Model selection and estimator selection for statistical learning

Sylvain Arlot

[1]CNRS

[2]École Normale Supérieure (Paris), LIENS, Équipe SIERRA

Scuola Normale Superiore di Pisa, 14–23 February 2011

Outline of the 5 lectures

1. Statistical learning
2. Model selection for least-squares regression
3. Linear estimator selection for least-squares regression
4. Resampling and model selection
5. Cross-validation and model/estimator selection

Part III

Linear estimator selection for least-squares regression

Outline

Outline

Linear estimators for regression

- Definition:

$$\widehat{F} = AY$$

with A deterministic (so, A can depend on X)

Linear estimators for regression

- Definition:

$$\widehat{F} = AY$$

with A deterministic (so, A can depend on X)

- Example: Least-squares estimator: A orthogonal projection matrix
- Example: "Regularized" estimator: A diagonalizable with eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq 0$

Linear estimators for regression

- Definition:

$$\widehat{F} = AY$$

with A deterministic (so, A can depend on X)

- Example: Least-squares estimator: A orthogonal projection matrix
- Example: "Regularized" estimator: A diagonalizable with eigenvalues $1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq 0$
- Assumptions:

$$\|A\| \leq 1 \qquad \text{or} \quad \|A\| \leq B$$

$$\mathrm{tr}(A^\top A) \leq \mathrm{tr}(A)$$

Kernel ridge regression (1/2)

- Idea: look for an estimator having a small empirical risk and a small norm in some functional space $\mathcal{F}$

- $\mathcal{F} \subset \mathbb{S}$ the Reproducing Kernel Hilbert Space (RKHS) associated with a positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

- Representer theorem $\Rightarrow \exists \widehat{\alpha} \in \mathbb{R}^n$ such that $\widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot)$

Kernel ridge regression (2/2)

$$\widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot) \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i) \right)^2 + \lambda \left\| f \right\|_{\mathcal{F}}^2 \right\}$$

- Fixed design: $\widehat{F} = (\widehat{f}(x_i))_{1 \leq i \leq n} = K\widehat{\alpha}$ with $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$ and

$$\widehat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \left\| Y - K\alpha \right\|^2 + \lambda \alpha^\top K \alpha \right\}$$

$$\Rightarrow \quad \widehat{F} = K(K + n\lambda I_n)^{-1} Y$$

Kernel ridge regression (2/2)

- Fixed design: $\widehat{F} = (\widehat{f}(x_i))_{1 \leq i \leq n} = K\widehat{\alpha}$ with $K = (k(x_i, x_j))_{1 \leq i,j \leq n}$ and

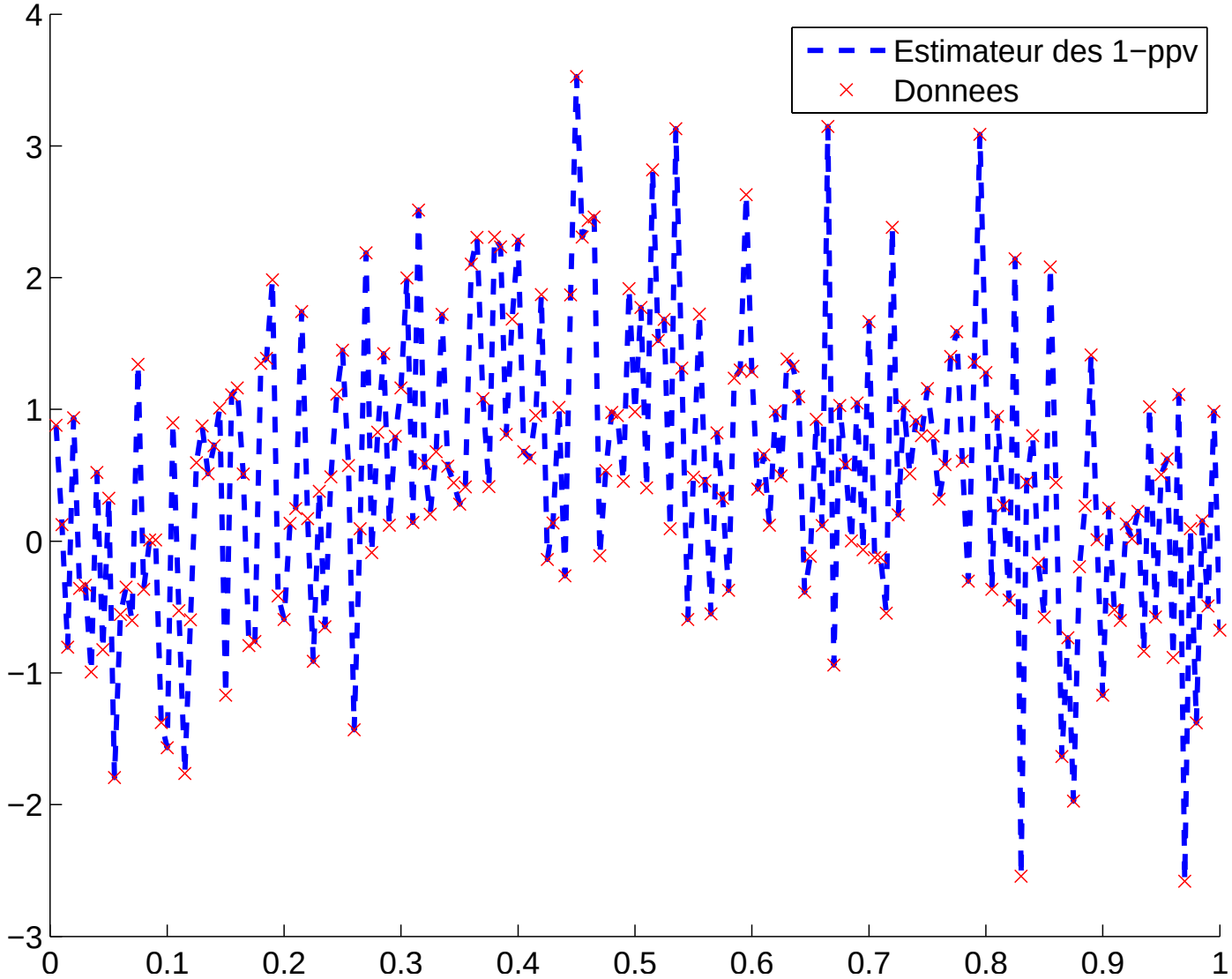
$$\widehat{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \|Y - K\alpha\|^2 + \lambda \alpha^\top K \alpha \right\}$$

