Sélection de modèles ou d'estimateurs: une introduction

Sylvain Arlot[1,2,3]

[1]Université Paris-Saclay

[2]Inria Saclay, équipe-projet Celeste

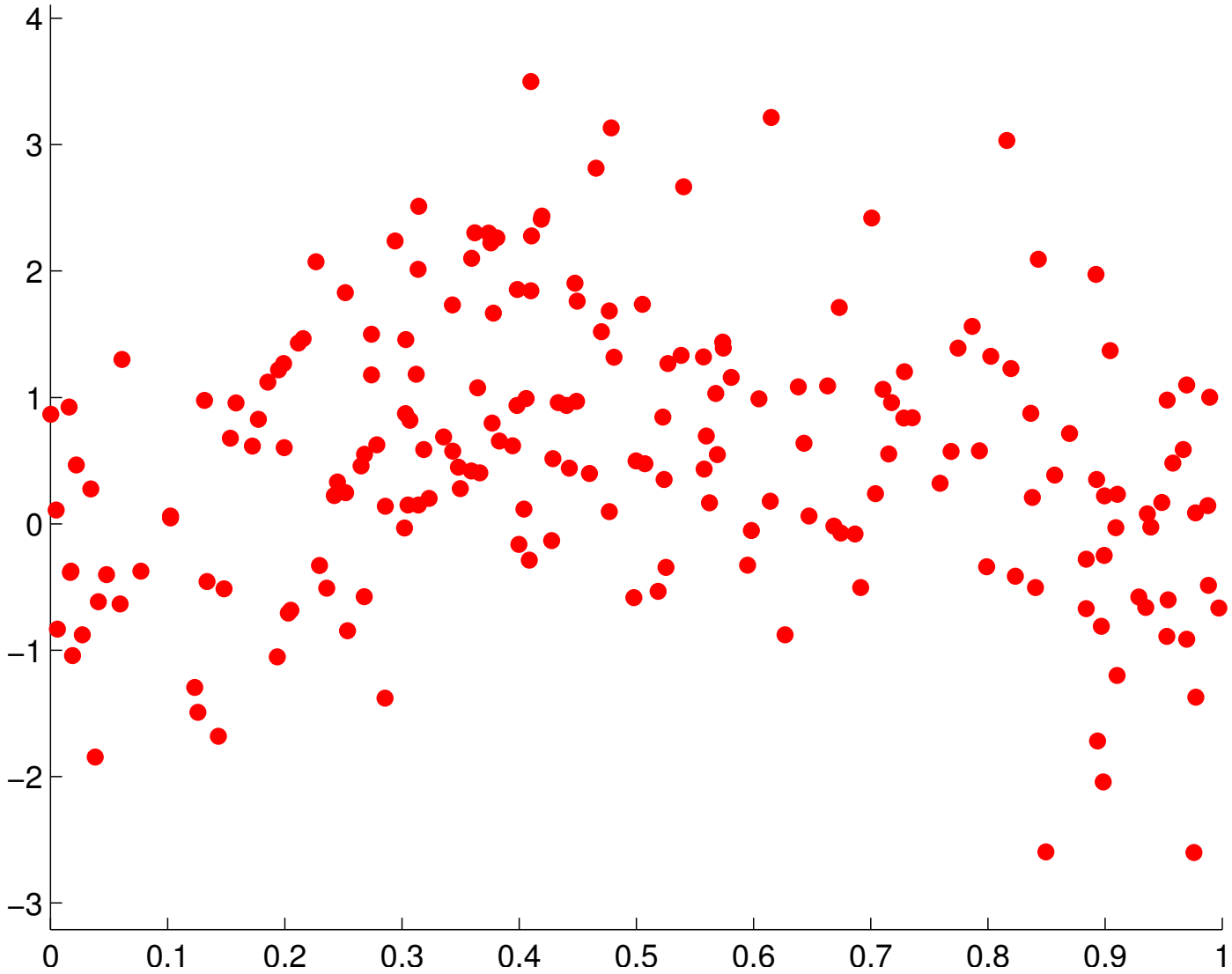
[3]Institut Universitaire de France

Séminaire Modal'X
23 Septembre 2021

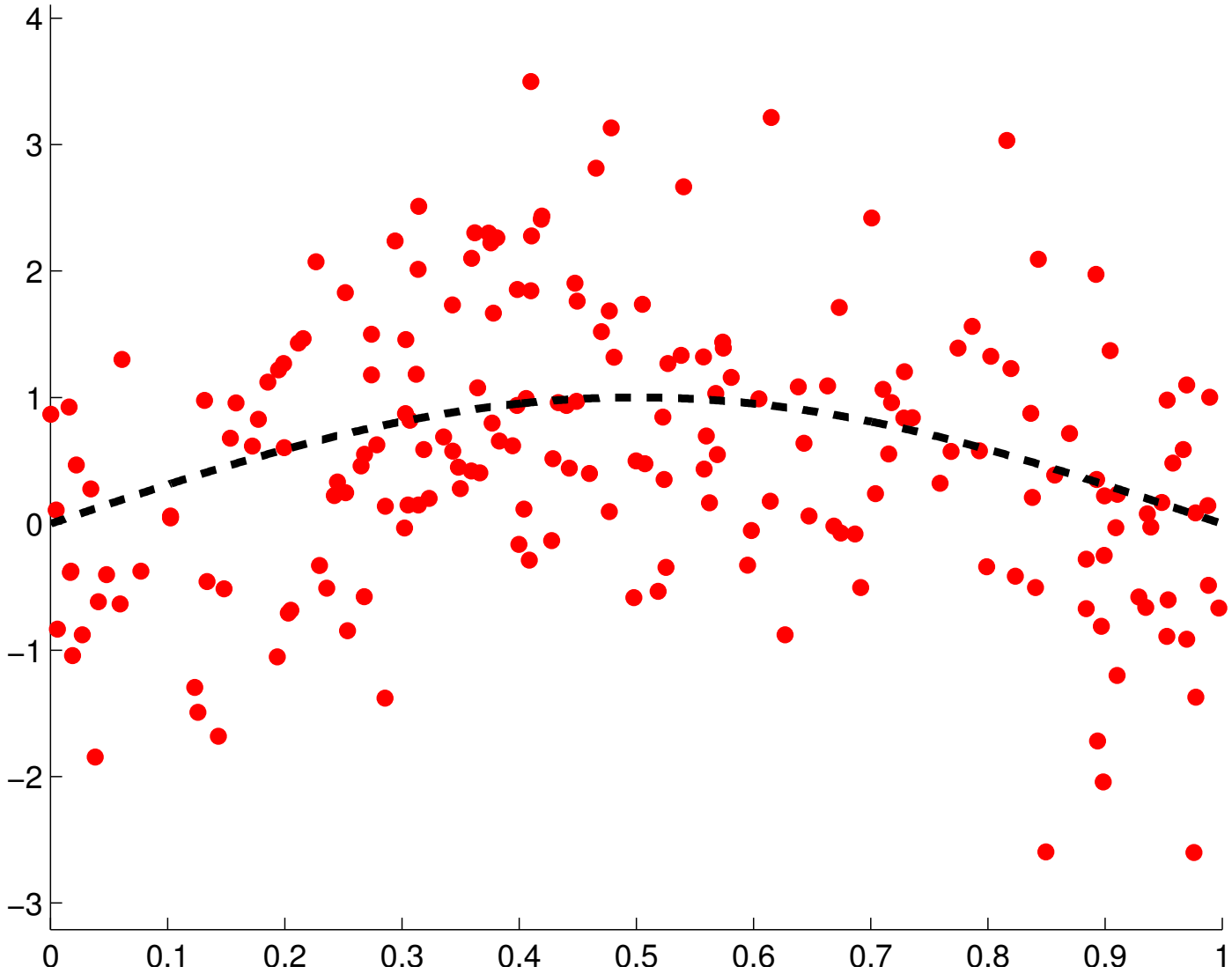
