Journal of Machine Learning Research 22 (2021) 1-59 Submitted 4/20; Revised 12/20; Published 1/21

Analyzing the discrepancy principle for kernelized spectral filter learning algorithms

Alain Celisse alain.celisse@univ-paris1.fr
Laboratoire SAMM
EA 4543, Université Paris 1 Panthéon-Sorbonne
90 rue de Tolbiac
75 634 Paris Cedex 13, France

Martin Wahl martin.wahl@math.hu-berlin.de
Institut für Mathematik
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany

Editor: Lorenzo Rosasco

Abstract

We investigate the construction of early stopping rules in the nonparametric regression problem where iterative learning algorithms are used and the optimal iteration number is unknown. More precisely, we study the discrepancy principle, as well as modifications based on smoothed residuals, for kernelized spectral filter learning algorithms including Tikhonov regularization and gradient descent. Our main theoretical bounds are oracle inequalities established for the empirical estimation error (fixed design), and for the prediction error (random design). From these finite-sample bounds it follows that the classical discrepancy principle is statistically adaptive for slow rates occurring in the hard learning scenario, while the smoothed discrepancy principles are adaptive over ranges of faster rates (resp. higher smoothness parameters). Our approach relies on deviation inequalities for the stopping rules in the fixed design setting, combined with change-of-norm arguments to deal with the random design setting.

Keywords: early stopping, discrepancy principle, non-parametric regression, spectral regularization, reproducing kernel Hilbert space, oracle inequality, effective dimension

1. Introduction

1.1 State-of-the-art

The present work addresses the problem of estimating a regression function in a nonparametric framework by means of iterative learning algorithms, which is an ubiquitous problem in the statistical and machine learning literature. Since it is out of the scope of the present introduction to review all of them, let us only mention a few contributions in machine learning such as the boosting strategies aiming at estimating a regression function from a set of weak learners by iteratively re-weighting them (Duffy and Helmbold, 2002; Bühlmann and Yu, 2003), or the more recent use of deep neural networks (Anthony and Bartlett, 1999; Goodfellow et al., 2016), where the iterative stochastic gradient descent algorithm is extensively applied (Jastrzebski et al., 2018; Li and Liang, 2018). Nonparametric regression is the topic of several monographs such as Györfi et al. (2002), Tsybakov (2009), or the more recent book by Giné and Nickl (2016) that provides a detailed account of classical techniques for the theoretical analysis of nonparametric models.

Our theoretical analysis applies to learning algorithms embedded in a reproducing kernel Hilbert space (RKHS) associated with a reproducing kernel (Aronszajn, 1950). Their use in machine learning traces back to Aizerman et al. (1964); Meisel (1969), and there is now an extensive literature on this topic. Among others, Cucker and Smale (2002) and Steinwart and Christmann (2008) describe the mathematical foundation of learning with reproducing kernels. Caponnetto and De Vito (2007) derive optimal convergence rates for the prediction error of the kernelized Tykhonov algorithm, while Jacot et al. (2018) connect the properties of a deep neural network during the training to a particular reproducing kernel called the neural tangent kernel (see Scholkopf and Smola (2001) and Shawe-Taylor and Cristianini (2004) for more applications of reproducing kernels).

The class of spectral filter algorithms (Bauer et al., 2007; Blanchard et al., 2018b; Lin et al., 2020) that is under consideration in the present work can be seen as a subset of the broader family of iterative algorithms. Iterative algorithms become ubiquitous in situations where some regularization is needed (Raskutti et al., 2014), or if no closed-form expressions are available for the estimator of interest. This typically arises for most of M-estimators (van der Vaart and Wellner, 1996) for which optimization algorithms such as gradient descent, coordinate descent, or Newton's method are used among others (Boyd and Vandenberghe, 2004). In practice using such iterative algorithms requires the knowledge of the best iteration number at which one should interrupt the process. This optimal iteration number actually reaches a crucial trade-off between the statistical precision output after some iterations and the computational resources induced by them. For instance, interrupting the process too early provides a poor statistical precision, whereas waiting for more iterations induces a higher computational price (and typically even worse performances) (Raskutti et al., 2014, Fig. 1).

The main focus here is given to the so-called *early stopping rules*, which are data-driven estimators of this usually unknown best iteration number. Designing such rules is all the more important as they are designed to output an efficient estimator while saving the computational resources. For instance, unlike Lepskii's method and similar model selection procedures (De Vito et al., 2010; Blanchard et al., 2019), early stopping rules avoid all pairwise comparisons between models, which turns out to be highly time consuming. The design and study of early stopping rules have received a lot of attention which can be traced back to the empirical work of Prechelt (1998) in the context of neural networks. A first line of research leads to *deterministic* stopping rules that only depend on the data through the sample size n and some smoothness parameters (see Zhang and Yu (2005) for the boosting, followed by Yao et al. (2007) and Lin et al. (2020) with spectral filter algorithms). A second strategy has been initiated by Raskutti et al. (2014) and then by Wei et al. (2019), which mainly relies on upper bounding with high probability the estimation error by means of the Rademacher complexity. The resulting stopping rules enjoy good convergence rates from an asymptotic perspective, but only depend on the data through the points of the design which limits their practical application. More recently, a new promising idea has been investigated by Blanchard et al. (2018b,a) in the context of the Gaussian sequence model where a stopping rule is suggested and analyzed which relies on the one hand on the discrepancy principle, and on the other hand on the estimation (rather than an upper bound) of the approximation error. While the resulting stopping rules still have some drawbacks compared to classical model selection procedures (such as Lepskii's method (Blanchard et al., 2019)) in terms of statistical optimality, they achieve good oracle properties in a computationally efficient way.

1.2 Contributions

From a practical perspective, our main contribution is the description of *data-driven* early stopping rules based on the discrepancy principle (Phillips, 1962). Unlike previous approaches, the dependence of our stopping rules with respect to the data is not limited to the sample size (Yao et al., 2007) nor to the design points (Raskutti et al., 2014; Wei et al., 2019). By contrast, the present work rather extends the results of Blanchard et al. (2018b,a) for inverse problems in the Gaussian sequence setting to the context of reproducing kernels and kernelized spectral filter estimators.

From a theoretical perspective our contributions are two-fold. On the one hand, we derive the first non-asymptotic theoretical analysis of these stopping rules applied to spectral filter algorithms combined with reproducing kernels. Firstly, this analysis relies on several new concentration inequalities in the fixed-design setting which lead to (non-asymptotic) oracle inequalities for two stopping rules based on the discrepancy principle. Secondly, we use a new change-of-norm argument which allow us to transfer these oracle inequalities to the random design setting. On the other hand, these finite-sample bounds from the random design case lead to establish that: (i) the classical discrepancy principle is statistically adaptive for slow rates occurring in the hard learning scenario (called outer case), and (ii) the smoothing-based discrepancy principles are adaptive over ranges of higher smoothness parameters (called inner case).

1.3 Outline

The remainder of the paper is organized as follows. Next Section 2 introduces the main notions used along the papers. It starts by describing the statistical model, the spectral filter learning algorithms, and reviewing previous works on optimal rates in the context of the present paper. The early stopping rule based on the discrepancy principle (DP) is then introduced and motivated in Section 2.4.

Our first main theoretical results are discussed in Section 3 which focuses on the DP stopping rule in the fixed-design setting. In particular, the main ingredients of the derivation are detailed in Section 3.1. The improved early stopping rule based on the smoothing of the residuals is then introduced and analyzed in Section 4 for the fixed-design case, while the random design framework is addressed in Section 5. A short illustration of the behaviour of the different stopping rules is provided in Section 6 by means of empirical simulations from synthetic data.

Finally, we provide proofs based on a unified analysis for both early stopping rules in Section 7 in the fixed-design, while proofs for the random design case are detailed in Section 8. The appendix collects some background material.

2. Spectral filters and discrepancy principle

2.1 Regression model and reproducing kernel

Let (X, Y) be a pair of random variables satisfying the regression equation

$$Y = f(X) + \epsilon, \tag{2.1}$$

where X is a random variable taking values in $\mathcal{X} \subseteq \mathbb{R}^d$, $f : \mathcal{X} \to \mathbb{R}$ is an unknown regression function, and ϵ is a real-valued random variable such that $\mathbb{E}(\epsilon|X) = 0$ and $\mathbb{E}(\epsilon^2|X) = \sigma^2$, with $\sigma^2 > 0$ *assumed to be known* as in Raskutti et al. (2014) for instance. Additionally, we suppose that ϵ is sub-Gaussian conditional on X, cf. Vershynin (2018).

Assumption 1 *There is a constant $A \geq 1$ such that*

$$\forall q \geq 1, \qquad q^{-1/2}(\mathbb{E}(|\epsilon|^q|X))^{1/q} \leq A\sigma. \tag{\textbf{SubGN}}$$

Let $k(\cdot,\cdot)$ be a continuous and positive kernel on $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathcal{H}$ be the reproducing kernel Hilbert space of k. We denote by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and $\|\cdot\|_{\mathcal{H}}$ the inner product in $\mathcal{H}$ and its corresponding norm. We also define the $\mathcal{H}$-valued random variable $k_X = k(X, \cdot)$ for which we make the following assumption.

Assumption 2 *There is a constant $M > 0$ such that*

$$\|k_X\|_{\mathcal{H}} \leq M \quad a.s. \tag{\textbf{BdK}}$$

For instance, (**BdK**) holds true if $\sup_{x\in\mathcal{X}} k(x, x) \leq M^2$ (from the reproducing property). This arises with any continuous kernel and a bounded domain $\mathcal{X}$, or with a bounded kernel and $\mathcal{X}$ unbounded (Gaussian kernel).

In particular, we can define the covariance operator

$$\Sigma = \mathbb{E}\left[\, k_X \otimes k_X \,\right],$$

where $a \otimes b \in \mathcal{L}(\mathcal{H})$ denotes the tensor product between elements $a, b \in \mathcal{H}$ such that $(a \otimes b)u = a\langle b, u\rangle_{\mathcal{H}}$, for every $u \in \mathcal{H}$. In the following, ρ represents the probability distribution of X, and

$$L_\rho : L^2(\rho) \to L^2(\rho), \quad L_\rho g(x) = \int k(x, y)g(y)\, d\rho(y)$$

denotes the integral operator associated with k and ρ. Let $\langle\cdot,\cdot\rangle_\rho$ and $\|\cdot\|_\rho$ denote the inner product in $L^2(\rho)$ and its corresponding norm. Under Assumption (**BdK**) we know that both, L_ρ and Σ are positive self-adjoint trace-class operators. Moreover, both operators L_ρ and Σ are intimately related, which can be seen by introducing the inclusion operator $S_\rho : \mathcal{H} \to L^2(\rho)$, mapping $h \in \mathcal{H}$ to its equivalence class in $L^2(\rho)$ (S_ρ is well-defined, because under Assumption (**BdK**) every $h \in \mathcal{H}$ is bounded a.s.). Then it is well-known that

$$S_\rho S_\rho^* = L_\rho \in \mathcal{L}(L^2(\rho)), \qquad S_\rho^* S_\rho = \Sigma \in \mathcal{L}(\mathcal{H}),$$

where S_ρ^* is the adjoint operator of S_ρ. For these and more information on the learning with kernels setting (see e.g. Cucker and Smale (2002) and De Vito et al. (2005)). By the spectral theorem, there exists a sequence $\lambda_1 \geq \lambda_2 \geq \cdots > 0$ of positive eigenvalues (which is either finite or converges to zero), together with an orthonormal system $u_1, u_2, \ldots$ of eigenvectors of the range of L_ρ such that $\Sigma = \sum_{j\geq 1} \lambda_j u_j \otimes u_j$.

We will assume that f satisfies a polynomial source condition (see Chap. 4 in Lu and Pereverzev (2013)) that is,

Assumption 3 *For some $r \geq 0$ and $R > 0$, we have*

$$f = L_\rho^r g, \quad \text{with } g \in L^2(\rho) \text{ and } \|g\|_\rho \leq R. \tag{$\mathbf{SC}$(r,R)}$$

Note that such source conditions are often written as $\|L_\rho^{-r} f\|_\rho \leq R$; see e.g. Smale and Zhou (2007).

Remark 1 (Inner and Outer cases) *On the one hand, if $r \geq 1/2$, then*

$$f = L_\rho^r g = S_\rho \Sigma^{r-1/2} \Sigma^{-1/2} S_\rho^* g = S_\rho f_{\mathcal{H}}, \tag{2.2}$$

where $f_{\mathcal{H}} = \Sigma^{r-1/2}(\Sigma^{-1/2} S_\rho^* g) \in \mathcal{H}$. This means that f (resp. its equivalence class) can be represented (through the inclusion operator S_ρ) as a function in $\mathcal{H}$. This case is then called the inner case*. Let us mention that one also recovers an alternative formulation of the source condition when $r \geq 1/2$ that is,*

$$f_{\mathcal{H}} = \Sigma^s h, \quad \text{where } h \in \mathcal{H} \text{ and } \|h\|_{\mathcal{H}} \leq R,$$

with $s = r - 1/2 \geq 0$ and $h = \Sigma^{-1/2} S_\rho^* g \in \mathcal{H}$, where $\|h\|_{\mathcal{H}} = \|\Sigma^{-1/2} S_\rho^* g\|_{\mathcal{H}} = \|g\|_\rho \leq R$. These results can be found in Cucker and Smale (2002), where it is shown how to characterize $\mathcal{H}$ through the eigenvalues of L_ρ.

On the other hand, if $r < 1/2$, then f can not be represented as a function in $\mathcal{H}$ in general, which justifies referring to this situation as the outer case.

In what follows, the outer and inner cases are respectively considered in Section 5.2 and Section 5.3.

We suppose that we observe n independent copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y). Let $K_n \in \mathbb{R}^{n\times n}$ be the kernel matrix defined by $(K_n)_{ij} = k(X_i, X_j)/n$ and Σ_n be the empirical covariance operator defined by

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^n k_{X_i} \otimes k_{X_i}.$$

Both operators K_n and Σ_n are strongly related, as can be seen by introducing the sampling operator S_n defined by $S_n : \mathcal{H} \to \mathbb{R}^n, h \mapsto (h(X_i))_{i=1}^n$ and its adjoint operator S_n^*, where $\mathbb{R}^n$ is endowed with the empirical inner product $\langle \cdot, \cdot \rangle_n$ and its corresponding *empirical norm* $\| \cdot \|_n$ such that $\langle a, b \rangle_n = (1/n) \sum_{i=1}^n a_i b_i$ and $\|a\|_n = \sqrt{\langle a, a \rangle_n}$ for every $a, b \in \mathbb{R}^n$. Then we have

$$S_n S_n^* = K_n, \qquad S_n^* S_n = \Sigma_n.$$

By the spectral theorem, there exists a sequence $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n$ of non-negative eigenvalues, together with an orthonormal system $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n$ in $\mathcal{H}$ and an orthonormal basis $\hat{v}_1, \ldots, \hat{v}_n$ of $(\mathbb{R}^n, \langle \cdot, \cdot \rangle_n)$ such that

$$S_n = \sum_{j=1}^{n} \hat{\lambda}_j^{1/2} \hat{v}_j \otimes \hat{u}_j. \tag{2.3}$$

In particular, we have $\Sigma_n = \sum_{j=1}^{n} \hat{\lambda}_j \hat{u}_j \otimes \hat{u}_j$ and $K_n = \sum_{j=1}^{n} \hat{\lambda}_j \hat{v}_j \hat{v}_j^T$. We write $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Moreover, for a function $g : \mathcal{X} \to \mathbb{R}$, we write $\mathbf{g} = (g(X_1), \ldots, g(X_n))^T \in \mathbb{R}^n$ for its evaluation at the design points. In particular, if $g \in \mathcal{H}$, then the function g and its bold version $\mathbf{g}$ are linked through the relation $\mathbf{g} = S_n g$. However note that if $g \notin \mathcal{H}$, then the latter equality is meaningless, which justifies the use of $\mathbf{g}$.

2.2 Spectral filter learning algorithms

Let us consider the problem of estimating f by means of spectral filter learning algorithms (see e.g. Bauer et al. (2007); Lu and Pereverzev (2013); Blanchard and Mücke (2018) and Lin et al. (2020)). For a function $g : [0, M^2] \times [0, \infty) \to \mathbb{R}$, let us write $g_t(\lambda) = g(\lambda, t)$.

Definition 2 (Regularizer) *A function $g : [0, M^2] \times [0, \infty) \to \mathbb{R}$ is called a regularizer if $(\lambda, t) \mapsto \lambda g_t(\lambda)$ is non-decreasing in t and λ, continuous in t, with $g_0(\lambda) = 0$ and $\lim_{t \to +\infty} \lambda g_t(\lambda) = 1$ for $\lambda > 0$, and if there is a constant $B > 0$ such that*

(i) For all $(\lambda, t) \in [0, M^2] \times [0, \infty)$, we have $0 \leq \lambda g_t(\lambda) \leq 1$; **(BdF)**

(ii) For all $(\lambda, t) \in [0, M^2] \times [0, \infty)$, we have $g_t(\lambda) \leq Bt$. **(LFU)**

Our definition of a regularizer is slightly stronger than the one used in Definition 1 of Bauer et al. (2007) or in Definition 2.13 of Blanchard and Mücke (2018). This owes to our continuity assumption with respect to t, which excludes the spectral cut-off algorithm (corresponding to the choice $g_t(\lambda) = \mathbb{1}_{(\lambda t \geq 1)}/\lambda$) from the present study. The continuity is not essential in our derivation but it greatly simplifies the analysis because it leads to continuous bias and variance terms as well. Note that Blanchard et al. (2018a) also originally derived results relying on continuity, and then extended them to the spectral cut-off algorithm in a second step (Blanchard et al. (2018b)).

Definition 3 (Spectral filter estimators) *For a given regularizer $g : [0, M^2] \times [0, \infty) \to \mathbb{R}$, a* spectral filter estimator *is an estimator given by*

$$\hat{f}^{(t)} = g_t(\Sigma_n) S_n^* \mathbf{Y}, \quad t \geq 0.$$

By (**BdK**), we have that $\max(\lambda_1, \hat{\lambda}_1) \leq M^2$ almost surely. This implies that the estimators $\hat{f}^{(t)}$ are indeed well-defined. The following examples provide several choices of spectral filter algorithms and regularizers.

Example 1 *The choice $g_t(\lambda) = (\lambda + t^{-1})^{-1}$ corresponds to Tikhonov regularization and Definition 2 holds with $B = 1$.*

Example 2 *Gradient descent with constant step size $\eta \in (0, 1/M^2)$ (also called Landweber iteration) corresponds to the sequence of iterations*

$$\hat{f}^{(0)} = 0, \quad \hat{f}^{(t)} = \hat{f}^{(t-1)} + \eta S_n^*(\mathbf{Y} - S_n \hat{f}^{(t-1)}), \quad t = 1, 2, \ldots.$$

It has the closed-form expression $\hat{f}^{(t)} = g_t(\Sigma_n) S_n^* \mathbf{Y}$ with $g_t(\lambda) = \lambda^{-1}(1 - (1 - \eta\lambda)^t)$. Interpolating, we may consider $g_t(\lambda) = \lambda^{-1}(1 - (1 - \eta\lambda)^t)$ for $t \geq 1$, and $g_t(\lambda) = \eta t$ for $t < 1$. In this case, Definition 2 holds with $B = \eta$.

Example 3 *The choice $g_t(\lambda) = \lambda^{-1}(1 - e^{-t\lambda})$ corresponds to Showalter's method. In this case, Definition 2 holds with $B = 1$.*

At some places, an additional assumption will turn to be useful in the analysis of spectral filter algorithms. It lower bounds the regularizer.

Assumption 4 *There is a constant $b > 0$ such that*

$$\text{for all } (\lambda, t) \in [0, M^2] \times [0, \infty), \text{ we have } \lambda g_t(\lambda) \geq b(1 \wedge \lambda t). \qquad (\mathbf{LFL})$$

For instance, this latter assumption holds true with Tikhonov regularization, Gradient descent and Showalter's method with $b = 1/2$.

Finally, when dealing with rates of convergence we will also need the following assumption on the qualification error.

Assumption 5 *There are constants $q, Q > 0$ such that*

$$\text{for all } (\lambda, t) \in [0, M^2] \times [0, \infty), \text{ we have } |r_t(\lambda)| \leq Q(\lambda t)^{-q}, \qquad (\mathbf{QuErr})$$

with $r_t(\lambda) = 1 - g_t(\lambda)\lambda$.

Remark 4 *Combining (**QuErr**) with (**BdF**), we have $r_t(\lambda) \leq 1 \wedge Q(t\lambda)^{-q}$ and thus also $r_t(\lambda) \leq 1 \wedge Q(t\lambda)^{-p}$ for each $p \leq q$, provided that $Q \geq 1$.*

It is well-known that Tikhonov regularization and gradient descent satisfy (**QuErr**) with respectively $q = 1$ and q arbitrary; see e.g. Blanchard and Mücke (2018) for more discussion.

Let us also introduce the g-effective dimension, which generalizes the classical notion of effective dimension (Zhang, 2003) to the case where g is not limited to the Tikhonov regularization.

Definition 5 (g-Effective dimension) *For every $t \geq 0$ and any regularizer g, the (population) g-effective dimension is defined by $\mathcal{N}^g(t) = \operatorname{tr}(\Sigma g_t(\Sigma))$, while the empirical effective dimension is defined by $\mathcal{N}_n^g(t) = \operatorname{tr}(\Sigma_n g_t(\Sigma_n))$.*

With Tikhonov regularization, that is $g_t(\lambda) = (\lambda + 1/t)^{-1}$, both the population and empirical g-effective dimension simply reduce to the usual population and empirical effective dimensions respectively given by $\mathcal{N}(t) = \operatorname{tr}(\Sigma(\Sigma + 1/t)^{-1})$ and $\mathcal{N}_n(t) = \operatorname{tr}(\Sigma_n(\Sigma_n + 1/t)^{-1})$. Note that most cited references consider the parameterization $\eta = t^{-1}$, i.e. they write $g_\eta(\lambda)$ and $\mathcal{N}(\eta)$ instead of $g_t(\lambda)$ and $\mathcal{N}(t)$ as in the present paper. Interestingly, it turns out that the effective and g-effective dimensions are closely related up to multiplicative constants as established by the next result.

Lemma 6 *Let g be a regularizer satisfying (**LFL**). Then for each $t \geq 0$,*

$$b\mathcal{N}_n(t) \leq \mathcal{N}_n^g(t) \leq 2(B \vee 1)\mathcal{N}_n(t).$$

Proof of Lemma 6 By (**BdF**) and (**LFU**) we have

$$\mathcal{N}_n^g(t) \leq (B \vee 1) \sum_{j=1}^{n} 1 \wedge \hat{\lambda}_j t \leq 2(B \vee 1) \sum_{j=1}^{n} \frac{\hat{\lambda}_j t}{\hat{\lambda}_j t + 1} = 2(B \vee 1)\mathcal{N}_n(t),$$

which gives the upper bound. The lower bound follows from (**LFL**) and the fact that $\lambda t/(\lambda t + 1) \leq 1 \wedge \lambda t$. □

2.3 Convergence rates in related works

The use of kernel-based spectral regularization in random regression problems (also known as "learning from examples") has been extensively studied in the literature; see e.g. Smale and Zhou (2005, 2007); Caponnetto and De Vito (2007) for Tikhonov regularization, Yao et al. (2007); Blanchard and Krämer (2016) for gradient descent methods and Bauer et al. (2007); Blanchard and Mücke (2018); Lin et al. (2020); Kriukova et al. (2016) for general spectral regularization schemes. Existing bounds are mostly established for the $L^2(\rho)$-error and the $\mathcal{H}$-error under (**SC**(r,R)) and a polynomial upper bound on the eigenvalues of L_ρ. They are usually used to construct deterministic early stopping rules (depending on the smoothness r and the eigenvalue decay); see e.g. Yao et al. (2007) for gradient descent, Blanchard and Krämer (2016) for conjugate gradient descent and Pillaud-Vivien et al. (2018) for stochastic gradient descent.