$$\Rightarrow \quad \widehat{F} = K(K + n\lambda I_n)^{-1} Y$$

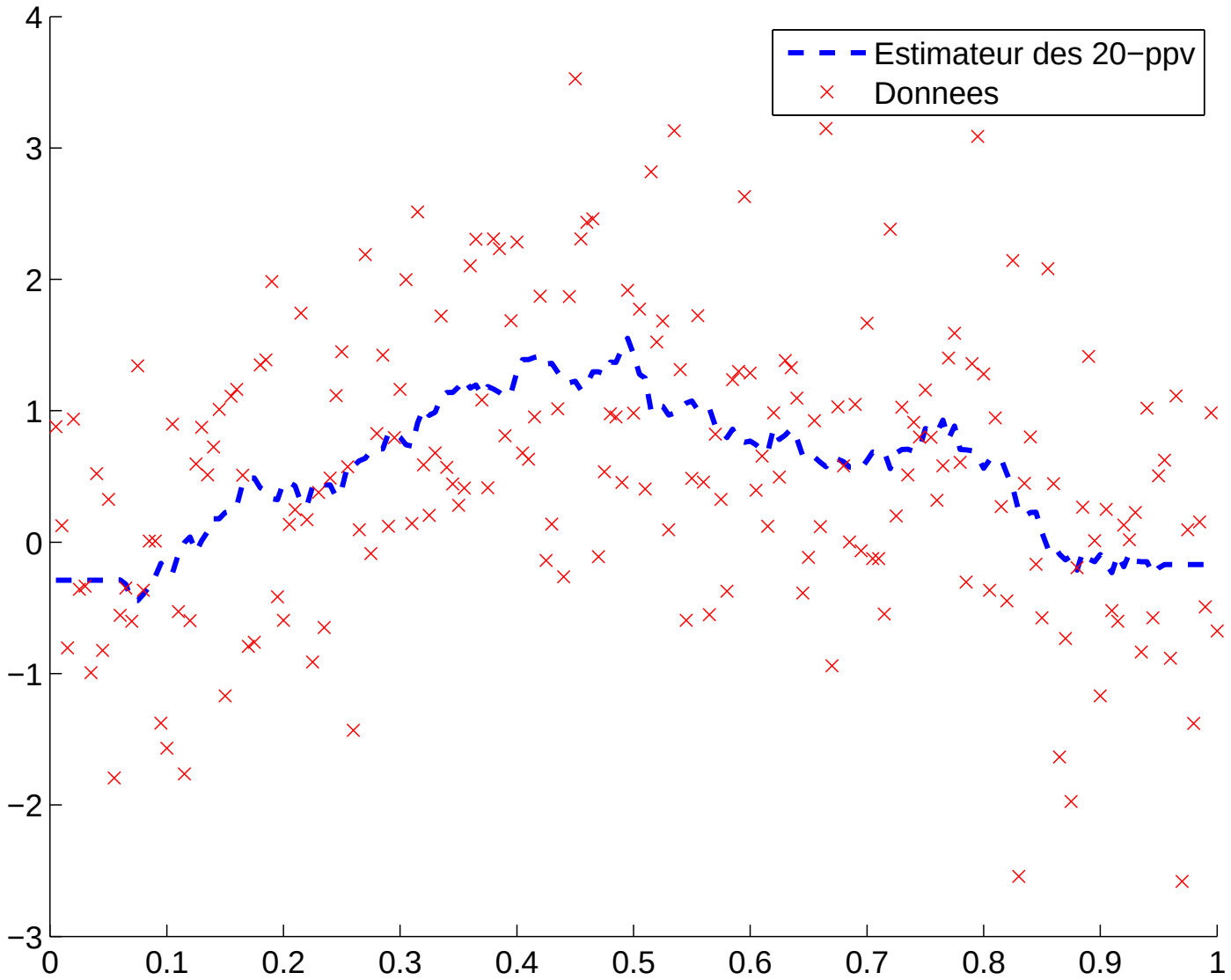
- If $(e_j)_{1 \leq j \leq n}$ orthonormal basis of eigenvectors of K (with eigenvalues $(\kappa_j)_{1 \leq j \leq n}$):

$$\widehat{F} e_j = \frac{\kappa_j}{\kappa_j + n\lambda} e_j$$

Nearest-neighbour estimator: $\widehat{F} = Y$



k-nearest-neighbours estimator ($k = 20$)



k-nearest-neighbours estimator

For all $i \in \{1, \ldots, n\}$,

$$\widehat{F}_i = \frac{1}{k} \sum_{x_j \in \{ k-\mathrm{NN} \text{ of } x_i \}} Y_j$$

$$= \sum_{1 \leq j \leq n} \left(\frac{1}{k} \mathbb{1}_{x_j \in \{ k-\mathrm{NN} \text{ of } x_i \}} Y_j \right)$$

k-nearest-neighbours estimator

For all $i \in \{1, \dots, n\}$,

$$\widehat{F}_i = \frac{1}{k} \sum_{x_j \in \{ k-\mathrm{NN} \text{ of } x_i \}} Y_j$$

$$= \sum_{1 \leq j \leq n} \left(\frac{1}{k} \mathbb{1}_{x_j \in \{ k-\mathrm{NN} \text{ of } x_i \}} Y_j \right)$$

$$\Rightarrow \qquad \widehat{F} = A(k) Y$$

$$\text{with} \quad A(k)_{i,j} = \frac{1}{k} \mathbb{1}_{x_j \in \{ k-\mathrm{NN} \text{ of } x_i \}}$$

Nadaraya-Watson estimator

Let $K : \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty)$ be some "kernel".

$$\forall i, j \in \{1, \ldots, n\}\ , \quad A(K)_{i,j} = \frac{K(x_i, x_j)}{\sum_{1 \leq \ell \leq n} K(x_i, x_\ell)}$$

$$\widehat{F} = A(K)Y$$

Nadaraya-Watson estimator

Let $K : \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty)$ be some "kernel".

$$\forall i, j \in \{1, \ldots, n\}\ , \quad A(K)_{i,j} = \frac{K(x_i, x_j)}{\sum_{1 \leq \ell \leq n} K(x_i, x_\ell)}$$

$$\widehat{F} = A(K)Y$$

- Typically, $K(x, y) = g(d(x, y)/h)$ for some distance d over $\mathcal{X}$, a bandwidth $h > 0$ and a non-increasing function $g : [0, +\infty) \mapsto [0, +\infty)$.

Nadaraya-Watson estimator

Let $K : \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty)$ be some "kernel".

$$\forall i, j \in \{1, \ldots, n\}\ , \quad A(K)_{i,j} = \frac{K(x_i, x_j)}{\sum_{1 \leq \ell \leq n} K(x_i, x_\ell)}$$

$$\widehat{F} = A(K)Y$$

- Typically, $K(x, y) = g(d(x, y)/h)$ for some distance d over $\mathcal{X}$, a bandwidth $h > 0$ and a non-increasing function $g : [0, +\infty) \mapsto [0, +\infty)$.
- Gaussian kernel $g(t) = \exp(-t^2)$.
- Window kernel $g(t) = \mathbb{1}_{t \in [0,1]}$

Outline

Risk, empirical risk, ideal penalty

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y$$

Risk, empirical risk, ideal penalty

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2 \operatorname{tr}(A_m^\top A_m)}{n}$$

Risk, empirical risk, ideal penalty

$$Y = F + \varepsilon \quad \text{with} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y$$

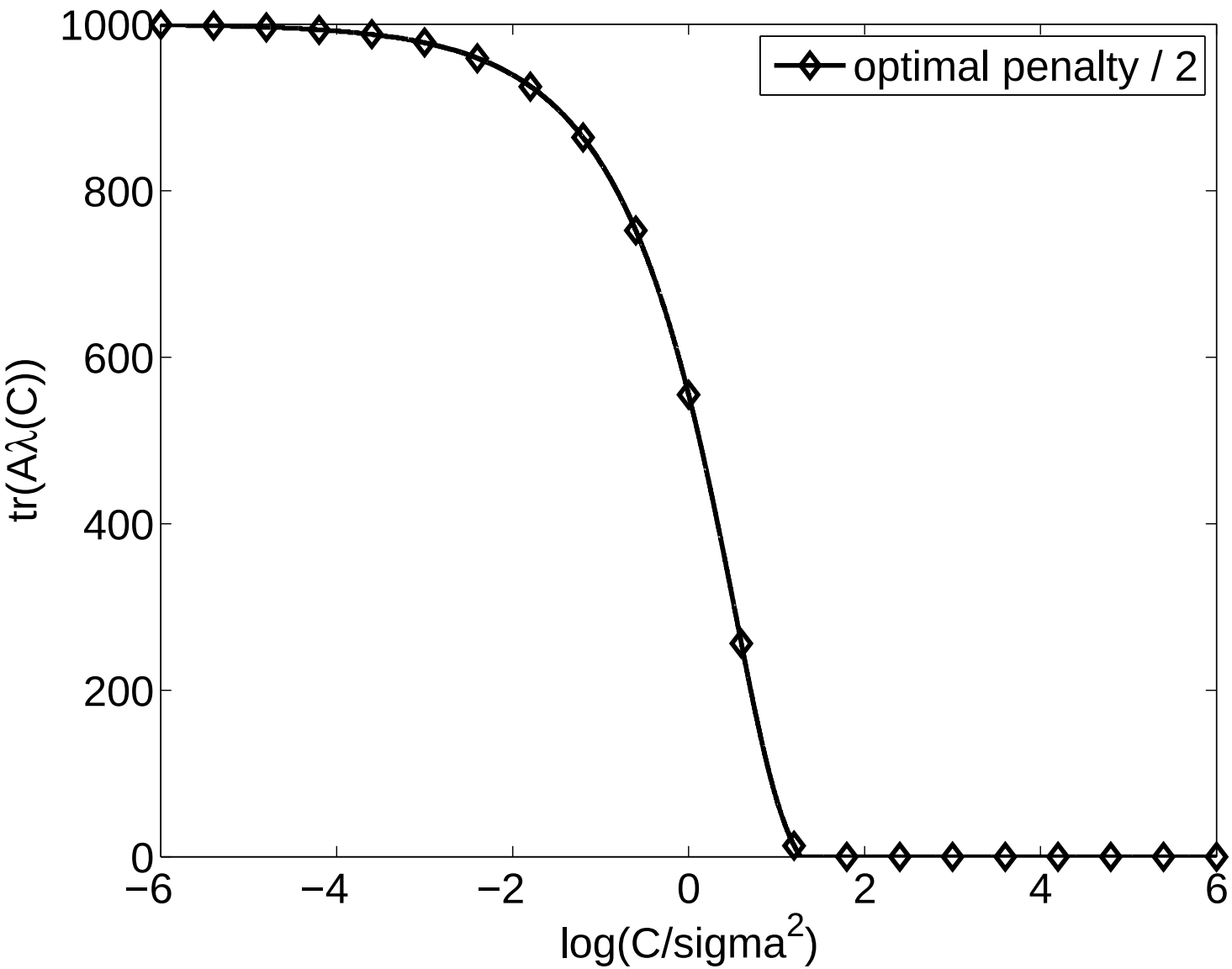
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2 \operatorname{tr}(A_m^\top A_m)}{n}$$