Référence principale: arXiv:1901.07277

Outline

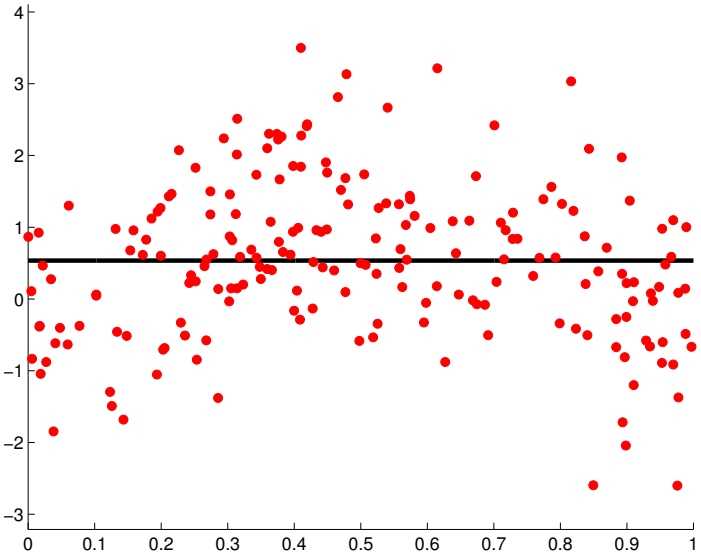
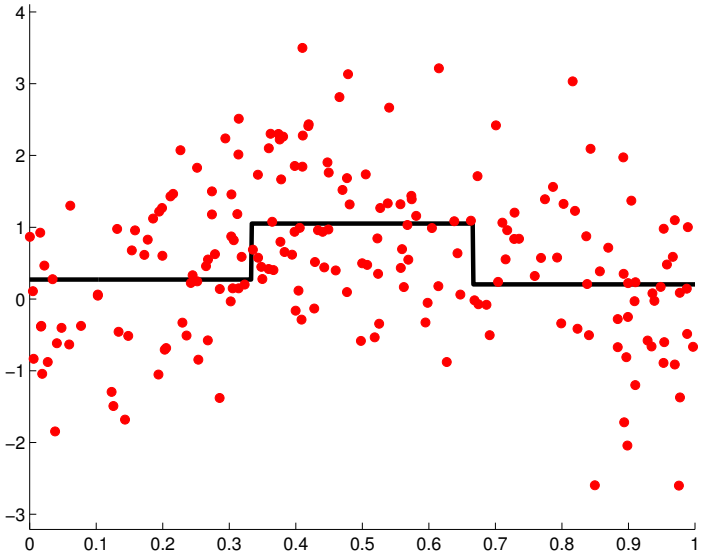
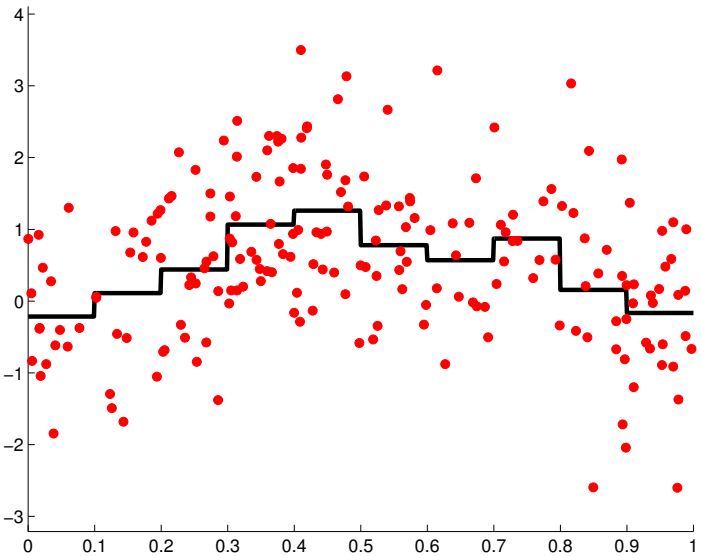
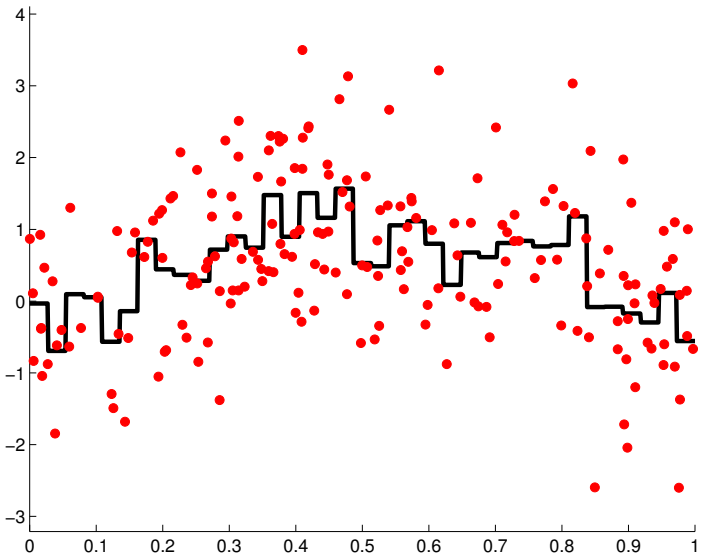
Regression: data $(X_1, Y_1), \ldots, (X_n, Y_n)$



