Electronic Journal of Statistics
Vol. 14 (2020) 3396–3428
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1747

Smoothed residual stopping for statistical inverse problems via truncated SVD estimation*

Bernhard Stankewitz

Institut für Mathematik
Humboldt-Universität zu Berlin, Germany
e-mail: stankebe@math.hu-berlin.de

Abstract: This work examines under what circumstances adaptivity for truncated SVD estimation can be achieved by an early stopping rule based on the smoothed residuals $\|(AA^\top)^{\alpha/2}(Y - A\widehat{\mu}^{(m)})\|^2$. Lower and upper bounds for the risk are derived, which show that moderate smoothing of the residuals can be used to adapt over classes of signals with varying smoothness, while oversmoothing yields suboptimal convergence rates. The range of smoothness classes for which adaptation is possible can be controlled via α. The theoretical results are illustrated by Monte-Carlo simulations.

MSC2020 subject classifications: 65J20, 62G05.
Keywords and phrases: Linear inverse problems, spectral cut-off, early stopping, discrepancy principle, adaptive estimation, oracle inequalities, weighed residuals.

Received September 2019.

1. Introduction

1.1. Preliminaries on early stopping

In machine learning and statistics, one of the central problems is that of coping with the generalisation error or, put another way, choosing the correct tuning parameter for an estimation procedure. For iterative procedures, the generalisation error typically decreases up to a point at which the algorithm begins to overfit. Hence, the problem becomes that of choosing a suitable iteration step. Classically, this problem would be addressed by model selection criteria such as *cross-validation*, *unbiased risk estimation* or *Lepski's balancing principle*. These criteria, however, require that all estimators we want to choose from be computed and then compared against each other. For high dimensional problems in particular, this may come at a computationally prohibitive cost. An alternative are *early stopping rules*, which halt the procedure at an iteration $\widehat{m}$ depending only on the iterates of index $m \leq \widehat{m}$ and potentially additional quantities computed up to that point. Since these require the computation of much fewer

*I am very grateful for long discussions with Markus Reiß and Martin Wahl and the comments of an associate editor and two anonymous referees.

iterates, they present the potential of simultaneously achieving computational and statistical efficiency.

In order to locate this work in the literature on early stopping, we shortly discuss three exemplary approaches: In practical machine learning applications, early stopping rules are widely adopted. They are usually based on a well founded heuristic understanding of the regularisation properties of early stopping. For example, the user may split the data into training and validation sets and iterate the learning algorithm on the training set until the validation error does not improve any further, see Chapter 7 in Goodfellow et al. [7]. However, proper theoretical results for such rules are lacking.

Some progress towards theoretical foundations of stopping rules has been made in the kernel learning literature. For the regression problem of learning f^* from data generated by $Y = f^*(X) + \varepsilon$, stopping rules have been suggested for gradient descent procedures, initially, via oracle stopping times, which cannot be computed from the data, see Bühlmann and Yu [4] and Caponetto et al. [5]. Later, these have been converted to data dependent rules using empirical versions of Gaussian and Rademacher complexities, see Raskutti et al. [11] and Yang et al. [14]. For example, in [11], the authors learn f^* by applying gradient descent to the problem $\min_{z\in\mathbb{R}^n} \|Y - \sqrt{K}z\|^2$, where Y is the vector of observations and K is the empirical kernel matrix. The procedure is stopped at

$$\widehat{T} := \inf\left\{t \in \mathbb{N} : \frac{1}{n}\sum_{i=1}^{n} \min\{\widehat{\lambda}_i, t^{-1/2}\} > (\sigma t)^{-1}\right\} - 1, \tag{1.1}$$

where the $(\widehat{\lambda}_i)$ are the scaled eigenvalues of K and σ is the noise level (up to a constant). This rule is computable from the data and allows to adapt to the complexity of the underlying kernel space. Yet, other than the heuristic stopping rule above, this rule structurally cannot adapt to the true data generating process. The kernel matrix and hence the sequence $(\widehat{\lambda}_i)_{i=1,\dots n}$ only depends on the design variables. Therefore, $\widehat{T}$ does not depend on f^* itself and will overfit when the true smoothness of f^* is larger than the minimal smoothness of functions from the kernel space.

Finally, additional progress has been made in the literature on statistical inverse problems, which is another important framework for learning, see e.g. Rosasco et al. [12]. Blanchard, Hoffmann and Reiß [2] consider early stopping for a D-dimensional discretisation of the inverse problem $Y = A\mu + \delta\dot{W}$ with white noise $\dot{W}$ and the sequence $(\widehat{\mu}^{(m)})_{m=1,\dots,D}$ of truncated SVD estimators. They analyse the stopping rule

$$\tau := \inf\{m \in \mathbb{N} \cup \{0\} : \|Y - A\widehat{\mu}^{(m)}\|^2 \le \delta^2 D\} \tag{1.2}$$

based on the *discrepancy principle*, which is well studied for deterministic inverse problems, see e.g. Engl et al. [6]. This problem is similar to [11] in that minimising $\|Y - \sqrt{K}z\|^2$ can also be understood as solving a finite dimensional inverse problem. The stopping rule τ, however, structurally differs from $\widehat{T}$ in that, via Y, it takes the true signal μ into account. Indeed, the authors prove

that, up to a dimension dependent error term, stopping according to τ satisfies an oracle inequality, which yields rate optimal adaptation simultaneously over a range of Sobolev-type ellipsoids of differing smoothness. Therefore, while the setting in [2] is less general than in the kernel literature, their version of early stopping is more comprehensive. In addition, their setting can be understood as a prototypical model of an iterative estimation procedure.

The analysis in this work is a continuation of the third approach above, where we stop using the (α)-smoothed residuals $\|(AA^\top)^{\alpha/2}(Y - A\widehat{\mu}^{(m)})\|^2$ for general $\alpha > 0$ instead. In the next section, we motivate in detail why this should be considered and what can be gained by it.

1.2. Model and problem formulation

We recall in detail the setting in Blanchard, Hoffmann and Reiß [2]: They consider problems of the form

$$Y = A\mu + \delta\dot{W}, \tag{1.3}$$

where $A : H_1 \to H_2$ is a linear bounded operator between real Hilbert spaces, $\mu \in H_1$ is the signal of interest, $\delta > 0$ is the noise level and $\dot{W}$ is a Gaussian white noise in H_2. In any practical application, the problem has to be discretised by the user. Therefore, we can assume that $H_1 = \mathbb{R}^D$ and $H_2 = \mathbb{R}^P$ for $D \leq P$, which both are possibly very large. Further, assume that $A : \mathbb{R}^D \to \mathbb{R}^P$ is one-to-one. By transforming (1.3), using the singular value decomposition (SVD) of A, we arrive at the Gaussian vector observation model

$$Y_i = \lambda_i\mu_i + \delta\varepsilon_i, \qquad i = 1, \dots, D. \tag{1.4}$$

$\lambda_1 \geq \lambda_2, \dots, \lambda_D > 0$ are the singular values of A, $(\mu_i)_{i\leq D}$ the coefficients of μ in the orthonormal basis of singular vectors and $(\varepsilon_i)_{i\leq D}$ are independent standard Gaussian random variables.

In order to recover the signal $\mu = (\mu_i)_{i\leq D}$ from the observation of (1.4), we use the *truncated SVD* (*cut-off*) estimators $\widehat{\mu}^{(m)}, m = 0, \dots, D$ given by

$$\widehat{\mu}_i^{(m)} := \mathbf{1}\{i \leq m\}\lambda_i^{-1}Y_i, \qquad i = 1, \dots, D. \tag{1.5}$$

For a fixed index m, the risk (expected squared Euclidean error) of $\widehat{\mu}^{(m)}$ can be decomposed into a bias and a variance term:

$$B_m^2(\mu) := \|\mathbb{E}\widehat{\mu}^{(m)} - \mu\|^2 = \sum_{i=m+1}^{D} \mu_i^2 \tag{1.6}$$

$$\text{and} \qquad V_m := \mathbb{E}\|\widehat{\mu}^{(m)} - \mathbb{E}\widehat{\mu}^{(m)}\|^2 = \sum_{i=1}^{m} \lambda_i^{-2}\delta^2. \tag{1.7}$$

In particular, the estimators are ordered with decreasing bias and increasing variance in m. We reemphasise the importance of this setting as a prototypical

model of an iterative method. Note that the truncated SVD-estimators are iterative in the sense that the SVD of the operator has to be computed alongside the estimators. This is the case, since in practice, we cannot expect the observation vector Y to be represented in an SVD basis, see also the detailed discussion in [2] and the references therein. Other iterative methods often share important qualitative features with cut-off estimation. Therefore, results from this simple framework typically carry over to more complex settings. For example, Blanchard, Hoffmann and Reiß [3] transfer the results of [2] to general regularisation schemes, including gradient descent.

In [2], the authors consider stopping according to the discrepancy principle, i.e. at the smallest m which satisfies

$$\|Y - A\widehat{\mu}^{(m)}\|^2 \leq \kappa \tag{1.8}$$

for a suitable critical value $\kappa > 0$. Their analysis shows that generally, stopping according to the condition in (1.8) is optimal (in terms of an oracle inequality) up to a dimension dependent error term, which stems from the variability of the residuals. For signals μ, which are not too smooth relative to the approximation dimension D, this term is of lower order. More precisely, (1.8) yields optimal results simultaneously for all signals satisfying $m^{\flat}(\mu) \gtrsim \sqrt{D}$, where

$$m^{\flat}(\mu) := \inf\{m \geq 0 : B_m^2(\mu) \leq V_m\} \tag{1.9}$$

is the index at which balance between the squared bias and variance is obtained. Otherwise, random deviations in the residuals systematically lead to stopping times which are too large.

Alternatively, Blanchard and Mathé [1] apply the discrepancy principle to the normal equation $A^\top Y = A^\top A\mu$ and stop according to

$$\|A^\top(Y - A\widehat{\mu}^{(m)})\|^2 = \|(AA^\top)^{1/2}(Y - A\widehat{\mu}^{(m)})\|^2 \leq \kappa, \tag{1.10}$$

i.e. the residuals are smoothed by $(AA^\top)^{1/2}$. This is motivated by the fact that in the infinite-dimensional problem, $A^*\dot{W}$ can be represented as an element of H_1 when A is Hilbert-Schmidt. The condition in (1.10) is able to to control the stochastic part of the residuals and avoid the dimension-dependency from [2]. Yet, it typically results in suboptimal convergence rates, since the variability of the residuals is reduced too much, which leads to stopping times which are too small.

These results raise the question of whether there is a stopping criterion in between (1.8) and (1.10) which is able to mitigate the dimension-dependency from [2] and thereby increase the range of signals for which adaptation is possible without slipping into the suboptimal regime discussed in [1]. A very natural consideration is to smooth the residuals by a general power $\alpha \geq 0$ of $(AA^\top)^{1/2}$ and stop at the smallest index m which satisfies

$$R_{m,\alpha}^2 := \|(AA^\top)^{\alpha/2}(Y - A\widehat{\mu}^{(m)})\|^2 \leq \kappa, \tag{1.11}$$

where $R^2_{m,\alpha}$ are the (α-)*smoothed residuals*. The main contribution of this paper is to answer the posed question in the affirmative for the criterion in (1.11), provided that the inverse problem is moderately ill-posed. Smoothing with $\alpha > 0$ reduces the variability of $R^2_{m,\alpha}$, which mitigates the constraint from [2]. For values of α which are small relative to the decay of the singular values of A, smoothing does not produce suboptimal rates. Additionally, it is possible to eliminate the dimension constraint entirely before the oversmoothing effect from [1] manifests. In order to further motivate stopping according to $R^2_{m,\alpha}$, we compare it to other possible generalisations of the discrepancy principle:

Remark 1.1 (Other discrepancy-type rules).

(a) Blanchard and Mathé [1] also choose a stopping criterion in between (1.8) and (1.10) in order to guarantee optimality. They weigh the residuals in (1.10) further by $\varrho_\lambda(A^\top A)$ for $\varrho_\lambda(t) := 1/\sqrt{t+\lambda}$, $t > 0$ and a tuning parameter λ. In their framework, however, the final choice of λ directly depends on the smoothness of the true signal and only yields optimal results for this smoothness class. Therefore, their stopping criterion will not adapt simultaneously to signals of varying smoothness, which is precisely the goal of our analysis.

(b) Other well founded variations of the discrepancy principle mostly take the form

$$\|H_m(AA^\top)(Y - A\widehat{\mu}^{(m)})\|^2 \leq \kappa, \tag{1.12}$$

i.e. the weight of $(AA^\top)$ depends on m, see e.g. Engl et al. [6]. Compared to the smoothed residuals, such a rule is computationally more expensive: In our setting, the computation of the first m estimators roughly requires $O(mD^2)$ operations, see [2]. With the update $R^2_{m+1,\alpha} = R^2_{m,\alpha} - \lambda^{2\alpha}_{m+1} Y^2_{m+1}$, the additional computational cost of the smoothed residuals is negligible. Note that the m-th eigenvalue λ_m already has to be computed for $\widehat{\mu}^{(m)}$. In contrast, computing (1.12) for $i = 0, \dots, m$ potentially requires $O(mD^2)$ operations itself. If we regard early stopping as a tool to treat the computational complexity of the problem, this provides further motivation for the (α)-smoothed residuals.

The remainder of the paper is structured as follows: In Section 2, we collect the structural assumptions of the analysis and provide an interpretation of the smoothed residual stopping procedure in (1.11) as estimating the bias of a smoothed version of the risk. At the end, we present the main results of the paper, which are derived in Section 3. Its constraints in terms of lower bounds are explored in Section 4. Finally, Section 5 discusses different choices for the smoothing parameter α and illustrates the results by Monte-Carlo simulations.

2. Framework for the analysis and main results

2.1. Structural assumptions

Throughout the paper, we assume that the inverse problem is moderately ill-posed, i.e. the singular values $(\lambda_i)_{i\le D}$ satisfy a *polynomial spectral decay* assumption of the form

$$C_A^{-1} i^{-p} \le \lambda_i \le C_A i^{-p}, \qquad i = 1, \dots, D \qquad (\mathrm{PSD}(p, C_A)) \tag{2.1}$$

for some $p \ge 0$ and $C_A \ge 1$. By dividing Equation (1.4) by λ_1, we can further assume that $\lambda_i \le 1$, $i = 1, \dots, D$. Additionally, we always require that the critical value κ satisfies

$$|\kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2| \le C_\kappa s_D \delta^2 \qquad \text{with} \qquad s_D^2 := 2\sum_{i=1}^{D} \lambda_i^{4\alpha} \tag{2.2}$$

for an absolute constant $C_\kappa > 0$.