$$\operatorname{pen}_{\mathrm{id}}(m) = \frac{2}{n}\left\langle A_m\varepsilon,\, \varepsilon\right\rangle + \frac{2}{n}\left\langle (A_m - I_n)F,\, \varepsilon\right\rangle$$

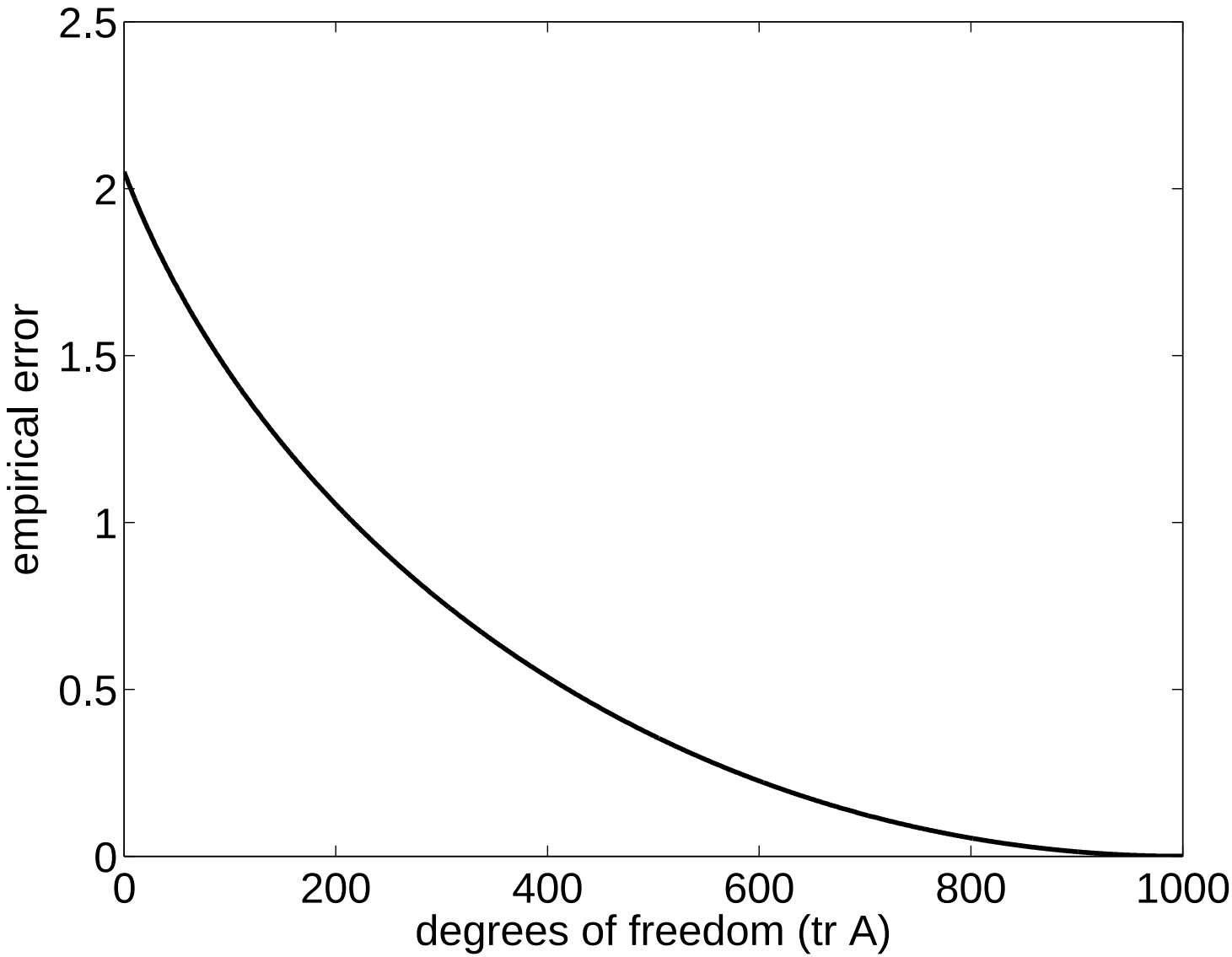
$$\mathbb{E}\left[\operatorname{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n} \quad \Rightarrow \quad C_L \text{ (Mallows, 1973)}$$

generalized degrees of freedom: $\operatorname{tr}(A_m)$

No dimension jump with a penalty $\propto \mathrm{tr}(A_m)$



The minimal penalty is not proportional to $\mathrm{tr}(A_m)$



What is the minimal penalty?

$$p_2(m) = \frac{1}{n}\left\|Y - F_m\right\|^2 - \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2$$
$$= \frac{2}{n}\left\langle \varepsilon,\, A_m\varepsilon \right\rangle - \frac{1}{n}\left\|A_m\varepsilon\right\|^2 - \frac{2}{n}\left\langle \varepsilon,\, A_m^\top (I_n - A_m) F \right\rangle$$

What is the minimal penalty?

$$p_2(m) = \frac{1}{n}\|Y - F_m\|^2 - \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2$$
$$= \frac{2}{n}\langle \varepsilon,\, A_m\varepsilon\rangle - \frac{1}{n}\|A_m\varepsilon\|^2 - \frac{2}{n}\left\langle \varepsilon,\, A_m^\top(I_n - A_m)F\right\rangle$$

so that

$$\mathrm{pen}_{\mathrm{min}}(m) = \mathbb{E}\left[p_2(m)\right] = \frac{\sigma^2\left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right)}{n}$$

What is the minimal penalty?

