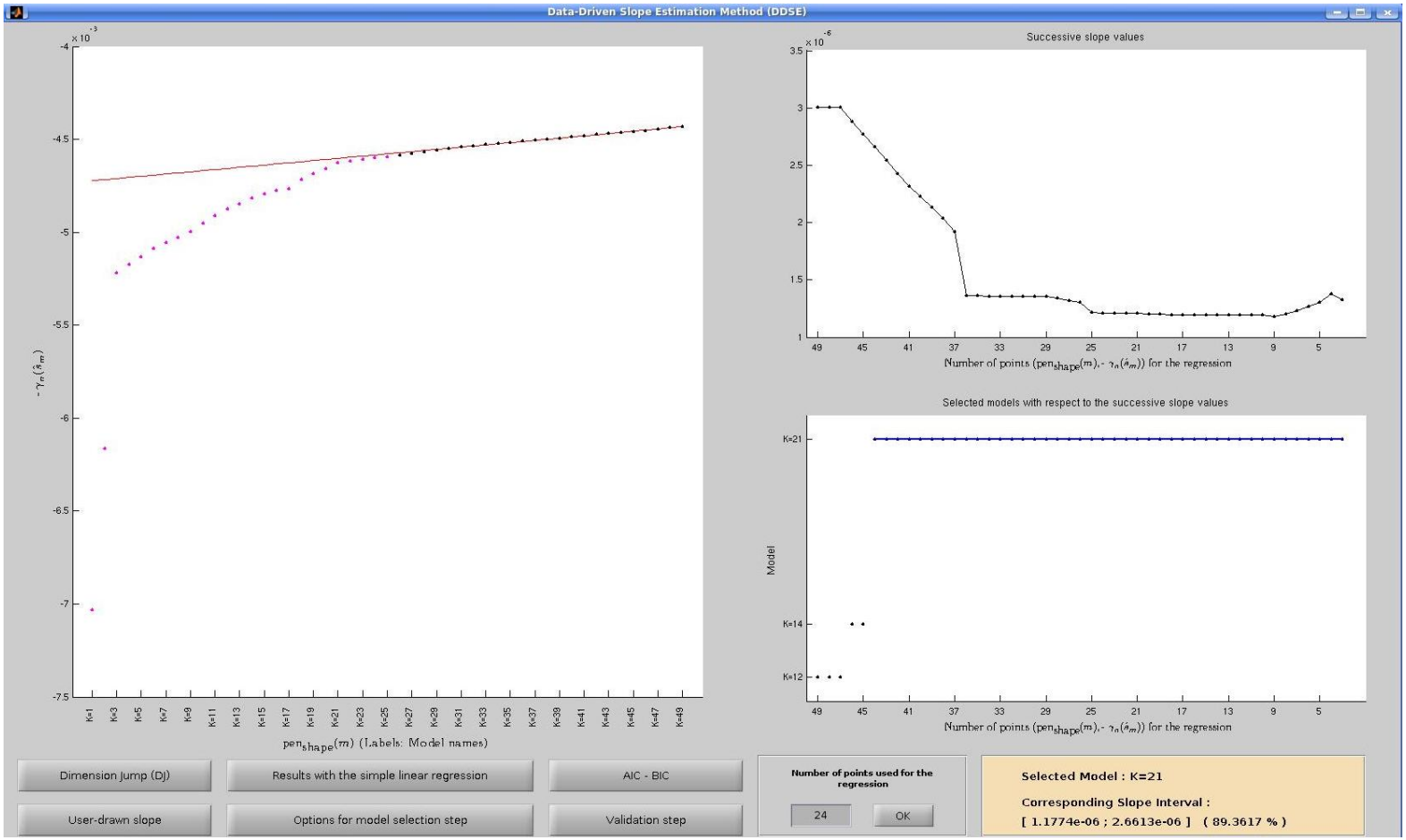
How to localize the jump in practice?

- Dimension jump: largest jump? jump on a geometrical window? complexity threshold?
- Estimation of the slope of the empirical risk as a function of the dimension:
  computed with which models? robust regression?
- Jump vs. slope? Take both!
  $\Rightarrow$ package CAPUSHE (Baudry, Maugis & Michel, 2011)
  `http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

CAPUSHE (Baudry, Maugis & Michel, 2011): jump



CAPUSHE (Baudry, Maugis & Michel, 2011): slope



Part II

Model/estimator selection in the prediction setting

Outline

General prediction setting

- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$

General prediction setting

- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- Predictor: $t : \mathcal{X} \to \mathcal{Y}$

  new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ "predicts" Y_{n+1}

General prediction setting

- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- Predictor: $t : \mathcal{X} \to \mathcal{Y}$

  new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ "predicts" Y_{n+1}
- Risk (prediction error): $\mathcal{R}(t) = \mathbb{E}\Big[c\big(t(X), Y\big)\Big]$ where $(X, Y) \sim P$

  minimal for $t = f^\star$ (Bayes predictor)

  $\Rightarrow$ Excess risk $\ell(t, f^\star) := \mathcal{R}(t) - \mathcal{R}(f^\star) \geqslant 0$.

General prediction setting

- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- Predictor: $t : \mathcal{X} \to \mathcal{Y}$
  new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ "predicts" Y_{n+1}
- Risk (prediction error): $\mathcal{R}(t) = \mathbb{E}\Big[c\big(t(X), Y\big)\Big]$ where $(X, Y) \sim P$
  minimal for $t = f^\star$ (Bayes predictor)
  $\Rightarrow$ Excess risk $\ell(t, f^\star) := \mathcal{R}(t) - \mathcal{R}(f^\star) \geqslant 0$.
- Goal: from D_n only, find t with $\mathcal{R}(t)$ minimal.

General prediction setting

- Data: $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n} \in (\mathcal{X} \times \mathcal{Y})^n$ assumed i.i.d. $\sim P$
- Predictor: $t : \mathcal{X} \to \mathcal{Y}$
  new data $X_{n+1} \rightsquigarrow t(X_{n+1})$ "predicts" Y_{n+1}
- Risk (prediction error): $\mathcal{R}(t) = \mathbb{E}\Big[c\big(t(X), Y\big)\Big]$ where $(X, Y) \sim P$
  minimal for $t = f^\star$ (Bayes predictor)
  $\Rightarrow$ Excess risk $\ell(t, f^\star) := \mathcal{R}(t) - \mathcal{R}(f^\star) \geqslant 0$.
- Goal: from D_n only, find t with $\mathcal{R}(t)$ minimal.
- More general setting possible, including density estimation with LS or KL risk.

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\left[(t(X) - f^\star(X))^2\right]$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\left[(t(X) - f^\star(X))^2\right]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\left[(t(X) - f^\star(X))^2\right]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$
  - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leqslant \delta \\ \delta\left(|y - y'| - \frac{\delta}{2}\right) & \text{otherwise} \end{cases}$$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\left[(t(X) - f^\star(X))^2\right]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$
  - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leqslant \delta \\ \delta\left(|y - y'| - \frac{\delta}{2}\right) & \text{otherwise} \end{cases}$$

- Binary classification: $\mathcal{Y} = \{0, 1\}$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\big[(t(X) - f^\star(X))^2\big]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$
  - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leqslant \delta \\ \delta\big(|y - y'| - \frac{\delta}{2}\big) & \text{otherwise} \end{cases}$$

- Binary classification: $\mathcal{Y} = \{0, 1\}$
  - 0–1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
    $\Rightarrow f^\star(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geqslant 1/2}$

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\big[(t(X) - f^\star(X))^2\big]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$
  - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leqslant \delta \\ \delta\big(|y - y'| - \frac{\delta}{2}\big) & \text{otherwise} \end{cases}$$

- Binary classification: $\mathcal{Y} = \{0, 1\}$
  - 0–1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
    $\Rightarrow f^\star(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geqslant 1/2}$
  - convex losses (hinge, logistic, exponential, ...)

Prediction setting: examples

- Regression: $\mathcal{Y} = \mathbb{R}$
  - least squares: $c(y, y') = (y - y')^2$
    $\Rightarrow f^\star(X) = \mathbb{E}[Y|X]$ and $\ell(t, f^\star) = \mathbb{E}\big[(t(X) - f^\star(X))^2\big]$
  - L^p loss: $c(y, y') = |y - y'|^p$, $p \geqslant 1$
  - Huber loss (robustness):

$$c(y, y') = \begin{cases} \frac{1}{2}(y - y')^2 & \text{if } |y - y'| \leqslant \delta \\ \delta\big(|y - y'| - \frac{\delta}{2}\big) & \text{otherwise} \end{cases}$$

- Binary classification: $\mathcal{Y} = \{0, 1\}$
  - 0–1 loss: $c(y, y') = \mathbb{1}_{t(x) \neq y}$
    $\Rightarrow f^\star(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geqslant 1/2}$
  - convex losses (hinge, logistic, exponential, ...)

- Multi-class classification: $\mathcal{Y} = \{0, \ldots, M - 1\}$

Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:
  $X_1, \ldots, X_n$ i.i.d.

- Fixed-design regression:
  $X_1, \ldots, X_n$ deterministic

Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:
  $X_1, \ldots, X_n$ i.i.d.
  target: $f^\star(X) = \mathbb{E}[Y|X]$

- Fixed-design regression:
  $X_1, \ldots, X_n$ deterministic
  target $\left(f^\star(X_i)\right)_{1 \leqslant i \leqslant n} = \left(\mathbb{E}[Y_i|X_i]\right)_{1 \leqslant i \leqslant n}$

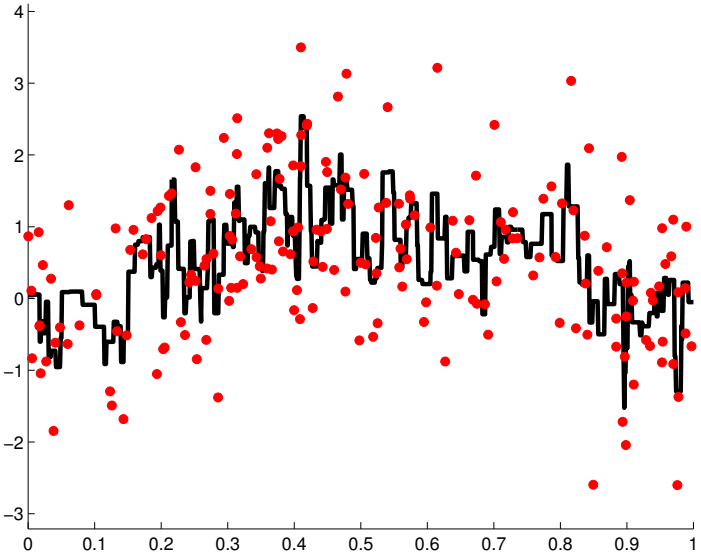
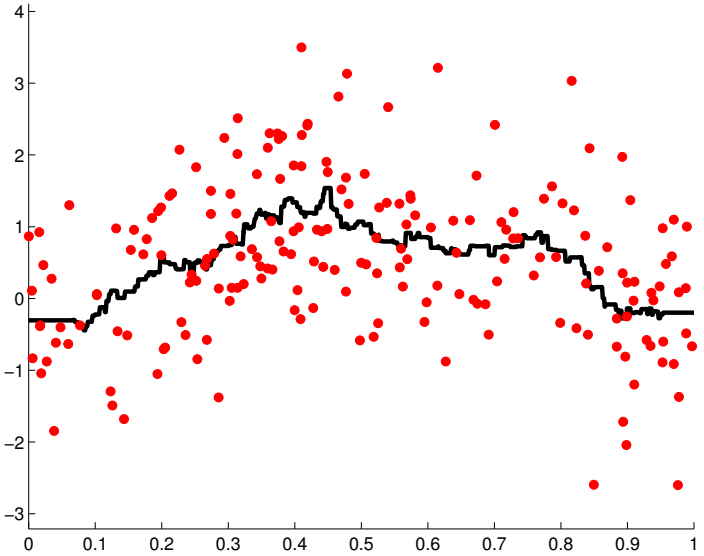
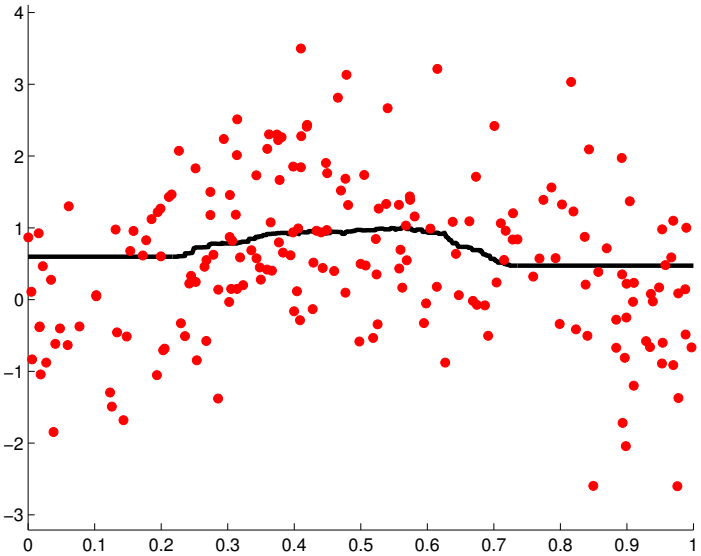
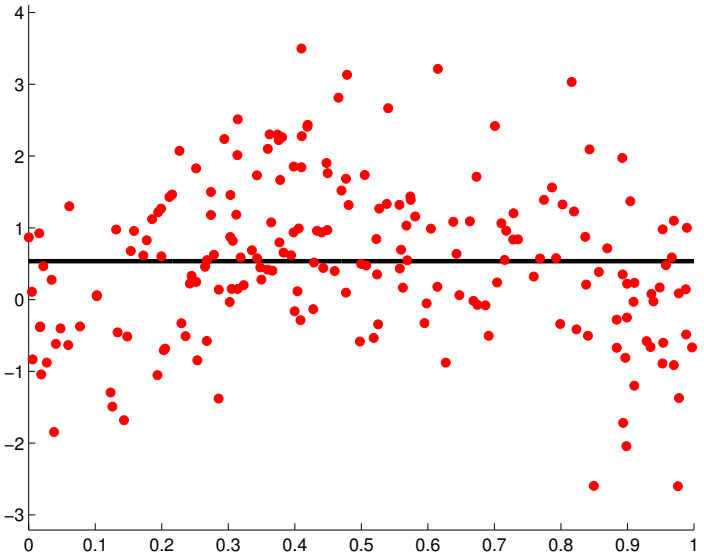
Link with part I (fixed-design regression)

- Random-design regression, least-squares loss:
  $X_1, \ldots, X_n$ i.i.d.
  target: $f^\star(X) = \mathbb{E}[Y|X]$
  excess risk: $\ell(t, f^\star) = \mathcal{R}(t) - \mathcal{R}(f^\star) = \mathbb{E}\Big[\big(t(X) - f^\star(X)\big)^2\Big]$