Surprisingly, while the inner case $r \geq 1/2$ is now well understood (Blanchard and Mücke, 2018; Lin et al., 2020), there remain some unsolved issues related to the outer case. The main difficulties arise in case of the so-called hard learning problems for which the optimal rates are achieved for very small regularization parameters (resp. a very large number of iterations, considerably exceeding the number of observations). In this direction, some improvements have been established e.g. in Fischer and Steinwart (2020); Pillaud-Vivien et al. (2018), based on more precise concentration inequalities for the eigenvalues of the kernel matrix (see Theorem 18).

Progress has also been made in the study of data-driven regularization parameter selection rules. Hold-out (that is, splitting the data into a training set and a validation set) and more general cross-validation procedures have been studied in Caponnetto and Yao (2010); Steinwart and Christmann (2008). Lepskii's balancing principle has been extended to the learning framework in De Vito et al. (2010); Lu and Pereverzev (2013); Blanchard et al. (2019). While the estimators from De Vito et al. (2010); Lu and Pereverzev (2013) are only adaptive with respect to the smoothness r, the estimator from Blanchard et al. (2019) achieves faster rates by also being adaptive with respect to the eigenvalue decay of the kernel integral operator. In slightly different directions, Page and Grünewälder (2018) studies the Goldenshluger-Lepskii method in a reproducing kernel framework, and Brunel et al. (2016) studies model selection for principal component regression in a functional regression model. While all these methods share good oracle properties (and thus minimax adaption over suitable smoothness classes), they all put no attention on computational issues. In fact, they require that all estimators up to some threshold have to be computed before a parameter with close-to-optimal performance is chosen.

In contrast, the question of data-driven early stopping rules remains widely open. Raskutti et al. (2014) suggest an early stopping rule for gradient descent that is adaptive to the decay rate of the eigenvalues but not to the smoothness r (assumed to be $r = 1/2$). They study the solution of a fixed-point equation corresponding to a bias-variance trade-off of the empirical norm and show that this rule leads to optimal rates for the prediction error. These results have been extended in Wei et al. (2019) to the L^2-boosting based on different loss functions. Our goal is to develop *data-driven* stopping rules based on the discrepancy principle which are *statistically adaptive* with respect to both the smoothness parameter r and the eigenvalue decay.

2.4 Early stopping and discrepancy principle: Motivation

As explained in the introduction, our goal is to make use of the discrepancy principle (DP) to find a value t having small excess risk. One of its main merits is the fact that it allows to start a search for the optimal "regularization parameter" from the easiest problem (Mathé and Pereverzyev, 2006). The discrepancy principle has been extensively studied in the context of inverse problems with deterministic noise, where it is also called Morozov's discrepancy principle, see e.g. Phillips (1962); Morozov (1966); Engl et al. (1996). Using $\hat{\mathbf{f}}^{(t)} = S_n g_t(\Sigma_n) S_n^* \mathbf{Y} = K_n g_t(K_n)\mathbf{Y}$ from Definition 3 with regularizer g from Definition 2, it is based on a comparison of the empirical risk $\|\mathbf{Y} - \hat{\mathbf{f}}^{(t)}\|_n^2$ (also called squared discrepancy or squared residual) with the noise level $\mathbf{E}_\epsilon \|\boldsymbol{\epsilon}\|_n^2 = \sigma^2$, where $\mathbf{E}_{\boldsymbol{\epsilon}}(\cdot) = \mathbb{E}(\cdot | X_1, \ldots, X_n)$ denotes the expectation with respect to $(X_1, Y_1), \ldots, (X_n, Y_n)$ conditional on the design $X_1, \ldots, X_n$). It then advocates taking a value t for which both quantities are of comparable size.

The discrepancy principle can also be motivated by considering the expected empirical risk $\mathbf{E}_\epsilon \|\mathbf{Y} - \hat{\mathbf{f}}^{(t)}\|_n^2 = \mathbf{E}_\epsilon \|r_t(K_n)\mathbf{Y}\|_n^2$. The first step consists in noticing that we have the following bias-variance decomposition of the excess risk

$$\mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2 = \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n} \operatorname{tr}(g_t^2(K_n) K_n^2).$$

Using (**BdF**) this identity implies

$$\mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2 \leq \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n} \mathcal{N}_n^g(t) \tag{2.4}$$

with g-effective dimension $\mathcal{N}_n^g(t)$.

The second step exploits Lemma 22 below, which reveals a close relation to (2.4) by showing that

$$\|r_t(K_n)\mathbf{f}\|_n^2 - 2\frac{\sigma^2}{n}\mathcal{N}_n^g(t) \leq \mathbf{E}_\epsilon \|r_t(K_n)\mathbf{Y}\|_n^2 - \sigma^2 \leq \|r_t(K_n)\mathbf{f}\|_n^2 - \frac{\sigma^2}{n}\mathcal{N}_n^g(t). \tag{2.5}$$

In particular, by defining $t_0 \geq 0$ such that

$$t_0 = \inf\left\{t \geq 0 \mid \mathbf{E}_\epsilon \|r_t(K_n)\mathbf{Y}\|_n^2 = \sigma^2\right\}, \tag{2.6}$$

it follows from (2.4) and (2.5) that

$$\mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t_0)}\|_n^2 \leq 3 \min_{t \geq 0} \left\{ \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t) \right\}, \tag{2.7}$$

where we also used that $\|r_t(K_n)\mathbf{f}\|_n^2$ and $\mathcal{N}_n^g(t)$ are respectively non-increasing and non-decreasing with respect to $t \geq 0$ (see Figure 1a). Let us mention that Ineq. (2.7) is called an *oracle-type* inequality in what follows. Similarly, we also have the next lower bound

$$\mathbf{E}_\epsilon\|\mathbf{f} - \hat{\mathbf{f}}^{(t_0)}\|_n^2 \geq \|r_{t_0}(K_n)\mathbf{f}\|_n^2 \geq \frac{1}{2}\min_{t\geq 0}\left\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t)\right\}.$$

The third step relies on the important consequence that these upper and lower bounds indicate that t_0 defined by Eq. (2.6) is the best choice (up to the proxy variance term and constants) one can make for *stopping early* the estimation process. In particular, this justifies the introduction of the following early stopping rule based on the discrepancy principle (DP), which should be seen as the empirical counterpart of Eq. (2.6).

Definition 7 (DP stopping rule) *For any estimator $\hat{f}^{(t)} = g_t(\Sigma_n)S_n^*\mathbf{Y}$ given by Definition 3, the DP-based stopping rule τ_{DP} is defined by*

$$\tau_{DP} = \tau_{DP}(\mathbf{Y}, \sigma^2, T) = \inf\{t \geq 0 : \|\mathbf{Y} - S_n\hat{f}^{(t)}\|_n^2 \leq \sigma^2\} \wedge T, \tag{2.8}$$

with the "emergency stop" $T \in [0, \infty]$.

In the context of inverse problems, see also Blanchard and Mathé (2012) with the conjugate gradient and Blanchard et al. (2018b) with the spectral cut-off.

The above definition depends on two parameters, the emergency stop T and the true noise level σ^2. In particular, in what follows, we assume that the noise variance σ^2 is known in order to avoid further technicalities. Yet from a practical perspective, it is still possible to plug an estimator $\hat{\sigma}^2$ of σ^2 in the above definition. For this purpose, many different estimators have been introduced in the literature Spokoiny (2002); Cai et al. (2009); Liitiäinen et al. (2010); Devroye et al. (2018). For a theoretical perspective, the analysis of the resulting fully data-driven stopping rule could be extended from the present material using the arguments outlined in Remark 26 below.

From Definition 2, the fact that $\lim_{t\to+\infty}\lambda g_t(\lambda) = 1$ implies that the empirical risk $\|\mathbf{Y} - S_n\hat{f}^{(t)}\|_n^2 = \|r_t(K_n)\mathbf{Y}\|_n^2$ converges to zero as $t \to +\infty$. This entails that the choice $T = \infty$ is admissible as well since we will interrupt the iterations after a finite number of them.

2.5 Further notation

The abbreviation $\mathbf{E}_{\boldsymbol{\epsilon}}(\cdot) = \mathbb{E}(\cdot|X_1, \ldots, X_n)$ denotes the expectation with respect to $(X_1, Y_1), \ldots, (X_n, Y_n)$ conditional on the design $X_1, \ldots, X_n$. This means a slight abuse of notation because in the present context, the distribution of ϵ_i is allowed to depend on X_i. We also write $\mathbf{P}_{\boldsymbol{\epsilon}}(\cdot) = \mathbb{P}(\cdot|X_1, \ldots, X_n)$.

Given a bounded operator A on $\mathcal{H}$ or a matrix $A \in \mathbb{R}^{n\times n}$, we write $\|A\|_{\mathrm{op}}$ for the operator norm. Given a Hilbert-Schmidt operator A on $\mathcal{H}$ or a matrix $A \in \mathbb{R}^{n\times n}$, we write $\|A\|_{\mathrm{HS}}$ for the Hilbert-Schmidt or Frobenius norm. Given a trace class operator A on $\mathcal{H}$ or a matrix $A \in \mathbb{R}^{n\times n}$, we denote the trace of A by $\mathrm{tr}(A)$.

Throughout the paper, we use the letters c, C for constants that may change from line to line. They are allowed to depend on A, B, b, Q, R, M and r. Apart from these dependencies, the constants are absolute and can be made explicit by considering the proofs. In Sections 5 and 8 they are also allowed to depend on L, α (introduced therein) and σ^2. Finally for any $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For $a \geq 0$ we denote by $\lfloor a \rfloor$ the largest natural number that is smaller than or equal to a.

3. DP and oracle inequality: Fixed-design

The goal of this section is to assess the statistical performance of the final estimator $\hat{f}^{(\tau_{DP})}$, where τ_{DP} is the early stopping rule defined by Eq. (2.8) and derived from the discrepancy principle (DP). We start by introducing new deviation inequalities for τ_{DP} and for bias and variance terms (Propositions 8 and 9), leading then to oracle-type inequalities (Proposition 10 and Theorem 12).

3.1 Preliminary results

3.1.1 Deviation inequalities for DP and main arguments

Our main deviation inequalities for the early stopping rules are developed in Section 7. For the sake of simplifications, let us specialize them to the classical discrepancy principle τ_{DP} with $T = \infty$. For this, we abbreviate the squared bias and the *proxy variance* as

$$b_t^2 = \|r_t(K_n)\mathbf{f}\|_n^2 \quad \text{and} \quad v_t = \frac{\sigma^2}{n}\mathcal{N}_n^g(t), \tag{3.1}$$

where $\mathcal{N}_n^g(t)$ denotes the empirical g-effective dimension from Definition 5. Moreover, we introduce the important balancing stopping rule

$$t_n^* = \inf\{t \geq 0 : b_t^2 = v_t\}.$$

For simplicity, we assume throughout Section 3.1 that such a t exists, meaning that $b_{t_n^*}^2 = v_{t_n^*}$, the general case is treated in Section 7. We start with a right-deviation inequality for τ_{DP} that can be alternatively expressed in terms of the proxy variance v_t.

Proposition 8 *If Assumption* (**SubGN**) *holds, then there is a constant $c > 0$ depending only on A such that for every $t > t_n^*$,*

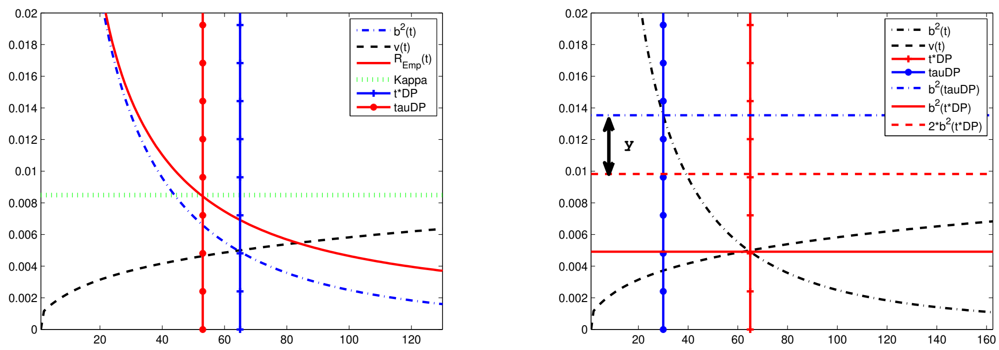
$$\mathbf{P}_\epsilon(\tau_{DP} > t) \leq 2\exp\Big(-cn\Big(\frac{y}{\sigma^2} \wedge \frac{y^2}{\sigma^4}\Big)\Big), \qquad y = v_t - v_{t_n^*}.$$

In particular, for every $y > 0$ we have

$$\mathbf{P}_\epsilon(v_{\tau_{DP}} > v_{t_n^*} + y) \leq 2\exp\Big(-cn\Big(\frac{y}{\sigma^2} \wedge \frac{y^2}{\sigma^4}\Big)\Big).$$

Both deviation inequalities are even equivalent if the proxy variance is strictly increasing. Proposition 8 is a simplified version of Proposition 24 below. The proof can be based on exploring Figure 1a in combination with concentration inequalities for the empirical risk. Here is an outline of the argument. Let us also mention that $t \mapsto b_t^2$ is continuous and non-increasing, while $t \mapsto v_t$ is continuous and non-decreasing. The definition of τ_{DP} yields $\mathbf{P}_\epsilon(\tau_{DP} > t) = \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 > \sigma^2)$. Subtracting $\mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2$ on both sides and invoking the upper bound in (2.5), we arrive at

$$\mathbf{P}_\epsilon(\tau_{DP} > t) \leq \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2 > v_t - b_t^2).$$



(a) (b)

Figure 1: Comparison of τ_{DP} and the balancing stopping time t_n^*. **(a):** The horizontal line defines $\kappa = \sigma^2$. The red plain decreasing curve crosses the horizontal line at τ_{DP}. The increasing curve crosses the blue dotted-dashed curve of the bias at t_n^*. **(b):** Illustration of Proposition 9. The red dashed horizontal line highlights the $2b^2(t_n^*)$ threshold to which $b^2(\tau_{DP})$ is compared.

By definition we have $v_t = v_{t_n^*} + y$. Moreover, from Figure 1a and the assumption on y, we get $b_t^2 \leq b_{t_n^*}^2 = v_{t_n^*}$. Hence, we conclude that

$$\mathbf{P}_\epsilon(\tau_{DP} > t) \leq \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2 > y),$$

and Proposition 8 follows from the Hanson-Wright inequality (see Lemma 27 below) and the fact that $b_t^2 \leq v_t \leq \sigma^2$ since $t \geq t_n^*$.

Next, we present a left-deviation inequality for τ_{DP} expressed in terms of the squared bias.

Proposition 9 *Suppose that Assumptions* (**SubGN**) *and* (**BdK**) *hold true. Then, for every $y > 0$, we have*

$$\mathbf{P}_\epsilon(b_{\tau_{DP}}^2 > 2b_{t_n^*}^2 + y) \leq 2\exp\Big(-cn\Big(\frac{y}{\sigma^2} \wedge \frac{y^2}{\sigma^4}\Big)\Big),$$

where $c > 0$ is a constant depending only on A.

Proposition 9 is a simplified version of Proposition 25 below, and follows similarly as Proposition 8 by exploiting the lower bound in (2.5) this time. As illustrated in Figure 1b, let $t < t_n^*$ be defined by $b_t^2 = 2b_{t_n^*}^2 + y$ (if such a t does not exist, then the claim is trivial). Then the definition of τ_{DP} yields $\mathbf{P}_\epsilon(b_{\tau_{DP}}^2 > b_t^2) \leq \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 \leq \sigma^2)$. Subtracting $\mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2$ on both sides and invoking the lower bound in (2.5), we arrive at

$$\mathbf{P}_\epsilon(b_{\tau_{DP}}^2 > 2b_{t_n^*}^2 + y) \leq \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2 \leq 2v_t - b_t^2).$$

By definition we have $b_t^2 = 2b_{t_n^*}^2 + y$. Moreover, from Figure 1b and the assumption on y, we get $v_t \leq v_{t_n^*} = b_{t_n^*}^2$. Hence, we conclude that

$$\mathbf{P}_\epsilon(b_{\tau_{DP}}^2 > 2b_{t_n^*}^2 + y) \leq \mathbf{P}_\epsilon(\|r_t(K_n)\mathbf{Y}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\mathbf{Y}\|_n^2 \leq -y),$$

and Proposition 9 follows from the Hanson-Wright inequality (see Lemma 27 below) and the fact that $b_t^2 \leq 2v_{t_n^*} + y \leq 2\sigma^2 + y$.

3.1.2 Non-asymptotic performance quantification

We are now in position to formulate our first upper bound for the estimation error in the empirical norm. It quantifies the statistical performance of the stopping rule based on the classical discrepancy principle (DP), namely τ_{DP}, in terms of an oracle-type inequality with high probability.

Proposition 10 *Suppose that Assumptions* (**SubGN**) *and* (**BdK**) *hold. Then the early stopping rule* τ_{DP} *based on the standard discrepancy principle* (2.8) *satisfies for each* $T \in [0, \infty]$ *and for every* $u > 0$*,*

$$\mathbf{P}_\epsilon\Big(\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{DP})}\|_n^2 \; > \; C\Big(\min_{0 \leq t \leq T} \Big\{ \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t) \Big\} + \frac{\sigma^2 \sqrt{u}}{\sqrt{n}} + \frac{\sigma^2 u}{n} \Big) \Big) \; \leq \; 5e^{-u},$$

where C is a constant depending only on A.

A proof of Proposition 10 is given in Section 7.3. The above result is established for spectral filter estimators with regularizer g, under mild assumptions on the noise (only required to be sub-Gaussian). Deriving this result under such mild assumptions has been made possible by introducing the proxy variance $v_t = \sigma^2 \mathcal{N}_n^g(t)/n$ (from Eq. (3.1)) instead of the more classical variance term in the r.h.s. of the inequality. Nevertheless, it is still possible to upper bound the proxy-variance by the classical one at the price of an additional assumption as will be done in the next section (Theorem 12).

3.2 Main oracle inequality

As explained earlier, the purpose of the present section is to establish an oracle inequality for τ_{DP}. Compared with Proposition 10, this is possible at the price of an additional assumption that we first motivate.

The desired derivation is made possible by connecting the proxy variance (that is, the g-effective dimension) to the classical variance. The key ingredient is that the g-effective dimension is typically dominated by the eigenvalues satisfying $t\hat{\lambda}_j > 1$ as highlighted by the proof of Lemma 11. For such eigenvalues, (**LFL**) yields $b \leq \hat{\lambda}_j g_t(\hat{\lambda}_j) \leq 1$, which leads to conclude that the proxy and true variances only differ by a constant. This argument can be made rigorous by means of the next (sufficient) condition.

Assumption 6 *There is a constant $E > 0$ such that for $k = 0$ and each $k \geq 1$ satisfying $\hat{\lambda}_k T \geq 1$, we have*

$$\hat{\lambda}_{k+1}^{-1} \sum_{j > k} \hat{\lambda}_j \leq E(k \vee 1). \tag{\textbf{EVBound}}$$

Considering this ratio between the tail series of eigenvalues and the kth largest one has already been made in the literature (see Definition 3 in Bartlett et al., 2019, for instance where this ratio is named the "effective rank"). It is noticeable that (**EVBound**) encompasses two classical assumptions on the decay rate of the eigenvalues, respectively called polynomial (**PolDecTS**) and exponential (**ExpDecTS**) decay.

Example 4 (Polynomial eigenvalues decay) *If there exist numeric constants $\ell, L > 0$, and $\alpha > 1$ such that*

$$\ell j^{-\alpha} \leq \hat{\lambda}_j \leq L j^{-\alpha}, \qquad 1 \leq j \leq n, \tag{\textbf{PolDecTS}}$$

then (**EVBound**) *holds true with $E = 1 + 2L\ell^{-1}(\alpha+1)^{-1}$.*

Example 5 (Exponential eigenvalues decay) *If there exist numeric constants $\ell, L > 0$, and $\alpha \in (0, 1]$ such that*

$$\ell e^{-j^{\alpha}} \leq \hat{\lambda}_j \leq L e^{-j^{\alpha}}, \qquad 1 \leq j \leq n, \tag{\textbf{ExpDecTS}}$$

then (**EVBound**) *holds true with*

$$E = 1 + \frac{2L}{\ell\alpha} \int_0^{\infty} (1+v)^{1/\alpha - 1} e^{-v} \, dv.$$

The previous two examples are provided for illustrative purposes only. A more general result will be proved under milder constraints on the empirical eigenvalues by means of (**EffRank**) combined with Lemma 35 which avoids requiring that (**PolDecTS**) or (**ExpDecTS**) hold true for all indices $1 \leq j \leq n$ (see Section 5.3 for more details).

The above assumptions are expressed in terms of the empirical eigenvalues of the Gram matrix. However such assumptions are usually easier to check for the population eigenvalues of the kernel operator. In our analysis, Lemmas 33 and 44 then translate bounds for the population eigenvalues of K into properties of the empirical eigenvalues of K_n. For instance, kernels with a polynomial decay are discussed in Raskutti et al. (2014) (Sobolev kernel), and Steinwart et al. (2009) (m-times differentiable kernels on Euclidean balls of $\mathbb{R}^d$). See also Steinwart and Christmann (2008) for a more extensive review of connections between the population eigenvalue decay and the entropy number of the corresponding RKHS. Typical kernels with an exponential decay are smooth radial kernels. For instance, Belkin (2018) derives tight exponential bounds (independent of any reference measure) on the population eigenvalue decay of smooth radial kernels.

We are now in position to explain how $\mathcal{N}_n^g(t)$ (resp. the proxy variance) connects to $\mathrm{tr}(g_t^2(K_n)K_n^2)$ (resp. the variance) by means of (**EVBound**).

Lemma 11 *Suppose that Assumptions* (**LFL**) *and* (**EVBound**) *hold. Then there is a constant $C > 0$ depending only on B, b and E such that*

$$\forall 0 \leq t \leq T, \quad \mathcal{N}_n^g(t) \leq C(\mathrm{tr}(g_t^2(K_n)K_n^2) + 1).$$

For the sake of comparison, let us mention that Lemma 11 shows that the constant C_{l^1,l^2} from Proposition 2.5 in (Blanchard et al., 2018a) does exist under mild assumptions on the decay rate of the eigenvalues.

Proof of Lemma 11 If $t\hat{\lambda}_1 < 1$, then (**LFU**) and (**EVBound**) imply

$$\mathcal{N}_n^g(t) \leq Bt \sum_{j\geq 1} \hat{\lambda}_j \leq B\hat{\lambda}_1^{-1} \sum_{j\geq 1} \hat{\lambda}_j \leq BE,$$

giving the claim with $C = BE$. On the other hand, if $t\hat{\lambda}_1 \geq 1$, then let $k \geq 1$ be defined by $t\hat{\lambda}_{k+1} < 1 \leq t\hat{\lambda}_k$. Applying (**LFU**), we have

$$\begin{aligned} \mathcal{N}_n^g(t) &= \sum_{j=1}^n \hat{\lambda}_j g_t(\hat{\lambda}_j) = \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j) + \sum_{j>k} \hat{\lambda}_j g_t(\hat{\lambda}_j) \\ &\leq \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j) + Bt \sum_{j>k} \hat{\lambda}_j. \end{aligned} \tag{3.2}$$

Now by the definition of k, (**EVBound**) and (**LFL**), we have

$$t \sum_{j>k} \hat{\lambda}_j \leq \hat{\lambda}_{k+1}^{-1} \sum_{j>k} \hat{\lambda}_j \leq Ek \leq Eb^{-1} \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j).$$

Inserting this into (3.2), we get

$$\sum_{j=1}^n \hat{\lambda}_j g_t(\hat{\lambda}_j) \leq C \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j) \leq b^{-1} C \sum_{j=1}^n \hat{\lambda}_j^2 g_t^2(\hat{\lambda}_j)$$

with $C = (1 + b^{-1}BE)$. □

Combining (**EVBound**) and Lemma 11 illustrates the way Proposition 10 can be transferred into a classical oracle inequality that is, involving bias and variance terms in the r.h.s., which is achieved by the next result.

Theorem 12 *Suppose that Assumptions* (**SubGN**), (**BdK**) *and* (**EVBound**) *hold and that the regularizer g satisfies* (**LFL**). *Then the early stopping rule τ_{DP} based on the standard discrepancy principle* (2.8) *satisfies for every $u > 1$ the bound*

$$\mathbf{P}_\epsilon \Big(\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{DP})}\|_n^2 > C \Big(\min_{0 \leq t \leq T} \mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2 + \frac{\sigma^2 \sqrt{u}}{\sqrt{n}} + \frac{\sigma^2 u}{n} \Big) \Big) \leq 5e^{-u},$$

where C is a constant depending only on A, b, B and E.