Note that $\sum_{i=1}^{D} \lambda_i^2 \delta^2$ is the expectation of the smoothed residuals for the zero signal at $m = 0$, since

$$R_{0,\alpha}^2 = \sum_{i=1}^{D} \left(\lambda_i^{2+2\alpha}\mu_i^2 + 2\lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i + \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 \right). \tag{2.3}$$

Similarly, $s_D\delta^2$ is the standard deviation of the dominant stochastic part of the term above. Therefore, (2.2) states that up to small deviations, κ should be chosen as the expectation of the smoothed residuals in the pure noise case.

In the following, we denote essential inequalities up to an absolute constant by "$\lesssim, \gtrsim, \sim$". Further dependencies on α, the operator A, i.e. p and C_A, and C_κ, are denoted by indices α, A and κ. Finally, we assume that all smoothing indices α are bounded from above by some $\bar\alpha > 0$. This guarantees that $\lambda_i^\alpha \sim_A i^{-\alpha p}$, $i \le D$. Under $(\mathrm{PSD}(p, C_A))$, the order of s_D is given by

$$s_D \sim_{\alpha,A} \begin{cases} D^{1/2-2\alpha p}, & \alpha p < 1/4, \\ \log D, & \alpha p = 1/4, \\ 1, & \alpha p > 1/4. \end{cases} \tag{2.4}$$

The fact that the order of s_D is decreasing in α will later allow to relax the constraint from Blanchard et al. [2]. The variance of $\widehat{\mu}^{(m)}$ is of order

$$V_m = \sum_{i=1}^{m} \lambda_i^{-2}\delta^2 \sim_A m^{2p+1}\delta^2. \tag{2.5}$$

For the analysis of lower bounds in Section 4, we consider signals from *Sobolev-type ellipsoids*

$$H^\beta(r, D) := \left\{ \mu \in \mathbb{R}^D : \sum_{i=1}^{D} i^{2\beta}\mu_i^2 \le r^2 \right\} \qquad \text{for some } \beta \ge 0, r > 0. \tag{2.6}$$

For $\mu \in H^\beta(r, D)$, we have the upper bound

$$B_m^2(\mu) = \sum_{i=m+1}^{D} \mu_i^2 \leq (m+1)^{-2\beta} r^2 \tag{2.7}$$

for the squared bias of $\widehat{\mu}^{(m)}$. The bounds in (2.5) and (2.7) are balanced at the order of the *minimax-truncation index*

$$t^{mm}_{\beta,p,r} = t^{mm}_{\beta,p,r}(\delta) := (r^2\delta^{-2})^{1/(2\beta+2p+1)}. \tag{2.8}$$

Taking the asymptotic view that $D = D(\delta) \to \infty$ for $\delta \to 0$, the rate v_δ^2 is optimal in the minimax sense if there exist estimators $(\widehat{\mu}_\delta)_{\delta>0}$ in the models corresponding to the ellipsoids $H^\beta(r, D(\delta))$ such that

$$\limsup_{\delta\to 0} v_\delta^{-2} \sup_{\mu\in H^\beta(r,D(\delta))} \mathbb{E}\|\widehat{\mu}_\delta - \mu\|^2 < \infty \tag{2.9}$$

and

$$\liminf_{\delta\to 0} v_\delta^{-2} \inf_{\widehat{\mu}} \sup_{\mu\in H^\beta(r,D(\delta))} \mathbb{E}\|\widehat{\mu} - \mu\|^2 > 0, \tag{2.10}$$

where the infimum is taken over all estimators $\widehat{\mu}$. A deterministic stopping index of the order of the minimax truncation index $t^{mm}_{\beta,p,r}$ in (2.8) yields the rate

$$\mathcal{R}^*_{\beta,p,r}(\delta) := r^2(r^{-2}\delta^2)^{2\beta/(2\beta+2p+1)}. \tag{2.11}$$

This is the minimax rate in the infinite-dimensional Gaussian sequence model. Note that lower bounding the minimax risk in the infinite-dimensional case, up to a constant, only requires to consider alternatives in the first $t^{mm}_{\beta,p,r}$ components, see e.g. Proposition 4.23 in Johnstone [8]. Therefore, if $D(\delta)$ is chosen at least of the order of $t^{mm}_{\beta,p,r}$, the rate $\mathcal{R}^*_{\beta,p,r}(\delta)$ is also minimax in our setting. In the asymptotic considerations, we will always assume that this is the case, since we can also think of $t^{mm}_{\beta,p,r}$ as the minimally sufficient approximation dimension. Indeed, the error of approximating a signal from an infinite-dimensional Sobolev ellipsoid of smoothness β by a signal from $H^\beta(r, D)$ will only be negligible if $D(\delta) \gtrsim t^{mm}_{\beta,p,r}$.

2.2. Smoothed residual stopping as bias estimation

For a clearer formulation of the results, we introduce continuous versions of the bias and the variance by linearly interpolating Equations (1.6) and (1.7). For $t \in [0, D]$, we set

$$B_t^2(\mu) := (\lceil t\rceil - t)\mu^2_{\lceil t\rceil} + \sum_{i=\lceil t\rceil+1}^{D} \mu_i^2 \tag{2.12}$$

$$\text{and} \qquad V_t := \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{-2}\delta^2 + (t - \lfloor t \rfloor)\lambda_{\lceil t \rceil}^{-2}\delta^2, \tag{2.13}$$

where $\lfloor t \rfloor$ and $\lceil t \rceil$ are the floor and ceiling functions, respectively. We can define a continuous cut-off estimator $\widehat{\mu}^{(t)}$ such that $\mathbb{E}\|\widehat{\mu}^{(t)} - \mu\|^2 = B_t^2(\mu) + V_t, t \in [0, D]$: By randomising between the discrete estimators with index $\lfloor t \rfloor$ and $\lceil t \rceil$, we set

$$\widehat{\mu}_i^{(t)} := (\mathbf{1}\{i \le \lfloor t \rfloor\} + \xi_t \mathbf{1}\{i = \lceil t \rceil\})\lambda_i^{-1} Y_i, \qquad i = 1, \dots, D, \tag{2.14}$$

where ξ_t are Bernoulli random variables with success probabilities $t - \lfloor t \rfloor$ independent of everything else. This also gives a continuous version of the smoothed residuals:

$$\begin{aligned} R_{t,\alpha}^2 &:= \|(AA^\top)^{\alpha/2}(Y - A\widehat{\mu}^{(t)})\|^2 \\ &= (\mathbf{1}\{t \ne \lceil t \rceil\} - \xi_t)\lambda_{\lceil t \rceil}^{2\alpha} Y_{\lceil t \rceil}^2 + \sum_{i=\lceil t \rceil + 1}^{D} \lambda_i^{2\alpha} Y_i^2 \end{aligned} \tag{2.15}$$

for $t \in [0, D]$. The (α-)*smoothed residual stopping time*

$$\tau_\alpha := \inf\{m \in \mathbb{N} \cup \{0\} : R_{m,\alpha}^2 \le \kappa\} \tag{2.16}$$

yet remains integer. In the following, integer indices are denoted by m and continuous indices are denoted by t.

Applying optional stopping to the martingale $M_m := \sum_{i=1}^m \lambda_i^{-2}(\varepsilon_i^2 - 1)$, $m \le D$, yields

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 = \mathbb{E}\Big(\sum_{i=\tau_\alpha+1}^{D} \mu_i^2 + \sum_{i=1}^{\tau_\alpha} \lambda_i^{-2}\delta^2\varepsilon_i^2 \Big) = \mathbb{E}\big(B_{\tau_\alpha}^2(\mu) + V_{\tau_\alpha}\big). \tag{2.17}$$

Therefore, at best, the risk at τ_α behaves like the risk at the *classical oracle index*

$$t^{\mathfrak{c}} = t^{\mathfrak{c}}(\mu) := \operatorname*{argmin}_{t \in [0,D]} \mathbb{E}\|\widehat{\mu}^{(t)} - \mu\|^2 \tag{2.18}$$

which minimises the risk over all deterministic stopping indices. There is, however, no direct connection between τ_α and $t^{\mathfrak{c}}$. This is intrinsic to the sequential nature of the analysis, since at truncation index t, we cannot say anything about the behaviour of the bias for larger indices.

For our purposes, we instead consider the *balanced oracle index*

$$t^{\mathfrak{b}} = t^{\mathfrak{b}}(\mu) := \inf\{t \ge 0 : B_t^2(\mu) \le V_t\}. \tag{2.19}$$

Due to the continuity of the functions $t \mapsto V_t$ and $t \mapsto B_t^2(\mu)$, we have that at $t^{\mathfrak{b}}$, squared bias and variance balance exactly, i.e. $B_{t^{\mathfrak{b}}}^2(\mu) = V_{t^{\mathfrak{b}}}$. Furthermore, the

balanced oracle risk is comparable to the classical oracle risk: The monotonicity of $t \mapsto V_t$ and $t \mapsto B_t^2(\mu)$ yields

$$\mathbb{E}\|\widehat{\mu}^{(t^\flat)} - \mu\|^2 = B_{t^\flat}^2(\mu) + V_{t^\flat} \leq 2\mathbb{E}\|\widehat{\mu}^{(t^\natural)} - \mu\|^2 \tag{2.20}$$

by distinguishing the cases $t^\natural \leq t^\flat$ and $t^\natural > t^\flat$. Assuming that the operator A and the noise level δ are known, knowledge of the bias is therefore enough to stop at an index at which the risk is of the order of the classical oracle risk.

The smoothed residuals $R_{t,\alpha}^2$ contain some information about the bias: We can write

$$\mathbb{E}R_{t,\alpha}^2 = B_{t,\alpha}^2(\mu) + \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 - V_{t,\alpha}, \qquad t \in [0, D], \tag{2.21}$$

where the α-bias and the α-variance

$$B_{t,\alpha}^2(\mu) := (\lceil t \rceil - t)\lambda_{\lceil t \rceil}^{2+2\alpha}\mu_{\lceil t \rceil}^2 + \sum_{i=\lceil t \rceil+1}^{D} \lambda_i^{2+2\alpha}\mu_i^2 \tag{2.22}$$

and

$$V_{t,\alpha} := \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{2\alpha}\delta^2 + (t - \lfloor t \rfloor)\lambda_{\lceil t \rceil}^{2\alpha}\delta^2 \tag{2.23}$$

are smoothed versions of $B_t^2(\mu)$ and V_t. Since $\lambda_i \leq 1$ for all $i = 1, \dots, D$, the smoothed quantities $B_{t,\alpha}^2$ and $V_{t,\alpha}$ are always smaller than their nonsmoothed counterparts. Analogously to $t^\flat$, we define the *α-balanced oracle*

$$t_\alpha^\flat = t_\alpha^\flat(\mu) := \inf\{t \geq 0 : B_{t,\alpha}^2(\mu) \leq V_{t,\alpha}\} \tag{2.24}$$

at which the squared α-bias and the α-variance balance.

The stopping condition $R_{m,\alpha}^2 \leq \kappa$ can be reformulated as

$$\widehat{B}_{m,\alpha}^2(\mu) := R_{m,\alpha}^2 + V_{m,\alpha} - \kappa \leq V_{m,\alpha}, \tag{2.25}$$

which yields

$$\tau_\alpha = \inf\{m \geq 0 : \widehat{B}_{m,\alpha}^2(\mu) \leq V_{m,\alpha}\}. \tag{2.26}$$

Due to (2.21), $\widehat{B}_{m,\alpha}^2(\mu)$ is an unbiased estimator of $B_{m,\alpha}^2(\mu)$ for $\kappa = \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2$. Therefore, stopping according to τ_α can be understood as estimating the α-bias and stopping when the estimate is smaller than the α-variance. For the specific choice of κ above, τ_α directly mimics $t_\alpha^\flat$. For other choices of κ, τ_α mimics the (α-)*oracle-proxy index*

$$t_\alpha^* = t_\alpha^*(\mu) := \inf\{t \geq 0 : \mathbb{E}\widehat{B}_{t,\alpha}^2 \leq V_{t,\alpha}\} = \inf\{t \geq 0 : \mathbb{E}R_{t,\alpha}^2 \leq \kappa\}. \tag{2.27}$$

This is illustrated in Figure 1. The oracle-proxy index satisfies

$$\begin{cases} t_\alpha^* > t_\alpha^\flat, & \kappa < \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2, \\ t_\alpha^* = t_\alpha^\flat, & \kappa = \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2, \\ t_\alpha^* < t_\alpha^\flat, & \kappa > \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2. \end{cases} \tag{2.28}$$

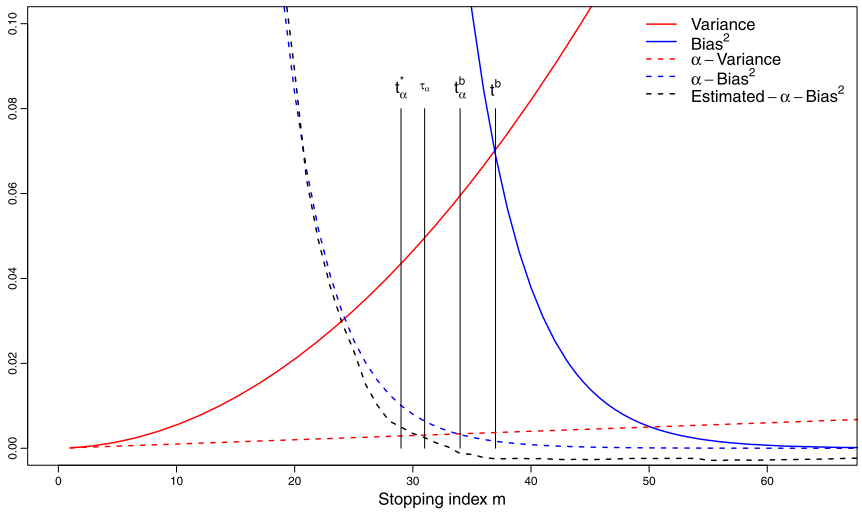

FIG 1. *Bias estimation with oracle indices. Here, $\alpha = 0$ to ensure that all curves fit into one plot.*

Assumption (2.2) can therefore be understood as a requirement on the difference between t^*_α and $t^\flat_\alpha$. So far, this yields the following picture: Approximately, τ_α is centred around the oracle proxy t^*_α, which is close to the α-balanced oracle $t^\flat_\alpha$ for an appropriate choice of κ. In turn, $t^\flat_\alpha$ is related to the balanced oracle $t^\flat$ due to the connection between the bias and the variance and their smoothed counterparts. Generally, we can therefore hope for adaptation as long as $t^\flat_\alpha$ and $t^\flat$ are of the same size.