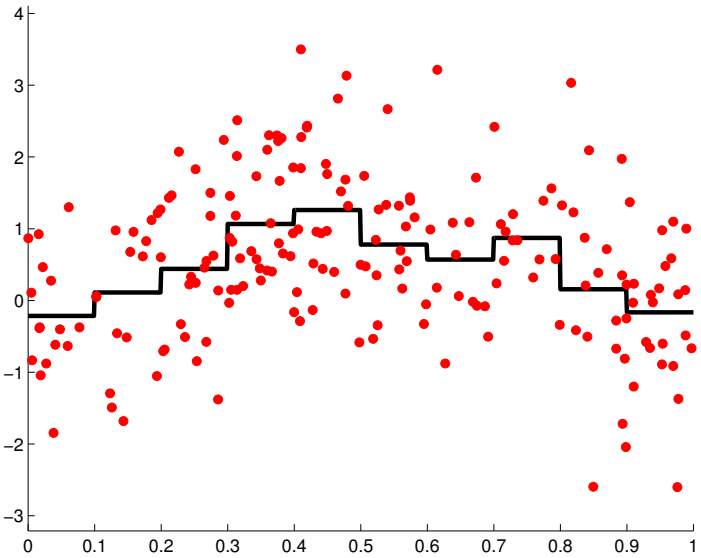
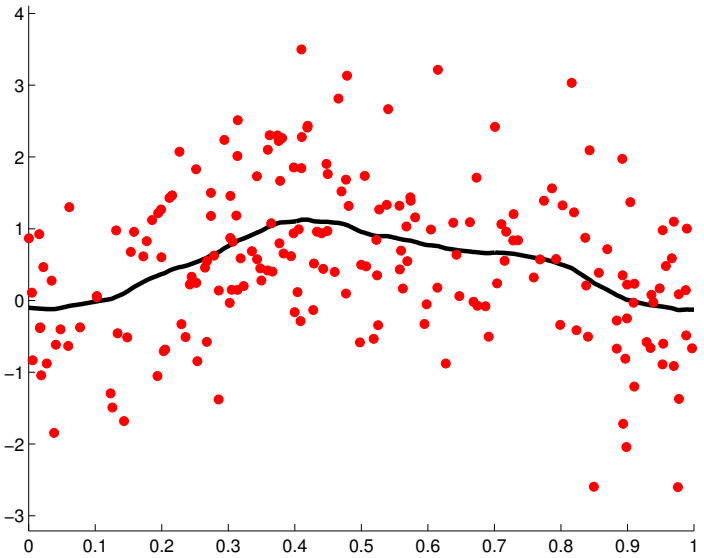
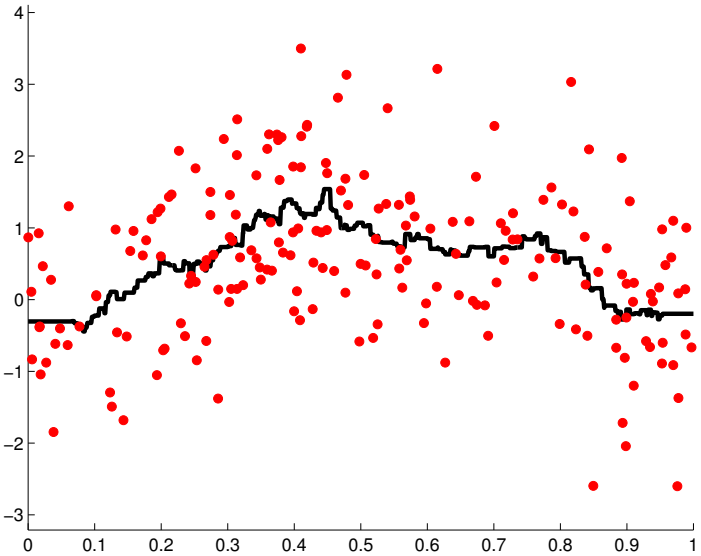
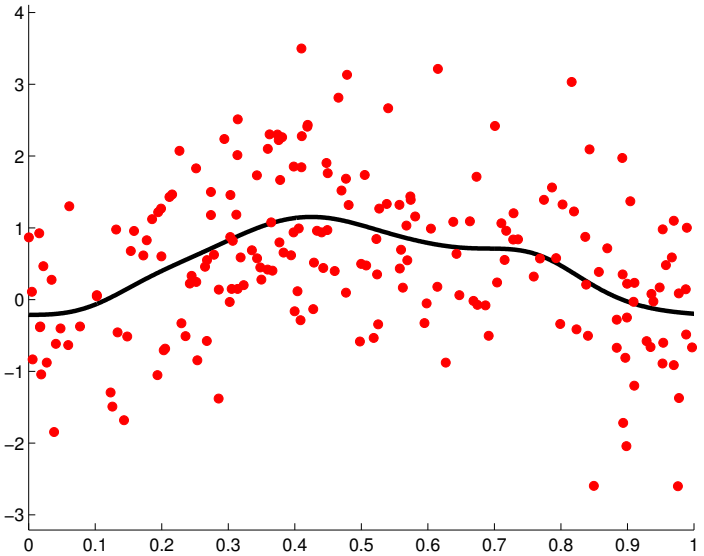
- Fixed-design regression:
  $X_1, \ldots, X_n$ deterministic
  target $\big(f^\star(X_i)\big)_{1 \leqslant i \leqslant n} = \big(\mathbb{E}[Y_i|X_i]\big)_{1 \leqslant i \leqslant n}$
  excess risk: $\frac{1}{n}\sum_{i=1}^n \big(t(X_i) - f^\star(X_i)\big)^2$
  $\leftrightsquigarrow$ "$X \sim \mathcal{U}(\{X_1, \ldots, X_n\})$"

Outline

Estimator selection: k nearest neighbours

Estimators: regressogram, ridge, k-NN, Nadaraya-Watson

Estimator selection

- Estimator collection $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(D_n)$?

Estimator selection

- Estimator collection $(\widehat{f}_m)_{m\in\mathcal{M}} \Rightarrow \widehat{m}(D_n)$?
- Examples:
  - model selection
  - parameter tuning (choosing k or the distance for k-NN, choice of a regularization parameter, choice of a kernel, etc.)
  - choice between different methods
    ex.: k-NN vs. kernel ridge?

Estimator selection

- Estimator collection $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow \widehat{m}(D_n)$?
- Examples:
  - model selection
  - parameter tuning (choosing k or the distance for k-NN, choice of a regularization parameter, choice of a kernel, etc.)
  - choice between different methods
    ex.: k-NN vs. kernel ridge?
- Goal: minimize the risk $\mathcal{R}\left(\widehat{f}_{\widehat{m}(D_n)}(D_n)\right)$
- Other possible goal: identify the "best" estimator (or the "true" model)

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error

- Sometimes possible:
  - empirical risk minimizer: $\widehat{f}_m \in \operatorname{argmin}_{t \in S_m} \widehat{\mathcal{R}}_n(t)$
  - "linear" estimators (fixed-design regression)
  - local averaging estimators

Approximation / estimation error decomposition?

- No general decomposition of the risk between approximation and estimation error
- Sometimes possible:
  - empirical risk minimizer: $\widehat{f}_m \in \operatorname{argmin}_{t \in S_m} \widehat{\mathcal{R}}_n(t)$
  - "linear" estimators (fixed-design regression)
  - local averaging estimators
- Always have to avoid overfitting and underfitting

Estimator selection: methods

- Classical approach:

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m)\}$$

Estimator selection: methods

- Classical approach:

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m)\}$$

- Examples:
  - penalization (Mallows' C_p, AIC, BIC, structural risk minimization, ...)
  - cross-validation
  - FPE, GCV, ...

Estimator selection: methods

- Classical approach:

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m)\}$$

- Examples:
  - penalization (Mallows' C_p, AIC, BIC, structural risk minimization, ...)
  - cross-validation
  - FPE, GCV, ...
- How to choose crit?

Estimator selection: methods

- Classical approach:

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \{\mathrm{crit}(m)\}$$

- Examples:
  - penalization (Mallows' C_p, AIC, BIC, structural risk minimization, ...)
  - cross-validation
  - FPE, GCV, ...

- How to choose crit?
  Idea: use the key lemma

The key lemma (revisited)

Lemma

Let $\mathrm{crit} : \mathcal{M} \to \mathbb{R}$ *be any function* (*possibly data-dependent*).
On the event Ω *on which,* $\forall m, m' \in \mathcal{M}$

$$\left[\mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m\right)\right] - \left[\mathrm{crit}(m') - \mathcal{R}\left(\widehat{f}_{m'}\right)\right] \leqslant A(m) + B(m')\,,$$

we have $\forall \widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{\mathrm{crit}(m)\right\}$,

$$\ell\left(\widehat{f}_{\widehat{m}}, f^\star\right) - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{\ell\left(\widehat{f}_m, f^\star\right) + A(m)\right\}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) \\
&= \mathrm{crit}(\widehat{m}) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&\leqslant \mathrm{crit}(m) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&= \mathcal{R}\left(\widehat{f}_{m}\right) + \mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_{m}\right) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&\leqslant \mathcal{R}\left(\widehat{f}_{m}\right) + A(m) + B(\widehat{m})
\end{aligned}$$

hence

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_{m}\right) + A(m) \right\}$$

Proof of the key lemma

For any $m \in \mathcal{M}$, we have

$$\begin{aligned}
&\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) \\
&= \mathrm{crit}(\widehat{m}) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&\leqslant \mathrm{crit}(m) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&= \mathcal{R}\left(\widehat{f}_{m}\right) + \mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_{m}\right) + \mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - \mathrm{crit}(\widehat{m}) \\
&\leqslant \mathcal{R}\left(\widehat{f}_{m}\right) + A(m) + B(\widehat{m})
\end{aligned}$$

hence

$$\mathcal{R}\left(\widehat{f}_{\widehat{m}}\right) - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_{m}\right) + A(m) \right\}$$

and $$\ell\left(\widehat{f}_{\widehat{m}}, f^\star\right) - B(\widehat{m}) \leqslant \inf_{m \in \mathcal{M}} \left\{ \ell\left(\widehat{f}_{m}, f^\star\right) + A(m) \right\}. \quad \square$$

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}\,, \qquad \mathbb{E}\big[\mathrm{crit}(m; D_n)\big] \approx \mathbb{E}\big[\mathcal{R}\left(\widehat{f}_m(D_n)\right)\big]$$

- Examples: C_p, AIC, cross-validation, FPE, ...

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}\,, \qquad \mathbb{E}\big[\mathrm{crit}(m; D_n)\big] \approx \mathbb{E}\big[\mathcal{R}\left(\widehat{f}_m(D_n)\right)\big]$$

- Examples: C_p, AIC, cross-validation, FPE, ...
- Concentration inequalities + $\mathcal{M}$ "not too large"
  $\Rightarrow$ with a large probability, $\forall m, m' \in \mathcal{M}$,

$$\left[\mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m\right)\right] - \left[\mathrm{crit}(m') - \mathcal{R}\left(\widehat{f}_{m'}\right)\right] \leqslant A(m) + B(m')\,,$$

with $\qquad A(m) = B(m) \leqslant \epsilon_1 \ell\left(\widehat{f}_m, f^\star\right) + \epsilon_2\,.$

Application 1: unbiased risk estimation

$$\forall m \in \mathcal{M}\,, \qquad \mathbb{E}\big[\mathrm{crit}(m; D_n)\big] \approx \mathbb{E}\big[\mathcal{R}\left(\widehat{f}_m(D_n)\right)\big]$$

- Examples: C_p, AIC, cross-validation, FPE, ...
- Concentration inequalities + $\mathcal{M}$ "not too large"
  $\Rightarrow$ with a large probability, $\forall m, m' \in \mathcal{M}$,

  $$\left[\mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m\right)\right] - \left[\mathrm{crit}(m') - \mathcal{R}\left(\widehat{f}_{m'}\right)\right] \leqslant A(m) + B(m')\,,$$

  with $\qquad A(m) = B(m) \leqslant \epsilon_1 \ell\left(\widehat{f}_m, f^\star\right) + \epsilon_2\,.$
- If $\epsilon_1 < 1$, by the key lemma

$$\ell\left(\widehat{f}_{\widehat{m}}, f^\star\right) \leqslant \frac{1+\epsilon_1}{1-\epsilon_1} \inf_{m \in \mathcal{M}} \left\{\ell\left(\widehat{f}_m(D_n), f^\star\right)\right\} + \frac{2\epsilon_2}{1-\epsilon_1}$$

$\Rightarrow$ oracle inequality, first-order optimal if $\epsilon_1 \ll 1$ and $\epsilon_2 \ll \inf_{m \in \mathcal{M}}\{\ell(f^\star, \widehat{f}_m(D_n))\}$.

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) \geqslant \mathcal{R}\left(\widehat{f}_m(D_n)\right) \quad \left(\text{or} \quad \ell\left(\widehat{f}_m(D_n), f^\star\right)\right)$$

(with a large probability)

- Examples: BIC, structural risk minimization, ...

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) \geqslant \mathcal{R}\left(\widehat{f}_m(D_n)\right) \quad \left(\text{or} \quad \ell\left(\widehat{f}_m(D_n), f^\star\right)\right)$$

(with a large probability)

- Examples: BIC, structural risk minimization, ...
- Then, $\forall m, m' \in \mathcal{M}$,

$$\left[\mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m\right)\right] - \left[\mathrm{crit}(m') - \mathcal{R}\left(\widehat{f}_{m'}\right)\right] \leqslant A(m) + B(m')\,,$$

with $\quad A(m) = 0 \quad$ and $\quad B(m) = \mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m(D_n)\right)$.

Application 2: upper bound on the risk

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) \geqslant \mathcal{R}\left(\widehat{f}_m(D_n)\right) \quad \left(\text{or} \quad \ell\left(\widehat{f}_m(D_n), f^\star\right)\right)$$

(with a large probability)

- Examples: BIC, structural risk minimization, ...
- Then, $\forall m, m' \in \mathcal{M}$,

$$\left[\mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m\right)\right] - \left[\mathrm{crit}(m') - \mathcal{R}\left(\widehat{f}_{m'}\right)\right] \leqslant A(m) + B(m')\,,$$

with $\quad A(m) = 0 \quad$ and $\quad B(m) = \mathrm{crit}(m) - \mathcal{R}\left(\widehat{f}_m(D_n)\right)$.

- By the key lemma

$$\ell\left(\widehat{f}_{\widehat{m}}, f^\star\right) \leqslant \inf_{m \in \mathcal{M}} \left\{ \ell\left(\widehat{f}_m(D_n), f^\star\right) + B(m) \right\}$$

$\Rightarrow$ oracle inequality, interesting if $B(m)$ small enough.

Estimator selection by penalization

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) = \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big) + \mathrm{pen}(m)$$

$$\text{where} \qquad \widehat{\mathcal{R}}_n(t) := \frac{1}{n}\sum_{i=1}^{n} c(t(X_i), Y_i) \qquad \text{(empirical risk)}$$

Estimator selection by penalization

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) = \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big) + \mathrm{pen}(m)$$

$$\text{where} \qquad \widehat{\mathcal{R}}_n(t) := \frac{1}{n}\sum_{i=1}^{n} c(t(X_i), Y_i) \qquad \text{(empirical risk)}$$

- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) := \mathcal{R}\Big(\widehat{f}_m(D_n)\Big) - \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big)\,.$$

Estimator selection by penalization

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) = \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big) + \mathrm{pen}(m)$$

$$\text{where} \qquad \widehat{\mathcal{R}}_n(t) := \frac{1}{n}\sum_{i=1}^{n} c(t(X_i), Y_i) \qquad \text{(empirical risk)}$$

- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) := \mathcal{R}\Big(\widehat{f}_m(D_n)\Big) - \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big)\,.$$

- Unbiased risk estimation
  $\Leftrightarrow \quad \forall m \in \mathcal{M}, \quad \mathbb{E}[\mathrm{pen}(m; D_n)] = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)].$

Estimator selection by penalization

$$\forall m \in \mathcal{M}\,, \qquad \mathrm{crit}(m; D_n) = \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big) + \mathrm{pen}(m)$$

$$\text{where} \qquad \widehat{\mathcal{R}}_n(t) := \frac{1}{n}\sum_{i=1}^{n} c(t(X_i), Y_i) \qquad \text{(empirical risk)}$$

- Ideal penalty:

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) := \mathcal{R}\Big(\widehat{f}_m(D_n)\Big) - \widehat{\mathcal{R}}_n\Big(\widehat{f}_m(D_n)\Big)\,.$$

- Unbiased risk estimation
  $\Leftrightarrow \quad \forall m \in \mathcal{M}, \quad \mathbb{E}[\mathrm{pen}(m; D_n)] = \mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m; D_n)].$
- Upper bound on the risk
  $\Leftrightarrow \quad \forall m \in \mathcal{M}, \quad \mathrm{pen}(m; D_n) \geqslant \mathrm{pen}_{\mathrm{id}}(m; D_n).$

Outline

Motivation: Penalties known up to a constant factor

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\mathrm{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n} \qquad \mathrm{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

Rk: various methods for estimating σ^2 or avoiding its estimation (FPE, Akaike, 1970; GCV, Craven & Wahba, 1978; Baraud, Giraud & Huet, 2009).

Motivation: Penalties known up to a constant factor

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \operatorname{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\operatorname{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n} \qquad \operatorname{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Optimal penalty known asymptotically (AIC; Akaike, 1973)
- Resampling-based penalties
- Optimal constant unknown even in theory (change-point detection, mixture models, global/local Rademacher complexities, ...)

Motivation: Penalties known up to a constant factor

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \operatorname{pen}(m) \right\}$$

- Optimal penalties depending on the noise level σ^2 (Mallows, 1973):

$$\operatorname{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n} \qquad \operatorname{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Optimal penalty known asymptotically (AIC; Akaike, 1973)
- Resampling-based penalties
- Optimal constant unknown even in theory (change-point detection, mixture models, global/local Rademacher complexities, ...)

