Data-driven calibration of linear estimators with minimal penalties

Sylvain Arlot, Francis Bach

Data-driven calibration of linear estimators with minimal penalties

Sylvain Arlot *
CNRS ; Sierra project-team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France
sylvain.arlot@ens.fr

Francis Bach †
INRIA ; Sierra project-team
Laboratoire d'Informatique de
l'Ecole Normale Superieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, CS 81321
75214 Paris Cedex 13, France
francis.bach@ens.fr

September 12, 2011

Abstract

This paper tackles the problem of selecting among several linear estimators in non-parametric regression; this includes model selection for linear regression, the choice of a regularization parameter in kernel ridge regression, spline smoothing or locally weighted regression, and the choice of a kernel in multiple kernel learning. We propose a new algorithm which first estimates consistently the variance of the noise, based upon the concept of minimal penalty, which was previously introduced in the context of model selection. Then, plugging our variance estimate in Mallows' C_L penalty is proved to lead to an algorithm satisfying an oracle inequality. Simulation experiments with kernel ridge regression and multiple kernel learning show that the proposed algorithm often improves significantly existing calibration procedures such as generalized cross-validation.

1 Introduction

Smoothing splines or kernel-based methods are now well-established tools for supervised learning, allowing to perform various tasks, such as regression or binary classification, with linear and non-linear predictors [37, 36]. A central issue common to all regularization frameworks is the choice of the regularization parameter: while most practitioners use cross-validation procedures to select such a parameter, data-driven procedures not based on cross-validation are rarely used. The choice of the kernel, a seemingly unrelated issue, is also important for good predictive performance: several techniques exist, either based on cross-validation, Gaussian processes or multiple kernel learning [13, 34, 5].

In this paper, we consider least-squares regression and cast these two problems as the problem of selecting among several *linear estimators*, where the goal is to choose an estimator with a quadratic risk which is as small as possible. As shown in Section 2, this problem includes for instance model selection for linear regression, the choice of a regularization parameter in

*http://www.di.ens.fr/∼arlot/
†http://www.di.ens.fr/∼fbach/

kernel ridge regression, spline smoothing, or locally weighted regression, the choice of a kernel in multiple kernel learning, the choice of k in k-nearest-neighbors regression, and the choice of a bandwidth of Nadaraya-Watson estimators.

Another motivation for studying linear estimators is their good theoretical properties. For instance, when the signal belongs to a Sobolev ball, it is known the Pinsker estimator (which is linear) is asymptotically minimax up to the optimal constant, while the best projection estimator is only rate-minimax [16, 43]. Furthermore, the set of signals that are well estimated by linear estimators is very rich: it contains, for instance, sampled smooth functions, sampled modulated smooth functions and sampled harmonic functions [20]. Finally, convergence rates of some linear spectral estimators have recently been proved optimal [12, 8].

The main contribution of the paper is to extend the notion of *minimal penalty* [9, 2] presented in Section 2 to all discrete classes of linear operators, and to use it for defining a fully data-driven selection algorithm satisfying a non-asymptotic oracle inequality. Our new theoretical results presented in Section 4 extend similar results which were limited to unregularized least-squares regression (i.e., projection operators). We also tackle continuously parameterized families of linear estimators which are typical in ridge regression and spline smoothing (where the one-dimensional parameter to be estimated is the regularization parameter). In order to do, we derive novel concentration inequalities which may be useful in other contexts (Section B). Our results also enlighten the classical elbow heuristics based algorithms—e.g., "L-curve maximum curvature criterion" for Tikhonov [35] and several others regularization problems [19, 18])—by providing theoretical grounds to another L-curve based calibration algorithm. Finally, in Section 5, we show that our algorithm improves the performances of classical selection procedures, such as GCV [14], for kernel ridge regression, nearest-neighbor regression or locally weighted regression.

2 Linear estimators

In this section, we define the problem we aim to solve and give several examples of linear estimators.

2.1 Framework and notation

Let us assume that one observes

$$Y_i = f(x_i) + \varepsilon_i \in \mathbb{R} \qquad \text{for} \quad i = 1, \dots, n \ ,$$

where $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered random variables with $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ unknown, f is an unknown measurable function $\mathcal{X} \mapsto \mathbb{R}$ and $x_1, \dots, x_n \in \mathcal{X}$ are deterministic design points. No assumption is made on the set $\mathcal{X}$. The goal is to reconstruct the signal $F = (f(x_i))_{1 \le i \le n} \in \mathbb{R}^n$, with some estimator $\widehat{F} \in \mathbb{R}^n$, depending only on $(x_1, Y_1), \dots, (x_n, Y_n)$, and having a small quadratic risk $n^{-1}\|\widehat{F} - F\|_2^2$, where $\forall t \in \mathbb{R}^n$, we denote by $\|t\|_2$ the ℓ_2-norm of t, defined as $\|t\|_2^2 := \sum_{i=1}^n t_i^2$.

In this paper, we focus on *linear estimators* $\widehat{F}$ that can be written as a linear function of $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, that is, $\widehat{F} = AY$, for some (deterministic) $n \times n$ matrix A. Here and in the rest of the paper, vectors such as Y or F are assumed to be column-vectors. We present in Section 2.2 several important families of estimators of this form. The matrix A may depend on $x_1, \dots, x_n$ (which are known and deterministic), but not on Y, and may be parameterized by certain quantities—usually regularization parameter or kernel combination weights.

Let us also define, for any matrix $A \in \mathcal{M}_n(\mathbb{R})$, the largest singular value of A:

$$|\!|\!| A |\!|\!| := \sup_{t \in \mathbb{R}^n, \, t \neq 0} \left\{ \frac{\|At\|_2}{\|t\|_2} \right\} \ .$$

2.2 Examples of linear estimators

In this paper, our theoretical results apply to matrices A such that

$$A \in \mathcal{M}_n(\mathbb{R}) \quad \|\!|A|\!\| \leq \mathbb{M} \quad \operatorname{tr}(A^\top A) \leq (2 - K_{\mathrm{df}}) \operatorname{tr}(A) \text{ with } K_{\mathrm{df}} \in (0,2) \ , \tag{1}$$

for some constants $\mathbb{M}$ and K_{df}. The main examples we have in mind are the following.

Ordinary least-squares regression / model selection. If we consider linear predictors (here, linear in the inputs $x_1, \dots, x_n$) from a design matrix $X \in \mathbb{R}^{n \times p}$, then $\widehat{F} = AY$ with $A = X(X^\top X)^{-1} X^\top$, which is a projection matrix (i.e., $A^\top A = A$); $\widehat{F} = AY$ is often called a *projection estimator*. In the variable selection setting, one wants to select a subset $J \subset \{1, \dots, p\}$, and matrices A are parameterized by J. If we denote X_J the matrix of size $n \times |J|$ composed of the columns of X indexed by J, then the matrix A_J is equal to $X_J (X_J^\top X_J)^{-1} X_J^\top$. For this matrix, we have $\operatorname{tr} A_J^\top A_J = \operatorname{tr} A_J^2 = \operatorname{tr} A_J$.

Kernel ridge regression / spline smoothing. We assume that a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given, and we are looking for a function $f : \mathcal{X} \to \mathbb{R}$ in the associated reproducing kernel Hilbert space (RKHS) $\mathcal{F}$, with norm $\| \cdot \|_{\mathcal{F}}$. If K denotes the $n \times n$ kernel matrix, defined by $K_{ab} = k(x_a, x_b)$, then the ridge regression estimator—a.k.a. spline smoothing estimator for spline kernels [44], or Tikhonov regularization [41]—is obtained by minimizing with respect to $f \in \mathcal{F}$ [36]:

$$\frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2 \ .$$

The unique solution is equal to $\widehat{f} = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$, where $\alpha = (K + n\lambda I)^{-1} Y$. This leads to the smoothing matrix $A_\lambda = K(K + n\lambda I_n)^{-1}$, parameterized by the regularization parameter $\lambda \in \mathbb{R}_+$. In this case, A is symmetric positive semi-definite, and we have $\operatorname{tr} A^2 \leqslant \operatorname{tr} A$.

Multiple kernel learning / Group Lasso / Lasso. We now assume that we have p different kernels k_j, feature spaces $\mathcal{F}_j$ and feature maps $\Phi_j : \mathcal{X} \to \mathcal{F}_j$, $j = 1, \dots, p$. The group Lasso [48] and multiple kernel learning [21, 5] frameworks consider the following objective function

$$J(f_1, \dots, f_p) = \tfrac{1}{n} \sum_{i=1}^n \big(y_i - \textstyle\sum_{j=1}^p \langle f_j, \Phi_j(x_i) \rangle \big)^2 + 2\lambda \displaystyle\sum_{j=1}^p \|f_j\|_{\mathcal{F}_j} = L(f_1, \dots, f_p) + 2\lambda \sum_{j=1}^p \|f_j\|_{\mathcal{F}_j} \ .$$

Note that when $\Phi_j(x)$ is simply the j-th coordinate of $x \in \mathbb{R}^p$, we get back the penalization by the ℓ^1-norm and thus the regular Lasso [40].

Following [32, 33], by using $a^{1/2} = \min_{b \geqslant 0} \frac{1}{2} \{ \frac{a}{b} + b \}$, we obtain a variational formulation of the sum of norms $2 \sum_{j=1}^p \|f_j\| = \min_{\eta \in \mathbb{R}_+^p} \sum_{j=1}^p \left\{ \frac{\|f_j\|^2}{\eta_j} + \eta_j \right\}$. Thus, minimizing $J(f_1, \dots, f_p)$ with respect to $(f_1, \dots, f_p)$ is equivalent to minimizing with respect to $\eta \in \mathbb{R}_+^p$ (see [5] for more details):

$$\min_{f_1, \dots, f_p} L(f_1, \dots, f_p) + \lambda \sum_{j=1}^p \frac{\|f_j\|^2}{\eta_j} + \lambda \sum_{j=1}^p \eta_j = \frac{1}{n} y^\top \big(\textstyle\sum_{j=1}^p \eta_j K_j + n\lambda I_n \big)^{-1} y + \lambda \displaystyle\sum_{j=1}^p \eta_j \ ,$$

where I_n is the $n \times n$ identity matrix. Moreover, given η, this leads to a smoothing matrix of the form

$$A_{\eta, \lambda} = (\textstyle\sum_{j=1}^p \eta_j K_j)(\sum_{j=1}^p \eta_j K_j + n\lambda I_n)^{-1} \ , \tag{2}$$

Method	A	parameter
Ridge regression	$K(K+\lambda I)^{-1}$	λ
Kernel learning	$K(K+I)^{-1}$	K
Nadaraya-Watson	$W\mathrm{Diag}(W\mathbf{1})^{-1}$	α where $W_{ij} = \exp(-\alpha\|x_i - x_j\|^2)$
Nearest-neighbor	$A \in \{0, 1/k\}^{N\times N}$	k

Table 1: Examples of linear estimators.

parameterized by the regularization parameter $\lambda \in \mathbb{R}_+$ and the kernel combinations in $\mathbb{R}_+^p$—note that it depends only on $\lambda^{-1}\eta$, which can be grouped in a single parameter in $\mathbb{R}_+^p$. Note that it corresponds to a specific parameterization of the kernel matrix K using η, and that it can be extended to other types of parameterization.

Thus, the Lasso/group lasso can be seen as particular (convex) ways of optimizing over η. In this paper, we propose a non-convex alternative with better statistical properties (oracle inequality in Theorem 3). Note that in our setting, finding the solution of the problem is hard in general since the optimization is not convex. However, while the model selection problem is by nature combinatorial, our optimization problems for multiple kernels are all differentiable and are thus amenable to gradient descent procedures—which only find local optima.

Nearest-neighbor regression. If we assume that we are given any similarity measure $d : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ and k a strictly positive integer, then, from n observations $x_1, \dots, x_n$, we may for each $i = 1, \dots, n$, find k-nearest neighbors of x_i, i.e., find any set J_i of k points x_j, $j \in \{1, \dots, n\}\backslash\{k\}$, which are among the k closest to x_i according to d (this definition takes into account possible ties). We can then build an $n \times n$ matrix A of nearest neighbors which is equal to $1/k$ for all pairs (i, j) such that $j \in J_i$ for all $i \in \{1, \dots, n\}$, and equal to zero otherwise.

Nadaraya-Watson estimators [30, 46]. We know assume that we are given a "window function" (not to be confused with a positive definite kernel) $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, from which we build the $n \times n$ matrix W of pairwise evaluations. The estimator correspond to the matrix A obtained by normalizing W to have unit row-sums, i.e., $A = WD^{-1}$, where $D = \mathrm{Diag}(W\mathbf{1})$ is the diagonal matrix of row sums. In this situation we have $\mathbb{M} \leq \sqrt{\max_i D_{ii}/\min_i D_{ii}}$. A typical example is the matrix W defined as $W_{ij} = \exp(-\alpha\|x_i - x_j\|^2)$ where x_i, $i = 1, \dots, n$, are the observed input data points, and α is the smoothing parameter to be learned.

Except for the bound on $|||A|||$ in the k-nearest neighbor and Nadaraya-Watson examples, Eq. (1) holds true with $\mathbb{M} = 1$ and $K_{\mathrm{df}} = 1$ for all the examples mentioned, as shown by the following result.

Proposition 1. *For any $n \geq 1$, $\mathrm{tr}(A^\top A) \leq \mathrm{tr}(A) \leq n$ for any matrix $A \in \mathcal{M}_n(\mathbb{R})$ among the following examples*:

(i) if A is symmetric with $\mathrm{Sp}(A) \subset [0, 1]$, for instance:

(ia) Ordinary least-squares regression: A is an orthogonal projection matrix.

(ib) Kernel ridge regression, Multiple kernel learning: $\exists x \in (0, +\infty)$ and $K \in \mathcal{M}_n(\mathbb{R})$ symmetric positive semi-definite such that $A = K(K + xI)^{-1}$.

(ii) if $\forall i, j$, $A_{i,i} \geq A_{i,j} \geq 0$ and $\sum_{k=1}^n A_{i,k} = 1$, for instance:

(iia) Nadaraya-Watson regression

(iib) k-nearest-neighbor regression, for some integer $k \in [1,n]$:

$$\left\{ \begin{array}{c} \forall 1 \leq i,j \leq n\,, \quad A_{i,j} \in \left\{0, \dfrac{1}{k}\right\} \quad \textit{with } k \in \{1,\dots,n\} \\ \forall 1 \leq i \leq n\,, \quad A_{i,i} = \dfrac{1}{k} \quad \textit{and} \quad \displaystyle\sum_{j=1}^{n} A_{i,j} = 1 \ . \end{array} \right. \tag{\textbf{kNN}}$$

In example (i), $|||A||| \leq 1$. *In examples (ia) and (iib),* $\mathrm{tr}(A) = \mathrm{tr}(A^\top A)$.

Proposition 1 is proved in Section C.

Other examples. Alternative linear estimators are classical in the statistical or learning literature:

- Pinsker filters [31, 16], that is, $A_{w,\alpha} = \mathrm{diag}\left((1 - (k^\alpha/w))_+ \,,\, k = 1 \dots n \right)$ for some parameters $w, \alpha > 0$. This example matches case (i) in Proposition 1.
- Linear spectral methods for statistical inverse problems [10, 26], such as spectral cut-off (or principal components regression) and ℓ_2-boosting.
- Symmetrized k-nearest neighbors [47].

More examples and references can be found in [45, Chapter 5] and [43, Chapter 3], for instance.

3 Linear estimator selection

In this section, we first describe the statistical framework of linear estimator selection, then introduce the notion of minimal penalty. Finally, we briefly review the related work on linear estimator selection.

3.1 Unbiased risk estimation heuristics

Usually, several estimators of the form $\widehat{F} = AY$ can be used. The problem that we consider in this paper is then to select one of them, that is, to choose a matrix A . Let us assume that a family of matrices $(A_\lambda)_{\lambda \in \Lambda}$ is given (examples are shown in Section 2.2), hence a family of estimators $(\widehat{F}_\lambda)_{\lambda \in \Lambda}$ can be used, with $\widehat{F}_\lambda := A_\lambda Y$. The goal is to choose *from data* some $\widehat{\lambda} \in \Lambda$, so that the quadratic risk of $\widehat{F}_{\widehat{\lambda}}$ is as small as possible.

The best choice would be the *oracle*:

$$\lambda^\star \in \arg\min_{\lambda \in \Lambda} \left\{ n^{-1} \|\widehat{F}_\lambda - F\|_2^2 \right\} \ ,$$

which cannot be used since it depends on the unknown signal F . Therefore, the goal is to define a data-driven $\widehat{\lambda}$ satisfying an *oracle inequality* of the form

$$n^{-1} \|\widehat{F}_{\widehat{\lambda}} - F\|_2^2 \leq C_n \inf_{\lambda \in \Lambda} \left\{ n^{-1} \|\widehat{F}_\lambda - F\|_2^2 \right\} + R_n \ , \tag{3}$$

with large probability, where the leading constant C_n should be close to 1 (at least for large n) and the remainder term R_n should be negligible compared to the risk of the oracle. Many classical

selection methods are built upon the "unbiased risk estimation" heuristics: If $\widehat{\lambda}$ minimizes a criterion $\mathrm{crit}(\lambda)$ such that

$$\forall \lambda \in \Lambda\,, \qquad \mathbb{E}\left[\mathrm{crit}(\lambda)\right] \approx \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] \ ,$$

then $\widehat{\lambda}$ satisfies an oracle inequality such as in Eq. (3) with large probability. For instance, cross-validation [1, 39] and generalized cross-validation (GCV) [14] are built upon this heuristics.

One way of implementing this heuristics is penalization, which consists in minimizing the sum of the empirical risk and a penalty term, i.e., using a criterion of the form:

$$\mathrm{crit}(\lambda) = n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + \mathrm{pen}(\lambda) \ .$$

The unbiased risk estimation heuristics, also called Mallows' heuristics, then leads to the *optimal (deterministic) penalty*

$$\mathrm{pen}_{\mathrm{id}}(\lambda) := \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] - \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - Y\|_2^2\right] \ .$$

When $\widehat{F}_\lambda = A_\lambda Y$, we have:

$$\|\widehat{F}_\lambda - F\|_2^2 = \|(A_\lambda - I_n)F\|_2^2 + \|A_\lambda\varepsilon\|_2^2 + 2\left\langle A_\lambda\varepsilon,\, (A_\lambda - I_n)F\right\rangle \ , \tag{4}$$

$$\|\widehat{F}_\lambda - Y\|_2^2 = \|\widehat{F}_\lambda - F\|_2^2 + \|\varepsilon\|_2^2 - 2\left\langle \varepsilon,\, A_\lambda\varepsilon\right\rangle + 2\left\langle \varepsilon,\, (I_n - A_\lambda)F\right\rangle \ , \tag{5}$$

where $\varepsilon = Y - F \in \mathbb{R}^n$ and $\forall t, u \in \mathbb{R}^n$, $\langle t,\, u\rangle = \sum_{i=1}^n t_i u_i$. Since ε is centered with covariance matrix $\sigma^2 I_n$, Eq. (4) and Eq. (5) imply that

$$\mathrm{pen}_{\mathrm{id}}(\lambda) = \frac{2\sigma^2\,\mathrm{tr}(A_\lambda)}{n} \ , \tag{6}$$

up to the term $-\mathbb{E}[n^{-1}\|\varepsilon\|_2^2] = -\sigma^2$, which can be dropped off since it does not vary with λ.

Note that $\mathrm{df}(\lambda) = \mathrm{tr}(A_\lambda)$ is called the *effective dimensionality* or *degrees of freedom* [49], so that the optimal penalty in Eq. (6) is proportional to the dimensionality associated with the matrix A_λ—for projection matrices, we get back the dimension of the subspace, which is classical in model selection.

The expression of the optimal penalty in Eq. (6) led to several selection procedures, in particular Mallows' C_L (called C_p in the case of projection estimators) [27], where σ^2 is replaced by some estimator $\widehat{\sigma^2}$. The estimator of σ^2 usually used with C_L is based upon the value of the empirical risk at some λ_0 with $\mathrm{df}(\lambda_0)$ large; it has the drawback of overestimating the risk, in a way which depends on λ_0 and F [17]. GCV, which implicitly estimates σ^2, has the drawback of overfitting if the family $(A_\lambda)_{\lambda\in\Lambda}$ contains a matrix too close to I_n [11], so that examples have been given where GCV is not asymptotically optimal [25]; GCV also overestimates the risk even more than C_L for most A_λ (see (7.9) and Table 4 in [17]).

In this paper, we define an estimator of σ^2 directly related to the selection task which does not have similar drawbacks. Our estimator relies on the concept of minimal penalty, introduced by Birgé and Massart [9] and further studied in [2].