With respect to the difference between $t^\flat_\alpha$ and $t^\flat$, we note:

Lemma 2.1. *The mapping $\alpha \mapsto t^\flat_\alpha, \alpha \geq 0$ is monotonously decreasing in α. Further, $t^\flat_\alpha \leq t^\flat$ for all $\alpha \geq 0$.*

Proof. Let $\alpha, \alpha' \geq 0$ with $\alpha \leq \alpha'$. Then, for any $t \in [0, D]$ which satisfies $B^2_{t,\alpha}(\mu) \leq V_{t,\alpha}$, we have

$$B^2_{t,\alpha'}(\mu) \leq \lambda^{2(\alpha'-\alpha)}_{\lceil t \rceil} B^2_{t,\alpha}(\mu) \leq \lambda^{2(\alpha'-\alpha)}_{\lceil t \rceil} V_{t,\alpha} \leq V_{t,\alpha'}. \tag{2.29}$$

Analogous reasoning yields $t^\flat_\alpha \leq t^\flat$ for all $\alpha \geq 0$. □

Therefore, smoothing increases the difference between $t^\flat_\alpha$ and $t^\flat$ and will generally induce smaller stopping times τ_α.

Under (PSD(p, C_A)), we also have essential upper bounds for $t^\flat_\alpha$ and $t^\flat_\alpha$: For $t^\flat$, the bounds on the size of the bias and the variance in (2.7) and (2.5) show that

$$t^\flat(\mu) \lesssim_A t^{mm}_{\beta,p,r}(\delta) = (r^2\delta^{-2})^{1/(2\beta+2p+1)} \qquad \text{for all } \mu \in H^\beta(r, D). \tag{2.30}$$

For $t_\alpha^\mathfrak{b}$, analogously to (2.7) and (2.5), we obtain

$$B_{m,\alpha}^2(\mu) \lesssim_A r^2 m^{-(2\beta+2p+2\alpha p)} \qquad \text{for all } \mu \in H^\beta(r,D) \tag{2.31}$$

$$\text{and} \quad V_{m,\alpha} \sim_A \begin{cases} m^{1-2\alpha p}\delta^2/(1-2\alpha p), & \alpha p < 1/2, \\ \log(m)\delta^2, & \alpha p = 1/2, \\ \delta^2, & \alpha p > 1/2 \end{cases} \tag{2.32}$$

for sufficiently large values of $m \geq 0$. Given that $t_\alpha^\mathfrak{b}$ is large enough, this gives the essential upper bound

$$t_\alpha^\mathfrak{b}(\mu) \lesssim_A t_{\beta,p,r,\alpha}^{mm}(\delta) \qquad \text{for all } \mu \in H^\beta(r,D), \tag{2.33}$$

where

$$t_{\beta,p,r,\alpha}^{mm} = t_{\beta,p,r,\alpha}^{mm}(\delta) := \begin{cases} ((1-2\alpha p) r^2\delta^{-2})^{1/(2\beta+2p+1)}, & \alpha p < 1/2, \\ (r^2\delta^{-2}/\log(r^2\delta^{-2}))^{1/(2\beta+2p+1)}, & \alpha p = 1/2, \\ (r^2\delta^{-2})^{1/(2\beta+2p+2\alpha p)}, & \alpha p > 1/2 \end{cases} \tag{2.34}$$

is the *α-minimax truncation* index.

For $\alpha p < 1/2$, $t_{\beta,p,r,\alpha}^{mm}$ is of the same order as $t_{\beta,p,r}^{mm}$, but smoothing shrinks $t_{\beta,p,r,\alpha}^{mm}$ by a power of $(1-2\alpha p)$. In the same way, we obtain that for $\alpha p \geq 1/2$, the α-balanced oracle is of order strictly smaller than the minimax-truncation index $t_{\beta,p,r}^{mm}(\delta)$. Since there are signals $\mu \in H^\beta(r,D)$, for which $t^\mathfrak{b}(\mu) \sim t_{\beta,p,r}^{mm}(\delta)$, we can therefore only expect to achieve adaptation on $H^\beta(r,D)$ as long as $\alpha p < 1/2$.

2.3. Main results

Based on the understanding of the stopping procedure developed in Sections 2.1 and 2.2, we can now formulate our main theorem. It provides an oracle inequality for the risk at τ_α in terms of the risk at the balanced oracle $t^\mathfrak{b}$.

Theorem 2.2 (Balanced oracle inequality)**.** *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then, there exists a constant $C_{\alpha,A,\kappa}$ depending on α, p, C_A and C_k such that*

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \leq C_{\alpha,A,\kappa}\big(\mathbb{E}\|\widehat{\mu}^{(t^\mathfrak{b})} - \mu\|^2 + s_D^{(2p+1)/(1-2\alpha p)}\delta^2\big).$$

For $t^\mathfrak{b} \gtrsim_{\alpha,A,\kappa} s_D^{1/(1-2\alpha p)}$, the risk of stopping at τ_α is of the order of the balanced-oracle risk.

Theorem 2.2 is derived in Section 3.

We comment on the result: $s_D^{(2p+1)/(1-2\alpha p)}\delta^2$ is a dimension-dependent error term. Since $V_t \sim_A t^{2p+1}\delta^2$, it is of order $V_{s_D^{1/(1-2\alpha p)}}$. Its existence stems from the

stochastic variability of the residuals, which is discussed in Section 4.1. Since the risk at $t^\flat$ is of the order of $V_{t^\flat}$, this error term is of lower order as long as

$$t^\flat \gtrsim_{\alpha,A,\kappa} s_D^{1/(1-2\alpha p)} \sim_{\alpha,A,\kappa} \begin{cases} D^{\frac{1/2-2\alpha p}{1-2\alpha p}}, & \alpha p < 1/4, \\ (\log D)^2, & \alpha p = 1/4, \\ 1, & \alpha p > 1/4. \end{cases} \tag{2.35}$$

Equation (2.35) determines for what signals we can obtain optimal estimation results and shows the advantage of smoothing: For $\alpha = 0$, we obtain the same result as in Blanchard et al. [2], i.e. we need to require $t^\flat \gtrsim_{A,\kappa} \sqrt{D}$. For values $\alpha > 0$, this constraint is weakened and thereby guarantees that the dimension dependent error term is of lower order for a larger class of signals. For $\alpha p = 1/4$, the error is only a log-term. For $\alpha p > 1/4$, it is of constant size.

Intuitively, under our assumptions, τ_α behaves like $t_\alpha^\flat$. As seen in Lemma 2.1, $t_\alpha^\flat$ is monotonously decreasing in α. While decreasing the variance, smoothing therefore increases the squared bias $B^2_{t_\alpha^\flat}(\mu)$. For $\alpha p < 1/2$, this results in an increase in the constant $C_{\alpha,A,\kappa}$. For $\alpha p \geq 1/2$, $t_\alpha^\flat$ and $t^\flat$ can be of different order such that the squared bias at $t_\alpha^\flat$ is strictly larger than the risk at $t^\flat$. Then, an oracle inequality is no longer possible. The details of this are further discussed in Section 4.2. One of the basic assumptions in Blanchard and Mathé [1] is that A is Hilbert-Schmidt. In our setting, this is the case when $p > 1/2$, which is the exact point when the discrepancy principle for the normal equation, i.e. $\alpha = 1$, loses the optimal rate. Therefore, the above reasoning provides a nice explanation for their nonoptimality result.

Finally, our result directly translates to an asymptotic minimax upper bound over the Sobolev-type ellipsoids $H^\beta(r, D)$: When $D = D(\delta) \to \infty$ for $\delta \to 0$, the risk at $t^\flat$ is of optimal order when $D(\delta)$ grows faster than the minimax truncation index $t^{mm}_{\beta,p,r}(\delta)$, see the discussion in Section 2.1. The same is true for the dimension-dependent error as long as $s_D^{1/(1-2\alpha p)} \lesssim_{\alpha,A,\kappa} t^{mm}_{\beta,p,r}$. Therefore, we obtain:

Corollary 2.3 (Adaptive rates for Sobolev ellipsoids)**.** *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then, there exists a constant $C_{\alpha,A,\kappa}$ depending on α, p, C_A and C_k such that*

$$\sup_{\mu \in H^\beta(r,D)} \mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \leq C_{\alpha,A,\kappa} \mathcal{R}^*_{\beta,p,r}(\delta)$$

for any $\beta, r > 0$ with $D \gtrsim t^{mm}_{\beta,p,r} \gtrsim_{\alpha,A,\kappa} s_D^{1/(1-2\alpha p)}$.

By comparing the size of $t^{mm}_{\beta,p,r} = (r^2\delta^{-2})^{1/(2\beta+2p+1)}$ with s_D, Corollary 2.3 yields a range of Sobolev-type ellipsoids $H^\beta(r, D)$ for which stopping according to the smoothed residual stopping time τ_α is simultaneously minimax adaptive. How this can used to choose a suitable smoothing parameter α is further discussed in Section 5.1.

3. Derivation of the main results

In this section, we derive the result in Theorem 2.2. By defining the stochastic error term

$$S_t := \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{-2}\delta^2\varepsilon_i^2 + (t - \lfloor t \rfloor)\lambda_{\lceil t \rceil}^{-2}\delta^2\varepsilon_{\lceil t \rceil}^2, \qquad t \in [0, D], \tag{3.1}$$

we obtain

$$\mathbb{E}\|\widehat{\mu}^{(t)} - \mu\|^2 = \mathbb{E}(B_t^2(\mu) + S_t), \quad t \in [0, D] \tag{3.2}$$

$$\text{and} \qquad \mathbb{E}\|\hat{\mu}^{\tau_\alpha} - \mu\|^2 = \mathbb{E}(B_{\tau_\alpha}^2(\mu) + S_{\tau_\alpha}). \tag{3.3}$$

This allows to decompose the difference between the risk at the smoothed residual stopping time τ_α and the risk at any deterministic index $t \in [0, D]$ into a bias part and a stochastic part:

$$\mathbb{E}\|\hat{\mu}^{(\tau_\alpha)} - \mu\|^2 - \mathbb{E}\|\hat{\mu}^{(t)} - \mu\|^2 \le \mathbb{E}\big(B_{\tau_\alpha}^2(\mu) - B_t^2(\mu)\big)^+ + \mathbb{E}(S_{\tau_\alpha} - S_t)^+. \tag{3.4}$$

3.1. An oracle-proxy inequality

Initially, we compare the risk at the smoothed residual stopping time τ_α with the risk at the oracle-proxy index t_α^*. For the bias part in (3.4), we can further decompose:

$$\begin{aligned}\mathbb{E}\big(B_{\tau_\alpha}^2(\mu) - B_t^2(\mu)\big)^+ &\le \lambda_{\lceil t \rceil}^{-(2+2\alpha)}\mathbb{E}\big(B_{\tau_\alpha,\alpha}^2(\mu) - B_{t,\alpha}^2(\mu)\big)^+ \\ &\le \lambda_{\lceil t \rceil}^{-(2+2\alpha)}\Big[\mathbb{E}\big(B_{\tau_\alpha,\alpha}^2(\mu) - B_{t_\alpha^*,\alpha}^2(\mu)\big)^+ + \big(B_{t_\alpha^*,\alpha}^2(\mu) - B_{t,\alpha}^2(\mu)\big)^+\Big].\end{aligned} \tag{3.5}$$

In Appendix A.1, we bound the probability $\mathbb{P}\{\tau_\alpha \le m\}$ for $m \ge 0$ to derive the following estimate for the first term in the square brackets:

Proposition 3.1. *For any signal $\mu \in \mathbb{R}^D$, we have*

$$\mathbb{E}\big(B_{\tau_\alpha,\alpha}^2(\mu) - B_{t_\alpha^*,\alpha}^2(\mu)\big)^+ \le C(B_{t_\alpha^*,\alpha}^2(\mu) + s_D\delta^2),$$

where $C \ge 1$ is an absolute constant.

Plugging the bound from Proposition 3.1 into (3.5) gives an inequality for the bias part in (3.4).

Corollary 3.2. *For any signal $\mu \in \mathbb{R}^D$ and $t \in [0, D]$, we have*

$$\mathbb{E}\big(B_{\tau_\alpha}^2(\mu) - B_t^2(\mu)\big)^+ \le C\lambda_{\lceil t \rceil}^{-(2+2\alpha)}\big(B_{t_\alpha^*,\alpha}^2(\mu) + s_D\delta^2\big),$$

where $C \ge 1$ is an absolute constant.

In Appendix A.1, we also bound the probability $\mathbb{P}\{\tau_\alpha \ge m\}$ for $m \ge 0$, which yields the following bound for the stochastic part in (3.4):

Proposition 3.3. *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then,*

$$\mathbb{E}(S_{\tau_\alpha} - S_{t^*_\alpha})^+ \le C_{\alpha,A,\kappa}\big(V_{t^*_\alpha} + s_D^{(2p+1)/(1-2\alpha p)}\delta^2\big),$$

where $C_{\alpha,A,\kappa} \ge 1$ is a constant depending on α, p, C_A and C_κ.

Together, Corollary 3.2 and Proposition 3.3 show that under a set of fairly general assumptions, the risk at the smoothed residual stopping time τ_α essentially behaves like the risk at the deterministic oracle-proxy index t^*_α. Note that the result holds for all signals $\mu \in \mathbb{R}^D$ and not only for Sobolev-type ellipsoids.

Theorem 3.4 (Oracle-proxy inequality)**.** *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then, there exists a constant $C_{\alpha,A,\kappa}$ depending on α, p, C_A and C_κ such that*

$$\mathbb{E}\|\hat\mu^{(\tau_\alpha)} - \mu\|^2 \le C_{\alpha,A,\kappa}\big(\mathbb{E}\|\hat\mu^{(t^*_\alpha)} - \mu\|^2 + s_D^{(2p+1)/(1-2\alpha p)}\delta^2\big).$$

For $t^*_\alpha \gtrsim_{\alpha,A,\kappa} s_D^{1/(1-2\alpha p)}$, the risk at τ_α is of the order of the risk at t^*_α.