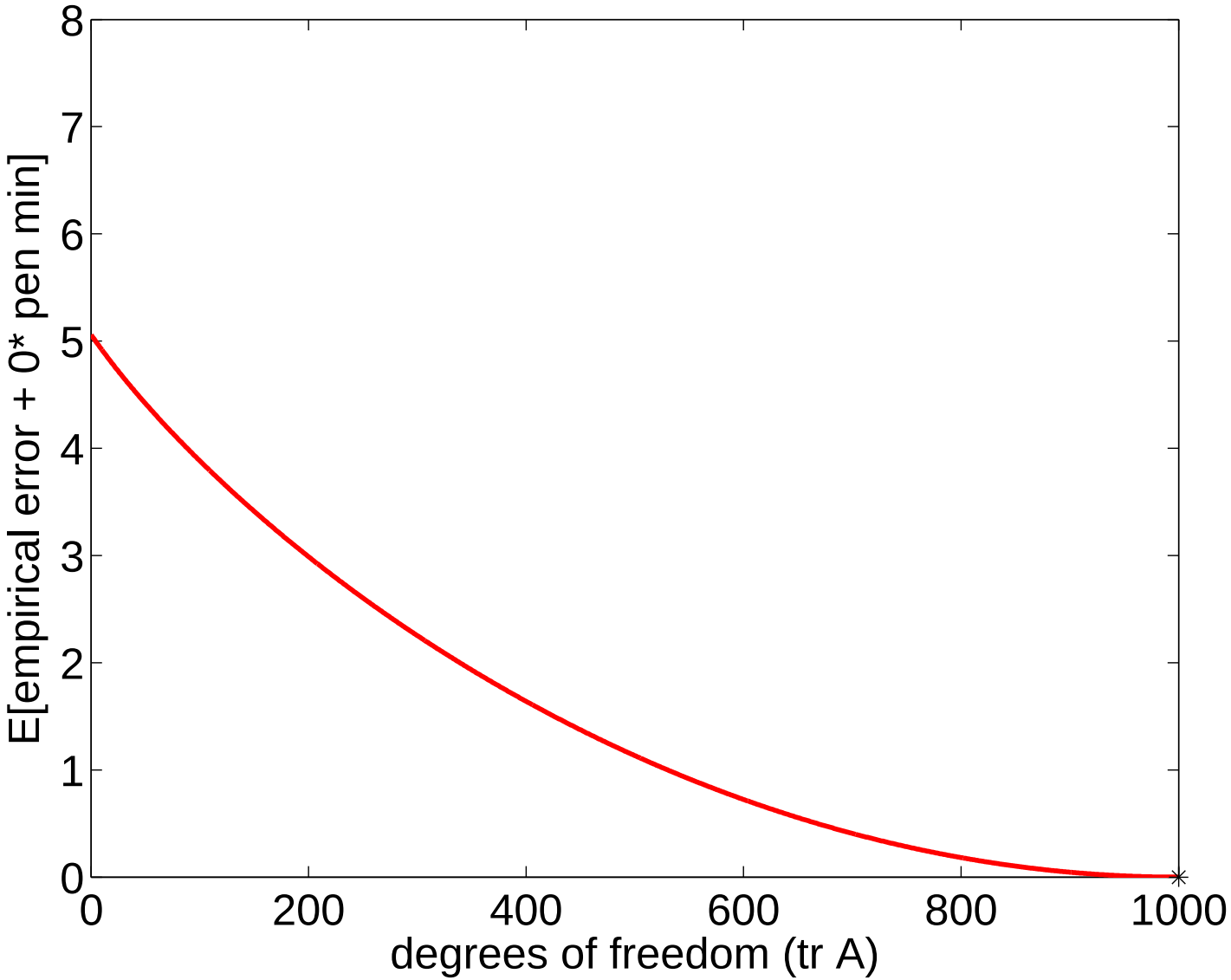
$$p_2(m) = \frac{1}{n}\|Y - F_m\|^2 - \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2$$
$$= \frac{2}{n}\langle \varepsilon,\, A_m\varepsilon\rangle - \frac{1}{n}\|A_m\varepsilon\|^2 - \frac{2}{n}\left\langle \varepsilon,\, A_m^\top(I_n - A_m)F\right\rangle$$

so that

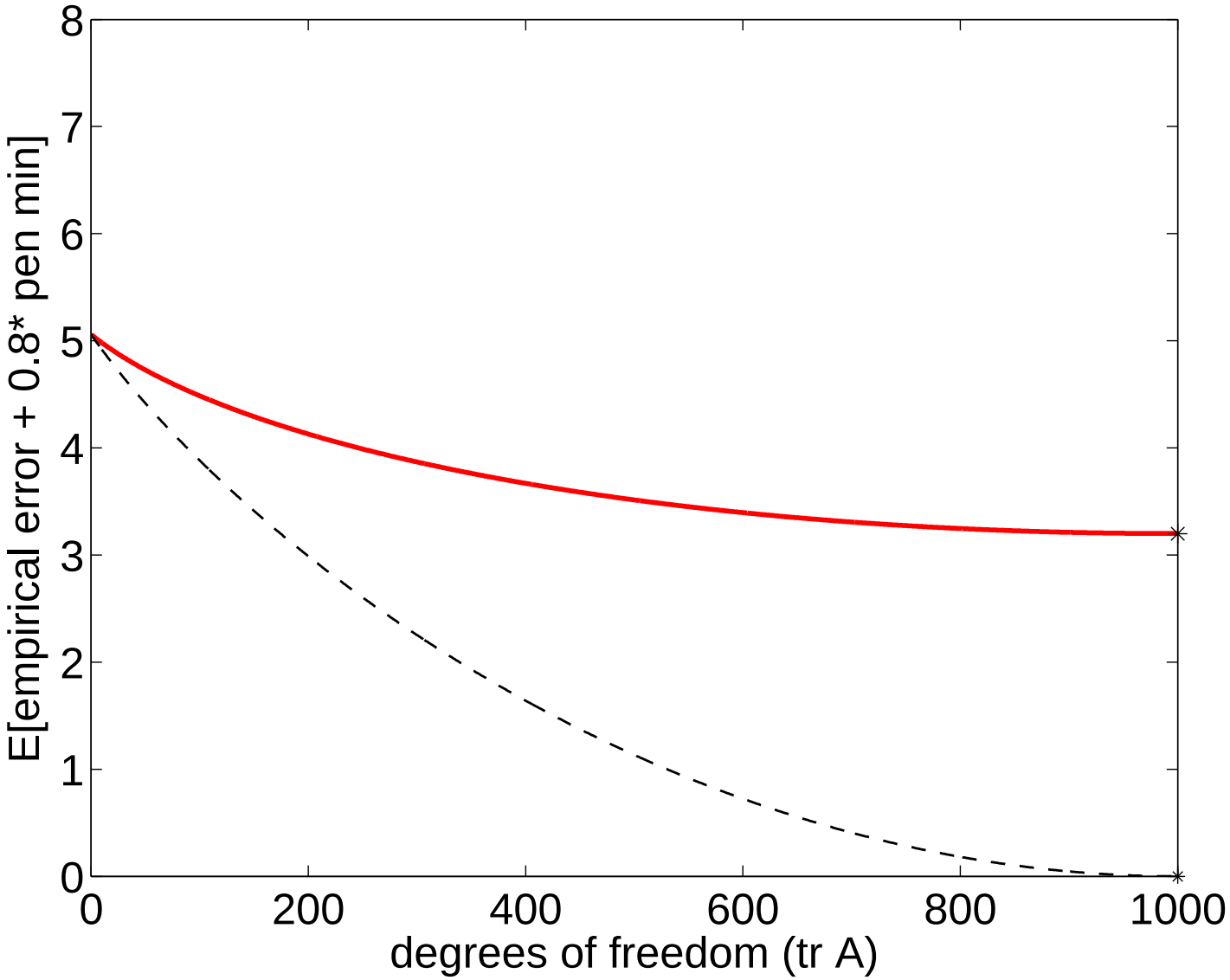
$$\mathrm{pen}_{\mathrm{min}}(m) = \mathbb{E}\left[p_2(m)\right] = \frac{\sigma^2\left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right)}{n}$$

$$\widehat{m}_{\mathrm{min}}(C) \in \arg\min_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + \frac{C\left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right)}{n}\right\}$$

