A Machine Learning Approach to Optimal Tikhonov Regularization I: Affine Manifolds

Ernesto De Vito (*corresponding author*) DEVITO@DIMA.UNIGE.IT
DIMA, Università di Genova, Via Dodecaneso 35, Genova, Italy

Massimo Fornasier MASSIMO.FORNASIER@MA.TUM.DE
Technische Universität München, Fakultät Mathematik, Boltzmannstrasse 3 D-85748, Garching bei München, Germany

Valeriya Naumova VALERIYA@SIMULA.NO
Simula Research Laboratory, Martin Linges vei 25, Fornebu, Norway

Abstract

Despite a variety of available techniques the issue of the proper regularization parameter choice for inverse problems still remains one of the biggest challenges. The main difficulty lies in constructing a rule, allowing to compute the parameter from given noisy data without relying either on a priori knowledge of the solution or on the noise level. In this paper we propose a novel method based on supervised machine learning to approximate the high-dimensional function, mapping noisy data into a good approximation to the optimal Tikhonov regularization parameter. Our assumptions are that solutions of the inverse problem are statistically distributed in a concentrated manner on (lower-dimensional) linear subspaces and the noise is sub-gaussian. One of the surprising facts is that the number of previously observed examples for the supervised learning of the optimal parameter mapping scales at most linearly with the dimension of the solution subspace. We also provide explicit error bounds on the accuracy of the approximated parameter and the corresponding regularization solution. Even though the results are more of theoretical nature, we present a recipe for the practical implementation of the approach and provide numerical experiments confirming the theoretical results. We also outline interesting directions for future research with some preliminary results, confirming their feasibility.

Keywords: Tikhonov regularization, parameter choice rule, sub-gaussian vectors, high dimensional function approximations, concentration inequalities.

1. Introduction

In many practical problems, one cannot observe directly the quantities of most interest; instead their values have to be inferred from their effects on observable quantities. When this relationship between observable Y and the quantity of interest X is (approximately) linear, as it is in surprisingly many cases, the situation can be modeled mathematically by the equation

$$Y = AX \tag{1}$$

for A being a linear operator model. If A is a "nice", easily invertible operator, and if the data Y are noiseless and complete, then finding X is a trivial task. Often, however, the mapping A is ill-conditioned or not invertible. Moreover, typically (1) is only an idealized

version, which completely neglects any presence of noise or disturbances; a more accurate model is

$$Y = AX + \eta, \tag{2}$$

in which the data are corrupted by an (unknown) noise. In order to deal with this type of reconstruction problem a regularization mechanism is required (Engl et al., 1996).

Regularization techniques attempt to incorporate as much as possible an (often vague) a priori knowledge on the nature of the solution X. A well-known assumption which is often used to regularize inverse problems is that the solution belongs to some ball of a suitable Banach space.

Regularization theory has shown to play its major role for solving infinite dimensional inverse problems. In this paper, however, we consider finite dimensional problems, since we intend to use probabilistic techniques for which the Euclidean space is the most standard setting. Accordingly, we assume the solution vector $X \in \mathbb{R}^d$, the linear model $A \in \mathbb{R}^{m\times d}$, and the datum $Y \in \mathbb{R}^m$. In the following we denote with $\|Z\|$ the Euclidean norm of a vector $Z \in \mathbb{R}^N$. One of the most widely used regularization approaches is realized by minimizing the following, so-called, Tikhonov functional

$$\min_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha \|z\|^2. \tag{3}$$

with $\alpha \in (0, +\infty)$. The *regularized solution* $Z^\alpha := Z^\alpha(Y)$ of such minimization procedure is unique. In this context, the regularization scheme represents a trade-off between the accuracy of fitting the data Y and the complexity of the solution, measured by a ball in $\mathbb{R}^d$ with radius depending on the *regularization parameter* α. Therefore, the choice of the regularization parameter α is very crucial to identify the best possible regularized solution, which does not overfit the noise. This issue still remains one of the most delicate aspects of this approach and other regularization schemes. Clearly the best possible parameter minimizes the discrepancy between Z^α and the solution X

$$\alpha^* = \arg \min_{\alpha\in(0,+\infty)} \|Z^\alpha - X\|.$$

Unfortunately, we usually have neither access to the solution X nor to information about the noise, for instance, we might not be aware of the noise level $\|\eta\|$. Hence, for determining a possible good approximation to the optimal regularization parameter several approaches have been proposed, which can be categorized into three classes

- A priori parameter choice rules based on the noise level and some known "smoothness" of the solution encoded in terms, e.g., of the so-called *source condition* (Engl et al., 1996);
- A posteriori parameter choice rules based on the datum Y and the noise level;
- A posteriori parameter choice rules based exclusively on the datum Y or, the so-called, heuristic parameter choice rules.

For the latter two categories there are by now a multitude of approaches. Below we recall the most used and relevant of them, indicating in square brackets their alternative names,

accepted in different scientific communities. In most cases, the names we provide are the descriptive names originally given to the methods. However, in a few cases, there was no original name, and, to achieve consistency in the naming, we have chosen an appropriate one, reflecting the nature of the method. We mention, for instance, (transformed/modified) discrepancy principle [Raus-Gfrerer rule, minimum bound method]; monotone error rule; (fast/hardened) balancing principle also for white noise; quasi-optimality criterion; L-curve method; modified discrepancy partner rule [Hanke-Raus rule]; extrapolated error method; normalized cumulative periodogram method; residual method; generalized maximum likelihood; (robust/strong robust/modified) generalized cross-validation. Considering the large number of available parameter choice methods, there are relatively few comparative studies and we refer to (Bauer and Lukas, 2011) for a rather comprehensive discussion on their differences, pros and contra. One of the features which is common to most of the a posteriori parameter choice rules is the need of solving (3) multiple times for different values of the parameters α, often selected out of a conveniently pre-defined grid.

In this paper, we intend to study a novel, fully data-driven, method for the determination of the optimal parameter in Tikhonov regularization. After an off-line learning phase, whose complexity scales at most algebraically with the dimensionality of the problem, our method does not require any additional knowledge of the noise level and the computation of a near-optimal regularization parameter can be performed very efficiently without the need of solving the regularization problem (3) multiple times. The approach aims at employing the framework of supervised machine learning to the problem of approximating the high-dimensional function, which maps noisy data into the corresponding optimal regularization parameter. More precisely, we assume that we are allowed to see a certain number n of examples of solutions X_i and corresponding noisy data $Y_i = AX_i + \eta_i$, for $i = 1, \ldots, n$. For all of these examples, we are clearly capable to compute the optimal regularization parameters as in the following scheme

$$\begin{aligned}
(X_1, Y_1) &\rightarrow \alpha_1^* = \arg \min_{\alpha \in (0,+\infty)} \|Z^\alpha(Y_1) - X_1\| \\
(X_2, Y_2) &\rightarrow \alpha_2^* = \arg \min_{\alpha \in (0,+\infty)} \|Z^\alpha(Y_2) - X_2\| \\
\ldots &\quad \ldots \\
(X_n, Y_n) &\rightarrow \alpha_n^* = \arg \min_{\alpha \in (0,+\infty)} \|Z^\alpha(Y_n) - X_n\| \\
(??, Y) &\rightarrow \bar{\alpha}
\end{aligned}$$

Denote μ the joint distribution of the empirical samples $(Y_1, \alpha_1^*), \ldots, (Y_n, \alpha_n^*)$. Were its conditional distribution $\mu(\cdot \mid Y)$ with respect to the first variable Y very much concentrated (for instance, when $\int_0^\infty (\alpha - \bar{\alpha})^q d\mu(\alpha \mid Y)$ is very small for $q \geq 1$ and for variable Y), then we could design a proper regression function

$$\mathcal{R} : Y \mapsto \bar{\alpha} := \mathcal{R}(Y) = \int_0^\infty \alpha d\mu(\alpha \mid Y).$$

Such a mapping would allow us, to a given new datum Y (without given solution!), to associate the corresponding parameter $\bar{\alpha}$ not too far from the true optimal one α^*, at least with high probability. We illustrate schematically this theoretical framework in Figure 1.

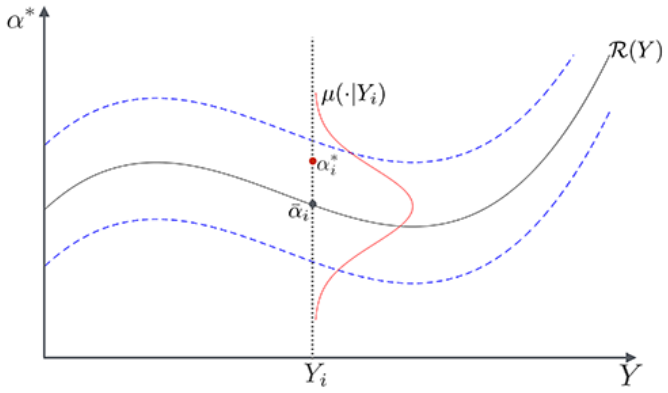

Figure 1: Learning optimal regularization parameters from previously observed samples by approximation of the regression function $\mathcal{R}$.

At a first glance, this setting may seem quite hopeless. First of all, one should establish the concentration of the conditional distribution generating α^* given Y. Secondly, even if we assume that the regression function $\mathcal{R}$ is very smooth, the vectors Y belong to the space $\mathbb{R}^m$ and the number of observations n required to learn such a function need to scale exponentially with the dimension m (Novak and Woźniakowski, 2009). It is clear that we cannot address neither of the above issues in general. The only hope is that the solutions are statistically distributed in a concentrated manner over smooth sets of lower dimension $h \ll m$ and the noise has also a concentrated distribution, so that the corresponding data Y are concentrated around lower-dimensional sets as well. And luckily these two assumptions are to a certain extent realistic.

By now, the assumption that the possible solutions belong to a lower-dimensional set of $\mathbb{R}^d$ has become an important prior for many signal and image processing tasks. For instance, were solutions natural images, then it is known that images can be represented as nearly-sparse coefficient vectors with respect to shearlets expansions (Kutyniok and Labate, 2012). Hence, in this case the set of possible solutions can be stylized as a union of lower-dimensional linear subspaces, consisting of sparse vectors. In other situations, it is known that the solution set can be stylized, at least locally, as a smooth lower-dimensional nonlinear manifold $\mathcal{V}$ (Chen et al., 2013). Also in this case, at least locally, it is possible to approximate the solution set by means of affine lower-dimensional sets, representing tangent spaces to the manifold. Hence, the a priori knowledge that the solution is belonging to some special (often nonlinear) set should also be taken into account when designing the regularization method.

In this paper, we want to show very rigorously how one can construct, from a relatively small number of previously observed examples, an approximation $\widehat{\mathcal{R}}$ to the regression function $\mathcal{R}$, which is mapping a noisy datum into a good approximation to the optimal Tikhonov regularization parameter. To this end, we assume the solutions to be distributed sub-gaussianly over a *linear subspace* $\mathcal{V} \subset \mathbb{R}^d$ of dimension $h \ll m$ and the noise η to be

also sub-gaussian. The first statistical assumption is perhaps mostly technical to allow us to provide rigorous estimates. Let us describe the method of computation as follows. We introduce the $d \times d$ noncentered covariance matrix built from the noisy measurements

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \otimes Y_i,$$

and we denote by $\widehat{\Pi}_n$ the projections onto the vector space spanned by the first most relevant eigenvectors of $\widehat{\Sigma}_n$. Furthermore, we set $\widehat{\alpha}_n \in (0, +\infty)$ as the minimizer of

$$\min_{\alpha\in(0,+\infty)} \|Z^\alpha - A^\dagger \widehat{\Pi}_n Y\|^2,$$

where $A^\dagger$ is the pseudo-inverse. We define

$$\widehat{\mathcal{R}}(Y) = \widehat{\alpha}_n$$

and we claim that this is actually a good approximation, up to noise level, to $\mathcal{R}$ as soon as n is large enough, without incurring in the curse of dimensionality. More precisely, we prove that, for a given $\tau > 0$, with probability greater than $1 - 6\mathrm{e}^{-\tau^2}$, we have that

$$\|Z^{\widehat{\alpha}_n} - X\| \leq \|Z^{\alpha^*} - X\| + \frac{1}{\sigma_d} B(n, \tau, \sigma),$$

where σ_d is the smallest singular value of A. Let us stress that $B(n, \tau, \sigma)$ gets actually small for small σ and for $n = \mathcal{O}(h)$ (see formula (29)), hence there is no curse of dimensionality and $\widehat{\mathcal{R}}(Y) = \widehat{\alpha}_n$ is quasi-optimal. We further provide an explicit expression for B in Proposition 5. In the special case where $A = I$ we derive a bound on the difference between the learned parameter $\widehat{\alpha}_n$ and the optimal parameter α^*, see Theorem 11, justifying even more precisely the approximation $\widehat{\mathcal{R}}(Y) = \widehat{\alpha}_n \approx \alpha^* = \mathcal{R}(Y)$.