3.2 Minimal and optimal penalties

We deduce from Eq. (4) the *bias-variance decomposition* of the risk:

$$\mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - F\|_2^2\right] = n^{-1}\,\|(A_\lambda - I_n)F\|_2^2 + \frac{\mathrm{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} = \mathrm{bias} + \mathrm{variance} \ , \tag{7}$$

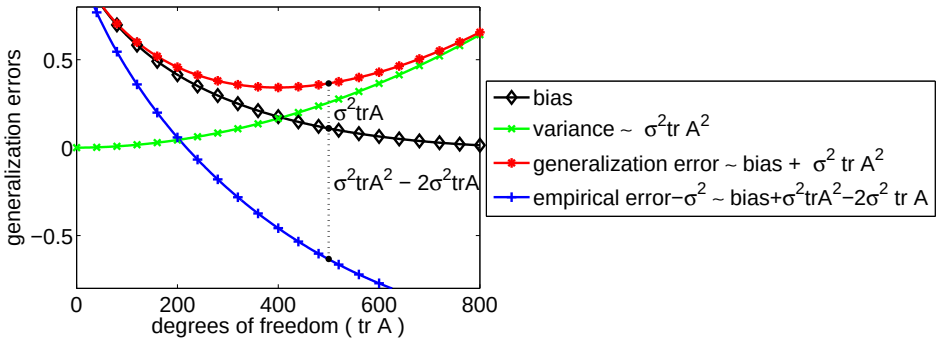

Figure 1: Bias-variance decomposition of the generalization error, and minimal/optimal penalties.

and from Eq. (5) the expectation of the empirical risk:

$$\mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 - \|\varepsilon\|_2^2\right] = n^{-1}\left\|(A_\lambda - I_n)F\right\|_2^2 - \frac{\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2}{n} \ . \tag{8}$$

Note that the variance term in Eq. (7) is not proportional to the effective dimensionality $\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda)$ but to $\operatorname{tr}(A_\lambda^\top A_\lambda)$. Although several papers argue these terms are of the same order (for instance, they are equal when A_λ is a projection matrix), this may not hold in general. If Eq. (1) holds for all matrices $A \in \{A_\lambda\}_{\lambda\in\Lambda}$, we only have

$$0 \leq \frac{\operatorname{tr}(A_\lambda^\top A_\lambda)}{2 - K_{\mathrm{df}}} \leq \operatorname{tr}(A_\lambda) \leq \frac{2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)}{K_{\mathrm{df}}} \leq \frac{2\operatorname{tr}(A_\lambda)}{K_{\mathrm{df}}} \ . \tag{9}$$

In order to give a first intuitive interpretation of Eq. (7) and Eq. (8), let us consider the kernel ridge regression example, where $A = K(K + \lambda I)^{-1}$, and assume that the risk and the empirical risk behave as their expectations in Eq. (7) and Eq. (8); see also Fig. 1. Completely rigorous arguments based upon concentration inequalities are developed in the Appendix and summarized in Section 4, leading to the same conclusions as the present informal reasoning.

First, as proved by Lemma 2 in Section B.2, the bias $n^{-1}\left\|(A_\lambda - I_n)F\right\|_2^2$ is a non-increasing function of the dimensionality $\mathrm{df}(\lambda) = \operatorname{tr}(A_\lambda)$, and the variance $\operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2 n^{-1}$ is an increasing function of $\mathrm{df}(\lambda)$, as well as $2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)$. Therefore, Eq. (7) shows that the optimal λ realizes the best trade-off between bias (which decreases with $\mathrm{df}(\lambda)$) and variance (which increases with $\mathrm{df}(\lambda)$), which is a classical fact in model selection (see Figure 1).

Second, the expectation of the empirical risk in Eq. (8) can be decomposed into the bias and a negative variance term which is the opposite of

$$\mathrm{pen}_{\mathrm{min}}(\lambda) := n^{-1}\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\sigma^2 \ . \tag{10}$$

As suggested by the notation $\mathrm{pen}_{\mathrm{min}}$, we will show it is a *minimal penalty* in the following sense. If

$$\forall D \geq 0, \qquad \widehat{\lambda}_{\mathrm{min}}(D) \in \arg\min_{\lambda\in\Lambda}\left\{n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + D\,\mathrm{pen}_{\mathrm{min}}(\lambda)\right\} \ ,$$

then, up to concentration inequalities that are detailed in Section 4.4, $\widehat{\lambda}_{\mathrm{min}}(D)$ behaves like a minimizer of

$$g_D(\lambda) = \mathbb{E}\left[n^{-1}\|\widehat{F}_\lambda - Y\|_2^2 + D\,\mathrm{pen}_{\mathrm{min}}(\lambda)\right] - \sigma^2 = n^{-1}\left\|(A_\lambda - I_n)F\right\|_2^2 + (D-1)\,\mathrm{pen}_{\mathrm{min}}(\lambda) \ .$$

Therefore, two main cases can be distinguished:

- if $D < 1$, then $g_D(\lambda)$ decreases with $\mathrm{df}(\lambda)$ so that $\mathrm{df}(\widehat{\lambda}_{\min}(D))$ is huge: $\widehat{\lambda}_{\min}(D)$ overfits.
- if $D > 1$, then $g_D(\lambda)$ increases with $\mathrm{df}(\lambda)$ when $\mathrm{df}(\lambda)$ is large enough, so that $\mathrm{df}(\widehat{\lambda}_{\min}(D))$ is much smaller than when $D < 1$.

As a conclusion, $\mathrm{pen}_{\min}(\lambda)$ is the minimal amount of penalization needed so that a minimizer $\widehat{\lambda}$ of a penalized criterion is not clearly overfitting.

Following an idea first proposed in [9] and further analyzed or used in several other papers such as [23, 2, 29], we now propose to use that $\mathrm{pen}_{\min}(\lambda)$ is a minimal penalty for estimating σ^2 and plug this estimator into Eq. (6). Indeed, if we penalize the empirical risk $n^{-1}\|\widehat{F}_\lambda - Y\|_2^2$ by $C\frac{\mathrm{pen}_{\min}(\lambda)}{\sigma^2}$ (which does not depend on σ^2), then the argument above suggests that around the value $D = C\sigma^{-2} = 1$ (i.e., around $C = \sigma^2$), we have a jump in the selected degrees of freedom. This leads to the algorithm described in Section 4.1.

Note that the minimal penalty given by Eq. (10) is new; it generalizes previous results [9, 2] where $\mathrm{pen}_{\min}(A_\lambda) = n^{-1}\operatorname{tr}(A_\lambda)\sigma^2$ because all A_λ were assumed to be projection matrices, i.e., $A_\lambda^\top A_\lambda = A_\lambda$. Furthermore, our results generalize the slope heuristics $\mathrm{pen}_{\mathrm{id}} \approx 2\,\mathrm{pen}_{\min}$ (only valid for projection estimators [9, 2]) to general linear estimators for which $\mathrm{pen}_{\mathrm{id}} / \mathrm{pen}_{\min} \in (1, 2]$.

3.3 Related work

Several procedures have been proposed in the literature for selecting among linear estimators. The most classical ones are Mallows' C_L [27] and GCV [14] (which have already been introduced) and cross-validation (see [4] for references).

Recently, Baraud, Giraud and Huet [7] proposed an estimator selection procedure via penalization, that applies in particular to linear estimator selection; a possible drawback of their procedure is that it strongly assumes the noise is Gaussian, since the Gaussian distribution explicitly appears in the definition of their penalty. Two penalized maximum likelihood criteria have also been proposed for selecting the ridge regression parameter [42], but they are only supported by simulation experiments.

Finally, let us mention here an aggregation procedure recently proposed by Dalalyan and Salmon [15] for affine estimators. Their goal is different from ours (aggregating instead of selecting), but still related since they prove some oracle inequalities for their final estimator.

4 Main results

In this section, we first describe our algorithm and then present our theoretical results.

4.1 Algorithm

The following algorithm first computes an estimator of $\widehat{C}$ of σ^2 using the minimal penalty in Eq. (10), then considers the optimal penalty in Eq. (6) for selecting λ.

Input: $Y \in \mathbb{R}^n$ and $\{A_\lambda\}_{\lambda\in\Lambda}$ a collection of matrices

1. $\forall C > 0$, compute $\widehat{\lambda}_0(C) \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + C\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)\right)\}$.
2. Find $\widehat{C}$ such that $\mathrm{df}(\widehat{\lambda}_0(\widehat{C})) \in [n/10, n/3]$.
3. Select $\widehat{\lambda} \in \arg\min_{\lambda\in\Lambda}\{\|\widehat{F}_\lambda - Y\|_2^2 + 2\widehat{C}\operatorname{tr}(A_\lambda)\}$.

In the steps 1 and 2 of the above algorithm, in practice, a grid in log-scale is used, and our theoretical results from the next section suggest to use a step-size of order $n^{-1/2}$. Step 1 can be solved efficiently (at least when Λ is finite or Λ is embedded with a total order) thanks to Algorithm 2 in [2].

Note that it may not be possible in all cases to find a C such that $\mathrm{df}(\widehat{\lambda}_0(C)) \in [n/10, n/3]$; therefore, our condition in step 2, could be relaxed to finding a $\widehat{C}$ such that for all $C > \widehat{C}(1+\delta)$, $\mathrm{df}(\widehat{\lambda}_0(C)) < n/10$ and for all $C < \widehat{C}/(1+\delta)$, $\mathrm{df}(\widehat{\lambda}_0(C)) > n/10$, with $\delta \propto \sqrt{\ln(n)/n}$.

Alternatively, using the same grid in log-scale, we can select $\widehat{C}$ with maximal jump between successive values of $\mathrm{df}(\widehat{\lambda}_0(C))$—note that our theoretical result then does not entirely hold, as we show the presence of a jump around σ^2, but do not show the absence of similar jumps elsewhere. See examples in Section 5.

4.2 Assumptions

Before stating our main results, let us state the main assumptions we make on $(A_\lambda)_{\lambda\in\Lambda}$ and on the distribution of the noise. We essentially consider a set of linear estimators which corresponds to a union of a discrete set and a union of matrices obtained from ridge regression (which are themselves parameterized by a single continuous parameter).

- Assumption on the matrices A_λ: some constants $K_{\mathrm{df}} \in (0,2)$ and $\mathbb{M} \geq 1$ exists such that

$$\left.\begin{array}{l} \forall \lambda \in \Lambda\,, \quad A_\lambda \in \mathcal{M}_n(\mathbb{R}) \text{ is deterministic}, \quad |\!|\!| A_\lambda |\!|\!| \leq \mathbb{M} \\ \mathrm{tr}\,(A_\lambda) \leq n \quad \text{and} \quad \mathrm{tr}(A_\lambda^\top A_\lambda) \leq (2 - K_{\mathrm{df}})\,\mathrm{tr}(A_\lambda)\ . \end{array}\right\} \qquad (\mathbf{HA}_\lambda)$$

- Assumption on Λ: $I \in \{\, A_\lambda \,\}_{\lambda\in\Lambda}$ and

$$\left.\begin{array}{l} \Lambda \subset \Lambda_0 \cup \displaystyle\bigcup_{j=1}^{N_\Lambda^r} (\,\{\, j \,\} \times [0,+\infty]\,) \quad \text{with} \quad \mathrm{Card}(\Lambda_0) \leq C_\Lambda^d n^{\alpha_\Lambda^d} \qquad N_\Lambda^r \leq C_\Lambda^r n^{\alpha_\Lambda^r} \\ \text{and} \quad \forall j \in \{\, 1, \ldots, N_\Lambda^r \,\}\,, \quad \forall x \in (0,+\infty)\,, \quad A_{(j,x)} = K_j (K_j + nx I_n)^{-1} \\ \qquad \text{with} \quad K_j \in \mathcal{M}_n(\mathbb{R}) \backslash \left\{ \mathbf{0}_{\mathcal{M}_n(\mathbb{R})} \right\} \quad \text{symmetric positive semi-definite} \\ \qquad \text{and} \quad A_{(j,0)} = I \quad A_{(j,+\infty)} = \mathbf{0}_{\mathcal{M}_n(\mathbb{R})} \end{array}\right\} \qquad (\mathbf{H\Lambda})$$

- Assumption on the noise:

$$\varepsilon_1, \ldots, \varepsilon_n \text{ are i.i.d. } \sim \mathcal{N}(0, \sigma^2) \qquad (\mathbf{H\mathcal{N}\sigma^2})$$

- Assumption on the bias:

$$\exists \lambda_1 \in \Lambda, \quad \mathrm{df}(\lambda_1) \leq \sqrt{n} \quad \text{and} \quad b(\lambda_1) \leq \sigma^2 \sqrt{n \ln(n)}\ . \qquad (\mathbf{Abias})$$

The above assumption set is discussed in details in Section 4.6.

Remark 1. *By Proposition 1, under assumption* $(\mathbf{H\Lambda})$*, assumption* $(\mathbf{HA}_\lambda)$ *holds with* $K_{\mathrm{df}} = 1$ *and* $\mathbb{M} = 1$ *for all* $\lambda \in \bigcup_j \{\, j \,\} \times [0,+\infty]$*. Furthermore, if all matrices* $(A_\lambda)_{\lambda\in\Lambda_0}$ *are among the examples of Proposition 1, assumption* $(\mathbf{HA}_\lambda)$ *holds with* $K_{\mathrm{df}} = 1$ *and* $\mathbb{M} = \sup_{\lambda\in\Lambda_0} |\!|\!| A_\lambda |\!|\!|$*, which happens to be close to 1 with large probability in all the examples we considered in our simulation experiments.*

4.3 Minimal penalty

Theorem 1. *Let* $\widehat{\lambda}_0$ *be defined by*

$$\forall C \geq 0, \qquad \widehat{\lambda}_0(C) \in \arg\min_{\lambda \in \Lambda} \left\{ \left\| \widehat{F}_\lambda - Y \right\|_2^2 + C \left(2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda) \right) \right\} \ . \tag{11}$$

Assume (**HΛ**), (**HA**$_\lambda$), (**HN**$\sigma^\mathbf{2}$) *and* (**Abias**) *hold true. Let*

$$\beta_1 = 27\mathbb{M} \quad \beta_2 = \frac{150\mathbb{M}}{K_{\mathrm{df}}} \qquad \widetilde{\alpha_\Lambda} = \max\left\{ \alpha_\Lambda^d \,,\, \alpha_\Lambda^r + 2 \right\} \quad \widetilde{C_\Lambda} = 6C_\Lambda^d + C_\Lambda^r \ .$$

Then, for every $\delta \geq 2$ *, a constant* $n_0(K_{\mathrm{df}}, \widetilde{\alpha_\Lambda} + \delta, \mathbb{M})$ *exists such that for every* $n \geq n_0$ *,*

$$\forall 0 \leq C < \left(1 - \beta_1(\widetilde{\alpha_\Lambda} + \delta)\sqrt{\frac{\ln(n)}{n}} \right) \sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \geq \frac{n}{3} \tag{12}$$

$$\textit{and} \quad \forall C > \left(1 + \beta_2(\widetilde{\alpha_\Lambda} + \delta)\sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \,, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \leq \frac{n}{10} \tag{13}$$

hold with probability at least $1 - \widetilde{C_\Lambda} n^{-\delta}$ *.*

Theorem 1 is proved in Section D.

The first important consequence of Theorem 1 is that under mild assumptions (see Section 4.6), with a large probability, the constant $\widehat{C}$ obtained by the algorithm of Section 4.1 satisfies

$$\left(1 - \beta_1(\widetilde{\alpha_\Lambda} + \delta)\sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \leq \widehat{C} \leq \left(1 + \beta_2(\widetilde{\alpha_\Lambda} + \delta)\sqrt{\frac{\ln(n)}{n}} \right) \sigma^2 \ .$$

In particular, $\widehat{C}$ is a consistent estimator of σ^2 , but Theorem 1 actually provides precise non-asymptotic *multiplicative* bounds for $\widehat{C}$ (with a convergence rate of order $\sqrt{\ln(n)/n}$), that are crucial for deriving a non-asymptotic oracle inequality for the algorithm of Section 4.1, as emphasized in Remark 8 below.

Compared to classical estimators of σ^2 , such as the one usually used with Mallows' C_L, $\widehat{C}$ does not depend on the choice of some model assumed to have almost no bias, which can lead to overestimating σ^2 by an unknown amount [17].

Remark 2. *On the same event and under the same assumptions as Theorem 1,*

$$\forall C > \left(1 + \frac{\beta_2}{10}(\widetilde{\alpha_\Lambda} + \delta)\frac{\sqrt{\ln(n)}}{n^{1/4}} \right) \sigma^2 \,, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \leq n^{3/4},$$

by taking $b_n = n^{3/4}$ *in Proposition 10. Therefore, a clearer jump can be observed by looking at* $\mathrm{df}(\widehat{\lambda}_0(C))$ *at a slightly less precise scale, which results in a possible loss in the estimation of* σ^2*.*

Remark 3. *The precise values* $n/3$ *and* $n/10$ *in Eq.* (12)–(13) *have no particular meaning*: $(n/3, n/10)$ *could be replaced by* $(n/\kappa, n/\kappa')$ *for any* $\kappa' > \kappa > 2$ *. If all matrices* A_λ *correspond to k-NN or OLS estimators, we can even take any* $\kappa' > \kappa > 1$ *, for instance,* $(9n/10, n/10)$ *.*

4.4 Oracle inequality

Define

$$\forall C \geq 0\,, \qquad \widehat{\lambda}_{\mathrm{opt}}(C) \in \arg\min_{\lambda\in\Lambda} \left\{ \left\| \widehat{F}_\lambda - Y \right\|_2^2 + 2C\,\mathrm{tr}(A_\lambda) \right\} \ . \tag{14}$$

We have the following general theorem.

Theorem 2. *Let $\widehat{\lambda}_{\mathrm{opt}}(C)$ be defined by Eq.* (14) *and assume that* ($\mathbf{H\Lambda}$), ($\mathbf{HA}_\lambda$) *and* ($\mathbf{H\mathcal{N}\sigma^2}$) *hold true. Let*

$$\beta_3 = \max\left\{ 32\mathbb{M}^2\,,\, 24\mathbb{M}^2 + 1225 \right\} \qquad \textit{and} \qquad \beta_4 = 9928\ .$$

Then, for every $\delta \geq 2$, if $n \geq n_1$ some absolute constant, with probability at least $1 - \widetilde{C_\Lambda} n^{-\delta}$, for every $C > 0$ and every $\eta \in (0,2)$,

$$\begin{aligned} n^{-1} \left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 \leq (1+\eta) \inf_{\lambda\in\Lambda} \left\{ n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 + \frac{2(C-\sigma^2)_+\,\mathrm{tr}(A_\lambda)}{n} \right\} \\ + \frac{32}{\eta} \left(C\sigma^{-2} - 1 \right)^2 \sigma^2 \mathbb{1}_{C\leq\sigma^2} + \left(\beta_3 + \frac{\beta_4}{\eta} \right) \frac{\ln(n)(\delta + \widetilde{\alpha_\Lambda})\sigma^2}{n} \end{aligned} \tag{15}$$

and

$$\begin{aligned} n^{-1} \left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 \leq (1+\eta) \inf_{\lambda\in\Lambda} \left\{ n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 \right\} + \frac{32}{\eta} \left(C\sigma^{-2} - 1 \right)^2 \sigma^2 \\ + \left(\beta_3 + \frac{\beta_4}{\eta} \right) \frac{\ln(n)(\delta + \widetilde{\alpha_\Lambda})\sigma^2}{n} \ . \end{aligned} \tag{16}$$

Theorem 2 is proved in Section E. Before applying it to the proposed algorithm, let us make a few remarks.