Goals: estimation of the optimal constant (e.g., σ^2) for estimator selection, under minimal assumptions, without overfitting

Motivation: what is the minimal penalization level?

$$\widehat{m}(D_n) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) + \mathrm{pen}(m) \right\}$$

- Optimal penalization level:

$$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)] = \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_m)\right) - \widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n)\right)\right]$$

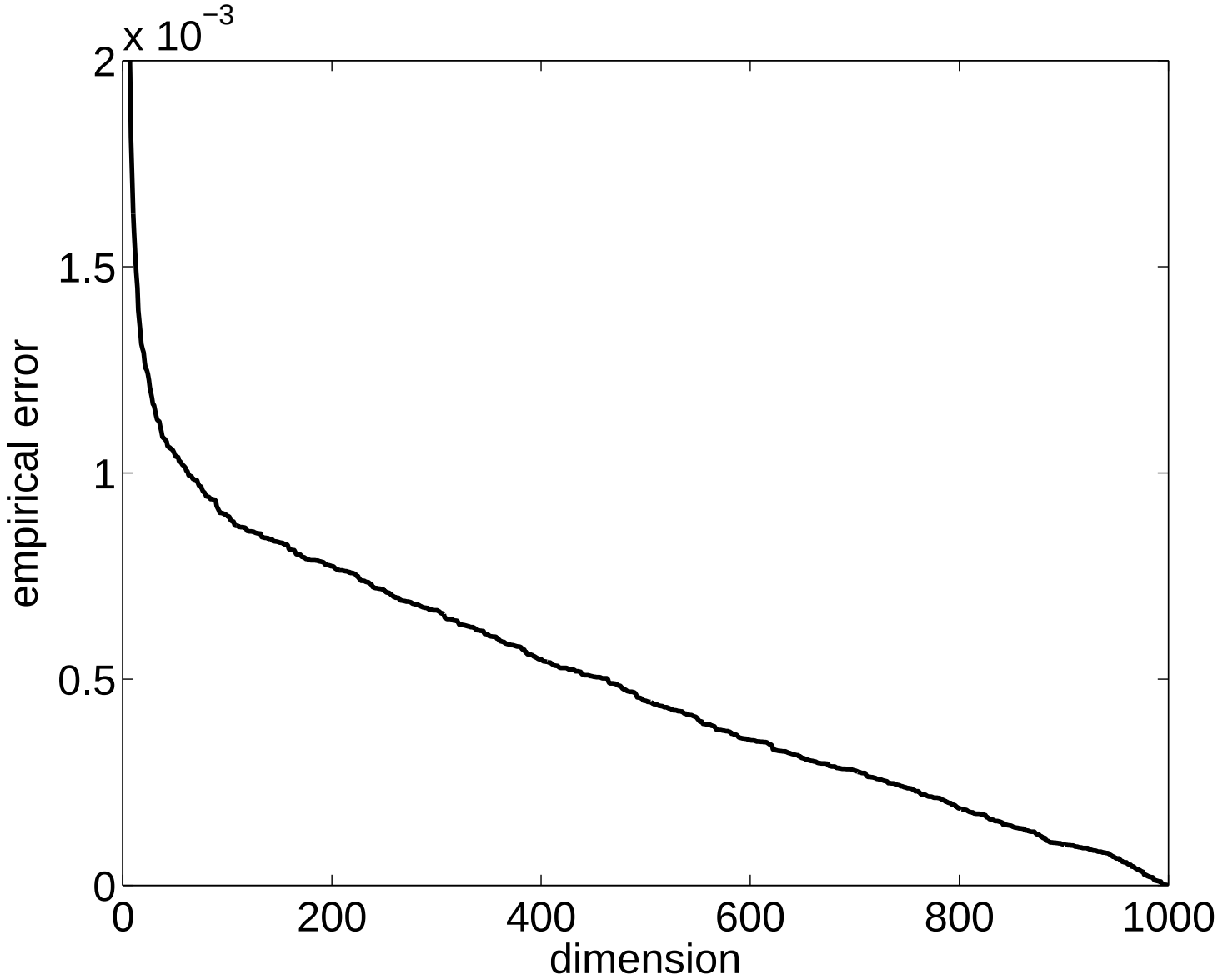
- Minimal penalization level?

  $\mathrm{pen}_{\mathrm{min}}(m) = C_{\mathrm{min}} \times \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$?

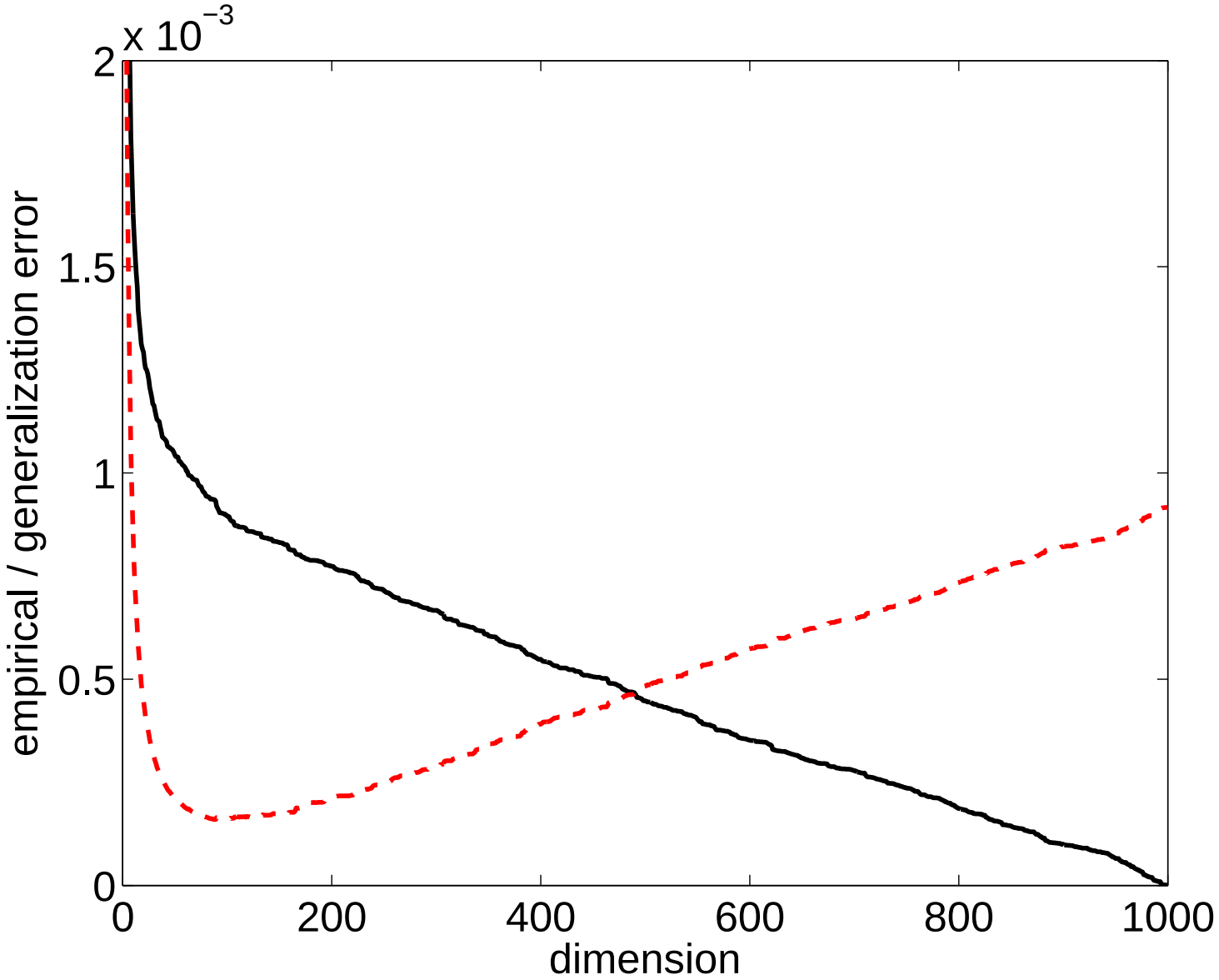
  other penalty shape?

  $\Rightarrow$ natural candidate: $\mathbb{E}[\widehat{\mathcal{R}}_n(f_m^\star) - \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right)]$ (up to the definition of $f_m^\star$)

Motivation: "L-curve" and elbow heuristics?



Motivation: "L-curve" and elbow heuristics?



General slope heuristics algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\mathrm{pen}_0(m)$ and $\mathcal{C}_m$

1. For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C\, \mathrm{pen}_0(m) \right\} .$$

2. Find $\widehat{C}_{\mathrm{jump}}$ such that $\mathcal{C}_{\widehat{m}_{\min}(C)}$ is "too large" when $C < \widehat{C}_{\mathrm{jump}}$ and "reasonably small" when $C > \widehat{C}_{\mathrm{jump}}$.

3. Select

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + 2\widehat{C}_{\mathrm{jump}}\, \mathrm{pen}_0(m) \right\} .$$

Slope heuristics: ideas for a proof

- $\exists C^\star > 0$, $C^\star \operatorname{pen}_0$ minimal penalty, $C^\star \operatorname{pen}_1$ optimal penalty.
- Decomposition of the ideal penalty:

$$\begin{aligned}\operatorname{pen}_{\mathrm{id}}(m) &= \mathcal{R}\left(\widehat{s}_m\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right) \\ &= \underbrace{\mathcal{R}\left(\widehat{s}_m\right) - \mathcal{R}\left(f_m^\star\right)}_{p_1(m)} + \underbrace{\mathcal{R}\left(f_m^\star\right) - \widehat{\mathcal{R}}_n\left(f_m^\star\right)}_{\delta(m)} + \underbrace{\widehat{\mathcal{R}}_n\left(f_m^\star\right) - \widehat{\mathcal{R}}_n\left(\widehat{s}_m\right)}_{p_2(m)}.\end{aligned}$$

- Good candidate for the minimal penalty: p_2 or $\mathbb{E}[p_2]$.
- Good candidate for the optimal penalty: $p_1 + p_2$ or $\mathbb{E}[p_1 + p_2] = \mathbb{E}[\operatorname{pen}_{\mathrm{id}}]$.
- Slope heuristics: $\mathbb{E}[p_1] \approx \mathbb{E}[p_2]$.

More details in:

S. A. *Minimal penalties and the slope heuristics: a survey*. Journal de la Société Française de Statistique, 2019, Vol 160, No 3. 1–106. arxiv:1901.07277

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) $\Rightarrow$ part I

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) $\Rightarrow$ part I
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares density estimation, i.i.d. (Lerasle, 2012) or mixing data (Lerasle, 2011)

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) $\Rightarrow$ part I
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares density estimation, i.i.d. (Lerasle, 2012) or mixing data (Lerasle, 2011)
- Density estimation, Kullback risk, maximum-likelihood estimators on histograms (Saumard, 2010)
- Minimum contrast estimator, regular contrast (Saumard, 2010)
- Specification probabilities in general random fields, least-squares/Kullback risks, empirical contrast minimizers (Lerasle & Takahashi, 2016)

Slope heuristics: theoretical results

- OLS, fixed-design regression, homoscedastic Gaussian noise (Birgé & Massart, 2007) $\Rightarrow$ part I
- OLS, random-design regression, heteroscedastic noise (regressograms, A. & Massart, 2009; piecewise polynomials, Saumard, 2013; localized basis, Navarro & Saumard 2017)
- Least-squares density estimation, i.i.d. (Lerasle, 2012) or mixing data (Lerasle, 2011)
- Density estimation, Kullback risk, maximum-likelihood estimators on histograms (Saumard, 2010)
- Minimum contrast estimator, regular contrast (Saumard, 2010)
- Specification probabilities in general random fields, least-squares/Kullback risks, empirical contrast minimizers (Lerasle & Takahashi, 2016)
- and many partial results, see arxiv:1901.07277

Slope heuristics: Empirical results

- Binary (supervised) classification (Zwald & Blanchard, 2005)
- Model-based clustering (Maugis & Michel 2011, Gallopin & Devijver 2018, and many others)
- Lasso (Connault, 2011)
- Large collection of models: Change-point detection (Lebarbier, 2005)

Slope heuristics: Empirical results

- Binary (supervised) classification (Zwald & Blanchard, 2005)
- Model-based clustering (Maugis & Michel 2011, Gallopin & Devijver 2018, and many others)
- Lasso (Connault, 2011)
- Large collection of models: Change-point detection (Lebarbier, 2005)
- and many others, see Baudry, Maugis & Michel (2011) and arXiv:1901.07277

General minimal penalty algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\mathrm{pen}_0(m)$, $\mathrm{pen}_1(m)$ and $\mathcal{C}_m$.

1. For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C\, \mathrm{pen}_0(m) \right\} .$$

2. Find $\widehat{C}_{\mathrm{jump}}$ such that $\mathcal{C}_{\widehat{m}_{\min}(C)}$ is "too large" when $C < \widehat{C}_{\mathrm{jump}}$ and "reasonably small" when $C > \widehat{C}_{\mathrm{jump}}$.

3. Select

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}}\, \mathrm{pen}_1(m) \right\} .$$

General minimal penalty algorithm

Input: $\forall m \in \mathcal{M}$, $\widehat{\mathcal{R}}_n(\widehat{s}_m)$, $\mathrm{pen}_0(m)$, $\mathrm{pen}_1(m)$ and $\mathcal{C}_m$.

1. For every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + C\,\mathrm{pen}_0(m) \right\} .$$

2. Find $\widehat{C}_{\mathrm{jump}}$ such that $\mathcal{C}_{\widehat{m}_{\min}(C)}$ is "too large" when $C < \widehat{C}_{\mathrm{jump}}$ and "reasonably small" when $C > \widehat{C}_{\mathrm{jump}}$.

3. Select

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n(\widehat{s}_m) + \widehat{C}_{\mathrm{jump}}\,\mathrm{pen}_1(m) \right\} .$$

Example (slope heuristics): $\mathrm{pen}_1 = 2\,\mathrm{pen}_0$.

General algorithm: results

Full theoretical results:

- Linear estimators, regression (OLS, k-NN, Nadaraya-Watson, kernel ridge, ...): (A. & Bach, 2009–2011)
  Fixed design regression, $\widehat{F}_m = A_m Y$
  $$\mathrm{pen}_0(m) = [2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)]/n$$
  $$\mathrm{pen}_1(m) = 2\,\mathrm{tr}(A_m)/n$$
- Linear estimators, least-squares density estimation (OLS, weighted least-squares, Parzen): Lerasle, Magalhães & Reynaud-Bouret (2016)

General algorithm: results

Full theoretical results:

- Linear estimators, regression (OLS, k-NN, Nadaraya-Watson, kernel ridge, ...): (A. & Bach, 2009–2011)
  Fixed design regression, $\widehat{F}_m = A_m Y$
  $$\mathrm{pen}_0(m) = [2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)]/n$$
  $$\mathrm{pen}_1(m) = 2\,\mathrm{tr}(A_m)/n$$
- Linear estimators, least-squares density estimation (OLS, weighted least-squares, Parzen): Lerasle, Magalhães & Reynaud-Bouret (2016)

Partial theory and empirical results for large collection of models (eg, change-point detection): Lebarbier (2005), Birgé & Massart (2007), Sorba (2017)...
$\Rightarrow \mathrm{pen}_1(m) = \kappa\,\mathrm{pen}_0(m)$ with $\kappa \neq 2$.

see arXiv:1901.07277 for details

More generalization?