The paper is organized as follows: After introducing some notation and problem set-up in the next section, we provide the accuracy bounds on the learned estimators with respect to their distribution dependent counterparts in Section 3. For the special case $A = I$ we provide an explicit bound on the difference between the learned and the optimal regularization parameter and discuss the amount of samples needed for an accurate learning in Section 4. We also exemplify the presented theoretical results with a few numerical illustrations. Section 5 provides explicit formulas by means of numerical linearization for the parameter learning. Section 6 offers a snapshot of the main contributions and presents a list of open questions for future work. Finally, Appendix A and Appendix B contain some background information on perturbation theory for compact operators, the sub-gaussian random variables, and proofs of some technical theorems, which are valuable for understanding the scope of the paper.

2. Setting

This section presents some background material and sets the notation for the rest of the work. First, we fix some notation. The Euclidean norm of a vector v is denoted by $\|v\|$ and the Euclidean scalar product between two vectors v, w by $\langle v, w\rangle$. We denote with S^{d-1}

the Euclidean unit sphere in $\mathbb{R}^d$. If M is a matrix, M^T denotes its transpose, $M^\dagger$ the pseudo-inverse, $M^{\dagger k} = (M^\dagger)^k$ and $\|M\|$ its spectral norm. Furthermore, $\ker M$ and $\operatorname{ran} M$ are the null space and the range of M respectively. For a square-matrix M, we use $\operatorname{Tr}(M)$ to denote its trace. If v and w are vectors (possibly of different length), $v \otimes w$ is the rank one matrix with entries $(v \otimes w)_{ij} = v_i w_j$.

Given a random vector $\xi \in \mathbb{R}^d$, its noncentered covariance matrix is denoted by

$$\Sigma_\xi = \mathbb{E}[\xi \otimes \xi],$$

which is a positive matrix satisfying the following property

$$\operatorname{ran} \Sigma_\xi = (\ker \Sigma_\xi)^\perp = \operatorname{span}\{x \in \mathbb{R}^d \mid \mathbb{P}[\xi \in B(x,r)] > 0 \ \forall r > 0\}, \tag{4}$$

here $B(x,r)$ denotes the ball of radius r with the center at x. A random vector ξ is called sub-gaussian if

$$\|\xi\|_{\psi_2} := \sup_{v \in S^{d-1}} \sup_{q \geq 1} q^{-\frac{1}{2}} \mathbb{E}[|\langle \xi, v\rangle|^q]^{\frac{1}{q}} < +\infty. \tag{5}$$

The value $\|\xi\|_{\psi_2}$ is called the sub-gaussian norm of ξ and the space of sub-gaussian vectors becomes a normed vector space (Vershynin, 2012). Appendix B reviews some basic properties about sub-gaussian vectors.

We consider the following class of inverse problems.

Assumption 1 *In the statistical linear inverse problem*

$$Y = AX + \sigma W,$$

the following conditions hold true:

a) A is an $m \times d$-matrix with norm $\|A\| = 1$;

b) the signal $X \in \mathbb{R}^d$ is a sub-gaussian random vector with $\|X\|_{\psi_2} = 1$;

c) the noise $W \in \mathbb{R}^m$ is a sub-gaussian centered random vector independent of X with $\|W\|_{\psi_2} = 1/\sqrt{2}$ and with the noise level $0 < \sigma < \sqrt{2}$;

d) the covariance matrix Σ_X of X has a low rank matrix, i.e.,

$$\operatorname{rank}(\Sigma_X) = h \ll d.$$

We add some comments on the above conditions. The normalisation assumptions on $\|A\|$, $\|X\|_{\psi_2}$ and $\|W\|_{\psi_2}$ are stated only to simplify the bounds. They can always be satisfied by rescaling A, X and W and our results hold true by replacing σ with $\sqrt{2}\|W\|_{\psi_2}\|A\|^{-1}\|X\|_{\psi_2}^{-1}\sigma$. The upper bound on σ reflects the intuition that σW is a small perturbation of the noiseless problem.

Condition d) means that X spans a low dimensional subspace of $\mathbb{R}^d$. Indeed, by (4) condition d) is equivalent to the fact that the vector space

$$\mathcal{V} = \operatorname{ran} \Sigma_X = \operatorname{span}\{x \in \mathbb{R}^d \mid \mathbb{P}[X \in B(x,r)] > 0 \text{ for all } r > 0\} \tag{6}$$

is an h-dimensional subspace and h is the dimension of the minimal subspace containing X with probability 1, *i.e.,*

$$h = \min_{\mathcal{K}} \dim \mathcal{K},$$

where the minimum is taken over all subspaces $\mathcal{K} \subset \mathbb{R}^d$ such that $\mathbb{P}[X \in \mathcal{K}] = 1$.

We write $a \lesssim b$ if there exists an absolute constant C such that $a \leq Cb$. By *absolute* we mean that it holds for all the problems $Y = AX + \sigma W$ satisfying Assumption 1, in particular, it is independent of d, m and h.

The datum Y depends only on the projection $X^\dagger$ of X onto $\ker A^\perp$ and Z^α as solutions of (3) also belong to $\ker A^\perp$. Therefore, we can always assume, without loss of generality, for the rest of the paper that A is *injective* by replacing X with $X^\dagger$, which is a sub-gaussian random vector, and $\mathbb{R}^d$ with $\ker A^\perp$.

Since A is injective, $\operatorname{rank}(A) = d$ and we define the singular value decomposition of A by $(u_i, v_i, \sigma_i)_{i=1}^d$, so that $A = UDV^T$ or

$$Av_i = \sigma_i u_i, \quad i = 1, \dots, d,$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0$. Since $\|A\| = 1$, clearly $\sigma_1 = 1$. Furthermore, let Q be the projection onto the $\operatorname{span}\{u_1, \dots, u_d\}$, so that $QA = A$, and we have the decomposition

$$Q = AA^\dagger. \tag{7}$$

Recalling (6), since A is now assumed injective and

$$\Sigma_{AX} = \mathbb{E}[AX \otimes AX] = A\Sigma_X A^T,$$

then

$$\mathcal{W} = \operatorname{ran} \Sigma_{AX} = (\ker \Sigma_{AX})^\perp = A\mathcal{V}, \tag{8}$$

and, by condition d) in Assumption 1, we have as well $\dim \mathcal{W} = h$.

We denote by Π the projection onto $\mathcal{W}$ and by

$$p = \max\{i \in \{1, \dots, d\} \mid \Pi u_i \neq 0\}, \tag{9}$$

so that, with probability 1,

$$\Pi AX = AX \quad \text{and} \quad X = \sum_{i=1}^{p} \langle X, v_i \rangle v_i. \tag{10}$$

Finally, the random vectors $\eta = \sigma W$, AX, and Y are sub-gaussian and take value in $\mathbb{R}^m$, $\mathcal{W}$, and $\mathbb{R}^m$, respectively, with

$$\|AX\|_{\psi_2} \leq \|A^T\| \|X\|_{\psi_2} = 1 \qquad \|Y\|_{\psi_2} \leq \|AX\|_{\psi_2} + \sigma \|W\|_{\psi_2} \leq 2 \tag{11}$$

since, by Assumption 1, $\|A\| = 1$ and $\sigma \leq \sqrt{2}$.

For any $t \in [0, 1]$ we set Z^t as the solution of the minimization problem

$$\min_{z \in \mathbb{R}^d} \left(t \|Az - Y\|^2 + (1 - t) \|z\|^2 \right), \tag{12}$$

which is the solution of the Tikhonov functional

$$\min_{z\in\mathbb{R}^d} \|Az - Y\|^2 + \alpha \|z\|^2.$$

with $\alpha = (1-t)/t \in [0, +\infty]$.

For $t < 1$, the solution is unique, for $t = 1$ the minimizer is not unique and we set

$$Z^1 = A^\dagger Y.$$

The explicit form of the solution of (12) is given by

$$\begin{aligned} Z^t &= t(tA^TA + (1-t)I)^{-1}A^TY \\ &= \sum_{i=1}^d \frac{t\sigma_i}{t\sigma_i^2 + (1-t)} \langle Y, u_i\rangle v_i \\ &= \sum_{i=1}^d \left(\frac{t\sigma_i^2}{t\sigma_i^2 + (1-t)} \langle X, v_i\rangle + \frac{t\sigma_i}{t\sigma_i^2 + (1-t)} \langle \eta, u_i\rangle \right) v_i, \end{aligned} \tag{13}$$

which shows that Z^t is also a sub-gaussian random vector.

We seek for the optimal parameter $t^* \in [0, 1]$ that minimizes the reconstruction error

$$\min_{t\in[0,1]} \|Z^t - X\|^2.$$

Since X is not known, the optimal parameter t^* can not be computed. We assume that we have at disposal a training set of n-independent noisy data

$$Y_1, \dots, Y_n,$$

where $Y_i = AX_i + \sigma W_i$, and each pair (X_i, W_i) is distributed as (X, W), for $i = 1, \dots, n$.

We introduce the $d \times d$ empirical covariance matrix

$$\widehat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n Y_i \otimes Y_i, \tag{14}$$

and we denote by $\widehat{\Pi}_n$ the projections onto the vector space spanned by the first h-eigenvectors of $\widehat{\Sigma}_n$, where the corresponding (repeated) eigenvalues are ordered in a nonincreasing way.

Remark 1 *The well-posedness of the empirical realization $\widehat{\Pi}_n$ in terms of spectral gap at the h-th eigenvalue will be given in Theorem 3, where we show that for n large enough the $h+1$-th eigenvalue is strictly smaller than the h-th eigenvalue.*

We define the empirical estimators of X and η as

$$\widehat{X} = A^\dagger \widehat{\Pi}_n Y \qquad \text{and} \qquad \hat{\eta} = (Y - \widehat{\Pi}_n Y), \tag{15}$$

so that, by Equation (7),

$$A\widehat{X} + Q\hat{\eta} = QY. \tag{16}$$

Furthermore, we set $\widehat{t}_n \in [0,1]$ as the minimizer of

$$\min_{t\in[0,1]} \|Z^t - \widehat{X}\|^2.$$

If $\widehat{X}$ is close to X, we expect that the solution $Z^{\widehat{t}_n}$ has a reconstruction error close to the minimum value. In the following sections, we study the statistical properties of $\widehat{t}_n$. However, we first provide some a priori information on the optimal regularization parameter t^*.

2.1 Distribution dependent quantities

We define the function $t \mapsto \|R(t)\|^2$, where

$$R(t) = Z^t - X \qquad t \in [0,1]$$

is the reconstruction error vector. Clearly, the function $t \mapsto \|R(t)\|^2$ is continuous, so that a global minimizer t^* always exists in the compact interval $[0,1]$.