Remark 4. *Note that the two inequalities in Eq. (15) and Eq. (16) differ from their treatment of underestimation and overestimation of C (compared to σ^2). As shown* explicitly *in Eq. (15), overestimation of C leads to a increase of generalization cost which grows as $\frac{2\,\mathrm{tr}(A_\lambda)}{n}$, which is small, while underestimation leads to a constant factor, which is more problematic, and requires greater care when estimating C from data, which we do in this paper.*

Remark 5. *When C is deterministic, we can deduce from Eq.* (15)–(16) *an oracle inequality in expectation; we refer to the proof of Theorem 3 for details.*

Remark 6. *If $\eta = (\ln(n))^{-1}$ and $\left|C\sigma^{-2} - 1\right| = O(\sqrt{\ln(n)/n})$, the remainder term in Eq.* (15)–(16) *is of order $(\ln(n))^3\sigma^2 n^{-1}$, which is negligible in front of the risk of the oracle provided that $v_2(\lambda^\star)$ grows with n faster than $(\ln(n))^3$, since the risk of $\widehat{F}_{\lambda^\star}$ is at least of order $v_2(\lambda^\star)n^{-1}$. This usually holds when the bias is not exactly zero for some $\lambda \in \Lambda$ with $\mathrm{tr}(A_\lambda^\top A_\lambda)$ too small, as often assumed in the model selection literature for proving asymptotic optimality results.*

Remark 7. *The term $\left(C\sigma^{-2} - 1\right)^2/\theta$ could be lowered with some assumption on $\sup_{\lambda\in\Lambda} \frac{\mathrm{tr}(A_\lambda)}{\mathrm{tr}(A_\lambda^\top A_\lambda)}$. Eq.* (15)–(16) *actually correspond to the worst-case situation, where no assumption is made. For instance, if all estimators are least-squares or k-nearest neighbours estimators, this term can be replaced by $\left|C\sigma^{-2} - 1\right|$.*

Remark 8. *When the noise-level σ^2 is known, taking $C = \sigma^2$ in Theorem 2 (that is, considering Mallows' C_L penalty) yields the following oracle inequality instead of Eq.* (16)*:*

$$n^{-1}\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 \leq (1+\eta)\inf_{\lambda\in\Lambda}\left\{n^{-1}\left\|\widehat{F}_\lambda - F\right\|_2^2\right\} + \left(\beta_3 + \frac{\beta_4}{\eta}\right)\frac{\ln(n)(\delta+\widetilde{\alpha_\Lambda})\sigma^2}{n} \ .$$

As underlined in previous works [6, 7], dealing with the case of unknown variance *is more challenging, even in the model selection case. In Theorem 2 above, the remainder term $\left(C\sigma^{-2}-1\right)^2\sigma^2$ underlines how important it is to have C close to σ^2. We conjecture this term is essentially unimprovable in the case where $\mathrm{tr}(A_\lambda^\top A_\lambda)$ is close to its lower bound $n^{-1}\mathrm{tr}(A_\lambda)^2$ for λ "close" to the oracle, since the penalty $2\sigma^2 n^{-1}\mathrm{tr}(A_\lambda)$ can then be an order of magnitude higher than the risk, which is the sum of the bias and of $\sigma^2 n^{-1}\mathrm{tr}(A_\lambda^\top A_\lambda)$. Therefore, estimating σ^2 with a precision as high as the one guaranteed in Theorem 1 is crucial to derive an oracle inequality valid for all linear estimators.*

4.5 Combined result

As a corollary of Theorems 1 and 2, we get the following non-asymptotic oracle inequality (with leading constant arbitrarily close to 1) for the algorithm proposed in Section 4.1.

Theorem 3. *Let $\widehat{C}$ and $\widehat{\lambda}$ be defined as in the algorithm of Section 4.1. Assume that* (**Abias**)*,* (**HΛ**)*,* (**HA$_\lambda$**) *and* (**H$\mathcal{N}\sigma^2$**) *hold true. Then, for every $\delta \geq 2$, a numerical constant β_5 and a constant $n_2(K_{\mathrm{df}}, \delta+\widetilde{\alpha_\Lambda}, \mathbb{M})$ exist such that if $n \geq n_2$, with probability at least $1-\widetilde{C_\Lambda}n^{-\delta}$, for every $\eta \in (0,2)$,*

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{\lambda}} - F\right\|_2^2 \leq (1+\eta)\inf_{\lambda\in\Lambda}\left\{\frac{1}{n}\left\|\widehat{F}_\lambda - F\right\|_2^2\right\} + \beta_5\mathbb{M}^2\left(\widetilde{\alpha_\Lambda}+\delta\right)^2\frac{\ln(n)\sigma^2}{\eta n} \ . \tag{17}$$

As a consequence, if $n \geq n_2(K_{\mathrm{df}}, \widetilde{\alpha_\Lambda}+2, \mathbb{M})$, for every $\eta \in (0,2)$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_{\widehat{\lambda}} - F\right\|_2^2\right] \leq (1+\eta)\,\mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{\frac{1}{n}\left\|\widehat{F}_\lambda - F\right\|_2^2\right\}\right] + \beta_5\mathbb{M}^2\left(\widetilde{\alpha_\Lambda}+2\right)^2\frac{\ln(n)\sigma^2}{\eta n} \\ + \frac{2\sqrt{\widetilde{C_\Lambda}}\left(2\sigma^2\mathbb{M}^2 + (1+\mathbb{M})^2 n^{-1}\|F\|^2\right)}{n} \ .\end{aligned} \tag{18}$$

Theorem 3 is proved in Section F. Its main consequences are detailed in Section 4.7.

Note that taking $\eta = \frac{\beta_5\mathbb{M}^2(\widetilde{\alpha_\Lambda}+2)^2\ln(n)\sigma^2}{\mathbb{E}\left[\inf_{\lambda\in\Lambda}\left\{\|\widehat{F}_\lambda - F\|_2^2\right\}\right]}$ in Eq. (18) yields a non-asymptotic oracle inequality with leading constant one, that directly implies an asymptotic optimality result when $\frac{\ln(n)\sigma^2}{n}$ is negligible in front of the risk of the oracle.

4.6 Discussion of the assumptions

Assumption (HA$_\lambda$). It holds true with $K_{\mathrm{df}} = 1$ in all the main examples detailed in Section 2.2, as proved by Proposition 1. Note that $|\!|\!|A_\lambda|\!|\!| \leq \mathbb{M}$ barely is an assumption. With $\mathbb{M} = 1$, it means that A_λ actually shrinks Y, which holds true for all examples except nearest-neighbors and Nadaraya-Watson. In general, since $|\!|\!|A_\lambda|\!|\!|$ is observable, one only has to check that $\mathbb{M} := \sup_{\lambda\in\Lambda}|\!|\!|A_\lambda|\!|\!|$ is not much larger than 1, as we observed in all our simulation experiments. Note that $|\!|\!|A_\lambda|\!|\!| \leq \mathbb{M} < +\infty$ is assumed in [24] for proving asymptotic optimality results when selecting among nearest-neighbors or Nadaraya-Watson estimators.

Assumption (Abias). It holds if $b(\lambda) \leq n\sigma^2 c\,\mathrm{df}(\lambda)^{-d}$ for some $c \geq 0$ and $d \geq 1$ (and some $\lambda \in \Lambda$ exists with $\mathrm{df}(\lambda) \leq \sqrt{n}$ close to $\sqrt{n}$), a standard assumption in the context of model selection. Besides, (**Abias**) is much less restrictive and can even be relaxed, see Appendix D. For instance, it is sufficient to have $b(\lambda) \leq n\sigma^2 c\,\mathrm{df}(\lambda)^{-d}$ for one among several families of estimators, without having to know which one it is.

Gaussian noise and (HΛ). For proving Theorems 1, 2 and 3, a key ingredient is a uniform concentration inequality for four functions of λ and ε, that is, a lower bound on the probability of the event $\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}(\Lambda)$ defined in Section A.2. In particular, assumptions (**HΛ**) and (**H$\mathcal{N}\sigma^2$**) are only used for proving this event has a probability at least $1 - \widetilde{C_\Lambda} n^{-\delta}$. So, any alternative assumption set under which a similar uniform concentration inequality could be proved could be used instead of (**HΛ**) and (**H$\mathcal{N}\sigma^2$**), leading to the same results (except for the precise values of the constants). We refer to the proofs for details.

For instance, kernel ridge regression could be replaced in assumption (**HΛ**) by Pinsker filters (see Section 2.2) with the one-dimensional parameter set $w \in (0, +\infty)$.

When ε is sub-Gaussian, the key concentration results (Lemma 8) can certainly be proved for $\xi = \sigma^{-1}\varepsilon$ (possibly with additional small deviation terms that do not change the core of the proof) at the price of additional technicalities (at least when $N_\Lambda^r = 0$, i.e., Λ is finite), which implies that Theorems 1, 2 and 3 would still be valid. Note that assuming the noise is Gaussian is classical when proving non-asymptotic oracle inequalities [11, 7]; only asymptotic results exist about linear estimators selection with moment conditions on the noise [25, 24]. Considering heavy-tailed noise is of clear interest but beyond the scope of this paper.

4.7 Main consequences of Theorem 3 and comparison with previous results

Oracle inequality. The algorithm of Section 4.1 satisfies a non-asymptotic oracle inequality with high probability, as shown by Eq. (17): The risk of the selected estimator $\widehat{F}_{\widehat{\lambda}}$ is close to the risk of the oracle, up to a remainder term which is negligible when the dimensionality $\mathrm{df}(\lambda^\star)$ of the oracle grows with n faster than $\ln(n)$, a typical situation when the bias is never equal to zero, for instance in kernel ridge regression.

Eq. (17) is *non-asymptotic*, meaning that it holds for every fixed n as soon as the assumptions explicitly made in Theorem 3 are satisfied. Most results (all but the more recent ones for linear estimator selection [11, 7]) are asymptotic, meaning that n is implicitly assumed to be larged compared to each parameter of the problem. This assumption can be problematic for several learning problems, for instance in multiple kernel learning when the number p of kernels may grow with n.

Another important feature of Eq. (17) is that it holds with high probability, which is stronger than most results only true in expectation, that is, similar to Eq. (18), or even weaker. As emphasized by [25], the difference is significant, since some examples exist where a procedure (GCV) is asymptotically optimal in expectation (like Eq. (18)) but not a.s. (like Eq. (17)).

Several oracle inequalities have been proved in the statistical literature for Mallows' C_L with σ^2 known, either asymptotic [25, 24] or non-asymptotic [11]. When σ^2 is unknown (and replaced by a consistent estimator), up to the best of our knowledge, guarantees for C_L are only available for the model selection problem (see [9, 2] and references therein).

Comparison with other procedures. Oracle inequalities or asymptotic optimality results have been proved for several other linear estimator selection procedures.

Generalized Cross Validation (GCV) [14] was mostly studied for model selection and (kernel) ridge regression. GCV is asymptotically optimal in expectation under mild assumptions [14], but additional restrictions are needed for its almost sure asymptotic optimality [25]: some ridge regression example exists where GCV is not asymptotically optimal whereas C_L is [25]. Up to the best of our knowkedge, except for model selection, non-asymptotic oracle inequalities for GCV only exist for (kernel) ridge regression [11]; their result requires a prior upper bound $\mathrm{tr}(A_\lambda) \leq n/5$ for all $\lambda \in \Lambda$, and it is only valid in expectation, so it is weaker than Eq. (17). Moreover, compared to [14], our results are applicable to all linear estimators and identify key assumptions regarding the bias and variance of our collection of linear estimators (i.e., Assumption (**Abias**)).

Asymptotic optimality results also exist for GCV for k-nearest neighbors regression [24], that require the same assumption $\|A_\lambda\| \leq \mathbb{M}$ than we have.

Cross-validation methods [4] also satisfy some asymptotic optimality results in the nearest-neighbor regression case [24]. Compared to GCV or to the algorithm of Section 4.1, cross-validation methods also suffer from a large computational cost, that can only be lowered by considering V-fold cross-validation with V rather small, at the price of an increased risk. See also the simulation study of Section 5 for a numerical comparison.

Baraud, Giraud and Huet's penalization procedure [7] also satisfies a non-asymptotic oracle inequality under mild assumptions (with a leading constant $C > 1$ and a remainder term that can be large) assuming the noise is Gaussian. Compared to our minimal penalty algorithm, their procedure is more general and can deal with arbitrarily large collections Λ (putting aside computational complexity). Nevertheless, their algorithm is slightly more complex, and probably more dependent on the Gaussian assumption on the noise, since the Gaussian distribution explicitly appears in the definition of their penalty.

No overfitting. Finally, let us mention a special feature of the minimal penalty algorithm that may not appear in the comparison of theoretical results, in particular because it plays an important role only at second order and when n is small. By construction, the algorithm of Section 4.1 selects $\widehat{\lambda}$ with an effective dimensionality larger than $\widehat{\lambda}_0(\widehat{C})$ at which the jump occurs. Therefore, our algorithm *never overfits too much*, in addition to the theoretical risk bounds we have proved. This is a quite interesting property compared for instance to GCV, which is likely to overfit if it is not corrected, because GCV minimizes a criterion proportional to the empirical risk.

5 Simulations

In this section, we report simulations experiments with examples of fixed design regression to illustrate our theoretical results. For simulations on random design regression, see [3].

We consider the following example: we take $n = 200$ and use n points uniformly spaced in $[0,1]$, i.e., $x_i = (i-1)/(n-1) \in [0,1]$. We then consider $Y_i = \sin(25\pi X_i) + \varepsilon_i$ or $Y_i = \sin(25\pi X_i^3) + \varepsilon_i$, where ε_i are independent standard Gaussian random variables.

We performed experiments with kernel ridge regression with kernel $k(x,y) = \prod_{i=1}^{d} e^{-|x_i - y_i|}$ (where the goal is to learn the regularization parameter), with nearest-neighbor regression (where the goal is to learn the number of neighbors), and with locally-weighted regression (where the goal is to learn the bandwidth of the kernel).

Jump. In Figure 2 we study the size of the jump for kernel ridge regression. With half the optimal penalty (which is used in traditional variable selection for linear regression), we do not get any jump, while with the minimal penalty we always do. Note that on the left plots, we may get sharp (but of lower amplitude) jumps away from $C = \sigma^2$. In this particular situation, this is due to the periodicity of the sine function.

Comparison of estimator selection methods. In Figure 3, we plot model selection results for 20 replications of data, comparing GCV [14] and Mallows C_L penalty (which assumes the knowledge of σ^2). For the minimal penalty, we consider two strategies for finding the largest jump. We select the set of parameters so that the degrees of freedom are integers in $[0, n]$. In the first strategy, we select the value C with the largest jump, while in the second strategy we select the first C so that the selected degrees of freedom goes below $n/2$.

We compare to the oracle (which can be computed because we can enumerate Λ). We see in Figure 3 that (a) the largest jump is not a good heuristic as spurrious jumps may occur, (b) the minimal penalty technique outperforms GCV, and (c) that in some cases, Mallows (which assumes more knowledge) is outperformed by our minimal penalty strategy. Indeed, when the signal-to-noise ratio is small, overpenalizing a bit (i.e., multiplying Mallows' penalty by a factor $\kappa > 1$) is often useful, which turns out to be done automatically with the minimal penalty.

6 Conclusion

A new light on the slope heuristics. Theorems 1, 2 and 3 generalize some results first proved in [9] where all A_λ are assumed to be projection matrices. To this extent, Birgé and Massart's slope heuristics has been modified in a way that sheds a new light on the "magical" factor 2 between the minimal and the optimal penalty, as proved in [9, 2]. Indeed, Theorems 1–2 show that for general linear estimators,

$$\frac{\mathrm{pen}_{\mathrm{id}}(\lambda)}{\mathrm{pen}_{\min}(\lambda)} = \frac{2\operatorname{tr}(A_\lambda)}{2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)} \ , \tag{19}$$

which can take any value in $(1, 2]$ in general (assuming $K_{\mathrm{df}} = 1$ as in all the major examples we have in mind); this ratio is only equal to 2 when $\operatorname{tr}(A_\lambda) \approx \operatorname{tr}(A_\lambda^\top A_\lambda)$, hence mostly when A_λ is a projection matrix or a k-NN matrix.

Covariance matrix estimation. A natural extension of the present work appears in the multitask regression example, when p regression problems (the tasks) are solved simultaneously. Then, a key quantity is the $p \times p$ covariance matrix of the tasks, which has to be estimated for an optimal selection of regularization parameters (for instance). As shown in [38], the minimal penalty strategy can be used successfully for estimating a covariance matrix under mild assumptions, by applying the algorithm of Section 4.1 to $p(p+1)/2$ well-chosen one-dimensional regression problems.

Future directions. The good empirical performances of elbow heuristics based algorithms (i.e., based on the sharp variation of a certain quantity around good hyperparameter values [19, 18, 35]) suggest that Theorem 3 can be generalized to many learning frameworks (and potentially to non-linear estimators), probably with small modifications in the algorithm, but always relying on the concept of minimal penalty.

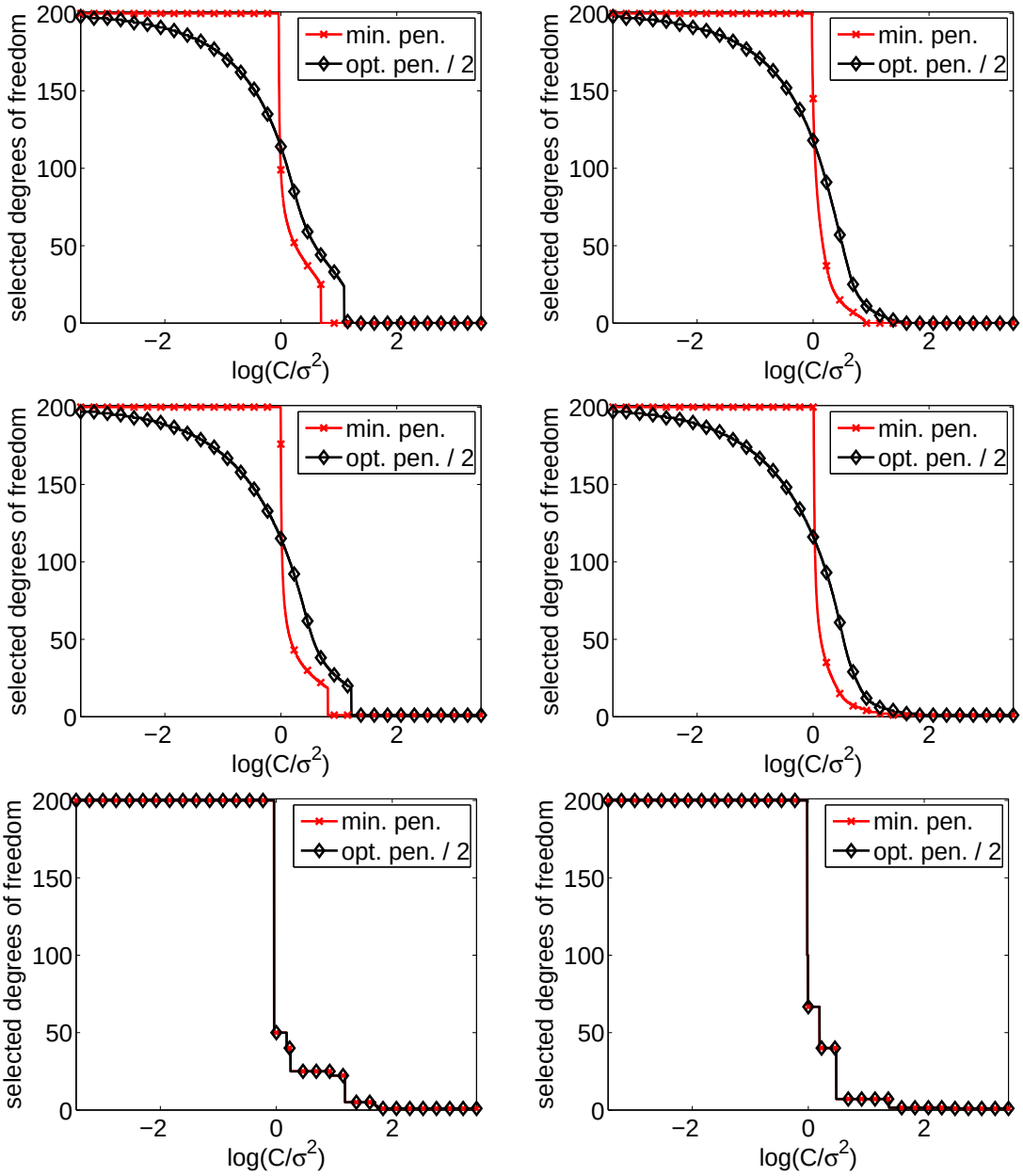

Figure 2: Selected degrees of freedom vs. penalty strength $\log(C/\sigma^2)$, for a a fixed design problem and kernel ridge regression (top), Nadaraya-Watson estimator (middle) and K-nearest neighbor regression: note that when penalizing by the minimal penalty, there is a strong jump at $C = \sigma^2$, while when using half the optimal penalty, this is not the case. Left: $Y_i = \sin(25\pi X_i) + \varepsilon_i$, Right: $Y_i = \sin(25\pi X_i^3) + \varepsilon_i$.