- Idea: use a visible phase transition (dimension/complexity jump) for making an optimal estimator selection

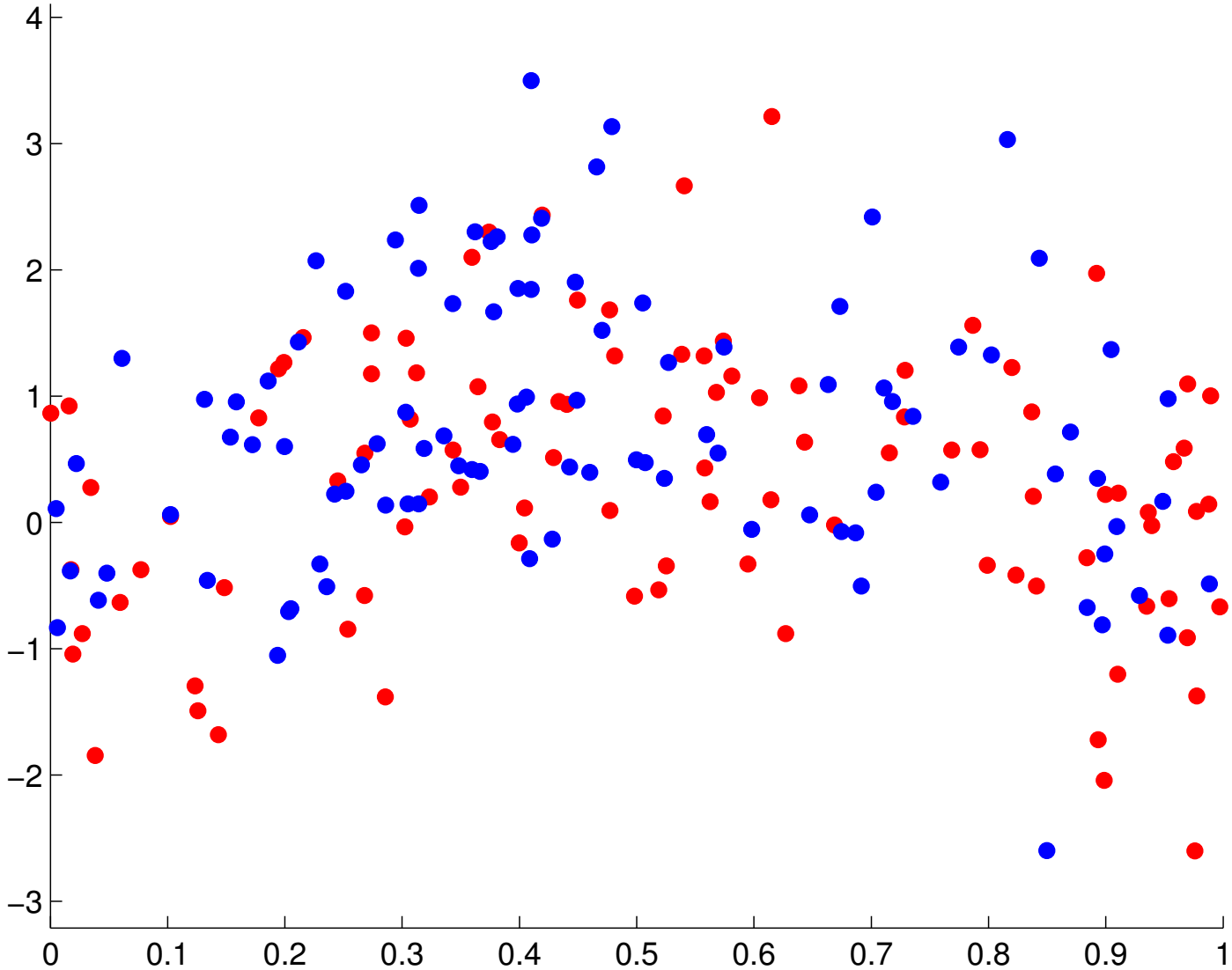
- Related procedures:
  - L-curve / elbow heuristics (no theory)
  - Scree test (no theory)
  - Thresholding under the null (requires to know null distribution)
  - Goldenshluger-Lepski's method, penalized comparison to overfitting (PCO): with theory!
  - ... (see arXiv:1901.07277)

Part III

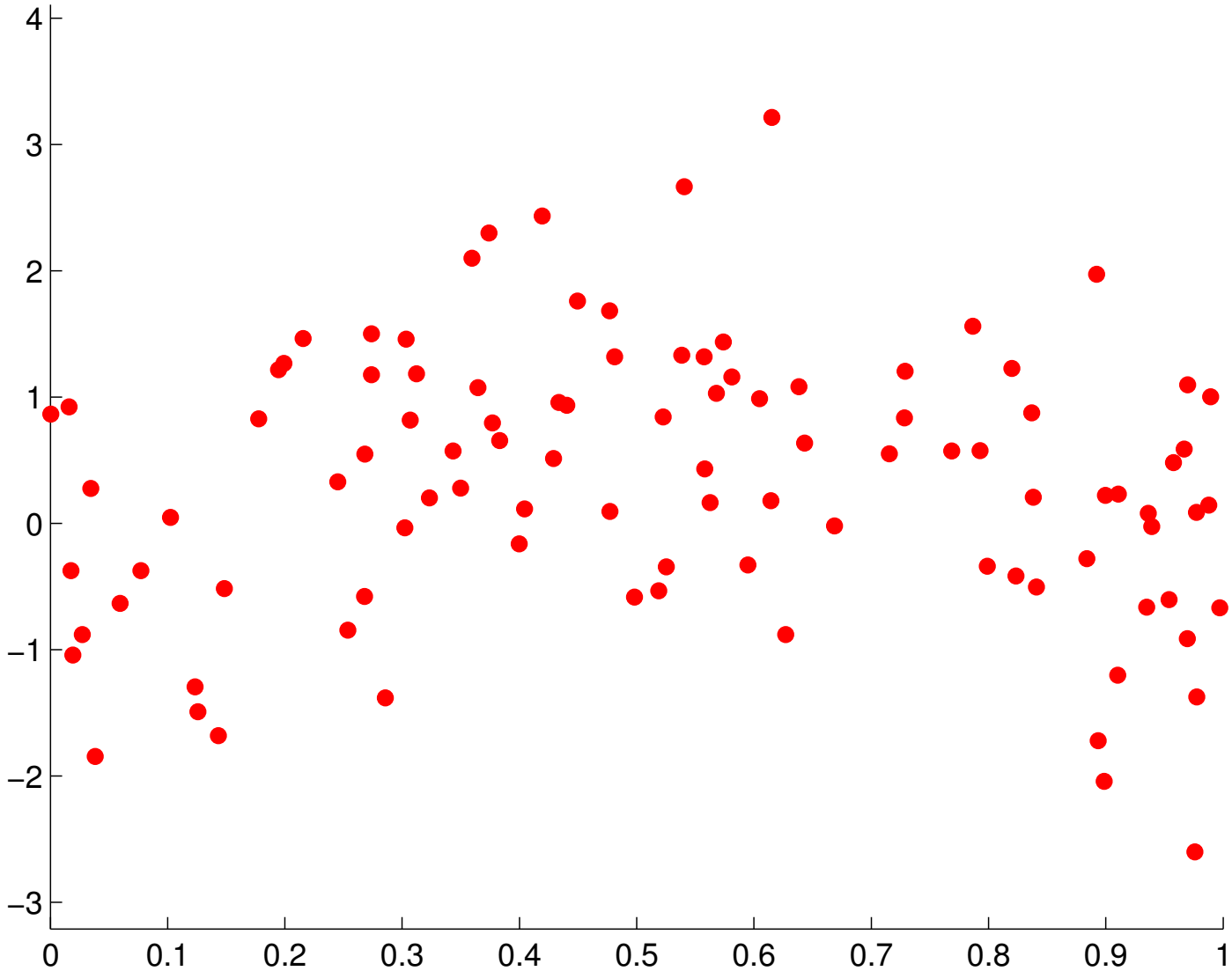
Cross-validation for estimator selection/aggregation

Outline

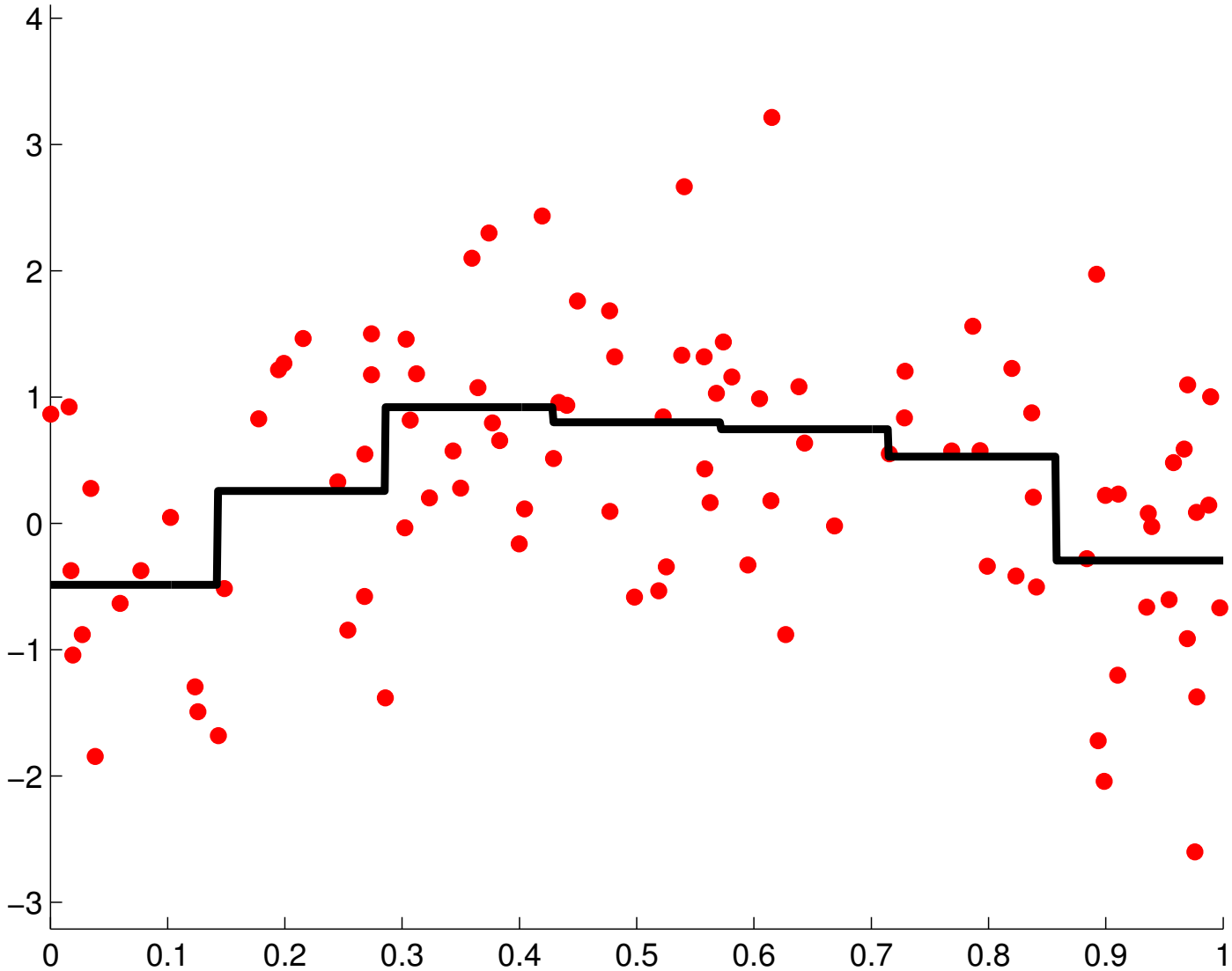
Validation principle: data splitting



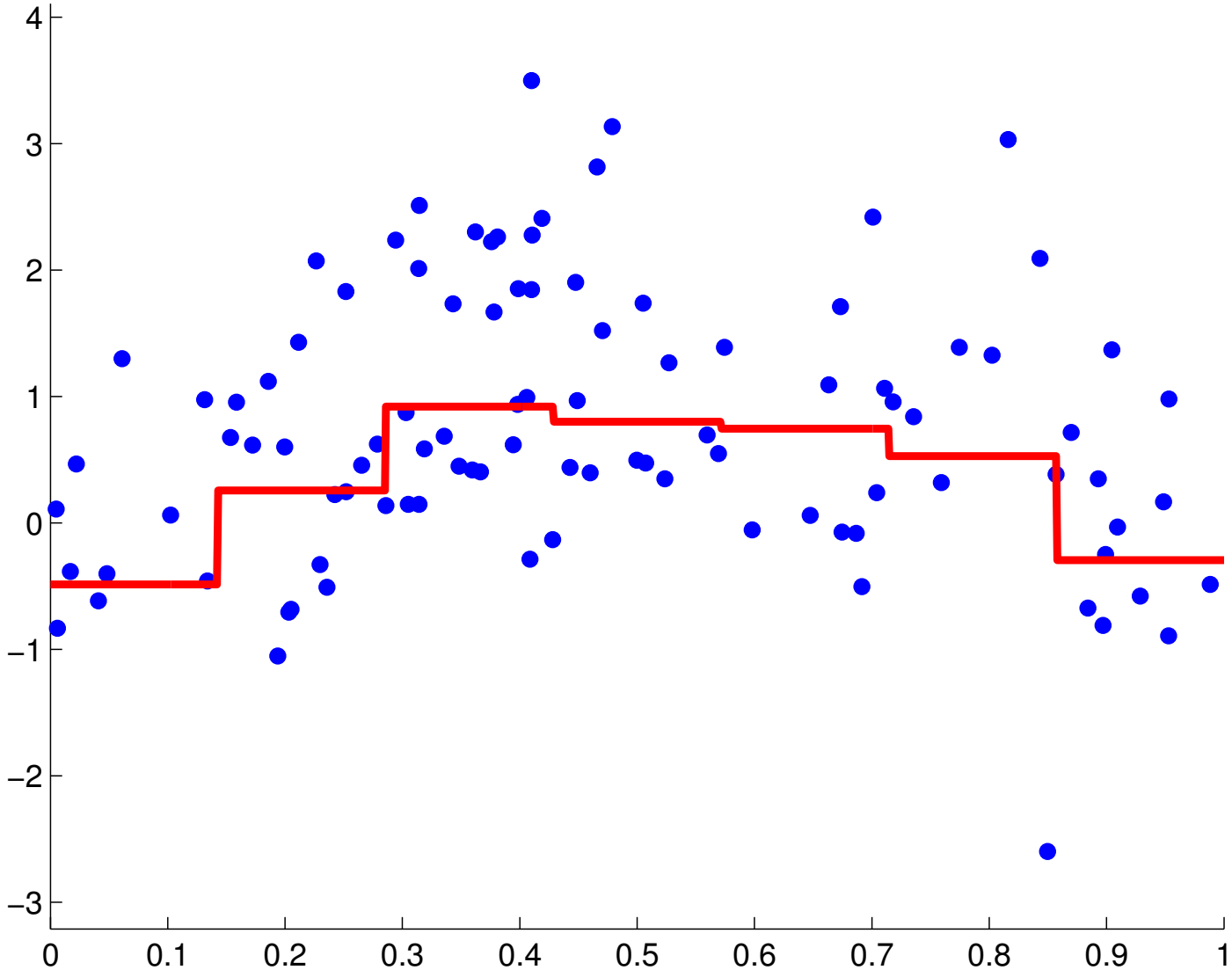
Validation principle: training/learning sample



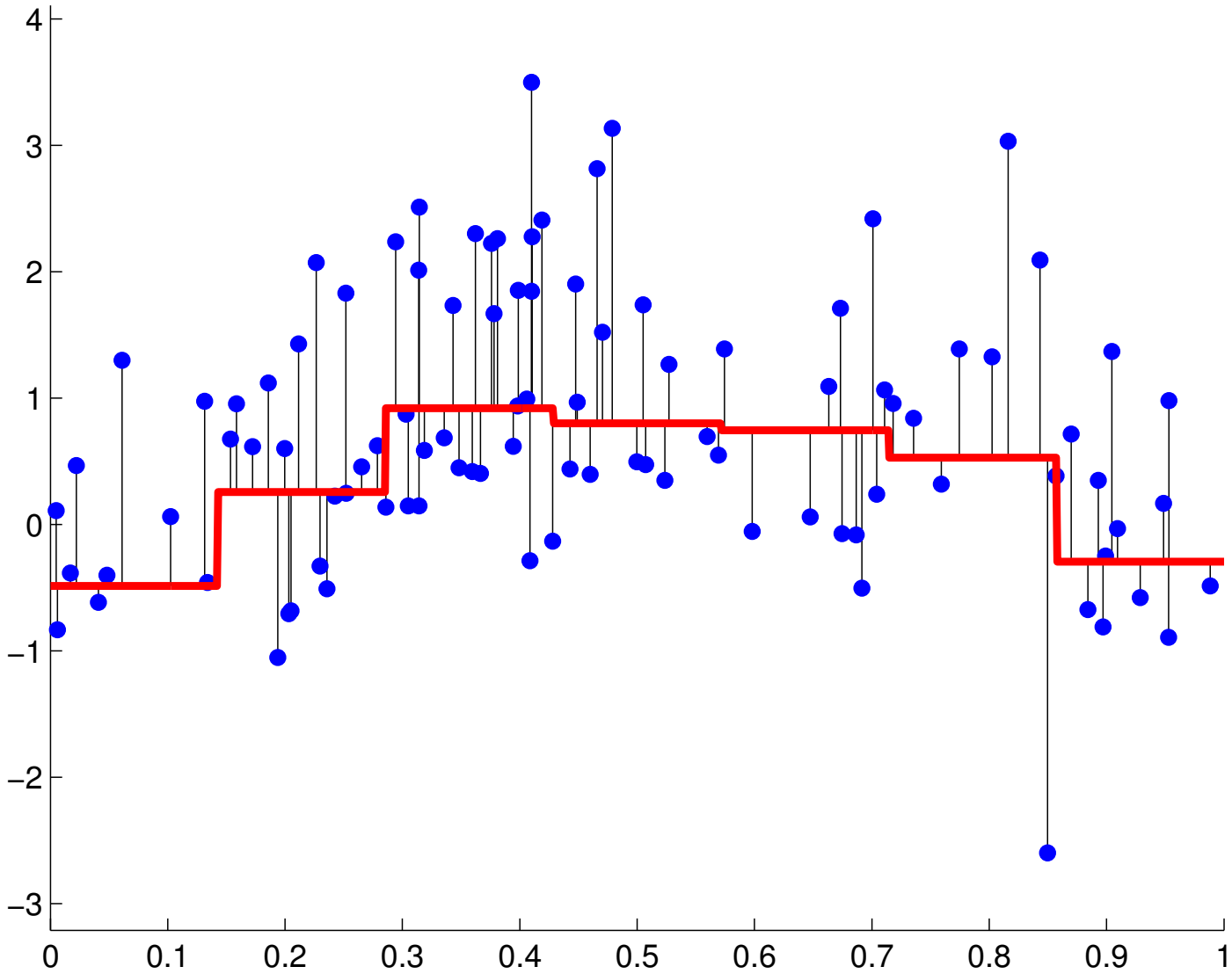
Validation principle: training/learning sample