The proof of Theorem 12 is deferred to Section 7.3. Theorem 12 yields a non-asymptotic result, which contrasts for instance with the one of Blanchard and Krämer (2016) where conjugate gradient descent and minimum discrepancy principle are analyzed. The above inequality is established with high probability, and it provides the precise sub-Gaussian and sub-exponential behaviours. This is a technical improvement compared to existing approaches where similar oracle inequalities in expectation are derived (Blanchard et al., 2018b,a).

The oracle performance in the r.h.s. of Theorem 12 is given through the expected excess risk (rather than the excess risk). This could be made at the price of an additional $\log T$ term, accounting for the uniform control of the discrepancy between the excess risk and its expectation over the first T iterations.

Let us also notice that Theorem 12 does not depend on any smoothness assumption on f. Making additional smoothness assumptions would immediately lead to a specific bound on $\min_{0 \leq t \leq T} \mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2$ expressed in terms of convergence rate. This will be done in the random design framework in Section 5.2, where it is shown that the classical discrepancy principle leads to optimal convergence rates whenever the latter rate is slower than the $n^{-1/2}$-rate. Such a situation can happen in the outer case $r < 1/2$.

In contrast, the $1/\sqrt{n}$-rate is not negligible whenever the minimal bias-variance trade-off is smaller than (or of same order as) $1/\sqrt{n}$. This holds true e.g. in the inner case $r \geq 1/2$. Compared to (Blanchard et al., 2018b) and (Blanchard et al., 2018a), the term $\sigma^2/\sqrt{n}$ corresponds to their term $\sqrt{D}\delta^2$ (with the analogy noise level $\delta^2 = \sigma^2/n$ and discretization dimension $D = n$). Moreover, in (Blanchard et al., 2018b) it has been shown for the specific case of spectral cut-off that such terms can not be avoided for early stopping rules based on the residual filtration. Hence, we conclude that the classical minimum discrepancy principle turns out to be useless when estimating smooth functions. This motivates considering smoothing-based strategies in Section 4.

3.3 Discussion

As earlier emphasized, the $\sigma^2/\sqrt{n}$ term in Theorem 12 cannot be improved. The reason for this term is the high variability in the stopping rule τ_{DP} and the empirical risk (see Figure 4a). To illustrate this further, let us consider the deviation inequality for τ_{DP} from Proposition 8 applied with t satisfying $\mathcal{N}_n^g(t) = (1+\delta)\mathcal{N}_n^g(t_n^*)$ with $\delta > 1$, leading to

$$\mathbf{P}_\epsilon(\tau_{DP} > t) \leq 2\exp\Big(- c\Big(\delta\mathcal{N}_n^g(t_n^*) \wedge \frac{(\delta\mathcal{N}_n^g(t_n^*))^2}{n}\Big)\Big). \tag{3.3}$$

If, for instance, (**PolDecTS**) and (**SC**(r,R)) hold, then $\mathcal{N}_n^g(t_n^*)$ is typically of order $n^{1/(2\alpha r+1)}$, meaning that the above (non-improvable) concentration bound becomes vacuous for $n^{1/(2\alpha r+1)} \lll n^{1/2}$. This is the case if r is larger than $1/(2\alpha)$. In such settings, the classical discrepancy principle will typically lead to stopping times that are too large with high probability. Interestingly, we prove in the random-design context of Section 5.2 that the discrepancy principle can nevertheless achieve state-of-the-art rates under the condition $r \leq 1/(2\alpha)$.

The limitation of τ_{DP} in the present context can be also interpreted as the consequence of trying to estimate a part of the signal that is smaller than the level of noise σ. This can be easily observed by computing the singular value decomposition (SVD) of the normalized Gram matrix K_n, and by computing the residuals in this new basis. Then a natural idea to overcome this problem is the smoothing of the residuals, then reducing the contribution of these "small coordinates" to the (smoothed) residuals. This strategy has been already explored in the literature (see for instance Blanchard and Krämer (2016) for the CGD). Studying how τ_{DP} can be improved when combined with the smoothing of the residuals is the purpose of Section 4.

4. SDP and oracle inequality: Fixed-design

We now turn to a modification of the discrepancy principle based on the smoothing of the residuals that is, on the smoothed empirical risk.

4.1 Smoothing-based discrepancy principle

As discussed in Section 3.3, the main drawback of the discrepancy principle-based rule τ_{DP} results from the large variance of the empirical risk, leading to the $\sigma^2/\sqrt{n}$ error term in Theorem 12.

The purpose of the present section is to show how this error term can be avoided by considering a modified stopping rule called τ_{SDP} based on the smoothing of residuals that is, the smoothed empirical risk. This can be encoded by considering the so-called smoothed empirical risk $\|L_n(\mathbf{Y} - S_n\hat{f}^{(t)})\|_n^2$ for some (smoothing) matrix $L_n \in \mathbb{R}^{n\times n}$. In what follows, we will restrict ourselves to the case where $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$ with regularizer $\tilde{g}$ (satisfying Definition 2) and consider

$$\tau_{SDP} = \inf\left\{t \geq 0 : \|\tilde{g}_T^{1/2}(K_n)K_n^{1/2}(\mathbf{Y} - S_n\hat{f}^{(t)})\|_n^2 \leq \frac{\sigma^2\mathcal{N}_n^{\tilde{g}}(T)}{n}\right\} \wedge T \tag{4.1}$$

with $T > 0$. the choice $\tilde{g}_T(\lambda) = (\lambda + T^{-1})^{-1}$ as Tikhonov regularization results in the early stopping rule earlier studied in Blanchard and Mathé (2012) in the statistical inverse problem setting. Different choices for L_n include $L_n = K_n^{s/2}$, $s \leq 1$, have been studied in (Stankewitz, 2020) for the spectral cut-off filter algorithm.

Note that, while τ_{SDP} is defined using the square-root of the kernel matrix, this dependence can be removed by invoking the identity $\|A^{1/2}x\|_n^2 = \langle Ax, x\rangle_n$ with $A = \tilde{g}_T(K_n)K_n$ and $x = \mathbf{Y} - S_n\hat{f}^{(t)}$. Hence, computing the SDP stopping rule does not necessarily require computing the singular value decomposition of K_n beforehand. Still, the factor $\tilde{g}_T(K_n)$ makes τ_{SDP} computationally more costly than τ_{DP}.

Then the goal in what follows is to assess the statistical performance of the final estimator $\hat{f}^{(\tau_{SDP})}$, where τ_{SDP} is obtained by the so-called *smoothed discrepancy principle* (SDP).

4.2 Main results

The present section follows the same structure as above Section 3 with firstly describing key deviation inequalities for τ_{SDP} and the related smoothed bias and variance terms, and secondly formulating our main improved oracle inequality for τ_{SDP}.

4.2.1 Deviation inequalities for the smoothed stopping rule

Let us now explain how deviation inequalities in the case of the classical DP (Section 3.1.1) can be extended to smoothed case. For simplicity of the present exposition, we restrict ourselves to τ_{SDP} applied with the Tikhonov smoothing $\tilde{g}_t(\lambda) = (\lambda + t^{-1})^{-1}$. However, the next results are not limited to this choice.

Following the analysis of the classical discrepancy principle in Section 3.1.1, it is easy to see that the expected smoothed empirical risk satisfies a basic inequality similar to (2.5). In fact introducing the *smoothed g-effective dimension* $\widetilde{\mathcal{N}}_n^g(t) = \mathrm{tr}((K_n + T^{-1})^{-1}K_n g_t(K_n)K_n)$, we have

$$\begin{aligned}&\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 - 2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t)\\ &\leq \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \frac{\sigma^2}{n}\mathcal{N}_n(T) \leq \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 - \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t), \quad t \geq 0,\end{aligned}$$

where $\tilde{\mathbf{a}} = (K_n + T^{-1})^{-1/2} K_n^{-1/2} \mathbf{a}$, for every $\mathbf{a} \in \mathbb{R}^n$. This allows us to carry out the same basic comparison between τ_{SDP} and the *smoothed balancing stopping rule*

$$\widetilde{t_n^*} = \inf \left\{ t \geq 0 : \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \leq \frac{\sigma^2}{n} \widetilde{\mathcal{N}}_n^g(t) \right\} \tag{4.2}$$

(with $\widetilde{t_n^*} = \infty$ if such a t does not exist). Our first result in this line is the next deviation inequality for τ_{SDP}, which should be seen as the smoothing-based counterpart of Proposition 8.

Proposition 13 *If* (**SubGN**) *holds, then there is a constant $c > 0$ depending only on A such that for every $t > \widetilde{t_n^*}$,*

$$\mathbf{P}_\epsilon(\tau_{SDP} > t) \leq 2 \exp\Big(- c\Big(y \wedge \frac{y^2}{\mathcal{N}_n(T)} \Big) \Big), \qquad y = \widetilde{\mathcal{N}}_n^g(t) - \widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}).$$

In particular, for every $y > 0$, we have

$$\mathbf{P}_\epsilon(\widetilde{\mathcal{N}}_n^g(\tau_{SDP}) > \widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}) + y) \leq 2 \exp\Big(- c\Big(y \wedge \frac{y^2}{\mathcal{N}_n(T)} \Big) \Big).$$

This is a simplified version of the deviation bound established in Proposition 23.

Let us make a few comments mainly emphasizing the differences with Proposition 8 established for τ_{DP}. Firstly, the former n at the denominator of the exponent is now replaced by the empirical effective dimension $\mathcal{N}_n(T)$, which allows for taking into account the decay rate of the eigenvalues of K_n. In particular, the condition for having this probability meaningful (that is, close to 0) is no longer $\sqrt{n} \ll y$ but instead $\sqrt{\mathcal{N}_n(T)} \ll y$, which is typically much weaker if one can exploit some knowledge on the decay rate of the eigenvalues. Secondly, the g-effective dimension in Proposition 8 is now replaced by its smoothed version $\widetilde{\mathcal{N}}_n^g(t)$. Since $\widetilde{\mathcal{N}}_n^g(t) \leq \mathcal{N}_n^g(t)$, this leads to a slightly weaker deviations in terms of y.

Let us emphasize that this deviation inequality of the $\widetilde{\mathcal{N}}_n^g(t)$ serves for controlling the variance of $\hat{f}^{(t)}$. This results from the key observation that the term $\operatorname{tr}(g_t^2(K_n)K_n^2)$ (appearing in the variance of $\hat{f}^{(t)}$) can be bounded by a constant times $\widetilde{\mathcal{N}}_n^g(t)$ (while in Section 2.4, we only used that it is bounded by the g-effective dimension).

Similarly, the squared bias $\|r_t(K_n)\mathbf{f}\|_n^2$ can be also related to its smoothed version $\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2$, where the latter term is dealt with in the following simplified version of the deviation bound in Proposition 25.

Proposition 14 *Suppose that Assumptions* (**SubGN**) *and* (**BdK**) *hold. Then, for every $y > 0$ such that $2\|r_{\widetilde{t_n^*}}(K_n)\tilde{\mathbf{f}}\|_n^2 + \sigma^2 n^{-1} y > \|r_T(K_n)\tilde{\mathbf{f}}\|_n^2$, we have*

$$\mathbf{P}_\epsilon\Big(\|r_{\tau_{SDP}}(K_n)\tilde{\mathbf{f}}\|_n^2 > 2\|r_{\widetilde{t_n^*}}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n} y \Big) \leq 2 \exp\Big(- c\Big(y \wedge \frac{y^2}{\mathcal{N}_n(T)} \Big) \Big).$$

If $\widetilde{t_n^*} = \infty$, we additionally assume that $\sigma^2 n^{-1} y \geq \|r_{\widetilde{t_n^*}}(K_n)\tilde{\mathbf{f}}\|_n^2$.

4.2.2 Improved oracle inequality

We are now in position to state an improved oracle inequality for the inner case that holds for the smoothed discrepancy principle (SDP), namely τ_{SDP}.

Theorem 15 *Suppose that* (**SubGN**)*,* (**BdK**)*,* (**EVBound**) *and* (**SC**(r,R)) *hold with* $s = r - 1/2 \geq 0$ *and the the regularizer* g *satisfies* (**LFL**)*. Moreover, suppose that* $\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\mathrm{op}} \leq 1/2$ *holds. Then early stopping rule* τ_{SDP} *based on the smoothed discrepancy principle from* (4.1) *with regularizer* $\tilde{g}$ *such that* (**LFL**) *holds satisfies the bound*

$$\mathbf{P}_\epsilon\Big(\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{SDP})}\|_n^2 > C\Big(\min_{0 \leq t \leq T} \mathbf{E}_\epsilon\|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2 + \frac{\sigma^2\sqrt{u\mathcal{N}_n(T)}}{n} + \frac{\sigma^2 u}{n} + T^{-(1+2s)} + T^{-1}\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}\Big)\Big) \leq 5e^{-u}, \quad u > 1,$$

where the term $T^{-1}\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}$ *can be dropped if* $s \leq 1/2$*.*

A proof of Theorem 15 is given in Section 7.3. Comparing this result to the oracle inequality in Theorem 12, we see that we replaced the term $\sigma^2/\sqrt{n}$ by $\sigma^2\sqrt{\mathcal{N}_n(T)}/n$. Under (**PolDecTS**), for instance, we have $\mathcal{N}_n(T) \leq CT^{1/\alpha}$, meaning that $\sqrt{\mathcal{N}_n(T)}/n \leq 1/\sqrt{n}$ as long as $T \leq n^\alpha$.

The event $\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\mathrm{op}} \leq 1/2$ is needed to apply the source condition (**SC**(r,R)) (formulated in terms of the population covariance operator) in the empirical world. It can be further weakened (there is e.g. no event in the case $s = 0$; see the proof of Lemma 29), but in its present form it is exactly the event needed to transfer the results from the fixed to the random design framework. This is the purpose of the next section. In particular, we will turn the above oracle inequality into a rate optimality statement, showing that the smoothed discrepancy principle is adaptive over a certain range of smoothness parameters and polynomial decay rates.

5. The random design framework

In this section we transfer our oracle inequalities from the fixed to the random design framework by means of a change-of-norm (or change of measure) argument exposed in Section 5.1. The purpose of Section 5.2 is the analysis of the stopping rule based on the discrepancy principle (DP) in the *outer case*, while Section 5.3 rather addresses its smoothed version (SDP) in the *inner case.*

To keep the exposition as simple as possible in what follows, we focus on results given in terms of expectations from now on. Similar results expressed “with high probability” can be derived from the technical material developed in Sections 7 and 8, but at the price of more involved expressions.

5.1 Change of norm argument

The first step in our analysis is a change of norm argument formulated by the next result, which controls the difference between the $L^2(\rho)$-norm ($\|\cdot\|_\rho$) and its empirical version, namely the n-th norm ($\|\cdot\|_n$).

Lemma 16 *Let $\delta \in (0,1)$ and $T > 0$. Then we have*

$$\forall h \in \mathcal{H}, \quad |\|S_n h\|_n^2 - \|S_\rho h\|_\rho^2| \le \delta(\|S_\rho h\|_\rho^2 + T^{-1}\|h\|_{\mathcal{H}}^2)$$

if and only if

$$\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\text{op}} \le \delta.$$

Lemma 16 establishes the equivalence between the uniform control of the difference between the squared ρ- and n-th norms and deriving an upper bound on the operator norm of the normalized difference between the empirical and population covariance operators. In particular if one of the assertion holds, then

$$\forall h \in \mathcal{H}, \quad \|S_\rho h\|_\rho^2 \le \frac{1}{1-\delta}\|S_n h\|_n^2 + \frac{\delta}{1-\delta}\frac{\|h\|_{\mathcal{H}}^2}{T}$$

gives rise to a natural strategy for upper bounding the ρ-norm of any function in $\mathcal{H}$. It consists first in upper bounding its n-th norm (which was the purpose of Sections 3.2 and 4.2.2), and then in controlling its $\mathcal{H}$-norm.

Proof of Lemma 16 Using the identities $\|S_n h\|_n^2 - \|S_\rho h\|_\rho^2 = \langle(\Sigma_n - \Sigma)h, h\rangle_{\mathcal{H}}$ and $\|S_\rho h\|_\rho^2 + T^{-1}\|h\|_{\mathcal{H}}^2 = \|(\Sigma + T^{-1})^{1/2}h\|_{\mathcal{H}}^2$, the first assertion is equivalent to

$$\forall h \in \mathcal{H}, \quad |\langle(\Sigma_n - \Sigma)h, h\rangle_{\mathcal{H}}| \le \delta\|(\Sigma + T^{-1})^{1/2}h\|_{\mathcal{H}}^2.$$

Since $(\Sigma + T^{-1})^{1/2}$ is self-adjoint and strictly positive definite, this is the case if and only if

$$\forall h \in \mathcal{H}, \quad |\langle(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}h, h\rangle_{\mathcal{H}}| \le \delta\|h\|_{\mathcal{H}}^2.$$

This gives the claim. □

5.2 DP performance: Outer case

5.2.1 Main result

We now turn to the classical discrepancy principle for which we formulate a result in the outer case.

Theorem 17 *Suppose that* (**SubGN**) *and* (**BdK**) *hold. Suppose that the source condition* (**SC**(r,R)) *holds with $r < 1/2$ and that f is bounded. Moreover, suppose that the regularizer g satisfies* (**QuErr**) *with $q \ge r$. Then there are constants $c, C > 0$ such that the standard discrepancy principle τ_{DP} with emergency stop $T = cn/\log n$, $n \ge 2$, satisfies*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \le C\Big(\min_{0<t\le c\frac{n}{\log n}}\Big\{t^{-2r} + \frac{\mathcal{N}(t)}{n}\Big\} + n^{-1/2}\Big).$$

The proof of Theorem 17 can be found in Section 8.4. Unlike the results from Sections 3.2 and 4.2.2 in the fixed design case, there is an additional constraint on the emergency stop T that has to be smaller than $cn/\log n$. This constraint is related to the control of the probability of the event $\{\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\text{op}} \le 1/2\}$.

Without any further assumption on the decay rate of the eigenvalues, the effective dimension $\mathcal{N}(t)$ can be upper bounded by $M^2 t$; see e.g. Appendix B. Theorem 17 thus gives

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \leq C \max\Big(n^{-\frac{2r}{2r+1}}, \Big(\frac{\log n}{n}\Big)^{2r}, n^{-1/2}\Big) \leq C n^{-\frac{2r}{2r+1}}.$$

As a consequence, the classical discrepancy principle leads to optimal rates of convergence throughout the whole range $r \in (0, 1/2)$ of the outer case (cf. Fischer and Steinwart (2020)).

5.2.2 Discussion and extensions for polynomial decay

For some $L > 0$ and $\alpha > 1$, suppose that

$$\forall j \geq 1, \qquad \lambda_j \leq L j^{-\alpha}. \tag{$\mathbf{Pol}(\alpha)$}$$

By Lemma 43(i) we have $\mathcal{N}(t) \leq C t^{1/\alpha}$ for all $t > L^{-1}$. Specialized to ($\mathbf{Pol}(\alpha)$), Theorem 17 thus gives

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \leq C \max\Big(n^{-\frac{2r}{2r+1/\alpha}}, \Big(\frac{\log n}{n}\Big)^{2r}, n^{-1/2}\Big).$$

In other words, we obtain up to some additional $\log n$ factors the following rates of convergence:

$$\begin{cases} n^{-\frac{2r}{2r+1/\alpha}}, & \text{if } 2r + 1/\alpha > 1, r \leq 1/(2\alpha), \\ n^{-2r}, & \text{if } 2r + 1/\alpha \leq 1, r \leq 1/4, \\ n^{-1/2}, & \text{if } r > 1/4, r > 1/(2\alpha). \end{cases}$$

We see that the classical discrepancy principle achieves the optimal rates of convergence in the hard learning scenario if $1/2 - 1/(2\alpha) < r \leq 1/(2\alpha)$.

In what follows we compare these rates to results from the literature that have been collected in Table 1, and we show how Theorem 17 can be improved under an additional condition on the kernel. Ignoring $\log n$ factors, Table 1 illustrates that the rate

$$n^{-\frac{2r}{2r+1/\alpha}}, \quad 1/2 - 1/(2\alpha) < r$$

is optimal (see the lower bound derived in Fischer and Steinwart (2020) for KRR with $\gamma = 0$). The classical discrepancy principle achieves the optimal rates of convergence in the present hard learning scenario if $1/2 - 1/(2\alpha) < r \leq 1/(2\alpha)$. By contrast,

$$n^{-2r}, \quad r \leq 1/2 - 1/(2\alpha). \tag{5.1}$$

is the state-of-the-art result in the outer case when only assuming that (**BdK**) holds; see e.g. Corollary 4.4 in Lin et al. (2020) and the lower bound from Fischer and Steinwart (2020) with $\gamma = 0$ which does not match the upper bound.

There are possible improvements under stronger boundedness assumptions such as the embedding assumption from Fischer and Steinwart (2020) parametrized by $0 < \mu \leq 1$ in

Paper/Criterion	$0 < r \leq \frac{\mu}{2} - \frac{1}{2\alpha}$	$\frac{\mu}{2} - \frac{1}{2\alpha} < r$	*Algorithms*	*Assumptions*
Lin et al. (2020) $\left\| L_\rho^{-a}\left(f_{\mathcal{H}} - S_\rho \hat{f}^{(t_n)}\right)\right\|_\rho^2$ $(0 \leq a \leq 1/2 \wedge r)$	$\leq Cn^{-(2r-2a)}$	$\leq Cn^{-\frac{2r-2a}{2r+1/\alpha}}$	Spectral filters	Qual(q) $(q \geq r)$, Pol(α) SC(ϕ) $\Rightarrow$ SC(r), $\mu = 1$
Pillaud-Vivien et al. (2018) $\left\| f - S_\rho \hat{f}^{(t_n)}\right\|_\rho^2$	$\leq Cn^{-\frac{2r}{\mu}}$	$\leq Cn^{-\frac{2r-2a}{2r+1/\alpha}}$	SGD	EMB(μ) $(0 \leq \mu \leq 1)$ Pol(α), SC(r)
Fischer and Steinwart (2020) $\left\| f - \hat{f}^{(t_n)}\right\|_\gamma^2$ $(0 \leq \gamma \leq 2r)$	$\leq Cn^{-\frac{2r-\gamma}{\mu}}$ $\geq cn^{-\frac{\mu-\gamma}{\mu+1/\alpha}}$ $(\mu/2 > r)$	$Cn^{-\frac{2r-\gamma}{2r+1/\alpha}}$ $\geq cn^{-\frac{2r-\gamma}{2r+1/\alpha}}$ $(\mu/2 \leq r)$	KRR	Qual(1) $(q = 1 \geq r)$ EMB(μ) $(0 \leq \mu \leq 1)$ Pol(α), SC(r)

Table 1: Convergence rates (upper and lower bounds up to logarithmic terms) derived with spectral filters under qualification $Qual(q)$, polynomial decay assumption $Pol(\alpha)$, general source conditions $SC(\phi)$ (Lin et al. (2020)) or polynomial one $SC(r)$, and embedding assumption $EMB(\mu)$, where $0 \leq \mu \leq 1$ (Pillaud-Vivien et al. (2018)). For each row, t_n denotes the optimized value of the parameter t such that the corresponding rate is achieved. Also, $c, C > 0$ denote generic constants independent of n that are different from cell to cell. Here $f_{\mathcal{H}}$ is the best approximation to f within the closure of $\mathcal{H}$ in $L^2(\rho)$. See also Section 2.3 for a more thorough overview.

Pillaud-Vivien et al. (2018). In fact, if there is a $\mu \leq 1$ such that $\|\Sigma^{\mu/2-1/2}k_X\|_{\mathcal{H}} \leq C_\mu M$ almost surely, then one can achieve (up to $\log n$ factors) the improved rate

$$\begin{cases} n^{-\frac{2r}{2r+1/\alpha}}, & 2r + 1/\alpha \geq \mu, \\ n^{-\frac{2r}{\mu}}, & 2r + 1/\alpha \leq \mu, \end{cases} \tag{5.2}$$

see e.g. Pillaud-Vivien et al. (2018) and Fischer and Steinwart (2020). Such improvements are also possible in our case, which is the purpose of the next result proved in Section 8.4.