Define for all $t \in [0,1]$ the $d \times d$ matrix

$$B(t) = tA^TA + (1-t)I = \sum_{i=1}^{d} (t\sigma_i^2 + 1 - t)\, v_i \otimes v_i,$$

which is invertible since A is injective and its inverse is

$$B(t)^{-1} = \sum_{i=1}^{d} \frac{1}{t\sigma_i^2 + 1 - t}\, v_i \otimes v_i.$$

Furthermore, $B(t)$ and $B(t)^{-1}$ are smooth functions of the parameter t and

$$B'(t) = (A^TA - I) \qquad (B(t)^{-1})' = -B(t)^{-2}(A^TA - I). \tag{17}$$

Since $Y = AX + \eta$, expression (13) gives

$$\begin{aligned} R(t) &= tB(t)^{-1}A^TY - X \\ &= tB(t)^{-1}A^T(AX + \eta) - X \\ &= B(t)^{-1}(tA^TAX - B(t)X + tA^T\eta) \\ &= B(t)^{-1}(-(1-t)X + tA^T\eta). \end{aligned} \tag{18}$$

Hence,

$$\begin{aligned} \|R(t)\|^2 &= \|B(t)^{-1}(-(1-t)X + tA^T\eta)\|^2 \\ &= \sum_{i=1}^{d} \left(\frac{-(1-t)\xi_i + t\sigma_i\nu_i}{t\sigma_i^2 + (1-t)} \right)^2, \end{aligned}$$

where for all $i = 1, \ldots, d$

$$\xi_i = \langle X, v_i \rangle \qquad \nu_i = \langle \eta, u_i \rangle.$$

In order to characterize t^* we may want to seek it among the zeros of the following function

$$H(t) = \frac{1}{2}\frac{d}{dt}\|Z^t - X\|^2 = \langle R(t), R'(t)\rangle.$$

Taking into account (17), the differentiation of (18) is given by

$$\begin{aligned} R'(t) &= B(t)^{-1}A^TY - tB(t)^{-2}(A^TA - I)A^TY \\ &= B(t)^{-2}(B(t) - t(A^TA - I))A^TY \\ &= B(t)^{-2}A^TY, \end{aligned} \tag{19}$$

so that

$$\begin{aligned} H(t) &= \langle AB(t)^{-3}(-(1-t)X + tA^T\eta), AX + \eta\rangle \\ &= \sum_{i=1}^d \sigma_i \frac{-(1-t)\xi_i + t\sigma_i\nu_i}{(t\sigma_i^2 + (1-t))^3}(\xi_i\sigma_i + \nu_i) \\ &= \sum_{i=1}^d \sigma_i\xi_i(\xi_i\sigma_i + \nu_i)\frac{(\sigma_i\nu_i\xi_i^{-1} + 1)t - 1}{(1-(1-\sigma_i^2)t)^3} \\ &= \sum_{i=1}^d \sigma_i\alpha_i h_i(t), \end{aligned} \tag{20}$$

where $\alpha_i = \xi_i(\sigma_i\xi_i + \nu_i)$ and $h_i(t) = \frac{(\sigma_i\nu_i\xi_i^{-1}+1)t-1}{(1-(1-\sigma_i^2)t)^3}$.

We observe that

a) if $t = 0$ (*i.e.,* $\alpha = +\infty$), $B(0) = I$, then

$$H(0) = -\|AX\|^2 + \langle AX, \eta\rangle,$$

which is negative if $\|\Pi\eta\| \le \|AX\|$, *i.e.,* for

$$\sigma \le \frac{\|AX\|}{\|\Pi W\|}.$$

Furthermore, by construction,

$$\mathbb{E}[H(0)] = -\operatorname{Tr}(\Sigma_{AX}) < 0;$$

b) if $t = 1$ (*i.e.,* $\alpha = 0$), $B(0) = A^TA$ and

$$\begin{aligned} H(1) &= \langle A(A^TA)^{-3}A^T\eta, AX + \eta\rangle \\ &= \|(AA^T)^\dagger\eta\|^2 + \langle (AA^T)^\dagger\eta, (A^T)^\dagger X\rangle, \end{aligned}$$

which is positive if $\|(AA^T)^\dagger\eta\| \ge \|(A^T)^\dagger X\|$, for example, when

$$\sigma \ge \sigma_d\frac{\|X\|}{|\langle W, u_d\rangle|}.$$

Furthermore, by construction,

$$\mathbb{E}[H(1)] = \operatorname{Tr}(\Sigma_{(AA^T)^\dagger\eta}) > 0.$$

Hence, if the noise level satisfies

$$\frac{\|X\|}{|\langle W, u_d\rangle|} \leq \sigma \leq \frac{\|AX\|}{\|\Pi W\|}$$

the minimizer t^* is in the open interval $(0,1)$ and it is a zero of $H(t)$. If σ is too small, there is no need of regularization since we are dealing with a finite dimensional problem. On the opposite side, if σ is too big, the best solution is the trivial one, *i.e.,* $Z^{t^*} = 0$.

2.2 Empirical quantities

We replace X and η with their empirical counterparts defined in (15). By Equation (16) and reasoning as in Equation (18), we obtain

$$\begin{aligned}\widehat{R}_n(t) &= Z^t - \widehat{X} \\ &= tB(t)^{-1}A^TQY - \widehat{X} \\ &= B(t)^{-1}(-(1-t)\widehat{X} + tA^T\hat{\eta}),\end{aligned}$$

and

$$\begin{aligned}\|\widehat{R}_n(t)\|^2 &= \|B(t)^{-1}(-(1-t)\widehat{X} + tA^T\hat{\eta})\|^2 \\ &= \sum_{i=1}^{d}\left(\frac{-(1-t)\hat{\xi}_i + t\sigma_i\hat{\nu}_i}{t\sigma_i^2 + (1-t)}\right)^2,\end{aligned}$$

where for all $i = 1, \ldots, d$

$$\hat{\xi}_i = \langle \widehat{X}, v_i\rangle \qquad \text{and} \qquad \hat{\nu}_i = \langle \hat{\eta}, u_i\rangle.$$

Clearly,

$$\widehat{R}'_n(t) = R'(t) = B(t)^{-2}A^TQY. \tag{21}$$

From (20), we get

$$\begin{aligned}\widehat{H}_n(t) &= \langle \widehat{R}_n(t), \widehat{R'}_n(t)\rangle \\ &= \langle B(t)^{-3}(-(1-t)\widehat{X} + tA^T\hat{\eta}), A^TA\widehat{X} + A^T\hat{\eta})\rangle \\ &= \sum_{i=1}^{d}\frac{-(1-t)\hat{\xi}_i + t\sigma_i\hat{\nu}_i}{(1-(1-\sigma_i^2)t)^3}(\hat{\xi}_i\sigma_i^2 + \sigma_i\hat{\nu}_i), \\ &= \sum_{i=1}^{d}\sigma_i\hat{\alpha}_i\hat{h}_i(t),\end{aligned} \tag{22}$$

where $\hat{\alpha}_i = \hat{\xi}_i(\sigma_i\hat{\xi}_i + \hat{\nu}_i)$ and $\hat{h}_i(t) = \frac{(\sigma_i\hat{\nu}_i\hat{\xi}_i^{-1}+1)t-1}{(1-(1-\sigma_i^2)t)^3}$.

An alternative form in terms of Y and $\widehat{\Pi}_n$, which can be useful as a different numerical implementation, is

$$\begin{aligned}\widehat{H}_n(t) &= \langle B(t)^{-1}(tA^T(Y - \widehat{\Pi}_nY) - (1-t)A^{-1}\widehat{\Pi}_nY), B(t)^{-2}A^TY\rangle \\ &= \langle tAA^T(Y - \widehat{\Pi}_nY) - (1-t)Q\widehat{\Pi}_nY, (tAA^T + (1-t)I)^{\dagger 3}QY\rangle \\ &= \langle tAA^T(Y - \widehat{\Pi}_nY) - (1-t)\widehat{\Pi}_nY, (tAA^T + (1-t)I)^{\dagger 3}QY\rangle.\end{aligned}$$

As for t^*, the minimizer $\widehat{t}_n$ of the function $t \mapsto \|\widehat{R}_n(t)\|^2$ always exists in $[0,1]$ and, for σ in the range of interest, it is in the open interval $(0,1)$, so that it is a zero of the function $\widehat{H}_n(t)$.

3. Concentration inequalities

In this section, we bound the difference between the empirical estimators and their distribution dependent counterparts.

By (8) and item d) of Assumption 1, the covariance matrix Σ_{AX} has rank h and, we set $\lambda_{\min}$ to be the smallest non-zero eigenvalue of Σ_{AX}. Furthermore, we denote by Π^Y the projection from $\mathbb{R}^m$ onto the vector space spanned by the eigenvectors of Σ_Y, whose eigenvalue is greater than $\lambda_{\min}/2$.

The following proposition shows that Π^Y is close to Π if the noise level is small enough.

Proposition 2 *If $\sigma^2 < \lambda_{\min}/4$, then $\dim \operatorname{ran} \Pi^Y = h$ and*

$$\|\Pi^Y - \Pi\| \leq \frac{2\sigma^2}{\lambda_{\min}}. \tag{23}$$

Proof Since AX and W are independent and W has zero mean, then

$$\Sigma_Y = \Sigma_{AX} + \sigma^2 \Sigma_W.$$

We apply Proposition 14 from Appendix A with $\mathcal{A} = \Sigma_{AX}$ and $\mathcal{B} = \Sigma_Y$, regarded as (compact) operators on $\mathbb{R}^m$. Since Σ_W is a positive matrix and W is a sub-gaussian vector satisfying (11), with the choice $q = 2$ in (5), we have

$$\|\Sigma_W\| = \sup_{v \in S^{m-1}} \langle \Sigma_W v, v \rangle = \sup_{v \in S^{m-1}} \mathbb{E}[\langle W, v \rangle^2] \leq 2\|W\|_{\psi_2}^2 = 1, \tag{24}$$

so that $\|\Sigma_Y - \Sigma_{AX}\| \leq \sigma^2 < \lambda_{\min}/4$.

By (8) the rank of Σ_{AX} is h and, with the notation of Proposition 14, we have that $N = h$. With the choice of $j = h$, so that $\Pi = P_h$ and $\lambda_{\min} = \alpha_j - \alpha_{j+1}$, there exists m such that β_m is the m-the largest eigenvalue[1] of Π^Y, $\dim Q_m = \dim P_h = h$ and $\beta_{m+1} < \lambda_{\min}/2 < \beta_m$. It follows that $\Pi^Y = Q_m$, $\dim \operatorname{ran} \Pi^Y = h$, and (37) implies (23) since $\lambda_{h+1} = 0$. ∎

Recall that $\widehat{\Pi}_n$ is the projection onto the vector space spanned by the first h-eigenvectors of $\widehat{\Sigma}_n$ defined by (14).