Validation principle: validation sample



Validation principle: validation sample



Cross-validation

$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}$ $\qquad$ $\underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}$

Training set $D_n^{(t)} \Rightarrow \widehat{f}_m^{(t)} = \widehat{f}_m(D_n^{(t)})$ $\qquad$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

Cross-validation

$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}$ $\qquad$ $\underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}$

Training set $D_n^{(t)} \Rightarrow \widehat{f}_m^{(t)} = \widehat{f}_m(D_n^{(t)})$ $\qquad$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i, Y_i) \in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; (I_j^{(t)})_{1 \leqslant j \leqslant V}\right) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{(v,j)}\left(\widehat{f}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (X_i, Y_i)_{i \in I_j^{(t)}}$$

Cross-validation

$\underbrace{(X_1, Y_1), \ldots, (X_{n_t}, Y_{n_t})}$ $\qquad$ $\underbrace{(X_{n_t+1}, Y_{n_t+1}), \ldots, (X_n, Y_n)}$

Training set $D_n^{(t)} \Rightarrow \widehat{f}_m^{(t)} = \widehat{f}_m(D_n^{(t)})$ $\qquad$ Validation set $D_n^{(v)} \Rightarrow$ evaluate risk

- hold-out estimator of the risk:

$$\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right) = \frac{1}{n_v} \sum_{(X_i,Y_i)\in D_n^{(v)}} c\left(\widehat{f}_m^{(t)}(X_i); Y_i\right) \qquad n_v = |D_n^{(v)}| = n - n_t$$

- cross-validation: average several hold-out estimators

$$\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right) = \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{(v,j)}\left(\widehat{f}_m^{(t,j)}\right) \qquad D_n^{(t,j)} = (X_i, Y_i)_{i\in I_j^{(t)}}$$

- estimator selection:

$$\widehat{m}^{cv}\left(D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n\right)\right\} \quad \Rightarrow \quad \widehat{f}_{\widehat{m}^{cv}(D_n;(I_j^{(t)})_{1\leqslant j\leqslant V})}(D_n)$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

- V-fold cross-validation: $\mathcal{B} = (B_j)_{1 \leqslant j \leqslant V}$ partition of $\{1, \ldots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) = \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n^j\left(\widehat{f}_m^{(-j)}\right)$$

Cross-validation: examples

- Exhaustive data splitting: all possible subsets of size n_t
  $\Rightarrow$ leave-one-out ($n_t = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}\left(\widehat{f}_m; D_n\right) = \frac{1}{n}\sum_{j=1}^{n} c\left(\widehat{f}_m^{(-j)}(X_j); Y_j\right)$$

  $\Rightarrow$ leave-p-out ($n_t = n - p$)

- V-fold cross-validation: $\mathcal{B} = (B_j)_{1 \leqslant j \leqslant V}$ partition of $\{1, \ldots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) = \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n^j\left(\widehat{f}_m^{(-j)}\right)$$

- Monte-Carlo CV / Repeated learning testing:

$$I_1^{(t)}, \ldots, I_V^{(t)} \text{ i.i.d. uniform}$$

Outline

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\left(\widehat{f}_m(D_n)\right)$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\left(\widehat{f}_m(D_n)\right)$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{CV}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant V}\right)\right] = \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_{n_t})\right)\right]$$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\left(D_n^{(t,j)}\right) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\left(\widehat{f}_m(D_n)\right)$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{CV}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant V}\right)\right] = \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_{n_t})\right)\right]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_n)\right)\right]$

Bias of cross-validation

- In this talk, we always assume: $\forall j$, $\mathrm{Card}\big(D_n^{(t,j)}\big) = n_t$
  For V-fold CV: $\mathrm{Card}(B_j) = n/V \Rightarrow n_t = n(V-1)/V$.
- Ideal criterion: $\mathcal{R}\big(\widehat{f}_m(D_n)\big)$
- General analysis for the bias:

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant V}\right)\right] = \mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_{n_t})\big)\Big]$$

$\Rightarrow$ everything depends on $n \to \mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_n)\big)\Big]$

- Note: bias can be corrected in some settings (Burman, 1989).
- Note: $D_n \to \widehat{f}_m(D_n)$ must be fixed before seeing any data; otherwise (e.g., data-driven model m), stronger bias.

Bias of cross-validation: generic example

Assume:

$$\mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

Bias of cross-validation: generic example

Assume:

$$\mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., LS/ridge/k-NN regression, LS/kernel density estimation).

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant V}\right)\right] = \alpha(m) + \frac{n}{n_t}\frac{\beta(m)}{n}$$

$\Rightarrow$ Bias:

- decreases as a function of n_t,
- minimal for $n_t = n-1$,
- negligible if $n_t \sim n$.

$\Rightarrow$ V-fold: bias decreases when V increases, vanishes as $V \to +\infty$.

Variance of cross-validation: general case

- Hold-out (Nadeau & Bengio, 2003):

$$\operatorname{var}\left(\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right)\right) = \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(c(f(X),Y)\,\middle|\, f=\widehat{f}_m^{(t)}\right)\right] + \operatorname{var}\left(\mathcal{R}\left(\widehat{f}_m(D_{n_t})\right)\right)$$

- Monte-Carlo CV and number of splits: $(p = n - n_t)$

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m; D_n; \left(I_j^{(t)}\right)_{1\leqslant j\leqslant V}\right)\right) = \operatorname{var}\left(\widehat{\mathcal{R}}^{\ell\mathrm{po}}\left(\widehat{f}_m; D_n\right)\right) + \frac{1}{V}\underbrace{\mathbb{E}\left[\operatorname{var}_{I^{(t)}}\left(\widehat{\mathcal{R}}_n^{(v)}\left(\widehat{f}_m^{(t)}\right)\,\middle|\, D_n\right)\right]}_{\text{permutation variance}}$$

- V-fold CV: V, n_t, n_v related
  leave-one-out: related to stability? (empirical results)

Variance of V-fold CV criterion

- Least-squares density estimation (A. & Lerasle, 2016), exact computation (non-asymptotic):

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right) = \frac{1+\mathcal{O}(1)}{n} \operatorname{var}_P(f_m^\star) + \frac{2}{n^2}\left[1 + \frac{4}{V-1} + \mathcal{O}\left(\tfrac{1}{V}+\tfrac{1}{n}\right)\right] A(m)$$

(simplified formula, histogram model with bin size d_m^{-1}, $A(m) \approx d_m$)

- Linear regression, asymptotic formula (Burman, 1989):

$$\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right)\right) = \frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + \mathrm{o}\left(n^{-2}\right)$$

$\Rightarrow$ decreasing with V, dependence only in second order terms.

Outline

Risk estimation and estimator selection are different goals

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}} \left(\widehat{f}_m \right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}\left(\widehat{f}_m(D_n) \right) \right\}$$

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}} \left(\widehat{f}_m \right) + Z \right\}$$

$\Rightarrow$ bias and variance meaningless.

Risk estimation and estimator selection are different goals

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}} \left(\widehat{f}_m \right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}\big(\widehat{f}_m(D_n)\big) \right\}$$

- For any Z (deterministic or random),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{cv}} \left(\widehat{f}_m \right) + Z \right\}$$

  $\Rightarrow$ bias and variance meaningless.

- Perfect ranking among $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\operatorname{sign}\big(\widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{cv}}(\widehat{f}_{m'})\big) = \operatorname{sign}\big(\mathcal{R}(\widehat{f}_m) - \mathcal{R}(\widehat{f}_{m'})\big)$$

$\Rightarrow$ $\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_{m'}\right) \right]$ should be of the good sign (unbiased risk estimation heuristic: AIC, C_p, leave-one-out...)

$\Rightarrow$ $\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{cv}}\left(\widehat{f}_{m'}\right) \right)$ should be minimal (detailed heuristic: A. & Lerasle, 2016)

CV with an estimation goal: the big picture ($\mathcal{M}$ "small")

- At first order, the bias drives the performance of:
  leave-p-out, V-fold CV,
  Monte-Carlo CV if $V \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_{n_t}) \right) \right] \right\}$$

CV with an estimation goal: the big picture ($\mathcal{M}$ "small")

- At first order, the bias drives the performance of:
  leave-p-out, V-fold CV,
  Monte-Carlo CV if $V \gg n^2$
  or if n_v large enough (including hold-out)
- CV performs similarly to

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_m(D_{n_t}) \right) \right] \right\}$$

$\Rightarrow$ first-order optimality if $n_t \sim n$

$\Rightarrow$ suboptimal otherwise
e.g., V-fold CV with V fixed.

- Theoretical results for least-squares regression and density estimation at least.

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) + \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) - \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n\left(\widehat{f}_m^{(-j)}\right)$$

$$= \widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n)\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B})}_{V\text{-fold penalty (A. 2008)}} .$$

Bias-corrected VFCV / V-fold penalization

- Bias-corrected V-fold CV (Burman, 1989):

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) := \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{f}_m; D_n; \mathcal{B}\right) + \widehat{\mathcal{R}}_n\left(\widehat{f}_m\right) - \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n\left(\widehat{f}_m^{(-j)}\right)$$

$$= \widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n)\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B})}_{V\text{-fold penalty (A. 2008)}} .$$

- In least-squares density estimation (A. & Lerasle, 2016):

$$\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{f}_m; D_n; \mathcal{B}) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \underbrace{\left(1 + \frac{1}{2(V-1)}\right)}_{\text{overpenalization factor}} \mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B})$$

$$\widehat{\mathcal{R}}^{\ell\mathrm{po}}(\widehat{f}_m; D_n; \mathcal{B}) = \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n)) + \overbrace{\left(1 + \frac{1}{2\left(\frac{n}{p}-1\right)}\right)}^{\text{overpenalization factor}} \mathrm{pen}_{\mathrm{VF}}(\widehat{f}_m; D_n; \mathcal{B}_{\mathrm{loo}}) .$$

Variance and estimator selection

$$\Delta(m, m', V) = \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{f}_{m'}\right)$$

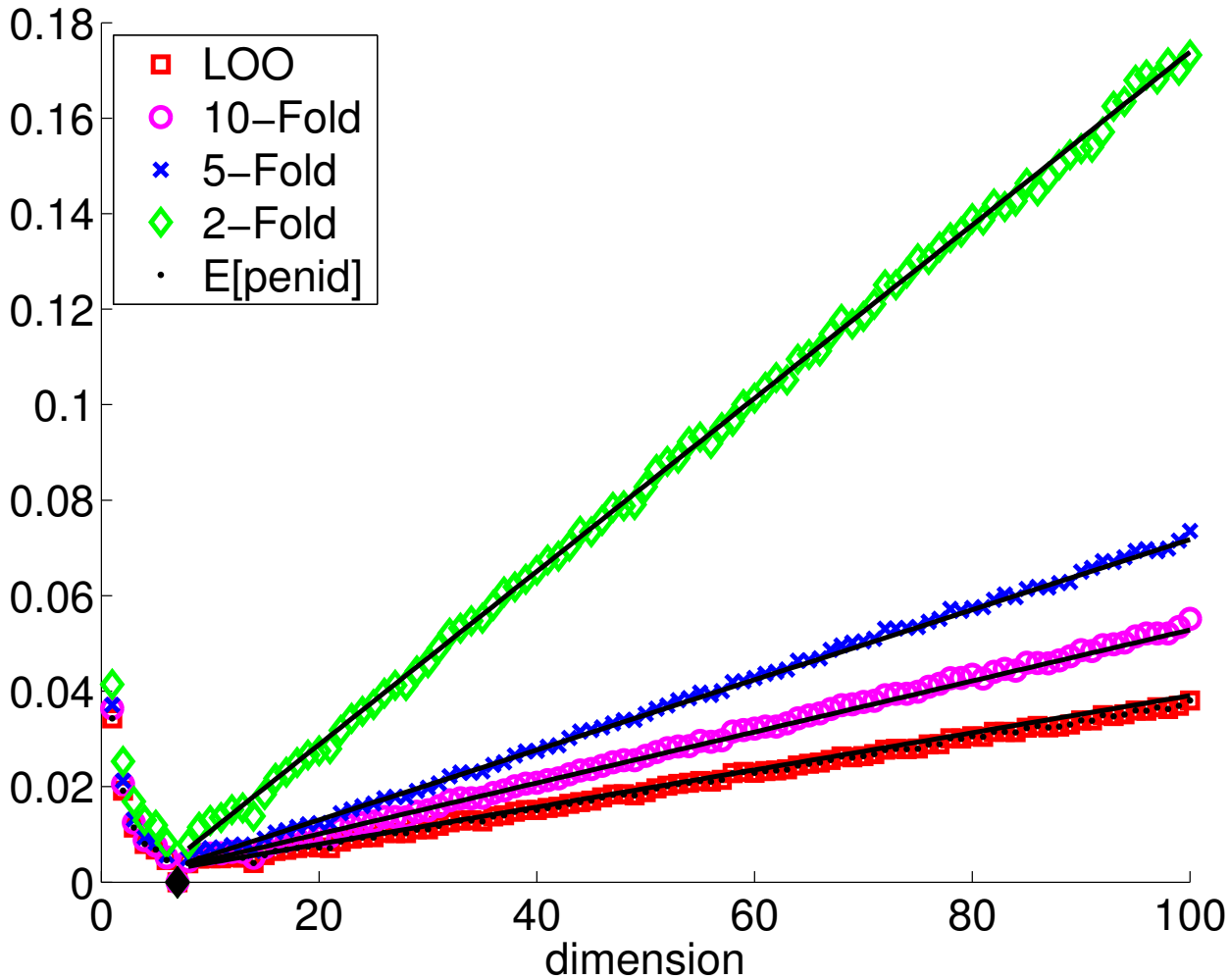
Theorem (A. & Lerasle, 2016, least-squares density estimation)

$$\mathrm{var}\left(\Delta(m, m', V)\right) = 4\left(1 + \frac{2}{n} + \frac{1}{n^2}\right)\frac{\mathrm{var}_P\left(f^\star_m - f^\star_{m'}\right)}{n} + 2\left(1 + \frac{4}{V-1} - \frac{1}{n}\right)\underbrace{\frac{B(m, m')}{n^2}}_{\geqslant 0}$$

If $S_m \subset S_{m'}$ are two histogram models with constant bin sizes $d_m^{-1}, d_{m'}^{-1}$, then, $B(m, m') \propto \|f^\star_m - f^\star_{m'}\|\, d_m$.