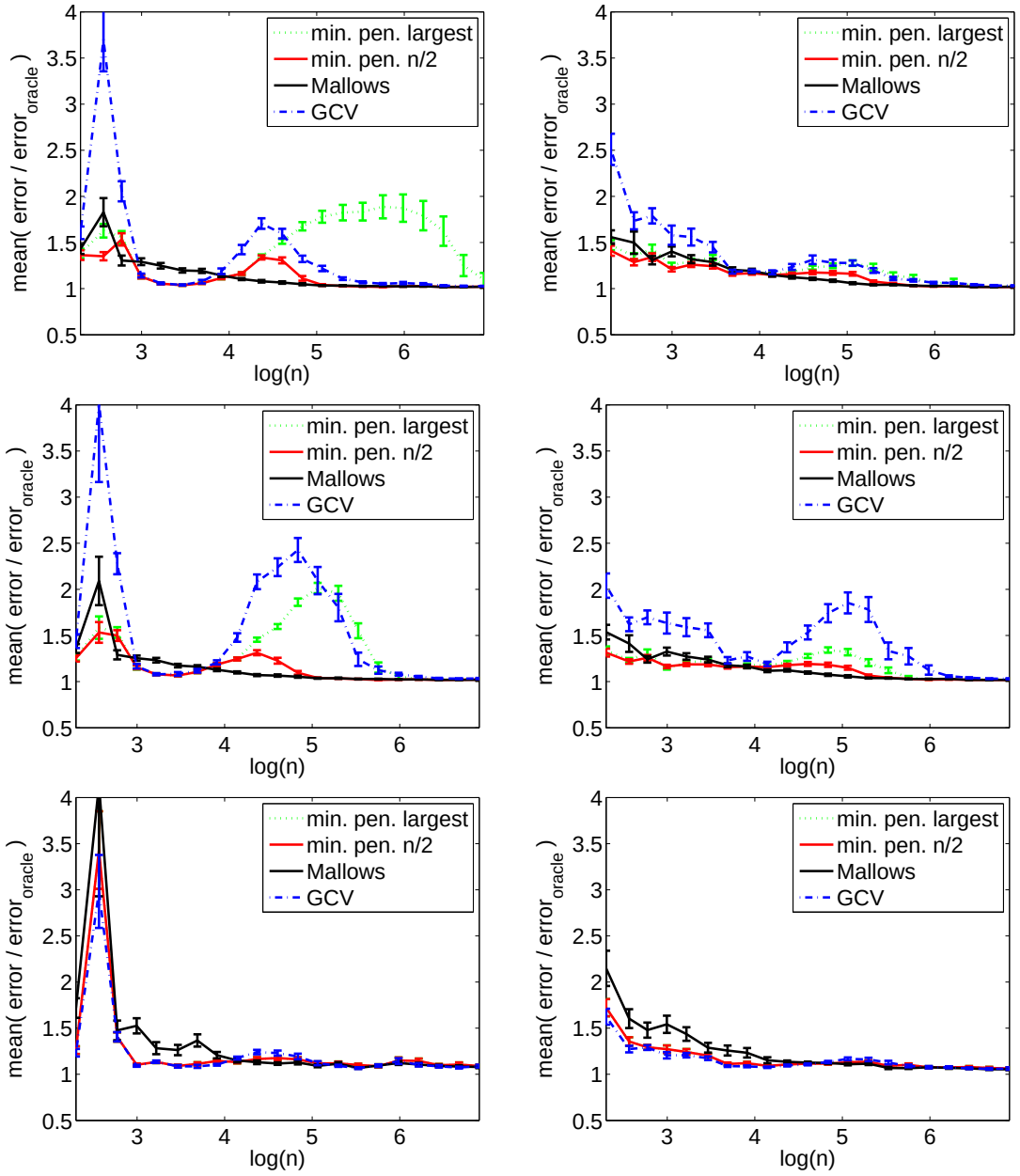

Figure 3: Comparison of various smoothing parameter selection (minimal with two types of jump selection, GCV, Mallows) for various values of numbers of observations. Left: $Y_i = \sin(25\pi X_i) + \varepsilon_i$, Right: $Y_i = \sin(25\pi X_i^3) + \varepsilon_i$.

In the case of projection estimators, the slope heuristics still holds when the design is random and data are heteroscedastic [2]; we conjecture a generalization of Eq. (19) is still valid for heteroscedastic data with some (but not all) linear estimators.

Another interesting open problem would be to extend the results of Section 4 to more general continuous sets Λ, such as the ones appearing naturally in multiple kernel learning. We conjecture that Theorem 3 is valid without modification for a "small" continuous Λ, that is, of "small" dimension. On the contrary, in applications such as the Lasso with $p \gg n$ variables, the natural set Λ cannot be well covered by a grid of cardinality n^α with α small, and our minimal penalty algorithm and Theorem 3 certainly have to be modified.

Appendix

A Notation and first computations

Recall that

$$Y = F + \varepsilon$$

where $F = (f(x_i))_{1 \leq i \leq n} \in \mathbb{R}^n$ is deterministic, $\varepsilon = (\varepsilon_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ is centered with covariance matrix $\sigma^2 I_n$ and I_n is the $n \times n$ identity matrix.

For every $x \in \mathbb{R}$, $x_+ = \max\{x, 0\}$ denotes the positive part of x.

In the proofs, we use repeatedly that

$$\forall a, b \geq 0, \forall \theta > 0, \quad 2\sqrt{ab} \leq \theta a + \theta^{-1} b \ , \tag{20}$$

with equality for $\theta = \sqrt{b/a}$.

A.1 General framework

For every $\lambda \in \Lambda$, $\widehat{F}_\lambda = A_\lambda Y$ for some $n \times n$ real-valued matrix A_λ, so that

$$\left\| \widehat{F}_\lambda - F \right\|_2^2 = \|(A_\lambda - I_n)F\|_2^2 + \|A_\lambda \varepsilon\|_2^2 + 2\langle A_\lambda \varepsilon, (A_\lambda - I_n)F \rangle \ , \tag{21}$$

$$\left\| \widehat{F}_\lambda - Y \right\|_2^2 = \left\| \widehat{F}_\lambda - F \right\|_2^2 + \|\varepsilon\|_2^2 - 2\langle \varepsilon, A_\lambda \varepsilon \rangle + 2\langle \varepsilon, (I_n - A_\lambda)F \rangle \ , \tag{22}$$

where $\forall t, u \in \mathbb{R}^n$, $\langle t, u \rangle = \sum_{i=1}^n t_i u_i$ and $\|t\|_2^2 = \langle t, t \rangle$.

Now, define, for every $\lambda \in \Lambda$,

$$\begin{aligned} b(\lambda) &= \|(A_\lambda - I_n)F\|_2^2 \qquad v_1(\lambda) = \operatorname{tr}(A_\lambda)\sigma^2 \qquad v_2(\lambda) = \operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2 \\ \delta_1(\lambda) &= \langle \varepsilon, A_\lambda \varepsilon \rangle - \operatorname{tr}(A_\lambda)\sigma^2 \qquad \delta_2(\lambda) = \|A_\lambda \varepsilon\|_2^2 - \operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2 \\ \delta_3(\lambda) &= 2\langle A_\lambda \varepsilon, (A_\lambda - I_n)F \rangle \qquad \delta_4(\lambda) = 2\langle \varepsilon, (I_n - A_\lambda)F \rangle \ , \end{aligned}$$

so that Eq. (21) and (22) can be rewritten

$$\left\| \widehat{F}_\lambda - F \right\|_2^2 = b(\lambda) + v_2(\lambda) + \delta_2(\lambda) + \delta_3(\lambda) \tag{23}$$

$$\left\| \widehat{F}_\lambda - Y \right\|_2^2 = \left\| \widehat{F}_\lambda - F \right\|_2^2 - 2v_1(\lambda) - 2\delta_1(\lambda) + \delta_4(\lambda) + \|\varepsilon\|_2^2 \ . \tag{24}$$

Note that $b(\lambda)$, $v_1(\lambda)$ and $v_2(\lambda)$ are deterministic, and for all $\lambda \in \Lambda$ and $i = 1 \dots 4$, $\delta_i(\lambda)$ is random with zero mean. In particular, we deduce the following expressions of the risk and the empirical risk of $\widehat{F}_\lambda$:

$$\mathbb{E}\left[n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 \right] = n^{-1} \left\| (A_\lambda - I_n) F \right\|_2^2 + \frac{\operatorname{tr}(A_\lambda^\top A_\lambda)\sigma^2}{n} \ , \tag{25}$$

$$\mathbb{E}\left[n^{-1} \left\| \widehat{F}_\lambda - Y \right\|_2^2 \right] - \sigma^2 = n^{-1} \left\| (A_\lambda - I_n) F \right\|_2^2 - \frac{\left(2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda) \right) \sigma^2}{n} \ . \tag{26}$$

A.2 The event Ω_x

In this section, we define the large probability event on which all our main results hold. Let $\mathcal{C}^\Omega \in [0, +\infty)^6$ be fixed. Then, for any $x \geq 0$, we define the event

$$\Omega_x = \Omega_x(\Lambda) = \Omega_x\left(\Lambda, \mathcal{C}^\Omega\right)$$

on which, for every $\lambda \in \Lambda$ and every $\theta_1, \theta_2, \theta_3, \theta_4 \in (0, 1]$,

$$|\delta_1(\lambda)| \leq \theta_1 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1} \right) x\sigma^2 \tag{27}$$

$$|\delta_2(\lambda)| \leq \theta_2 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(\mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega \theta_2^{-1} \right) x\sigma^2 \tag{28}$$

$$|\delta_3(\lambda)| \leq \theta_3 \left\| (I_n - A_\lambda) F \right\|_2^2 + \mathcal{C}_5^\Omega \theta_3^{-1} x\sigma^2 \tag{29}$$

$$|\delta_4(\lambda)| \leq \theta_4 \left\| (I_n - A_\lambda) F \right\|_2^2 + \mathcal{C}_6^\Omega \theta_4^{-1} x\sigma^2 \ . \tag{30}$$

A.3 Splitting assumption (**HΛ**)

We actually deal with assumption (**HΛ**) by considering separately the case of discrete Λ (assumption (**HΛdis**)) and the case of ridge regression with a continuous one-dimensional parameter (assumption (**Hridge**)):

- The set Λ is finite (with a polynomial size w.r.t. n):

$$\operatorname{Card}(\Lambda) \leq C_\Lambda^d n^{\alpha_\Lambda^d} \tag{**HΛdis**}$$

- In the kernel ridge regression example:

$$\left.\begin{array}{c} \Lambda = [0, +\infty] \qquad A_0 = I_n \qquad A_\infty = \mathbf{0}_{\mathcal{M}_n(\mathbb{R})} \\ \exists K \in \mathcal{M}_n(\mathbb{R}) \backslash \left\{ \mathbf{0}_{\mathcal{M}_n(\mathbb{R})} \right\} \text{ symmetric positive semi-definite} \\ \text{such that} \quad \forall \lambda \in (0, +\infty), \quad A_\lambda = K(K + n\lambda I_n)^{-1} \end{array}\right\} \tag{**Hridge**}$$

Once the concentration results will be proved under each assumption, an union bound will yield the desired results under the composite assumption (**HΛ**).

A.4 Kernel ridge regression

Let us now prove a few useful elementary results that are specific to kernel ridge regression, that is, when assumption (**Hridge**) holds true.

Under assumption (**Hridge**), a key remark is the following. Since K is symmetric positive, non-negative numbers $\mu_1, \ldots, \mu_n$ and some orthogonal matrix $P \in \mathcal{M}_n(\mathbb{R})$ exists such that $K = P^\top \operatorname{diag}(\mu_1, \ldots, \mu_n) P$. Then,

$$\forall \lambda \in \Lambda\,, \quad A_\lambda = P^\top D_\lambda P \quad \text{where} \quad D_\lambda = \operatorname{diag}\left(\left(\frac{\mu_j}{\mu_j + n\lambda} \right)_{1 \le j \le n} \right) \ , \tag{31}$$

with the convention $D_0 = I_n$ and $D_\infty = \mathbf{0}_{\mathcal{M}_n(\mathbb{R})}$. We also define

$$(f_j)_{1 \le j \le n} = PF \ .$$

Lemma 2. *If assumption* (**Hridge**) *holds true, then,*

$$\operatorname{tr}(A_\lambda) \ , \quad \operatorname{tr}(A_\lambda^\top A_\lambda) \quad \textit{and} \quad 2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda)$$

are decreasing continuous functions of λ over $[0, +\infty]$, all equal to n for $\lambda = 0$ and to 0 for $\lambda = +\infty$; $b(\lambda)$ is a nondecreasing continuous function of λ, with $b(0) = 0$ and $b(+\infty) = \|F\|^2$.

Proof of Lemma 2. According to Eq. (31), for every $\lambda \in [0, +\infty)$,

$$\begin{aligned}
\operatorname{tr}(A_\lambda) = \operatorname{df}(\lambda) &= \sum_{j=1}^n \left(\frac{\mu_j}{\mu_j + n\lambda} \right) \\
\operatorname{tr}(A_\lambda^\top A_\lambda) &= \sum_{j=1}^n \left(\frac{\mu_j}{\mu_j + n\lambda} \right)^2 \\
2\operatorname{tr}(A_\lambda) - \operatorname{tr}(A_\lambda^\top A_\lambda) &= \sum_{j=1}^n \left[\frac{2\mu_j}{\mu_j + n\lambda} - \left(\frac{\mu_j}{\mu_j + n\lambda} \right)^2 \right] = \sum_{j=1}^n \left[\frac{\mu_j(\mu_j + 2n\lambda)}{(\mu_j + n\lambda)^2} \right] \\
b(\lambda) = \|(A_\lambda - I_n)F\|_2^2 &= \sum_{j=1}^n \left(1 - \frac{\mu_j}{\mu_j + n\lambda} \right)^2 f_j^2 \ ,
\end{aligned}$$

The result follows since $\mu_j \ge 0$ for every j. □

A.5 A useful lemma

Lemma 3. *Let $n \ge 1$ an integer. Then, for any matrix $A \in \mathcal{M}_n(\mathbb{R})$,*

$$\operatorname{tr}(A) \le \sqrt{n \operatorname{tr}(A^\top A)} \ , \tag{32}$$

from which we get

$$\forall A \in \mathcal{M}_n(\mathbb{R})\,,\, x \ge 0\,,\, \theta > 0\,, \quad x \operatorname{tr}(A) \le \theta \operatorname{tr}(A^\top A) + \frac{x^2 n}{4\theta} \ . \tag{33}$$

Proof of Lemma 3. First, since $(A, B) \mapsto \operatorname{tr}(A^\top B)$ is a scalar product on $\mathcal{M}_n(\mathbb{R})$, by Cauchy-Schwarz inequality, for every $A \in \mathcal{M}_n(\mathbb{R})$,

$$(\operatorname{tr}(A))^2 = \left(\operatorname{tr}(I^\top A) \right)^2 \le \operatorname{tr}\left(A^\top A \right) \operatorname{tr}\left(I^\top I \right) = n \operatorname{tr}\left(A^\top A \right) \ .$$

Therefore, for every $x \ge 0$, $\theta > 0$,

$$x \operatorname{tr}(A) \le 2\sqrt{\operatorname{tr}(A^\top A) \frac{x^2 n}{4}} \le \theta \operatorname{tr}(A^\top A) + \frac{x^2 n}{4\theta} \qquad \text{by Eq. (20).}$$

□

B Key concentration inequalities

In this section, we state the concentration inequalities showing the event Ω_x has a large probability. Our main contributions are the following. Proposition 6 slightly extends a result by Laurent and Massart [22] for the concentration of quadratic forms of a Gaussian vector. In the kernel ridge regression example, we prove uniform concentration inequalities (over a continuous set) for some linear forms of a Gaussian vector (Proposition 5 and Lemma 13), and for some quadratic forms of a random vector (Proposition 7). All results are proved in Section G.

B.1 Linear functions of ε

In the Gaussian case, concentration inequalities for $\delta_3(\lambda)$ and $\delta_4(\lambda)$ come from the following standard result.

Proposition 4. *Let ξ be a standard Gaussian vector in $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$ and $Z = \langle \xi, \alpha \rangle = \sum_{j=1}^n \alpha_j \xi_j$. Then, for every $x \geq 0$,*

$$\mathbb{P}\left(|Z| \leq \sqrt{2x}\, \|\alpha\|_2 \right) \geq 1 - e^{-x} \ .$$

Proof of Proposition 4. The result is clear when $\alpha = 0$. Otherwise, $\langle \xi, \alpha \rangle \sim \mathcal{N}\left(0, \|\alpha\|^2\right)$ which gives the result. □

In the kernel ridge regression case, we get a similar result (with larger constants) uniformly over $\Lambda = [0, +\infty]$. Up to the best of our knowledge, Proposition 5 is a new result that could be useful for studying kernel ridge regression in a more general framework.

Proposition 5. *Assume that* (**Hridge**) *and* (**H$\mathcal{N}\sigma^2$**) *hold true. Then, for every $x \geq 0$, an event Ω_x^R of probability at least $1 - \exp(-x + 1026 + \ln(n))$ exists on which for every $\lambda \in \Lambda$,*

$$|\delta_3(\lambda)| \leq 35\sigma \, \|(I_n - A_\lambda)F\| \sqrt{x}, \tag{34}$$
$$|\delta_4(\lambda)| \leq 35\sigma \, \|(I_n - A_\lambda)F\| \sqrt{x} \ . \tag{35}$$

Proposition 5 is proved in Section G.1.

B.2 Quadratic functions of ε

In the Gaussian case, concentration inequalities for $\delta_1(\lambda)$ and $\delta_2(\lambda)$ come from the following result.

Proposition 6. *Let ξ be a standard Gaussian vector in $\mathbb{R}^n$, M a real-valued $n \times n$ matrix and $Z := \langle \xi, M\xi \rangle - \operatorname{tr}(M) = \xi^\top M \xi - \operatorname{tr}(M)$. Then, $\mathbb{E}[Z] = 0$ and for every $x \geq 0$,*

$$\mathbb{P}\left(Z \leq \sqrt{2x\left(\operatorname{tr}(M^2) + \operatorname{tr}(M^\top M)\right)} + 2\,|||M|||\, x \right) \geq 1 - \exp(-x) \ . \tag{36}$$

Since $\operatorname{tr}(M^2) \leq \operatorname{tr}(M^\top M)$ (by Lemma 15), we get that for every $x \geq 0$,

$$\mathbb{P}\left(\langle \xi, M\xi \rangle \leq \operatorname{tr}(M) + 2\sqrt{x \operatorname{tr}(M^\top M)} + 2\,|||M|||\, x \right) \geq 1 - \exp(-x) \ . \tag{37}$$

Proposition 6 extends [22, Lemma 1]; it is proved in Section G.2. The main deviation term in Eq. (36) is optimal since $\operatorname{var}(Z) = \operatorname{tr}(M^2) + \operatorname{tr}(M^\top M)$ as shown in Section H.2.

In the kernel ridge regression case, we get a similar result uniformly over $\Lambda = [0, +\infty]$. Up to the best of our knowledge, Proposition 7 is a new result that could be useful for studying kernel ridge regression in a more general framework.

Proposition 7. *Assume that* (**Hridge**) *holds true. Then, assumption* $(\mathbf{HA}_\lambda)$ *is satisfied with* $\mathbb{M} = K_{\mathrm{df}} = 1$*, and some* $\Lambda_1 \subset \Lambda$ *exists such that* $\mathrm{Card}(\Lambda_1) \leq 2n$ *and for every* $\mathcal{C}^\Omega \in [0, +\infty)^6$ *and* $x \geq 0$*, on* $\Omega_x\left(\Lambda_1, \mathcal{C}^\Omega\right)$*, for every* $\theta_1, \theta_2 \in (0, 1]$*, for every* $\lambda \in \Lambda$*,*

$$|\delta_1(\lambda)| \leq \theta_1 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1}\right) x\sigma^2 + 2\sigma^2 \tag{38}$$

$$|\delta_2(\lambda)| \leq \theta_2 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(\mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega \theta_2^{-1}\right) x\sigma^2 + 2\sigma^2 \ . \tag{39}$$

Remark 9. *Proposition 7 does not rely on any assumption on the distribution of the noise* ε*. Therefore, it can be used in the Gaussian case (combined with Lemma 8), but also under any alternative assumption for which concentration inequalities for* $\delta_1(\lambda)$ *and* $\delta_2(\lambda)$ *can be proved.*

Proposition 7 is proved in Section G.3.

B.3 Lower bounds on the probability of $\Omega_x(\Lambda)$

We are now in position to state lower bounds on the probability of $\Omega_x(\Lambda, \mathcal{C}^\Omega)$ provided $\mathcal{C}^\Omega$ is well-chosen. First, we consider the case of a finite Λ.

Lemma 8. *Under assumptions* (**HΛdis**), $(\mathbf{HA}_\lambda)$ *and* $(\mathbf{H}\mathcal{N}\sigma^\mathbf{2})$, $\mathbb{P}(\Omega_x(\Lambda, \mathcal{C}^\Omega)) \geq 1 - 6\,\mathrm{Card}(\Lambda) e^{-x}$

$$\textit{with} \quad \mathcal{C}_1^\Omega = 2\mathbb{M} \quad \mathcal{C}_2^\Omega = 1 \quad \mathcal{C}_3^\Omega = 2\mathbb{M}^2 \quad \mathcal{C}_4^\Omega = \mathbb{M}^2 \quad \mathcal{C}_5^\Omega = 2\mathbb{M}^2 \quad \textit{and} \quad \mathcal{C}_6^\Omega = 2 \ .$$

Lemma 8, a consequence of Propositions 4 and 6, is proved in Section G.4.