Theorem 3 *Given $\tau > 0$ with probability greater than $1 - 2\mathrm{e}^{-\tau^2}$, $\widehat{\Pi}_n$ coincides with the projection onto the vector space spanned by the eigenvectors of $\widehat{\Sigma}_n$, whose eigenvalues are greater than $\lambda_{\min}/2$. Furthermore*

$$\|\widehat{\Pi}_n - \Pi\| \lesssim \frac{1}{\lambda_{\min}} \left(\sqrt{\frac{m}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2 \right), \tag{25}$$

1. In the statement of Proposition 14 the eigenvalues are counted without their multiplicity.

provided that

$$n \gtrsim (\sqrt{m}+\tau)^2 \max\left\{\frac{64}{\lambda_{\min}^2}, 1\right\} \qquad (26)$$
$$\sigma^2 < \frac{\lambda_{\min}}{8}.$$

Proof Observe that Theorem 19 applied with $\xi_i = Y_i$ and (24) imply that, with probability greater than $1 - 2\mathrm{e}^{-\tau^2}$,

$$\begin{aligned}\|\widehat{\Sigma}_n - \Sigma_{AX}\| &\leq \|\widehat{\Sigma}_n - \Sigma_Y\| + \|\Sigma_Y - \Sigma_{AX}\| \\ &\leq C\left(\sqrt{\frac{m}{n}} + \frac{\tau}{\sqrt{n}}\right) + \sigma^2 \\ &\leq \frac{\lambda_{\min}}{8} + \frac{\lambda_{\min}}{8} = \frac{\lambda_{\min}}{4},\end{aligned}$$

where the last inequality follows by the assumptions on n, σ, and

$$C\sqrt{\frac{m}{n}} + \frac{\tau}{\sqrt{n}} \leq \min\{1, \lambda_{\min}/8\} \leq 1.$$

Proposition 14 with $\mathcal{A} = \Sigma_{AX}$, $\mathcal{B} = \widehat{\Sigma}_n$, and $N = h$ tells that β_m, the m largest eigenvalue of $\widehat{\Sigma}_n$, satisfies $\beta_{m+1} < \lambda_{\min}/2 < \beta_m$. Let Q_m the orthogonal projection onto the vector space spanned by the eigenvectors of $\widehat{\Sigma}_n$, whose eigenvalues are greater or equal than $\lambda_{\min}/2$. Then, $\dim \operatorname{ran} Q_m = \dim \operatorname{ran} \Pi = h$, so that $Q_m = \widehat{\Pi}_n$. Finally, (37) implies (25). Note that the constant C depends on $\|Y\|_{\psi_2} \leq 2$ by (11), so that it becomes an absolute constant, when considering the worst case $\|Y\|_{\psi_2} = 2$. ∎

If n and σ satisfy (26), the above proposition shows that the empirical covariance matrix $\widehat{\Sigma}_n$ has a spectral gap around the value $\lambda_{\min}/2$ and the number of eigenvectors, whose eigenvalues are greater than $\lambda_{\min}/2$, is precisely h, so that $\widehat{\Pi}_n$ is uniquely defined. Furthermore, the dimension h can be estimated by observing spectral gaps in the singular value decomposition of $\widehat{\Sigma}_n$.

We need the following technical lemma.

Lemma 4 *Given $\tau > 0$, with probability greater than $1 - 4\mathrm{e}^{-\tau^2}$, simultaneously it holds*

$$\|X\| \lesssim (\sqrt{h}+\tau) \qquad \|Y\| \lesssim (\sqrt{h}+\sigma\sqrt{m}+\tau) \qquad \|\Pi W\| \lesssim (\sqrt{h}+\tau). \qquad (27)$$

Proof Since X is a sub-gaussian random vector taking values in $\mathcal{V}$ with $h = \dim \mathcal{V}$, taking into account that $\|X\|_{\psi_2} = 1$, bound (39) gives

$$\|X\| \leq 9(\sqrt{h}+\tau),$$

with probability greater than $1 - 2\mathrm{e}^{-\tau^2}$. Since W is a centered sub-gaussian random vector taking values in $\mathbb{R}^m$ and $\|W\|_{\psi_2} \leq 1$, by (40)

$$\|W\| \leq 16(\sqrt{m}+\tau),$$

with probability greater than $1-\mathrm{e}^{-\tau^2}$. Since $\|A\|=1$ and

$$\|Y\| \le \|AX\| + \sigma\|W\| \le \|X\| + \sigma\|W\|,$$

the first two bounds in (27) hold true with probability greater than $1-3\mathrm{e}^{-\tau^2}$. Since ΠW is a centered sub-gaussian random vector taking values in $\mathcal{W}$ with $h=\dim\mathcal{W}$, and $\|\Pi W\|_{\psi_2}\le 1$, by (40)

$$\|\Pi W\| \le 16(\sqrt{h}+\tau),$$

with probability greater than $1-\mathrm{e}^{-\tau^2}$. ■

As a consequence, we have the following bound.

Proposition 5 *Given $\tau>0$, if n and σ satisfy* (26)*, then with probability greater than* $1-6\mathrm{e}^{-\tau^2}$

$$\|(\Pi-\widehat{\Pi}_n)Y-\Pi\eta\| \lesssim B(n,\tau,\sigma), \tag{28}$$

where

$$\begin{aligned} B(n,\tau,\sigma) = {} & \frac{1}{\lambda_{\min}}\sqrt{\frac{hm}{n}} + \sigma\left(\sqrt{h}+\frac{1}{\lambda_{\min}}\frac{m}{\sqrt{n}}\right) + \frac{\sigma^2}{\lambda_{\min}}\sqrt{h} + \frac{\sigma^3}{\lambda_{\min}}\sqrt{m} + \\ & + \tau\left(\frac{1}{\lambda_{\min}}\sqrt{\frac{m}{n}} + \sigma\left(1+\frac{1}{\lambda_{\min}}\sqrt{\frac{m}{n}}\right) + \frac{\sigma^2}{\lambda_{\min}}\right) + \tau^2\frac{1}{\lambda_{\min}}\frac{1}{\sqrt{n}}. \end{aligned} \tag{29}$$

Proof Clearly,

$$\|(\Pi-\widehat{\Pi}_n)Y-\Pi\eta\| \le \|\Pi-\widehat{\Pi}_n\|\|Y\| + \sigma\|\Pi W\|.$$

If (26) holds true, bounds (25) and (27) imply

$$\|(\Pi-\widehat{\Pi}_n)Y-\Pi\eta\| \lesssim \frac{1}{\lambda_{\min}}\left(\sqrt{\frac{m}{n}}+\frac{\tau}{\sqrt{n}}+\sigma^2\right)(\sqrt{h}+\sigma\sqrt{m}+\tau)+\sigma(\sqrt{h}+\tau),$$

with probability greater than $1-6\mathrm{e}^{-\tau^2}$. By developing the brackets and taking into account that $\sqrt{h+m}\le\sqrt{2m}$, (28) holds true. ■

Remark 6 *Usually in machine learning bounds of the type* (28) *are considered in terms of their expectation, e.g., with respect to (X,Y). In our framework, this would amount to the following bound*

$$\begin{aligned} & \mathbb{E}\left[\|(\Pi-\widehat{\Pi}_n)Y-\Pi\eta\| \,\middle|\, Y_1,\dots,Y_n\right] \lesssim \\ \lesssim \quad & \frac{1}{\lambda_{\min}}\left(\sqrt{\frac{m}{n}}+\frac{\tau}{\sqrt{n}}+\sigma^2\right)(\sqrt{h}+\sigma\sqrt{m})+\sigma\sqrt{h}, \end{aligned}$$

obtained by observing that $\mathbb{E}[\|Y\|]\le\mathbb{E}[\|A\|\|X\|]+\sigma\mathbb{E}[\|W\|]$,

$$\mathbb{E}[\|X\|]^2 \le \mathbb{E}[\|X\|^2] = \mathrm{Tr}(\Sigma_X) \le 2h\|X\|_{\psi_2}^2 \lesssim h,$$

and, by a similar computation,

$$\mathbb{E}[\|W\|]\lesssim\sqrt{m} \qquad \mathbb{E}[\|\Pi W\|]\lesssim\sqrt{h}.$$

Our bound (28) *is much stronger and it holds in probability with respect to both the training set $Y_1,\dots Y_n$ and the new pair (X,Y).*

Our first result is a direct consequence of the estimate (28).

Theorem 7 *Given $\tau > 0$, with probability greater than $1 - 6\mathrm{e}^{-\tau^2}$,*

$$\begin{aligned}
\|\widehat{X} - X\| &\lesssim \frac{1}{\sigma_d} B(n, \tau, \sigma) \\
\|Q\hat{\eta} - Q\eta\| &\lesssim B(n, \tau, \sigma) \\
\|Z^{\widehat{t}_n} - X\| - \|Z^{t^*} - X\| &\lesssim \frac{1}{\sigma_d} B(n, \tau, \sigma) \\
\sup_{0 \le t \le 1} |\|\widehat{R}_n(t)\| - \|R(t)\|| &\lesssim \frac{1}{\sigma_d} B(n, \tau, \sigma)
\end{aligned}$$

provided that n and σ satisfy (26).

Proof By the first identity of (10)

$$\begin{aligned}
X - \widehat{X} &= A^\dagger \Pi A X - A^\dagger \widehat{\Pi}_n (AX + \eta) \\
&= A^\dagger (\Pi - \widehat{\Pi}_n) AX + A^\dagger (\Pi - \widehat{\Pi}_n)\eta - A^\dagger \Pi \eta \\
&= A^\dagger \left((\Pi - \widehat{\Pi}_n) Y - \Pi \eta \right),
\end{aligned} \tag{30}$$

so that

$$\|X - \widehat{X}\| \le \frac{1}{\sigma_d} \|(\Pi - \widehat{\Pi}_n) Y - \Pi \eta\|.$$

An application of (28) to the previous estimate gives the first bound of the statement. Similarly, we derive the second bound as follows. Equations (16), (7), and (30) yield

$$\begin{aligned}
Q\eta - Q\hat{\eta} &= A(X - \widehat{X}) \\
&= Q \left((\Pi - \widehat{\Pi}_n) Y - \Pi \eta \right).
\end{aligned}$$

The other bounds follow by estimating them by multiples of $\|X - \widehat{X}\|$ as we show below. By definition of $\widehat{t}_n$

$$\begin{aligned}
\|Z^{\widehat{t}_n} - X\| &\le \|Z^{\widehat{t}_n} - \widehat{X}\| + \|X - \widehat{X}\| \\
&\le \|Z^{t^*} - \widehat{X}\| + \|X - \widehat{X}\|, \\
&\le \|Z^{t^*} - X\| + 2\|X - \widehat{X}\|.
\end{aligned}$$

Furthermore,

$$\widehat{R}_n(t) - R(t) = X - \widehat{X} = A^\dagger \left((\Pi - \widehat{\Pi}_n) Y - \Pi \eta \right), \tag{31}$$

and triangle inequality gives

$$\sup_{0 \le t \le 1} |\|\widehat{R}_n(t)\| - \|R(t)\|| \le \sup_{0 \le t \le 1} \|\widehat{R}_n(t) - R(t)\| = \|X - \widehat{X}\|.$$

All the remaining bounds in the statement of the theorem are now consequences of (28). ■

Remark 8 *To justify and explain the consistency of the sampling strategy for approximation of the optimal regularization parameter t^*, let us assume that n goes to infinity and σ vanishes. Under this theoretical assumption, Theorem 7 shows that $\|\widehat{R}_n(t)\|$ convergences uniformly to $\|R(t)\|$ with high probability. The uniform convergence implies the Γ-convergence (see Braides, 2001), and, since the domain $[0,1]$ is compact, Theorem 1.22 in Braides (2001) ensures that*

$$\lim_{\substack{n\to+\infty\\ \sigma\to 0}} \left(\inf_{0\le t\le 1} \|\widehat{R}_n(t)\| - \inf_{0\le t\le 1} \|R(t)\| \right) = 0.$$

While the compactness given by the Γ-convergence guarantees the consistency of the approximation to an optimal parameter, it is much harder for arbitrary A to provide an error bound, depending on n. For the case $A = I$ in Section 4 we are able to establish very precise quantitative bounds with high probability.

Remark 9 *Under the conditions of Theorem 7 for all $i = 1, \dots, d$ it holds as well*

$$|\xi_i - \hat{\xi}_i| \lesssim \frac{1}{\sigma_i} B(n,\tau,\sigma)$$
$$|\nu_i - \hat{\nu}_i| \lesssim B(n,\tau,\sigma).$$

These bounds are a direct consequence of Theorem 7.