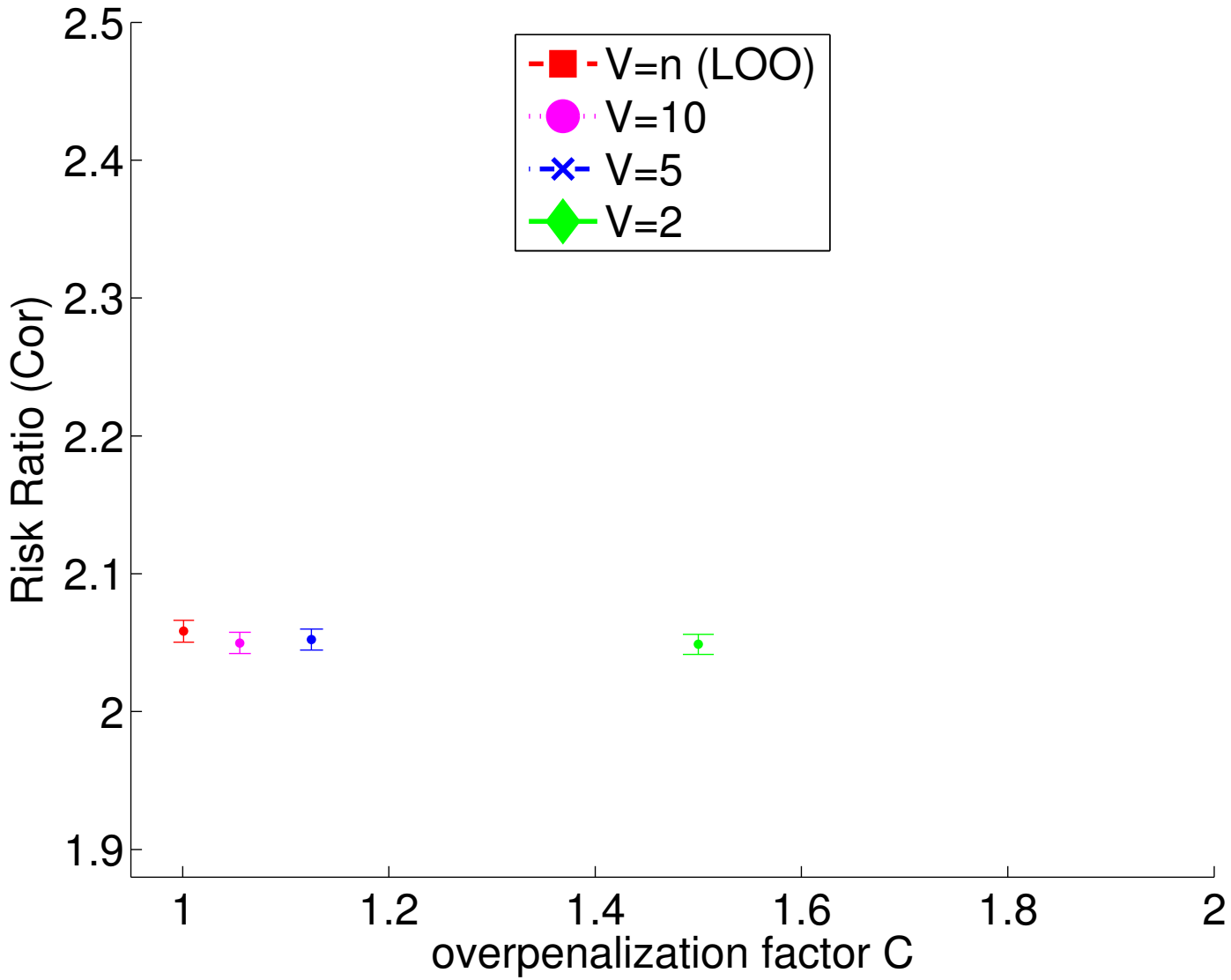
The two terms are of the same order if $\|f^\star_m - f^\star_{m'}\| \approx d_m/n$.

Variance of $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{f}_{m^\star})$ vs. (d_m, V)

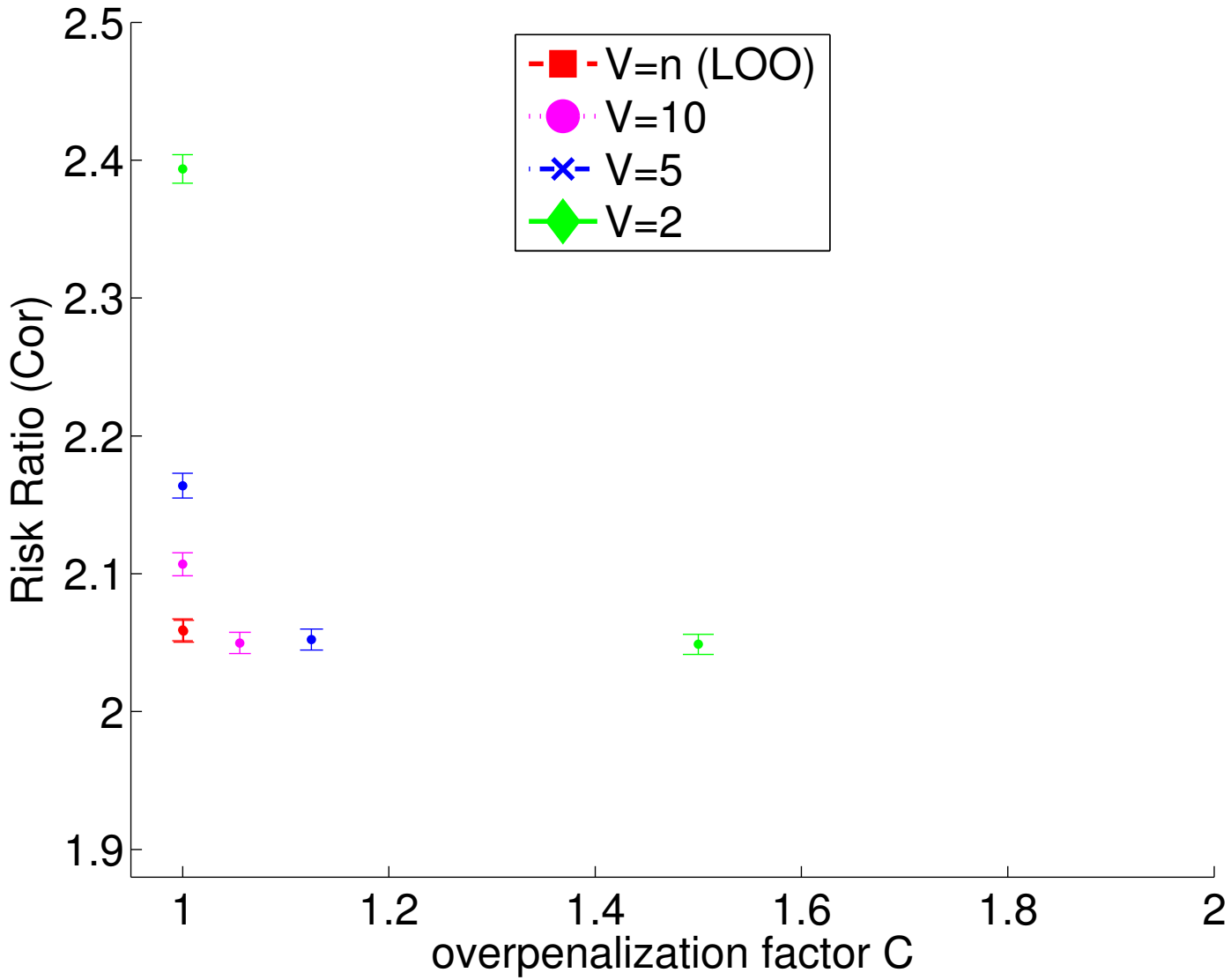


$$\mathrm{var}(\Delta(m, m', V)) \approx n^{-2}\left[29\left(1 + \tfrac{0.8}{V-1}\right) + 3.7\left(1 + \tfrac{3.8}{V-1}\right)(d_m - d_{m^\star})\right]$$

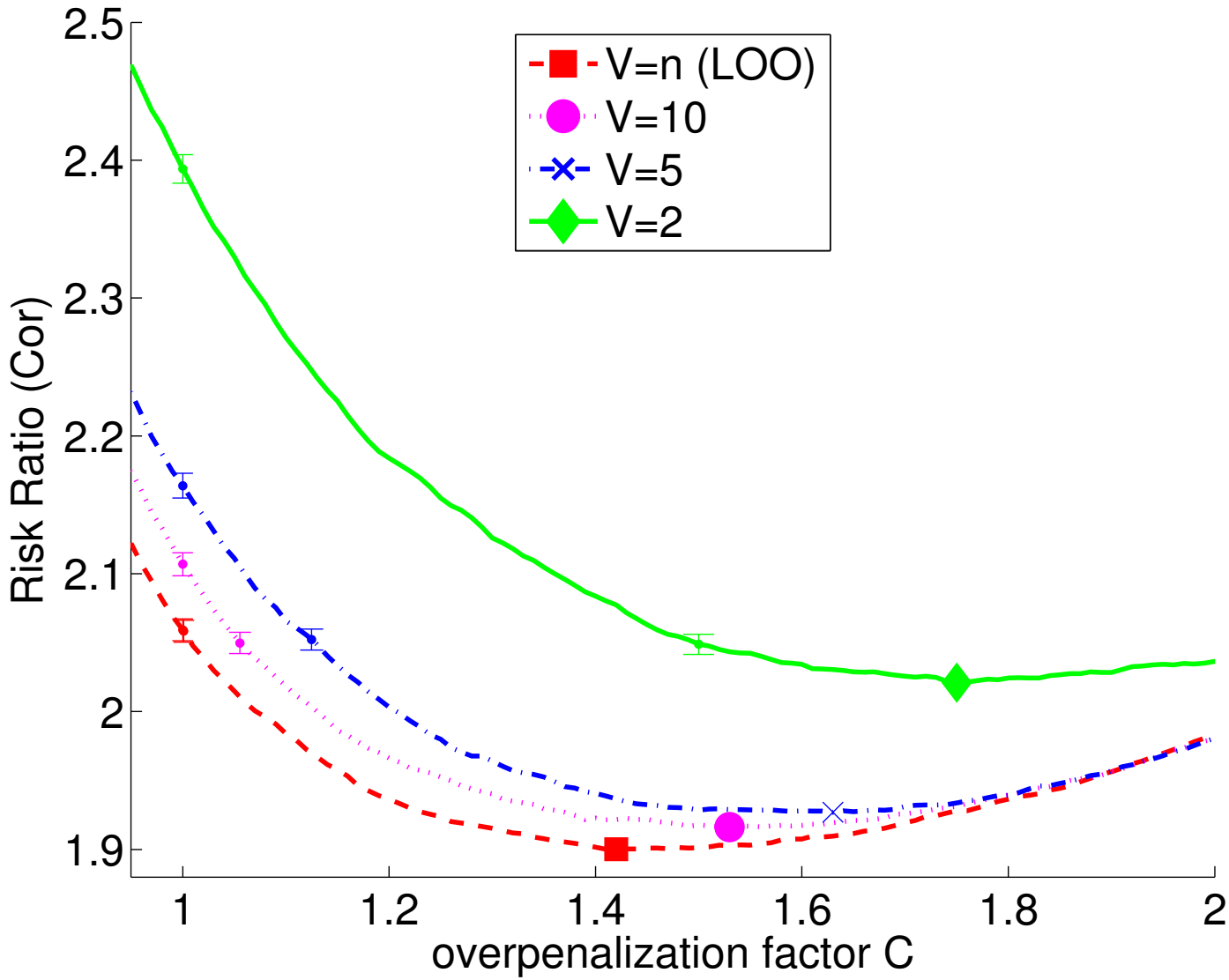
Experiment (LS density estimation): V-fold CV



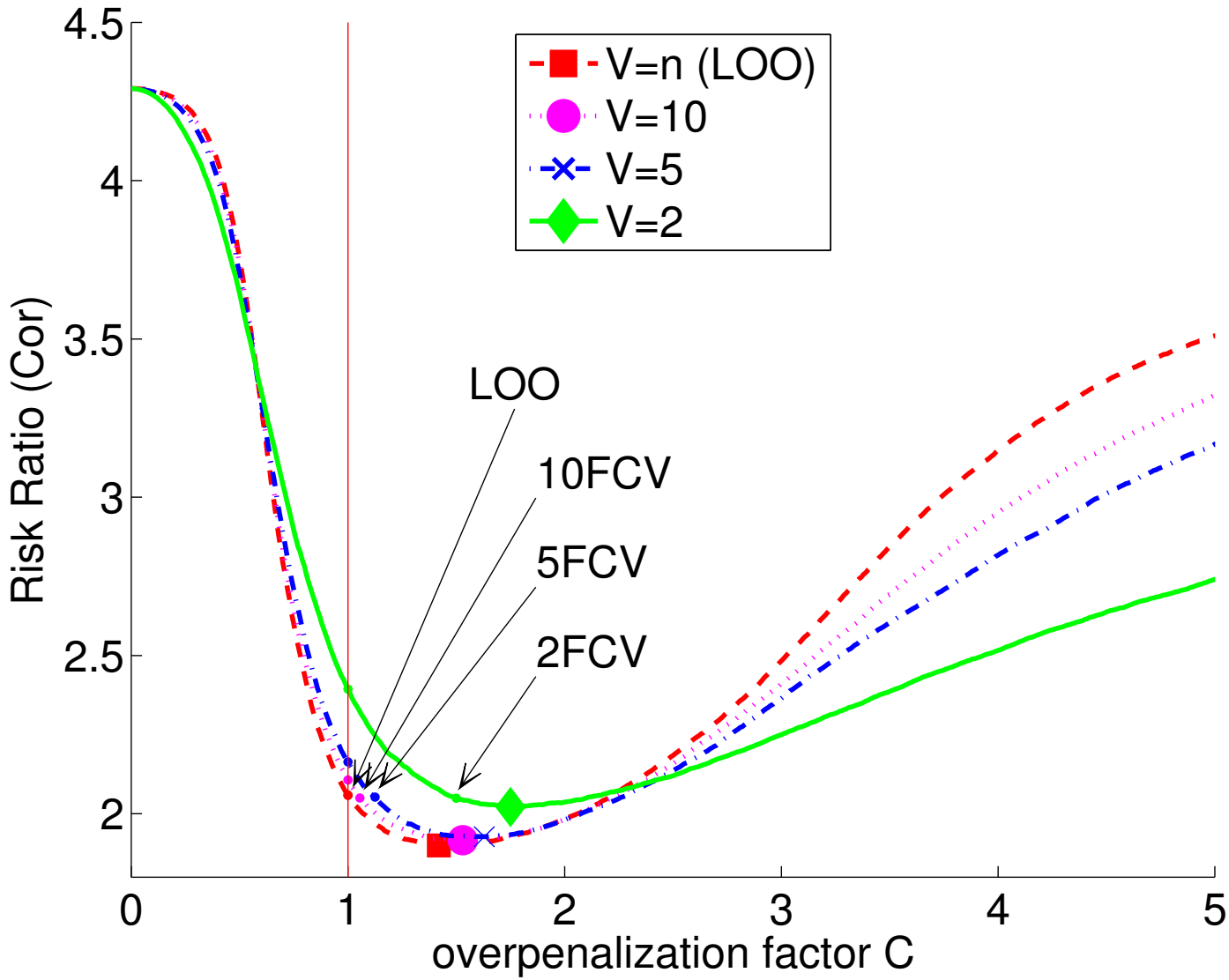
Experiment (LS density estimation): V-fold penalization



Experiment (LS density estimation): overpenalization



Experiment (LS density estimation): conclusion



Outline

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
  - Bias: decreases with V / can be removed
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$

  $\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
  - Bias: decreases with V / can be removed
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$

  $\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...

  ... if optimal overpenalization factor $C^\star \approx 1$ (various behaviours possible).

Estimator selection with V-fold: conclusion

- Computational complexity: $\mathcal{O}(V)$ in general
- V-fold cross-validation:
  - Bias: decreases with V / can be removed
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$

  $\Rightarrow$ best performance for the largest V and almost optimal with $V = 10$...
  ... if optimal overpenalization factor $C^\star \approx 1$ (various behaviours possible).
- V-fold penalization:
  - Decoupling of bias and variance $\Rightarrow$ easier to understand.
  - Bias: chosen directly through C, without any constraint.
  - Variance: decreases with V / almost minimal with $V \in [5, 10]$.

How general are these conclusions? (i.i.d. case)

- At least valid for least-square regression / density estimation, kernel density estimation.
- Bias-correction / V-fold penalization: valid if

$$\mathbb{E}\Big[(\mathcal{R} - \widehat{\mathcal{R}}_n)(\widehat{f}_m)\Big] \approx \frac{\gamma(m)}{n} \quad .$$

  Otherwise: use repeated V-fold or Monte-Carlo CV with a well-chosen n_t.
- Variance: different behaviours can occur in other settings (experiments).
- Everything can be checked on synthetic data: plot

$$n \to \mathbb{E}\Big[\mathcal{R}\big(\widehat{f}_m(D_n)\big)\Big] \qquad \text{and} \qquad m \to \operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{CV}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{CV}}(\widehat{f}_{m^\star})\Big) \, .$$

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent $\Rightarrow$ CV heuristic fails!