Theorem 18 *Suppose that* (**SubGN**), (**BdK**), (**SC**(r,R)) *and* (**Pol**(α)) *holds with* $r < 1/2$ *and that* f *is bounded. Suppose that there is a* $\mu \in [0, 1)$ *and a constant* $C_\mu > 0$ *such that* $\|\Sigma^{\mu/2-1/2}k_X\|_{\mathcal{H}} \leq C_\mu M$. *Finally, suppose that the regularizer* g *satisfies* (**QuErr**) *with* $q \geq r$. *Then there are constants* $c, C > 0$ *such that the standard discrepancy principle* τ_{DP}

with emergency stop $T = c(n/\log n)^{1/\mu}$, $n \geq 2$, *satisfies*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \leq C\Big(\min_{0<t\leq c(\frac{n}{\log n})^{1/\mu}} \Big\{ t^{-2r} + \frac{t^{1/\alpha}}{n} \Big\} + n^{-1/2} \Big).$$

Let us first notice that introducing the stronger assumption involving the parameter $0 \leq \mu \leq 1$ allows to enlarge the emergency stop T and thus the range of values of t over which the minimum in the r.h.s. is computed since $1/\mu \geq 1$. By the arguments from above Table 1 also proves that the classical discrepancy principle achieves the optimal rates of convergence in the hard learning scenario if $\mu/2 - 1/(2\alpha) < r \leq 1/(2\alpha)$. In the setting of Sobolev spaces any $\mu > 1/\alpha$ is admissible (see Example 2 in Pillaud-Vivien et al. (2018)), leading to the adaptation interval $r \in (\delta, 1/(2\alpha)]$, $\delta > 0$ arbitrary.

5.3 SDP performance: Inner case

In the present section, we establish two inequalities in the inner case for τ_{SDP}. The main difference between these results lies in the use of different emergency stopping times T. In the first one (Theorem 19), a deterministic emergency stop of size at most $n/\log n$ is used, while the second result (Theorem 20) allows for using a more sophisticated *data-driven* emergency stop defined as the solution of a fixed-point equation, which gives rise to an optimal (leading to statistical adaptivity) early stopping rule that can be applied in practice.

5.3.1 Main result

The transfer from the fixed design to the random design cases requires first an additional assumption on the effective rank, which is the population version of the former (**EVBound**) assumption earlier introduced in the fixed design case.

Assumption 7 *There exists a constant* $E' > 0$ *such that, for each* $k \geq 0$*, we have*

$$\lambda_{k+1}^{-1} \sum_{j>k} \lambda_j \leq E'(k \vee 1). \qquad \textbf{(EffRank)}$$

This assumption is a population version of (**EVBound**) and Lemma 35 specifies an event on which it indeed implies (**EVBound**). Similarly as in Section 3.2, (**EVBound**) is needed to bound the proxy variance term in terms of the smoothed proxy variance term (cf. Lemma 36). Under this additional assumption the smoothed discrepancy principle from Section 4 achieved the following bound.

Theorem 19 *Suppose that Assumptions* (**SubGN**)*,* (**SC**(r,R))*,* (**BdK**)*,* (**LFL**) *and* (**EffRank**) *hold with* $s = r - 1/2 \geq 0$*. Moreover, suppose that the regularizer* g *satisfies* (**QuErr**) *with* $q \geq r$*. Then there are constants* $c, C > 0$ *such that the smoothed discrepancy principle* τ_{SDP} *from* (4.1) *with* $\tilde{g}_t(\lambda) = (\lambda + t^{-1})^{-1}$ *and* $T \leq cn/(\log n)$*,* $n \geq 2$*, achieves the bound*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C\Big(\min_{0<t\leq T} \Big\{ t^{-2r} + \frac{\mathcal{N}(t)}{n} \Big\} + \frac{\sqrt{\mathcal{N}(T)}}{n} \Big).$$

The proof of Theorem 19 can be found in Section 8.3.1. Note that the condition $q \geq r$ on the qualification error of g can be dropped by introducing slower rates depending also on q.

Without any further assumption on the decay rate of the eigenvalues (except of (**EffRank**)), Theorem 19 gives

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \leq C \max\left(n^{-\frac{2r}{2r+1}}, T^{-2r}, \frac{\sqrt{T}}{n}\right).$$

Let us now assume that a lower bound $r_0 \geq 1/2$ is known on the smoothness parameter r, which means that (**SC**(r,R)) holds with $r \geq r_0$. Then using this side information, the choice $T = c_1 n^{1/(2r_0+1)}$ with $c_1 > 0$ sufficiently small leads to

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C \max\left(n^{-\frac{2r}{2r+1}}, n^{-\frac{4r_0+1}{4r_0+2}}\right).$$

This entails that the smoothed discrepancy principle τ_{SDP} reaches optimal rates of convergence throughout the range

$$r \in \left[r_0, 2r_0 + \frac{1}{2}\right].$$

For instance with $r_0 = 1/2$ (that is the inner case without additional smoothness information), τ_{SDP} is optimal over the range $r \in [\,1/2, 3/2\,]$.

Instead of choosing $T = n^{1/(2r_0+1)}$, one might also define T as the solution to the fixed-point equation $c_0 T^{-2r_0} = \mathcal{N}(T)/n$ with $c_0 = 1$, which corresponds to a bias-variance trade-off in the case $r = r_0$. This would lead to the same adaptation interval $[r_0, 2r_0 + 1/2]$. Such and related fixed-point equations play a central role in empirical risk minimization problems; see e.g. Bartlett et al. (2005) and Koltchinskii (2006), and it is easy to see, using the proof of Lemma 6 and Proposition 3.3 in Koltchinskii (2011), that the effective dimension $\mathcal{N}(t)$ can be bounded from below and above in terms of local Rademacher averages.

With an additional assumption such as a polynomial eigenvalue decay, the previous analysis can be further applied. If (**Pol**(α)) and (**SC**(r,R)) hold with $r \geq r_0$, then the choice $T^{2r_0}\mathcal{N}(T) = n$ leads to

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C \max\left(n^{-\frac{2r\alpha}{2r\alpha+1}}, n^{-\frac{4\alpha r_0+1}{4\alpha r_0+2}}\right),$$

meaning that the smoothed discrepancy principle leads to optimal rates of convergence throughout the range

$$r \in \left[r_0, 2r_0 + \frac{1}{2\alpha}\right]. \tag{5.3}$$

Let us emphasize that this range of values is narrower than the previous one derived without any assumption on the eigenvalue decay ($\alpha > 1$). This owes to the fact that, by specifying an eigenvalue decay assumption (that is, by choosing a given kernel), we restrict the smoothness of the functions in the induced Hilbert space that can be well approximated.

5.3.2 Improvement towards data-driven emergency stops

In previous Section 5.3.1, we have chosen a (deterministic) T as the solution of the equation $t^{2r_0}\mathcal{N}(t) = c_0 n$ by taking advantage of the prior knowledge of a lower bound r_0 on the smoothness parameter. Without such an a priori knowledge on r, the equation $T\mathcal{N}(T) = c_0 n$ provides a natural choice for T. Yet, such a choice is not achievable in practice since $\mathcal{N}(t)$ is not known.

In this section we show that similar bounds hold true if $T = T(X_1, \dots, X_n)$ is allowed to depend on the covariates $X_1, \dots, X_n$ (but not on the responses). This is possible since all results established in the fixed design case continue to hold. The following result focuses on the choice $T\mathcal{N}_n(T) = c_0 n$.

Theorem 20 *Suppose that Assumptions* (**SubGN**)*,* (**SC**(r,R))*,* (**BdK**)*,* (**LFL**) *and* (**EffRank**) *hold with $s = r - 1/2 \geq 0$. Moreover, suppose that the regularizer g satisfy* (**QuErr**) *with $q \geq r$. Let $\hat{T} > 0$ be the solution of $\hat{T}\mathcal{N}_n(\hat{T}) = n$ (set $\hat{T} = \infty$ if such a solution does not exist). Then the smoothed discrepancy principle τ_{SDP} from* (4.1) *with $\tilde{g}_t(\lambda) = (\lambda + t^{-1})^{-1}$ and $T = \min(\hat{T}, cn/\log n)$, $n \geq 2$ and $c > 0$ sufficiently small, achieves the bound*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C\Big(\min_{t>0} \Big\{ t^{-2r} + \frac{\mathcal{N}(t)}{n} \Big\} + \sqrt{\frac{1}{n} \min_{t>0} \Big\{ t^{-1} + \frac{\mathcal{N}(t)}{n} \Big\}} + \frac{\log n}{n} \Big).$$

The proof of Theorem 20 can be found in Section 8.3.2. Compared to the statement in Theorem 19, the term $\sqrt{\mathcal{N}(T)}/n$ has disappeared. Actually it has been replaced by the square-root on the r.h.s. of the above inequality due to the control of $\sqrt{\mathcal{N}(T)}$ with the present (random) choice of $T = \min(\hat{T}, cn/\log n)$. As can be easily checked from the proof, the control of this term is also responsible for the additional $(\log n)/n$, which does not really influence our conclusion regarding convergence rates.

Let us also remark that the above definition $\hat{T}$ with $c_0 = 1$ does not take into account constants such as the variance σ^2 or $\|f\|_{\mathcal{H}}$ for instance that should arise from the upper bounds on the variance or bias terms. Obviously introducing these constants in the fixed-point equation would not modify our conclusion regarding the convergence rates and the statistical adaptivity property, which is the main achievement of the present analysis. In practice, one could replace these constants in the upper bound on the bias term by upper bounds with high probability derived from the empirical risk evaluated at 0.

Illustration on two classical eigenvalue decay assumptions Since the interpretation in terms of convergence rates is not easy from the statement in Theorem 20, let us now consider two illustrative examples allowing for drawing further insightful conclusions.

Example 6 (Polynomial decay) *Under the assumptions of Theorem 20 and* (**Pol**(α))*, we get*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C \max\Big(n^{-\frac{2\alpha r}{2\alpha r + 1}}, n^{-\frac{2\alpha + 1}{2\alpha + 2}} \Big). \tag{5.4}$$

This means that, by including the data-driven *choice of $\hat{T}$, the smoothed discrepancy principle τ_{SDP} still reaches statistical adaptivity (that is, automatically enjoys optimal rates of convergence) throughout the range $r \in [1/2, 1 + 1/(2\alpha)]$.*

Note that this choice for $\hat{T}$ corresponds to the stopping rule defined by Eq. (6) in Raskutti et al. (2014). The striking remark is that Raskutti et al. (2014) establishes the rate $n^{-\alpha/(\alpha+1)}$, while we obtain an estimator automatically achieving the optimal rate $n^{-(2\alpha r)/(2\alpha r+1)}$ throughout $r \in [1/2, 1+1/(2\alpha)]$ and the rate $n^{-(\alpha+1/2)/(\alpha+1)}$ otherwise. This proves that τ_{SDP} is uniformly better than the stopping rule of Raskutti et al. (2014) in the inner case ($r \geq 1/2$) and under a polynomial decay assumption.

Example 7 (Exponential decay) *For some $L > 0$ and $\alpha > 1$, suppose that $\lambda_j \leq e^{-Lj^{\alpha}}$ for every $j \geq 1$. Applying Theorem 20 and Lemma 43(ii), we get*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq C\frac{(\log n)^{1/\alpha}}{n}.$$

5.4 Extensions to RKHS-norm and related works

While it is well known that there is a regularization parameter which is simultaneously optimal for the prediction risk ($L^2(\rho)-$norm) and the RKHS-norm risk (at least under a polynomial source condition; see Lu and Pereverzev (2013); Blanchard and Mücke (2018)), extensions to τ_{DP} or τ_{SDP} are not straightforward. Indeed any early stopping rule focusing on one risk might still be sub-optimal for the other. For instance, simultaneous adaptation has been proved from two main strategies where (i) both norms are taken into account, or (ii) stronger model assumptions are imposed to make a change-of-norm argument more accessible. The first case has been addressed with the balancing principle in De Vito et al. (2010); Lu and Pereverzev (2013) by considering both the empirical and the RKHS-norm, while Blanchard et al. (2019) study a modified balancing principle based on a varying norm. The second case is illustrated by the recent strong norm bounds derived for early stopping rules by Blanchard et al. (2018b,a); Stankewitz (2020) in a Gaussian sequence model. Oracle inequalities are established e.g. under a polynomial decay assumption and for specific signal classes.

From a more general perspective on bounds for spectral filter algorithms, optimal rates have been recently derived by Blanchard and Mücke (2018) and Lin et al. (2020) which simultaneously hold for different norms between the RKHS-norm and the $L^2(\rho)$-norm. Designing a fully data-driven stopping rule that would enjoy simultaneous optimality properties for such different norms remains a challenge to be addressed in future work.

6. Simulation experiments

The goal of the present section is to illustrate the main behaviors of the stopping rules under consideration, as predicted from a theoretical perspective, respectively in Sections 5.2 and 5.3.

6.1 Simulation design: Generating synthetic data

The present simulation experiments are carried out with the Landweber algorithm (that is, Gradient descent with constant step-size $\eta > 0$ along the iterations) as described in Section 2.2. The sample size n varies within $\{200, 400, 600, 800, 1\,000\}$ and the number of replicates in all the experiments is $N = 200$. In all the simulation experiments, when applying the smoothed discrepancy principle rule τ_{SDP} (see Eq. (4.1)).

The data are drawn from the model described by (2.1) with the variance σ^2 of the Gaussian noise to be equal to 1, and where the deterministic vector $(x_1, \ldots, x_n)$ is defined by $x_i = i/n$ for $1 \leq i \leq n$. Two distinct settings have been considered with specific tuning of the related parameters.

- Outer case (see also Section 5.2): The regression function f to be estimated is given for all $x \in [0,1]$ by

$$f(x) = 2\mathbb{1}_{[0.15,0.3[}(x) - \mathbb{1}_{[0.3,0.5[}(x) + \mathbb{1}_{[0.5,0.85[}(x) - \mathbb{1}_{[0.85,1[}(x).$$

  The results are only reported for the Sobolev kernel ($k_S(x,y) = \min(x,y)$, for $x, y \in [0,1]$). The maximum number of iterations, called $T_{\max}$ is respectively equal to 500 if $n \leq 400$, 1 000 if $n = 600$, 2 000 if $n = 800$, and 3 000 if $n = 1\,000$. The step-size of the Landweber algorithm is $\eta = 2.4$, and the emergency stopping time T is chosen such that $T = 2n/\log n$ for τ_{SDP} (see Theorem 17), and $T = T_{\max}$ for τ_{DP}.

- Inner case (see also Section 5.3):

$$f(x) = \frac{1+x}{2}\sin(2\pi x(1+x)).$$

  For the inner case, two reproducing kernels are used: the Sobolev kernel (see above) and the Gaussian kernel ($k_G(x,y) = \exp\left((x-y)^2/w^2\right)$, with bandwidth $w = 0.02$). The maximum number of iterations is $T_{\max} = 500$. The step-size of the Landweber algorithm is respectively set at $\eta = 2.4$ for the Sobolev kernel, and $\eta = 0.5$ for the Gaussian kernel. The emergency stopping time T is chosen such that $T = 4\sqrt{n}$ for τ_{SDP} (see the discussion following Theorem 19 with $r_0 = 1/2$) and $T = T_{\max}$ for τ_{DP}.

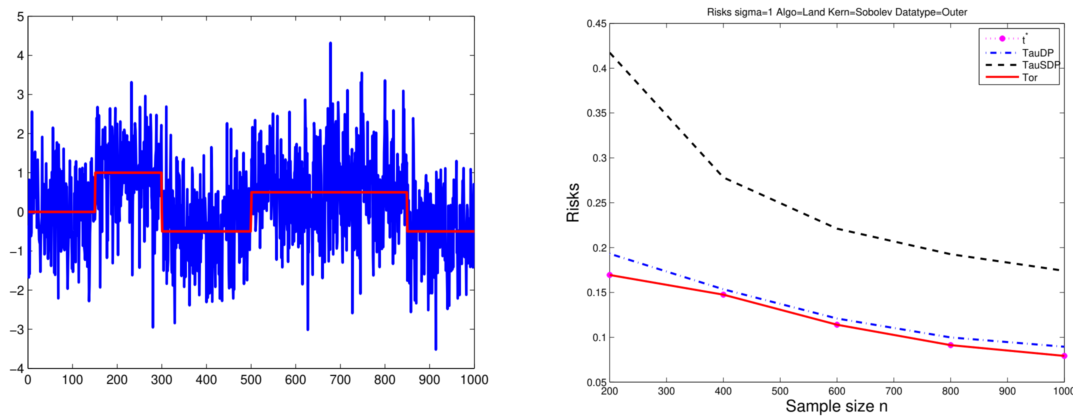
For any given stopping rule $\hat{t}$, its performance is measured by means of the squared empirical norm $\|\mathbf{f} - \hat{\mathbf{f}}^{(\hat{t})}\|_n^2$ averaged over the $N = 200$ replications, which is called the (averaged) "loss" for short in what follows.

6.2 The outer case

Figure 2a displays an example of signal generated from the outer case framework. The piecewise-constant regression function (red curve) makes the estimation problem a difficult task as long as one is limited to using functions from the reproducing kernel Hilbert space (RKHS) generated by the Sobolev kernel k_S. This justifies calling this situation the outer case. For increasing sample sizes, Figure 2b displays the empirical performances (measured in terms of the averaged loss) of several stopping rules, namely τ_{DP}, and τ_{SDP}. They are compared to the performance of the so-called *oracle stopping rule* denoted by t_{or} and defined as a global minimum location of the risk that is,

$$t_{or} = \underset{0 < t \leq T_{\max}}{\operatorname{argmin}} \mathbf{E}_\epsilon \|\mathbf{f} - \hat{\mathbf{f}}^{(t)}\|_n^2. \tag{6.1}$$

Although all the performances improve as the sample size grows, the performance of τ_{DP} still remains uniformly closer to that of t_{or}, than the one of τ_{SDP}. Keeping in mind that τ_{SDP} is known to improve upon τ_{DP} in the case of smooth regression functions (inner



(a) (b)

Figure 2: **(a):** Realization of the Outer case model. Instance of signal generated from the Outer case model. **(b):** Averaged loss performances versus the increasing sample size. Averaged losses of t_{or}, τ_{DP}, and τ_{SDP} in the Outer case. The number of replications is $N = 200$.

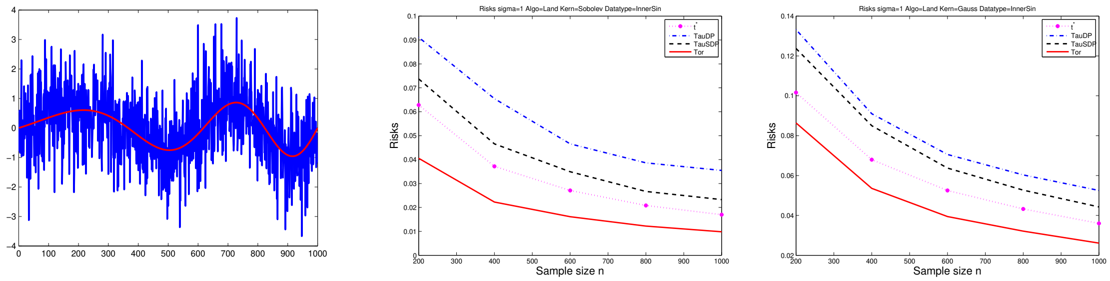
case that is, $r \geq 1/2$), it confirms that the present situation is by contrast a true instance of an outer case ($r < 1/2$), meaning that f is outside the RKHS.

More precisely, since f lies outside the RKHS, the expected number of iterations required for achieving a reliable estimator of f is large. This is what we observe with the oracle stopping rule t_{or} which remains close to the maximum number of iterations $T_{\max}$ as n grows. One main feature in designing τ_{SDP} is the smoothing of the residuals as a means for avoiding too large values of the stopping rule (compared to τ_{DP}). Therefore the present situation is one typical instance where the trend of τ_{DP} to take large values (unlike τ_{SDP}) makes this stopping rule a better candidate.

6.3 The inner case

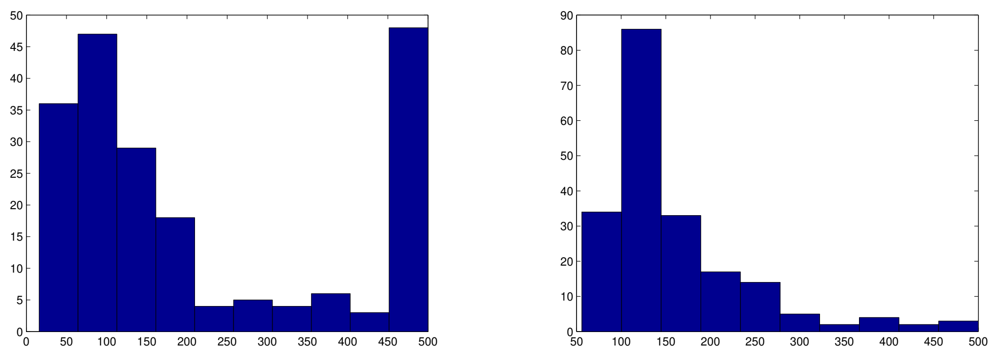
Figure 3a displays an example of signal generated in the inner case. By contrast with the previous example (outer case), the smoothness of the regression function f allows for using both the Gaussian and the Sobolev kernels, respectively denoted by k_G and k_S. Their respective performance are summarized in Figures 3b and 3c, where the different curves display the averaged loss for several stopping rules, namely t_n^*, τ_{DP}, τ_{SDP}, and the oracle stopping rule t_{or} (see Eq. (6.1)).

All the curves from Figures 3b and 3c decrease as n grows. The best performance is uniformly achieved by t_n^*, which is the stopping rule reaching the trade-off between the bias and the (proxy-)variance term (see also Figure 1a). From an asymptotic perspective, this is the best choice one can make in the present early stopping context. In particular, the data-drive stopping rules such as τ_{DP} and τ_{SDP} are estimating t_n^*. It is then consistent that their respective performances are worse than that of t_n^*.



(a) (b) (c)

Figure 3: Averaged losses of t_{or}, t_n^*, τ_{DP}, and τ_{SDP} in the Inner case. The number of replications is $N = 200$. **(a):** Realization of the Inner case model. **(b):** Sobolev kernel. **(c):** Gaussian kernel.



(a) (b)

Figure 4: Empirical distribution of τ_{DP} and τ_{SDP} over $N = 200$ replications for the Sobolev kernel with $n = 800$ in the Inner case. **(a):** Empirical distribution of τ_{DP}. **(b):** Empirical distribution of τ_{SDP}.

For both the kernels k_G and k_S, the worst performance is achieved by τ_{DP}. This sub-optimal behaviour in terms of averaged loss results from the higher variability of τ_{DP} compared to τ_{SDP}, as it can be observed from the histograms of Figures 4a and 4b obtained with $n = 800$ and $T_{\max} = 500$. In particular, this emphasizes that the residual smoothing encoded within the τ_{SDP} stopping rule induces a considerable variance reduction, which avoids stopping too late (and then wasting time).

7. Proofs for fixed-design results

In this section, we analyze the discrepancy principle conditional on the design.

7.1 A unified framework

A linearly transformed model is now introduced for simultaneously dealing with the smoothed and non-smoothed cases.

$$\tilde{\mathbf{Y}} = L_n\mathbf{Y} = L_n\mathbf{f} + L_n\boldsymbol{\epsilon} = \tilde{\mathbf{f}} + \tilde{\boldsymbol{\epsilon}}, \qquad L_n \in \mathbb{R}^{n\times n},$$

with L_n satisfying $\|L_n\|_{\mathrm{op}} \le 1$. The new noise variable $\tilde{\boldsymbol{\epsilon}}$ is mean-zero and has covariance matrix $\sigma^2 L_n L_n^T$. For a regularizer g in the sense of Definition 2, our main goal is to analyze the stopping rule τ defined by

$$\tau = \inf\left\{t \ge 0 : \|\tilde{\mathbf{Y}} - K_n g_t(K_n)\tilde{\mathbf{Y}}\|_n^2 = \|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \le \frac{\sigma^2 \operatorname{tr}(L_n L_n^T)}{n}\right\} \wedge T \tag{7.1}$$

with $T \in (0,\infty]$. For $L_n = I_n$ the stopping rule in (7.1) coincides with τ_{DP} from (2.8), while for $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$ with regularizer $\tilde{g}$ it coincides with τ_{SDP} from (4.1).