The following theorem is about the uniform approximation to the derivative function $H(t)$.

Theorem 10 *Given $\tau > 0$, with probability greater than $1 - 10\mathrm{e}^{-\tau^2}$,*

$$\sup_{0\le t\le 1} |\widehat{H}_n(t) - H(t)| \lesssim B(n,\tau,\sigma) \left(\frac{1}{\sigma_p^3}(\sqrt{h} + \tau) + \frac{\sigma}{\sigma_d^4}(\sqrt{d} + \tau) \right)$$

provided that n and σ satisfy (26), where p is defined in (9).

Proof Equations (21) and (31) give

$$\begin{aligned} \widehat{H}_n(t) - H(t) &= \langle \widehat{R}_n(t) - R(t), R'(t) \rangle \\ &= \langle (\Pi - \widehat{\Pi}_n)Y - \Pi\eta, (A^T)^{-1} B(t)^{-2} A^T Y \rangle \\ &= \langle (\Pi - \widehat{\Pi}_n)Y - \Pi\eta, (tAA^T + (1-t)I)^{-2} QY \rangle, \end{aligned}$$

where we observe that $tAA^T + (1-t)I$ is invertible on $\operatorname{ran} Q$. Hence,

$$\begin{aligned} |\widehat{H}_n(t) - H(t)| \le \|(\Pi - \widehat{\Pi}_n)Y - \Pi\eta\| \times \\ \times \left(\|(tAA^T + (1-t)I)^{-2} AX\| + \sigma \|(tAA^T + (1-t)I)^{-2} QW\| \right). \end{aligned}$$

Furthermore, recalling that $AX = \Pi AX$ and $\Pi u_i = 0$ for all $i > p$, (27) implies that

$$\|(tAA^T + (1-t)I)^{-2} AX\| \le \frac{\sigma_p}{(t\sigma_p^2 + (1-t))^2} \|X\| \lesssim \frac{1}{\sigma_p^3}(\sqrt{h} + \tau)$$
$$\|(tAA^T + (1-t)I)^{-2} QW\| \le \frac{1}{(t\sigma_d^2 + (1-t))^2} \|QW\| \lesssim \frac{1}{\sigma_d^4}(\sqrt{d} + \tau)$$

hold with probability greater than $1 - 4\mathrm{e}^{-\tau^2}$. Bound (28) provides the desired claim. ■

The uniform approximation result of Theorem 10 allows us to claim that any $\widehat{t}_n \in [0,1]$ such that $\widehat{H}_n(\widehat{t}_n) = 0$ can be attempted as a proxy for the optimal parameter t^*, especially if it is the only root in the interval $(0,1)$.

Nevertheless, being $\widehat{H}_n$ a sum of d rational functions of polynomial numerator of degree 1 and polynomial denominator of degree 3, the computation of its zeros in $[0,1]$ is equivalent to the computation of the roots of a polynomial of degree $3(d-1)+1 = 3d-2$. The computation cannot be done analytically for $d > 2$, because it would require the solution of a polynomial equation of degree larger than 4. For $d > 2$, we are forced to use numerical methods, but this is not a great deal as by now there are plenty of stable and reliable routines to perform such a task (for instance, Newton method, numerical computation of the eigenvalues of the companion matrix, just to mention a few).

We provide below relatively simple numerical experiments to validate the theoretical results reported above. In Figure 2 we show optimal parameters t^* and corresponding approximations $\widehat{t}_n$ (computed by numerical solution to the scalar nonlinear equation $\widehat{H}_n(t) = 0$ on $[0,1]$), for n different data $Y = AX + \eta$. The accordance of the two parameters t^* and $\widehat{t}_n$ is visually very convincing and their statistical (empirical) distributions reported in Figure 3 are also very close.

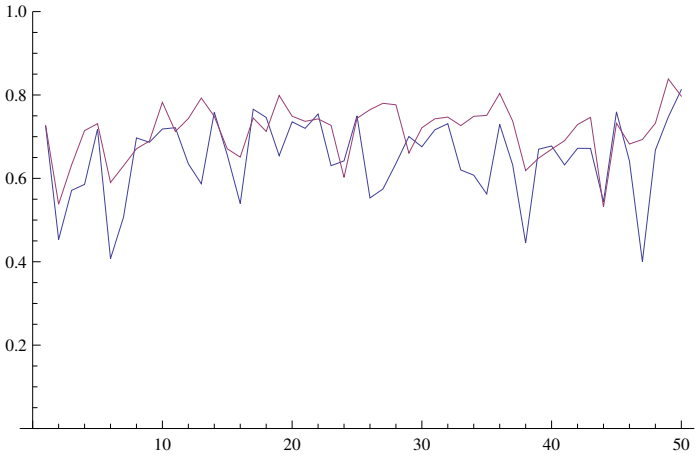

Figure 2: Optimal parameters t^* and corresponding approximations $\widehat{t}_n$ for 50 different data $Y = AX + \eta$ for $X \in \mathbb{R}^d$ and $\eta \in \mathbb{R}^m$ Gaussian vectors, $d = 200$, $m = 60$ and $A \in \mathbb{R}^{60\times 200}$. Here we assumed that $X \in \mathcal{V}$ for $\mathcal{V} = \mathrm{span}\{\mathrm{e}_1, \ldots, \mathrm{e}_5\}$. We designed the matrix in such a way that the spectrum is vanishing, *i*.e., $\sigma_{\min} \approx 0$. Here we considered as noise level $\sigma = 0.03$, so that the optimal parameter t^* is rather concentrated around 0.7. The accordance of the two parameters t^* and $\widehat{t}_n$ is visually very convincing.

In the next two sections we discuss special cases where we can provide even more precise statements and explicit bounds.

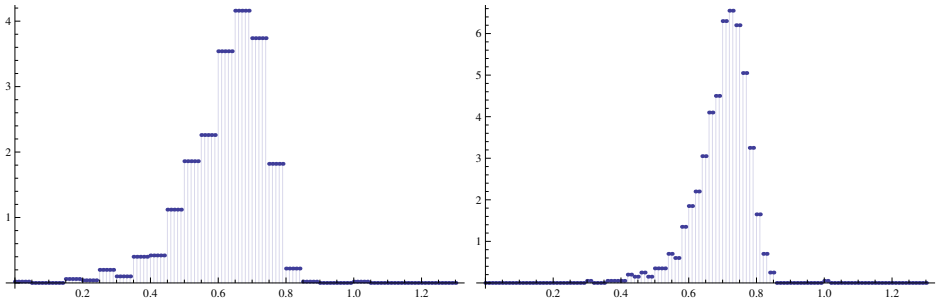

Figure 3: Empirical distribution of the optimal parameters t^* (left) and the corresponding empirical distribution of the approximating parameters $\widehat{t}_n$ (right) for 1000 randomly generated data $Y = AX + \eta$ with the same noise level. The statistical accordance of the two parameters t^* and $\widehat{t}_n$ is shown.

4. The case $A = I$

As an example, we consider the simple case where $m = d$ and $A = I$, so that $\mathcal{W} = \mathcal{V}$. In this case, we get that

$$R(t) = -(1-t)X + t\eta.$$

If $Y \neq 0$, an easy computation shows that the minimizer of the reconstruction error $\|R(t)\|^2$ is

$$t^* = t^*(Y, X) = \varphi\left(\frac{\langle Y, X\rangle}{\langle Y, Y\rangle}\right), \tag{32}$$

where

$$\varphi(s) = \begin{cases} 0 & \text{if } s \leq 0 \\ s & \text{if } 0 < s < 1 \\ 1 & \text{if } s \geq 1 \end{cases}.$$

If $Y = 0$, the solution Z^t does not depend on t, so that there is not a unique optimal parameter and we set $t^* = 0$.

We further assume that X is bounded from 0 with high probability, more precisely,

$$\mathbb{P}[\|X\| < r] \leq 2\exp\left(-\frac{1}{r^2}\right). \tag{33}$$

This assumption is necessary to avoid that the noise is much bigger than the signal.

Theorem 11 *Given $\tau \geq 1$, with probability greater than $1 - 5\mathrm{e}^{-\tau^2}$*

$$|\widehat{t}_n - t^*| \leq \frac{1}{\lambda_{\min}}\left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2\right) + \sigma \ln\left(\frac{\mathrm{e}}{\sigma}\right)(\sqrt{h} + \tau), \tag{34}$$

provided that

$$n \gtrsim (\sqrt{d}+\tau)^2 \max\left\{\frac{64}{\lambda_{\min}^2}, 1\right\} \tag{35a}$$

$$\sigma < \min\left\{\sqrt{\frac{\lambda_{\min}}{8}}, \mathrm{e}^{1-16\tau^2}\right\}. \tag{35b}$$

Proof Without loss of generality, we assume that $\lambda_{\min} \leq 8$. Furthermore, on the event $\{Y = 0\}$, by definition $t^* = \widehat{t}_n = 0$, so that we can further assume that $Y \neq 0$.

Since φ is a Lipschitz continuous function with Lipschitz constant 1,

$$\begin{aligned}
|\widehat{t}_n - t^*| &\leq \frac{|\langle Y, \widehat{X} - X\rangle|}{\|Y\|^2} \\
&\leq \frac{\|(\Pi - \widehat{\Pi}_n)Y - \Pi\eta\|}{\|Y\|} \\
&\leq \|(\Pi - \widehat{\Pi}_n)\| + \sigma\frac{\|\Pi W\|}{\|Y\|},
\end{aligned}$$

where the second inequality is consequence of (30). Since (35a) and (35b) imply (26) and $m = d$, by (25) we get

$$\|\widehat{\Pi}_n - \Pi\| \lesssim \frac{1}{\lambda_{\min}}\left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2\right)$$

with probability greater than $1 - 2\mathrm{e}^{-\tau^2}$. It is now convenient to denote the probability distribution of X as ρ_X, *i.e.*, $X \sim \rho_X$. Fixed $r > 0$, set

$$\Omega = \{\|X\| < r\} \cup \{2\sigma\langle X, W\rangle < -\|X\|^2/2\},$$

whose probability is bounded by

$$\begin{aligned}
\mathbb{P}[\Omega] &\leq \mathbb{P}[\|X\| < r] + \mathbb{P}[4\sigma\langle X, W\rangle < -\|X\|^2, X \geq r] \\
&= \mathbb{P}[\|X\| < r] + \int_{\|x\|\geq r} \mathbb{P}[4\sigma\langle x, W\rangle < -\|x\|^2]\, d\rho_X(x) \\
&\leq \mathbb{P}[\|X\| < r] + \int_{\|x\|\geq r} \exp\left(-\frac{\|x\|^2}{256\sigma^2}\right) d\rho_X(x) \\
&\leq \mathbb{P}[\|X\| < r] + \exp\left(-\frac{r^2}{256\sigma^2}\right),
\end{aligned}$$

where we use (38c) with $\xi = -W$ (and the fact that W and X are independent), $\tau = \|x\|/(16\sigma)$ and $\|W\|_{\psi_2} = 1/\sqrt{2}$. With the choice $r = 16\tau/\ln(\mathrm{e}/\sigma)$, we obtain

$$\begin{aligned}
\mathbb{P}[\Omega] &\leq \mathbb{P}[\|X\| < 16\tau/\ln(\mathrm{e}/\sigma)] + \exp\left(-\frac{\tau^2}{\sigma^2\ln^2(\mathrm{e}/\sigma)}\right) \\
&\leq \mathbb{P}[\|X\| < 16\tau/\ln(\mathrm{e}/\sigma)] + \exp(-\tau^2),
\end{aligned}$$

where $\sigma \mapsto \sigma \ln(\mathrm{e}/\sigma)$ is an increasing positive function on $(0,1]$, so that it is bounded by 1. Furthermore, by (35b), *i.e.,* $16\tau/\ln(\mathrm{e}/\sigma) \leq \frac{1}{\tau}$, we have

$$\mathbb{P}[\|X\| < 16\tau/\ln(\mathrm{e}/\sigma)] \leq P[\|X\| < \frac{1}{\tau}] \leq 2\exp(-\tau^2),$$

by Assumption (33).