Proof. After plugging the inequalities from Corollary 3.2 and Proposition 3.3 into Equation (3.4) with $t = t^*_\alpha$, only the remaining bias part has to be estimated. For any $m \ge 0$, however, we have $\lambda_m^{-(2+2\alpha)} B^2_{m,\alpha}(\mu) \le B^2_m(\mu)$. This yields

$$\begin{aligned}\lambda_{\lceil t^*_\alpha\rceil}^{-(2+2\alpha)}(B^2_{t^*_\alpha,\alpha}(\mu) + s_D\delta^2) &\lesssim_A B^2_{t^*_\alpha}(\mu) + (t^*_\alpha)^{2p+2\alpha p} s_D\delta^2 \\ &\lesssim_A B^2_{t^*_\alpha}(\mu) + V_{t^*_\alpha} + s_D^{(2p+1)/(1-2\alpha p)}\delta^2\end{aligned} \tag{3.6}$$

by distinguishing the cases where t^*_α is smaller or greater than $s_D^{1/(1-2\alpha p)}$. □

The proof of Proposition 3.3 relies on the growth of $m \mapsto V_{m,\alpha} - B^2_{m,\alpha}(\mu)$ for $m \ge \lceil t^*_\alpha\rceil$, which can be insufficient for $\alpha p \ge 1/2$ even if we assume that $\mu \in H^\beta(r, D)$. This suggests that a result as in Theorem 3.4 for $\alpha p \ge 1/2$ requires additional assumptions on the decay of $m \mapsto B^2_{m,\alpha}(\mu)$. We note a sufficient condition from the literature, see e.g. Kindermann and Neubauer [9] or Szabó et al. [13].

Remark 3.5 (Oracle-proxy inequality under polished tails). Assume that the signal μ is not only an element of $H^\beta(r, D)$ but additionally the projection onto the first D components of an infinite-dimensional signal $\tilde\mu$, which satisfies a *polished tail condition* of the form

$$\sum_{i=m}^{\infty} \tilde\mu_i^2 \le C_0 \sum_{i=m}^{\rho m} \tilde\mu_i^2 \qquad \text{for all } m \ge 1 \tag{3.7}$$

for an integer constant $\rho \ge 2$ and $C_0 > 0$. Then, we have $\mathbb{E}(S_{\tau_\alpha} - S_{t^*_\alpha})^+ \lesssim_{A,\kappa} (t^*_\alpha)^{2p+1}\delta^2$ also when $\alpha p \ge 1/2$ and $\kappa \ge \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2$, see Proposition A.1(i) in Appendix A.2. Under this condition, we obtain $\mathbb{E}\|\hat\mu^{(\tau_\alpha)} - \mu\|^2 \lesssim_A \mathbb{E}\|\hat\mu^{(t^*_\alpha)} - \mu\|^2$ the same way as in Theorem 3.4.

3.2. Comparison of the oracle risks

In this section, we derive the balanced oracle inequality in Theorem 2.2 from the oracle-proxy inequality in Theorem 3.4. We do this by comparing the different bias and variance quantities at $t^*_\alpha, t^\mathfrak{b}_\alpha$ and $t^\mathfrak{b}$. Initially, we bound the difference between the α-risk terms at t^*_α and $t^\mathfrak{b}_\alpha$.

Lemma 3.6. *We have*

$$(V_{t^*_\alpha,\alpha} - V_{t^\mathfrak{b}_\alpha,\alpha})^+ \le \Big(\sum_{i=1}^D \lambda_i^{2\alpha}\delta^2 - \kappa\Big)^+$$

and

$$\big(B^2_{t^*_\alpha,\alpha}(\mu) - B^2_{t^\mathfrak{b}_\alpha,\alpha}(\mu)\big)^+ \le \Big(\kappa - \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2\Big)^+.$$

Proof. For the first inequality, we assume without loss of generality that $t^\mathfrak{b}_\alpha < t^*_\alpha$. The monotonicity of $t \mapsto B^2_{t,\alpha}(\mu)$, the fact that $\mathbb{E}R^2_{t^*_\alpha,\alpha} = \kappa$ and Equation (2.21) yield

$$\begin{aligned} V_{t^*_\alpha,\alpha} &= B^2_{t^*_\alpha,\alpha}(\mu) + \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2 - \kappa \le B^2_{t^\mathfrak{b}_\alpha,\alpha}(\mu) + \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2 - \kappa \qquad (3.8)\\ &\le V_{t^\mathfrak{b}_\alpha,\alpha}(\mu) + \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2 - \kappa. \end{aligned}$$

For the second inequality, we analogously assume without loss of generality that $t^*_\alpha < t^\mathfrak{b}_\alpha$. The monotonicity of $t \mapsto V_{t,\alpha}$, the fact that $\mathbb{E}R^2_{t^*_\alpha,\alpha} \le \kappa$ and Equation (2.21) then yield

$$\begin{aligned} B^2_{t^*_\alpha,\alpha}(\mu) &\le V_{t^*_\alpha,\alpha} + \kappa - \sum_{i=1}^D \lambda_i^2\delta^2 \le V_{t^\mathfrak{b}_\alpha,\alpha} + \kappa - \sum_{i=1}^D \lambda_i^2\delta^2 \qquad (3.9)\\ &\le B^2_{t^\mathfrak{b}_\alpha,\alpha}(\mu) + \kappa - \sum_{i=1}^D \lambda_i^2\delta^2. \end{aligned}$$

□

Under our assumptions, the first inequality in Lemma 3.6 allows to bound the size of t^*_α:

Corollary 3.7. *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then,*

$$t^*_\alpha \lesssim_{\alpha,A,\kappa} t^\mathfrak{b}_\alpha + s_D^{1/(1-2\alpha p)} \le t^\mathfrak{b} + s_D^{1/(1-2\alpha p)}.$$

Proof. Under (PSD(p, C_A)) with $\alpha p < 1/2$, we have

$$\delta^{-2}(V_{t^*_\alpha,\alpha} - V_{t^\mathfrak{b}_\alpha}) \gtrsim_A \int_{t^\mathfrak{b}_\alpha}^{t^*_\alpha} t^{-2\alpha p}\,dt = \frac{(t^*_\alpha)^{1-2\alpha p} - (t^\mathfrak{b}_\alpha)^{1-2\alpha p}}{1-2\alpha p}. \qquad (3.10)$$

Now, the result follows from Lemma 3.6 and assumption (2.2). □

We can now essentially compare the order of the risk at t^*_α, $t^\flat_\alpha$ and $t^\flat$.

Proposition 3.8 (Comparison of the oracle risks)**.** *Assume (PSD(p, C_A)) with $\alpha p < 1/2$ and (2.2). Then,*

$$\mathbb{E}\|\hat\mu^{(t^\flat_\alpha)} - \mu\|^2 \sim_{\alpha,A,\kappa} \mathbb{E}\|\hat\mu^{(t^\flat)} - \mu\|^2$$
$$\text{and} \qquad \mathbb{E}\|\hat\mu^{(t^*_\alpha)} - \mu\|^2 \lesssim_{\alpha,A,\kappa} \mathbb{E}\|\hat\mu^{(t^\flat)} - \mu\|^2 + s_D^{(2p+1)/(1-2\alpha p)}\delta^2.$$

Proof. For the second statement, we note that, as in (3.4), we can write

$$\mathbb{E}\|\hat\mu^{(t^*_\alpha)} - \mu\|^2 - \mathbb{E}\|\hat\mu^{(t^\flat)} - \mu\|^2 \le \big(B^2_{t^*_\alpha}(\mu) - B^2_{t^\flat}(\mu)\big)^+ + (V_{t^*_\alpha} - V_{t^\flat})^+. \quad (3.11)$$

We treat the two terms on the right-hand side separately. For the bias part, we can assume $t^*_\alpha \le t^\flat$. Analogously to (3.5), we have

$$\begin{aligned} B^2_{t^*_\alpha}(\mu) - B^2_{t^\flat}(\mu) &\le \lambda_{t^\flat}^{-(2+2\alpha)} B^2_{t^*_\alpha,\alpha}(\mu) \qquad (3.12)\\ &\le \lambda_{t^\flat}^{-(2+2\alpha)}\big(V_{t^\flat,\alpha} + \kappa - \sum_{i=1}^{D}\lambda_i^{2\alpha}\delta^2\big)\\ &\lesssim_{\alpha,A,\kappa} \big((t^\flat)^{2p+1} + (t^\flat)^{2p+2\alpha p}s_D\big)\delta^2, \end{aligned}$$

since $V_{t^\flat,\alpha} \lesssim_{\alpha,A,\kappa} (t^\flat)^{1-2\alpha p}\delta^2$.

For the variance part, we can assume $t^*_\alpha \ge t^\flat$ and obtain

$$\begin{aligned} V_{t^*_\alpha} - V_{t^\flat} &\le \lambda^{-(2+2\alpha)}_{\lceil t^*_\alpha\rceil}(V_{t^*_\alpha,\alpha} - V_{t^\flat,\alpha}) \lesssim_{A,\kappa} (t^*_\alpha)^{2p+2\alpha p}s_D\delta^2 \qquad (3.13)\\ &\lesssim_{\alpha,A,\kappa} ((t^\flat)^{2p+2\alpha p}s_D + s_D^{(2p+1)/(1-2\alpha p)})\delta^2 \end{aligned}$$

using Lemma 3.6 and Corollary 3.7. The intended inequality now follows from

$$(t^\flat)^{2p+1}\delta^2 \sim_A V_{t^\flat} \sim \mathbb{E}\|\hat\mu^{(t^\flat)} - \mu\|^2 \quad (3.14)$$

and distinguishing the cases where $t^\flat$ is smaller or greater than $s_D^{1/(1-2\alpha p)}$.

The essential inequality "$\lesssim_{\alpha,A,\kappa}$" in the first statement follows by replacing t^*_α with $t^\flat_\alpha$ in (3.12) and noting both that $B^2_{t^\flat_\alpha,\alpha}(\mu) = V_{t^\flat_\alpha,\alpha}$ and $V_{t^\flat_\alpha} \le V_{t^\flat}$, since $t^\flat_\alpha \le t^\flat$. The reverse direction "$\gtrsim_{\alpha,A,\kappa}$" follows immediately from the fact that the risk at $t^\flat$ is always of smaller order than the risk at any other $t \in [0, D]$. □

Together with Theorem 3.4, Proposition 3.8 yields the result in Theorem 2.2.

4. Constraints in terms of lower bounds

4.1. Undersmoothing for $\alpha p \le 1/4$

The first constraint in Theorem 2.2 is the dimension-dependent error term

$$s_D^{(2p+1)/(1-2\alpha p)}\delta^2 \sim_A V_{s_D^{1/(1-2\alpha p)}}. \quad (4.1)$$

We show that an error of this order is unavoidable: From the identity in (2.17) and the monotonicity of $t \mapsto V_t$, we obtain that for any $i_0 \in \{0, \dots, D\}$,

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \geq \mathbb{E}(V_{i_0} \mathbf{1}\{\tau_\alpha \geq i_0\}) \geq \mathbb{P}\{\tau_\alpha \geq i_0\} V_{i_0}. \tag{4.2}$$

By considering the zero signal $\mu = 0$, we can isolate the error, which stems directly from the stochastic variability of the smoothed residuals. In Appendix A.2, we show that for $\alpha p \leq 1/4$, we stop later than $i_0 = s_D^{1/(1-2\alpha p)}$ with nonvanishing probability for $D = D_\delta \to \infty$ when $\delta \to 0$. This causes a dimension-dependent error of the size $V_{s_D^{1/(1-2\alpha p)}}$. Since this reasoning can be extended to $\mu \neq 0$, we obtain:

Proposition 4.1 (Dimension-dependent lower bound)**.** *Assume (PSD(p, C_A)) with $\alpha p \leq 1/4$ and (2.2). Then, we have for any $\mu \in \mathbb{R}^D$ that*

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \geq C s_D^{(2p+1)/(1-2\alpha p)} \delta^2$$

with an absolute constant $C > 0$, provided that δ is sufficiently small and $D = D_\delta \to \infty$ for $\delta \to 0$.

The proof of Proposition 4.1 shows that decreasing the admissible order of $|\kappa - \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2|$ beyond $s_D\delta^2$ does not decrease the order of the lower bound. At the same time, increasing the admissible order of $|\kappa - \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2|$ may increase the order of the lower bound. This further motivates assumption (2.2).

4.2. Oversmoothing for $\alpha p \geq 1/2$

The second constraint in Theorem 2.2 is $\alpha p < 1/2$. We already anticipated in Section 2.3 that for $\alpha p \geq 1/2$, an oracle inequality is no longer possible, since $t_\alpha^\mathfrak{b}$ can be of strictly smaller order than $t^\mathfrak{b}$. We make this precise by providing a lower bound. Analogously to (4.2), the monotonicity of $t \mapsto B_t^2(\mu)$ yields that for any $i_0 \in \{0, \dots, D\}$, we have

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \geq \mathbb{E}(B_{i_0}^2(\mu) \mathbf{1}\{\tau_\alpha \leq i_0\}) \geq \mathbb{P}\{\tau_\alpha \leq i_0\} B_{i_0}^2(\mu). \tag{4.3}$$

Intuitively, τ_α centres around t_α^*. Therefore, we can hope to bound the probability in (4.3) from below against a constant when i_0 is of the order of t_α^*. If $t_\alpha^* \leq t_\alpha^\mathfrak{b}$, this gives a bound in terms of $B_{t_\alpha^\mathfrak{b}}^2(\mu)$. From (2.28), we have that $t_\alpha^* \leq t_\alpha^\mathfrak{b}$ exactly when $\kappa \geq \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2$. Under this assumption, we obtain:

Proposition 4.2 (α-balanced oracle lower bounds)**.** *Assume (PSD(p, C_A)) and $\kappa \geq \sum_{i=1}^D \lambda_i^{2\alpha}\delta^2$. Then, there exists a constant $C_A' > 0$ depending on p and C_A such that*

$$\sup_{\mu \in H^\beta(r,D)} \mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \geq C_A' \mathcal{R}_{\beta,r,p,\alpha}(\delta)$$

with

$$\mathcal{R}_{\beta,r,p,\alpha}(\delta) := \begin{cases} r^2\big(r^{-2}\delta^2/(1-2\alpha p)\big)^{2\beta/(2\beta+2p+1)}, & \alpha p < 1/2, \\ r^2(r^{-2}\delta^2 \log(r^2\delta^{-2}))^{2\beta/(2\beta+2p+1)}, & \alpha p = 1/2, \\ r^2(r^{-2}\delta^2)^{2\beta/(2\beta+2p+2\alpha p)}, & \alpha p > 1/2, \end{cases}$$

provided that δ is sufficiently small and $t^{mm}_{\beta,p,r}(\delta) = o(D)$.

The proof is postponed to Appendix A.2.