Moreover, the stopping rule (7.1) can be interpreted as applying the classical discrepancy principle to the smoothed data $\tilde{\mathbf{Y}}$ and the class of estimators $K_n g_t(K_n)\tilde{\mathbf{Y}} = S_n g_t(\Sigma_n) S_n^* \tilde{\mathbf{Y}}$ where spectral regularization is applied to the smoothed data.

Definition 21 *For every $t \ge 0$ and every regularizer g, we define*

$$\widetilde{\mathcal{N}}_n^g(t) = \operatorname{tr}(L_n L_n^T K_n g_t(K_n)).$$

Lemma 22 (Basic inequality) *Assumption* (**BdF**) *yields, for every $t \ge 0$,*

$$\begin{aligned}&\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 - 2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t)\\ &\le \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) \le \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 - \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t).\end{aligned}$$

Since g is a regularizer, the term

$$\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 = \sum_{j=1}^n r_t^2(\hat{\lambda}_j)\langle \hat{v}_j, \tilde{\mathbf{f}}\rangle_n^2$$

is continuous and non-increasing in $t \ge 0$, while the term

$$\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) = \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T g_t(K_n)K_n) = \frac{\sigma^2}{n}\sum_{j=1}^n \|L_n^T \hat{v}_j\|_2^2 \hat{\lambda}_j g_t(\hat{\lambda}_j)$$

is continuous, non-decreasing in $t \ge 0$ and equal to zero for $t = 0$. Hence, we can define the following balancing stopping rule

$$\widetilde{t}_n^* = \inf\left\{t \ge 0 : \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 = \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t)\right\}. \tag{7.2}$$

If such a t exists, then we have $\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 = \sigma^2 n^{-1}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*)$. Otherwise, we set $\widetilde{t}_n^* = \infty$ in which case we still have $\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 = \lim_{t\to\infty}\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \ge \lim_{t\to\infty}\sigma^2 n^{-1}\widetilde{\mathcal{N}}_n^g(t) = \sigma^2 n^{-1}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*)$.

Proof of Lemma 22 We have

$$\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 = \|r_t(K_n)\tilde{\mathbf{f}} + r_t(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2$$

and thus, using $\tilde{\boldsymbol{\epsilon}} = L_n \boldsymbol{\epsilon}$ and $r_t(K_n) = I - K_n g_t(K_n)$,

$$\begin{aligned}\mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 &= \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T (I - K_n g_t(K_n))^2) \\ &= \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) \\ &- 2\frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T g_t(K_n) K_n) + \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T g_t^2(K_n) K_n^2).\end{aligned}$$

The lower bound follows from the fact that the last term is non-negative, while the upper bound follows from (**BdF**). □

7.2 Deviation inequality for the variance and bias parts

The results of this section are improvements over previous results from (Blanchard et al., 2018b) and (Blanchard et al., 2018a), differentiating more precisely between the sub-Gaussian and sub-exponential behaviours. Surprisingly, these improvements result from different arguments based on a more basic comparison between the discrepancy principle and its reference balancing stopping rule $\widetilde{t}_n^*$.

7.2.1 Deviation inequality for the variance part

Our first main result is a deviation inequality for τ from (7.1).

Proposition 23 *Suppose that Assumptions* (**SubGN**) *and* (**BdF**) *hold. Then, for every* $t > \widetilde{t}_n^*$, *we have*

$$\mathbf{P}_\epsilon(\tau > t) \le 2\exp\Big(-c\Big(y \wedge \frac{y^2}{\operatorname{tr}(L_n L_n^\top)}\Big)\Big), \qquad y = \widetilde{\mathcal{N}}_n^g(t) - \widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*),$$

where $c > 0$ *is a constant depending only on* A.

Proof of Proposition 23 Since the claim is trivial for $\widetilde{t}_n^* = \infty$ it remains to consider the case that $\widetilde{t}_n^*$ is finite. Inserting the definition of the discrepancy principle in (7.1), we have

$$\mathbf{P}_\epsilon(\tau > t) \le \mathbf{P}_\epsilon\Big(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 > \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T)\Big). \tag{7.3}$$

By Lemma 22, we have

$$\frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \ge \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) - \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2.$$

Since $t > \widetilde{t}_n^*$ implies that

$$\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \le \|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 = \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*),$$

we arrive at

$$\frac{\sigma^2}{n}\operatorname{tr}(L_nL_n^T) - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \geq \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) - \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) = \frac{\sigma^2}{n}y.$$

Inserting this into (7.3), we get

$$\mathbf{P}_\epsilon(\tau > t) \leq \mathbf{P}_\epsilon\Big(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 > \frac{\sigma^2}{n}y\Big).$$

Applying Lemma 27, using also that $t > \widetilde{t}_n^*$ and (**BdF**) imply $\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \leq \sigma^2 n^{-1}\widetilde{\mathcal{N}}_n^g(t) \leq \sigma^2 n^{-1}\operatorname{tr}(L_nL_n^T)$, the claim follows. □

Our next main result is a deviation inequality for the variance part.

Proposition 24 *Suppose that* (**SubGN**) *holds true. Then, for every $y > 0$, we have*

$$\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_\tau^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + \frac{\sigma^2}{n}2y\Big)\Big) \leq 3\exp\Big(-c\Big(y \wedge \frac{y^2}{\operatorname{tr}(L_nL_n^\top)}\Big)\Big)$$

with constant $c > 0$ depending only on A.

Proof of Proposition 24 By Definition 2, the term $\|K_n^{1/2}g_t^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2$ is non-decreasing in $t \geq 0$. Now, if

$$\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + y > \widetilde{\mathcal{N}}_n^g(T), \tag{7.4}$$

then

$$\begin{aligned}&\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_\tau^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + 2\frac{\sigma^2}{n}y\Big)\\ &\leq \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_T^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(T) + \frac{\sigma^2}{n}y\Big),\end{aligned}$$

and the claim follows from Lemma 28. On the other hand, if (7.4) does not hold, then we can define $\widetilde{t}_n^* < t \leq T$ by

$$\widetilde{\mathcal{N}}_n^g(t) = \widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + y. \tag{7.5}$$

In this case we have

$$\begin{aligned}&\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_\tau^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + 2\frac{\sigma^2}{n}y\Big)\\ &\leq \mathbf{P}_\epsilon\Big(\{\tau \leq t\} \cap \Big\{\|K_n^{1/2}g_\tau^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + 2\frac{\sigma^2}{n}y\Big\}\Big) + \mathbf{P}_\epsilon(\tau > t)\\ &\leq \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_t^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) + \frac{\sigma^2}{n}y\Big) + \mathbf{P}_\epsilon(\tau > t),\end{aligned}$$

and the claim follows from applying Lemma 28 to the second last term and Proposition 23 to the last term, using that $t > \widetilde{t}_n^*$ and $y = \widetilde{\mathcal{N}}_n^g(t) - \widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*)$. □

7.2.2 Deviation inequality for the bias part

Proposition 25 *Suppose that Assumptions* (**SubGN**) *and* (**BdK**) *hold true. Then, for every $y > 0$ such that $2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \sigma^2 n^{-1} y > \|r_T(K_n)\tilde{\mathbf{f}}\|_n^2$, we have*

$$\mathbf{P}_\epsilon\Big(\|r_\tau(K_n)\tilde{\mathbf{f}}\|_n^2 > 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n} y\Big) \leq 2\exp\Big(-c\Big(\frac{y^2}{\operatorname{tr}(L_n L_n^T)} \wedge y\Big)\Big) \tag{7.6}$$

with constant $c > 0$ depending only on A. If $\widetilde{t}_n^* = \infty$ we additionally assume that $\sigma^2 n^{-1} y \geq \|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2$.

Proof of Proposition 25 From $2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \sigma^2 n^{-1} y > \|r_T(K_n)\tilde{\mathbf{f}}\|_n^2$ it follows that, under the event considered in (7.6), the stopping rule τ has to be smaller than T. This means that in the definition of τ in (7.1), we can ignore the minimum with T in what follows.

If $\|r_0(K_n)\tilde{\mathbf{f}}\|_n^2 < 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \sigma^2 n^{-1} y$, then the claim is clear because the probability on the left-hand side of (7.6) is equal to zero. Otherwise, we define $0 \leq t < \widetilde{t}_n^*$ by

$$2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n} y = \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2,$$

leading to

$$\begin{aligned}&\mathbf{P}_\epsilon\Big(\|r_\tau(K_n)\tilde{\mathbf{f}}\|_n^2 > 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n} y\Big)\\ &\leq \mathbf{P}_\epsilon(\tau < t) \leq \mathbf{P}_\epsilon\Big(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq \frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T)\Big).\end{aligned} \tag{7.7}$$

By Lemma 22, we have

$$\frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq 2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) - \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2.$$

Since $t < \widetilde{t}_n^*$, (7.2) implies

$$2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) \leq 2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) \leq 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2.$$

Thus we get

$$\frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 - \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 = -\frac{\sigma^2}{n} y.$$

Inserting this into (7.7), we get

$$\mathbf{P}_\epsilon\Big(\|r_\tau(K_n)\tilde{\mathbf{f}}\|_n^2 > \frac{\sigma^2}{n} y\Big) \leq \mathbf{P}_\epsilon\Big(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq -\frac{\sigma^2}{n} y\Big),$$

In the case that $\widetilde{t}_n^*$ is finite, using that

$$\begin{aligned}\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 &= 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n} y\\ &= 2\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + \frac{\sigma^2}{n} y \leq 2\frac{\sigma^2}{n}\operatorname{tr}(L_n L_n^T) + \frac{\sigma^2}{n} y,\end{aligned}$$

the claim follows from Lemma 27. On the other hand, if $\widetilde{t}_n^* = \infty$, then we use $\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 = 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \sigma^2 n^{-1} y \leq 3\sigma^2 n^{-1} y$ instead. □

7.3 Proofs of oracle inequalities (fixed-design)

The present section gathers proofs of main oracle inequalities established in the fixed-design setting. They mainly follow from the results from Section 7.2. In each of these proofs, notations are used according to the context where the theorem has been stated.

Proof of Proposition 10 The proof follows from Sections 7.1 and 7.2 applied with $L_n = I_n$, in which case τ_{DP} coincides with τ from (7.1) and t_n^* coincides with $\widetilde{t_n^*}$ from (7.2).

By (**BdF**) and using that $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} \|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{DP})}\|_n^2 &\leq 2\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2 + 2\|K_n g_{\tau_{DP}}(K_n)\boldsymbol{\epsilon}\|_n^2 \\ &\leq 2\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2 + 2\|K_n^{1/2} g_{\tau_{DP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2. \end{aligned} \tag{7.8}$$

Proposition 24 with $L_n = I_n$ yields that, for every $u > 0$,

$$\mathbf{P}_\epsilon\Big(\|K_n^{1/2} g_{\tau_{DP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\mathcal{N}_n^g(t_n^*) + C\Big(\frac{\sigma^2\sqrt{u}}{\sqrt{n}} + \frac{\sigma^2 u}{n}\Big)\Big) \leq 3e^{-u}. \tag{7.9}$$

On the other hand, from Proposition 25 with $L_n = I_n$, it follows that

$$\mathbf{P}_\epsilon\Big(\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2 > 2\|r_{t_n^*\wedge T}(K_n)\mathbf{f}\|_n^2 + C\Big(\frac{\sigma^2\sqrt{u}}{\sqrt{n}} + \frac{\sigma^2 u}{n}\Big)\Big) \leq 2e^{-u}. \tag{7.10}$$

By the definition of t_n^*, we have

$$\|r_{t_n^*\wedge T}(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t_n^*) \leq 2\min_{0\leq t\leq T}\Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t)\Big\}. \tag{7.11}$$

Using (7.8) and (7.11) combined with (7.9) and (7.10), and the union bound, the claim now follows. □

Proof of Theorem 12 The claim follows from inserting Lemma 11 into Theorem 12. □

Proof of Theorem 15 The result follows from Sections 7.1 and 7.2 applied with $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$, in which case τ_{SDP} from (4.1) coincides with τ from (7.1).

A key remark is that, since the regularizer $\tilde{g}$ satisfies (**LFL**), we have $\lambda g_T(\lambda) \leq (B \vee 1)b^{-1}\lambda\tilde{g}_T(\lambda)$. Thus $\tau_{SDP} \leq T$ implies

$$\begin{aligned} &\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{SDP})}\|_n^2 \\ &\leq 2\|r_{\tau_{SDP}}(K_n)\mathbf{f}\|_n^2 + 2\|K_n g_{\tau_{SDP}}(K_n)\boldsymbol{\epsilon}\|_n^2 \\ &\leq 2\|r_{\tau_{SDP}}(K_n)\mathbf{f}\|_n^2 + 2(B\vee 1)b^{-1}\|K_n^{1/2} g_{\tau_{SDP}}^{1/2}(K_n)K_n^{1/2}\tilde{g}_T^{-1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 \\ &= 2\|r_{\tau_{SDP}}(K_n)\mathbf{f}\|_n^2 + 2(B\vee 1)b^{-1}\|K_n^{1/2} g_{\tau_{SDP}}^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2, \end{aligned} \tag{7.12}$$

where $\boldsymbol{\epsilon}$ has been replaced by $\tilde{\boldsymbol{\epsilon}}$ in the last inequality. Invoking the first claim of Lemma 29 we get

$$\begin{aligned} &\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{SDP})}\|_n^2 \\ &\leq C\Big(\|r_{\tau_{SDP}}(K_n)\tilde{\mathbf{f}}\|_n^2 + \|K_n^{1/2} g_{\tau_{SDP}}^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 + \frac{1}{T^{2s+1}} + \frac{\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}}{T}\Big), \end{aligned} \tag{7.13}$$

where the last term $CT^{-1}\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}$ can be dropped if $s \leq 1/2$. On the one hand, Proposition 25 with $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$ and Lemma yields that

$$\mathbf{P}_\epsilon\Big(\|r_{\tau_{SDP}}(K_n)\tilde{\mathbf{f}}\|_n^2 > 2\|r_{\widetilde{t_n^*}\wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + C\frac{\sigma^2}{n}\Big(\sqrt{u\mathcal{N}_n^{\tilde{g}}(T)} + u\Big)\Big) \leq 2e^{-u}, \quad u > 0. \quad (7.14)$$

On the other hand, from Proposition 24 with $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$, we get

$$\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}) + C\frac{\sigma^2}{n}\Big(\sqrt{u\mathcal{N}_n^{\tilde{g}}(T)} + u\Big)\Big) \leq 3e^{-u}, \quad u > 0. \quad (7.15)$$

The definition of $\widetilde{t_n^*}$ and (**BdF**) lead to

$$\|r_{\widetilde{t_n^*}\wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}) \leq 2\min_{0\leq t\leq T}\Big\{\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t)\Big\}.$$

Now, using (**BdF**), we have $\widetilde{\mathcal{N}}_n^g(t) \leq \mathcal{N}_n^g(t)$ and $\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \leq \|r_t(K_n)\mathbf{f}\|_n^2$. Thus combining everything together yields

$$\|r_{\widetilde{t_n^*}\wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}) \leq 2\min_{0\leq t\leq T}\Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n^g(t)\Big\}. \quad (7.16)$$

Using (7.13) and (7.16) combined with (7.14) and (7.15), and the union bound, we get for every $u > 0$

$$\mathbf{P}_\epsilon\Big(\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau_{SDP})}\|_n^2 > C\Big(\min_{0\leq t\leq T}\Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\sigma^2}{n}\mathcal{N}_n(t)\Big\} + \frac{\sigma^2\sqrt{u\mathcal{N}_n^{\tilde{g}}(T)}}{n} + \frac{\sigma^2 u}{n} + \frac{1}{T^{2s+1}} + \frac{\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}}{T}\Big)\Big) \leq 5e^{-u} \quad (7.17)$$

The desired inequality now follows from inserting Lemmas 6 and 11. □

Remark 26 *Let us briefly outline one possible strategy for extending our oracle inequalities to an early stopping rule*

$$\tau = \tau(\mathbf{Y}, \hat{\sigma}^2, L_n, T) = \inf\Big\{t \geq 0 : \|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq \frac{\hat{\sigma}^2\,\mathrm{tr}(L_nL_n^T)}{n}\Big\} \wedge T \quad (7.18)$$

which is based on an estimator $\hat{\sigma}^2$ of σ^2.

First, focusing on the event that the inequalities $\kappa_1 \leq \hat{\sigma}^2 \leq \kappa_2$ hold for some fixed $\kappa_1, \kappa_2 > 0$, it is easy to see (using (7.8) and the monotonicity of the bias and variance terms) that the squared empirical norm $\|\mathbf{f} - \hat{\mathbf{f}}^{(\tau)}\|_n^2$ is bounded by $2\|r_{\tau_2}(K_n)\mathbf{f}\|_n^2 + 2\|K_ng_{\tau_1}(K_n)\boldsymbol{\epsilon}\|_n^2$, where the stopping rules $\tau_2 \leq \tau \leq \tau_1$ are defined by

$$\tau_i = \tau_i(\mathbf{Y}, \kappa_i, L_n, T) = \inf\Big\{t \geq 0 : \|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \leq \frac{\kappa_i\,\mathrm{tr}(L_nL_n^T)}{n}\Big\} \wedge T.$$

The main observation now is that the deviation inequalities from Section 7.2 can be extended to the stopping rules τ_i, provided that the κ_i are sufficiently close to σ^2.

In fact, following the same line of arguments, one can show that if $|\kappa_i - \sigma^2| \leq C\sigma^2$ *for some constant* $C > 0$*, then Proposition 24 continues to hold for* τ_i *with*

$$\frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + \frac{\sigma^2}{n}2y \quad \textit{replaced by} \quad \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + \frac{\sigma^2}{n}2y + \frac{\max(\sigma^2 - \kappa_i, 0)}{n}\operatorname{tr}(L_n L_n^T),$$

while Proposition 25 continues to hold for τ_i *with*

$$2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}y \quad \textit{replaced by} \quad 2\|r_{\widetilde{t}_n^*}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}y + \frac{\max(\kappa_i - \sigma^2, 0)}{n}\operatorname{tr}(L_n L_n^T).$$

Hence, the stopping rules τ_i *also lead to oracle inequalities, but, compared to the stopping rule based on* σ^2*, with an additional error term* $|\sigma^2 - \kappa_i|n^{-1}\operatorname{tr}(L_n L_n^T)$*. In particular, Proposition 10, Theorem 12 and Theorem 15 continue to hold with the additional remainder terms* $|\sigma^2 - \kappa_i|$ *and* $|\sigma^2 - \kappa_i|n^{-1}\mathcal{N}_n^{\tilde{g}}(T)$*, respectively. In summary, we get oracle inequalities for the stopping rule* τ *based on an estimator* $\hat{\sigma}^2$*, provided that we have sufficiently sharp high probability upper and lower bounds for* $\hat{\sigma}^2$*.*

7.4 Key technical results

In order to prove Propositions 24 and 25, we need the following two concentration inequalities, namely Lemmas 27 and 28.

Lemma 27 *Suppose that Assumption* (**SubGN**) *holds. Then, for every* $t \geq 0$ *and every* $y > 0$*, we have*

$$\begin{aligned}
&\mathbf{P}_\epsilon(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 > y) \\
&\leq \exp\Big(- c\Big(\frac{n^2y^2}{\sigma^4\operatorname{tr}(L_n L_n^T)} \wedge \frac{ny}{\sigma^2}\Big)\Big) + \exp\Big(- \frac{cny^2}{\sigma^2\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2}\Big)
\end{aligned}$$

and the same upper bound holds for $\mathbf{P}_\epsilon(\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 < -y)$*.*

Proof of Lemma 27 We have

$$\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 = \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + \langle r_t(K_n)\tilde{\mathbf{f}}, r_t(K_n)\tilde{\boldsymbol{\epsilon}}\rangle_n + \|r_t(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2$$

and thus

$$\begin{aligned}
&\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)\tilde{\mathbf{Y}}\|_n^2 \\
&= \langle L_n^T r_t^2(K_n)\tilde{\mathbf{f}}, \boldsymbol{\epsilon}\rangle_n + \|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2.
\end{aligned}$$

By (**SubGN**) and a general Hoeffding inequality for sub-Gaussian random variables (cf. (Vershynin, 2018, Theorem 2.6.3)), we have for all $y > 0$,

$$\begin{aligned}
\mathbf{P}_\epsilon(\langle L_n^T r_t^2(K_n)\tilde{\mathbf{f}}, \boldsymbol{\epsilon}\rangle_n > y) &\leq \exp\Big(- \frac{cn^2y^2}{\sigma^2\|L_n^T r_t^2(K_n)\tilde{\mathbf{f}}\|_2^2}\Big) \\
&\leq \exp\Big(- \frac{cny^2}{\sigma^2\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2}\Big),
\end{aligned}$$

where we used the fact that $\|L_n^T\|_{\text{op}} = \|L_n\|_{\text{op}} \le 1$ and (**BdF**) in the second inequality. Moreover, an application of the Hanson-Wright inequality (cf. (Vershynin, 2018, Theorem 6.2.1)) gives for all $y > 0$,

$$\begin{aligned} &\mathbf{P}_\epsilon(\|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2 > y) \\ &\le \exp\Big(-c\Big(\frac{n^2y^2}{\sigma^4\|L_n^T r_t^2(K_n)L_n\|_{\text{HS}}^2} \wedge \frac{ny}{\sigma^2\|L_n^T r_t^2(K_n)L_n\|_{\text{op}}}\Big)\Big). \end{aligned}$$

By Assumption (**BdF**) and the fact that $\|L_n\|_{\text{op}} = \|L_n^T\|_{\text{op}} \le 1$, we have

$$\|L_n^T r_t^2(K_n)L_n\|_{\text{op}} \le 1 \quad \text{and} \quad \|L_n^T r_t^2(K_n)L_n\|_{\text{HS}}^2 \le \|L_n\|_{\text{HS}}^2 = \operatorname{tr}(L_nL_n^T).$$

We thus obtain that

$$\begin{aligned} &\mathbf{P}_\epsilon(\|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2 - \mathbf{E}_\epsilon\|r_t(K_n)L_n\boldsymbol{\epsilon}\|_n^2 > y) \\ &\le \exp\Big(-c\Big(\frac{n^2y^2}{\sigma^4\operatorname{tr}(L_nL_n^T)} \wedge \frac{ny}{\sigma^2}\Big)\Big). \end{aligned}$$

This completes the proof of the right-deviation inequality. The left-deviation inequality follows analogously. □

Lemma 28 *Suppose that Assumption* (**SubGN**) *holds. Then, for every $t \ge 0$ and every $y > 0$, we have*

$$\begin{aligned} \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_t^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(t) + y\Big) &\le \exp\Big(-c\Big(\frac{n^2y^2}{\sigma^4\widetilde{\mathcal{N}}_n^g(t)} \wedge \frac{ny}{\sigma^2}\Big)\Big) \\ &\le \exp\Big(-c\Big(\frac{n^2y^2}{\sigma^4\operatorname{tr}(L_nL_n^T)} \wedge \frac{ny}{\sigma^2}\Big)\Big). \end{aligned}$$

Proof of Lemma 28 First, note that $\sigma^2 n^{-1}\widetilde{\mathcal{N}}_n^g(t) = \mathbf{E}_\epsilon\|K_n^{1/2}g_t^{1/2}(K_n)\tilde{\boldsymbol{\epsilon}}\|_n^2$. Moreover, by the Hanson-Wright inequality (cf. (Vershynin, 2018, Theorem 6.2.1)), we have for all $y > 0$,

$$\begin{aligned} &\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_t^{1/2}(K_n)L_n\boldsymbol{\epsilon}\|_n^2 > \mathbf{E}_\epsilon\|K_n^{1/2}g_t^{1/2}(K_n)L_n\boldsymbol{\epsilon}\|_n^2 + y\Big) \\ &\le \exp\Big(-c\Big(\frac{n^2y^2}{\sigma^4\|L_n^TK_ng_t(K_n)L_n\|_{\text{HS}}^2} \wedge \frac{ny}{\sigma^2\|L_n^TK_ng_t(K_n)L_n\|_{\text{op}}}\Big)\Big). \end{aligned}$$

The claims now follow from inserting $\|L_n^TK_ng_t(K_n)L_n\|_{\text{op}} \le 1$ as well as $\|L_n^TK_ng_t(K_n)L_n\|_{\text{HS}}^2 \le \operatorname{tr}(L_nL_n^TK_ng_t(K_n)) \le \operatorname{tr}(L_nL_n^T)$. □

Lemma 29 *Let $L_n = \tilde{g}_T^{1/2}(K_n)K_n^{1/2}$ with regularizer $\tilde{g}$ satisfying* (**LFL**)*. If* (**SC**(r,R)) *holds with $s = r - 1/2 \ge 0$ and if $\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\text{op}} \le 1/2$, then there is a constant $C > 0$ depending only on s, R and M such that for every $0 < t \le T$,*

$$\|r_t(K_n)\mathbf{f}\|_n^2 \le \frac{1}{b}\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + C\Big(\frac{1}{T^{2s+1}} + \frac{\|\Sigma_n - \Sigma\|_{\text{op}}^{2\wedge 2s}}{T}\Big),$$

where the last term in the upper bound $CT^{-1}\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}$ *can be dropped if* $s \le 1/2$. *Moreover, if* (**SC**(r,R)) *and* (**QuErr**) *hold with* $s = r - 1/2 \ge 0$ *and* $q \ge r$ *and if* $\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_{\mathrm{op}} \le 1/2$, *then we have for every* $0 < t \le T$,

$$\|r_t(K_n)\mathbf{f}\|_n^2 \le C\Big(\frac{1}{t^{2s+1}} + \frac{\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}}{t}\Big),$$

where the second term in the upper bound can be dropped if $s \le 1/2$.