In the kernel ridge regression example, we prove a similar concentration result (with slightly larger constants) uniformly over the continuous set $\Lambda = [0, +\infty]$.

Lemma 9. *Under assumptions* (**Hridge**) *and* $(\mathbf{H}\mathcal{N}\sigma^\mathbf{2})$, $\mathbb{P}(\Omega_x(\Lambda, \mathcal{C}^\Omega)) \geq 1 - \exp\left(1027 + \ln(n)\right) e^{-x}$

$$\textit{if} \quad \mathcal{C}_1^\Omega = 2 \quad \mathcal{C}_2^\Omega = 1 \quad \mathcal{C}_3^\Omega = 2 \quad \mathcal{C}_4^\Omega = 1 \quad \mathcal{C}_5^\Omega = 306.25 \quad \textit{and} \quad \mathcal{C}_6^\Omega = 306.25 \ .$$

Lemma 9, a consequence of Propositions 5, 6 and 7, is proved in Section G.5.

C Proof of Proposition 1

We consider separately the examples to which Proposition 1 applies.

C.1 Case (i)

Since A is symmetric, it can be diagonalized in an orthonormal basis, with eigenvalues $a_1, \ldots, a_n \in [0, 1]$ (by assumption), so that

$$\operatorname{tr}(A^\top A) = \sum_{i=1}^n a_i^2 \leq \sum_{i=1}^n a_i = \operatorname{tr}(A) \leq n \ .$$

In particular, in example (ia), A is an orthogonal projection matrix, so A is symmetric with $\mathrm{Sp}(A) \subset [0, 1]$ and $A^\top A = A$ implies $\operatorname{tr}(A^\top A) = \operatorname{tr}(A)$. For example (ib), by Eq. (31), A is symmetric and $\mathrm{Sp}(A) \subset [0, 1]$.

C.2 Case (ii)

$$\mathrm{tr}(A^\top A) = \sum_{1\le i,j\le n} A_{i,j}^2 \le \sum_{i=1}^n \sum_{j=1}^n A_{i,j}A_{i,i} = \sum_{i=1}^n A_{i,i} = \mathrm{tr}(A) \le \sum_{i=1}^n \sum_{k=1}^n A_{i,k} = n \ .$$

In particular, if (**kNN**) holds true, then $\mathrm{tr}(A) = n/k$ and

$$\mathrm{tr}(A^\top A) = \sum_{1\le i,j\le n} A_{i,j}^2 = \frac{1}{k}\sum_{1\le i,j\le n} A_{i,j} = \frac{n}{k} \ . \quad \square$$

D Minimal penalty (proof of Theorem 1)

Let us recall the definition (11) of $\widehat{\lambda}_0(C)$:

$$\forall C \ge 0, \qquad \widehat{\lambda}_0(C) \in \arg\min_{\lambda\in\Lambda} \left\{ \left\| \widehat{F}_\lambda - Y \right\|_2^2 + C\left(2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda) \right) \right\} \ .$$

We prove in this section Theorem 1, which is actually a corollary of the following proposition:

Proposition 10. *Let $\widehat{\lambda}_0$ be defined by Eq. (11). Assume that $I \in \{ A_\lambda \}_{\lambda\in\Lambda}$, (**HA**$_\lambda$) holds true, and let $c_n \in [0,n]$ and $\beta_n \ge \sqrt{n\ln(n)}$ be such that*

$$\exists \lambda_1 \in \Lambda, \quad \mathrm{df}(\lambda_1) \le c_n \quad \textit{and} \quad b(\lambda_1) \le \sigma^2\beta_n \ . \qquad (\textbf{Abiais}')$$

Let $\mathcal{C}^\Omega \in [0,+\infty)^6$ and define

$$K_A^\Omega := 2\mathcal{C}_1^\Omega + \mathcal{C}_3^\Omega + 3\mathcal{C}_5^\Omega + 3\mathcal{C}_6^\Omega \qquad \textit{and} \qquad K_B^\Omega := 6\mathcal{C}_2^\Omega + 3\mathcal{C}_4^\Omega \ .$$

Then, for every $\gamma \ge 2$, $L_1 > 0$, a constant $n_3 > 0$ only depending on $K_A^\Omega, K_B^\Omega, \gamma, L_1$ exists such that if $n \ge n_3$, for every $a_n \in [0, \frac{n}{2})$ and $b_n \in [\frac{10}{K_{\mathrm{df}}} \max\{K_B^\Omega \gamma \ln(n), 2c_n\}, n]$

$$\forall 0 \le C < \left[1 - \left(2\sqrt{2K_B^\Omega} + \frac{1}{L_1} \right) \frac{\gamma\sqrt{n\ln(n)}}{n-2a_n} \right] \sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \ge a_n \qquad (40)$$

$$\forall C > \left[1 + \left(\frac{25}{27} + \frac{1}{L_1} + \frac{40}{9}\sqrt{K_B^\Omega} \right) \frac{\gamma\beta_n}{K_{\mathrm{df}} b_n} \right] \sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \le b_n \ . \qquad (41)$$

hold on the event $\Omega_{\gamma\ln(n)}(\Lambda, \mathcal{C}^\Omega)$.

*In particular, under assumptions (**HΛ**) and (**HNσ^2**), if $n \ge n_4(\gamma, \mathbb{M})$, with probability at least $1 - (\,6\,\mathrm{Card}(\Lambda_0) + N_\Lambda^r \exp(\,1027 + \ln(n)\,)\,)\, n^{-\gamma}$,*

$$\forall a_n \in \left[0, \frac{n}{2}\right), \forall 0 \le C < \left[1 - \frac{9\mathbb{M}\gamma\sqrt{n\ln(n)}}{n-2a_n} \right] \sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \ge a_n \qquad (42)$$

$$\forall b_n \in \left[\frac{10}{K_{\mathrm{df}}} \max\left\{ K_B^\Omega \gamma \ln(n), 2c_n \right\}, n \right], \forall C > \left[1 + \frac{15\gamma\beta_n\mathbb{M}}{K_{\mathrm{df}} b_n} \right] \sigma^2, \quad \mathrm{df}(\widehat{\lambda}_0(C)) \le b_n \ . \qquad (43)$$

Proof of Theorem 1. Taking $c_n = \sqrt{n}$ and $\beta_n = \sqrt{n\ln(n)}$, assumption (**Abiais**′) becomes (**Abias**). Let $a_n = n/3$, $b_n = n/10$ (which is possible in Proposition 10 as soon as $n/\ln(n) \ge 100K_B^\Omega\gamma$ and $n \ge (200/K_{\mathrm{df}})^2$). Then, choosing $\gamma = \widetilde{\alpha_\Lambda} + \delta$, Eq. (40) becomes Eq. (12), Eq. (41) becomes Eq. (13), and they hold with probability at least

$$
\begin{aligned}
& 1 - \left(6 \operatorname{Card}(\Lambda_0) + N_\Lambda^r \exp\left(1027 + \ln(n) \right) \right) n^{-\widetilde{\alpha_\Lambda} - \delta} \\
& \geq 1 - \left(6 C_\Lambda^d n^{\alpha_\Lambda^d} + C_\Lambda^r \exp\left(1027 + (\alpha_\Lambda^r + 1) \ln(n) \right) \right) n^{-\widetilde{\alpha_\Lambda} - \delta} \\
& = 1 - \left(6 C_\Lambda^d n^{\alpha_\Lambda^d - \widetilde{\alpha_\Lambda}} + C_\Lambda^r \exp\left(1027 + (\alpha_\Lambda^r + 1 - \widetilde{\alpha_\Lambda}) \ln(n) \right) \right) n^{-\delta} \\
& \geq 1 - \left(6 C_\Lambda^d + C_\Lambda^r \exp\left(1027 - \ln(n) \right) \right) n^{-\delta} \\
& \geq 1 - \left(6 C_\Lambda^d + C_\Lambda^r \right) n^{-\delta}
\end{aligned}
$$

as soon as $\ln(n) \geq 1027$, since $\widetilde{\alpha_\Lambda} \geq 2 + \alpha_\Lambda^r$ and $\widetilde{\alpha_\Lambda} \geq \alpha_\Lambda^d$. □

Remark 10. *On the event $\Omega_{\gamma \ln(n)}(\Lambda)$ where Eq. (40) and (41) hold and under the same assumptions, we can derive from Eq. (23), (28) with $\theta_2 = 1/2$, (29) with $\theta_3 = 1$, that*

$$
\forall \lambda \in \Lambda, \quad n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 \geq \frac{\operatorname{tr}(A_\lambda^\top A_\lambda) \sigma^2}{2n} - \frac{(\mathcal{C}_3^\Omega + 2\mathcal{C}_4^\Omega + \mathcal{C}_5^\Omega) \gamma \ln(n) \sigma^2}{n} \ .
$$

Since $\operatorname{tr}(A_\lambda^\top A_\lambda) \geq n^{-1} \left(\operatorname{df}(\lambda) \right)^2$ we deduce that for all $\lambda \in \Lambda$:

$$
\operatorname{df}(\lambda) \geq \frac{n}{\ln(n)} \ \Rightarrow \ n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 \geq \sigma^2 \left(\frac{1}{2 \left(\ln(n) \right)^2} - \frac{(\mathcal{C}_3^\Omega + 2\mathcal{C}_4^\Omega + \mathcal{C}_5^\Omega) \gamma \ln(n)}{n} \right) \ .
$$

Hence, the blow up of $\operatorname{df}(\widehat{\lambda}_0(C))$ holding when the penalty is below the minimal penalty also implies a blow up of the risk $n^{-1} \left\| \widehat{F}_{\widehat{\lambda}_0(C)} - F \right\|_2^2$.

Remark 11. *Using assumption ($\mathbf{HA}_\lambda$) and Lemma 3, we deduce that for every $\lambda \in \Lambda$,*

$$
\frac{\left(\operatorname{df}(\lambda) \right)^2}{n} \leq \operatorname{tr}(A_\lambda^\top A_\lambda) \leq 2 \operatorname{tr}(A_\lambda) = 2 \operatorname{df}(\lambda) \ .
$$

Therefore, on the event defined by Proposition 10, a jump in $\operatorname{tr}\left(A_{\widehat{\lambda}_0(C)}^\top A_{\widehat{\lambda}_0(C)} \right)$ also occurs when C goes through σ^2. Indeed,

$$
\operatorname{df}(\widehat{\lambda}_0(C)) \geq a_n \quad \textit{implies that} \quad \operatorname{tr}\left(A_{\widehat{\lambda}_0(C)}^\top A_{\widehat{\lambda}_0(C)} \right) \geq a_n^2
$$

and this lower bound becomes $n/9$ if $a_n = n/3$. Furthermore,

$$
\operatorname{df}(\widehat{\lambda}_0(C)) \leq b_n \quad \textit{implies that} \quad \operatorname{tr}\left(A_{\widehat{\lambda}_0(C)}^\top A_{\widehat{\lambda}_0(C)} \right) \leq 2 b_n
$$

which is equal to $2n^{3/4}$ when $b_n = n^{3/4}$, for instance.

Let us now prove Proposition 10. The proof is organized is as follows:

1. Section D.1 makes use of the definition of the event $\Omega_{\gamma \ln(n)}(\Lambda, \mathcal{C}^\Omega)$ for controlling uniformly over C and $\lambda \in \Lambda$ the criterion crit_C minimized by $\widehat{\lambda}_0(C)$.
2. Section D.2 considers the case $C < \sigma^2$, showing that if $A_{\lambda_2} = I$ and $\lambda \in \Lambda$ satisfies $\operatorname{df}(\lambda) \leq a_n$, then $\operatorname{crit}_C(\lambda_2) \leq \operatorname{crit}_C(\lambda)$, hence $\widehat{\lambda}_0(C) \geq a_n$.
3. Section D.3 considers the case $C > \sigma^2$, showing that if $\lambda \in \Lambda$ satisfies $\operatorname{df}(\lambda) \geq b_n$, then $\operatorname{crit}_C(\lambda_1) < \operatorname{crit}_C(\lambda)$, hence $\widehat{\lambda}_0(C) \leq b_n$.

D.1 General starting point

Combining Eq. (11) with Eq. (23) and (24), for every $C \geq 0$, $\widehat{\lambda}_0(C)$ also minimizes over $\lambda \in \Lambda$

$$\begin{aligned}\mathrm{crit}_C(\lambda) &:= \left\| \widehat{F}_\lambda - Y \right\|_2^2 - \|\varepsilon\|_2^2 + C\left(2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)\right) \\ &= b(\lambda) + (\sigma^{-2}C - 1)\,(2v_1(\lambda) - v_2(\lambda)\,) - 2\delta_1(\lambda) + \delta_2(\lambda) + \delta_3(\lambda) + \delta_4(\lambda)\ .\end{aligned}$$

On the event $\Omega_{\gamma \ln(n)}(\Lambda, \mathcal{C}^\Omega)$, using Eq. (27), (28), (29) and (30), we get for every $\lambda \in \Lambda$

$$\begin{aligned}\mathrm{crit}_C(\lambda) \geq\ & (1 - \theta_3 - \theta_4)b(\lambda) + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) - (2\theta_1 + \theta_2)v_2(\lambda) \\ & - \left(2\mathcal{C}_1^\Omega + 2\mathcal{C}_2^\Omega\theta_1^{-1} + \mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega\theta_2^{-1} + \mathcal{C}_5^\Omega\theta_3^{-1} + \mathcal{C}_6^\Omega\theta_4^{-1}\right)\gamma\ln(n)\sigma^2\ ,\end{aligned}$$

and

$$\begin{aligned}\mathrm{crit}_C(\lambda) \leq\ & (1 + \theta_3 + \theta_4)b(\lambda) + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) + (2\theta_1 + \theta_2)v_2(\lambda) \\ & + \left(2\mathcal{C}_1^\Omega + 2\mathcal{C}_2^\Omega\theta_1^{-1} + \mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega\theta_2^{-1} + \mathcal{C}_5^\Omega\theta_3^{-1} + \mathcal{C}_6^\Omega\theta_4^{-1}\right)\gamma\ln(n)\sigma^2\ .\end{aligned}$$

Taking $\theta_1 = \theta_2 = \theta/3$ for some $\theta \in (0, 3]$, and $\theta_3 = \theta_4 = 1/3$, this implies:

$$\mathrm{crit}_C(\lambda) \geq \frac{b(\lambda)}{3} + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) - \theta v_2(\lambda) - \left(K_A^\Omega + K_B^\Omega\theta^{-1}\right)\gamma\ln(n)\sigma^2 \tag{44}$$

$$\mathrm{crit}_C(\lambda) \leq \frac{5b(\lambda)}{3} + (\sigma^{-2}C - 1)(2v_1(\lambda) - v_2(\lambda)) + \theta v_2(\lambda) + \left(K_A^\Omega + K_B^\Omega\theta^{-1}\right)\gamma\ln(n)\sigma^2\ . \tag{45}$$

D.2 Below the minimal penalty

We assume in this subsection that $C \in [0, \sigma^2)$ and $a_n \in [0, n/2)$. Let $\lambda \in \Lambda$. Three cases can be distinguished:

1. If $K_B^\Omega\gamma\ln(n)/2 \leq \mathrm{df}(\lambda) \leq a_n$, then Eq. (44) yields

$$\begin{aligned}\mathrm{crit}_C(\lambda) &\geq 2(C - \sigma^2)\,\mathrm{df}(\lambda) - 2\theta\,\mathrm{df}(\lambda)\sigma^2 - \theta^{-1}K_B^\Omega\gamma\ln(n)\sigma^2 - K_A^\Omega\gamma\ln(n)\sigma^2 \\ &\geq 2(C - \sigma^2)\,\mathrm{df}(\lambda) - 2\sqrt{2K_B^\Omega\gamma\ln(n)\,\mathrm{df}(\lambda)}\sigma^2 - K_A^\Omega\gamma\ln(n)\sigma^2 \\ &\geq 2(C - \sigma^2)a_n - 2\sqrt{2K_B^\Omega\gamma\ln(n)a_n}\sigma^2 - K_A^\Omega\gamma\ln(n)\sigma^2\ ,\end{aligned} \tag{46}$$

   by taking $\theta = \sqrt{K_B^\Omega\gamma\ln(n)/(2\,\mathrm{df}(\lambda))} \leq 1$.

2. If $\mathrm{df}(\lambda) \leq K_B^\Omega\gamma\ln(n)/2$, taking $\theta = 1$ in Eq. (44) yields

$$\begin{aligned}\mathrm{crit}_C(\lambda) &\geq 2(C - \sigma^2)\,\mathrm{df}(\lambda) - 2\sigma^2\,\mathrm{df}(\lambda) - (K_A^\Omega + K_B^\Omega)\gamma\ln(n)\sigma^2 \\ &\geq -4\,\mathrm{df}(\lambda)\sigma^2 - \left(K_B^\Omega + K_A^\Omega\right)\gamma\ln(n)\sigma^2 \\ &\geq -\left(3K_B^\Omega + K_A^\Omega\right)\gamma\ln(n)\sigma^2\ .\end{aligned} \tag{47}$$

3. For $\lambda_2 \in \Lambda$ such that $A_{\lambda_2} = I$, $\mathrm{df}(\lambda_2) = n$ and $b(\lambda_2) = 0$. So, Eq. (45) yields

$$\begin{aligned}\mathrm{crit}_C(\lambda_2) &\leq (C - \sigma^2 + \theta\sigma^2)n + \left(K_A^\Omega + K_B^\Omega\theta^{-1}\right)\gamma\ln(n)\sigma^2 \\ &= (C - \sigma^2)n + 2\sqrt{nK_B^\Omega\gamma\ln(n)}\sigma^2 + K_A^\Omega\gamma\ln(n)\sigma^2\ ,\end{aligned} \tag{48}$$

   by taking $\theta = \sqrt{K_B^\Omega\gamma\ln(n)/n} \leq 1$, assuming $n/\ln(n) \geq K_B^\Omega\gamma$.

First condition on C: case 1 vs. case 3. Comparing Eq. (46) and Eq. (48), we get that

$$\mathrm{crit}_C(\lambda_2) < \inf_{\lambda\in\Lambda,\, K_B^\Omega\gamma\ln(n)/2\le \mathrm{df}(\lambda)\le a_n} \{\,\mathrm{crit}_C(\lambda)\,\} \tag{49}$$

if

$$\begin{aligned} &2(C-\sigma^2)a_n - 2\sqrt{2K_B^\Omega\gamma\ln(n)a_n}\sigma^2 - K_A^\Omega\gamma\ln(n)\sigma^2 \\ &> (C-\sigma^2)n + 2\sqrt{nK_B^\Omega\gamma\ln(n)}\sigma^2 + K_A^\Omega\gamma\ln(n)\sigma^2 \end{aligned}$$

which holds if

$$(1-\sigma^{-2}C)(n-2a_n) > 2\left(\sqrt{n}+\sqrt{2a_n}\right)\sqrt{K_B^\Omega\gamma\ln(n)} + 2K_A^\Omega\gamma\ln(n)\ .$$

The right-hand side of the above inequality is smaller than

$$\gamma\sqrt{n\ln(n)}\left[2\sqrt{2K_B^\Omega} + \frac{2K_A^\Omega}{L_0}\right]$$

where we used that $a_n < n/2$ and $\gamma \ge 2$, and we assumed that $\sqrt{n/\ln(n)} \ge L_0$ for some $L_0 > 0$ to be chosen later. Hence, Eq. (49) holds as soon as $\sqrt{n/\ln(n)} \ge L_0$ and

$$1-\sigma^{-2}C > 2\left[\sqrt{2K_B^\Omega} + \frac{K_A^\Omega}{L_0}\right]\frac{\gamma\sqrt{n\ln(n)}}{n-2a_n}\ . \tag{50}$$

Second condition on C: case 2 vs. case 3. Comparing Eq. (47) and Eq. (48), we get that

$$\mathrm{crit}_C(\lambda_2) < \inf_{\lambda\in\Lambda,\, \mathrm{df}(\lambda)\le K_B^\Omega\gamma\ln(n)/2} \{\,\mathrm{crit}_C(\lambda)\,\} \tag{51}$$

if

$$-\left(3K_B^\Omega + K_A^\Omega\right)\gamma\ln(n)\sigma^2 > (C-\sigma^2)n + 2\sqrt{nK_B^\Omega\gamma\ln(n)}\sigma^2 + K_A^\Omega\gamma\ln(n)\sigma^2$$

which holds if

$$(1-\sigma^{-2}C)n > 2\sqrt{nK_B^\Omega\gamma\ln(n)} + \left(3K_B^\Omega + 2K_A^\Omega\right)\gamma\ln(n)\ .$$

The right-hand side of the above inequality is smaller than

$$\left(\sqrt{2K_B^\Omega} + \frac{3K_B^\Omega + 2K_A^\Omega}{L_0}\right)\gamma\sqrt{n\ln(n)}$$

where we used that $\gamma \ge 2$ and we assumed that $\sqrt{n/\ln(n)} \ge L_0$. Hence, Eq. (51) holds as soon as $\sqrt{n/\ln(n)} \ge L_0$ and

$$1-\sigma^{-2}C > \left(\sqrt{2K_B^\Omega} + \frac{3K_B^\Omega + 2K_A^\Omega}{L_0}\right)\gamma\sqrt{\frac{\ln(n)}{n}}\ . \tag{52}$$

Combining the two conditions. Finally, we have proved that $\mathrm{df}(\widehat{\lambda}_0(C)) > a_n$ if $\sqrt{n/\ln(n)} \ge L_0$ and conditions (50) and (52) are both satisfied, hence if

$$1-\sigma^{-2}C > \left(2\sqrt{2K_B^\Omega} + \frac{1}{L_1}\right)\frac{\gamma\sqrt{n\ln(n)}}{n-2a_n}$$

and $\sqrt{n/\ln(n)} \ge L_0 = \left(2K_A^\Omega + 3K_B^\Omega\right)L_1$.