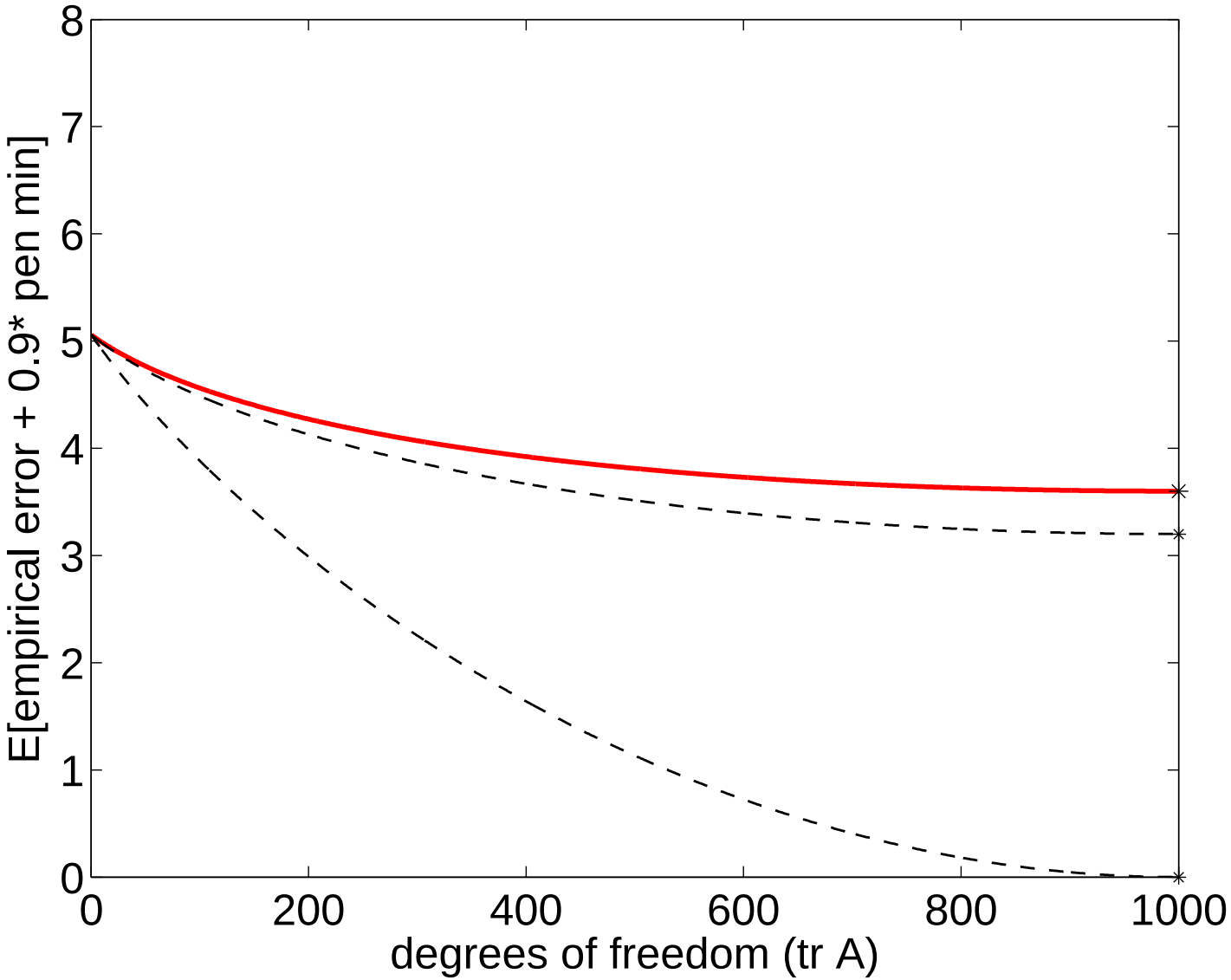
$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



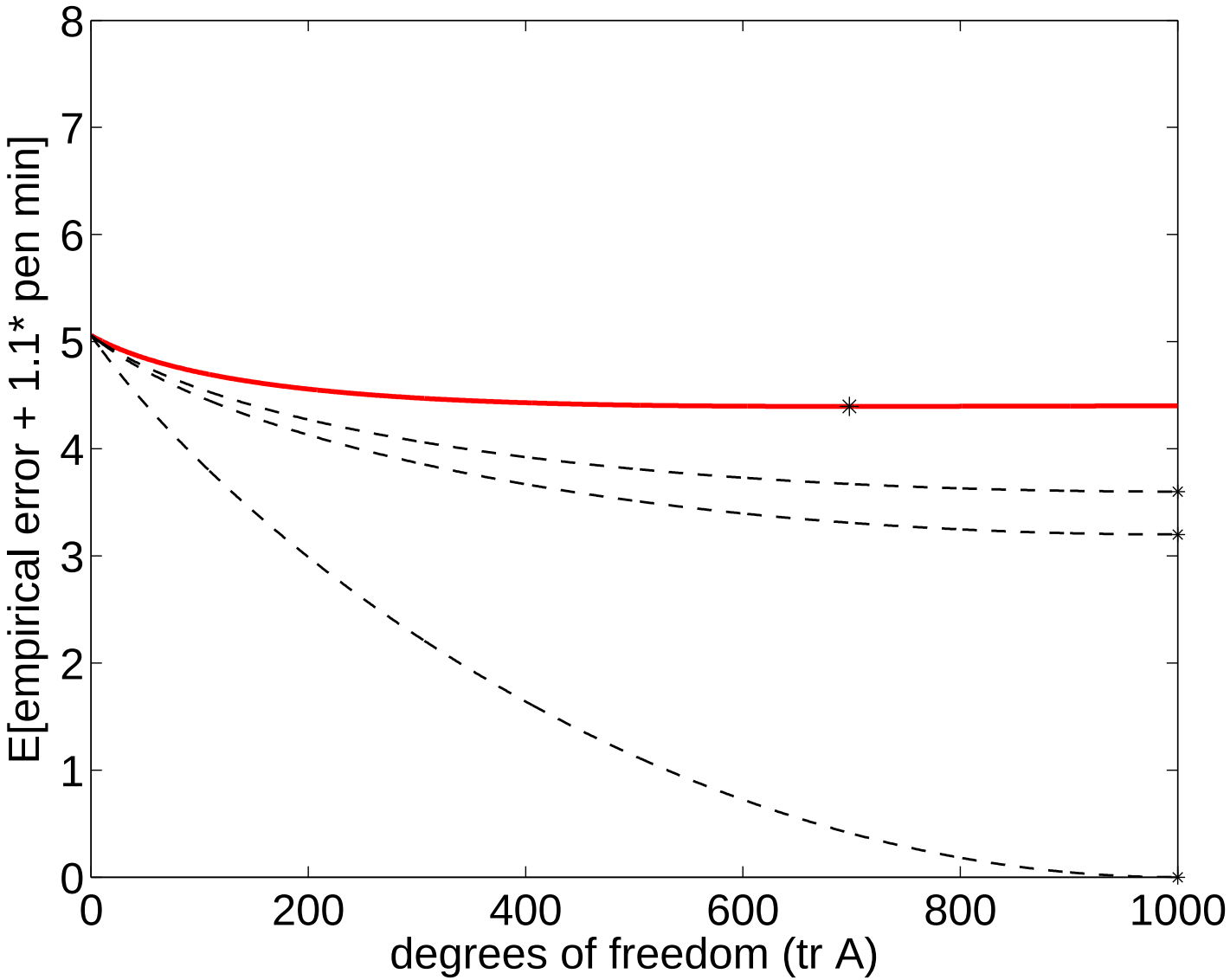
$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0.8 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