Proposition 4.2 directly reflects the bound on $t^{\flat}_{\alpha}$ from (2.33). As long as $\alpha p < 1/2$, the lower bound is of the order of the minimax rate $\mathcal{R}^*_{\beta,r,p}(\delta)$, however, we lose a power of $1/(1-2\alpha p)$ in the constant. This is exactly what would be expected from the possible loss of smoothing in the size of $t^{\flat}_{\alpha}$ deduced in (2.33). Note that this result also implies that the constant in Theorem 2.2 grows at least this fast in α. For $\alpha p \geq 1/2$, the balanced oracles $t^{\flat}_{\alpha}$ and $t^{\flat}$ are of different order. Since τ_α reflects the size of $t^{\flat}_{\alpha}$ rather than $t^{\flat}$, we oversmooth and stop too early such that rate optimal adaptation is no longer possible.

For $\alpha = 1$ and $p > 1/2$, the lower bound for $\alpha p > 1/2$ in Proposition 4.2 is the same rate that Blanchard and Mathé [1] achieve via the discrepancy principle for the normal equation (up to a log-factor). In our setting, this also is the correct rate. In Appendix A.2, we separately control the stochastic error for $\alpha p \geq 1/2$. We can then prove:

Proposition 4.3. *Assume (PSD(p, C_A)) with $\alpha p \geq 1/2$, $\kappa \geq \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2$ and (2.2). Then, there exists a constant $C_{A,\kappa}$ depending on p, C_A and C_κ such that*

$$\sup_{\mu \in H^\beta(r,D)} \mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \leq C_{A,\kappa}\mathcal{R}_{\beta,r,p,\alpha}(\delta) \tag{4.4}$$

with $\mathcal{R}_{\beta,r,p,\alpha}(\delta)$ from Proposition 4.2.

Proof. From (3.4) with $t = t^{mm}_{\beta,p,r,\alpha}$ and Proposition A.1(ii) in Appendix A.2, we obtain

$$\begin{aligned} &\mathbb{E}\|\hat{\mu}^{(\tau_\alpha)} - \mu\|^2 - \mathbb{E}\|\hat{\mu}^{(t^{mm}_{\beta,p,r,\alpha})} - \mu\|^2 \\ &\lesssim_{A,\kappa} \lambda^{-(2+2\alpha)}_{\lceil t^{mm}_{\beta,p,r,\alpha}\rceil}\big(B^2_{t^*_\alpha,\alpha}(\mu) + s_D\delta^2\big) + (t^{mm}_{\beta,p,r,\alpha})^{2p+1}\delta^2 \\ &\lesssim_{A,\kappa} \lambda^{-(2+2\alpha)}_{\lceil t^{mm}_{\beta,p,r,\alpha}\rceil}\big(V_{t^*_\alpha,\alpha} + s_D\delta^2\big) + (t^{mm}_{\beta,p,r,\alpha})^{2p+1}\delta^2 \\ &\lesssim_{A,\kappa} (t^{mm}_{\beta,p,r,\alpha})^{2p+2\alpha p} V_{t^{mm}_{\beta,p,r,\alpha},\alpha} + (t^{mm}_{\beta,p,r,\alpha})^{2p+1}\delta^2. \end{aligned} \tag{4.5}$$

Since $V_{t^{mm}_{\beta,p,r,\alpha}} \sim_A \log(t^{mm}_{\beta,p,r,\alpha})\delta^2$ for $\alpha p = 1/2$ and $V_{t^{mm}_{\beta,p,r,\alpha}} \sim_A \delta^2$ for $\alpha p > 1/2$, this gives the result. □

Note that in addition to (2.2), an assumption on κ such as $\kappa \geq \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2$ is necessary for $\alpha p \geq 1/2$. Otherwise, $\kappa = 0$ satisfies (2.2), which yields that $\tau_\alpha = D$.

TABLE 1
Overview of the smoothing regimes.

$\alpha p < 1/4$	$1/4 \leq \alpha p < 1/2$	$1/2 \leq \alpha p$
Risk at $t^{\flat}_{\alpha} \sim$ Risk at $t^{\flat}$	Risk at $t^{\flat}_{\alpha} \sim$ Risk at $t^{\flat}$	Risk at $t^{\flat}_{\alpha} \not\lesssim$ Risk at $t^{\flat}$, $t^{\flat}_{\alpha} \not\gtrsim t^{\flat}$, stop too early
Dimension error	No dimension error	Dimension error
Stop too late for $t^{\flat} \not\gtrsim s_D^{1/(1-2\alpha p)}$		
Undersmoothing	Loss in the constant	Oversmoothing

5. Discussion and simulations

The results from Sections 2, 3 and 4 reveal three different smoothing regimes: For $\alpha p \leq 1/4$, the risk at $t^{\flat}_{\alpha}$ is of the same order as the risk at $t^{\flat}$. There is, however, an dimension-dependent error present and we potentially stop too late when $t^{\flat} \not\gtrsim_{\alpha,A,\kappa} s_D^{1/(1-2\alpha p)}$, i.e. we undersmooth. For $1/4 < \alpha p < 1/2$, the risk at $t^{\flat}_{\alpha}$ is still of the same order as the risk at $t^{\flat}$ and the dimension-dependent error disappears. Note, however, that we lose in the constant $C_{\alpha,A,\kappa}$ from Theorem 2.2, which was discussed in detail after Proposition 4.2. For $1/2 \leq \alpha p$, the risk at $t^{\flat}_{\alpha}$ can be of smaller order than the risk at $t^{\flat}$. We potentially stop too early, i.e. we oversmooth. This is summarised in Table 1.

In Section 5.2, we discuss particular choices of α and in Section 5.2, we compare our theoretical results with the estimation results for simulated data.

5.1. *Choosing the smoothing parameter α*

We consider the problem of choosing a suitable smoothing parameter $\alpha \geq 0$ in order to adaptively estimate signals from $H^{\beta}(r, D)$ for fixed $r > 0$ and a range of smoothness levels β in $[\beta_{\min}, \infty)$. Here, we assume that $\beta_{\min}$ is a minimal a priori smoothness available to the user. This yields the minimally sufficient approximation dimension $D \sim t^{mm}_{\beta_{\min},p,r} = (r^2\delta^{-2})^{1/(2\beta_{\min}+2p+1)}$, see the discussion in Section 2.1. Note that the choice $\beta_{\min} = 0$, which provides a sufficient approximation for any degree of smoothness $\beta \geq 0$, may already be computationally feasible. For $D \sim t^{mm}_{\beta_{\min},p,r}$, the size of the standard deviation term is of order

$$s_D \sim_{\alpha,A} \begin{cases} (t^{mm}_{\beta_{\min},p,r})^{1/2-2\alpha p}, & \alpha p < 1/4, \\ \log t^{mm}_{\beta_{\min},p,r}, & \alpha p = 1/4, \\ 1, & \alpha p > 1/4. \end{cases} \tag{5.1}$$

When a maximal degree of smoothness $\beta_{\max}$ is known, the user may consider the tradeoff between the smoothing parameter α and the constant $C_{\alpha,A,\kappa}$ in Theorem 2.2. The optimal smoothing index is then given by the smallest α, which guarantees adaptation over all $\beta \in [\beta_{\min}, \beta_{\max}]$. By Corollary 2.3, this index is given by the smallest $\alpha \in [0, 1/(4p))$ such that

$$t^{mm}_{\beta_{\max},p,r} \gtrsim_{\alpha,A,\kappa} (t^{mm}_{\beta_{\min},p,r})^{\frac{1/2-2\alpha p}{1-2\alpha p}}, \tag{5.2}$$

$$\text{i.e.} \quad 2\beta_{\max} + 2p + 1 \leq \frac{1 - 2\alpha p}{1/2 - 2\alpha p}(2\beta_{\min} + 2p + 1).$$

When no such $\beta_{\max}$ is known, the natural choice for the smoothing index is $\alpha = 1/(4p)$, which is the smallest index at which the dimension dependent error is of lower order for any $\beta \geq \beta_{\min}$: Theorem 2.2 together with (2.20) yields that for all $\beta \geq \beta_{\min}$,

$$\mathbb{E}\|\hat{\mu}^{(\tau_{1/(4p)})} - \mu\|^2 \leq C_{A,\kappa}\Big(\min_{t\in[0,D]} \mathbb{E}\|\hat{\mu}^{(t)} - \mu\|^2 + (\log t^{mm}_{\beta_{\min},p,r})^{2(2p+1)}\delta^2 \Big) \quad (5.3)$$

with a constant $C_{A,\kappa}$ depending on p, C_A and C_κ. For any $\beta \geq \beta_{\min}$, $t^{mm}_{\beta,p,r}$ is essentially larger than $\log t^{mm}_{\beta_{\min},p,r}$ up to a constant depending on β. Therefore, for any $\beta \geq \beta_{\min}$,

$$\mathbb{E}\|\hat{\mu}^{(\tau_{1/(4p)})} - \mu\|^2 \leq C_{A,\kappa,\beta}\mathcal{R}^*_{\beta,p,r}(\delta) \qquad \text{for all } \mu \in H^\beta(r, D) \quad (5.4)$$

with a constant $C_{A,\kappa,\beta} > 0$ which depends on p, C_A, C_κ and β.

This clearly shows the advantage of smoothing compared to no smoothing: We can directly influence the range of adaptation, whereas whithout smoothing, the range is fixed and we cannot expect to adapt to signals of smoothness greater than $2\beta_{\min} + p + 1/2$. Additionally, the discussion above yields a natural choice for α, i.e. $\alpha = 1/(4p)$, which in particular depends only on the degree of the polynomial spectral decay p. This choice can further be optimised given additional information about $\beta_{\max}$.

Finally, we may not have access to arbitrary powers of $(AA^\top)$ and only be able to choose between $\alpha = 0$ and $\alpha = 1$. For the direct comparison of nonsmoothed residual stopping and the discrepancy principle for the normal equation, our results show the following: As long as $p < 1/2$, we should clearly prefer the $\alpha = 1$. When p is only slightly larger than $1/2$, no method is clearly better than the other and our choice should depend on the size of D and possibly additional prior knowledge about the signals we want to estimate. Finally, when p is substantially larger than $1/2$, we should prefer nonsmoothed residual stopping. In particular, the two-step procedure from Blanchard et al. [2] – when computationally affordable – should produce uniformly better results, since we neither pay in the rate nor in the constant.

5.2. Estimation results for simulated data

In this section, the properties of smoothed residual stopping, which have been analysed in the previous sections are illustrated by Monte Carlo simulations. Analogous to the simulations in Blanchard et al. [2], we set

$$\delta = 0.01, \quad p = 0.5, \quad \lambda_i = i^{-p}, \quad i = 1, \dots D \qquad \text{and} \qquad \kappa = \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 \quad (5.5)$$

such that $t^*_\alpha = t^{\mathfrak{b}}_\alpha$. In this setting, the natural parameter choice from Section 5.1 is $\alpha = 1/(4p) = 0.5$. The threshhold at which we enter the oversmoothing regime

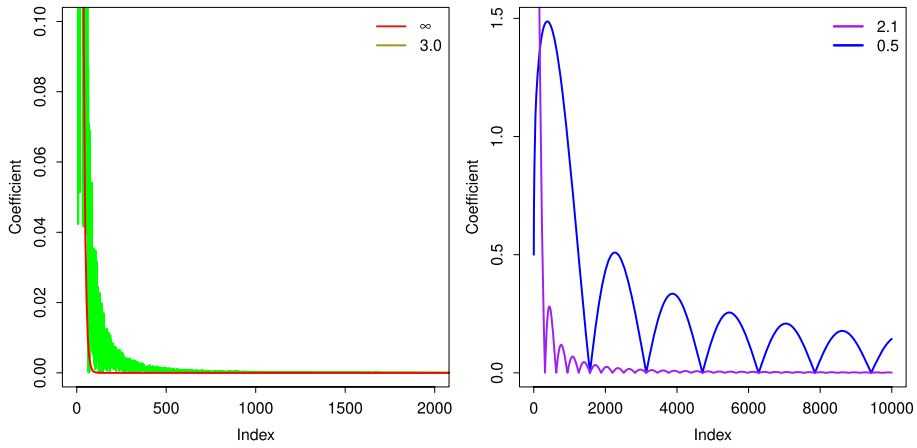

FIG 2. *SVD coefficients for four signals of different smoothness.*

is $\alpha = 1/(2p) = 1$. We consider the signals $\mu^{(\infty)}, \mu^{(3.0)}, \mu^{(2.1)}$ and $\mu^{(0.5)}$ defined by

$$\begin{aligned} \mu_i^{(\infty)} &= 5\exp(-0.1i), & \mu_i^{(3.0)} &= 500|U_i|i^{-2.05}, \\ \mu_i^{(2.1)} &= 5000|\sin(0.01i)|i^{-1.6}, & \mu_i^{(0.5)} &= 250|\sin(0.002i)|i^{-0.8}, \end{aligned} \tag{5.6}$$

with $(U_i)_{i\le D}$ independent standard uniform random variables. $\mu^{(\infty)}, \mu^{(2.1)}$ and $\mu^{(0.5)}$ are the *supersmooth*, *smooth* and *rough* signals from [2], respectively. The random signal $\mu^{(3.0)}$ will further illustrate the effect of gradually increasing the smoothing index α. All signals are indexed by their smoothness parameter 2β for the corresponding Sobolev-type ellipsoid $H^\beta(r, D)$, i.e. they are ordered $(\mu^{(\infty)}, \mu^{(3.0)}, \mu^{(2.1)}, \mu^{(0.5)})$ from smooth to rough. The SVD coefficients $(\mu_i)_{i\le D}$ of the signals and their decay are illustrated in Figure 2.

Initially, we set $D = D_\delta = 10000$ to make our results directly comparable with [2]. In this setting, the integer valued classical oracle indices of $(\mu^{(\infty)}, \mu^{(3.0)}, \mu^{(2.1)}, \mu^{(0.5)})$ are given by $(43, 58, 504, 1331)$. The balanced counterparts are $(37, 52, 445, 2379)$. For any of the signals, 1000 realisations of the model

$$Y_i = \lambda_i \mu_i + \delta\varepsilon_i, \qquad i = 1, \dots D \tag{5.7}$$

are simulated. For each of these, we calculate the smoothed residual stopping time τ_α for smoothing parameters $\alpha \in \{0, 0.2, 0.5, 1, 1.5\}$. As in [2], we compute the *relative efficiency*

$$\big(\min_{m\le D} \mathbb{E}\|\widehat{\mu}^{(m)} - \mu\|^2\big)^{1/2} / \|\widehat{\mu}^{(\tau_\alpha)} - \mu\|, \tag{5.8}$$

which serves as an estimate for the inverse of the square root of the constant between $\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2$ and $\mathbb{E}\|\widehat{\mu}^{(t^\epsilon)} - \mu\|^2$. Additionally, we determine the *relative stopping time* $\lceil t_\alpha^\flat \rceil / \tau_\alpha$. Boxplots of these quantities are presented in Figure 3.