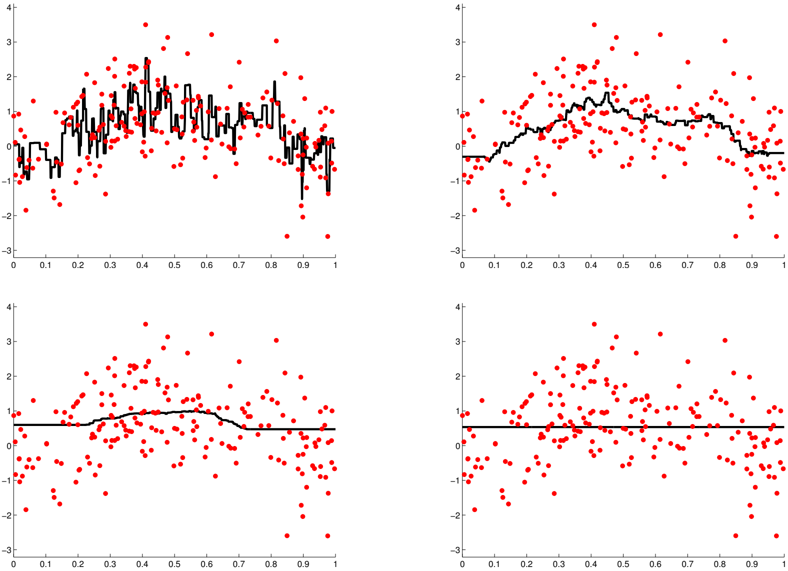
Goal: find the signal (denoising)



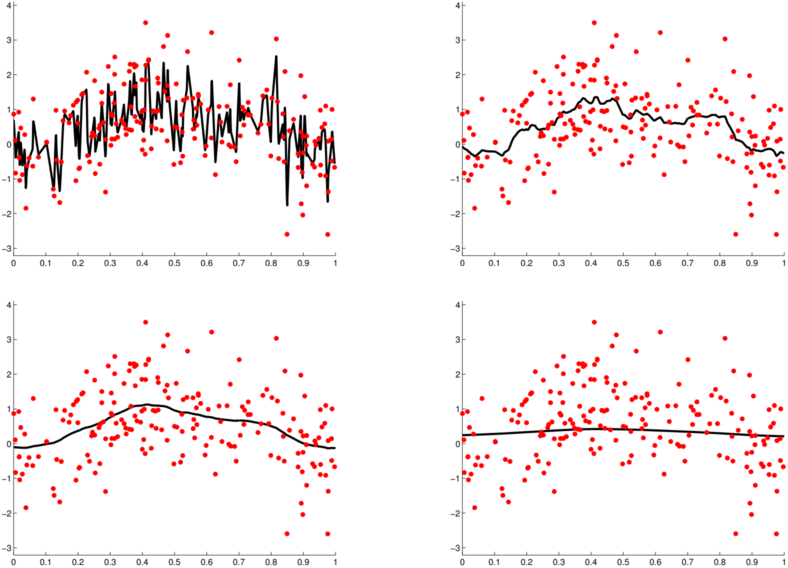
Model selection: regular regressograms, choose D?