D.3 Above the minimal penalty

We assume in this subsection that $C > \sigma^2$. Let $\lambda \in \Lambda$. As in Section D.2, we consider three cases.

1. If $K_B^\Omega \gamma \ln(n)/2 \leq \mathrm{df}(\lambda) \leq c_n$, then, using assumption $(\mathbf{HA}_\lambda)$ in Eq. (45) yields

$$\begin{aligned}
\mathrm{crit}_C(\lambda) &\leq \frac{5b(\lambda)}{3} + (C-\sigma^2)(2\,\mathrm{df}(\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)) + \theta\,\mathrm{tr}(A_\lambda^\top A_\lambda)\sigma^2 + \left(K_A^\Omega + K_B^\Omega \theta^{-1}\right)\gamma \ln(n)\sigma^2 \\
&\leq \frac{5b(\lambda)}{3} + 2(C-\sigma^2)\,\mathrm{df}(\lambda) + 2\theta\,\mathrm{df}(\lambda)\sigma^2 + \left(K_A^\Omega + K_B^\Omega \theta^{-1}\right)\gamma \ln(n)\sigma^2 \\
&= \frac{5b(\lambda)}{3} + 2(C-\sigma^2)\,\mathrm{df}(\lambda) + 2\sqrt{2\,\mathrm{df}(\lambda) K_B^\Omega \gamma \ln(n)}\sigma^2 + K_A^\Omega \gamma \ln(n)\sigma^2 \\
&\leq \frac{5b(\lambda)}{3} + 2(C-\sigma^2)c_n + 2\sqrt{2c_n K_B^\Omega \gamma \ln(n)}\sigma^2 + K_A^\Omega \gamma \ln(n)\sigma^2 \ , \qquad (53)
\end{aligned}$$

   by taking $\theta = \sqrt{K_B^\Omega \gamma \ln(n)/(2\,\mathrm{df}(\lambda))} \leq 1$.

2. If $\mathrm{df}(\lambda) \leq K_B^\Omega \gamma \ln(n)/2$, then, using assumption $(\mathbf{HA}_\lambda)$ in Eq. (45) with $\theta = 1$ yields

$$\begin{aligned}
\mathrm{crit}_C(\lambda) &\leq \frac{5b(\lambda)}{3} + 2(C-\sigma^2)\,\mathrm{df}(\lambda) + (2\sigma^2 - C)\,\mathrm{tr}(A_\lambda^\top A_\lambda) + \left(K_A^\Omega + K_B^\Omega\right)\gamma \ln(n)\sigma^2 \\
&\leq \frac{5b(\lambda)}{3} + 2(C-\sigma^2)\,\mathrm{df}(\lambda) + 2\,\mathrm{df}(\lambda)\sigma^2 + \left(K_A^\Omega + K_B^\Omega\right)\gamma \ln(n)\sigma^2 \\
&\leq \frac{5b(\lambda)}{3} + (C-\sigma^2)K_B^\Omega \gamma \ln(n) + \left(K_A^\Omega + 2K_B^\Omega\right)\gamma \ln(n)\sigma^2 \ . \qquad (54)
\end{aligned}$$

3. If $\mathrm{df}(\lambda) \geq b_n$, then, using assumption $(\mathbf{HA}_\lambda)$ in Eq. (44) yields

$$\begin{aligned}
\mathrm{crit}_C(\lambda) &\geq (C-\sigma^2)(2\,\mathrm{df}(\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)) - \theta\,\mathrm{tr}(A_\lambda^\top A_\lambda)\sigma^2 - \left(K_A^\Omega + \theta^{-1}K_B^\Omega\right)\gamma \ln(n)\sigma^2 \\
&\geq K_{\mathrm{df}}(C-\sigma^2)\,\mathrm{df}(\lambda) - 2\theta\,\mathrm{df}(\lambda)\sigma^2 - \left(K_A^\Omega + \theta^{-1}K_B^\Omega\right)\gamma \ln(n)\sigma^2 \\
&= K_{\mathrm{df}}(C-\sigma^2)\,\mathrm{df}(\lambda) - 2\sqrt{2K_B^\Omega \gamma \ln(n)\,\mathrm{df}(\lambda)}\sigma^2 - K_A^\Omega \gamma \ln(n)\sigma^2 \\
&\geq K_{\mathrm{df}}(C-\sigma^2)b_n - 2\sqrt{2K_B^\Omega \gamma \ln(n) n}\sigma^2 - K_A^\Omega \gamma \ln(n)\sigma^2 \ , \qquad (55)
\end{aligned}$$

   by taking $\theta = \sqrt{K_B^\Omega \gamma \ln(n)/(2\,\mathrm{df}(\lambda))} \leq 1$ since $b_n \geq K_B^\Omega \gamma \ln(n)/2$, and using that $(\mathbf{HA}_\lambda)$ implies $\mathrm{df}(\lambda) \leq n$.

Eq. (55) implies

$$\inf_{\lambda \in \Lambda,\, \mathrm{df}(\lambda) \geq b_n} \{\mathrm{crit}_C(\lambda)\} \geq K_{\mathrm{df}}(C-\sigma^2)b_n - 2\sqrt{2K_B^\Omega \gamma \ln(n) n}\sigma^2 - K_A^\Omega \gamma \ln(n)\sigma^2 \ . \qquad (56)$$

Let $\lambda = \lambda_1$ given by assumption $(\mathbf{Abiais}')$. Two cases can occur:

If λ_1 matches case 1: If $c_n \geq \mathrm{df}(\lambda_1) \geq K_B^\Omega \gamma \ln(n)/2$, taking $\lambda = \lambda_1$ in Eq. (53) implies

$$\mathrm{crit}_C(\lambda_1) \leq \frac{5\sigma^2 \beta_n}{3} + 2(C-\sigma^2)c_n + 2\sqrt{2c_n K_B^\Omega \gamma \ln(n)}\sigma^2 + K_A^\Omega \gamma \ln(n)\sigma^2 \ . \qquad (57)$$

Comparing Eq. (57) and Eq. (56), we get that

$$\mathrm{crit}_C(\lambda_1) < \inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)\geq b_n} \{\,\mathrm{crit}_C(\lambda)\,\} \tag{58}$$

hence $\mathrm{df}(\widehat{\lambda}_0(C)) < b_n$ if

$$(C-\sigma^2)\,(\,K_{\mathrm{df}}b_n - 2c_n\,) > \frac{5\sigma^2\beta_n}{3} + 2\left(n^{1/2}+\sqrt{c_n}\right)\sqrt{2K_B^\Omega\gamma\ln(n)\sigma^2} + 2K_A^\Omega\gamma\ln(n)\sigma^2 \ ,$$

which holds if

$$(\sigma^{-2}C-1)\,(\,K_{\mathrm{df}}b_n - 2c_n\,) > \frac{5\beta_n}{3} + 4\sqrt{2K_B^\Omega\gamma n\ln(n)} + 2K_A^\Omega\gamma\ln(n)$$

since $c_n \leq n$. Using in addition that $\gamma \geq 2$ and $\beta_n \geq \sqrt{n\ln(n)}$, the right-hand side of the above equation is smaller than

$$\begin{aligned}\frac{5\beta_n}{3} + 4\sqrt{K_B^\Omega}\gamma\sqrt{n\ln(n)} + 2K_A^\Omega\gamma\ln(n) &\leq \frac{5\beta_n}{3} + \left(4\sqrt{K_B^\Omega} + 2K_A^\Omega L_2^{-1}\right)\gamma\sqrt{n\ln(n)} \\ &\leq \left[\frac{5}{6} + 4\sqrt{K_B^\Omega} + 2K_A^\Omega L_2^{-1}\right]\gamma\beta_n \ ,\end{aligned}$$

assuming that $\sqrt{n/\ln(n)} \geq L_2$ for some $L_2 > 0$ to be chosen later. Now, since $b_n \geq 20c_n/K_{\mathrm{df}}$, $K_{\mathrm{df}}b_n - 2c_n \geq 9K_{\mathrm{df}}b_n/10$ and Eq. (58) holds as soon as

$$\sigma^{-2}C - 1 > \frac{5}{9}\left[\frac{5}{3} + 8\sqrt{K_B^\Omega} + \frac{4K_A^\Omega}{L_2}\right]\frac{\gamma\beta_n}{K_{\mathrm{df}}b_n} \ . \tag{59}$$

If λ_1 matches case 2: If $\mathrm{df}(\lambda_1) \leq K_B^\Omega\gamma\ln(n)/2$, taking $\lambda = \lambda_1$ in Eq. (54) implies

$$\mathrm{crit}_C(\lambda_1) \leq \frac{5\beta_n\sigma^2}{3} + (C-\sigma^2)K_B^\Omega\gamma\ln(n) + \left(K_A^\Omega + 2K_B^\Omega\right)\gamma\ln(n)\sigma^2 \ . \tag{60}$$

Comparing Eq. (60) and Eq. (56), we get that

$$\mathrm{crit}_C(\lambda_1) < \inf_{\lambda\in\Lambda,\,\mathrm{df}(\lambda)\geq b_n} \{\,\mathrm{crit}_C(\lambda)\,\} \tag{61}$$

hence $\mathrm{df}(\widehat{\lambda}_0(C)) < b_n$ if

$$\begin{aligned}&\frac{5\beta_n\sigma^2}{3} + (C-\sigma^2)K_B^\Omega\gamma\ln(n) + \left(K_A^\Omega + 2K_B^\Omega\right)\gamma\ln(n)\sigma^2 \\ &< K_{\mathrm{df}}(C-\sigma^2)b_n - 2\sqrt{2K_B^\Omega\gamma\ln(n)n\sigma^2} - K_A^\Omega\gamma\ln(n)\sigma^2 \ ,\end{aligned}$$

which holds if

$$(\sigma^{-2}C-1)\left(K_{\mathrm{df}}b_n - K_B^\Omega\gamma\ln(n)\right) > \frac{5\beta_n}{3} + \left(2K_A^\Omega + 2K_B^\Omega\right)\gamma\ln(n) + 2\sqrt{2K_B^\Omega\gamma\ln(n)n} \ .$$

The right-hand side of the above equation is smaller than

$$\left[\frac{5}{6} + 2\sqrt{K_B^\Omega} + \frac{2\left(K_A^\Omega + K_B^\Omega\right)}{L_2}\right]\gamma\beta_n$$

since $\gamma \geq 2$, $\beta_n \geq \sqrt{n\ln(n)}$ and $\sqrt{n/\ln(n)} \geq L_2 > 0$. Now, $K_{\mathrm{df}}b_n - K_B^\Omega\gamma\ln(n) \geq 9K_{\mathrm{df}}b_n/10$ since $b_n \geq 10K_B^\Omega\gamma\ln(n)/K_{\mathrm{df}}$, so that Eq. (61) holds as soon as

$$\sigma^{-2}C - 1 > \left[\frac{5}{3} + 4\sqrt{K_B^\Omega} + \frac{4(K_A^\Omega + K_B^\Omega)}{L_2}\right]\frac{5}{9}\frac{\gamma\beta_n}{K_{\mathrm{df}}b_n} \ . \tag{62}$$

Combining the two conditions. Finally, we have proved that whatever the value of $\mathrm{df}(\lambda_1)$, $\mathrm{df}(\widehat{\lambda}_0(C)) < b_n$ holds if conditions (59) and (62) are both satisfied and if $\sqrt{n/\ln(n)} \geq L_2$, $b_n \geq \frac{10}{K_{\mathrm{df}}} \max\{K_B^\Omega \gamma \ln(n), 2c_n\}$. Eq. (41) follows by choosing $L_2 = 3(K_A^\Omega + K_B^\Omega)L_1$. Merging all assumptions on n made in the proof, the constant n_3 is defined by

$$\inf_{n \geq n_3} \left\{ \frac{n}{\ln(n)} \right\} \geq \max \left\{ K_B^\Omega \gamma, 16 \left(K_A^\Omega + K_B^\Omega \right)^2 L_1^2 \right\} \ .$$

D.4 Second part of Proposition 10

We apply Lemmas 8 and 9 and the union bound, taking $L_1 = 3$ and noticing that

$$\begin{aligned} K_A^\Omega &= 4\mathbb{M} + 2\mathbb{M}^2 + 3\left(\max\{2\mathbb{M}^2, \frac{35^2}{4}\} + \frac{35^2}{4} \right) \\ &\leq 6\mathbb{M}^2 + \max\{6\mathbb{M}^2, 918.75\} + 918.75 \quad \text{and} \quad K_B^\Omega = 6 + 3\mathbb{M}^2 \leq 9\mathbb{M}^2 \ . \end{aligned}$$ □

E Oracle inequality (proof of Theorem 2)

Recall the definition (14) of $\widehat{\lambda}_{\mathrm{opt}}(C)$:

$$\forall C \geq 0\,, \qquad \widehat{\lambda}_{\mathrm{opt}}(C) \in \arg\min_{\lambda \in \Lambda} \left\{ \left\| \widehat{F}_\lambda - Y \right\|_2^2 + 2C \operatorname{tr}(A_\lambda) \right\} \ .$$

We prove in this section Theorem 2, as a corollary of the following proposition.

Proposition 11. *Let $\widehat{\lambda}_{\mathrm{opt}}(C)$ be defined by Eq. (14). Assume that $(\mathbf{HA}_\lambda)$ holds true. Let $\mathcal{C}^\Omega \in [0, +\infty)^6$ and define*

$$K_C^\Omega := 4\mathcal{C}_1^\Omega \qquad K_D^\Omega := 8\left(\mathcal{C}_3^\Omega + 2\mathcal{C}_4^\Omega + 2\mathcal{C}_5^\Omega \right) \quad \text{and} \quad K_E^\Omega := 2\left(4\mathcal{C}_2^\Omega + \mathcal{C}_6^\Omega \right) \ .$$

Then, for every $\gamma \geq 2$, on the event $\Omega_{\gamma \ln(n)}(\Lambda, \mathcal{C}^\Omega)$, for every $\theta \in (0, 1/4)$,

$$\begin{aligned} n^{-1} \left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 \leq \frac{1+2\theta}{1-4\theta} \inf_{\lambda \in \Lambda} \left\{ n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 + \frac{2(C-\sigma^2)_+ \operatorname{tr}(A_\lambda)}{n} \right\} \\ + \frac{\left(C\sigma^{-2} - 1 \right)^2 \mathbb{1}_{C \leq \sigma^2} \sigma^2}{\theta(1-4\theta)} + \frac{\left(K_C^\Omega + \frac{3}{4} K_D^\Omega \theta + K_E^\Omega \theta^{-1} \right)}{1-4\theta} \frac{\ln(n)\gamma\sigma^2}{n} \ . \end{aligned} \tag{63}$$

and

$$\begin{aligned} n^{-1} \left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 \leq \frac{1+4\theta}{1-4\theta} \inf_{\lambda \in \Lambda} \left\{ n^{-1} \left\| \widehat{F}_\lambda - F \right\|_2^2 \right\} + \frac{\left(C\sigma^{-2} - 1 \right)^2 \sigma^2}{\theta(1-4\theta)} \\ + \frac{\left(K_C^\Omega + K_D^\Omega \theta + K_E^\Omega \theta^{-1} \right)}{1-4\theta} \frac{\ln(n)\gamma\sigma^2}{n} \ . \end{aligned} \tag{64}$$

Corollary 12. *Under assumptions $(\mathbf{HA}_\lambda)$, $(\mathbf{H\Lambda})$ and $(\mathbf{H}\mathcal{N}\boldsymbol{\sigma^2})$, for every $\gamma \geq 2$, with probability at least $1 - \left(6\,\mathrm{Card}(\Lambda_0) + N_\Lambda^r \exp\left(1027 + \ln(n) \right) \right) n^{-\gamma}$, for every $C > 0$, for every $\theta \in (0, 1/4)$, Eq. (63) and (64) hold true, with $(K_C^\Omega, K_D^\Omega, K_E^\Omega)$ replaced by $(\beta_{11}, \beta_{12}, \beta_{13})$ where*

$$\beta_{11} = 8\mathbb{M} \qquad \beta_{12} = 8\left(4\mathbb{M}^2 + \max\left\{ 4\mathbb{M}^2, 612.5 \right\} \right) \quad \text{and} \quad \beta_{13} = 620.5 \ .$$

We can now prove Theorem 2:

Proof of Theorem 2. The proof is similar to the one of Theorem 1, starting from Corollary 12 instead of Proposition 10. In addition, we remark that $\forall \theta \in (0, 1/8)$,

$$\frac{1+2\theta}{1-4\theta} \leq \frac{1+4\theta}{1-4\theta} \leq 1 + 16\theta$$

and we take $\theta = \eta/16$. Choosing $\gamma = \widetilde{\alpha_\Lambda} + \delta \geq 2$ in Corollary 12, Eq. (63) and (64) become Eq. (15) and (16), which holds with probability at least $1 - \left(6C_\Lambda^r + C_\Lambda^d\right) n^{-\delta}$, assuming that n is larger than some numerical constant $n_1 = \exp(1027)$. □

Let us now prove Proposition 11 and Corollary 12. The proof is organized is as follows:

1. In Section E.1, standard algebraic manipulations reduce the problem to bounding $\widehat{\Delta}(\lambda) := -2\delta_1(\lambda) + \delta_4(\lambda)$ uniformly over $\lambda \in \Lambda$.
2. Section E.2 make use of the definition of the event $\Omega_{\gamma \ln(n)}(\Lambda, \mathcal{C}^\Omega)$ for bounding $\widehat{\Delta}(\lambda)$.
3. In Section E.3, remainder terms proportional to $\operatorname{tr}(A_\lambda)$ are upper bounded in terms of $\operatorname{tr}(A_\lambda^\top A_\lambda)$, so they can be compared to the risk of $\widehat{F}_\lambda$.

E.1 General starting point

Combining Eq. (24) and (14), we obtain that for every $C > 0$ and every $\lambda \in \Lambda$,

$$\begin{aligned} &\left\| \widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F \right\|_2^2 + 2(C - \sigma^2) \operatorname{tr}(A_{\widehat{\lambda}_{\mathrm{opt}}(C)}) + \widehat{\Delta}(\widehat{\lambda}_{\mathrm{opt}}(C)) \\ &\qquad \leq \inf_{\lambda \in \Lambda} \left\{ \left\| \widehat{F}_\lambda - F \right\|_2^2 + 2(C - \sigma^2) \operatorname{tr}(A_\lambda) + \widehat{\Delta}(\lambda) \right\} \ . \end{aligned} \tag{65}$$

where

$$\forall \lambda \in \Lambda, \quad \widehat{\Delta}(\lambda) := -2\delta_1(\lambda) + \delta_4(\lambda) \ .$$

Inequality (65) implies an oracle inequality as soon as $\widehat{\Delta}(\lambda)$ is small compared to $\|\widehat{F}_\lambda - F\|_2^2$ and $C - \sigma^2$ is small enough.