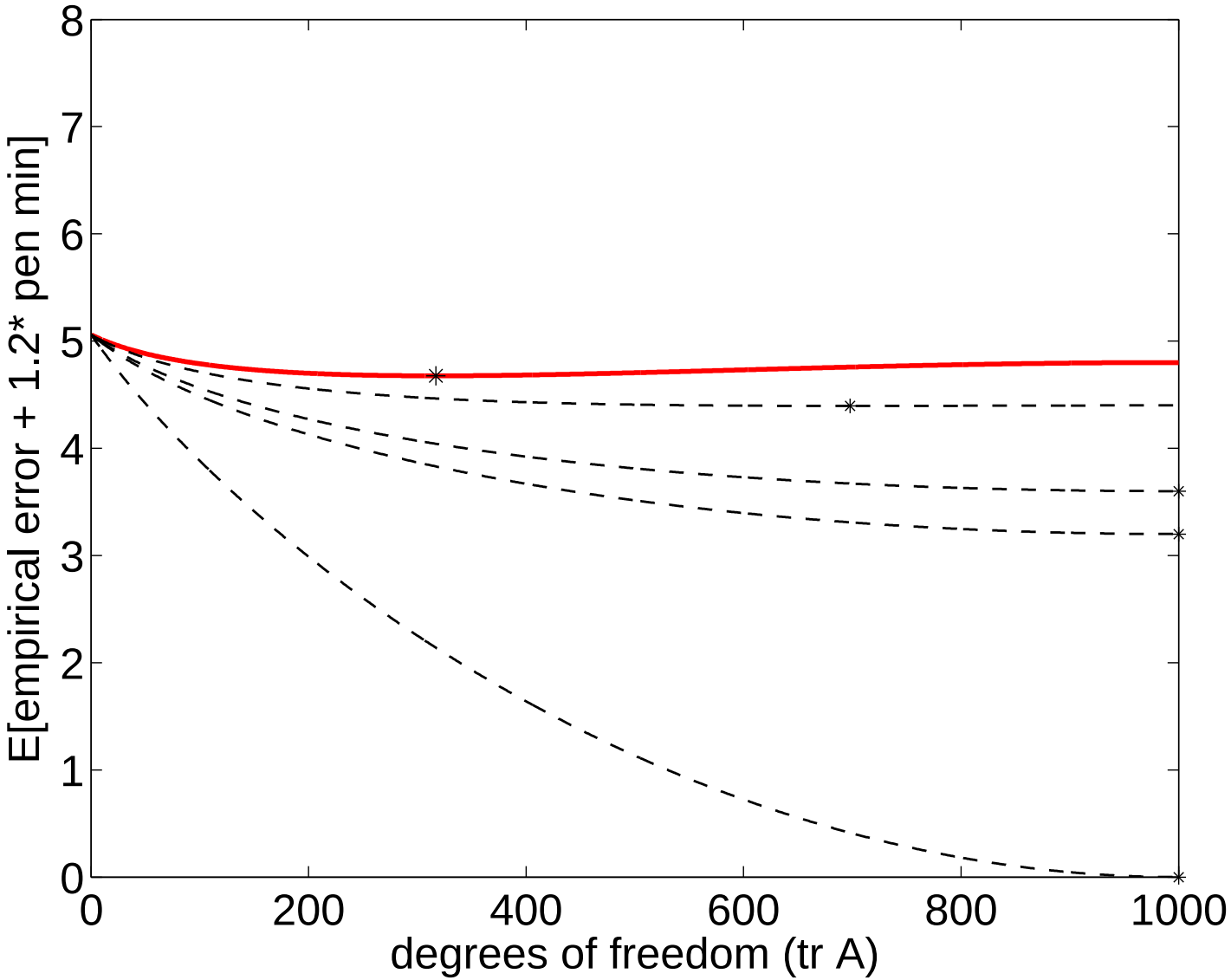
$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 0.9 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



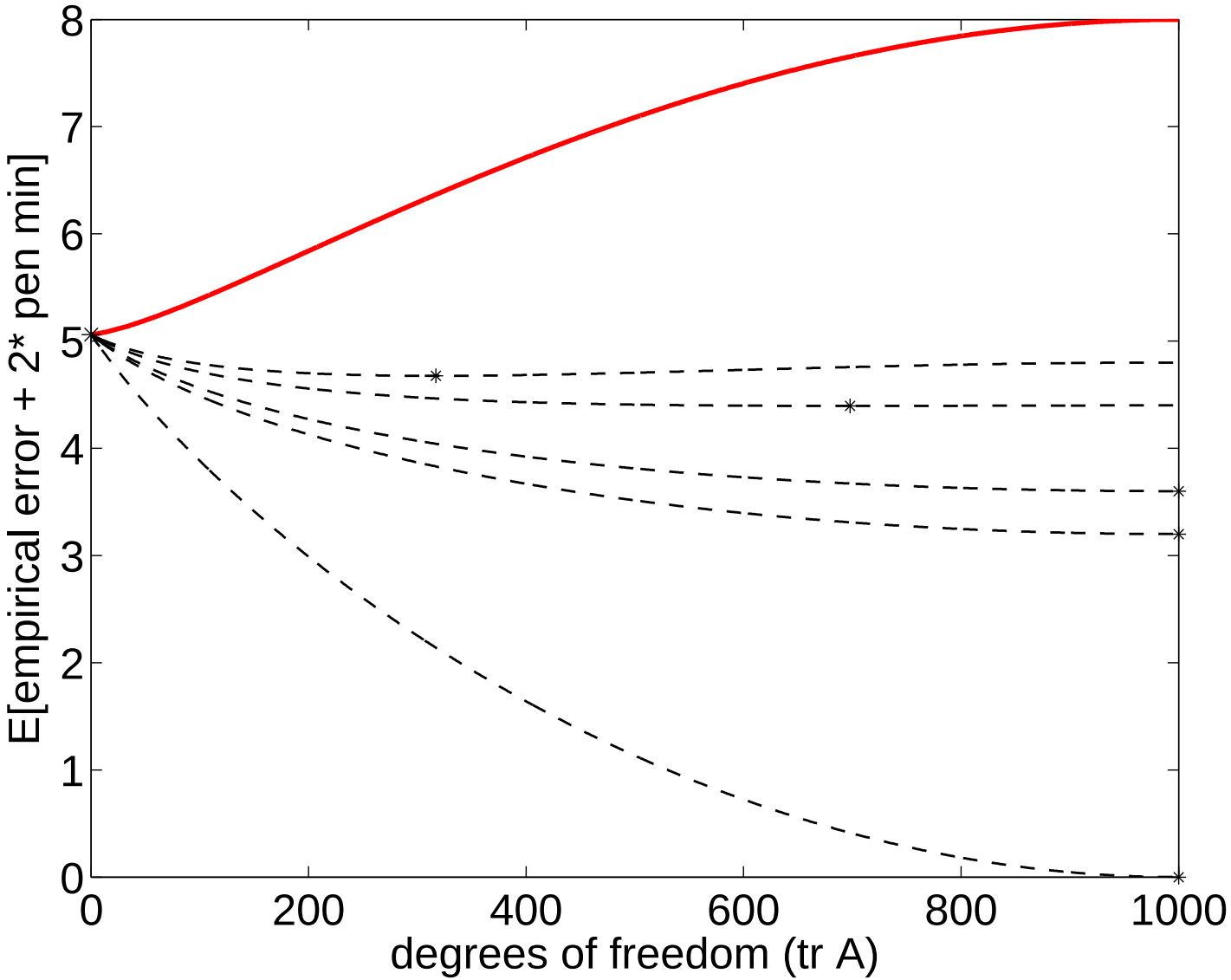
$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 1.1 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