Proof of Lemma 29 Using the identity $\mathbf{f} = S_n f$ and the singular value decomposition in (2.3), we have

$$\begin{aligned}
\|r_t(K_n)\mathbf{f}\|_n^2 &= \sum_{j\ge 1} \hat\lambda_j r_t^2(\hat\lambda_j)\langle f, \hat u_j\rangle^2 \\
&\le \frac{1}{b}\sum_{\hat\lambda_j T > 1} \lambda_j \tilde g_T(\hat\lambda_j)\hat\lambda_j r_t^2(\hat\lambda_j)\langle f, \hat u_j\rangle^2 + \frac{1}{T}\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2 \\
&= \frac{1}{b}\|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{1}{T}\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2,
\end{aligned}$$

where we applied (**LFL**) and (**BdF**) in the inequality. To see the first claim, we have show that

$$\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2 \le C(T^{-2s} + \|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}), \tag{7.19}$$

where the second term $\|\Sigma_n - \Sigma\|_{\mathrm{op}}^{2\wedge 2s}$ can be dropped if $s \le 1/2$. By assumption $\|\Sigma - \Sigma_n\|_{\mathrm{op}} \le (\lambda_1 + T^{-1})/2$. By assumption, we have $f = \Sigma^s g$ with $\|g\|_{\mathcal{H}} \le R$ and $s = r - 1/2 \ge 0$. Hence,

$$\begin{aligned}
\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2 &\le 2\sum_{\hat\lambda_j T\le 1}\langle \Sigma_n^s g, \hat u_j\rangle^2 + 2\sum_{\hat\lambda_j T\le 1}\langle (\Sigma^s - \Sigma_n^s) g, \hat u_j\rangle^2 \\
&\le 2\sum_{\hat\lambda_j T\le 1}\hat\lambda_j^{2s}\langle g, \hat u_j\rangle^2 + 2\|(\Sigma^s - \Sigma_n^s)g\|_{\mathcal{H}}^2 \\
&\le 2T^{-2s}\|g\|_{\mathcal{H}}^2 + C\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}\|g\|_{\mathcal{H}}^2,
\end{aligned}$$

where we applied (A.1) and (A.2) in the last inequality and where $C > 0$ is a constant depending only on s and M. If $s \le 1/2$, then we have

$$\begin{aligned}
&\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2 = \sum_{\hat\lambda_j T\le 1}\langle (\Sigma_n + T^{-1})^s(\Sigma_n + T^{-1})^{-s}\Sigma^s g, \hat u_j\rangle^2 \\
&\le (2T^{-1})^{2s}\|(\Sigma_n + T^{-1})^{-s}\Sigma^s\|_{\mathrm{op}}^2 R^2 \le (2T^{-1})^{2s}\|(\Sigma_n + T^{-1})^{-1/2}\Sigma^{1/2}\|_{\mathrm{op}}^{2s} R^2,
\end{aligned}$$

where we applied (A.3) in the last inequality and where $C > 0$ is a constant depending only on s, M and R. Hence, the second part of the claim follows from

$$\|(\Sigma_n + T^{-1})^{-1/2}\Sigma^{1/2}\|_{\mathrm{op}}^2 \le \|(\Sigma_n + T^{-1})^{-1/2}(\Sigma + T^{-1})^{1/2}\|_{\mathrm{op}}^2$$

$$
\begin{aligned}
&= \|(\Sigma + T^{-1})^{1/2}(\Sigma_n + T^{-1})^{-1}(\Sigma + T^{-1})^{1/2}\|_{\mathrm{op}} \\
&= \|((\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2} + 1)^{-1}\|_{\mathrm{op}} \leq 2.
\end{aligned} \tag{7.20}
$$

The proof of the last claim is very similar. Using (**QuErr**) with $q \geq r$, (A.1) and (A.2), we get

$$
\begin{aligned}
\|\Sigma_n^{1/2} r_t(\Sigma_n) f\|_{\mathcal{H}}^2 &\leq 2\|\Sigma_n^{1/2} r_t(\Sigma_n)\Sigma_n^s g\|_{\mathcal{H}}^2 + 2\|\Sigma_n^{1/2} r_t(\Sigma_n)(\Sigma_n^s - \Sigma^s) g\|_{\mathcal{H}}^2 \\
&\leq C(t^{-1-2s} + t^{-1}\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}),
\end{aligned}
$$

and the second part of the last claim follows. On the other hand, if $s \leq 1/2$, then we have

$$
\begin{aligned}
&\|\Sigma_n^{1/2} r_t(\Sigma_n) f\|_{\mathcal{H}}^2 \leq \|\Sigma_n^{1/2} r_t(\Sigma_n)(\Sigma_n + T^{-1})^s (\Sigma_n + T^{-1})^{-s}\Sigma^s g\|_{\mathcal{H}}^2 \\
&\leq C_1 \|\Sigma_n^{1/2} r_t(\Sigma_n)(\Sigma_n + t^{-1})^s\|_{\mathrm{op}}^2 \|(\Sigma_n + t^{-1})^{-1/2}(\Sigma + t^{-1})^{1/2}\|_{\mathrm{op}}^{2s} \leq C_2 t^{-1-2s},
\end{aligned}
$$

where we applied (**QuErr**) and (7.20). □

8. Proofs for random design results

8.1 Concentration inequalities

In this section, we provide concentration and deviation inequalities needed to transfer our results from the fixed to the random design setting. We start with a deviation inequality dealing with the change of norm event from Lemma 16. The next lemma follows from an extension of Tropp (2015) obtained in Minsker (2017) and further simplified by Dicker et al. (2017) (see Lemma 45).

Lemma 30 *Suppose that* (**BdK**) *holds. For $t > 0$, let $\mathcal{E}_t$ be the event defined by*

$$
\mathcal{E}_t = \{\|(\Sigma + t^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + t^{-1})^{-1/2}\|_{\mathrm{op}} \leq 1/2\}.
$$

Then there are constants $c_1, c_2, C_1 > 0$ depending only on M such that, for every $0 < t \leq c_2 n$,

$$
\mathbb{P}(\mathcal{E}_t^c) \leq C_1 t \exp(-c_1 n/t).
$$

Proof of Lemma 30 The proof consists in checking the assumptions of Lemma 45 from the Appendix. This justifies introducing constants R, V, and D from Lemma 45. In particular

$$
\xi_i = (\Sigma + t^{-1})^{-1/2} k_{X_i} \otimes (\Sigma + t^{-1})^{-1/2} k_{X_i} - (\Sigma + t^{-1})^{-1}\Sigma.
$$

Then $\|\xi_1\|_{\mathrm{op}} \leq 2\|(\Sigma + t^{-1})^{-1/2} k_{X_1}\|_{\mathcal{H}}^2 \leq 2M^2 t = R$. Moreover, we have

$$
\begin{aligned}
\|\mathbb{E}\xi_1^2\|_{\mathrm{op}} &\leq \left\|\mathbb{E}\Big((\Sigma + t^{-1})^{-1/2} k_X \otimes (\Sigma + t^{-1})^{-1/2} k_X\Big)^2\right\|_{\mathrm{op}} \\
&\leq \left\|\mathbb{E}\langle(\Sigma + t^{-1})^{-1} k_{X_1}, k_X\rangle_{\mathcal{H}} (\Sigma + t^{-1})^{-1/2} k_X \otimes (\Sigma + t^{-1})^{-1/2} k_X\right\|_{\mathrm{op}} \\
&\leq tM^2 \left\|\mathbb{E}(\Sigma + t^{-1})^{-1/2} k_X \otimes (\Sigma + t^{-1})^{-1/2} k_X\right\|_{\mathrm{op}}
\end{aligned}
$$

$$= tM^2 \left\| (\Sigma + t^{-1})^{-1}\Sigma \right\|_{\mathrm{op}} \leq tM^2 = V.$$

Similarly with $D = \mathcal{N}(t)$, we have

$$\operatorname{tr}(\mathbb{E}\xi_1^2) \leq tM^2 \operatorname{tr}((\Sigma + t^{-1})^{-1}\Sigma) = tM^2\mathcal{N}(t) = V \cdot D.$$

Then, for every $t > 0$ such that $V^{1/2}n^{-1/2} + (3n)^{-1}R \leq 1/2$,

$$\begin{aligned} \mathbb{P}\left[\left\| \frac{1}{n}\sum_{i=1}^{n} \xi_i \right\|_{\mathrm{op}} \geq \frac{1}{2} \right] &\leq 4\mathcal{N}(t) \exp\left[-\frac{n}{8\,(M^2 + (2/6)M^2)\,t} \right] \\ &\leq 4M^2 t \exp\left[-\frac{n}{(32/3)M^2 t} \right], \end{aligned}$$

where the last inequality results from (**BdK**), which yields the claim with $C_1 = 4M^2$, $c_1 = (32/3)M^2$, and $c_2 = (3/4)^2(\sqrt{7/3} - 1)^2/M^2$. □

Next, we establish a concentration inequality for the empirical effective dimension. Interestingly, the event $\mathcal{E}_T$ again plays a key role. Besides, we will apply exponential-type inequalities in Hilbert spaces.

Lemma 31 *Suppose that* (**BdK**) *holds. Then there is a constant $C > 0$ depending only on M and $\lambda_1 = \|\Sigma\|_{\mathrm{op}}$ such that, for every $1 \leq t \leq T$,*

$$\mathbb{P}\Big(\mathcal{E}_T \cap \Big\{ \mathcal{N}_n(t) > C\mathcal{N}(t) \Big\} \Big) \leq e^{-n/t}.$$

In particular, for every $1 \leq t \leq T$, we have

$$\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\mathcal{N}_n(t) \leq C\mathcal{N}(t) + ne^{-n/t}.$$

Remark 32 *Lemma 31 deals only with the case $t \geq 1$. The reason for this is that for $0 < t \leq 1$, the trivial bound $\mathcal{N}_n(t) \leq M^2 t \leq M^2$ will be sufficient for our purposes.*

Proof of Lemma 31 Setting

$$A_t = (\Sigma + t^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + t^{-1})^{-1/2}, \tag{8.1}$$

we have

$$(\Sigma_n + t^{-1})^{-1} = (\Sigma + t^{-1})^{-1/2}(I + A_t)^{-1}(\Sigma + t^{-1})^{-1/2}.$$

Hence,

$$\begin{aligned} \mathcal{N}_n(t) &= \operatorname{tr}(\Sigma_n(\Sigma_n + t^{-1})^{-1}) \\ &= \operatorname{tr}\left[\Sigma_n(\Sigma + t^{-1})^{-1/2}(I + A_t)^{-1}(\Sigma + t^{-1})^{-1/2} \right] \\ &= \operatorname{tr}\left[(\Sigma + t^{-1})^{-1/2}\Sigma_n(\Sigma + t^{-1})^{-1/2}(I + A_t)^{-1} \right]. \end{aligned}$$

Since $\mathcal{E}_T$ holds and $t \leq T$, we have $\|A_t\|_{\mathrm{op}} \leq \|A_T\|_{\mathrm{op}} \leq 1/2$ by using

$$A_t = (\Sigma + t^{-1})^{-1/2}(\Sigma + T^{-1})^{1/2} A_T (\Sigma + T^{-1})^{1/2}(\Sigma + t^{-1})^{-1/2},$$

which implies that $\|(I + A_t)^{-1}\|_{\mathrm{op}} \leq 2$.

Then, the von Neumann trace inequality applied to non-negative symmetric operators on the event $\mathcal{E}_T$ leads to

$$\begin{aligned} \mathcal{N}_n(t) &\leq \left\|(I + A_t)^{-1}\right\|_{\mathrm{op}} \operatorname{tr}\left[(\Sigma + t^{-1})^{-1/2}\Sigma_n(\Sigma + t^{-1})^{-1/2} \right] \\ &\leq 2 \operatorname{tr}\left[(\Sigma + t^{-1})^{-1/2}\Sigma_n(\Sigma + t^{-1})^{-1/2} \right] \\ &\leq 2\left[\mathcal{N}(t) + \operatorname{tr}(A_t) \right]. \end{aligned} \tag{8.2}$$

Using the definition of the empirical covariance operator, we have

$$\operatorname{tr}(A_t) = \frac{1}{n}\sum_{i=1}^{n} \|(\Sigma + t^{-1})^{-1/2} k_{X_i}\|_{\mathcal{H}}^2 - \mathbb{E}\|(\Sigma + t^{-1})^{-1/2} k_X\|_{\mathcal{H}}^2.$$

In addition since $\|(\Sigma + t^{-1})^{-1/2} k_{X_1}\|_{\mathcal{H}}^2 \leq M^2 t$, and $\mathbb{E}\|(\Sigma + t^{-1})^{-1/2} k_{X_1}\|_{\mathcal{H}}^4 \leq M^2 t \mathcal{N}(t)$, Bernstein's inequality yields

$$\mathbb{P}\Big(\operatorname{tr}(A_t) > \sqrt{\frac{2uM^2 t\mathcal{N}(t)}{n}} + \frac{M^2}{3}\frac{ut}{n} \Big) \leq e^{-u}.$$

Inserting

$$\sqrt{\frac{2uM^2 t\mathcal{N}(t)}{n}} \leq \mathcal{N}(t) + M^2\frac{tu}{n}, \tag{8.3}$$

we get for every $u > 0$,

$$\mathbb{P}\Big(\operatorname{tr}(A_t) > \mathcal{N}(t) + \frac{4M^2}{3}\frac{ut}{n} \Big) \leq e^{-u}.$$

Finally setting $u = n/t$ and using $\mathcal{N}(t) \geq \lambda_1/(\lambda_1 + 1)$ for $t \geq 1$, it results

$$\mathbb{P}\Big(\operatorname{tr}(A_t) > \Big(1 + \frac{4M^2}{3}\Big(1 + \frac{1}{\lambda_1}\Big)\Big)\mathcal{N}(t) \Big) \leq e^{-n/t}.$$

Combining this with (8.2), the first claim follows with $C = 4(1 + 2(1 + \lambda_1^{-1})M^2/3)$. The second claim follows from the first one, using also that $\mathcal{N}_n(t) \leq n$. $\square$

Finally, we establish the following deviation bound for remainder traces.

Lemma 33 *Suppose that* (**BdK**) *holds. Then, for each $u > 0$ and any $0 \leq k \leq n$, we have*

$$\mathbb{P}\Big(\sum_{j>k} \hat{\lambda}_j > 2\sum_{j>k} \lambda_j + 2M^2\frac{u}{n} \Big) \leq e^{-u}.$$

In particular, defining

$$\mathcal{A}(t, K) = \Big\{ \forall 0 \leq k \leq K : \sum_{j>k} \hat{\lambda}_j \leq 2\sum_{j>k} \lambda_j + 2M^2\Big(\frac{1}{t} + \frac{\log(K+1)}{n}\Big) \Big\}$$

with $0 \leq K \leq n$ and $t > 0$, we have

$$\mathbb{P}(\mathcal{A}(t, K)) \geq 1 - e^{-n/t}.$$

Proof of Lemma 33 Let Π_k be the orthogonal projection from $\mathcal{H}$ onto the span of the (population) eigenvectors $(u_j : j > k)$. Then, by the variational characterization of partial traces, we have $\sum_{j>k} \lambda_j = \operatorname{tr}(\Pi_k \Sigma)$ and $\sum_{j>k} \hat{\lambda}_j \leq \operatorname{tr}(\Pi_k \hat{\Sigma})$. We conclude that

$$\sum_{j>k} \hat{\lambda}_j - \sum_{j>k} \lambda_j \leq \operatorname{tr}(\Pi_k(\hat{\Sigma} - \Sigma)\Pi_k) = \frac{1}{n}\sum_{i=1}^{n} \|\Pi_k k_{X_i}\|_{\mathcal{H}}^2 - \mathbb{E}\|\Pi_k k_X\|_{\mathcal{H}}^2.$$

Since $\|\Pi_k k_{X_i}\|_{\mathcal{H}}^2 \leq \|k_{X_i}\|_{\mathcal{H}}^2 \leq M^2$, and $\mathbb{E}\|\Pi_k k_{X_i}\|_{\mathcal{H}}^4 \leq M^2 \mathbb{E}\|\Pi_k k_{X_i}\|_{\mathcal{H}}^2 = M^2 \sum_{j>k} \lambda_j$, Bernstein's inequality yields

$$\mathbb{P}\Big(\sum_{j>k} \hat{\lambda}_j > \sum_{j>k} \lambda_j + \sqrt{\frac{2uM^2(\sum_{j>k} \lambda_j)}{n}} + \frac{M^2}{n} u\Big) \leq e^{-u}.$$

Inserting

$$\sqrt{\frac{2uM^2(\sum_{j>k} \lambda_j)}{n}} \leq \sum_{j>k} \lambda_j + \frac{M^2}{n} u,$$

the first claim follows. The second claim follows from the first one with $u = n/t + \log(K+1)$ in combination with the union bound. □

8.2 Bounds for the variance and bias parts

We also use the notation of Section 7 with $L_n = \tilde{g}_T^{1/2}(K_n) K_n^{1/2}$. In particular, we abbreviate $\tilde{\mathbf{f}} = \tilde{g}_T^{1/2}(K_n) K_n^{1/2} \mathbf{f}$ and $\tilde{\boldsymbol{\epsilon}} = \tilde{g}_T^{1/2}(K_n) K_n^{1/2} \boldsymbol{\epsilon}$. Moreover, we write $\widetilde{\mathcal{N}}_n^g(t) = \operatorname{tr}(\tilde{g}_T(K_n) K_n g_t(K_n) K_n)$ for the smoothed g-effective dimension and $\widetilde{t}_n^* = \inf\{t \geq 0 : \|r_t(K_n)\tilde{\mathbf{f}}\|_n^2 \leq \sigma^2 n^{-1} \widetilde{\mathcal{N}}_n^g(t)\}$ for the balancing stopping rule defined from the smoothed bias and proxy variance terms.

8.2.1 A bound for the variance part

Proposition 34 *Under the assumptions of Theorem 19, we have on the event* $\mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2 T \rfloor)$,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho g_{\tau_{SDP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 > y(u)) \leq 3e^{-u}, \quad u > 0,$$

with

$$y(u) = C\frac{\sigma^2}{n}(\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + \sqrt{u \mathcal{N}_n(T)} + u + 1).$$

The proof of Proposition 34 will be based on a series of lemmas successively detailed in what follows.

The following lemma provides a slightly weaker version of Assumption (**EVBound**) that is implied by the population variant (**EffRank**).

Lemma 35 *Suppose that* (**EffRank**) *and* (**BdK**) *hold. Let* $T > 0$ *be such that* $T \log(\lfloor M^2 T \rfloor + 1) \leq n$*. Then, on the event* $\mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2 T \rfloor)$*, we have*

$$\forall 0 < t \leq T, \qquad t \sum_{j : t\hat{\lambda}_j < 1} \hat{\lambda}_j \leq E(|\{j : t\hat{\lambda}_j \geq 1\}| \vee 1)$$

with $E = 6E' + 4M^2$*.*

Proof of Lemma 35 Firstly by (**BdK**) we have $k\hat{\lambda}_k \leq \sum_{j \leq k} \hat{\lambda}_j \leq \operatorname{tr}(\hat{\Sigma}) \leq M^2$ and thus $\hat{\lambda}_k \leq M^2 k^{-1}$ for every $k \geq 1$.

For $0 < t \leq T$ define now $k \geq 0$ such that $t\hat{\lambda}_k \geq 1 > t\hat{\lambda}_{k+1}$ (with the convention that $k = 0$ if $t\hat{\lambda}_1 < 1$). Then it follows from the above that $k \leq \lfloor M^2 T \rfloor$. Let us now consider the event $\mathcal{A}(T, \lfloor M^2 T \rfloor) \cap \mathcal{E}_T$. We have

$$\begin{aligned} t \sum_{j > k} \hat{\lambda}_j &\leq 2t \sum_{j > k} \lambda_j + 2M^2 + \frac{2M^2 T \log(\lfloor M^2 T \rfloor + 1)}{n} \\ &\leq 2tE' \lambda_{k+1}(k \vee 1) + 4M^2, \end{aligned} \tag{8.4}$$

where we applied (**EffRank**) and $T \log(\lfloor M^2 T \rfloor + 1) \leq n$ in the second inequality. Using the lower bound in Lemma 44, we have $\lambda_{k+1} \leq 2\hat{\lambda}_{k+1} + 1/T$. Inserting this into (8.4), we get

$$t \sum_{j > k} \hat{\lambda}_j \leq 4E'(k \vee 1) + 2E'(k \vee 1) + 4M^2 \leq (6E' + 4M^2)(k \vee 1),$$

and the claim follows with $E = 6E' + 4M^2$. □

Lemma 36 *Suppose that* (**EffRank**)*,* (**BdK**) *and* (**LFL**) *hold. Let* $T > 0$ *be such that* $T \log(\lfloor M^2 T \rfloor + 1) \leq n$*. Then, on the event* $\mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2 T \rfloor)$*, we have*

$$\forall 1 \leq t \leq T, \qquad \mathcal{N}_n^g(t) \leq C_1 \widetilde{\mathcal{N}}_n^g(t) + C_2$$

with $C_1 = 2(1 + b^{-1}EB)$*,* $C_2 = BE$ *and* $E = 6E' + 4M^2$*.*

Proof of Lemma 36 If $t\hat{\lambda}_1 < 1$, then (**LFU**) and Lemma 35 imply

$$\mathcal{N}_n^g(t) \leq Bt \sum_{j \geq 1} \hat{\lambda}_j \leq BE,$$

yielding the claim in this case. On the other hand, if $t\hat{\lambda}_1 \geq 1$, then let $k \geq 1$ be defined by $t\hat{\lambda}_{k+1} < 1 \leq t\hat{\lambda}_k$. By (3.2), we have

$$\mathcal{N}_n^g(t) \leq \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j) + Bt \sum_{j > k} \hat{\lambda}_j. \tag{8.5}$$

Now by the definition of k, Lemma 35 and (**LFL**), we have

$$t \sum_{j > k} \hat{\lambda}_j \leq Ek \leq Eb^{-1} \sum_{j \leq k} \hat{\lambda}_j g_t(\hat{\lambda}_j).$$

Inserting this into (8.5), we get

$$\sum_{j=1}^{n}\hat{\lambda}_j g_t(\hat{\lambda}_j) \leq C\sum_{j\leq k}\hat{\lambda}_j g_t(\hat{\lambda}_j) \leq \tilde{b}^{-1}C\sum_{j=1}^{n} g_t(\hat{\lambda}_j)\hat{\lambda}_j\tilde{g}_T(\hat{\lambda}_j)\hat{\lambda}_j$$

with $C = (1 + b^{-1}BE)$ and $\tilde{b} = 1/2$. □

Lemma 37 *Suppose that* (**EffRank**) *and* (**BdK**) *hold. Let* $T > 0$ *be such that* $T\log(\lfloor M^2T\rfloor + 1) \leq n$. *Then, on the event* $\mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2T\rfloor)$, *we have*

$$\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\Lambda(y)\Big) \leq 3\exp\Big(-c\Big(y \wedge \frac{y^2}{\mathcal{N}_n(T)}\Big)\Big), \quad y > 0,$$

with

$$\Lambda(y) = C\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + (C+1)y + BE,$$

where $C = 2(1 + b^{-1}BE)$, $E = 6E' + 4M^2$ *and* $c > 0$ *is a constant depending only on* A.