E.2 With concentration inequalities

On the event $\Omega_{\gamma \ln(n)}$, using Eq. (27) and (30) with $2\theta_1 = \theta_4 = \theta \in (0, 1]$, we get that for every $\theta \in (0, 1]$ and $\lambda \in \Lambda$,

$$\left| \widehat{\Delta}(\lambda) \right| \leq \theta \left[b(\lambda) + v_2(\lambda) \right] + \left(\frac{K_C^\Omega}{2} + \frac{K_E^\Omega}{2\theta} \right) \gamma \ln(n) \sigma^2 \tag{66}$$

Using also Eq. (23), (28) and (29) with $\theta_2 = \theta_3 = 1/2$, we get that for every $\lambda \in \Lambda$,

$$\begin{aligned} \left\| \widehat{F}_\lambda - F \right\|^2 &= b(\lambda) + v_2(\lambda) + \delta_2(\lambda) + \delta_3(\lambda) \\ &\geq \frac{1}{2} \left(b(\lambda) + v_2(\lambda) \right) - \frac{K_D^\Omega}{8} \gamma \ln(n) \sigma^2 \end{aligned} \tag{67}$$

so that

$$b(\lambda) + v_2(\lambda) \leq 2 \left\| \widehat{F}_\lambda - F \right\|^2 + \frac{K_D^\Omega}{4} \gamma \ln(n) \sigma^2 \ . \tag{68}$$

Combining Eq. (66) and (68), we get

$$\left|\widehat{\Delta}(\lambda)\right| \leq 2\theta \left\|\widehat{F}_\lambda - F\right\|^2 + \left(\frac{K_C^\Omega}{2} + \frac{\theta K_D^\Omega}{4} + \frac{K_E^\Omega}{2\theta}\right)\gamma \ln(n)\sigma^2$$

so that Eq. (65) implies that

$$\begin{aligned}
&(1-2\theta)\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 + 2(C-\sigma^2)\operatorname{tr}(A_{\widehat{\lambda}_{\mathrm{opt}}(C)}) \\
&\leq \inf_{\lambda\in\Lambda}\left\{(1+2\theta)\left\|\widehat{F}_\lambda - F\right\|_2^2 + 2(C-\sigma^2)\operatorname{tr}(A_\lambda)\right\} + \left(K_C^\Omega + \frac{\theta K_D^\Omega}{2} + \frac{K_E^\Omega}{\theta}\right)\gamma\ln(n)\sigma^2 \ .
\end{aligned} \tag{69}$$

E.3 Handling the small terms proportional to $\operatorname{tr}(A_\lambda)$

We will now make use of Lemma 3 for handling the terms $(C-\sigma^2)\operatorname{tr}(A_\lambda)$ that appear in Eq. (69). By Eq. (33) with $x = 2\left|C\sigma^{-2}-1\right|\mathbb{1}_{C\leq\sigma^2}$ and Eq. (68), for every $\lambda\in\Lambda$,

$$\begin{aligned}
2\left|C-\sigma^2\right|\operatorname{tr}(A_\lambda)\mathbb{1}_{C\leq\sigma^2} &\leq \theta v_2(\lambda) + \frac{n\left(C\sigma^{-2}-1\right)^2\mathbb{1}_{C\leq\sigma^2}\sigma^2}{\theta} \\
&\leq 2\theta\left\|\widehat{F}_\lambda - F\right\|_2^2 + \frac{\theta K_D^\Omega}{4}\gamma\ln(n)\sigma^2 + \frac{n\left(C\sigma^{-2}-1\right)^2\sigma^2}{\theta}\mathbb{1}_{C\leq\sigma^2} \ .
\end{aligned} \tag{70}$$

Applying Eq. (70) with $\lambda = \widehat{\lambda}_{\mathrm{opt}}(C)$, Eq. (69) implies

$$\begin{aligned}
(1-4\theta)\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 &\leq \inf_{\lambda\in\Lambda}\left\{(1+2\theta)\left\|\widehat{F}_\lambda - F\right\|_2^2 + 2(C-\sigma^2)\operatorname{tr}(A_\lambda)\right\} \\
&+\frac{n\left(C\sigma^{-2}-1\right)^2\sigma^2}{\theta}\mathbb{1}_{C\leq\sigma^2} + \left(K_C^\Omega + \frac{3\theta K_D^\Omega}{4} + \frac{K_E^\Omega}{\theta}\right)\gamma\ln(n)\sigma^2 \ ,
\end{aligned} \tag{71}$$

which proves Eq. (63) since $\theta\in(0,1/4)$. Applying again Eq. (68) and Eq. (33) with $x = 2\left|C\sigma^{-2}-1\right|\mathbb{1}_{C\geq\sigma^2}$, Eq. (71) implies

$$\begin{aligned}
(1-4\theta)\left\|\widehat{F}_{\widehat{\lambda}_{\mathrm{opt}}(C)} - F\right\|_2^2 &\leq \inf_{\lambda\in\Lambda}\left\{(1+4\theta)\left\|\widehat{F}_\lambda - F\right\|_2^2\right\} \\
&+\frac{n\left(C\sigma^{-2}-1\right)^2\sigma^2}{\theta} + \left(K_C^\Omega + \theta K_D^\Omega + \frac{K_E^\Omega}{\theta}\right)\gamma\ln(n)\sigma^2 \ ,
\end{aligned}$$

which proves Eq. (64) since $\theta\in(0,1/4)$. □

E.4 Proof of Corollary 12

The reasoning is the same as for proving the second part of Proposition 10: we take $\mathcal{C}^\Omega$ according to Lemmas 8 and 9, so that the probability of $\Omega(\Lambda,\mathcal{C}^\Omega)$ can be lower-bounded by the union bound. □

F Proof of Theorem 3

Theorem 3 is a straightforward consequence of Theorems 1 and 2. Indeed, let us first remark the events defined by Theorems 1 and 2 are the same, namely $\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}(\Lambda,\mathcal{C}^\Omega)$ for some

well-chosen $\mathcal{C}^\Omega$. So, on $\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}(\Lambda,\mathcal{C}^\Omega)$, by Eq. (12) and (13),

$$\left|C\sigma^{-2}-1\right| \leq \max\{\beta_1\,,\,\beta_2\}\,(\widetilde{\alpha_\Lambda}+\delta)\sqrt{\frac{\ln(n)}{n}}\ .$$

Since Eq. (16) also holds on $\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}(\Lambda,\mathcal{C}^\Omega)$, we get Eq. (17) with some numerical constant

$$\beta_5 \geq \mathbb{M}^{-2}\left(32\max\{\beta_1,\beta_2\}^2+\beta_3+\frac{\beta_4}{2}\right)\ .$$

We deduce from Eq. (17) an oracle inequality in expectation by noting that if $n^{-1}\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^2 \leq R_{n,\delta}$ on $\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}$, then

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^2\right] &= \mathbb{E}\left[\frac{\mathbb{1}_{\Omega_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}}}{n}\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^2\right]+\mathbb{E}\left[\frac{\mathbb{1}_{\Omega^c_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)}}}{n}\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^2\right]\\
&\leq \mathbb{E}\left[R_{n,\delta}\right]+\frac{1}{n}\sqrt{\mathbb{P}(\Omega^c_{(\widetilde{\alpha_\Lambda}+\delta)\ln(n)})}\sqrt{\mathbb{E}\left[\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^4\right]}\\
&\leq \mathbb{E}\left[R_{n,\delta}\right]+\frac{1}{n}\sqrt{\widetilde{C_\Lambda}n^{-\delta}}\sqrt{\mathbb{E}\left[\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^4\right]} \qquad (72)
\end{aligned}$$

by Cauchy-Schwarz inequality. Now, remark that for every $\lambda\in\Lambda$,

$$\left\|\widehat{F}_\lambda-F\right\|_2^2 \leq 2\left\|A_\lambda\varepsilon\right\|^2+2\left\|(I-A_\lambda)F\right\|^2 \leq 2\mathbb{M}^2\left\|\varepsilon\right\|^2+2\left(1+\mathbb{M}\right)^2\left\|F\right\|^2$$

where we used that $|\!|\!|A_\lambda|\!|\!| \leq \mathbb{M}$ by assumption ($\mathbf{HA}_\lambda$). So,

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{F}_{\widehat{\lambda}}-F\right\|_2^4\right] &\leq \mathbb{E}\left[\sup_{\lambda\in\Lambda}\left\|\widehat{F}_\lambda-F\right\|_2^4\right]\\
&\leq 4\mathbb{E}\left[\left(\mathbb{M}^2\left\|\varepsilon\right\|^2+\left(1+\mathbb{M}\right)^2\left\|F\right\|^2\right)^2\right]\\
&= 4\left(\mathbb{M}^4\mathbb{E}\left[\left\|\varepsilon\right\|^4\right]+2\left(1+\mathbb{M}\right)^2\mathbb{M}^2\left\|F\right\|^2\mathbb{E}\left[\left\|\varepsilon\right\|^2\right]+\left(1+\mathbb{M}\right)^4\left\|F\right\|^4\right)\\
&= 4\left(\mathbb{M}^4(n^2+2n)\sigma^4+2\left(1+\mathbb{M}\right)^2\mathbb{M}^2\left\|F\right\|^2 n\sigma^2+\left(1+\mathbb{M}\right)^4\left\|F\right\|^4\right)\\
&\leq 4\left((n+1)\sigma^2\mathbb{M}^2+\left(1+\mathbb{M}\right)^2\left\|F\right\|^2\right)^2\ .
\end{aligned}$$

Using also Eq. (72) and (17), Eq. (18) follows, taking $\delta=2$. □

G Proof of the concentration inequalities

G.1 Proof of Proposition 5

The proof of Proposition 5 relies on Corollary 14, which is itself a consequence of a general concentration result (Lemma 13). Both results, which are proved at the end of the subsection, rely on a rather classical argument: a concentration inequality for the supremum of a Gaussian process, and an upper bound for its expectation in terms of entropy. Then, the proof reduces to bounding the length of a $\mathcal{C}^1$ path inside the unit Euclidean ball, which is done in Eq. (74) under the assumptions of Lemma 13.

Lemma 13. *Let $n \in \mathbb{N}\setminus\{0\}$, $Z \geq 0$. Let ξ be a standard Gaussian vector in $\mathbb{R}^n$, and $u : (0,+\infty) \mapsto \mathbb{R}^n$ be some function such that*

$$\left.\begin{array}{r} \forall t \in (0,+\infty)\,, \quad \|u(t)\|^2 = \displaystyle\sum_{j=1}^{n} u_j(t)^2 \leq 1 \qquad \forall j \in \{1,\dots,n\}\,, \quad u_j \in \mathcal{C}^1((0,+\infty)) \\ \text{and either} \quad u_j' \equiv 0 \text{ or } \operatorname{Card}\left\{ t \in (0,+\infty) \text{ s.t. } u_j'(t) = 0 \right\} \leq Z \end{array}\right\} \quad \textbf{(Hp)}$$

Then, u admits a continuous extension to $[0,+\infty]$, and for every $x \geq 0$,

$$\mathbb{P}\left(\sup_{t\in[0,+\infty]} |\langle \xi,\, u(t)\rangle| \leq \sqrt{2x} + 12\sqrt{\ln(2+4n(Z+1))} + 6\sqrt{\pi} \right) \geq 1 - e^{-x}\ . \tag{73}$$

Corollary 14. *Let $D_R, D_S \in \mathbb{N}$, and for every $j \in \{1,\dots,n\}$, let $R_j \in \mathbb{R}[X]$ be a polynomial of degree at most D_R, and $S_j \in \mathbb{R}[X]$ be a polynomial of degree at most D_S with no positive root. If the polynomials $(R_j)_{1\leq j\leq n}$ have no common positive root, then*

$$\forall t \in (0,+\infty)\,, \quad x(t) = \left(\frac{R_j(t)}{S_j(t)} \right)_{1\leq j\leq n}$$

is well-defined and non-zero, so that $u : t \to \|x(t)\|^{-1} x(t)$ is well-defined on $(0,+\infty)$. Moreover, u satisfies the assumption (**Hp**) *with $Z = (3D_R + (3n-2)D_S - 1)_+$, so that Eq.* (73) *holds true.*

We can now prove Proposition 5.

Proof of Proposition 5. Let P be an orthogonal matrix such that Eq. (31) holds true (that is, P is an orthogonal matrix which diagonalizes K), $\xi = \sigma^{-1} P\varepsilon$ and $PF = (f_j)_{1\leq j\leq n}$. Then, ξ is a standard Gaussian vector in $\mathbb{R}^n$ and for every $\lambda \in (0,+\infty)$,

$$|\delta_3(\lambda)| = 2\left|\left\langle \varepsilon,\, A_\lambda^\top (A_\lambda - I_n) F \right\rangle\right| = 2\sigma\,|\langle \xi,\, D_\lambda (D_\lambda - I_n) PF\rangle| = 2\sigma\lambda\,|\langle \xi,\, x(\lambda)\rangle|$$
$$\text{with} \qquad x(\lambda) = \lambda^{-1}\left(\frac{\mu_j}{\mu_j + n\lambda}\left(\frac{\mu_j}{\mu_j + n\lambda} - 1 \right) f_j \right)_{1\leq j\leq n} = \left(\frac{-n\mu_j f_j}{(\mu_j + n\lambda)^2} \right)_{1\leq j\leq n}\ .$$

Therefore, Corollary 14 can be applied with $D_R = 0$ and $D_S = 2$, hence $Z = 6n-5$: for every $y \geq 0$, an event of probability at least $1 - e^{-y}$ exists on which $\forall \lambda \in [0,+\infty]$,

$$\begin{aligned} |\delta_3(\lambda)| &\leq 2\sigma \left\| A_\lambda^\top (A_\lambda - I_n) F \right\| \left(\sqrt{2y} + 12\sqrt{\ln(2+4n(6n-4))} + 6\sqrt{\pi} \right) \\ &\leq 2\sigma\, \|A_\lambda\|\, \|(A_\lambda - I_n)F\| \left(\sqrt{2}\sqrt{y} + 12\sqrt{2\ln(n)} + \left(6\sqrt{\pi} + 12\sqrt{\ln(24)} \right) \right) \\ &\leq 2\sigma\, \|(A_\lambda - I_n)F\|\, \sqrt{2+288+1}\sqrt{y + \ln(n) + \left(6\sqrt{\pi} + 12\sqrt{\ln(24)} \right)^2} \\ &\leq 35\sigma\, \|(A_\lambda - I_n)F\|\, \sqrt{y + \ln(n) + 1025.8} \end{aligned}$$

by Cauchy-Schwarz inequality. So, for every $x \geq 0$, taking $y = x - \ln(n) - 1025.8$, an event of probability at least $1 - \exp(\ln(n) + 1025.8)\exp(-x)$ exists on which Eq. (34) holds true for every $\lambda \in [0,+\infty]$.

Similarly, for every $\lambda \in [0,+\infty]$,

$$|\delta_4(\lambda)| = 2\,|\langle \varepsilon,\, (I_n - A_\lambda) F\rangle| = 2\sigma\,|\langle \xi,\, (I_n - D_\lambda) PF\rangle| = 2\sigma\lambda\,|\langle \xi,\, x(\lambda)\rangle|$$
$$\text{with} \qquad x(\lambda) = \lambda^{-1}\left(\left(1 - \frac{\mu_j}{\mu_j + n\lambda} \right) f_j \right)_{1\leq j\leq n} = \left(\frac{n f_j}{\mu_j + n\lambda} \right)_{1\leq j\leq n}\ ,$$

so that Corollary 14 can be applied with $D_R = 0$ and $D_S = 1$, hence $Z = 3n - 3$: for every $y \geq 0$, an event of probability at least $1 - e^{-y}$ on which $\forall \lambda \in [0, +\infty]$,

$$\begin{aligned}
|\delta_4(\lambda)| &\leq 2\sigma \, \|(I_n - A_\lambda)F\| \left(\sqrt{2y} + 12\sqrt{\ln(2 + 4n(3n-2))} + 6\sqrt{\pi} \right) \\
&\leq 2\sigma \, \|(A_\lambda - I_n)F\| \left(\sqrt{2}\sqrt{y} + 12\sqrt{2\ln(n)} + \left(6\sqrt{\pi} + 12\sqrt{\ln(12)} \right) \right) \\
&\leq 2\sigma \, \|(A_\lambda - I_n)F\| \sqrt{2 + 288 + 1} \sqrt{y + \ln(n) + \left(6\sqrt{\pi} + 12\sqrt{\ln(12)} \right)^2} \\
&\leq 35\sigma \, \|(A_\lambda - I_n)F\| \sqrt{y + \ln(n) + 874} \ .
\end{aligned}$$

So, for every $x \geq 0$, taking $y = x - \ln(n) - 874$, an event of probability at least $1 - \exp(\ln(n) + 874)\exp(-x)$ exists on which Eq. (35) holds true for every $\lambda \in [0, +\infty]$.

The result follows by taking an union bound since

$$\exp(1025.8 + \ln(1 + \exp(874 - 1025.8))) \leq \exp(1025.8 + e^{874 - 1025.8}) \leq e^{1026} \ .$$

□

Proof of Lemma 13. Let $T = \{ u(t) \text{ s.t. } t \in (0, +\infty) \}$. Since T is a $\mathcal{C}^1$ path, its length $L(T)$ is well-defined (possibly infinite) and satisfies

$$L(T) = \int_0^{+\infty} \sqrt{\sum_{j=1}^{n} \left(u_j'^2(t) \right)} dt \leq \sum_{j=1}^{n} \int_0^{+\infty} \left| u_j'(t) \right| dt \leq 2n(Z+1) \ . \tag{74}$$

Indeed, for every $j = 1, \ldots, n$, either $u_j' \equiv 0$ so that $\int_0^{+\infty} \left| u_j'(t) \right| dt = 0$, or u_j' has at most Z zeros, so $\int_0^{+\infty} \left| u_j'(t) \right| dt$ is the sum of the amplitudes of variation of u_j over at most $(Z+1)$ intervals, each term being smaller or equal to 2 since $|u_j(t)| \leq 1$ for every $t > 0$ by assumption.

Eq. (74) also implies $u_j(t)$ has finite limits when $t \to 0$ and when $t \to +\infty$ for every j, so $u(0)$ and $u(+\infty)$ can be defined by continuity and $L(\overline{T}) = L(T)$ with $\overline{T} := \{ u(t) \text{ s.t. } t \in [0, +\infty] \}$.