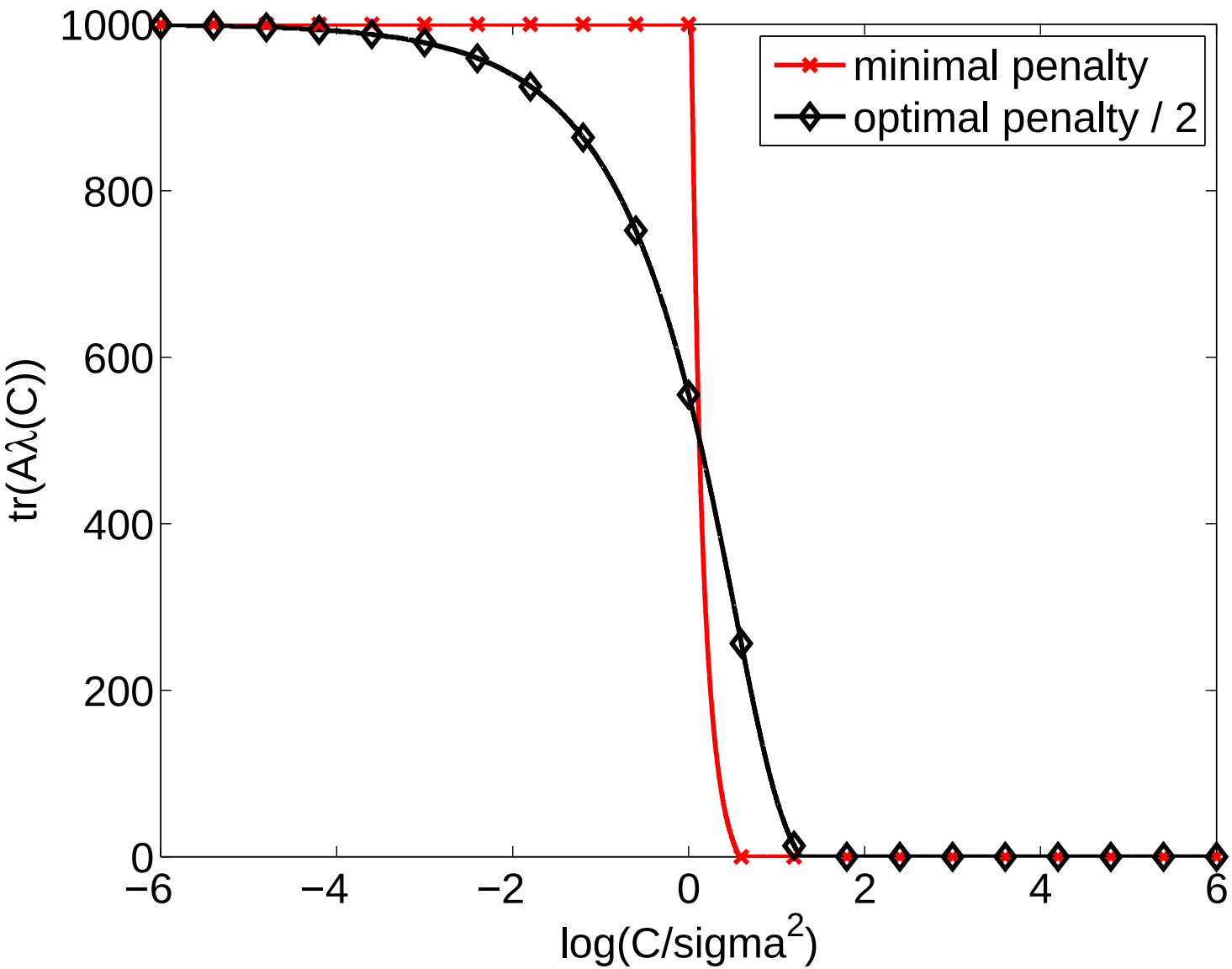
$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 1.2 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



$$\mathbb{E}[n^{-1}\|Y - \widehat{F}_m\|^2] + 2 \times \sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))n^{-1}$$



"Dimension" jump (ridge regression)



Penalty calibration algorithm (A. & Bach 2009)

1. for every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C \left(2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m) \right)}{n} \right\}$$

2. find $\widehat{C}_{\min}$ such that $\operatorname{tr}(A_{\widehat{m}_{\min}(C)})$ is "too large" when $C < \widehat{C}_{\min}$ and "reasonably small" when $C > \widehat{C}_{\min}$,

3. select

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2 \widehat{C}_{\min} \operatorname{tr}(A_m)}{n} \right\}$$

Comparison with least-squares

- Linear estimators:

$$\mathrm{pen}_{\mathrm{min}}(m) = \frac{\sigma^2 \left(2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m) \right)}{n}$$

$$\mathrm{pen}_{\mathrm{opt}}(m) = \frac{\sigma^2 \left(2 \operatorname{tr}(A_m) \right)}{n}$$

$$\frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\mathrm{min}}(m)} = \frac{2 \operatorname{tr}(A_m)}{2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)} \in (1, 2]$$

Comparison with least-squares

- Linear estimators:

$$\mathrm{pen}_{\mathrm{min}}(m) = \frac{\sigma^2 \left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m) \right)}{n}$$

$$\mathrm{pen}_{\mathrm{opt}}(m) = \frac{\sigma^2 \left(2\,\mathrm{tr}(A_m) \right)}{n}$$

$$\frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\mathrm{min}}(m)} = \frac{2\,\mathrm{tr}(A_m)}{2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)} \in (1, 2]$$

- Least-squares case:

$$A_m^\top A_m = A_m \quad \Rightarrow \quad \frac{\mathrm{pen}_{\mathrm{opt}}(m)}{\mathrm{pen}_{\mathrm{min}}(m)} = 2 \quad \Rightarrow \quad \text{Slope heuristics}$$

The k-nearest neighbours case

$$\forall i,j \in \{1,\dots,n\}\ , \quad A_{i,j} \in \left\{0, \frac{1}{k}\right\}$$

$$\forall i \in \{1,\dots,n\}\ , \quad A_{i,i} = \frac{1}{k} \quad \text{and} \quad \sum_{j=1}^{n} A_{i,j} = 1$$

The k-nearest neighbours case

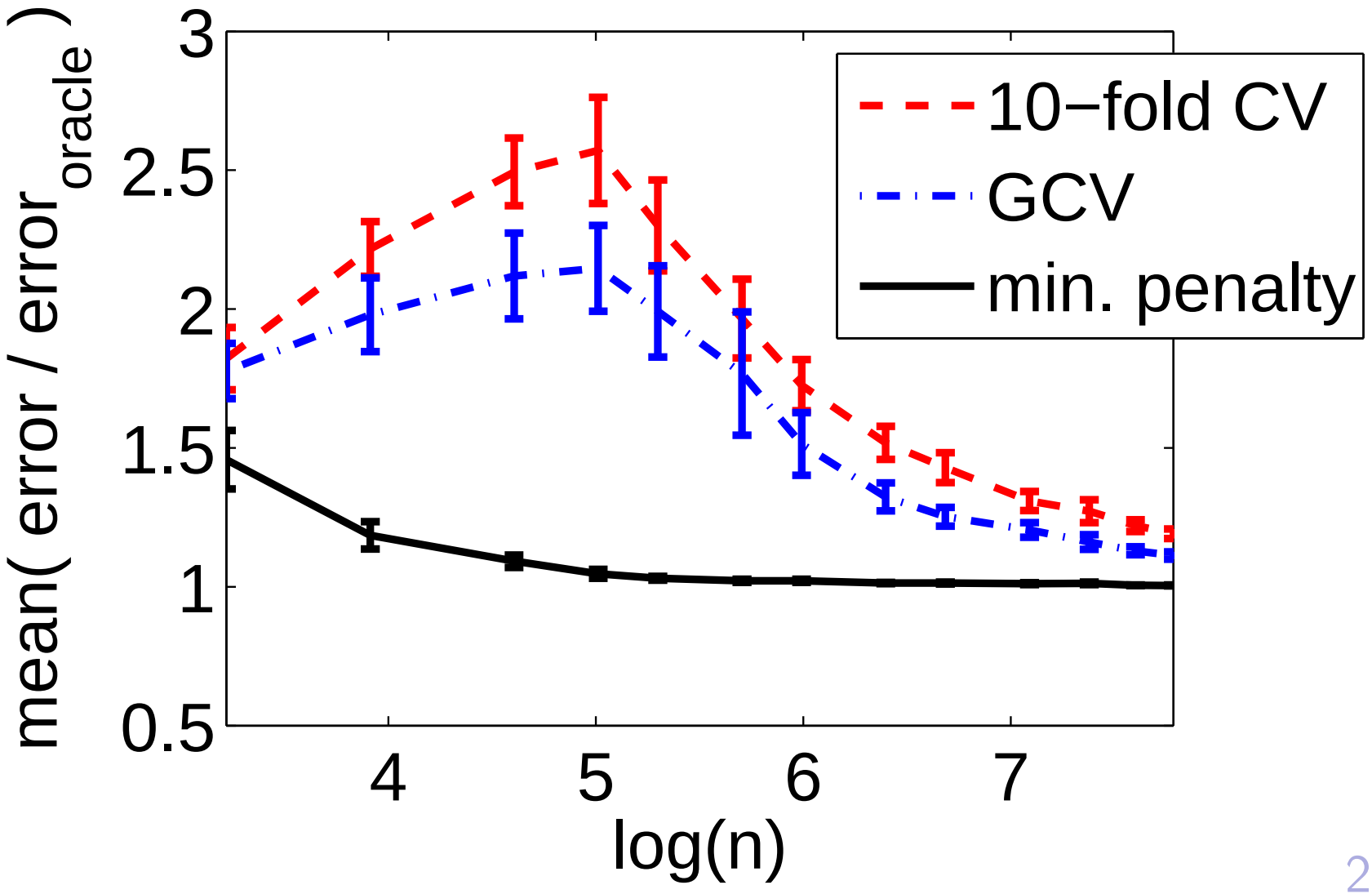
$$\forall i, j \in \{1, \ldots, n\}\ , \quad A_{i,j} \in \left\{0, \frac{1}{k}\right\}$$