Proof of Lemma 37 If (7.4) holds, that is if $\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + y > \widetilde{\mathcal{N}}_n^g(T)$, then Lemma 36 implies that on $\mathcal{E}_T \cap \mathcal{A}(n/T, K)$,

$$\Lambda(y) \geq C\widetilde{\mathcal{N}}_n^g(T) + y + BE \geq \mathcal{N}_n^g(T) + y.$$

Hence,

$$\begin{aligned}
&\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\Lambda(y)\Big)\\
&\leq \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_T^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(T) + \frac{\sigma^2}{n}y\Big)
\end{aligned}$$

and the claim follows from Lemma 28 and Lemma 6. On the other hand, if (7.4) does not hold, then we can define $\widetilde{t}_n^* < t \leq T$ by $\widetilde{\mathcal{N}}_n^g(t) = \widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) + y$. On $\mathcal{E}_T \cap \mathcal{A}(n/T, K)$, Lemma 36 implies

$$\Lambda(y) = C\widetilde{\mathcal{N}}_n^g(t) + y + BE \geq \mathcal{N}_n^g(t) + y.$$

Hence,

$$\begin{aligned}
&\mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\Lambda(y)\Big)\\
&\leq \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_t^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\Lambda(y)\Big) + \mathbf{P}_\epsilon(\tau_{SDP} > t)\\
&\leq \mathbf{P}_\epsilon\Big(\|K_n^{1/2}g_t^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > \frac{\sigma^2}{n}\mathcal{N}_n^g(t) + \frac{\sigma^2}{n}y\Big) + \mathbf{P}_\epsilon(\tau_{SDP} > t),
\end{aligned}$$

and the claim follows from applying Lemma 28 and Lemma 6 to the second last term and Proposition 23 and Lemma 6 to the last term, using that $t > \widetilde{t}_n^*$ and $y = \widetilde{\mathcal{N}}_n^g(t) - \widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*)$. □

Proof of Proposition 34 First, by Lemma 16, we have on the event $\mathcal{E}_T$,

$$\|S_\rho g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_\rho^2 \le 2\|S_n g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_n^2 + T^{-1}\|g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_{\mathcal{H}}^2$$

Applying (**LFU**) and the fact that $\tau_{SDP} \le T$, and then (**BdF**), we get

$$\begin{aligned}\|S_\rho g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_\rho^2 &\le 2\|S_n g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_n^2 + T^{-1}\|g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_{\mathcal{H}}^2\\ &\le 2\|S_n g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_n^2 + B\|g_{\tau_{SDP}}^{1/2}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_{\mathcal{H}}^2\\ &= 2\|K_n g_{\tau_{SDP}}(K_n)\boldsymbol{\epsilon}\|_n^2 + B\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2\\ &\le (2+B)\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2. \end{aligned} \tag{8.6}$$

Hence, on the event $\mathcal{E}_T$,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_\rho^2 > y(u)) \le \mathbf{P}_{\boldsymbol{\epsilon}}((2+B)\|K_n^{1/2}g_{\tau_{SDP}}^{1/2}(K_n)\boldsymbol{\epsilon}\|_n^2 > y(u)),$$

and the claim follows from Lemma 37 applied with $y = C(\sqrt{\mathcal{N}_n(T)u} + u)$ and the fact that the assumption $T \le cn/(\log n)$ with c small enough implies that $T\log(\lfloor M^2T\rfloor + 1) \le n$. □

8.2.2 A bound for the bias part

Proposition 38 *Under the assumptions of Theorem 19, we have on the event* $\mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2T\rfloor)$,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 > z(u)) \le 2e^{-u}, \quad , u > 0,$$

with

$$z(u) = C\Big(\|r_{\tilde{t}_n^*\wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sqrt{u\mathcal{N}_n(T)}+u}{n} + \frac{1}{T^{1+2s}} + \frac{\|\Sigma-\Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}}{T}\Big),$$

If $s \le 1/2$*, then the last term in the definition of* $z(u)$ *can be dropped.*

Proof of Proposition 38 First, note that under $s = r - 1/2 \ge 0$ the regression function f can be represented as a function in $\mathcal{H}$. By Lemma 16, we have on the event $\mathcal{E}_T$,

$$\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 \le 2\|S_n r_{\tau_{SDP}}(\Sigma_n)f\|_n^2 + T^{-1}\|r_{\tau_{SDP}}(\Sigma_n)f\|_{\mathcal{H}}^2.$$

Using this and (**BdF**), we get

$$\begin{aligned}\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 &\le \sum_{j\ge 1}(2\hat\lambda_j + T^{-1})r_{\tau_{SDP}}^2(\hat\lambda_j)\langle f, \hat u_j\rangle^2\\ &\le 3b^{-1}\sum_{\hat\lambda_j T>1}\hat\lambda_j r_{\tau_{SDP}}^2(\hat\lambda_j)\tilde g_T(\hat\lambda_j)\hat\lambda_j\langle f, \hat u_j\rangle^2 + 3T^{-1}\sum_{\hat\lambda_j T\le 1}\langle f, \hat u_j\rangle^2.\end{aligned}$$

Using (7.19), we get on $\mathcal{E}_T$,

$$\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 \le 3b^{-1}\|r_{\tau_{SDP}}(K_n)\tilde{\mathbf{f}}\|_n^2 + z(u)/2,$$

provided that the constant C in the definition of $z(u)$ is six times as big as the constant in (7.19). Hence, on the event $\mathcal{E}_T$,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 > z(u)) \le \mathbf{P}_{\boldsymbol{\epsilon}}(6b^{-1}\|r_{\tau_{SDP}}(K_n)\tilde{\mathbf{f}}\|_n^2 > z(u)),$$

and the claim follows from (7.14), provided that C in the definition of $z(u)$ is chosen large enough. □

8.3 Proofs of oracle inequalities (inner case)

8.3.1 Proof of Theorem 19

Since $s = r - 1/2 \geq 0$, f can be represented as a function in $\mathcal{H}$. In particular, we can write $\mathbf{Y} = S_n f + \boldsymbol{\epsilon}$, leading to

$$\begin{aligned} f - \hat{f}^{(\tau_{SDP})} &= f - g_{\tau_{SDP}}(\Sigma_n)\Sigma_n f - g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon} \\ &= r_{\tau_{SDP}}(\Sigma_n)f - g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}. \end{aligned}$$

Hence,

$$\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \leq 2\|S_\rho r_{\tau_{SDP}}(\Sigma_n)f\|_\rho^2 + 2\|S_\rho g_{\tau_{SDP}}(\Sigma_n)S_n^*\boldsymbol{\epsilon}\|_\rho^2.$$

The last but one term is addressed by Lemma 38 and the last one by Proposition 34. Combining these estimates with (7.16), introducing the event $\Omega_T = \mathcal{E}_T \cap \mathcal{A}(T, \lfloor M^2 T \rfloor)$, we get on the event Ω_T,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 > x(u)) \leq 5e^{-u}, \quad u > 0,$$

with

$$\begin{aligned} x(u) = C\Big(&\min_{0 \leq t \leq T} \Big\{ \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n^g(t)}{n} \Big\} \\ &+ \frac{\sqrt{u\mathcal{N}_n(T)} + u + 1}{n} + \frac{1}{T^{1+2s}} + \frac{\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}}{T} \Big), \end{aligned}$$

where the last term in the definition of $x(u)$ can be dropped if $s \leq 1/2$. Invoking the last claim in Lemma 29 and Lemma 6, we get on the event Ω_T,

$$\mathbf{P}_{\boldsymbol{\epsilon}}(\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 > \tilde{x}(u)) \leq 5e^{-u}, \quad u > 0.$$

with

$$\tilde{x}(u) = C\Big(\min_{0 < t \leq T} \Big\{ \frac{1}{t^{1+2s}} + \frac{\mathcal{N}_n(t)}{n} + \frac{\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}}{t} \Big\} + \frac{\sqrt{u\mathcal{N}_n(T)} + u}{n} \Big),$$

where the last term in the curly brackets in the definition of $\tilde{x}(u)$ can be dropped if $s \leq 1/2$. Integrating this inequality on the event Ω_T, we get

$$\begin{aligned} &\mathbb{E}\mathbb{1}_{\Omega_T}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 = \mathbb{E}\mathbb{1}_{\Omega_T}\mathbf{E}_{\boldsymbol{\epsilon}}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \\ &\leq C\Big(\min_{1 \leq t \leq T} \Big\{ \frac{1}{t^{1+2s}} + \frac{\mathbb{E}\mathbb{1}_{\Omega_T}\mathcal{N}_n(t)}{n} + \frac{\mathbb{E}\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}}{t} \Big\} + \frac{\mathbb{E}\mathbb{1}_{\Omega_T}\sqrt{\mathcal{N}_n(T)} + 1}{n} \Big), \end{aligned}$$

where the last term in the curly brackets can be dropped if $s \leq 1/2$. Here, we have replaced the minimum over $0 < t \leq T$ by $1 \leq t \leq T$ since the range $t \in (0, 1]$ does not yield any improvement. Focusing now on $\mathbb{E}\mathbb{1}_{\Omega_T}\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s}$, this latter term can be tackled by first

$$\mathbb{E}\|\Sigma - \Sigma_n\|_{\mathrm{op}}^{2\wedge 2s} \leq (\mathbb{E}\|\Sigma - \Sigma_n\|_{\mathrm{op}}^2)^{1\wedge s} \leq (\mathbb{E}\|\Sigma - \Sigma_n\|_{\mathrm{HS}}^2)^{1\wedge s}.$$

Then, since the random variables $k_{X_i} \otimes k_{X_i} - \Sigma$ are centered and independent, we have

$$\mathbb{E}\|\Sigma - \Sigma_n\|_{\mathrm{HS}}^2 \leq \frac{1}{n}\mathbb{E}\|k_X \otimes k_X\|_{\mathrm{HS}}^2 = \frac{1}{n}\mathbb{E}\|k_X\|_{\mathcal{H}}^4 \leq \frac{M^4}{n}. \tag{8.7}$$

Using the Cauchy-Schwarz inequality, the second claim in Lemma 31 and the previous bound, we get

$$\begin{aligned}&\mathbb{E}\mathbb{1}_{\Omega_T}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \\ &\leq C\Big(\min_{1\leq t\leq T}\Big\{\frac{1}{t^{1+2s}} + \frac{1}{tn^{1\wedge s}} + \frac{\mathcal{N}(t) + ne^{-n/t}}{n}\Big\} + \frac{\sqrt{\mathcal{N}(T) + ne^{-n/T}}}{n}\Big),\end{aligned}$$

where the second term $t^{-1}n^{-(1\wedge s)}$ is only present for $s > 1/2$.

We now show that this term can also be dropped for $s > 1/2$. If $s \geq 1$, this is clear using $t^{-1}n^{-1\wedge s} \leq n^{-1}$. Assume now that $s \in (1/2, 1)$. If $t \leq \sqrt{n}$, then $t^{-1}n^{-s} \leq t^{-1-2s}$, while if $t > \sqrt{n}$ then $t^{-1}n^{-s} \leq n^{-1/2-s} \leq n^{-1}$. Moreover, the terms $ne^{-n/t}$ and $ne^{-n/T}$ can also be dropped using the condition $1 \leq t \leq T \leq c_1 n/(\log n)$ with c_1 small enough. We thus get

$$\mathbb{E}\mathbb{1}_{\Omega_T}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \leq C\Big(\min_{1\leq t\leq T}\Big\{\frac{1}{t^{1+2s}} + \frac{\mathcal{N}(t)}{n}\Big\} + \frac{\sqrt{\mathcal{N}(T)}}{n}\Big).$$

The last part of the proof consists in analyzing the prediction error on the complement of the event Ω_T. Since $\|\hat{f}^{(t)}\|_{\mathcal{H}}^2$ is non-decreasing in $t \geq 1$, we have $\|\hat{f}^{(\tau_{SDP})}\|_{\mathcal{H}}^2 \leq \|\hat{f}^{(T)}\|_{\mathcal{H}}^2$. Moreover, applying (**BdF**) and (**LFU**), we get $\|\hat{f}^{(T)}\|_{\mathcal{H}}^2 \leq BT\|\mathbf{Y}\|_n^2$. Hence, $\|S_\rho\hat{f}^{(\tau_{SDP})}\|_\rho^2 \leq \lambda_1 BT\|\mathbf{Y}\|_n^2$ and

$$\begin{aligned}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 &\leq 2\|S_\rho f\|_\rho^2 + 2\lambda_1 BT\|\mathbf{Y}\|_n^2 \\ &\leq 2\|S_\rho f\|_\rho^2 + 4M^2\lambda_1 BT\|f\|_{\mathcal{H}}^2 + 4\lambda_1 BT\|\boldsymbol{\epsilon}\|_n^2 \leq C(1 + T\|\boldsymbol{\epsilon}\|_n^2),\end{aligned} \tag{8.8}$$

where we applied $\|S_n f\|_n^2 = (1/n)\sum_{i=1}^n \langle f, k_{X_i}\rangle_{\mathcal{H}}^2 \leq M^2\|f\|_{\mathcal{H}}^2$ in the second inequality. Using $T \leq c_1 n/(\log n)$ with c_1 small enough, we get $\mathbb{P}(\Omega_T^c) \leq \mathbb{P}(A(n/T, 3M^2T)^c) + \mathbb{P}(\mathcal{E}_T^c) \leq 2C_1 Te^{-c_2 n/T} \leq 2C_2 n^{-C_3}$ with $C_3 > 4$. Using the Cauchy-Schwarz inequality and (**SubGN**) it follows that

$$\mathbb{E}\mathbb{1}_{\Omega_T^c}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \leq Cn^{-1} \tag{8.9}$$

and the claim follows.

8.3.2 Proof of Theorem 20

We prove the result in the case $s \leq 1/2$, the other case follows similarly. From previous Section 8.3.1, let us consider the event $\Omega_{T_1} = \mathcal{E}_{T_1} \cap \mathcal{A}(n/T_1, 3M^2T_1)$, where $T_1 = c_1 n/\log n$ and c_1 is sufficiently small such that $\mathbb{P}(\Omega_{T_1}) \leq n^{-4}$ (such a choice is possible by Lemma 30 and Lemma 33).

We first show that with $T = \min(T_1, \hat{T})$, we have on the event Ω_{T_1}

$$\|r_{\widetilde{t}_n^* \wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t}_n^*) \leq C\Big(\min_{t>0}\Big\{t^{-2r} + \frac{\mathcal{N}_n(t)}{n}\Big\} + \frac{\log n}{n}\Big) \tag{8.10}$$

By the definition of $\widetilde{t_n^*}$ (Eq. (4.2)), Eq. (7.16) and Lemma 29, we have on the event Ω_{T_1}

$$\|r_{\widetilde{t_n^*}\wedge T}(K_n)\tilde{\mathbf{f}}\|_n^2 + \frac{\sigma^2}{n}\widetilde{\mathcal{N}}_n^g(\widetilde{t_n^*}) \leq C \min_{0<t\leq T}\Big\{t^{-2r} + \frac{\mathcal{N}_n(t)}{n}\Big\}. \tag{8.11}$$

On the one hand, if $\hat{T} > T_1$, then $T = T_1$ and $T_1\mathcal{N}_n(T_1) < n$ and thus (since $2r \geq 1$)

$$\min_{0<t\leq T}\Big\{t^{-2r} + \frac{\mathcal{N}_n(t)}{n}\Big\} \leq \frac{1}{T_1} + \frac{\mathcal{N}_n(T_1)}{n} < \frac{2}{T_1} \leq \frac{2}{c_1}\frac{\log n}{n}\cdot$$

On the other hand, if $\hat{T} \leq T_1$, then $T = \hat{T}$ and t_n defined by $t_n^{2r}\mathcal{N}_n(t_n) = n$ satisfies either $1 \leq t_n \leq \hat{T}$ or $0 < t_n < 1$. In the former case the right-hand side of (8.11) is bounded by $2C\min_{t>0}\{t^{-2r}+n^{-1}\mathcal{N}_n(t)\}$, where the constraint that $t \leq T$ has been removed, while in the latter case the bound (8.10) is trivial since $2\min_{t>0}\{t^{-2r}+n^{-1}\mathcal{N}_n(t)\} \geq t_n^{-2r}+n^{-1}\mathcal{N}_n(t_n) \geq 1$ in this case. This completes the proof of (8.10).

Similarly, by the definition of T, we have

$$\frac{\sqrt{\mathcal{N}_n(T)}}{n} = \frac{1}{\sqrt{n}}\sqrt{\frac{\mathcal{N}_n(T)}{n}} \leq C\Big(\sqrt{\frac{1}{n}\min_{t>0}\Big\{t^{-1} + \frac{\mathcal{N}_n(t)}{n}\Big\}} + \frac{\log n}{n}\Big)$$

We can now proceed as in Proposition 34 and Proposition 38 to obtain on the event Ω_{T_1}

$$\begin{aligned}&\mathbf{E}_\epsilon\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2\\ &\leq C\Big(\min_{t>0}\Big\{t^{-2r} + \frac{1}{n}\mathcal{N}_n(t)\Big\} + \sqrt{\frac{1}{n}\min_{t>0}\Big\{t^{-1} + \frac{\mathcal{N}_n(t)}{n}\Big\}} + \frac{\log n}{n}\Big).\end{aligned}$$

Here we used that T does only depend on the design and is thus fixed conditional on the design. Hence, taking expectation and using Lemma 31 and Remark 32, we conclude

$$\begin{aligned}&\mathbb{E}\mathbb{1}_{\Omega_{T_1}}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2\\ &\leq C\Big(\min_{t>0}\Big\{t^{-2r} + \frac{\mathcal{N}(t)}{n}\Big\} + \sqrt{\frac{1}{n}\min_{t>0}\Big\{t^{-1} + \frac{\mathcal{N}(t)}{n}\Big\}} + \frac{\log n}{n}\Big).\end{aligned}$$

The claim follows from the final arguments in the proof of Theorem 19, showing that

$$\mathbb{E}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2 \leq \mathbb{E}\Big[\mathbb{1}_{\Omega_{T_1}}\|S_\rho(f - \hat{f}^{(\tau_{SDP})})\|_\rho^2\Big] + Cn^{-1}.$$

8.4 Proofs of oracle inequalities (outer case)

8.4.1 Proof of Theorem 17

For simplicity, we prove Theorem 17 only in the case of Tikhonov regularization. Throughout the proof, we set $T = cn/(\log n)$ with c sufficiently small such that

$$\mathbb{P}(\mathcal{E}_T^c) \leq n^{-C}, \qquad C > 4. \tag{8.12}$$

Such a choice is possible by Lemma 30.

Lemma 39 *Suppose that* ($\mathbf{SC}$(r,R)) *holds with* $0 < r \leq 1/2$. *For* $t \geq 1$, *let* $f^{(t)} = (\Sigma + t^{-1})^{-1} S_\rho^* f \in \mathcal{H}$. *Then we have*

(i) $\|f - S_\rho f^{(t)}\|_\rho^2 \leq t^{-2r} R^2$,

(ii) $\|f^{(t)}\|_{\mathcal{H}}^2 \leq t^{-2r+1} R^2$.

Part (i) follows from Theorem 4 in Smale and Zhou (2005) applied with $\lambda = t^{-1}$. Part (ii) can be proved analogously; see e.g. Proposition 3 in Caponnetto (2006).

Lemma 40 *Under the assumptions of Theorem 17, we have on* $\mathcal{E}_T$,

$$\mathbf{E}_\epsilon \|S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 \leq C\Big(\min_{0<t\leq c\frac{n}{\log n}} \Big\{ \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n(t)}{n} \Big\} + \frac{1}{\sqrt{n}} \Big).$$

Proof of Lemma 40 By (8.6) with τ_{SDP} replaced by τ_{DP}, we have on the event $\mathcal{E}_T$,

$$\|S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 \leq (2+B) \|K_n^{1/2} g_{\tau_{DP}}^{1/2}(K_n) \boldsymbol{\epsilon}\|_n^2.$$

Applying (7.9), we get on the event $\mathcal{E}_T$ and for every $u > 0$,

$$\begin{aligned} &\mathbf{P}_\epsilon\Big(\|S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 > C\Big(\frac{\mathcal{N}_n(t_n^*)}{n} + \frac{\sqrt{u}}{\sqrt{n}} + \frac{u}{n} \Big) \Big) \\ &\leq \mathbf{P}_\epsilon\Big((2+B) \|K_n^{1/2} g_{\tau_{DP}}^{1/2}(K_n) \boldsymbol{\epsilon}\|_n^2 > C\Big(\frac{\mathcal{N}_n(t_n^*)}{n} + \frac{\sqrt{u}}{\sqrt{n}} + \frac{u}{n} \Big) \Big) \leq 3e^{-u} \end{aligned}$$

with C sufficiently large. Integrating this inequality and inserting (7.11), the claim follows.
□

Lemma 41 *Under the assumptions of Theorem 17, we have*

$$\mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \leq C\Big(\mathbb{E}\mathbb{1}_{\mathcal{E}_T} \min_{0<t\leq c\frac{n}{\log n}} \Big\{ \|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n(t)}{n} \Big\} + \frac{1}{\sqrt{n}} + \Big(\frac{\log n}{n} \Big)^{2r} \Big).$$

Proof of Lemma 41 We have

$$\begin{aligned} \mathbb{E}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 &\leq \mathbb{E}\mathbb{1}_{\mathcal{E}_T} \|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 + 2\mathbb{E}\mathbb{1}_{\mathcal{E}_T^c} \|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \\ &\leq \mathbb{E}\mathbb{1}_{\mathcal{E}_T} \|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 + Cn^{-1}, \end{aligned}$$

where the second inequality follows by the same line of arguments as at the end of the proof of Theorem 19 (cf. (8.8) and (8.9)), using that f is bounded this time which implies $\|g_{\tau_{DP}}^{1/2}(K_n) K_n^{1/2} \mathbf{f}\|_n \leq \|f\|_\infty$.

Let us now introduce, for $t_1 > 0$ to be chosen later,

$$f - S_\rho \hat{f}^{(\tau_{DP})} = f - S_\rho f^{(t_1)} + S_\rho f^{(t_1)} - S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \mathbf{f} - S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon},$$

where $f^{(t_1)} = (\Sigma + t_1^{-1})^{-1} S_\rho^* f$. It results that

$$\begin{aligned}
&\frac{1}{3}\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|f - S_\rho \hat{f}^{(\tau_{DP})}\|_\rho^2 \\
&\le \|f - S_\rho f^{(t_1)}\|_\rho^2 + \mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 + \mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|S_\rho f^{(t_1)} - S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \mathbf{f}\|_\rho^2 \\
&=: I_1 + I_2 + I_3.
\end{aligned}$$

Form Lemma 39(i), we get $I_1 \le R^2 t_1^{-2r}$, and Lemma 40 provides

$$\begin{aligned}
I_2 &= \mathbb{E}\mathbb{1}_{\mathcal{E}_T}\mathbf{E}_\epsilon\|S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \boldsymbol{\epsilon}\|_\rho^2 \\
&\le C\Big(\mathbb{E}\mathbb{1}_{\mathcal{E}_T} \min_{0<t\le c\frac{n}{\log n}} \Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n(t)}{n}\Big\} + \frac{1}{\sqrt{n}}\Big).
\end{aligned}$$

The remainder of this proof consists in considering the term I_3.