On the event Ω^c

$$\begin{aligned}\|Y\|^2 &= \|X\|^2 + 2\sigma\langle X, W\rangle + \sigma^2\|W\|^2 \\ &\geq \|X\|^2 + 2\sigma\langle X, W\rangle \geq \|X\|^2/2 \\ &\geq r^2/2 \simeq \tau^2/\ln^2(\mathrm{e}/\sigma) \geq 1/\ln^2(\mathrm{e}/\sigma)\end{aligned}$$

since $\tau \geq 1$. Finally, (40) with $\xi = \Pi W \in \mathcal{W}$ yields

$$\|\Pi W\| \lesssim (\sqrt{h} + \tau)$$

with probability greater than $1 - \exp(-\tau^2)$. Taking into account the above estimates, we conclude with probability greater than $1 - 3\exp(-\tau^2)$ that

$$\frac{\|\Pi W\|}{\|Y\|} \lesssim \ln\frac{\mathrm{e}}{\sigma}(\sqrt{h} + \tau).$$

Then, with probability greater than $1 - 5\exp(-\tau^2)$, we conclude the estimate

$$|\widehat{t}_n - t^*| \lesssim \frac{1}{\lambda_{\min}}\left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2\right) + \sigma\ln(\mathrm{e}/\sigma)(\sqrt{h} + \tau).$$

■

Remark 12 *The function* $\ln(\mathrm{e}/\sigma)$ *can be replaced by any positive function* $f(\sigma)$ *such that* $\sigma f(\sigma)$ *is an infinitesimal function bounded by 1 in the interval* $(0,1]$. *The condition* (35b) *becomes*

$$\sigma < \min\left\{\sqrt{\frac{\lambda_{\min}}{8}}, 1\right\} \qquad f(\sigma) \geq 16\tau^2,$$

and, if f *is strictly decreasing,*

$$\sigma < \min\left\{\sqrt{\frac{\lambda_{\min}}{8}}, 1, f^{-1}(16\tau^2)\right\}.$$

Theorem 11 shows that if the number n of examples is large enough and the noise level is small enough, the estimator $\widehat{t}_n$ is a good approximation of the optimal value t^*. Let us stress very much that the number n of samples needed to achieve a good accuracy depends at most algebraically on the dimension d, more precisely $n = \mathcal{O}(d)$. Hence, in this case one does not incur in the *curse of dimensionality*. Moreover, the second term of the error estimate (34) gets smaller for smaller dimensionality h.

Remark 13 *If there exists an orthonormal basis* $(e_i)_i$ *of* $\mathbb{R}^d$*, such that the random variables* $\langle W, e_1\rangle, \dots, \langle W, e_d\rangle$ *are independent with* $\mathbb{E}[\langle W, e_1\rangle^2] = 1$*, then Rudelson and Vershynin (2013, Theorem 2.1) showed that* $\|W\|$ *concentrates around* $\sqrt{d}$ *with high probability. Reasoning as in the proof of Theorem 11, by replacing* $\|X\|^2$ *with* $\sigma^2\|W\|^2$*, with high probability it holds that*

$$|\widehat{t}_n - t^*| \lesssim \frac{1}{\lambda_{\min}} \left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} + \sigma^2 \right) + \frac{1}{\sqrt{d}} (\sqrt{h} + \tau) \frac{\tau}{\sqrt{d}}$$

without assuming condition (33).

In Figures 4–7, we show examples of numerical accordance between optimal and estimated regularization parameters. In this case, the agreement between optimal parameter t^* and learned parameter $\widehat{t}_n$ is overwhelming.

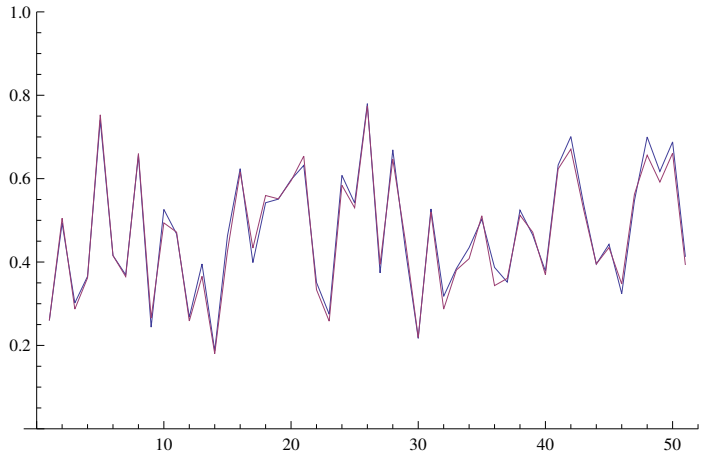

Figure 4: Optimal parameters t^* and corresponding approximations $\widehat{t}_n$ for 50 different data $Y = X + \eta$ for X and η generated randomly with Gaussian distributions in $\mathbb{R}^d$ for $d = 1000$. We assume that $X \in \mathcal{V}$ for $\mathcal{V} = \mathrm{span}\{\mathrm{e}_1, \dots, \mathrm{e}_5\}$.

5. An explicit formula by linearization

While it is not possible to solve the equation $H(t) \approx \widehat{H}_n(t) = 0$ by analytic methods for $d > 2$ in the general case, one might attempt a linearization of this equation in certain regimes. It is well-known that the optimal Tikhonov regularization parameter $\alpha^* = (1-t^*)/t^*$ converges to 0 for vanishing noise level and this means that $t^* = t^*(\sigma) \to 1$ as $\sigma \to 0$. Hence, if the matrix A has a significant spectral gap, *i.e.,* $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \gg 0$ and $\sigma \approx 0$ is small enough, then

$$\sigma_i \gg (1 - t^*), \tag{36}$$

and in this case

$$\hat{h}_i(t) = \frac{(\sigma_i \hat{\nu}_i \hat{\xi}_i^{-1} + 1)t - 1}{((1-t) + t\sigma_i^2)^3} \approx \frac{(\sigma_i \hat{\nu}_i \hat{\xi}_i^{-1} + 1)t - 1}{\sigma_i^6}, \quad t \approx t^*.$$

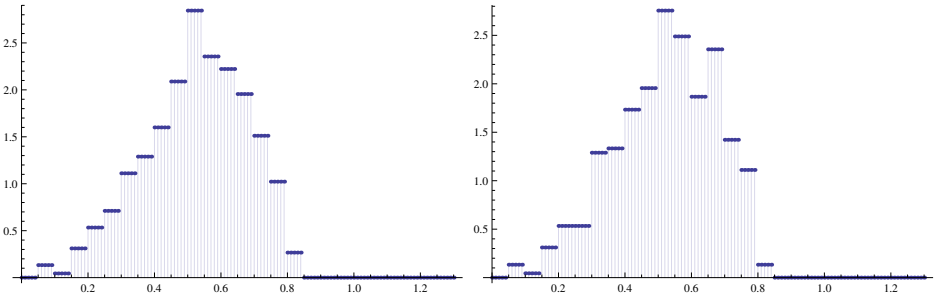

Figure 5: Empirical distribution of the optimal parameters t^* (left) and the corresponding empirical distribution of the learned parameters $\widehat{t}_n$ (right) for 500 randomly generated data $Y = X + \eta$ with the same noise level.

The above linear approximation is equivalent to replacing $B(t)^{-1}$ with $B(1)^{-1} = (A^TA)^{-1}$ and Equation (22) is replaced by the following proxy (at least if $t \approx t^*$)

$$\begin{aligned}
\widehat{H}_n^{\mathrm{lin}}(t) &= \langle tAA^T(Y - \widehat{\Pi}_n Y) - (1-t)\widehat{\Pi}_n Y, (AA^T)^{\dagger 3} Y\rangle \\
&= \langle A(A^TA)^{-3}(-(1-t)\widehat{X} + tA^T\hat{\eta}), A\widehat{X} + \hat{\eta})\rangle \\
&= \left(\langle A(A^TA)^{-3}(\widehat{X} + A^T\hat{\eta}), A\widehat{X} + \hat{\eta}\rangle\right) t - \langle A(A^TA)^{-3}\widehat{X}, A\widehat{X} + \hat{\eta}\rangle \\
&= \left(\sum_{i=1}^d \frac{\hat{\alpha}_i}{\sigma_i^5}(\sigma_i\hat{\nu}_i\hat{\xi}_i^{-1} + 1)\right) t - \sum_{i=1}^d \frac{\hat{\alpha}_i}{\sigma_i^5}.
\end{aligned}$$

The only zero of $\widehat{H}_n^{\mathrm{lin}}(t)$ is

$$\begin{aligned}
\widehat{t}_n^{\mathrm{lin}} &= \frac{\langle \widehat{\Pi}_n Y, (AA^T)^{\dagger 3} Y\rangle}{\langle AA^T(Y - \widehat{\Pi}_n Y) + \widehat{\Pi}_n Y, (AA^T)^{\dagger 3} Y\rangle} \\
&= 1 - \frac{\langle Y - \widehat{\Pi}_n Y, (AA^T)^{\dagger 2} Y\rangle}{\langle AA^T(Y - \widehat{\Pi}_n Y) + \widehat{\Pi}_n Y, (AA^T)^{\dagger 3} Y\rangle} \\
&= \frac{\langle A(A^TA)^{-3}\widehat{X}, A\widehat{X} + \hat{\eta}\rangle}{\langle A(A^TA)^{-3}(\widehat{X} + A^T\hat{\eta}), A\widehat{X} + \hat{\eta}\rangle} \\
&= 1 - \frac{\langle (AA^T)^{\dagger 2}\hat{\eta}, A\widehat{X} + \hat{\eta}\rangle}{\langle A(A^TA)^{-3}(\widehat{X} + A^T\hat{\eta}), A\widehat{X} + \hat{\eta}\rangle} \\
&= \frac{\sum_{i=1}^d \sigma_i^{-5}\hat{\alpha}_i}{\sum_{i=1}^d \sigma_i^{-5}\hat{\alpha}_i(\sigma_i\hat{\nu}_i\hat{\xi}_i^{-1} + 1)} \\
&= 1 - \frac{\sum_{i=1}^d \sigma_i^{-4}\hat{\alpha}_i\hat{\nu}_i}{\sum_{i=1}^d \sigma_i^{-5}\hat{\alpha}_i(\sigma_i\hat{\nu}_i\hat{\xi}_i^{-1} + 1)}.
\end{aligned}$$

In Figure 6, we present the comparison between optimal parameters t^* and their approximations $\widehat{t}_n^{\mathrm{lin}}$. Despite the fact that the gap between σ_d and $1 - t^*$ is not as large as requested in (36), the agreement between t^* and $\widehat{t}_n^{\mathrm{lin}}$ keeps rather satisfactory. In Figure 7, we report the empirical distributions of the parameters, showing essentially their agreement.