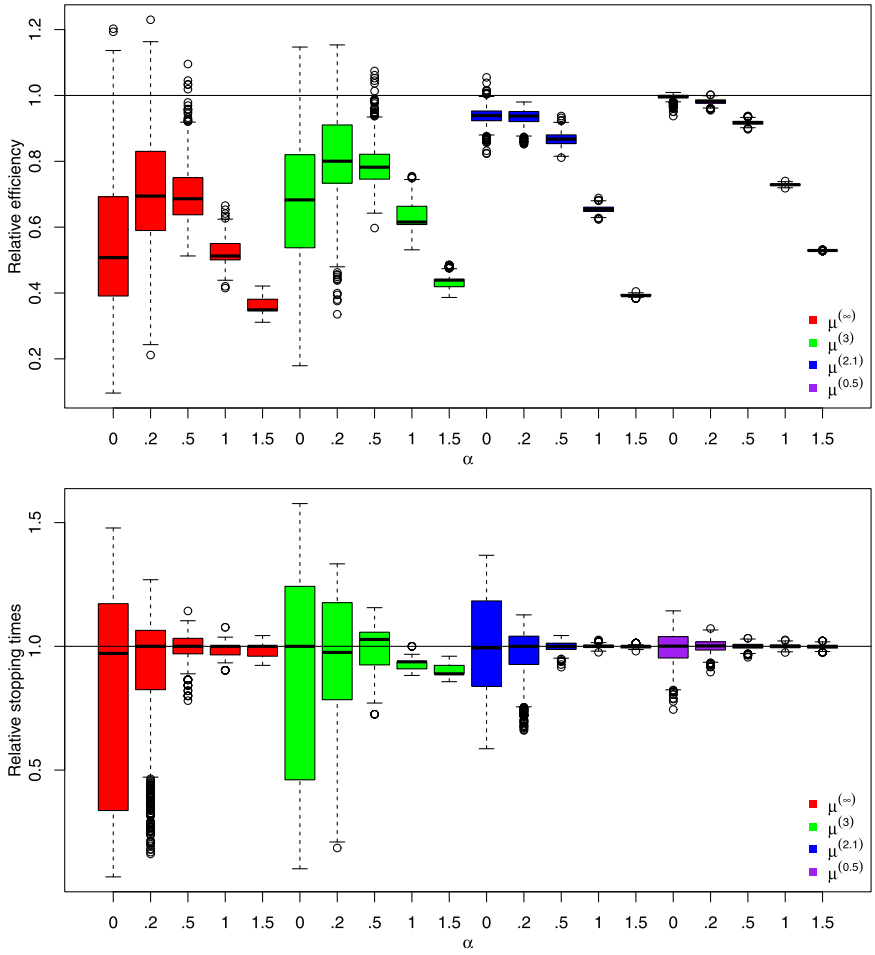

FIG 3. *Boxplots of simulation results for* $D = 10000$.

The simulation of the relative efficiency closely matches the theoretical results. For no to little smoothing of the residuals, i.e. $\alpha \in \{0, 0.2\}$, the risk of estimating the smooth signals $\mu^{(\infty)}$ and $\mu^{(3)}$ is clearly dominated by the dimension dependent error term in Theorem 2.2, i.e. we are in the undersmoothing regime, see Table 1.

This is evident, since the relative efficiency does not concentrate well and can take values close to zero, i.e. the loss at the stopping time can be much larger than the oracle risk. Smoothing is able to mitigate this. Indeed, for the natural parameter choice $\alpha = 1/(4p) = 0.5$, the relative efficiency concentrates around a reasonable constant across all signals. Note, however, that for the rougher signals $\mu^{(2.1)}$ and $\mu^{(0.5)}$, smoothing has worsened the constant. This shows that the tradeoff between the range of adaptation and the constant dis-

cussed in Sections 4.2 and 5.1 cannot be neglected in practice. Finally, we observe a clear dropoff in the quality of estimation over all signals for $a \geq 1/(2p) = 1$, which is also expected from Table 1, since we are entering the oversmoothing regime.

The same effects are illustrated by the behaviour of the stopping time itself. The boxplots of $\lceil t_\alpha^\flat \rceil / \tau_\alpha$ reflect our findings from Section 3.1 that τ_α centers around t_α^*, which is equal to $t_\alpha^\flat$ in our case. For $\alpha \in \{0, 0.2\}$, we are in the undersmoothing regime and large deviations from $t_\alpha^\flat$ are possible due to the result in Proposition 4.1. By gradually increasing α, these vanish and for $\alpha \geq 1$, τ_α evermore resembles the deterministic stopping time $t_\alpha^\flat$. Numerical evaluation of $t_\alpha^\flat$ shows that for $\alpha \geq 1$, $t_\alpha^\flat$ itself rapidly decreases for all signals considered, resulting in stopping times which are substantially too early. This increases the bias of $\hat{\mu}^{(\tau_\alpha)}$, which explains the loss in the relative efficiency. The size of the loss suggests that for $\alpha \geq 1$, we are indeed in the oversmoothing regime.

Finally, we directly illustrate the behaviour of convergence rates in the asymptotical setting where $D_\delta \to \infty$ for $\delta \to 0$. We consider the estimation for the super-smooth signal $\mu^{(\infty)}$ and the rough signal $\mu^{(0.5)}$. For different smoothing indices α, these already display all three possible regimes for the convergence rate. In the simulations, we use values of $D = D_k = 100 \cdot 2^k$ for $k = 0, \dots, 10$ with corresponding noise levels

$$\delta_k = \sqrt{r_{\max}^2 / D_k^{2\beta_{\min}+2p+1}}, \qquad k = 0, \dots, 10 \tag{5.9}$$

where $r_{\max} = 1000$ and $2\beta_{\min} = 0.5$. In this scenario, $D_{0.01} = 10000$ as before and D_{δ_k} grows as the minimax truncation $t_{0.5,p,r}^{mm}$ index of the rough signal $\mu^{(0.5)}$, i.e. we assume that we want to be able to cover signals up to at least this roughness. Again, we simulate 1000 realisations from (5.7) and consider the stopped estimator for smoothing indices $\alpha \in \{0, 0.2, 0.5, 1, 1.5\}$. We take the mean squared loss as an estimate for the risk and compare the convergence behaviour of the stopped estimator with the optimal rate, which is achieved by stopping at $t^\natural$ and the rate of stopping deterministically at $\sqrt{D}$, which gives the dimension-dependent rate $D^{(2p+1)/2}\delta^2 = D\delta^2$ from Proposition 4.1 for no smoothing. The results are displayed in Figure 4.

We consider the results for $\mu^{(\infty)}$. For $\alpha = 0$, we are in the undersmoothing regime and obtain the $D\delta^2$-rate, i.e. we do about as good as stopping at a deterministic index of size $\sqrt{D}$. This is exactly what we would expect from the lower bound in Proposition 4.1. Smoothing of the residuals improves the rate. Numerical calculations show that the simulated behaviour for $\alpha = 1/(4p) = 0.5$ is optimal up to a factor of 2.5. Note, however, that for $\alpha = 1/(2p) = 1$, the results already deteriorate again, which is consistent with the fact that this is the threshhold case from Table 1 at which we should lose rate optimality. Finally, for $\alpha = 1.5$, we are deep into the oversmoothing regime and obtain substantially suboptimal behaviour.

For $\mu^{(0.5)}$, the picture is different. Since $\mu^{(0.5)}$ is particularly rough, the risk initially increases with the approximation dimension, simply because a larger part of the signal is considered. As $t^\flat$ is always substantially greater than $\sqrt{D}$, we

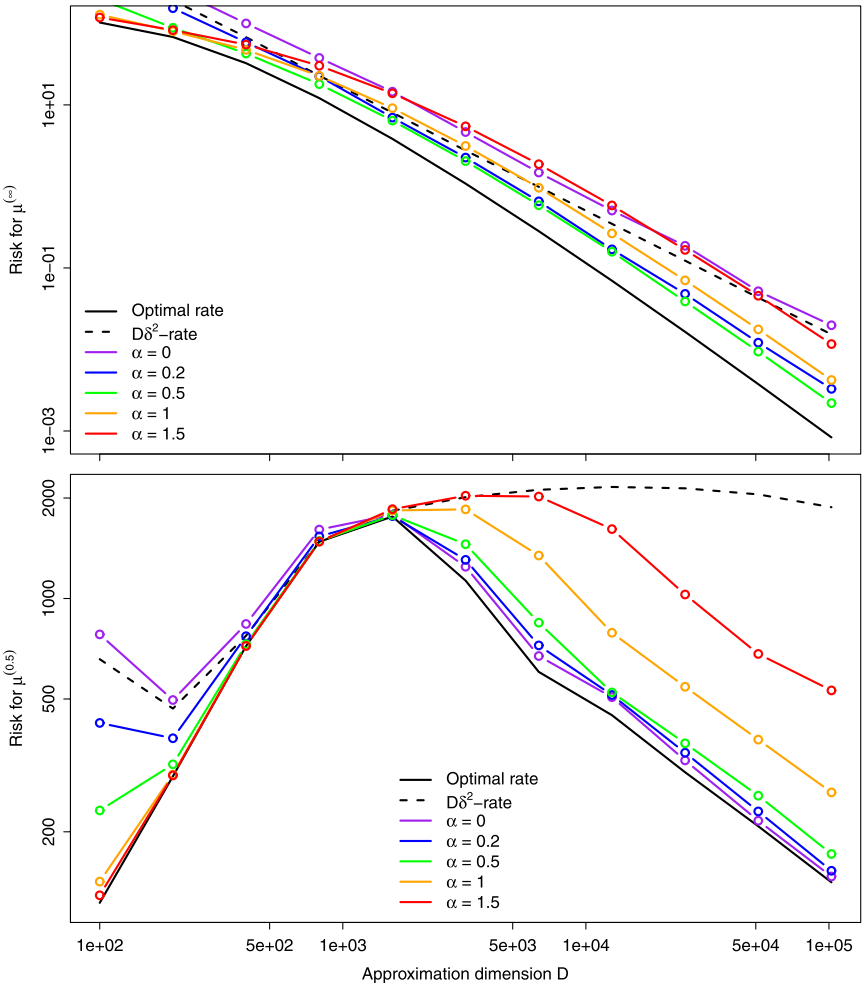

FIG 4. *Log-log plot of the risk rates for $\mu^{(\infty)}$ and $\mu^{(0.5)}$ and different smoothing indices.*

never suffer from undersmoothing due to the stochastic variability of the residuals. Therefore, $\alpha = 0$ outperforms all other indices. As predicted by Proposition 4.2, the results deteriorate with increasing α. For $\alpha \in \{0, 0.2, 0.5\}$, however, they group tightly together. This is exactly what is expected from the theoretical results, since for smoothing up to the natural choice $\alpha = 1/(4p) = 0.5$, we should only observe a loss in the constant but not in the rate. For values of α greater than the threshhold $1/(2p) = 1$, we clearly observe oversmoothing.

Summarising, the simulations reiterate the theoretical results from Sections 2.3 and 4 as well as the discussion in Section 5.1. In particular, the parameter choice $\alpha = 1/(4p)$ yields reasonable estimation results across the board for all signals. At the same time, the tradeoff between the range of adap-

tation and the constant in front of the rate is important. Therefore, if prior information about the maximal possible smoothness is available to the user, it should be incorporated to further optimise the choice of α.

Appendix A

A.1. Proof appendix for the main result

Proof of Proposition 3.1. If $B^2_{\tau_\alpha,\alpha}(\mu) > B^2_{t^*_\alpha,\alpha}(\mu)$, then we have that $\tau_\alpha \le \lfloor t^*_\alpha \rfloor$. This yields

$$\mathbb{E}\big(B^2_{\tau_\alpha,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu)\big)^+ = \sum_{m=0}^{\lfloor t^*_\alpha\rfloor - 1} \lambda_{m+1}^{2+2\alpha}\mu_{m+1}^2 \mathbb{P}\{\tau_\alpha \le m\} + (t^*_\alpha - \lfloor t^*_\alpha\rfloor)\lambda^{2+2\alpha}_{\lceil t^*_\alpha\rceil}\mu^2_{\lceil t^*_\alpha\rceil}\mathbb{P}\{\tau_\alpha \le \lfloor t^*_\alpha\rfloor\}. \tag{A.1}$$

For a fixed $m \le \lfloor t^*_\alpha\rfloor$, we consider the event $\{\tau_\alpha \le m\} = \{R^2_{m,\alpha} \le \kappa\}$. The probability of this event can be bounded by

$$\mathbb{P}\{R^2_{m,\alpha} \le \kappa\} = \mathbb{P}\Big\{\sum_{i=m+1}^{D} \lambda_i^{2\alpha}(\lambda_i^2\mu_i^2 + 2\lambda_i\mu_i\delta\varepsilon_i + \delta^2\varepsilon_i^2) \le \kappa\Big\} \tag{A.2}$$

$$= \mathbb{P}\Big\{\sum_{i=m+1}^{D} \lambda_i^{2\alpha}\big(2\lambda_i\mu_i\delta\varepsilon_i + \delta^2(\varepsilon_i^2-1)\big) \le -(\mathbb{E}R^2_{m,\alpha} - \kappa)\Big\} \tag{A.3}$$

$$\le \mathbb{P}\Big\{\sum_{i=m+1}^{D} \lambda_i^{2\alpha}\delta^2(\varepsilon_i^2-1) \le \frac{-(\mathbb{E}R^2_{m,\alpha} - \kappa)}{2}\Big\} \tag{A.4}$$

$$+ \mathbb{P}\Big\{\sum_{i=m+1}^{D} \lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i \le \frac{-(\mathbb{E}R^2_{m,\alpha} - \kappa)}{4}\Big\}$$

$$\le \exp\left(\frac{-(\mathbb{E}R^2_{m,\alpha} - \kappa)^2}{16 s_D^2\delta^4}\right) + \exp\left(\frac{-(\mathbb{E}R^2_{m,\lambda} - \kappa)^2}{32\delta^2\sum_{i=m+1}^{D}\lambda_i^{2+4\alpha}\mu_i^2}\right) \tag{A.5}$$

$$\le \exp\left(\frac{-\big(B^2_{m,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu)\big)^2}{16 s_D^2\delta^4}\right) + \exp\left(\frac{-\big(B^2_{m,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu)\big)^2}{32\delta^2 B^2_{m,\alpha}(\mu)}\right). \tag{A.6}$$

In order to obtain (A.5), we use Lemma 1 from Laurent and Massart [10] and the Gaussian tail bound $\mathbb{P}\{Z \le -t\} \le e^{-t^2/(2\sigma^2)}$, $t > 0$, for a random variable Z distributed according to $N(0,\sigma^2)$. Further, we use that for $m \le \lfloor t^*_\alpha\rfloor$,

$$\begin{aligned}\mathbb{E}R^2_{m,\alpha} - \kappa &= \mathbb{E}R^2_{m,\alpha} - \mathbb{E}R^2_{t^*_\alpha,\alpha} \\ &= B^2_{m,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu) + V_{t^*_\alpha,\alpha} - V_{m,\alpha} \\ &\ge B^2_{m,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu)\end{aligned} \tag{A.7}$$

to obtain (A.6).