$\Rightarrow$ possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

Dependent data

- $D_n^{(t)}, D_n^{(v)}$ dependent $\Rightarrow$ CV heuristic fails!

$\Rightarrow$ possible troubles for risk estimation (Hart & Wehrly, 1986; Opsomer et al., 2001).

- Solution for short-term dependence:
  remove some data at each split $\Rightarrow$ gap between training and validation samples.

Cross-validation with an identification goal

- Main change: value of the optimal overpenalization factor $C^\star$, often $C^\star \to +\infty$ when $n \to +\infty$.

$\Leftrightarrow$ Cross-validation paradox (Yang, 2006, 2007): $n_t \ll n$ can be necessary!

- Why? Smaller $n_t \Rightarrow$ easier to distinguish the two best procedures... if n_t large enough (asymptotic regime).

- Remark: estimation goal, parametric setting $\Rightarrow$ similar behaviour.

Large collection of estimators/models

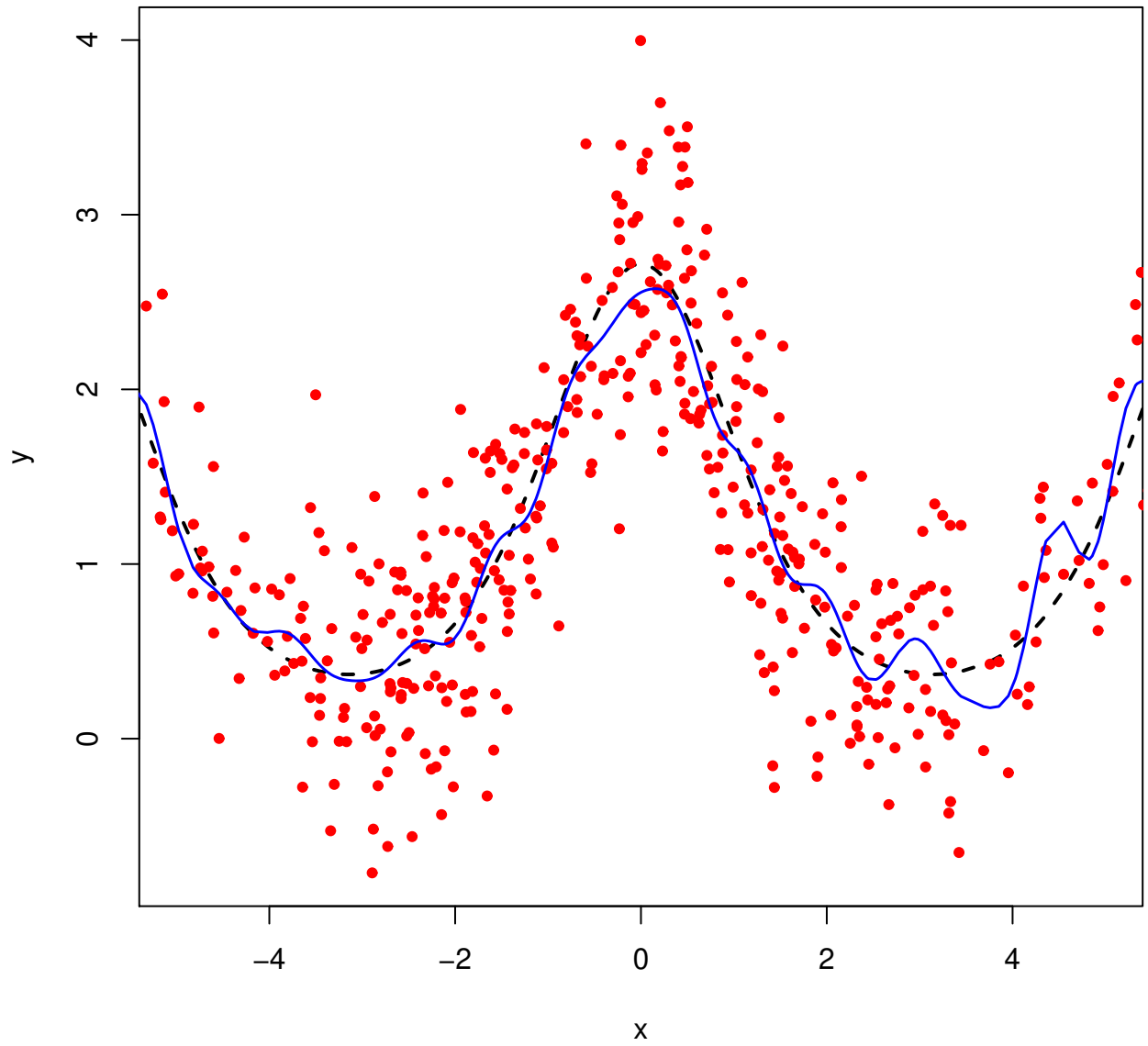
- Estimator/model selection with an "exponential collection" (implicitly excluded in all results above).
  $\Rightarrow$ Expectations do not drive the first order!

- Examples: variable selection with $p \geqslant n$ variables, change-point detection.

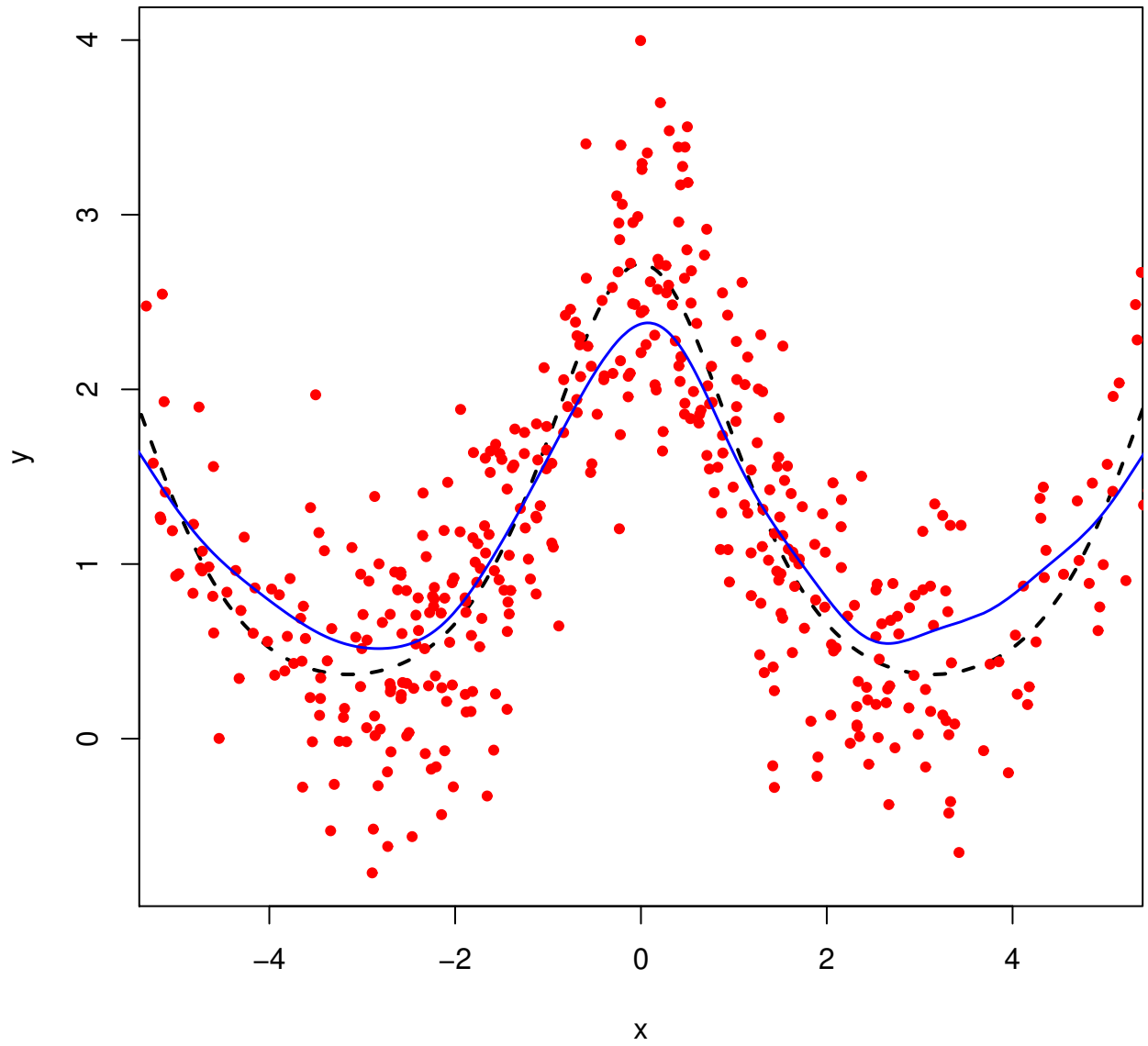
- Solution: group the models $\Rightarrow$ one estimator per "dimension" (e.g., empirical risk minimizer)
  works for change-point detection (A. & Celisse, 2010).

Outline

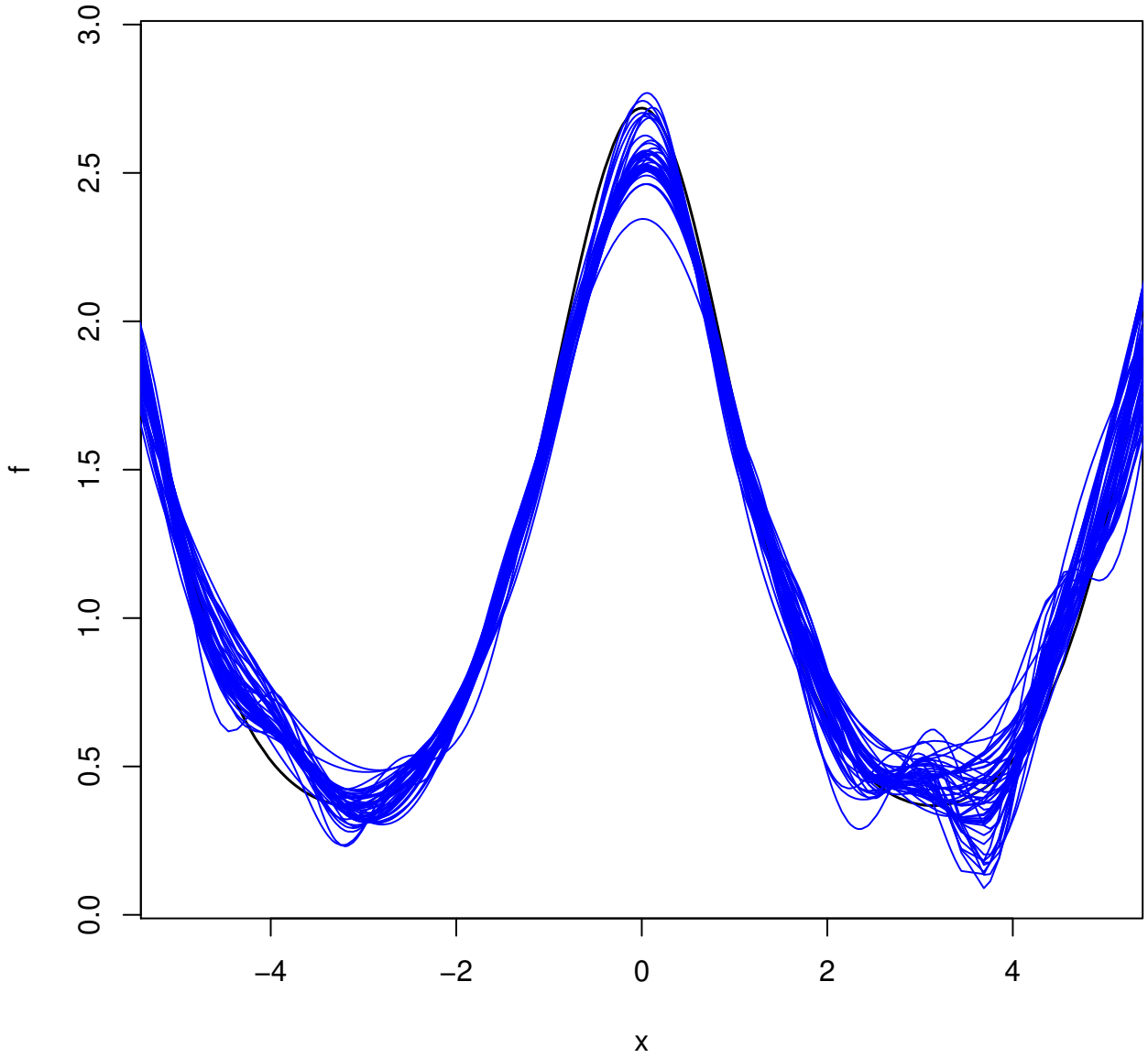
Example: regression, ϵ-SVM estimator (undersmoothed)



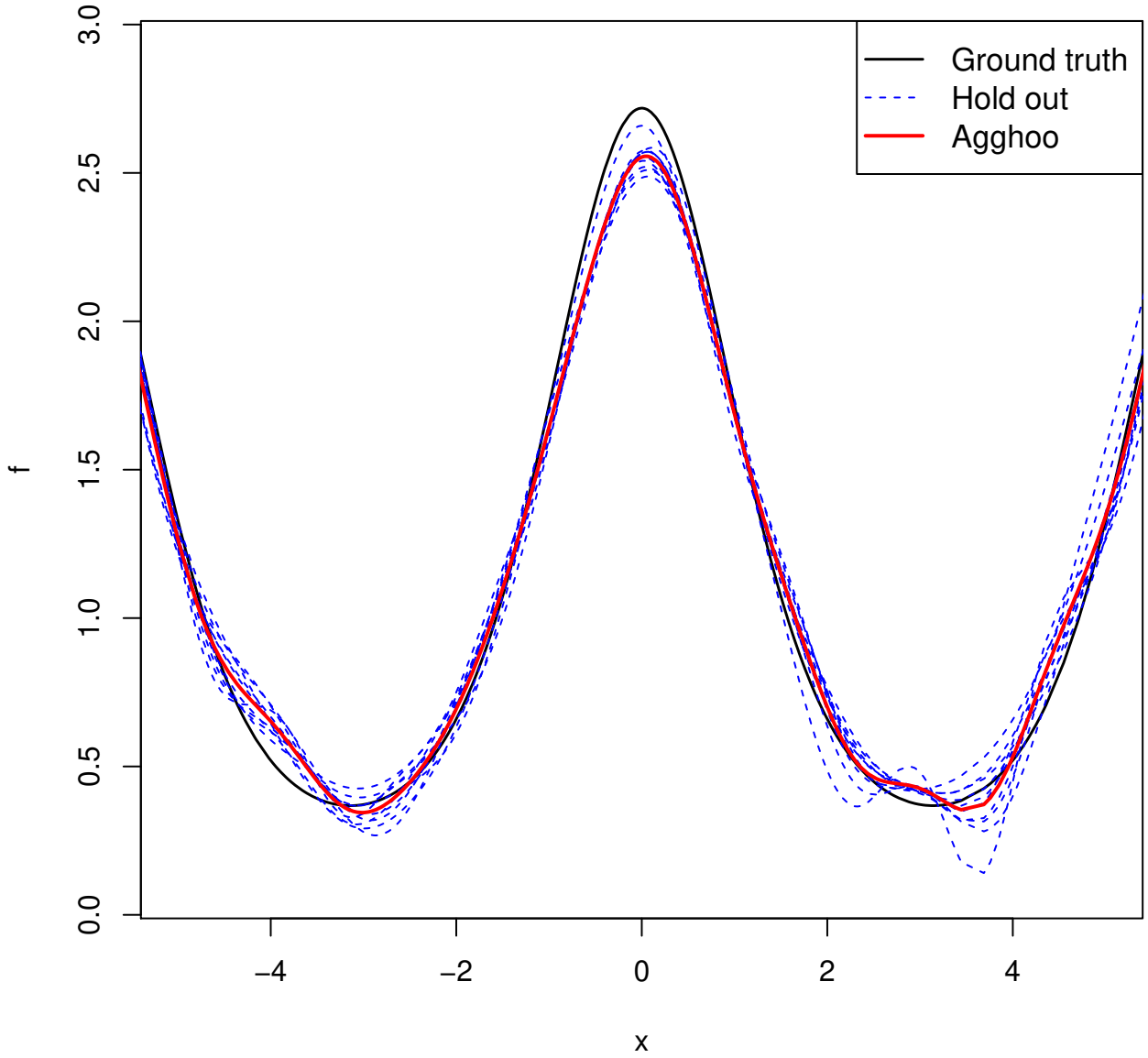
Example: regression, ϵ-SVM estimator (oversmoothed)