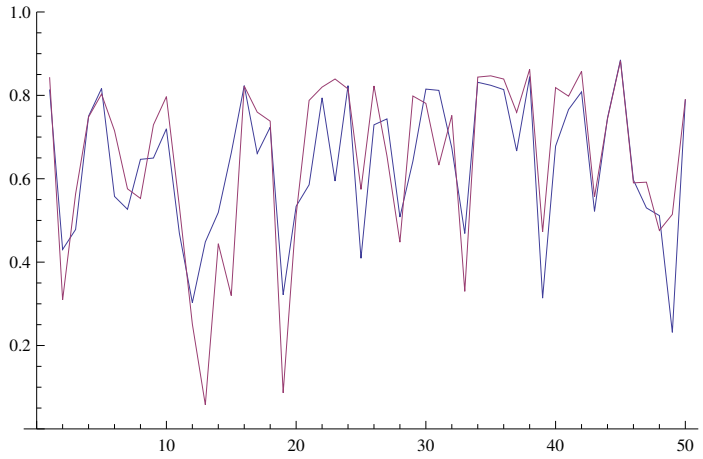

Figure 6: Optimal parameters t^* and corresponding approximations $\widehat{t}_n$ for 50 different data $Y = AX + \eta$ for X and η generated as for the experiment of Figure 2. Here we considered a noise level $\sigma = 0.006$, so that the optimal parameter t^* can be very close to 0.5 and the minimal singular value of A is $\sigma_d \approx 0.7$.

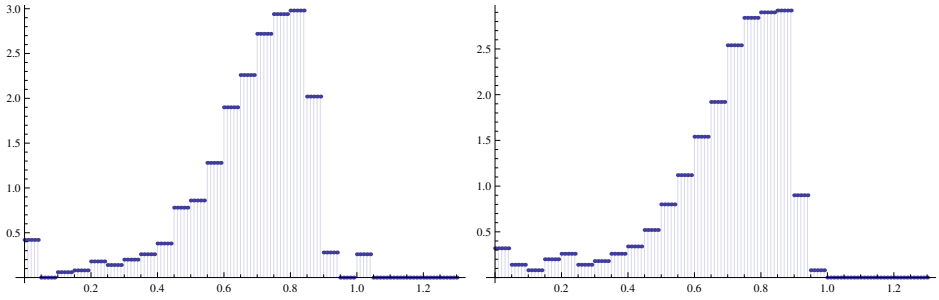

Figure 7: Empirical distribution of the optimal parameters t^* (left) and the corresponding empirical distribution of the approximating parameters $\widehat{t}_n^{\text{lin}}$ (right) for 1000 randomly generated data $Y = AX + \eta$ with the same higher noise level. The statistical accordance of the two parameters t^* and $\widehat{t}_n^{\text{lin}}$ is shown.

6. A glimps to future directions

Motivated by the challenge of the regularization parameter choice/learning relevant for many inverse problems in real-life, in this paper we presented a method to determine the parameter, based on the usage of a supervised machine learning framework. Under the assumption that the solution of the inverse problem is distributed sub-gaussianly over a small dimensional linear subspace $\mathcal{V}$ and the noise is also sub-gaussian, we provided a rigorous theoretical justification for the learning procedure of the function, which maps given noisy data into an optimal Tikhonov regularization parameter. We also presented

and discussed explicit bounds for special cases and provided techniques for the practical implementation of the method.

Our current efforts are devoted to the extension of the analysis to the case where the underlying space $\mathcal{V}$ is actually a smooth lower-dimensional *nonlinear* manifold. This extension will be realized by firstly approximating the nonlinear manifold locally on a proper decomposition by means of affine spaces as proposed in (Chen et al., 2013) and then applying our presented results on those local *linear* approximations.

Another interesting future direction consists of extending the approach to sets $\mathcal{V}$, unions of linear subspaces as in the case of solutions expressible sparsely the respect to certain dictionaries. In this situation, one would need to consider different regularization techniques and possibly non-convex non-smooth penalty quasi-norms.

For the sake of providing a first glimps on the feasibility of the latter possible extension, we consider below the problem of image denoising. In particular, as a simple example of the image denoising algorithm, we consider the wavelet shrinkage (Donoho and Johnstone, 1994): given a noisy image $Y = X + \sigma W$ (already expressed in wavelet coordinates), where σ is the level of Gaussian noise, the denoised image is obtained by

$$Z^\alpha = \mathbb{S}_\alpha(X) = \arg\min_Z \|Z - Y\|^2 + 2\alpha\|Z\|_{\ell_1},$$

where $\|\cdot\|_{\ell_1}$ denotes the ℓ_1 norm, which promotes a sparse representation of the image with respect to a wavelet decomposition. Here, as earlier, we are interested in learning the high-dimensional function mapping noisy images X into their optimal shrinkage parameters, *i.e.,* an optimal solution of $\|Z^\alpha - X\|^2 \to \min_\alpha$.

Employing a properly modified version of the procedure described in this paper, we are obtaining very exciting and promising results, in particular that the optimal shrinkage parameter α essentially depends nonlinearly on very few (actually 1 or 2) linear evaluations of Y. This is not a new observation and it is a data-driven verification of the well-known results of (Donoho and Johnstone, 1994) and (Chambolle et al., 1998), establishing that the optimal parameter depends essentially on two meta-features of the noisy image, *i.e.,* the noise level and its Besov regularity. In Figure 8 and Figure 9 we present the numerical results for wavelet shrinkage, which show that our approach chooses a nearly optimal parameter in terms of peak signal-to-noise ratio (PSNR) and visual quality of the denoising.

Acknowledgments

E. De Vito is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). M. Fornasier acknowledges the financial support of the ERC-Starting Grant HDSPCONTR "High-Dimensional Sparse Optimal Control". V. Naumova acknowledges the support of project "Function-driven Data Learning in High Dimension" (FunDaHD) funded by the Research Council of Norway.

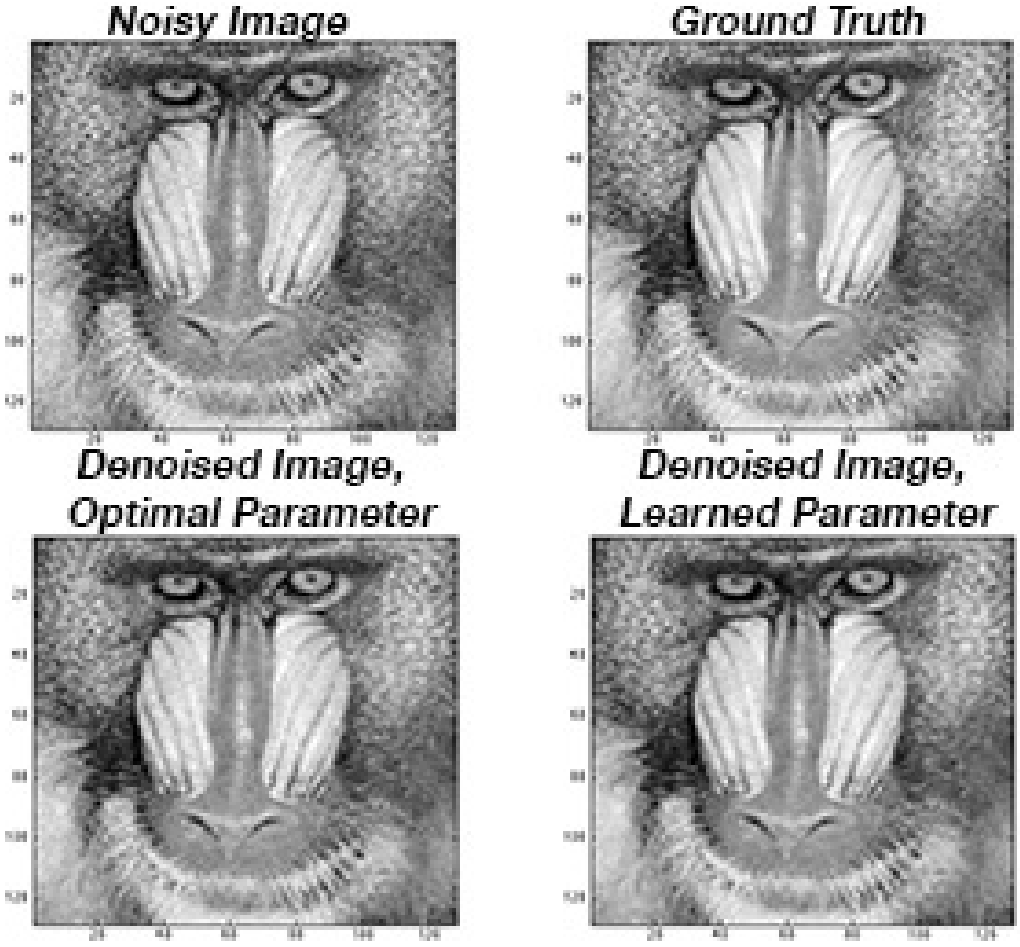

Figure 8: Numerical Experiments for Wavelet Shrinkage

Appendix A. Perturbation result for compact operators

We recall the following perturbation result for compact operators in Hilbert spaces (Anselone, 1971) and (Zwald and Blanchard, 2006, Theorem 3), whose proof also holds without the assumption that the spectrum is simple (see Rosasco et al., 2010, Theorem 20).

Proposition 14 *Let $\mathcal{A}$ and $\mathcal{B}$ be two compact positive operators on a Hilbert space $\mathcal{H}$ and denote by $(\alpha_j)_{j=1}^N$ and $(\beta_m)_{m=1}^M$ the corresponding families of (distinct) strictly positive eigenvalues of $\mathcal{A}$ and $\mathcal{B}$ ordered in a decreasing way. For all $1 \leq j \leq N$, denote by P_j (resp. Q_m with $1 \leq m \leq M$) the projection onto the vector space spanned by the eigenvectors of $\mathcal{A}$ (resp. $\mathcal{B}$) whose eigenvalues are greater or equal than α_j (respect. β_m). Let $j \leq N$ such*

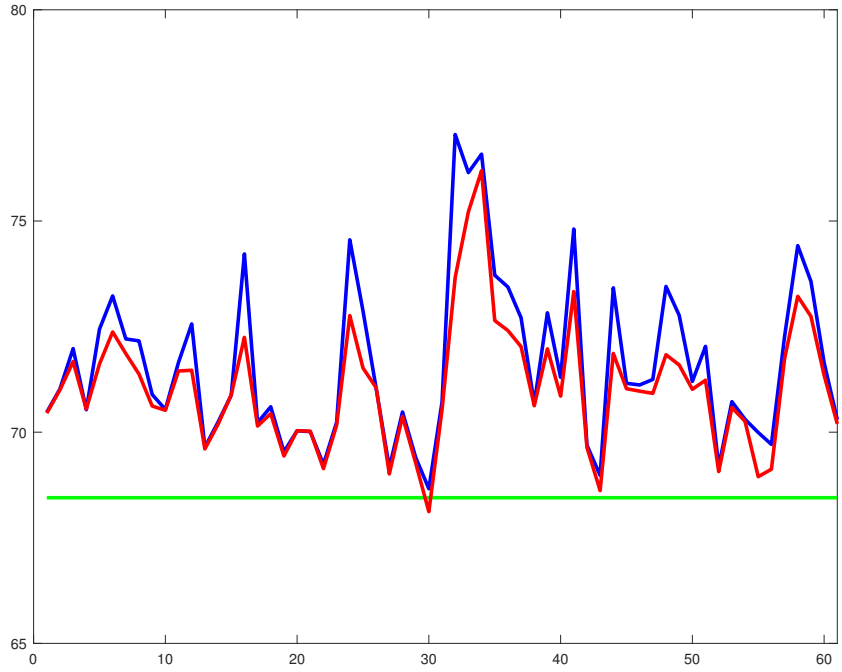

Figure 9: PSNR between the ground truth image and its noisy version (green line); the ground truth and its denoised version with the optimal parameter (blue line) and the learned parameter (red line). The results are presented for 60 random images

that $\|\mathcal{A}-\mathcal{B}\| < \frac{\alpha_j-\alpha_{j+1}}{4}$*, then there exists* $m \le M$ *so that*

$$\begin{aligned} &\beta_{m+1} < \frac{\alpha_j+\alpha_{j+1}}{2} < \beta_m \\ &\|Q_j - P_m\| \le \frac{2}{\alpha_j-\alpha_{j+1}}\|\mathcal{A}-\mathcal{B}\| \\ &\dim P_j\mathcal{H} = \dim Q_m\mathcal{H}. \end{aligned} \tag{37}$$

If $\mathcal{A}$ *and* $\mathcal{B}$ *are Hilbert-Schmidt, the operator norm in the above bound can be replaced by the Hilbert-Schmidt norm.*

In the above proposition, if N or M are finite, $\alpha_{N+1} = 0$ or $\beta_{M+1} = 0$.