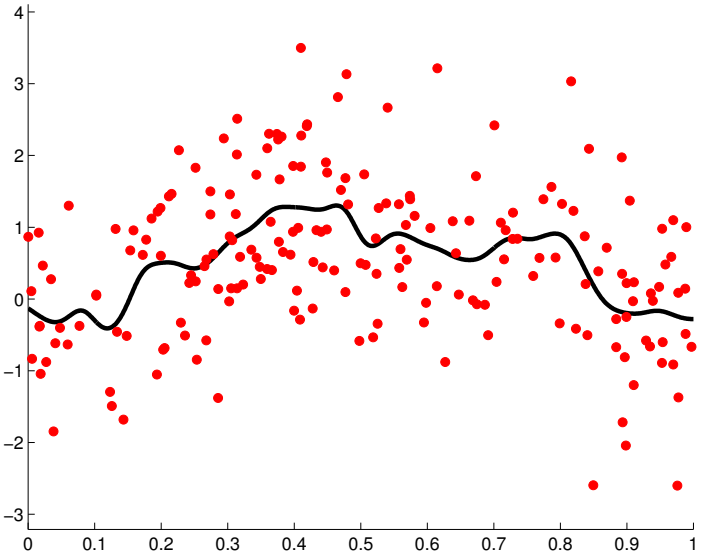
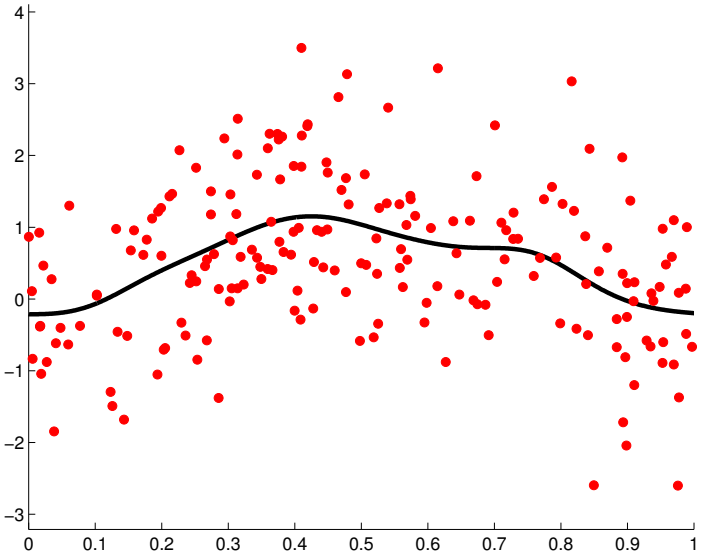
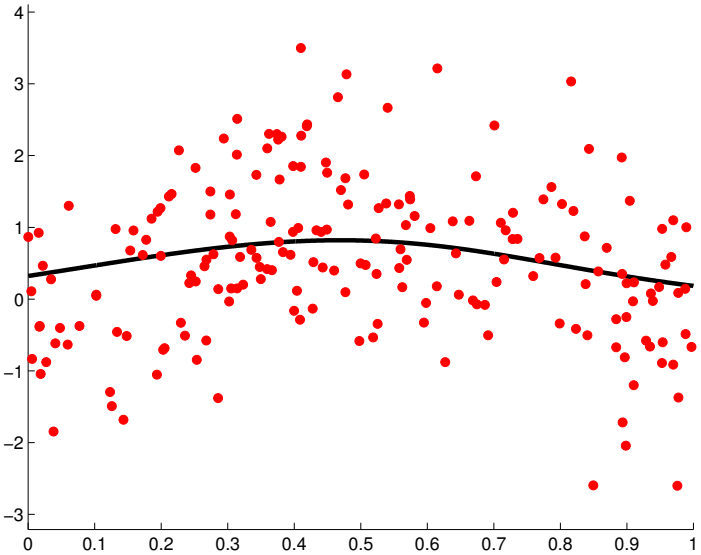
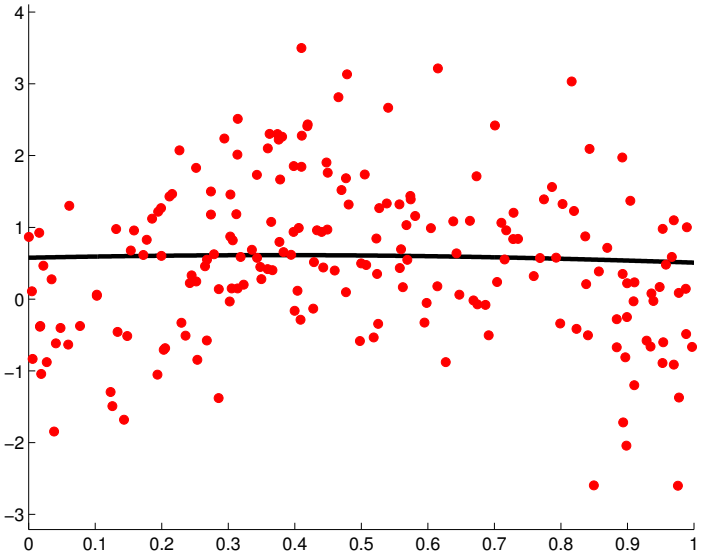
Estimator selection: k nearest neighbours, choose k?



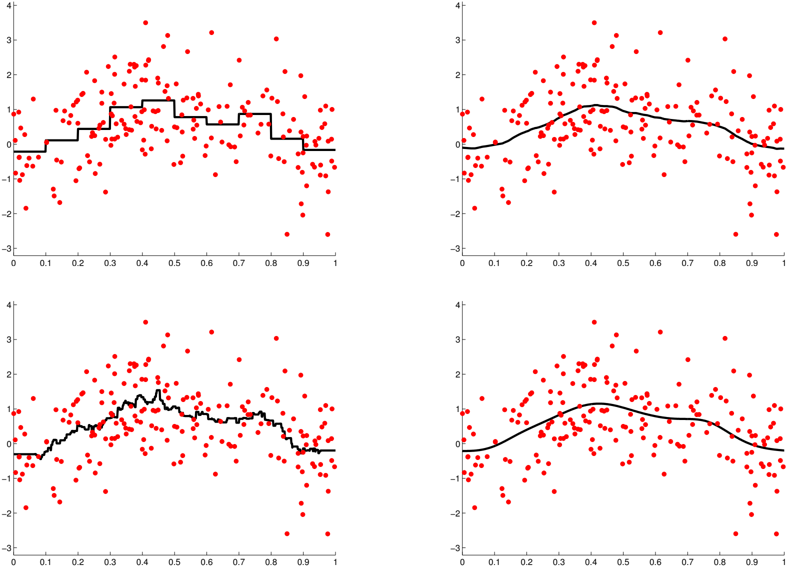
Estimator selection: kernel ridge, choose λ?



Nadaraya-Watson, choose bandwidth h?

Regressogram, ridge, k-NN or Nadaraya-Watson?



Outline

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

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:

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

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

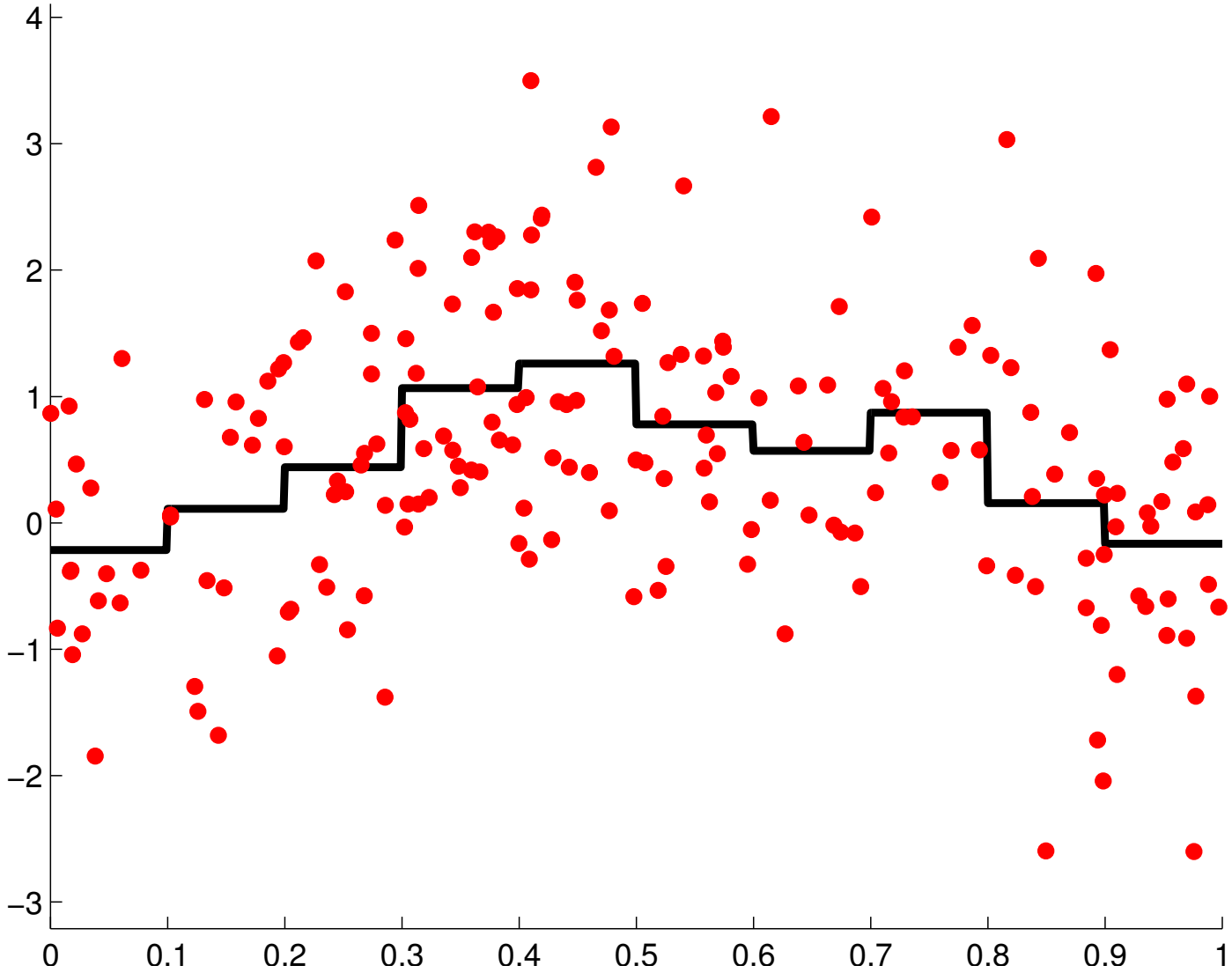
- Model: $S \subset \mathbb{R}^n \Rightarrow$ Least-squares estimator on S:

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

so that

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

Estimators: example: regressogram



Model examples

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

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

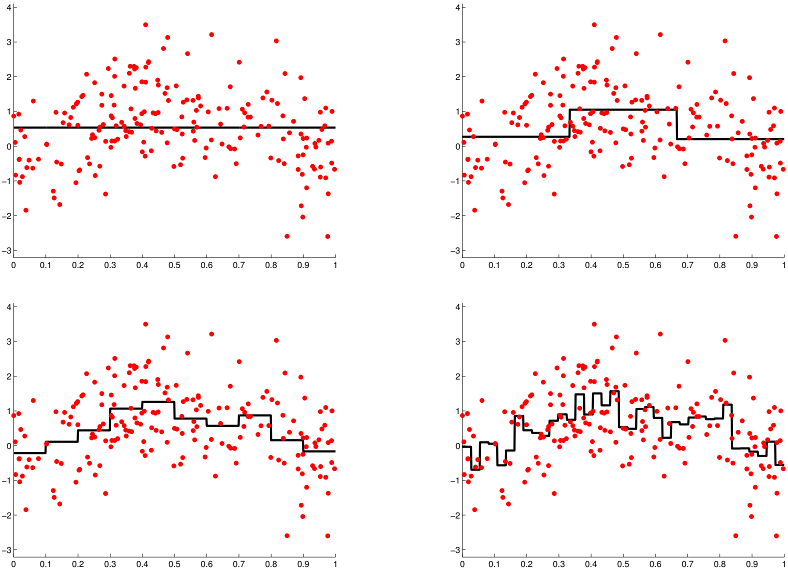
- variable selection: $x_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in \mathbb{R}^p$ gathers p variables that can (linearly) explain Y_i

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

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

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):

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

- Other possible goal: identify the "true" model

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 $\widehat{F}_m$:

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

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

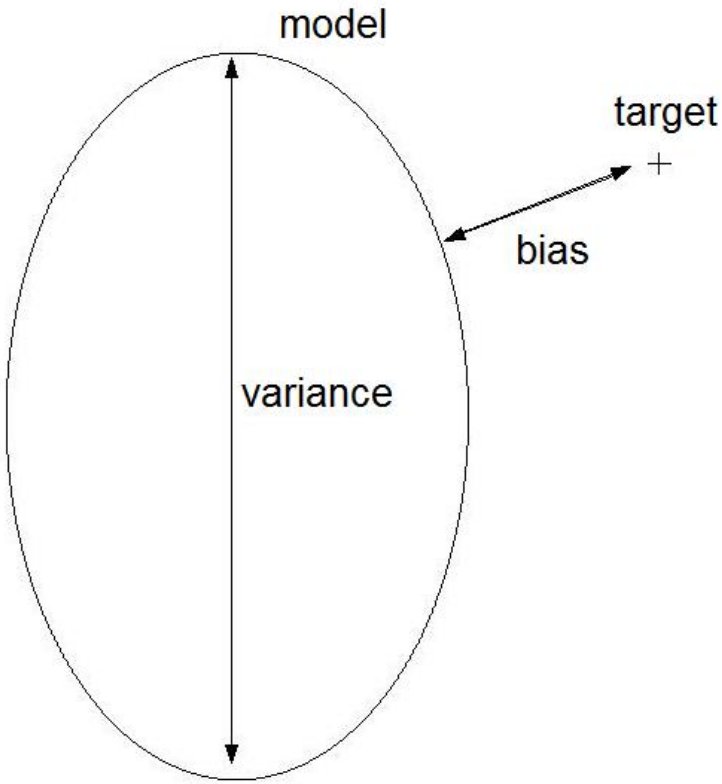
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|F_m - F\right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error / Bias:

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

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

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



Bias-variance trade-off

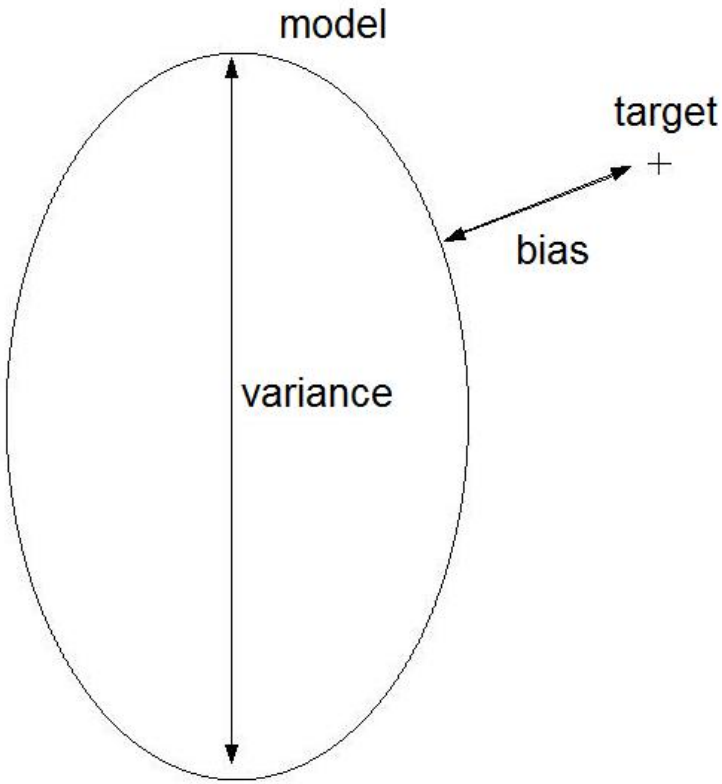
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|F_m - F\right\|^2 + \frac{\sigma^2 D_m}{n}$$

Approximation error / Bias:

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

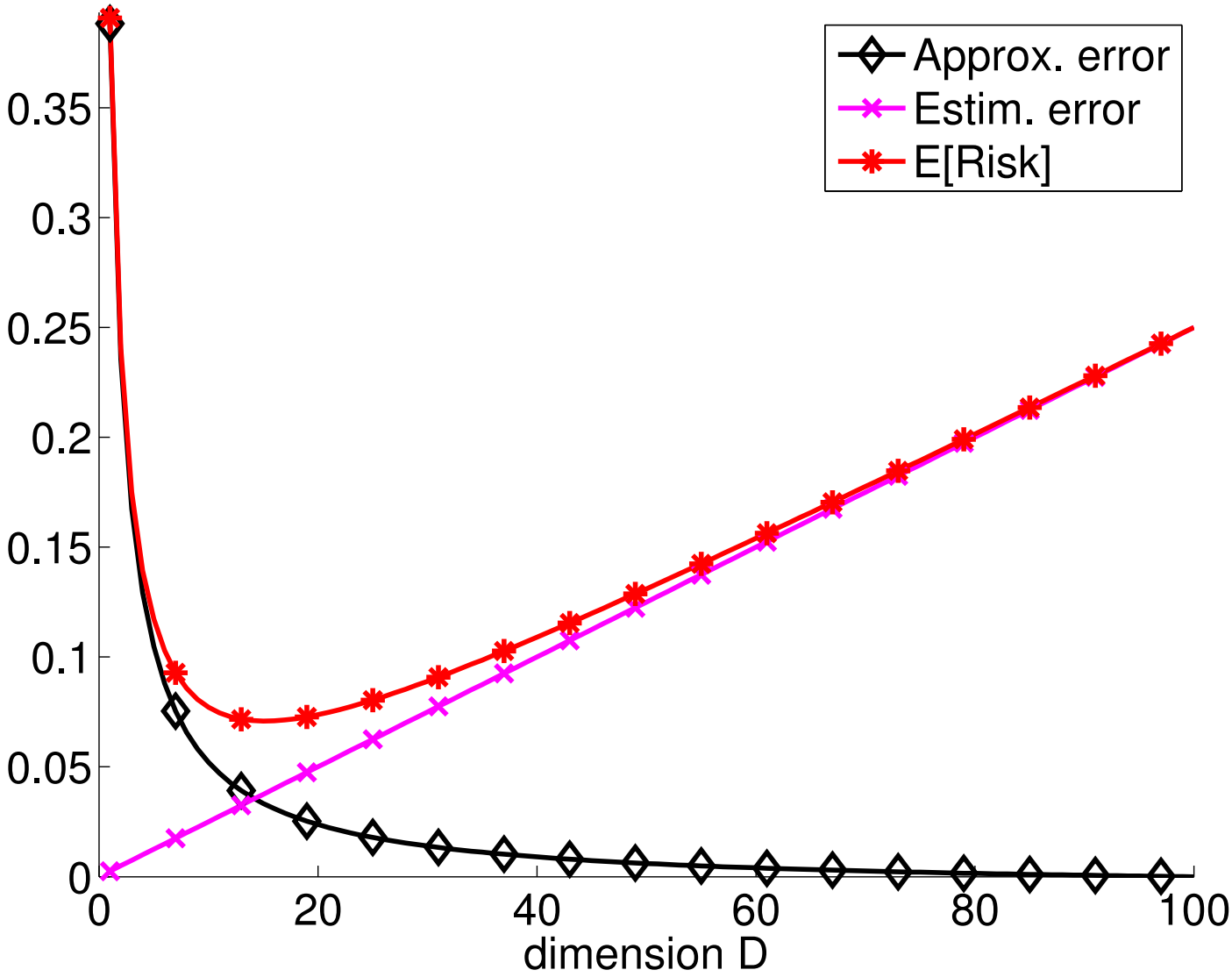
$\mathbb{E}$[Estimation error] / Variance:

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

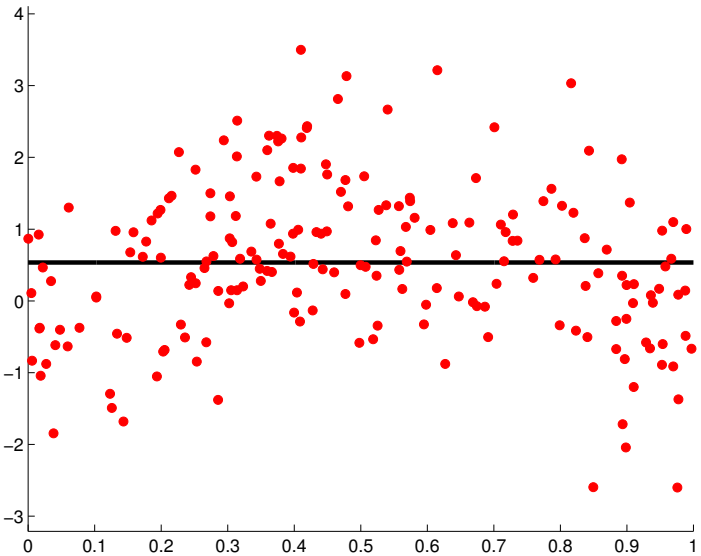


Bias-variance trade-off
$\Leftrightarrow$ avoid overfitting and underfitting

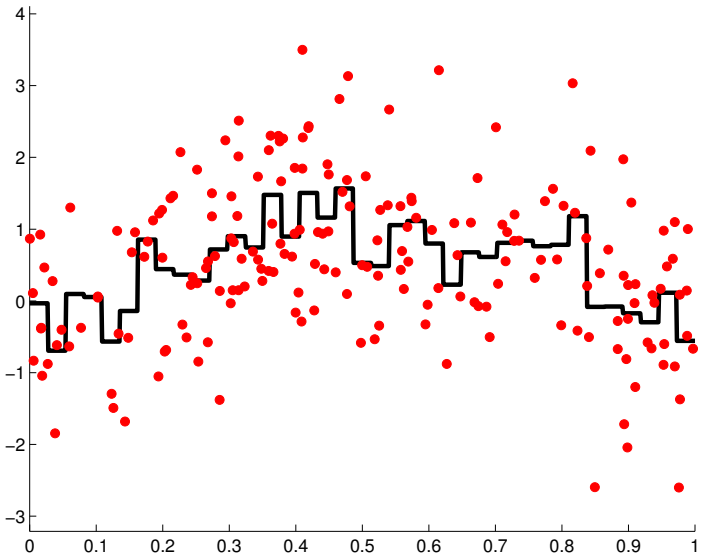
Bias-variance trade-off



Overfitting/underfitting: regressograms

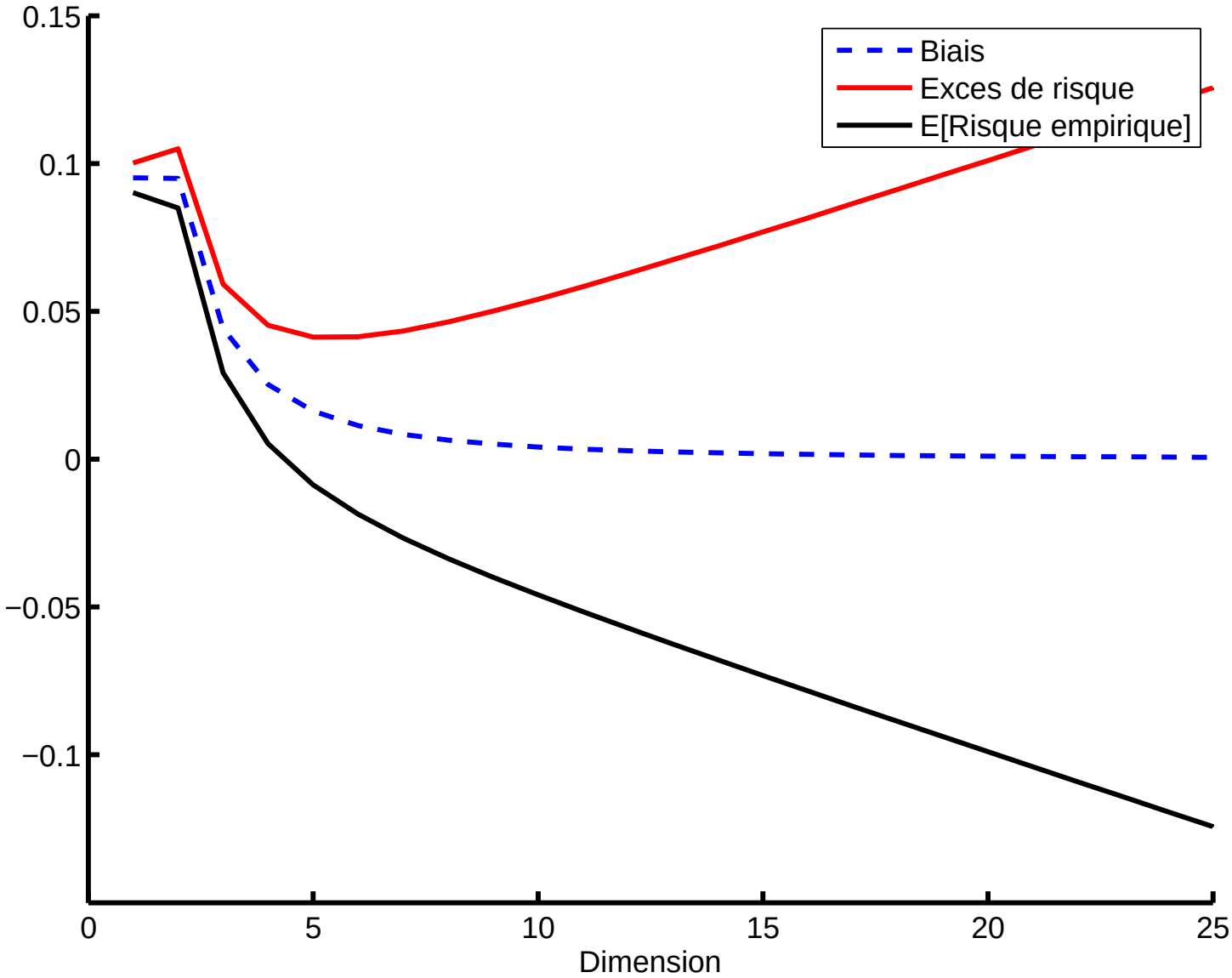


Underfitting
$D = 1$ (too small)



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

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

$\Rightarrow$ Expectation of the ideal penalty

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

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,* $\forall m, m' \in \mathcal{M}$

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

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

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

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

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

Proof of the key lemma

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

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

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,* $\forall m, m' \in \mathcal{M}$