We set

$$F(t) := \exp\Big(\frac{-t^2}{16 s_D^2 \delta^4}\Big) + \exp\Big(\frac{-t^2}{32\delta^2 (B^2_{t^*_\alpha,\alpha}(\mu) + t)}\Big), \qquad t \geq 0. \tag{A.8}$$

The monotonicity of $t \mapsto B^2_{t,\alpha}$ and F and a Riemann sum approximation yield

$$\begin{aligned}
&\mathbb{E}\big(B^2_{\tau_\alpha,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu)\big)^+ \qquad (A.9)\\
&\leq \sum_{m=0}^{\lfloor t^*_\alpha \rfloor - 1} \lambda_{m+1}^{2+2\alpha} \mu_{m+1}^2 F(B^2_{m,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu))\\
&+ (t^*_\alpha - \lfloor t^*_\alpha \rfloor) \lambda^{2+2\alpha}_{\lceil t^*_\alpha \rceil} \mu^2_{\lceil t^*_\alpha \rceil} F(B^2_{\lfloor t^*_\alpha \rfloor,\alpha}(\mu) - B^2_{t^*_\alpha,\alpha}(\mu))\\
&\leq \int_{B^2_{t^*_\alpha,\alpha}(\mu)}^{\infty} F(t - B^2_{t^*_\alpha,\alpha}(\mu))\, dt \leq \int_0^\infty F(t) t\, dt\\
&\leq \frac{1}{2}\sqrt{2\pi 8 s_D^2 \delta^4} + \int_0^{B^2_{t^*_\alpha,\alpha}(\mu)} \exp\Big(\frac{-t^2}{64\delta^2 B^2_{t^*_\alpha,\alpha}(\mu)}\Big)\, dt + \int_{B^2_{t^*_\alpha,\alpha}(\mu)}^\infty \exp\Big(\frac{-t}{64\delta^2}\Big)\, dt\\
&\leq \sqrt{4\pi} s_D \delta^2 + \frac{1}{2}\sqrt{2\pi \cdot 32 \delta^2 B^2_{t^*_\alpha,\alpha}(\mu)} + 64\delta^2 \leq 74 s_D \delta^2 + 2 B^2_{t^*_\alpha,\alpha}(\mu).
\end{aligned}$$

For the last inequality, we use the binomial identity to obtain $\sqrt{\pi\delta^2 B^2_{t^*_\alpha,\alpha}(\mu)} \leq (\pi\delta^2 + B^2_{t^*_\alpha,\alpha}(\mu))/2$ and the estimate $\sqrt{4\pi} + 2\pi \leq 10$. □

Proof of Proposition 3.3. The Cauchy-Schwarz inequality and $\mathbb{E}\varepsilon_m^4 = 3$ yield

$$\begin{aligned}
\mathbb{E}(S_{\tau_\alpha} - S_{\lceil t^*_\alpha \rceil})^+ &= \delta^2 \sum_{m=\lceil t^*_\alpha \rceil + 1}^{D} \lambda_m^{-2} \mathbb{E}(\varepsilon_m^2 \mathbf{1}\{\tau_\alpha \geq m\}) \qquad (A.10)\\
&\leq \delta^2 \sum_{m=\lceil t^*_\alpha \rceil + 1}^{D} \lambda_m^{-2} \sqrt{\mathbb{E}\varepsilon_m^4} \sqrt{\mathbb{P}\{\tau_\alpha \geq m\}}\\
&\leq \sqrt{3}\delta^2 \sum_{m=\lceil t^*_\alpha \rceil + 1}^{D} \lambda_m^{-2} \sqrt{\mathbb{P}\{\tau_\alpha \geq m\}}.
\end{aligned}$$

The smoothed residual stopping time satisfies $\tau_\alpha \geq m$ exactly when $R^2_{m-1,\alpha} > \kappa$. For $m \geq \lceil t^*_\alpha \rceil + 1$, the probability above can therefore be estimated by

$$\begin{aligned}
\mathbb{P}\{R^2_{m-1,\alpha} > \kappa\} &= \mathbb{P}\Big\{ \sum_{i=m}^D \lambda_i^{2\alpha} (\lambda_i \mu_i + \delta\varepsilon_i)^2 > \kappa \Big\} \qquad (A.11)\\
&= \mathbb{P}\Big\{ \sum_{i=m}^D \lambda_i^{2+2\alpha} \mu_i^2 + 2\lambda_i^{1+2\alpha} \mu_i \delta\varepsilon_i + \lambda_i^{2\alpha} \delta^2 \varepsilon_i^2 > \kappa \Big\} \qquad (A.12)\\
&\leq \mathbb{P}\Big\{ \sum_{i=m}^D \lambda_i^{2\alpha} \delta^2 (\varepsilon_i^2 - 1) > \frac{\kappa - \mathbb{E}R^2_{m-1,\alpha}}{2} \Big\} \qquad (A.13)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\Big\{\sum_{i=m}^{D}\lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i > \frac{\kappa-\mathbb{E}R_{m-1,\alpha}^2}{4}\Big\}\\
&\le \exp\left(\frac{-(\kappa-\mathbb{E}R_{m-1,\alpha}^2)^2}{16\sum_{i=m}^{D}\lambda_i^{4\alpha}\delta^4+8\delta^2\lambda_m^{2\alpha}(\kappa-\mathbb{E}R_{m-1,\alpha}^2)}\right) \qquad \text{(A.14)}\\
&+\exp\left(\frac{-(\kappa-\mathbb{E}R_{m-1,\alpha}^2)^2}{32\delta^2\sum_{i=m}^{D}\lambda_i^{2+4\alpha}\mu_i^2}\right).
\end{aligned}
$$

The last inequality follows again from Lemma 1 in [10] and the Gaussian tail bound $\mathbb{P}\{Z \le -t\} \le e^{-t^2/(2\sigma^2)}$, $t > 0$, for a random variable Z distributed according to $N(0,\sigma^2)$.

Since $\alpha p < 1/2$, we have the following essential lower bound for the numerator in the exponential terms in (A.14):

$$
\begin{aligned}
\kappa-\mathbb{E}R_{m-1,\alpha}^2 &\ge \mathbb{E}R_{t_\alpha^*,\alpha}^2-\mathbb{E}R_{m-1,\alpha}^2 \qquad \text{(A.15)}\\
&= B_{t_\alpha^*,\alpha}^2(\mu)-V_{t_\alpha^*,\alpha}+V_{m-1,\alpha}-B_{m-1,\alpha}^2(\mu)\\
&\gtrsim_A \sum_{i=\lceil t_\alpha^*\rceil+1}^{m-1} i^{-2\alpha p}\delta^2 \ge \delta^2\int_{\lceil t_\alpha^*\rceil+1}^{m} t^{-2\alpha p}\,dt\\
&\ge \frac{\delta^2}{1-2\alpha p}\big(m^{1-2\alpha p}-\lceil t_\alpha^*\rceil^{1-2\alpha p}\big).
\end{aligned}
$$

For the denominators, we use the upper bounds

$$
\sum_{i=m}^{D}\lambda_i^{4\alpha} \le s_D^2, \qquad \text{(A.16)}
$$

$$
\begin{aligned}
\lambda_m^{2\alpha}(\kappa-\mathbb{E}R_{m-1,\alpha}^2) &\le \lambda_m^{2\alpha}\Big(\sum_{i=1}^{D}\lambda_i^{2\alpha}\delta^2+C_\kappa s_D\delta^2-\sum_{i=m}^{D}\lambda_i^{2\alpha}\delta^2\Big) \qquad \text{(A.17)}\\
&\le (1+C_\kappa)s_D^2\delta^2,
\end{aligned}
$$

and

$$
\begin{aligned}
\sum_{i=m}^{D}\lambda_i^{2+4\alpha}\mu_i^2 &\le \lambda_m^{2\alpha}B_{m-1,\alpha}^2(\mu) \qquad \text{(A.18)}\\
&\le \lambda_m^{2\alpha}\Big(V_{m,\alpha}+\kappa-\sum_{i=1}^{D}\lambda_i^{2\alpha}\delta^2\Big)\\
&\le (1+C_\kappa)s_D^2\delta^2,
\end{aligned}
$$

for $m \ge \lceil t_\alpha^*\rceil + 1$.

Together, this yields

$$
\mathbb{E}(S_{\tau_\alpha}-S_{\lceil t_\alpha^*\rceil})^+ \lesssim \delta^2\sum_{m=\lceil t_\alpha^*\rceil}^{D} m^{2p}\exp\left(\frac{-(m^{1-2\alpha p}-\lceil t_\alpha^*\rceil^{1-2\alpha p})^2}{C_{A,\kappa}(1-2\alpha p)^2 s_D^2}\right) \qquad \text{(A.19)}
$$

for a constant $C_{A,\kappa} > 0$ depending on p, C_A and κ. By a Riemann sum approximation, the sum in (A.19) can essentially be estimated from above by

$$\begin{aligned}
&\int_{\lceil t_\alpha^* \rceil}^{\infty} t^{2p} \exp\left(\frac{-(t^{1-2\alpha p} - \lceil t_\alpha^* \rceil^{1-2\alpha p})^2}{C_{A,\kappa}(1-2\alpha p)^2 s_D^2}\right) dt \qquad \text{(A.20)}\\
&\sim_{\alpha,A} \int_{\lceil t_\alpha^* \rceil^{1-2\alpha p}}^{\infty} u^{\frac{2p+2\alpha p}{1-2\alpha p}} \exp\left(\frac{-(u - \lceil t_\alpha^* \rceil^{1-2\alpha p})^2}{C_{A,\kappa}(1-2\alpha p)^2 s_D^2}\right) du\\
&\lesssim_{\alpha,A} \int_0^{\infty} (u + \lceil t_\alpha^* \rceil^{1-2\alpha p})^{\frac{2p+2\alpha p}{1-2\alpha p}} \exp\left(\frac{-u^2}{C_{A,\kappa}(1-2\alpha p)^2 s_D^2}\right) du\\
&\lesssim_{\alpha,A,\kappa} \lceil t_\alpha^* \rceil^{2p+2\alpha p} s_D + s_D^{(2p+1)/(1-2\alpha p)}.\\
&\lesssim_{\alpha,A,\kappa} (t_\alpha^*)^{2p+1} + s_D^{(2p+1)/(1-2\alpha p)}.
\end{aligned}$$

Noting that $\mathbb{E}(S_{\lceil t_\alpha^* \rceil} - S_{t_\alpha^*})^+ \lesssim_A (t_\alpha^*)^{2p}\delta^2$ and $V_t \sim_A t^{2p+1}\delta^2$ yields the result. □

A.2. Proof appendix for supplementary results

Proof of Proposition 4.1. For $\mu = 0$ and a fixed i_0, we have $\tau_\alpha \geq i_0$ if and only if

$$R_{i_0-1,\alpha}^2 = \sum_{i=i_0}^{D} \lambda_i^{2\alpha}(\lambda_i \cdot 0 + \delta\varepsilon_i)^2 = \sum_{i=i_0}^{D} \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 > \kappa. \qquad \text{(A.21)}$$

This condition can be reformulated to

$$\sum_{i=i_0}^{D} \lambda_i^{2\alpha}(\varepsilon_i^2 - 1) - \sum_{i=1}^{i_0-1} \lambda_i^{2\alpha} > \delta^{-2}\kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}. \qquad \text{(A.22)}$$

Assumption (2.2) and the fact that $\sum_{i=1}^{i_0-1} \lambda_i^{2\alpha} \lesssim_{\alpha,A} i_0^{1-2\alpha p}$ imply that there exists a constant $C_{\alpha,A,\kappa} > 0$ depending only on α, p, C_A and C_κ such that for $i_0 \sim s_D^{1/(1-2\alpha p)}$,

$$s_D^{-1} \sum_{i=i_0}^{D} \lambda_i^{2\alpha}(\varepsilon_i^2 - 1) > C_{\alpha,A,\kappa} \qquad \text{(A.23)}$$

is sufficient for (A.22). Since $\alpha p \leq 1/4$, the left-hand side normalises: We have

$$\begin{aligned}
\tilde{s}_D^2 &:= \operatorname{Var} \sum_{i=i_0}^{D} \lambda_i^{2\alpha}(\varepsilon_i^2 - 1) \qquad \text{(A.24)}\\
&= 2\sum_{i=i_0}^{D} \lambda_i^{4\alpha} \gtrsim_A \begin{cases} D^{1-4\alpha p} - i_0^{1-4\alpha p}, & \alpha p < 1/4, \\ \log D - \log i_0, & \alpha p = 1/4, \end{cases}
\end{aligned}$$

which implies that $\tilde{s}_D^2 \to \infty$ for $\delta \to 0$, since $i_0 = o(D)$. This yields that the sum in (A.23) satisfies Lindeberg's condition. By Slutzky's Lemma, the left-hand side

in (A.23) then converges in distribution to a centred Gaussian random variable Z and

$$\mathbb{P}\Big\{ s_D^{-1} \sum_{i=i_0+2}^{D} \lambda_i^{2\alpha}(\varepsilon_i^2 - 1) > C_{\alpha,A,\kappa} \Big\} \xrightarrow{\delta\to 0} \mathbb{P}\{Z > C_{\alpha,A,\kappa}\} > 0. \tag{A.25}$$

This implies that $\mathbb{P}\{\tau_\alpha \geq i_0\} \geq C$ for some constant $C > 0$ and δ sufficiently small. Together with (4.2), this gives

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - 0\|^2 \geq \mathbb{P}\{\tau_\alpha \geq i_0\} V_{i_0} \geq C s_D^{(2p+1)/(1-2\alpha p)} \delta^2, \tag{A.26}$$

since $V_{i_0} \sim_A i_0^{2p+1}\delta^2$.