For any set S, let $H(\delta, S) = \ln(N(\delta, S))$ be the metric entropy of S w.r.t. $\|\cdot\|$. Since $\overline{T}$ is a continuous path,

$$\forall \delta > 0 \, , \quad N(\delta, \overline{T} \cup (-\overline{T})) \leq 2N(\delta, \overline{T}) \leq 2 \left\lceil \frac{L(\overline{T})}{\delta} \right\rceil \leq 2 + \frac{2L(\overline{T})}{\delta} \ . \tag{75}$$

In particular, $\sqrt{H(\cdot, \overline{T} \cup (-\overline{T}))}$ is integrable at 0, so Theorem 3.18 in [28] yields

$$\mathbb{E}\left[\sup_{z \in \overline{T}} |\langle \xi, z \rangle| \right] = \mathbb{E}\left[\sup_{z \in \overline{T} \cup (-\overline{T})} \{ \langle \xi, z \rangle \} \right] \leq 12 \int_0^1 \sqrt{H(\delta, \overline{T} \cup (-\overline{T}))} d\delta \ . \tag{76}$$

Combining Eq. (74), (75) and (76), we get that

$$\begin{aligned}
\mathbb{E}\left[\sup_{z \in \overline{T}} |\langle \xi, z \rangle| \right] &\leq 12 \int_0^1 \sqrt{\ln\left(2 + \frac{4n(Z+1)}{\delta} \right)} d\delta \leq 12\sqrt{\ln(2 + 4n(Z+1))} + 12 \int_0^1 \sqrt{\ln\left(\frac{1}{\delta} \right)} d\delta \\
&\leq 12\sqrt{\ln(2 + 4n(Z+1))} + 6\sqrt{\pi}
\end{aligned}$$

since

$$\int_0^1 \sqrt{\ln\left(\frac{1}{\delta}\right)}d\delta = \int_0^{+\infty} 2x^2\exp(-x^2)dx = \int_0^{+\infty}\exp(-x^2)dx = \frac{\sqrt{\pi}}{2}\ .$$

Finally, by assumption (**Hp**), $\sup_{z\in\overline{T}}\|z\| \le 1$, so by Proposition 3.19 in [28], with probability $1-e^{-x}$,

$$\sup_{z\in\overline{T}}|\langle\xi,\,z\rangle| \le \mathbb{E}\left[\sup_{z\in\overline{T}}|\langle\xi,\,z\rangle|\right] + \sqrt{2x}\ , \tag{77}$$

hence the result. □

Proof of Corollary 14. First, $x(t)$ is well-defined for every $t>0$ since S_j has no positive root for all $j=1,\dots,n$. Second, $x(t)\neq 0$ for every $t>0$ since the R_j have no common positive root, so that u is well-defined on $(0,+\infty)$. For every $t>0$, let $N(t) := \|x(t)\| > 0$. Each coordinate x_j of x is of class $\mathcal{C}^1$ because it is a well-defined rational fraction, so N also is of class $\mathcal{C}^1$, as well as each coordinate u_j of u and for every $t\in(0,+\infty)$ and $j\in\{1,\dots,n\}$,

$$\begin{aligned}
N'(t) &= \frac{\langle x(t),\,x'(t)\rangle}{N(t)}\\
u_j'(t) &= \frac{x_j'(t)N(t) - x_j(t)N'(t)}{N(t)^2} = \frac{1}{N(t)^3}\left(x_j'(t)\,(N(t))^2 - \langle x(t),\,x'(t)\rangle\,x_j(t)\right)\\
&= \frac{1}{N(t)^3}\left(x_j'(t)\sum_{k=1}^n\left(x_k(t)^2\right) - x_j(t)\sum_{k=1}^n\left(x_k(t)x_k'(t)\right)\right)\\
&= \frac{1}{N(t)^3}\left[\frac{R_j'(t)S_j(t)-R_j(t)S_j'(t)}{S_j(t)^2}\sum_{k=1}^n\left(\frac{R_k(t)^2}{S_k(t)^2}\right) - \frac{R_j(t)}{S_j(t)}\sum_{k=1}^n\left(\frac{R_k(t)R_k'(t)S_k(t) - (R_k(t))^2S_k'(t)}{S_k(t)^3}\right)\right]\\
&= \frac{P_j(t)}{(N(t))^3\,S_j(t)\,(\prod_{k=1}^n S_k(t))^3}
\end{aligned}$$

for some polynomial $P_j\in\mathbb{R}[X]$, either equal to the null polynomial, or of degree smaller or equal to

$$Z = \max\{0,\,3D_R + (3n-2)D_S - 1\}\ ,$$

which proves (**Hp**) holds true. □

G.2 Proof of Proposition 6

The case where M is a diagonal matrix is Lemma 1 in [22]. Let us prove how the general case can be reduced to the diagonal case. Let $B = \frac{1}{2}\left(M^\top + M\right)$ so that $Z = \langle X,\,BX\rangle - \mathrm{tr}(B)$. Since B is symmetric, it can be diagonalized in an orthonormal basis:

$$\exists P\in O(n)\ \text{ s.t. }\ B = P^\top DP \quad\text{with}\quad D = \mathrm{diag}\,(d_1,\dots,d_n)\ .$$

Hence,

$$Z = X^\top BX - \mathrm{tr}(B) = (PX)^\top D(PX) - \mathrm{tr}(D) = \sum_{i=1}^n d_i\left(\xi_i^2 - 1\right)$$

where $\xi = PX\in\mathbb{R}^n$ is a standard Gaussian vector. By Lemma 1 in [22], we get that for every $x\ge 0$,

$$\mathbb{P}\left(Z \ge 2\sqrt{\mathrm{tr}(B^\top B)x} + 2\,|||B|||\,x\right) \le e^{-x}\ .$$

(It is assumed that $d_i \geq 0$ in [22], but the proof actually does not use it for proving the above inequality.) The result follows since

$$\begin{aligned}
|\!|\!| B |\!|\!| &= \left|\!\left|\!\left| \frac{1}{2}\left(M^\top + M\right) \right|\!\right|\!\right| \leq \frac{1}{2}\left(\left|\!\left|\!\left| M^\top \right|\!\right|\!\right| + |\!|\!| M |\!|\!| \right) = |\!|\!| M |\!|\!| \\
\text{and} \quad \operatorname{tr}(B^\top B) &= \frac{1}{4}\operatorname{tr}\left[\left(M^\top + M\right)\left(M^\top + M\right) \right] \\
&= \frac{1}{4}\left[\operatorname{tr}\left(\left(M^\top\right)^2 \right) + \operatorname{tr}\left(M^2\right) + 2\operatorname{tr}\left(M^\top M\right) \right] \\
&= \frac{1}{2}\left[\operatorname{tr}\left(M^2\right) + \operatorname{tr}\left(M^\top M\right) \right] \quad . \quad \square
\end{aligned}$$

G.3 Proof of Proposition 7

Proposition 1 shows that (**Hridge**) implies ($\mathbf{HA}_\lambda$) with $\mathbb{M} = K_{\mathrm{df}} = 1$. Let us now consider the concentration inequalities (38) and (39) for δ_1 and δ_2. By Lemma 2, a sequence $\lambda_1^a > \cdots > \lambda_{n-1}^a \in \Lambda$ exists such that for every $j \in \{1, \ldots, n-1\}$, $\operatorname{tr}(A_{\lambda_j^a}) = j$. Similarly, a sequence $\lambda_1^b > \cdots > \lambda_{n-1}^b \in \Lambda$ exists such that for every $j \in \{1, \ldots, n-1\}$, $\operatorname{tr}(A_{\lambda_j^b}^\top A_{\lambda_j^b}) = j$. Let

$$\Lambda_1 = \left\{ 0, \lambda_1^a, \ldots, \lambda_{n-1}^a, \lambda_1^b, \ldots, \lambda_{n-1}^b \right\} \quad ,$$

so that $\operatorname{Card}(\Lambda_1) \leq 2n$ and for every $\lambda \in \Lambda$, some $\lambda_+, \lambda_- \in \Lambda_1$ exist such that

$$\lambda_- \leq \lambda \leq \lambda_+ \tag{78}$$

$$\operatorname{tr}(A_{\lambda_-}) - 1 \leq \operatorname{tr}(A_{\lambda_+}) \leq \operatorname{tr}(A_\lambda) \leq \operatorname{tr}(A_{\lambda_-}) \leq \operatorname{tr}(A_{\lambda_+}) + 1 \tag{79}$$

$$\operatorname{tr}(A_{\lambda_-}^\top A_{\lambda_-}) - 1 \leq \operatorname{tr}(A_{\lambda_+}^\top A_{\lambda_+}) \leq \operatorname{tr}(A_\lambda^\top A_\lambda) \leq \operatorname{tr}(A_{\lambda_-}^\top A_{\lambda_-}) \leq \operatorname{tr}(A_{\lambda_+}^\top A_{\lambda_+}) + 1 \quad . \tag{80}$$

Using the notation introduced in Section A.4, let $\xi = P\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, so that

$$\|A_\lambda \varepsilon\|^2 = \sum_{j=1}^n \left(\xi_j^2 \frac{\mu_j^2}{(\mu_j + n\lambda)^2} \right) \quad \text{and} \quad \langle \varepsilon,\, A_\lambda \varepsilon \rangle = \sum_{j=1}^n \left(\xi_j^2 \frac{\mu_j}{\mu_j + n\lambda} \right)$$

both are non-increasing functions of λ since $\mu_j \geq 0$.

Let us now assume the event $\Omega_x\left(\Lambda_1, \mathcal{C}^\Omega\right)$ is realized, so that Eq. (27) and (28) hold for any $\lambda \in \Lambda_1$. For any $\lambda \in \Lambda$, let $\lambda_-, \lambda_+ \in \Lambda_1$ such that Eq. (78), (79) and (80) hold true. Then, since $\lambda \mapsto \|A_\lambda \varepsilon\|^2$ is non-increasing, for every $\theta_1 \in (0, 1]$,

$$\begin{aligned}
\delta_1(\lambda) &= \|A_\lambda \varepsilon\|^2 - \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) \\
&\geq \left\|A_{\lambda_+}\varepsilon\right\|^2 - \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) \qquad \text{by Eq. (78)} \\
&\geq \left\|A_{\lambda_+}\varepsilon\right\|^2 - \sigma^2 \operatorname{tr}(A_{\lambda_+}^\top A_{\lambda_+}) - \sigma^2 \qquad \text{by Eq. (80)} \\
&= \delta_1(\lambda_+) - \sigma^2 \\
&\geq -\theta_1 \sigma^2 \operatorname{tr}(A_{\lambda_+}^\top A_{\lambda_+}) - \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1}\right) x\sigma^2 - \sigma^2 \qquad \text{by Eq. (27)} \\
&\geq -\theta_1 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) - \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1}\right) x\sigma^2 - \sigma^2 \qquad \text{by Eq. (80)}
\end{aligned}$$

and

$$\begin{aligned}
\delta_1(\lambda) &= \|A_\lambda \varepsilon\|^2 - \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) \\
&\le \left\|A_{\lambda_-}\varepsilon\right\|^2 - \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) \qquad \text{by Eq. (78)} \\
&\le \left\|A_{\lambda_-}\varepsilon\right\|^2 - \sigma^2 \operatorname{tr}(A_{\lambda_-}^\top A_{\lambda_-}) + \sigma^2 \qquad \text{by Eq. (80)} \\
&= \delta_1(\lambda_-) + \sigma^2 \\
&\le \theta_1 \sigma^2 \operatorname{tr}(A_{\lambda_-}^\top A_{\lambda_-}) + \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1}\right) x\sigma^2 + \sigma^2 \qquad \text{by Eq. (27)} \\
&\le \theta_1 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + (1+\theta_1)\sigma^2 + \left(\mathcal{C}_1^\Omega + \mathcal{C}_2^\Omega \theta_1^{-1}\right) x\sigma^2 \qquad \text{by Eq. (80).}
\end{aligned}$$

So, for every $\theta_1 \in (0,1]$, Eq. (38) holds true.

Similarly, since $\lambda \mapsto \langle \varepsilon,\, A_\lambda \varepsilon \rangle$ is non-increasing, for every $\theta_2 \in (0,1]$,

$$\begin{aligned}
&\delta_2(\lambda) \ge -\theta_2 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) - \left(\mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega \theta_2^{-1}\right) x\sigma^2 - \sigma^2 \\
\text{and} \quad &\delta_2(\lambda) \le \theta_2 \sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + (1+\theta_2)\sigma^2 + \left(\mathcal{C}_3^\Omega + \mathcal{C}_4^\Omega \theta_2^{-1}\right) x\sigma^2
\end{aligned}$$

So, for every $\theta_2 \in (0,1]$, Eq. (39) holds true. □

G.4 Proof of Lemma 8

Since Λ is finite by assumption (**HΛdis**), we use a union bound over $\lambda \in \Lambda$. For each $\lambda \in \Lambda$, using assumption (**HNσ^2**), $\sigma^{-1}\varepsilon$ is a standard Gaussian vector.

- By Proposition 6 in Section B.2 with $M = \pm\sigma^2 A_\lambda$ and $M = \pm\sigma^2 A_\lambda^\top A_\lambda$, we deduce that for every $x \ge 0$,

$$\begin{aligned}
&\mathbb{P}\left(\forall \theta > 0,\ |\delta_1(\lambda)| \le \theta\sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(2\,|||A_\lambda||| + \theta^{-1}\right) x\sigma^2\right) \ge 1 - 2e^{-x} \\
&\mathbb{P}\left(\forall \theta > 0,\ |\delta_2(\lambda)| \le \theta\sigma^2 \operatorname{tr}(A_\lambda^\top A_\lambda) + \left(2 + \theta^{-1}\right) |||A_\lambda|||^2 x\sigma^2\right) \ge 1 - 2e^{-x}\ ,
\end{aligned}$$

  where we used Eq. (20), $\left|\left|\left|A_\lambda^\top A_\lambda\right|\right|\right| \le |||A_\lambda|||^2$, and that $\operatorname{tr}((A_\lambda^\top A_\lambda)^2) \le |||A_\lambda|||^2 \operatorname{tr}(A_\lambda^\top A_\lambda)$. Eq. (27) and (28) follow, using that $|||A_\lambda||| \le \mathbb{M}$ by assumption (**HA**$_\lambda$).

- Since $\delta_3(\lambda) = \left\langle \sigma^{-1}\varepsilon,\, 2\sigma A_\lambda^\top (I_n - A_\lambda) F \right\rangle$ and $\delta_4(\lambda) = \left\langle \sigma^{-1}\varepsilon,\, 2\sigma (I_n - A_\lambda) F \right\rangle$, Proposition 4 in Section B.1 shows, for every $x \ge 0$,

$$\begin{aligned}
&\mathbb{P}\left(\forall \theta > 0,\ |\delta_3(\lambda)| \le \theta\, \|(I_n - A_\lambda)F\|_2^2 + \frac{2\,|||A_\lambda|||^2\, x\sigma^2}{\theta}\right) \ge 1 - e^{-x} \\
\text{and} \quad &\mathbb{P}\left(\forall \theta > 0,\ |\delta_4(\lambda)| \le \theta\, \|(I_n - A_\lambda)F\|_2^2 + \frac{2x\sigma^2}{\theta}\right) \ge 1 - e^{-x}\ ,
\end{aligned}$$

  using Eq. (20). Eq. (29) and (30) follow, using that $|||A_\lambda||| \le \mathbb{M}$ by assumption (**HA**$_\lambda$).

□

G.5 Proof of Lemma 9

Since (**Hridge**) holds true, we can apply Proposition 7: some Λ_1 exists such that $\operatorname{Card}(\Lambda_1) \le 2n$ and for every $x \ge 2/3$,

$$\Omega_{x-2/3}\left(\Lambda_1, \mathcal{C}^\Omega\right) \subset \Omega_x\left(\Lambda, \mathcal{C}^\Omega\right)\ . \tag{81}$$

Since Λ_1 is finite, and assumptions $(\mathbf{H}\mathcal{N}\boldsymbol{\sigma}^{\mathbf{2}})$ and $(\mathbf{HA}_\lambda)$ hold true with $\mathbb{M} = K_{\mathrm{df}} = 1$, we get by the first part of proof of Lemma 8 that for every $x \geq 2/3$,

$$\mathbb{P}\left(\Omega_{x-2/3}\left(\Lambda_1, (2,1,2,1,+\infty,+\infty)\right)\right) \geq 1 - 6e^{-x+2/3} \ . \tag{82}$$

Now, by Proposition 5 and Eq. (20),

$$\mathbb{P}\left(\Omega_x\left(\Lambda_1, (+\infty,+\infty,+\infty,+\infty,306.25,306.25)\right)\right) \geq 1 - e^{-x+1026+\ln(n)} \ . \tag{83}$$

Combining Eq. (81), (82), (83) and an union bound, we get for every $x \geq 2/3$,

$$\mathbb{P}\left(\Omega_x\left(\Lambda_1, (2,1,2,1,306.25,306.25)\right)\right) \geq 1 - e^{-x+\ln(n)}\left(e^{1026} + 6e^{2/3}\right) \geq 1 - e^{-x+\ln(n)+1027} \ ,$$

a bound also valid when $0 \leq x < 2/3$ (since it is negative). □

Acknowledgments

We would like to thank Matthieu Solnon for helping us to improve an earlier version of the paper. We also acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01 (DETECT project) as well as the European Research Council (SIERRA starting grant 239993).

References

[1] D. M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[3] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems (NIPS)*, December 2009.

[4] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4:40–79, 2010.

[5] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[6] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[7] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Estimator selection in the gaussian setting, 2010. arXiv:1007.2096.

[8] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[9] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[10] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM J. Numer. Anal.*, 45(6):2610–2636 (electronic), 2007.

[11] Y. Cao and Y. Golubev. On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15(4):398–414 (2007), 2006.

[12] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

[13] O. Chapelle and V. Vapnik. Model selection for support vector machines. In *Advances in Neural Information Processing Systems (NIPS)*, 1999.

[14] P. Craven and G. Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4):377–403, 1978/79.

[15] Arnak S. Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators, 2011. arXiv:1104.3969.

[16] Sam Efromovich and Mark Pinsker. Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica*, 6(4):925–942, 1996.

[17] B. Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470, 1986.

[18] Oleg Grodzevich and Henry Wolkowicz. Regularization using a parameterized trust region subproblem. *Math. Program.*, 116(1-2):193–220, 2009.

[19] Per Christian Hansen and Dianne Prost O'Leary. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, 14(6):1487–1503, 1993.

[20] Anatoli Juditsky and Arkadi Nemirovski. Nonparametric denoising of signals with unknown local structure. I. Oracle inequalities. *Appl. Comput. Harmon. Anal.*, 27(2):157–179, 2009.

[21] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72 (electronic), 2003/04.

[22] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[23] É. Lebarbier. Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

[24] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987.

[25] Ker-Chau Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.*, 14(3):1101–1112, 1986.

[26] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20(7):1873–1897, 2008.

[27] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[28] P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[29] C. Maugis and B. Michel. Slope heuristics for variable selection and clustering via gaussian mixtures. Technical Report 6550, INRIA, 2008.

[30] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

[31] M. S. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Peredachi Inf.*, 16(2):52–68, 1980.

[32] M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

[33] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[34] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[35] Mansoor Rezghi and S. Mohammad Hosseini. A new variant of L-curve for Tikhonov regularization. *J. Comput. Appl. Math.*, 231(2):914–924, 2009.

[36] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

[37] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[38] Matthieu Solnon, Sylvain Arlot, and Francis Bach. Multi-task regression using minimal penalties, 2011. Work in progress.

[39] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

[40] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

[41] A. N. Tikhonov and V. A. Morozov. Methods for the regularization of ill-posed problems. *Vychisl. Metody i Programmirovanie*, (35):3–34, 1981.

[42] Minh Ngoc Tran. Penalized maximum likelihood principle for choosing ridge parameter. *Comm. Statist. Simulation and Computation*, 38(9):1610–1624, 2009.

[43] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[44] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

[45] Larry Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

[46] Geoffrey S. Watson. Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372, 1964.

[47] Shie-Shien Yang. Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.*, 76(375):658–662, 1981.

[48] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

[49] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.*, 17(9):2077–2098, 2005.

H Supplementary material

H.1 Technical lemmas

Lemma 15. *For any $M \in \mathcal{M}_n(\mathbb{R})$,*

$$\operatorname{tr}(M^2) \leq \operatorname{tr}(M^\top M) \tag{84}$$

Proof of Lemma 15.

$$\operatorname{tr}(M^2) = \sum_{i=1}^n \sum_{j=1}^n M_{i,j} M_{j,i} \leq \sum_{i=1}^n \sum_{j=1}^n \frac{M_{i,j}^2 + M_{j,i}^2}{2} = \sum_{i=1}^n \sum_{j=1}^n M_{i,j}^2 = \operatorname{tr}(M^\top M)$$

□

H.2 About the concentration of quadratic forms of Gaussian vectors

Remark 12. *By Lemma 16 below (with $m = 2$), when $\xi_1, \ldots, \xi_n$ are i.i.d. standard Gaussian variables,*

$$\operatorname{var}(\langle \xi,\, M\xi \rangle) = \operatorname{tr}(M^2) + \operatorname{tr}(M^\top M) \leq 2 \operatorname{tr}(M^\top M) \ .$$

Therefore, the deviation term $\sqrt{2x(\operatorname{tr}(M^2) + \operatorname{tr}(M^\top M))}$ cannot be improved in Eq. (36).

Lemma 16. *Let $\xi_1, \ldots, \xi_n$ be independent random variables such that for every $i \in \{1, \ldots, n\}$, $\mathbb{E}[\xi_i] = 0$, $\mathbb{E}\left[\xi_i^2\right] = 1$ and $\operatorname{var}(\xi_i^2) = \mathbb{E}\left[\xi_i^4\right] - 1 = m$. Let $M \in \mathcal{M}_n(\mathbb{R})$. Then,*

$$\operatorname{var}\left(\langle \xi,\, M\xi \rangle\right) = (m-2) \sum_{i=1}^n M_{i,i}^2 + \operatorname{tr}(M^2) + \operatorname{tr}(M^\top M) \leq \left(2 + (m-2)_+\right) \operatorname{tr}(M^\top M) \ . \tag{85}$$

Proof of Lemma 16. On the one hand,

$$\langle \xi,\, M\xi \rangle = \sum_{1 \leq i,j \leq n} M_{i,j} \xi_i \xi_j$$

so that $\mathbb{E}[\langle \xi,\, M\xi \rangle] = \operatorname{tr}(M)$. On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\langle \xi,\, M\xi \rangle^2\right] &= \sum_{1 \leq i,j,k,\ell \leq n} M_{i,j} M_{k,\ell} \mathbb{E}[\xi_i \xi_j \xi_k \xi_\ell] \\
&= \sum_{i=1}^n M_{i,i}^2 \mathbb{E}\left[\xi_i^4\right] + \sum_{1 \leq i \neq j \leq n} \left(M_{i,j} M_{j,i} + M_{i,j}^2 + M_{i,i} M_{j,j}\right) \\
&= (m-2) \sum_{i=1}^n M_{i,i}^2 + \operatorname{tr}\left(M^2\right) + \operatorname{tr}\left(M^\top M\right) + \left(\operatorname{tr}(M)\right)^2
\end{aligned}$$

so that Eq. (85) holds true (using Lemma 15). □