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

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

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

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 $\mathrm{pen}(m) = \mathbb{E}\big[\mathrm{pen}_{\mathrm{id}}(m)\big]$ (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)

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

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* $x \geqslant 0$,

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

$\Rightarrow$ 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 $\theta > 0$,

$$\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2\right] \right|$$

$$\leqslant \frac{\theta}{n}\|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right)\frac{\sigma^2 x}{n}$$

$$= \theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \left(6 + \frac{4}{\theta}\right)\frac{x\sigma^2}{n}$$

Concentration of the ideal penalty: summary

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

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

Concentration of the ideal penalty: summary

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

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

$\Rightarrow$ 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 $\theta \in (0,1)$,

$$\left| \mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2 - \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m) + \frac{1}{n}\|\varepsilon\|^2\right] \right|$$

$$\leqslant \frac{\theta}{n}\|\Pi_m F - F\|^2 + \frac{2x\sigma^2}{\theta n} + \frac{\theta\sigma^2 D_m}{n} + \left(4 + \frac{4}{\theta}\right)\frac{\sigma^2 x}{n}$$

$$= \theta\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] + \left(6 + \frac{4}{\theta}\right)\frac{x\sigma^2}{n}$$

$$\leqslant \frac{\theta}{1-\theta}\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + L(\theta)\frac{x\sigma^2}{n} =: A(m) =: B(m)\,.$$

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

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

Application of the key lemma (step 4)

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

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

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

Application of the key lemma (step 4)

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

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

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

Application of the key lemma (step 4)

- With probability $\geqslant 1 - 4\,\mathrm{Card}(\mathcal{M})e^{-x}$, $\forall \theta \in (0, 1/2)$,

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

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

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* $\theta \in (0, 1/6)$,

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

Conclusion: oracle inequality for C_p

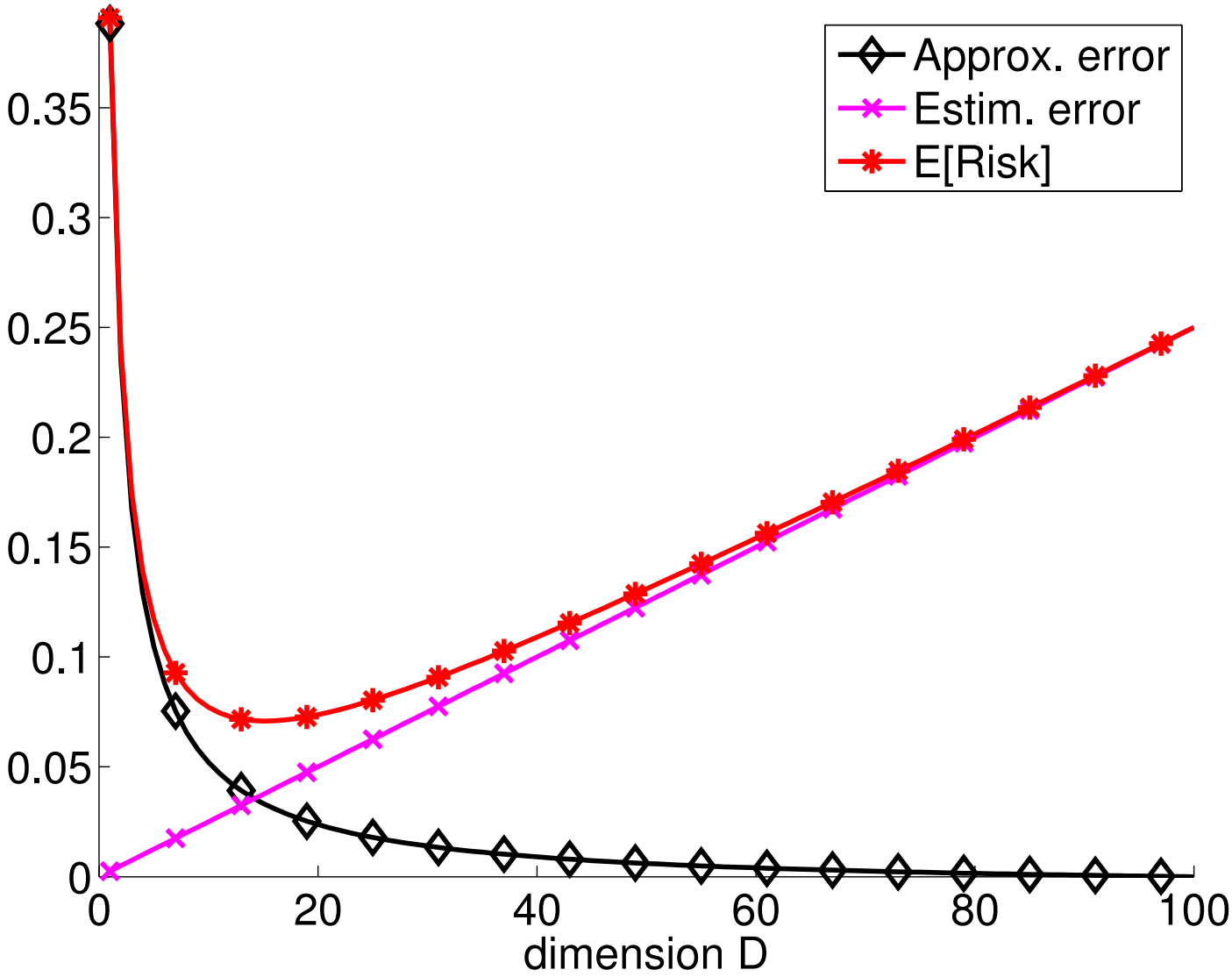
Theorem (Birgé & Massart 2007, reformulated = Theorem 1 in arXiv:1901.07277)

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* $\theta \in (0, 1/6)$,

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

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

Meaning of an oracle inequality



Outline

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

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

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: $c(y, y') = \mathbb{1}_{y \neq y'}$
    $\Rightarrow f^\star(X) = \mathbb{1}_{\mathbb{E}[Y|X] \geqslant 1/2}$
  - convex losses (hinge, logistic, exponential, ...)

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

Link with 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 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 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

- 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 $\forall \widehat{m} \in \arg\min_{m \in \mathcal{M}} \{\mathrm{crit}(m)\}$,

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

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

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