Finally, we note that for $\mu \neq 0$,

$$\begin{aligned}
\mathbb{P}\{R_{i_0,\alpha}^2 \geq \kappa\} &= \mathbb{P}\Big\{ \sum_{i=i_0}^{D} \lambda_i^{2+2\alpha}\mu_i^2 + 2\lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i + \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 > \kappa \Big\} \\
&\geq \mathbb{P}\Big\{ \sum_{i=i_0}^{D} 2\lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i + \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 > \kappa \Big\} \\
&\geq \mathbb{P}\Big\{ \sum_{i=i_0}^{D} 2\lambda_i^{1+2\alpha}\mu_i\delta\varepsilon_i \geq 0, \sum_{i=i_0}^{D} \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 > \kappa \Big\} \\
&= \frac{1}{2}\mathbb{P}\Big\{ \sum_{i=i_0}^{D} \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 \geq \kappa \Big\},
\end{aligned} \tag{A.27}$$

which shows that (A.26) also holds for $\mu \neq 0$. □

Proof of Proposition 4.2. We consider a signal $\mu = \mu(\delta) \in H^\beta(r, D)$ with only one nonzero coefficient at position $i_0 + 1$ given by

$$\mu_{i_0+1}^2 := \lambda_{i_0+1}^{-(2+2\alpha)} \sum_{i=2}^{i_0} \lambda_i^{2\alpha}\delta^2 \quad \text{and} \quad \mu_i := 0 \quad \text{for all } i \neq i_0 + 1. \tag{A.28}$$

Note that the coefficient μ_{i_0+1} is chosen in a way that the α-balanced oracle $t_\alpha^\flat$ is slightly smaller than i_0 but of the same order. Under the assumption on κ, a sufficient condition for the stopping criterion $R_{i_0,\alpha}^2 \leq \kappa$ is given by

$$\sum_{i=2}^{i_0} \lambda_i^{2\alpha}\delta^2 + 2\lambda_{i_0+1}^{1+2\alpha}\mu_{i_0+1}\delta\varepsilon_{i_0+1} + \sum_{i=i_0+1}^{D} \lambda_i^{2\alpha}\delta^2\varepsilon_i^2 \leq \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2. \tag{A.29}$$

We consider the different regimes of αp:

(a) If $\alpha p \leq 1/4$, then we consider the condition

$$\varepsilon_{i_0+1} \in [-1, 0] \qquad \text{and} \qquad \sum_{i=i_0+2}^{D} \lambda_i^{2\alpha}(\varepsilon_i^2 - 1) \leq 0, \tag{A.30}$$

which is sufficient for (A.29). Due to the independence of the $(\varepsilon_i)_{i\le D}$, we only have to control the second part of the event defined by (A.30). If we choose $i_0 = i_0(\delta) \lesssim t^{mm}_{\beta,p,r}(\delta)$, then the standardisation of this term normalises in the same way as in the proof of Proposition 4.1 due to the growth condition on D. We have

$$\lambda_{i_0+1}^{-(2+2\alpha)} \sum_{i=2}^{i_0} \lambda_i^{2\alpha} \sim_A \frac{1}{1-2\alpha p}(i_0+1)^{2p+1} \tag{A.31}$$

for i_0 sufficiently large. Therefore, we can choose

$$i_0 \sim ((1-2\alpha p)\delta^{-2}r^2)^{1/(2\beta+2p+1)} \tag{A.32}$$

when δ is sufficiently small while still maintaining $\mu \in H^\beta(r,D)$. This yields

$$\begin{aligned}\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 &\gtrsim_A (1-2\alpha p)^{-1} i_0^{2p+1}\delta^2 \\ &\gtrsim_A r^2(r^{-2}\delta^2/(1-2\alpha p))^{2\beta/(2\beta+2p+1)}.\end{aligned} \tag{A.33}$$

(b) If $1/4 < \alpha p < 1/2$, then we rearrange (A.29) to

$$2\lambda_{i_0+1}^{\alpha}\sqrt{\sum_{i=2}^{i_0}\lambda_i^{2\alpha}}\varepsilon_{i_0+1} + \sum_{i=i_0+1}^{D}\lambda_i^{2\alpha}(\varepsilon_i^2-1) \le \lambda_1^{2\alpha}. \tag{A.34}$$

If we choose $i_0 = i_0(\delta) \lesssim t^{mm}_{\beta,p,r}(\delta)$, both terms on the left-hand side of (A.34) converge to zero in probability, since their variances are multiples of

$$\lambda_{i_0+1}^{2\alpha}\sum_{i=2}^{i_0}\lambda_i^{2\alpha} \lesssim_A (i_0+1)^{1-4\alpha p} \quad \text{and} \quad \sum_{i=i_0+1}^{D}\lambda_i^{4\alpha} \lesssim_A \sum_{i=i_0+1}^{D} i^{-4\alpha p}, \tag{A.35}$$

which both vanish for $\delta \to 0$. Since $\lambda_1^{2\alpha} > 0$, this yields $\mathbb{P}\{\tau_\alpha \le i_0\} \to 1$ for $\delta \to 0$, which gives the same result as in (a).

(c) If $\alpha p \ge 1/2$, the same reasoning as in (b) allows to bound the probability $\mathbb{P}\{\tau_\alpha \le i_0\}$ from below for $\delta \to 0$. Since

$$\lambda_{i_0+1}^{-(2+2\alpha)}\sum_{i=2}^{i_0}\lambda_i^{2\alpha} \sim_A \begin{cases} i_0^{2p+1}\log(i_0), & \alpha p = 1/2, \\ i_0^{2p+2\alpha p}, & \alpha p > 1/2, \end{cases} \tag{A.36}$$

we can choose i_0 of order $t^{mm}_{\beta,p,r,\alpha}(\delta)$ while still maintaining $\mu \in H^\beta(r,D)$. This yields the bound

$$\mathbb{E}\|\widehat{\mu}^{(\tau_\alpha)} - \mu\|^2 \gtrsim_A \begin{cases} (t^{mm}_{\beta,p,r,\alpha})^{2p+1}\log(t^{mm}_{\beta,p,r,\alpha})\delta^2, & \alpha p = 1/2, \\ r^2(r^{-2}\delta^2)^{2\beta/(2\beta+2p+2\alpha p)}, & \alpha p > 1/2. \end{cases} \tag{A.37}$$

This finishes the result. □

Proposition A.1 (Control of the stochastic error for $\alpha p \geq 1/2$)**.** *Assume (PSD(p, C_A)) with $\alpha p \geq 1/2$, $\kappa \geq \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2$ and (2.2). Then, we have the following control over the stochastic error:*

(i) For any $\mu \in H^\beta(r, D)$ which is the D-dimensional projection of a signal satisfying the polished tail condition (3.7), there exists a constant $C_{A,\kappa} > 0$ depending on p, C_A and C_κ such that

$$\mathbb{E}(S_{\tau_\alpha} - S_{t_\alpha^*})^+ \leq C_{A,\kappa}(t_\alpha^*)^{2p+1}\delta^2.$$

(ii) For any $\mu \in H^\beta(r, D)$, there exists a constant $C_{A,\kappa} > 0$ depending on p, C_A and C_κ such that

$$\mathbb{E}(S_{\tau_\alpha} - S_{t^{mm}_{\beta,p,r,\alpha}})^+ \leq C_{A,\kappa}(t^{mm}_{\beta,p,r,\alpha})^{2p+1}\delta^2.$$

Proof. We proceed as in the proof of Proposition 3.3 up to the inequality in (A.14). We split the two exponential terms in three and estimate from above with

$$\exp\left(\frac{-(\kappa - \mathbb{E}R^2_{m-1,\alpha})^2}{32\sum_{i=m}^{D}\lambda_i^{4\alpha}\delta^4}\right) + \exp\left(\frac{-(\kappa - \mathbb{E}R^2_{m-1,\alpha})}{16\delta^2\lambda_m^{2\alpha}}\right) + \exp\left(\frac{-(\kappa - \mathbb{E}R^2_{m-1,\alpha})^2}{32\delta^2\sum_{i=m}^{D}\lambda_i^{2+4\alpha}\mu_i^2}\right). \tag{A.38}$$

For (i), we have

$$\begin{aligned} B^2_{\lceil t_\alpha^*\rceil,\alpha}(\mu) &= \sum_{i=\lceil t_\alpha^*\rceil+1}^{D} \lambda_i^{2+2\alpha}\mu_i^2 \\ &\lesssim_A \lceil t_\alpha^*\rceil^{-(2p+2\alpha p)} \sum_{i=\lceil t_\alpha^*\rceil+1}^{\rho(\lceil t_\alpha^*\rceil+1)} \mu_i^2 \\ &\lesssim_A \sum_{i=\lceil t_\alpha^*\rceil+1}^{\rho(\lceil t_\alpha^*\rceil+1)} \lambda_i^{2+2\alpha}\mu_i^2. \end{aligned} \tag{A.39}$$

Choosing $m - 1 \geq \rho(\lceil t_\alpha^*\rceil + 1)$, we obtain that

$$B^2_{m-1,\alpha}(\mu) = \sum_{i=m}^{D} \lambda_i^{2+2\alpha}\mu_i^2 \leq c_A B^2_{\lceil t_\alpha^*\rceil,\alpha}(\mu) \tag{A.40}$$

for a constant $c_A < 1$ depending on p, C_A.

For (ii), choosing $m - 1 \geq C'_A t^{mm}_{\beta,p,r,\alpha}$ for a constant $C'_A > 1$ depending on p and C_A yields

$$B^2_{m-1,\alpha}(\mu) \lesssim_A r^2(C'_A t^{mm}_{\beta,p,r,\alpha})^{-(2\beta+2p+2\alpha p)}. \tag{A.41}$$

Setting $\bar{t} := t^*_\alpha$ for (i) or $\bar{t} := t^{mm}_{\beta,p,r,\alpha}$ for (ii), we can therefore choose a constant $C'_A > 1$ such that for $m - 1 \geq C'_A \lceil \bar{t} \rceil$,

$$\begin{aligned}
\kappa - \mathbb{E}R^2_{m-1,\alpha} &= \kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{m-1,\alpha} - B_{m-1,\alpha}(\mu) \qquad \text{(A.42)}\\
&\geq \begin{cases} \kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{m-1,\alpha} - c_A B_{t^*_\alpha,\alpha}(\mu), & \bar{t} = t^*_\alpha, \\ \kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{m-1,\alpha} - B^2_{m-1,\alpha}(\mu) & \bar{t} = t^{mm}_{\beta,p,r,\alpha}, \end{cases}\\
&\geq \begin{cases} (1-c_A)\Big(\kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{t^*_\alpha,\alpha}\Big), & \bar{t} = t^*_\alpha, \\ \kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{m-1,\alpha} - B^2_{m-1,\alpha}(\mu) & \bar{t} = t^{mm}_{\beta,p,r,\alpha}, \end{cases}\\
&\gtrsim_A \delta^2,
\end{aligned}$$

where we have used (A.41) and the definition of $t^{mm}_{\beta,p,r,\alpha}$ from (2.34) for the last inequality. Additionally, we have the estimates

$$\sum_{i=m}^{D} \lambda_i^{4\alpha} \lesssim_A \lambda_m^{\alpha} \qquad \text{(A.43)}$$

and

$$\begin{aligned}
\sum_{i=m}^{D} \lambda_i^{2+4\alpha}\mu_i^2 &\leq \lambda_m^{2\alpha} B^2_{m-1,\alpha}(\mu) \qquad \text{(A.44)}\\
&\leq \lambda_m^{2\alpha}\Big(\kappa - \sum_{i=1}^{D} \lambda_i^{2\alpha}\delta^2 + V_{m-1,\alpha}\Big)\\
&\lesssim_{A,\kappa} \lambda_m^{2\alpha} \log(m)\delta^2,
\end{aligned}$$

where we have used Equation (2.21), assumption (2.2) and that without loss of generality, $m \geq t^*_\alpha$. Note that the log factor occurs only for $\alpha p = 1/2$. We therefore obtain that for a constant $C''_{A,\kappa} > 0$ depending on p, C_A and C_κ,

$$\begin{aligned}
\mathbb{E}(S_{\tau_\alpha} - S_{\lceil \bar{t} \rceil})^+ &\lesssim_A \delta^2 \sum_{m=\lceil \bar{t} \rceil}^{\lceil C'_A \bar{t} \rceil} m^{2p} + \delta^2 \sum_{m=\lceil C'_A \bar{t} \rceil + 1}^{D} m^{2p} \exp(-m^{\alpha p}/(C''_{A,\kappa} \log m))\\
&\lesssim_{A,\kappa} \bar{t}^{2p+1}\delta^2. \qquad \text{(A.45)}
\end{aligned}$$

Noting that $\mathbb{E}(S_{\lceil \bar{t} \rceil} - S_{\bar{t}})^+ \lesssim_A \bar{t}^{2p}\delta^2$ finishes the proof. □

References

[1] G. Blanchard and P. Mathé. Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems*, 28(11):115011/1–115011/23, 2012. MR2992966

[2] G. Blanchard, M. Hoffmann, and M. Reiß. Early stopping for statistical inverse problems via truncated SVD estimation. *Electronic Journal of Statistics*, 12(2):3204–3231, 2018a. MR3859376

[3] G. Blanchard, M. Hoffmann, and M. Reiß. Optimal adaptation for early stopping in statistical inverse problems. *SIAM/ASA Journal of Uncertainty Quantification*, 6(3):1043–1075, 2018b. MR3829522

[4] P. Bühlmann and B. Yu. Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003. MR1995709

[5] A. Caponetto, L. Rosasco, and Y. Yao. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007. MR2327601

[6] H. Engl, M. Hanke, and A. Neubauer. *Regularisation of Inverse Problems*, volume 375 of *Mathematics and Its Applications*. Kluwer Academic Publishers, Dordrecht, 1996. MR1408680

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL http://www.deeplearningbook.org. MR3617773

[8] I. Johnstone. Gaussian estimation: Sequence and wavelet models, draft of a monograph, 2017. URL https://statweb.stanford.edu/~imj/GE_08_09_17.pdf.

[9] S. Kindermann and A. Neubauer. On the convergence of the quasioptimality criterion for (iterated) Tikhonov regularization. *Inverse Problems & Imaging*, 2(2):291–299, 2008. MR2395145

[10] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. MR1805785

[11] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and nonparametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014. MR3190843

[12] L. Rosasco, E. De Vito, A. Caponetto, U. De Giovannini, and F. Odone. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6:883–904, 2005. MR2249842

[13] B. Szabó, A. van der Vaart, and J. Zanten. Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics*, 43(4):1391–1428, 2015. MR3357861

[14] F. Yang, Y. Wei, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 2019. MR4009223