$$\forall i \in \{1, \ldots, n\}\ , \quad A_{i,i} = \frac{1}{k} \quad \text{and} \quad \sum_{j=1}^{n} A_{i,j} = 1$$

$$\Rightarrow \quad \operatorname{tr}(A) = \frac{k}{n} = \operatorname{tr}(A^\top A)$$

$$\Rightarrow \quad \mathrm{pen}_{\mathrm{opt}} = 2\,\mathrm{pen}_{\mathrm{min}}$$

Simulation study (ridge regression, choice of λ)



Theorem (1): dimension jump (A. & Bach, 2009)

- polynomial complexity: $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha}$
- homoscedastic Gaussian noise, fixed design
- $\exists m_1,\, m_2 \in \mathcal{M}_n$ s.t. $\mathrm{tr}(A_{m_1}) \geq n/2$, $\mathrm{tr}(A_{m_2}) \leq \sqrt{n}$ and $\forall i \in \{1,2\}$, $n^{-1} \left\| (I - A_{m_i}) F \right\|^2 \leq \sigma^2 \sqrt{\ln(n)/n}$

Theorem (Minimal penalty; A. & Bach 2009)

With probability at least $1 - 8 C_{\mathcal{M}} n^{-2}$, *if* $n \geq n_0(\alpha)$,

$$\forall C < \left(1 - L_{\alpha} \sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \;, \qquad \mathrm{tr}\left(A_{\widehat{m}_{\min}(C)} \right) \geq \frac{n}{3}$$

$$\forall C > \left(1 + L_{\alpha} \frac{\sqrt{\ln(n)}}{n^{1/4}} \right) \sigma^2 \;, \qquad \mathrm{tr}\left(A_{\widehat{m}_{\min}(C)} \right) \leq n^{3/4}$$

Theorem (2): oracle inequality (A. & Bach, 2009)

Additional assumption:

$$\exists \kappa \geq 1 \ , \ \forall m \in \mathcal{M}_n \ , \qquad \mathrm{tr}(A_m)\sigma^2 \leq \kappa \mathbb{E}\left[\left\| F - \widehat{F}_m \right\|^2 \right]$$

Theorem (Oracle inequality; A. & Bach 2009)

Then, with probability at least $1 - 8 C_{\mathcal{M}} n^{-2}$, *si* $n \geq n_0(\alpha)$,

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

$$\frac{1}{n} \left\| F - \widehat{F}_{\widehat{m}} \right\|^2 \leq \left(1 + \frac{40\kappa}{\ln(n)} \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - \widehat{F}_m \right\|^2 \right\} + \frac{36(\kappa + \alpha + 2)\ln(n)\sigma^2}{n}$$

Outline

Minimal penalties: general case

$$\widehat{m}_{\min}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \operatorname{pen}_0(m) \right\}$$

- risk burst for $C < C^\star_{\min}$
- oracle inequality for $C > C^\star_{\min}$

Minimal penalties: general case

$$\widehat{m}_{\text{min}}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \, \text{pen}_0(m) \right\}$$

- risk burst for $C < C^\star_{\text{min}}$
- oracle inequality for $C > C^\star_{\text{min}}$
- jump of some complexity $\mathcal{C}_{\widehat{m}(C)}$ around $C = C^\star_{\text{min}}$
- $\text{pen}_{\text{min}} = C^\star_{\text{min}} \, \text{pen}_0$ with pen_0 known (or can be estimated)

Minimal penalties: general case

$$\widehat{m}_{\text{min}}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \, \text{pen}_0(m) \right\}$$

- risk burst for $C < C^\star_{\text{min}}$
- oracle inequality for $C > C^\star_{\text{min}}$
- jump of some complexity $\mathcal{C}_{\widehat{m}(C)}$ around $C = C^\star_{\text{min}}$
- $\text{pen}_{\text{min}} = C^\star_{\text{min}} \, \text{pen}_0$ with pen_0 known (or can be estimated)
- optimal oracle inequality when $\text{pen} = \text{pen}_{\text{opt}}$
- pen_{opt} known (that can be estimated) up to $C^\star$

Minimal penalties: general case

$$\widehat{m}_{\text{min}}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + C \, \text{pen}_0(m) \right\}$$

- risk burst for $C < C^\star_{\text{min}}$
- oracle inequality for $C > C^\star_{\text{min}}$
- jump of some complexity $\mathcal{C}_{\widehat{m}(C)}$ around $C = C^\star_{\text{min}}$
- $\text{pen}_{\text{min}} = C^\star_{\text{min}} \, \text{pen}_0$ with pen_0 known (or can be estimated)
- optimal oracle inequality when $\text{pen} = \text{pen}_{\text{opt}}$
- pen_{opt} known (that can be estimated) up to $C^\star$
- known relationship between $C^\star_{\text{min}}$ and $C^\star$

Calibration with minimal penalties

Inputs: $(\mathrm{pen}_0(m))_{m\in\mathcal{M}_n}$ $(\mathrm{pen}_1(m))_{m\in\mathcal{M}_n}$ $(\mathcal{C}_m)_{m\in\mathcal{M}_n}$ f

Assumptions: $\mathrm{pen}_0 = \frac{1}{C^\star_{\min}}\,\mathrm{pen}_{\min}$ $\mathrm{pen}_1 = \frac{1}{f(C^\star_{\min})}\,\mathrm{pen}_{\mathrm{opt}}$

1. for every $C > 0$, compute

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + C\,\mathrm{pen}_0(m) \right\}$$

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

3. select

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + f\left(\widehat{C}_{\min}\right)\mathrm{pen}_1(m) \right\}$$

Known mathematical results

- Concentration of p_2: minimum contrast estimators with bounded contrast (Boucheron & Massart, 2010)
- Burst of the risk if $C < C^\star_{\text{min}}$:
  - OLS, Gaussian noise, exponential complexity of $\mathcal{M}_n$ (Birgé & Massart, 2007)
  - OLS, Gaussian noise, medium range complexity of $\mathcal{M}_n$ and $s^\star = 0$ (Birgé & Massart, 2007)
  - Density estimation, Dantzig estimator, special case (Bertin, Le Pennec & Rivoirard, 2009)
  - Estimation of intensity of a Poisson process by thresholding (Reynaud-Bouret & Rivoirard, 2009), with $C^\star / C^\star_{\text{min}} \in [1, 12]$
- Burst of the dimension if $C < C^\star_{\text{min}}$: OLS, Gaussian noise, multiplicative penalties, polynomial or exponential complexity of $\mathcal{M}_n$ and $s^\star = 0$ (Baraud, Giraud & Huet, 2009)