By the change of norm argument of Lemma 16 applied to functions belonging to $\mathcal{H}$, on the event $\mathcal{E}_T$, we have

$$\begin{aligned}
&\|S_\rho f^{(t_1)} - S_\rho g_{\tau_{DP}}(\Sigma_n) S_n^* \mathbf{f}\|_\rho^2 \\
&\le \|\mathbf{f}^{(t_1)} - g_{\tau_{DP}}(K_n) K_n \mathbf{f}\|_n^2 + T^{-1}\|f^{(t_1)} - g_{\tau_{DP}}(\Sigma_n) S_n^* \mathbf{f}\|_{\mathcal{H}}^2. \qquad (8.13)
\end{aligned}$$

Empirical norm in (8.13)**:** Integrating yields

$$\begin{aligned}
\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|\mathbf{f}^{(t_1)} - g_{\tau_{DP}}(K_n) K_n \mathbf{f}\|_n^2 &\le 2\mathbb{E}\|\mathbf{f} - \mathbf{f}^{(t_1)}\|_n^2 + 2\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2 \\
&= 2\|f - S_\rho f^{(t_1)}\|_\rho^2 + 2\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2.
\end{aligned}$$

The first term in the r.h.s. is addressed by Lemma 39(i), leading to the upper bound $2R^2 t_1^{-2r}$. For the second one, integrating (7.10) with $T = cn/\log n$ and inserting (7.11), we get

$$\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|r_{\tau_{DP}}(K_n)\mathbf{f}\|_n^2 \le C\Big(\mathbb{E}\mathbb{1}_{\mathcal{E}_T} \min_{0<t\le c\frac{n}{\log n}} \Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n(t)}{n}\Big\} + \frac{1}{\sqrt{n}}\Big).$$

Hilbert norm in (8.13)**:** We have

$$\begin{aligned}
\|f^{(t_1)} - g_{\tau_{DP}}(\Sigma_n) S_n^* \mathbf{f}\|_{\mathcal{H}}^2 &= \|f^{(t_1)} - g_{\tau_{DP}}(\Sigma_n)\Sigma_n f^{(t_1)} + g_{\tau_{DP}}(\Sigma_n) S_n^*(\mathbf{f}^{(t_1)} - \mathbf{f})\|_{\mathcal{H}}^2 \\
&\le 2\|r_{\tau_{DP}}(\Sigma_n) f^{(t_1)}\|_{\mathcal{H}}^2 + 2\|g_{\tau_{DP}}(\Sigma_n) S_n^*(\mathbf{f}^{(t_1)} - \mathbf{f})\|_{\mathcal{H}}^2 \\
&\le 2R^2 t_1^{1-2r} + 2BT\|\mathbf{f}^{(t_1)} - \mathbf{f}\|_n^2, \qquad (8.14)
\end{aligned}$$

where we applied (**BdF**) and Lemma 39(ii) to the first term and (**BdF**), (**LFU**) and the inequality $\tau_{DP} \le T$ to the second term.

Collecting these bounds and using $T = cn/(\log n)$ and Lemma 39(i), we get

$$I_3 \le C\Big(\mathbb{E}\mathbb{1}_{\mathcal{E}_T} \min_{0<t\le c\frac{n}{\log n}} \Big\{\|r_t(K_n)\mathbf{f}\|_n^2 + \frac{\mathcal{N}_n(t)}{n}\Big\} + \frac{1}{\sqrt{n}} + t_1^{-2r} + \frac{\log n}{n} t_1^{1-2r}\Big).$$

The claim now follows from these bounds for $I_1 - I_3$ by setting $t_1 = cn/(\log n)$. $\square$

Lemma 42 *For $t > 0$ let $g_t(\lambda) = (\lambda + t^{-1})^{-1}$, and let $T = cn/(\log n)$. Suppose that* (**BdK**) *holds. Then we have*

$$\forall 0 < t \le T, \qquad \mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|r_t(K_n)\mathbf{f}\|_n^2 \le C\Big(t^{-2r} + \frac{\mathcal{N}(t)}{n}\Big).$$

Moreover, we have

$$\forall 0 < t \le T, \qquad \mathbb{E}\mathbb{1}_{\mathcal{E}_T}\mathcal{N}_n(t) \le C_2(\mathcal{N}(t) + 1).$$

Proof of Lemma 42 The second claim directly follows from Lemma 31 in combination with Remark 32.

For the first claim, set $f^{(t)} = (\Sigma + t^{-1})^{-1} S_\rho^* f$. By Lemma 39, we have

$$\begin{aligned}
&\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|\mathbf{f} - S_n(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_n^2 \\
&\le 2\|f - S_\rho f^{(t)}\|_\rho^2 + 2\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|S_n f^{(t)} - S_n(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_n^2 \\
&\le 2R^2 t^{-2r} + 2\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|S_n f^{(t)} - S_n(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_n^2.
\end{aligned}$$

It remains to analyze the last term. Using Lemma 16 (change of norm), we have on $\mathcal{E}_T$,

$$\begin{aligned}
&\|S_n f^{(t)} - S_n(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_n^2 \\
&\le 2\|S_\rho f^{(t)} - S_\rho(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_\rho^2 + C\frac{\log n}{n}\|f^{(t)} - (\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_{\mathcal{H}}^2.
\end{aligned}$$

By (8.14) (where τ_{DP} is replaced by t), the $\mathcal{H}$-norm is bounded by $C(t^{1-2r} + t\|\mathbf{f} - \mathbf{f}^{(t)}\|_n^2)$ and thus on $\mathcal{E}_T$,

$$\begin{aligned}
&\|S_n f^{(t)} - S_n(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_n^2 \\
&\le 2\|S_\rho f^{(t)} - S_\rho(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_\rho^2 + C(t^{-2r} + \|\mathbf{f} - \mathbf{f}^{(t)}\|_n^2),
\end{aligned}$$

where we also used that $t \le T = cn/(\log n)$. Since $\mathbb{E}\|\mathbf{f} - \mathbf{f}^{(t)}\|_n^2 = \|f - S_\rho f^{(t)}\|_\rho^2 \le R^2 t^{-2r}$, as can be seen from Lemma 39(i), it remains to bound the term

$$\begin{aligned}
&2\|S_\rho f^{(t)} - S_\rho(\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f}\|_\rho^2 \\
&\le 2\|(\Sigma + t^{-1})^{1/2}((\Sigma + t^{-1})^{-1}S_\rho^* f - (\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f})\|_{\mathcal{H}}^2,
\end{aligned}$$

where we used $\|S_\rho h\|_\rho^2 = \|\Sigma^{1/2}h\|_{\mathcal{H}}^2 \le \|(\Sigma + t^{-1})^{1/2}h\|_{\mathcal{H}}^2$, $h \in \mathcal{H}$, in the inequality. Inserting

$$\begin{aligned}
&(\Sigma + t^{-1})^{-1}S_\rho^* f - (\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f} \\
&= (\Sigma_n + t^{-1})^{-1}(S_\rho^* f - S_n^*\mathbf{f}) - (\Sigma_n + t^{-1})^{-1}(\Sigma_n - \Sigma)(\Sigma + t^{-1})^{-1}S_\rho^* f
\end{aligned}$$

and

$$(\Sigma_n + t^{-1})^{-1} = (\Sigma + t^{-1})^{-1/2}(I + A_t)^{-1}(\Sigma + t^{-1})^{-1/2}$$

with A_t from (8.1), we get

$$\|(\Sigma + t^{-1})^{1/2}((\Sigma + t^{-1})^{-1}S_\rho^* f - (\Sigma_n + t^{-1})^{-1}S_n^*\mathbf{f})\|_{\mathcal{H}}^2$$

$$\leq 2\|(I+A_t)^{-1}(\Sigma+t^{-1})^{-1/2}(S_\rho^* f - S_n^* \mathbf{f})\|_{\mathcal{H}}^2 \\ + 2\|(I+A_t)^{-1}(\Sigma+t^{-1})^{-1/2}(\Sigma_n - \Sigma)f^{(t)}\|_{\mathcal{H}}^2.$$

In the proof of Lemma 31, we have shown that on the event $\mathcal{E}_T$ we have $\|(I+A_t)^{-1}\|_{\mathrm{op}} \leq 2$. Hence, on $\mathcal{E}_T$,

$$\|(\Sigma+t^{-1})^{1/2}((\Sigma+t^{-1})^{-1}S_\rho^* f - (\Sigma_n+t^{-1})^{-1}S_n^*\mathbf{f})\|_{\mathcal{H}}^2 \\ \leq 4\|(\Sigma+t^{-1})^{-1/2}(S_\rho^* f - S_n^*\mathbf{f})\|_{\mathcal{H}}^2 + 4\|(\Sigma+t^{-1})^{-1/2}(\Sigma_n-\Sigma)f^{(t)}\|_{\mathcal{H}}^2.$$

We conclude that

$$\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\|r_t(K_n)\mathbf{f}\|_n^2 \leq 8\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}(S_\rho^* f - S_n^*\mathbf{f})\|_{\mathcal{H}}^2 \\ + 8\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}(\Sigma_n-\Sigma)f^{(t)}\|_{\mathcal{H}}^2 + Ct^{-2r}.$$

By construction $S_n^*\mathbf{f} - S_\rho^* f$ is a sum of independent, zero-mean random variables. To see the second claim, use that for every $h \in \mathcal{H}$, we have $\mathbb{E}f(X)\langle k_X, h\rangle_{\mathcal{H}} = \langle f, S_\rho h\rangle_\rho = \langle S_\rho^* f, h\rangle_{\mathcal{H}}$, and thus $\mathbb{E}f(X)k_X = S_\rho^* f$. Now, using the fact that f is bounded, we have

$$\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}(S_\rho^* f - S_n^*\mathbf{f})\|_{\mathcal{H}}^2 \leq \frac{1}{n}\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}k_X f(X)\|_{\mathcal{H}}^2 \\ \leq \frac{1}{n}\|f\|_\infty^2 \mathbb{E}\|(\Sigma+t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2 = \|f\|_\infty^2 \frac{\mathcal{N}(t)}{n}.$$

Similarly, we have

$$\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}(\Sigma_n-\Sigma)f^{(t)}\|_{\mathcal{H}}^2 \\ \leq \frac{1}{n}\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}k_X\langle k_X, f^{(t)}\rangle_{\mathcal{H}}\|_{\mathcal{H}}^2 \\ \leq \frac{2}{n}\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2((f(X))^2 + (f^{(t)}(X) - f(X))^2).$$

Using that that f is bounded, the fact that $\|(\Sigma+t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2 \leq M^2 t$ and Lemma 39(i), we get

$$\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}(\Sigma_n-\Sigma)f^{(t)}\|_{\mathcal{H}}^2 \\ \leq \frac{2\|f\|_\infty}{n}\mathbb{E}\|(\Sigma+t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2 + M^2 t\|f - S_\rho f^{(t)}\|_\rho^2 \\ \leq 2\|f\|_\infty \frac{\mathcal{N}(t)}{n} + R^2M^2\frac{t^{-2r+1}}{n} \leq C\Big(\frac{\mathcal{N}(t)}{n} + t^{-2r}\Big),$$

where the last inequality follows from $t \leq c_1 n/(\log n)$. This completes the proof. □

Proof of Theorem 17 (End) The claim follows from inserting Lemma 42 into Lemma 41. □

8.4.2 Sketch of proof of Theorem 18

The proof of Theorem 18 follows from the arguments of the proof of Theorem 17. The improvement is based on the fact that if additionally $\|\Sigma^{\mu/2-1/2}k_X\|_{\mathcal{H}} \leq C_\mu M$ holds for some $\mu \in [0,1)$, then one can improve the concentration and deviation bounds in Lemma 30 and Lemma 31 accordingly. First, Lemma 30 can be improved to $\mathbb{P}(\mathcal{E}_T^c) \leq C_1 T^\mu \exp(-c_1 n/T^\mu)$, since now $\|(\Sigma + t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2$ can be bounded by $C_\mu^2 M^2 t^\mu$. Similarly Lemma 31 can be improved to $\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\mathcal{N}_n(t) \leq C\mathcal{N}(t) + 2ne^{-n/t^\mu}$. In particular, setting $T = c(n/(\log n))^{1/\mu}$ with c sufficiently small, we get $\mathbb{P}(\mathcal{E}_T^c) \leq n^{-4}$ and $\mathbb{E}\mathbb{1}_{\mathcal{E}_T}\mathcal{N}_n(t) \leq C_2(\mathcal{N}(t)+1)$. We can now follow the same line of arguments from above to obtain Theorem 18. Only at the end of proof of Lemma 42, we have to apply $\|(\Sigma + t^{-1})^{-1/2}k_X\|_{\mathcal{H}}^2 \leq C_\mu^2 M^2 t^\mu$ once more.

Acknowledgments

The authors thank Markus Reiß for helpful discussions and his great support for making this work possible.
The research of Martin Wahl has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901.

References

M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. Method of potential functions in the problem of restoration of a functional converter characteristic by means of points observed randomly. *Avtomatika i Telemekhanika*, 25(12), 1964.

Martin Anthony and Peter L. Bartlett. *Neural network learning: theoretical foundations.* Cambridge University Press, Cambridge, 1999.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*, 2019.

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

Mikhail Belkin. Approximation beats concentration? an approximation view on inference with smooth radial kernels. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1348–1361, 2018.

Rajendra Bhatia. *Matrix analysis.* Springer-Verlag, New York, 1997.

Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl. (Singap.)*, 14(6):763–794, 2016.

Gilles Blanchard and Peter Mathé. Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems*, 28(11):115011, 23, 2012.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18(4):971–1013, 2018.

Gilles Blanchard, Marc Hoffmann, and Markus Reiß. Optimal adaptation for early stopping in statistical inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 6(3):1043–1075, 2018a.

Gilles Blanchard, Marc Hoffmann, and Markus Reiß. Early stopping for statistical inverse problems via truncated SVD estimation. *Electron. J. Stat.*, 12(2):3204–3231, 2018b.

Gilles Blanchard, Peter Mathé, and Nicole Mücke. Lepskii principle in supervised learning. *arXiv preprint arXiv:1905.10764*, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

Élodie Brunel, André Mas, and Angélina Roche. Non-asymptotic adaptive prediction in functional linear models. *J. Multivariate Anal.*, 143:208–232, 2016.

Peter Bühlmann and Bin Yu. Boosting with the L_2 loss: regression and classification. *J. Amer. Statist. Assoc.*, 98(462):324–339, 2003.

T. Tony Cai, Michael Levine, and Lie Wang. Variance function estimation in multivariate nonparametric regression with fixed design. *J. Multivariate Anal.*, 100(1):126–136, 2009.

Andrea Caponnetto. Optimal rates for regularization operators in learning theory. Technical report, MIT, 2006.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl. (Singap.)*, 8(2):161–183, 2010.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49, 2002.

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

Ernesto De Vito, Sergei Pereverzyev, and Lorenzo Rosasco. Adaptive kernel methods using the balancing principle. *Foundations of Computational Mathematics*, 10(4):455–479, 2010.

Luc Devroye, László Györfi, Gábor Lugosi, and Harro Walk. A nearest neighbor estimate of the residual variance. *Electron. J. Stat.*, 12(1):1752–1778, 2018.

Lee H. Dicker, Dean P. Foster, and Daniel Hsu. Kernel ridge vs. principal component regression: minimax bounds and the qualification of regularization operators. *Electron. J. Stat.*, 11(1):1022–1047, 2017.

Nigel Duffy and David Helmbold. Boosting methods for regression. *Machine Learning*, 47 (2-3):153–200, 2002.

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

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, New York, 2016.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT Press, Cambridge, MA, 2016.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Stanislaw Jastrzebski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of DNN loss and the SGD step length. *arXiv preprint arXiv:1807.05031*, 2018.

Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics.* Springer, Heidelberg, 2011.

Galyna Kriukova, Oleksandra Panasiuk, Sergei V. Pereverzyev, and Pavlo Tkachenko. A linear functional strategy for regularized ranking. *Neural Networks*, 73:26–35, 2016.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8157–8166, 2018.

Elia Liitiäinen, Francesco Corona, and Amaury Lendasse. Residual variance estimation using a nearest neighbor statistic. *J. Multivariate Anal.*, 101(4):811–823, 2010.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Appl. Comput. Harmon. Anal.*, 48(3):868–890, 2020.

Shuai Lu and Sergei V. Pereverzev. *Regularization theory for ill-posed problems.* De Gruyter, Berlin, 2013. Selected topics.

Peter Mathé and Sergei V. Pereverzyev. The discretized discrepancy principle under general source conditions. *Journal of Complexity*, 22(3):371–381, 2006.

William S. Meisel. Potential functions in mathematical pattern recognition. *IEEE Transactions on Computers*, 100(10):911–918, 1969.

Cassandra Milbradt and Martin Wahl. High-probability bounds for the reconstruction error of PCA. *Statist. Probab. Lett.*, 161:108741, 2020.

Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statist. Probab. Lett.*, 127:111–119, 2017.

Vladimir Alekseevich Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

Stephen Page and Steffen Grünewälder. The Goldenshluger-Lepski method for constrained least-squares estimators over RKHSs. Bernoulli, to appear, 2018.

David L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. Assoc. Comput. Mach.*, 9:84–97, 1962.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 1998.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15:335–366, 2014.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge University Press, USA, 2004.

Steve Smale and Ding-Xuan Zhou. Shannon sampling. II. Connections to learning theory. *Appl. Comput. Harmon. Anal.*, 19(3):285–302, 2005.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

Vladimir Spokoiny. Variance estimation for high-dimensional regression models. *J. Multivariate Anal.*, 82(1):111–133, 2002.

Bernhard Stankewitz. Smoothed residual stopping for statistical inverse problems via truncated SVD estimation. *Electron. J. Stat.*, 14(2):3396–3428, 2020.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT*, 2009.

Joel Tropp. An introduction to matrix concentration inequalities. Available at arxiv.org/abs/1501.01571, 2015.

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.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer-Verlag, New York, 1996.

Roman Vershynin. *High-dimensional probability*. Cambridge University Press, Cambridge, 2018.

Yuting Wei, Fanny Yang, and Martin J. Wainwright. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. *IEEE Trans. Inform. Theory*, 65(10):6685–6703, 2019.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

Tong Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 471–478, 2003.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *Ann. Statist.*, 33(4):1538–1579, 2005.

Appendix A. Some useful operator bounds

Let A, B be two positive, compact operators A and B on $\mathcal{H}$. Then we have

$$\|A^s - B^s\|_{\mathrm{op}} \leq \|A - B\|_{\mathrm{op}}^s, \quad 0 \leq s \leq 1, \tag{A.1}$$

and

$$\|A^s - B^s\|_{\mathrm{op}} \leq C_s(\|A\|_{\mathrm{op}} + \|A - B\|_{\mathrm{op}})^{s-1}\|A - B\|_{\mathrm{op}}, \quad s > 1. \tag{A.2}$$

Moreover, we have

$$\|A^s B^s\|_{\mathrm{op}} \leq \|AB\|_{\mathrm{op}}^s, \quad 0 \leq s \leq 1. \tag{A.3}$$

For a proof of the first and the third claim see Theorem X.1.1 and Theorem IX.2.1 in Bhatia (1997), for a proof of the second claim see e.g. Blanchard and Mücke (2018).

Appendix B. Effective dimension and eigenvalue bounds

The effective dimension $\mathcal{N}(t)$ of a positive self-adjoint trace-class operator Σ is a continuous and non-decreasing function in $t \geq 0$. Moreover, under (**BdK**), we have $\mathrm{tr}(\Sigma) = \mathbb{E}\|k_X\|_{\mathcal{H}}^2 \leq M^2$, leading to $\mathcal{N}(t) \leq M^2 t$ for all $t \geq 0$. Under additional assumption on the decay of the eigenvalues, this bound can be further improved.

Lemma 43 *(i) Suppose that for some $\alpha > 1$ and $L > 0$, we have $\lambda_j \leq Lj^{-\alpha}$ for all $j \geq 1$. Then there is a constant $C > 0$ depending only on α and L such that $\mathcal{N}(t) \leq Ct^{1/\alpha}$ for all $t \geq L^{-1}$.*

(ii) Suppose that for some $\alpha \in (0, 1]$ and $L > 0$, we have $\lambda_j \leq e^{-Lj^\alpha}$ for all $j \geq 1$. Then there is a constant $C > 0$ depending only on α and L such that $\mathcal{N}(t) \leq C(\log t)^{1/\alpha}$ for all $t \geq e^L$.

Proof of Lemma 43 Part (i) is proved in Proposition 3 in Caponnetto and De Vito (2007), see also Lemma 5.1 in Blanchard and Mücke (2018). In order to get part (ii), we use that $\lambda/(\lambda + 1/t)$ is increasing in λ, such that

$$\mathcal{N}(t) \leq \sum_{j \geq 1} \frac{Le^{-Lj^\alpha}}{Le^{-Lj^\alpha} + 1/t}.$$

Defining $k \geq 1$ by $e^{-L(k+1)^\alpha} < 1/t \leq e^{-Lk^\alpha}$ (using that $te^{-L} \geq 1$), we have

$$\begin{aligned}\mathcal{N}(t) &\leq \sum_{j \leq k} \frac{Le^{-Lj^\alpha}}{Le^{-Lj^\alpha} + 1/t} + \sum_{j > k} \frac{Le^{-Lj^\alpha}}{Le^{-Lj^\alpha} + 1/t} \\ &\leq k + t\sum_{j > k} e^{-Lj^\alpha} \leq k + Ct(k+1)^{1-\alpha}e^{-L(k+1)^\alpha} \leq k + C(k+1)^{1-\alpha},\end{aligned} \tag{B.1}$$

where we applied Equation (5.1) in Milbradt and Wahl (2020) in the third inequality. Now $1/t \leq e^{-Lk^\alpha}$ implies $k \leq (L^{-1}\log t)^{1/\alpha}$ and inserting this into (B.1) gives the claim. □

Lemma 44 *If $\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_\infty \leq 1/2$ holds, then*

$$\forall j \geq 1, \quad \lambda_j/2 - 1/(2T) \leq \hat{\lambda}_j \leq 3\lambda_j/2 + 1/(2T).$$

Proof of Lemma 44 We have

$$\|(\Sigma + T^{-1})^{-1/2}(\Sigma_n - \Sigma)(\Sigma + T^{-1})^{-1/2}\|_\infty \le 1/2$$

if and only if

$$-(1/2)(\langle h, \Sigma h\rangle_{\mathcal{H}} + T^{-1}) \le \langle h, (\Sigma_n - \Sigma)h\rangle_{\mathcal{H}} \le (1/2)(\langle h, \Sigma h\rangle_{\mathcal{H}} + T^{-1})$$

for every $h \in \mathcal{H}$ such that $\|h\|_{\mathcal{H}} = 1$. Rearranging the terms this is equivalent to

$$(1/2)\langle h, \Sigma h\rangle_{\mathcal{H}} - 1/(2T) \le \langle h, \Sigma_n h\rangle_{\mathcal{H}} \le (3/2)\langle h, \Sigma h\rangle_{\mathcal{H}} + 1/(2T)$$

for every $h \in \mathcal{H}$ such that $\|h\|_{\mathcal{H}} = 1$. The claim now follows from the minimax characterization of eigenvalues. □

Appendix C. Concentration inequalities

The following lemma is taken from (Dicker et al., 2017). It is an extension of (Tropp, 2015) from self-adjoint matrices to self-adjoint Hilbert-Schmidt operators.

Lemma 45 (From Lemma 5 in Dicker et al. (2017)) *Let $\xi_1, \dots, \xi_n$ be a sequence of independently and identically distributed self-adjoint Hilbert-Schmidt operators on a separable Hilbert space. Suppose that $\mathbb{E}\xi_1 = 0$ and $\|\xi_1\|_{\mathrm{op}} \le R$ almost surely for some constant $R > 0$. Moreover, suppose that there are constants $V, D > 0$ satisfying $\|\mathbb{E}\xi_1^2\|_{\mathrm{op}} \le V$ and $\mathrm{tr}(\mathbb{E}\xi_1^2) \le VD$. Then, for all $u \ge V^{1/2}n^{-1/2} + (3n)^{-1}R$,*

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n \xi_i\Big\|_{\mathrm{op}} \ge u\Big) \le 4D\exp\Big(-\frac{nu^2}{2V + (2/3)uR}\Big).$$