Appendix B. Sub-gaussian vectors

We recall some facts about sub-gaussian random vectors and we follow the presentation in (Vershynin, 2012), which provides the proofs of the main results in Lemma 5.5.

Proposition 15 *Let ξ be a sub-gaussian random vector in $\mathbb{R}^d$. Then for all $\tau > 0$*

$$\mathbb{P}[|\langle \xi, v\rangle| > 3\|\xi\|_{\psi_2}\|v\|\tau] \leq 2\exp(-\tau^2). \tag{38a}$$

Under the further assumption that ξ is centered, then

$$\mathbb{E}[\exp(\tau\langle \xi, v\rangle)] \leq \exp\left(8\tau^2\|\xi\|_{\psi_2}^2\right), \tag{38b}$$

and for all $\tau > 0$

$$\mathbb{P}[\langle \xi, v\rangle > 4\sqrt{2}\|\xi\|_{\psi_2}\|v\|\tau] \leq \exp(-\tau^2). \tag{38c}$$

Proof We follow the idea in Vershynin (2012, Lemma 5.5) of explicitly computing the constants. By rescaling ξ to $\xi/\|\xi\|_{\psi_2}$, we can assume that $\|\xi\|_{\psi_2} = 1$.

Let $c > 0$ be a small constant to be fixed. Given, $v \in S^{d-1}$, set $\chi = \langle \xi, v\rangle$, which is a real sub-gaussian vector. By Markov inequality,

$$\begin{aligned}\mathbb{P}[|\chi| > \tau] &= \mathbb{P}[c|\chi|^2 > c\tau^2] = \mathbb{P}[\exp(c|\chi|^2) > \exp(c\tau^2)] \\ &\leq \mathbb{E}[\exp(c|\chi|^2)]\mathrm{e}^{-c\tau^2}.\end{aligned}$$

By (5), we get

$$\mathbb{E}[\exp(c|\chi|^2)] = 1 + \sum_{k=1}^{+\infty}\frac{c^k}{k!}\mathbb{E}[|\chi|^{2k}] \leq 1 + \sum_{k=1}^{+\infty}\frac{(2ck)^k}{k!} \leq 1 + \frac{1}{\mathrm{e}}\sum_{k=1}^{+\infty}(2c\mathrm{e})^k = 1 + \frac{2c}{1-2c\mathrm{e}},$$

where we use the estimate $k! \geq \mathrm{e}(k/\mathrm{e})^k$ for $k \geq 1$. Setting $c = 1/9$, $1 + \frac{2c}{1-2c\mathrm{e}} < 2$, so that

$$\mathbb{P}[|\chi| > 3\tau] \leq 2\exp(-\tau^2).$$

Assume now that $\mathbb{E}[\xi] = 0$. By (5.8) in Vershynin (2012)

$$\mathbb{E}[\exp(\frac{\tau}{\mathrm{e}}\chi)] \leq 1 + \sum_{k=2}^{+\infty}\left(\frac{|\tau|}{\sqrt{k}}\right)^k,$$

and by (5.9) in Vershynin (2012)

$$\exp\left(\frac{\tau^2M^2}{2\mathrm{e}^2}\right) = \exp\left(\tau^2C^2\right) \geq 1 + \sum_{h=2}\left(\frac{C|\tau|}{\sqrt{h}}\right)^{2h} \qquad M = \sqrt{2}\mathrm{e}C.$$

If $|\tau| \leq 1$, fix an even $k \geq 2$ and set $h = k/2$, then the k-th and $(k+1)$-th terms of the first series

$$\left(\frac{|\tau|}{\sqrt{k}}\right)^k + \left(\frac{|\tau|}{\sqrt{k+1}}\right)^{k+1} \leq 2\left(\frac{|\tau|}{\sqrt{k}}\right)^k$$

can be bounded by the h-th term of the second series

$$\left(\frac{C|\tau|}{\sqrt{h}}\right)^{2h} = \left(\frac{\sqrt{2}C|\tau|}{\sqrt{k}}\right)^k,$$

provided that $C > 1$. If $|\tau| \geq 1$, fix an odd $k \geq 3$ and set $h = (k+1)/2$, then the k-th and $(k+1)$-th terms of the first series

$$\left(\frac{|\tau|}{\sqrt{k}}\right)^k + \left(\frac{|\tau|}{\sqrt{k+1}}\right)^{k+1} \leq 2\left(\frac{|\tau|}{\sqrt{k+1}}\right)^{k+1}$$

can be bounded by the h-th term of the second series

$$\left(\frac{C|\tau|}{\sqrt{h}}\right)^{2h} = \left(\frac{\sqrt{2}C|\tau|}{\sqrt{k+1}}\right)^{k+1},$$

provided that

$$C \geq \frac{2^{\frac{1}{4}}}{2}\sqrt{\frac{4}{3}} \geq \frac{2^{\frac{1}{k+1}}}{2}\sqrt{\frac{k+1}{k}},$$

which means that we can use $M = 4 \geq (\sqrt{2}2^{\frac{1}{4}}\mathrm{e})/\sqrt{3}$.

Finally, reasoning as in the proof of (38a) and by using (38b)

$$\begin{aligned}\mathbb{P}[\chi > \tau] &= P[\exp(c\chi) > \exp(c\tau)] \\ &\leq \mathbb{E}[\exp(c\chi)]\mathrm{e}^{-c\tau} \\ &\leq \exp(8c^2 - c\tau),\end{aligned}$$

which takes the minimum at $c = \tau/16$. Hence,

$$\mathbb{P}[\chi > \tau] = \exp(-\frac{\tau^2}{32}).$$

■

Remark 16 *Both* (38a) *and* (38b) *(for a suitable constants instead of* $\|\xi\|_{\psi_2}$*) are sufficient conditions for sub-gaussianity and* (38b) *implies that* $\mathbb{E}[\xi] = 0$*, see (Vershynin, 2012, Lemma 5.5).*

The following proposition bounds the Euclidean norm of a sub-gaussian vector. The proof is standard and essentially based on the results in Vershynin (2012), but we were not able to find the precise reference. The centered case is done in Rigolet (2015).

Proposition 17 *Let* ξ *a sub-gaussian random vector in* $\mathbb{R}^d$*. Given* $\tau > 0$ *with probability greater than* $1 - 2\mathrm{e}^{-\tau^2}$

$$\|\xi\| \leq 3\|\xi\|_{\psi_2}(\sqrt{6d} + 2\tau) \leq 9\|\xi\|_{\psi_2}(\sqrt{d} + \tau). \tag{39}$$

If $\mathbb{E}[\xi] = 0$*, then with probability greater than* $1 - \mathrm{e}^{-\tau^2}$

$$\|\xi\| \leq 8\|\xi\|_{\psi_2}(\sqrt{3d} + \sqrt{2}\tau) \leq 16\|\xi\|_{\psi_2}(\sqrt{d} + \tau). \tag{40}$$

Proof As usual we assume that $\|\xi\|_{\psi_2} = 1$. Let $\mathcal{N}$ be a 1/2-net of S^{d-1}. Lemmas 5.2 and 5.3 in Vershynin (2012) give

$$\|\xi\| \leq 2 \max_{v \in \mathcal{N}} \langle \xi, v \rangle \qquad |\mathcal{N}| \leq 5^d.$$

Fixed $v \in \mathcal{N}$, (38a) gives that

$$\mathbb{P}[|\langle \xi, v \rangle| > 3t] \leq 2 \exp(-t^2).$$

By union bound

$$\mathbb{P}[\|\xi\| > 6t] \leq \mathbb{P}[\max_{v \in \mathcal{N}} |\langle \xi, v \rangle| > 3t] \leq 2|\mathcal{N}| \exp(-t^2) \leq 2 \exp(\ln(5)d - t^2).$$

Setting $t = \sqrt{\tau} + \sqrt{3d/2}$, so that $t^2 - \ln(5) > \tau^2$, we prove the claim.

Assume that ξ is centered and use (38c) instead of (38a). Then

$$\mathbb{P}[\|\xi\| > 8\sqrt{2}t] \leq \mathbb{P}[\max_{v \in \mathcal{N}} |\langle \xi, v \rangle| > 3t] \leq |\mathcal{N}| \exp(-t^2) \leq \exp(\ln(5)d - t^2).$$

As above $t = \sqrt{\tau} + \sqrt{3d/2}$ provides the claim. ■

Remark 18 *Compare with Theorem 1.19 in Rigolet (2015), noting that by* (38b) *the parameter σ in Definition 1.2 of Rigolet (2015) is bounded by $4\|\xi\|_{\psi_2}$.*

The following result is a concentration inequality for the second momentum of sub-gaussian random vector, see Theorem 5.39 and Remark 5.40 in Vershynin (2012) and footnote 20.

Theorem 19 *Let $\xi \in \mathbb{R}^d$ be a sub-gaussian vector random vector in $\mathbb{R}^d$. Given a family $\xi_1, \ldots, \xi_n$ of random vectors independent and identically distributed as ξ, then for $\tau > 0$*

$$\mathbb{P}[\|\frac{1}{n} \sum_{i=1}^{n} \xi_i \otimes \xi_i - \mathbb{E}[\xi \otimes \xi]\| > \max\{\delta, \delta^2\}] \leq 2\mathrm{e}^{-\tau^2}$$

where

$$\delta = C_\xi \left(\sqrt{\frac{d}{n}} + \frac{\tau}{\sqrt{n}} \right),$$

and C_ξ is a constant depending only on the sub-gaussian norm $\|\xi\|_{\psi_2}$.

References

Philip M. Anselone. *Collectively compact operator approximation theory and applications to integral equations.* Prentice-Hall Inc., Englewood Cliffs, N. J., 1971.

Frank Bauer and Mark A. Lukas. Comparing parameter choice methods for regularization of ill-posed problems. *Math. Comput. Simul.*, 81(9):1795–1841, 2011.

Andrea Braides. Gamma-convergence for beginners. Lecture notes. Available on line http://www.mat.uniroma2.it/ braides/0001/dotting.html, 2001.

Antonin Chambolle, Ronald DeVore, Nam-yong Lee, and Bradley Lucier. Nonlinear wavelet image processing: variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Trans. Image Process.*, 7(3):319–335, 1998.

Guangliang Chen, Anna V. Little, and Mauro Maggioni. Multi-resolution geometric analysis for data in high dimensions. In *Excursions in harmonic analysis. Volume 1*, pages 259–285. New York, NY: Birkhäuser/Springer, 2013.

David Donoho and Iain Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–55, 1994.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems.* Dordrecht: Kluwer Academic Publishers, 1996.

Gitta Kutyniok and Demetrio Labate, editors. *Shearlets. Multiscale analysis for multivariate data.* Boston, MA: Birkhäuser, 2012.

Erich Novak and Henryk Woźniakowski. Optimal order of convergence and (in)tractability of multivariate approximation of smooth functions. *Constr. Approx.*, 30(3):457–473, 2009.

Philippe Rigolet. 18.S997: High dimensional statistics. Lecture notes. Available on line http://www-math.mit.edu/ rigollet/PDFs/RigNotes15.pdf, 2015.

Lorenzo Rosasco, Mikhail Belkin, and Ernesto De Vito. On learning with integral operators. *J. Mach. Learn. Res.*, 11:905–934, 2010.

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.*, 18(82):1–9, 2013.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1649–1656. MIT Press, Cambridge, MA, 2006.