Example: regression, ϵ-SVM: several hold-out estimators



Example: regression, ϵ-SVM: aggregated hold-out



Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(D_n^{(t,j)}\right)$$

Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(D_n^{(t,j)}\right)$$

- If $\mathcal{Y}$ is finite (classification):

$$\widehat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(x; D_n^{(t,j)}\right) / \, j = 1, \ldots, V \right\}$$

Aggregated hold-out (Agghoo): definition

- Idea: aggregate several hold-out estimators.
- If $\mathcal{Y}$ is convex (e.g., regression):

$$\widehat{f}^{agghoo} = \frac{1}{V} \sum_{j=1}^{V} \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(D_n^{(t,j)}\right)$$

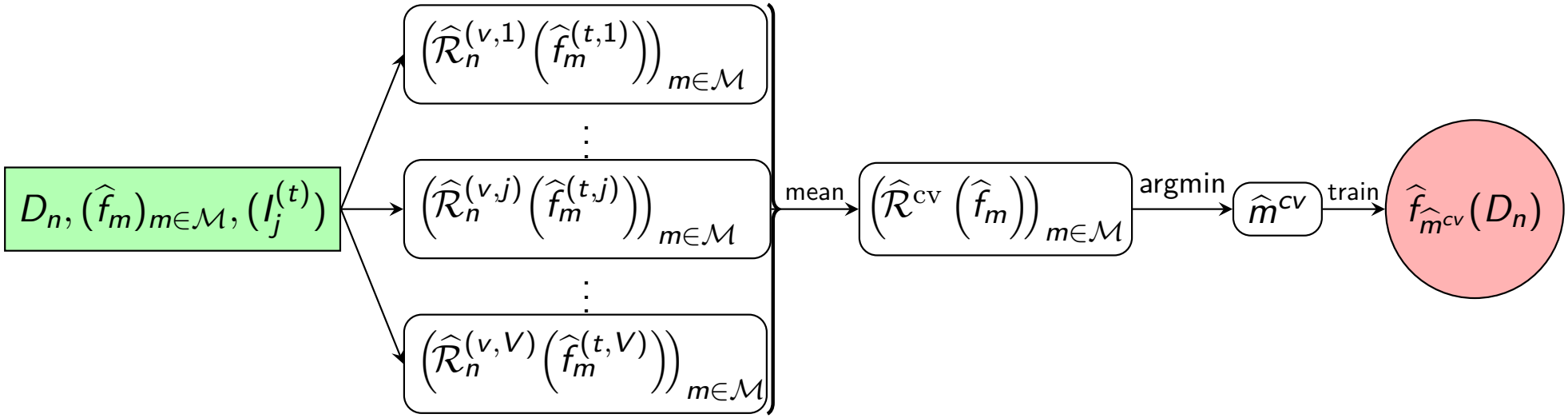
- If $\mathcal{Y}$ is finite (classification):

$$\widehat{f}^{agghoo} : x \mapsto \text{majority vote among } \left\{ \widehat{f}_{\widehat{m}^{ho}(I_j^{(t)})}\left(x; D_n^{(t,j)}\right) / \, j = 1, \ldots, V \right\}$$

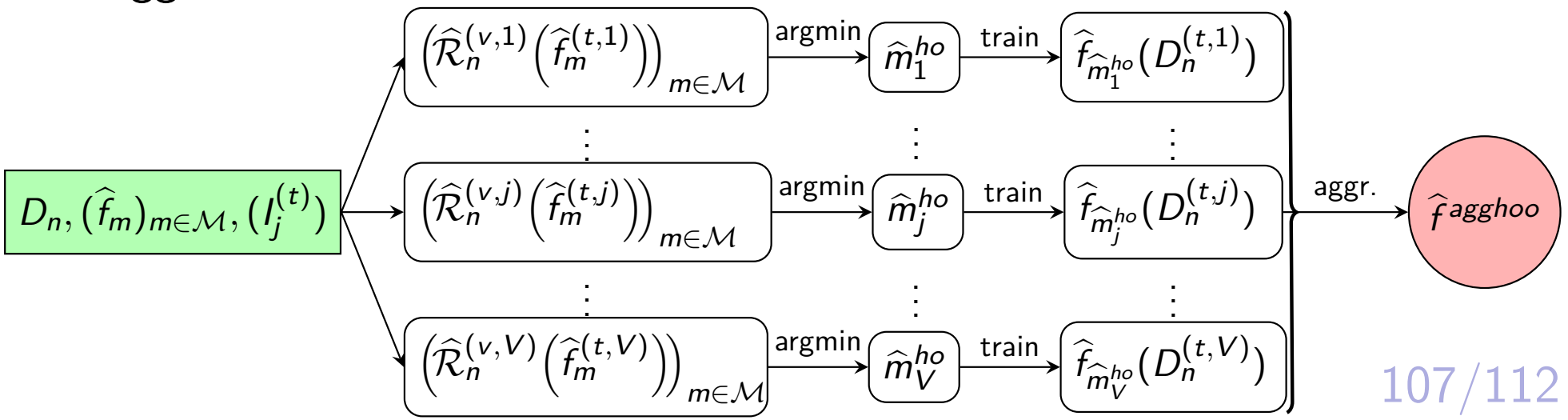
- Usual assumption: $\forall j \in \{1, \ldots, V\}$, $\mathrm{Card}(I_j^{(t)}) = \tau n$.
- Remark: $V = 1 \Rightarrow \widehat{f}^{agghoo} = \widehat{f}_{\widehat{m}^{ho}}(D_n^t) \approx$ hold-out estimator

Agghoo and cross-validation

Cross-validation:



Agghoo:



Related procedures

- CV bagging (data science folklore)
  - hold-out + subagging $\neq$ agghoo: for $j = 1, \ldots, V$,

hold-out + subagging = hold-out on subsamples

$$\underbrace{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k})}_{\text{train}}, \underbrace{(X_{i_{k+1}}, Y_{i_{k+1}}), \ldots, (X_{i_\ell}, Y_{i_\ell})}_{\text{validation}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \ldots, (X_{i_n}, Y_{i_n})}_{\text{unused}}$$

agghoo = hold-out on different splits

$$\underbrace{(X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k}), (X_{i_{k+1}}, Y_{i_{k+1}}), \ldots, (X_{i_\ell}, Y_{i_\ell})}_{\text{train}}, \underbrace{(X_{i_{\ell+1}}, Y_{i_{\ell+1}}), \ldots, (X_{i_n}, Y_{i_n})}_{\text{validation}}$$

  - "CV bagging" also used for procedures close to agghoo

- Averaging of the chosen *parameters* $\widehat{m}_j^{ho}$:
  - K-fold averaging cross-validation (ACV; Jung and Hu, 2015)
  - efficient K-fold cross-validation (EKCV; Jung, 2016)

Performance of agghoo: theory

- Regression with c convex: if $\forall j$, $\mathrm{Card}(I_j^{(t)}) = \tau n$,

$$\forall V \geqslant 1, \qquad \mathbb{E}\left[\mathcal{R}\left(\widehat{f}^{agghoo}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right)\right)\right]$$
$$\leqslant \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_{\widehat{m}^{ho}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; I_1^{(t)}\right)}(D_n^{t,1})\right)\right]$$

  Corollary: oracle inequalities for the hold-out $\Rightarrow$ oracle inequalities for agghoo (sanity check)
- Binary classification, 0–1 risk (Maillard, A. & Lerasle, 2017): same with an additional factor 2

Performance of agghoo: theory

- Regression with c convex: if $\forall j$, $\mathrm{Card}(I_j^{(t)}) = \tau n$,

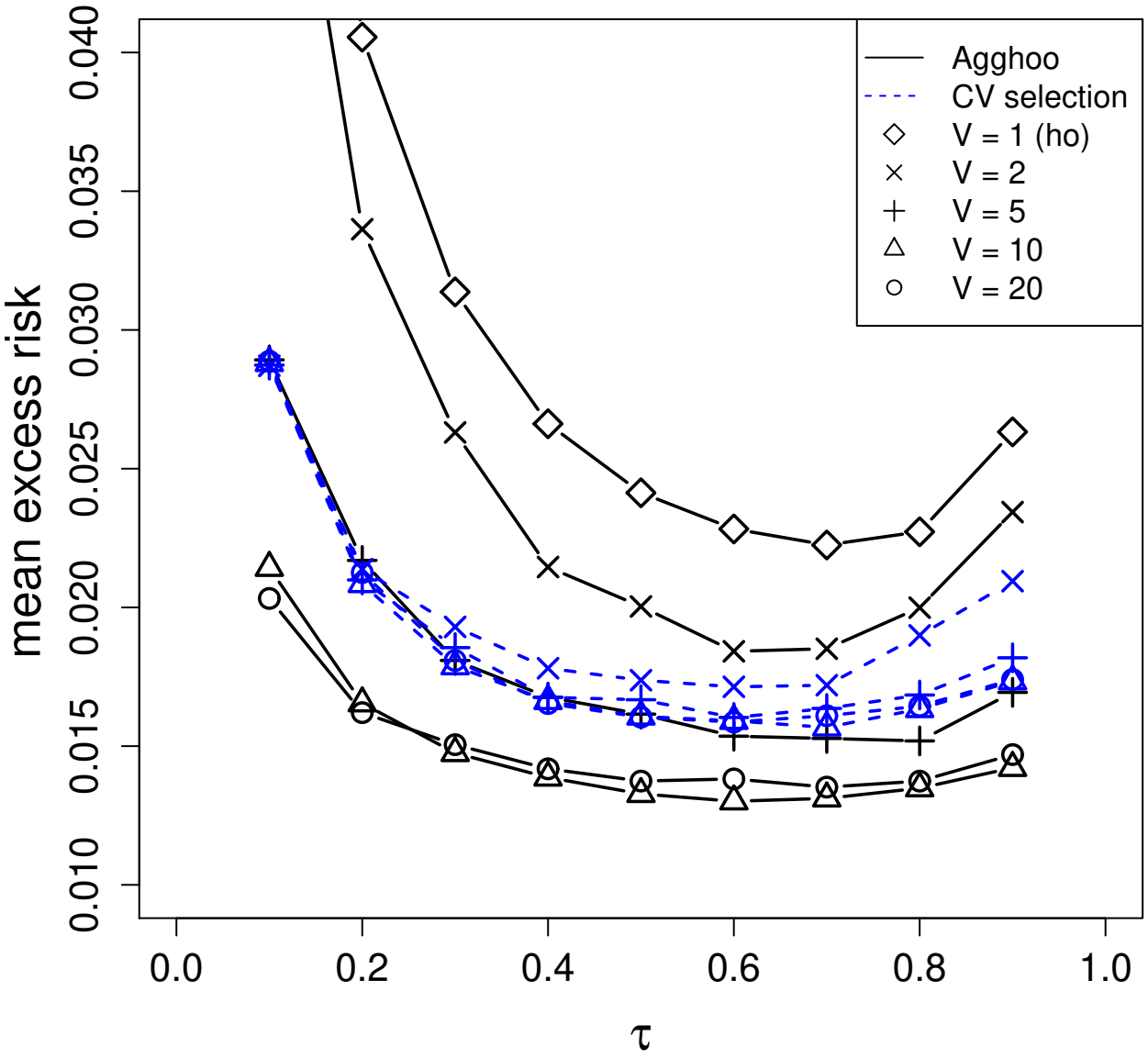
$$\forall V \geqslant 1, \qquad \mathbb{E}\left[\mathcal{R}\left(\widehat{f}^{agghoo}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; (I_j^{(t)})_{1\leqslant j\leqslant V}\right)\right)\right]$$
$$\leqslant \mathbb{E}\left[\mathcal{R}\left(\widehat{f}_{\widehat{m}^{ho}\left((\widehat{f}_m)_{m\in\mathcal{M}}; D_n; I_1^{(t)}\right)}(D_n^{t,1})\right)\right]$$

  Corollary: oracle inequalities for the hold-out $\Rightarrow$ oracle inequalities for agghoo (sanity check)
- Binary classification, 0–1 risk (Maillard, A. & Lerasle, 2017): same with an additional factor 2
- Maillard (2020): in a specific setting (projection estimators on a trigonometric basis, density estimation), proof that

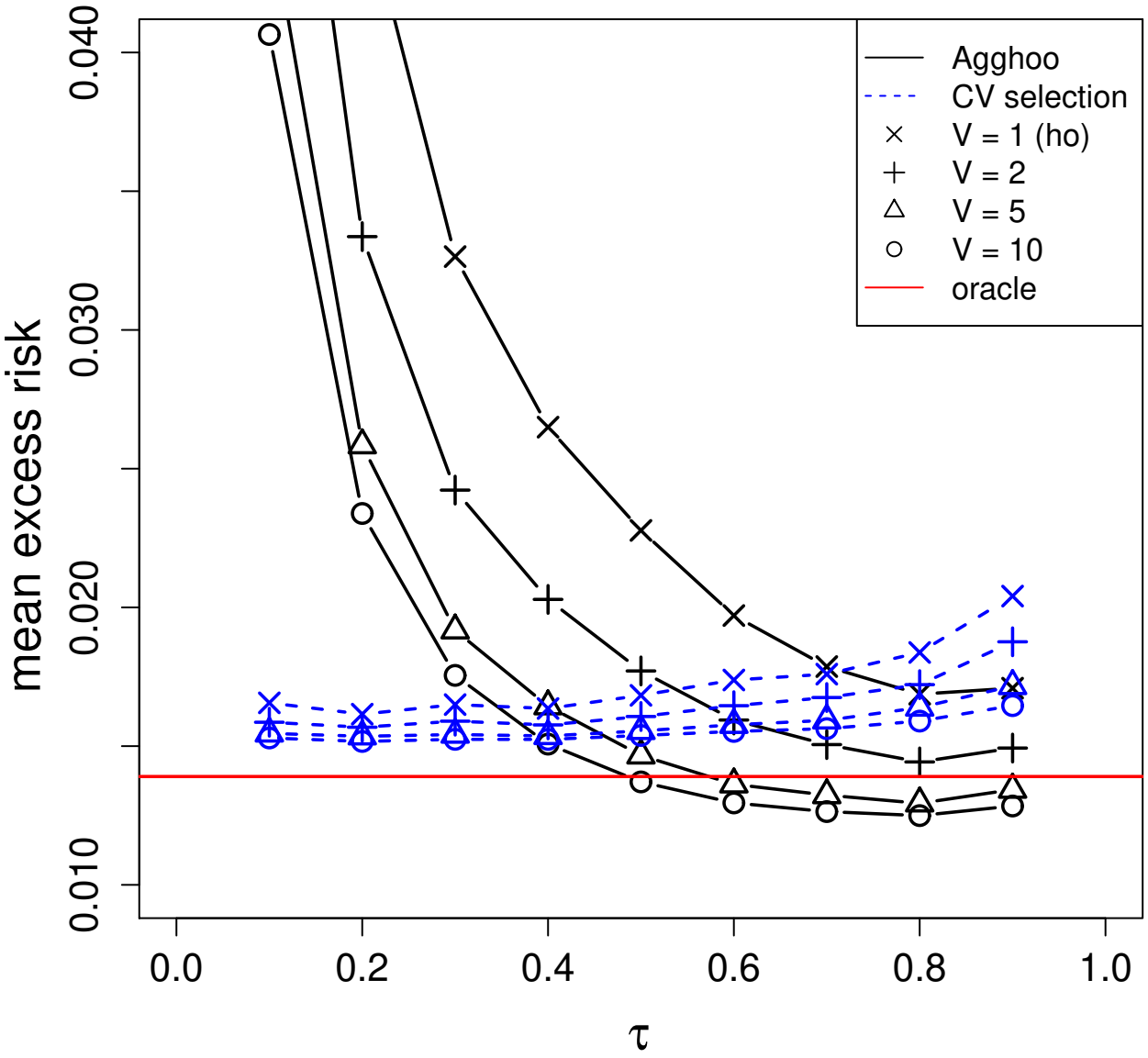
$$\ell\left(\widehat{f}^{agghoo}, f^\star\right) \leqslant C \inf_{m\in\mathcal{M}}\left\{\ell\left(\widehat{f}_m, f^\star\right)\right\} \qquad \text{with} \quad C < 1$$

  $\Rightarrow$ better than any selection procedure, even the oracle!

Numerical experiments: 0–1 binary classification, k-NN



Numerical experiments: regression, L^1 loss, ϵ-SVM



Questions?