Early stopping for statistical inverse problems via truncated SVD estimation

Gilles Blanchard

Institut für Mathematik
Universität Potsdam, Germany
e-mail: gilles.blanchard@math.uni-potsdam.de

Marc Hoffmann

CEREMADE
Université Paris-Dauphine & PSL, France
e-mail: hoffmann@ceremade.dauphine.fr

Markus Reiß

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

Abstract: We consider truncated SVD (or spectral cut-off, projection) estimators for a prototypical statistical inverse problem in dimension D. Since calculating the singular value decomposition (SVD) only for the largest singular values is much less costly than the full SVD, our aim is to select a data-driven truncation level $\widehat{m} \in \{1, \ldots, D\}$ only based on the knowledge of the first $\widehat{m}$ singular values and vectors.

We analyse in detail whether sequential *early stopping* rules of this type can preserve statistical optimality. Information-constrained lower bounds and matching upper bounds for a residual based stopping rule are provided, which give a clear picture in which situation optimal sequential adaptation is feasible. Finally, a hybrid two-step approach is proposed which allows for classical oracle inequalities while considerably reducing numerical complexity.

1. Introduction and overview of results

1.1. Model

A classical model for statistical inverse problems is the observation of

$$Y = A\mu + \delta\dot{W} \tag{1.1}$$

where $A : H_1 \to H_2$ is a linear, bounded operator between real Hilbert spaces H_1, H_2, $\mu \in H_1$ is the signal of interest, $\delta > 0$ is the noise level and $\dot{W}$ is a Gaussian white noise in H_2, see *e.g.* Bissantz *et al.* [1], Cavalier [5] and the references therein. In any concrete situation the problem is discretised, for instance by using a Galerkin scheme projecting on finite element or other approximation spaces. Therefore we can assume $H_1 = \mathbb{R}^D$, $H_2 = \mathbb{R}^P$ with possibly very large D and P. Since the discretisation of μ is at our choice, we assume $D \leq P$, and that $A : \mathbb{R}^D \to \mathbb{R}^P$ is one-to-one. We transform (1.1) by the singular value decomposition (SVD) of A into the Gaussian vector observation model

$$Y_i = \lambda_i \mu_i + \delta \varepsilon_i, \quad i = 1, \ldots, D, \tag{1.2}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D > 0$ are the nonzero singular values of A, $(\mu_i)_{1 \leq i \leq D}$ the coefficients of μ in the orthonormal basis of singular vectors and $(\varepsilon_i)_{1 \leq i \leq D}$ are independent standard Gaussian random variables. The results will easily extend to subgaussian errors as discussed below.

Working in the SVD representation (1.2), the objective is to recover the signal $\mu = (\mu_i)_{1 \leq i \leq D}$ with best possible accuracy from the data $(Y_i)_{1 \leq i \leq D}$. A classical method is to use the truncated SVD estimators (also called projection or spectral cut-off estimators) $\widehat{\mu}^{(m)}$, $0 \leq m \leq D$, given by

$$\widehat{\mu}_i^{(m)} = \mathbf{1}(i \leq m) \lambda_i^{-1} Y_i, \qquad i = 1, \ldots, D, \tag{1.3}$$

which are ordered with decreasing bias and increasing variance (w.r.t. m). Choosing a suitable truncation index $\widehat{m} = \widehat{m}(Y)$ from the observed data is the genuine problem of adaptive model selection. Typical methods use (generalized) cross validation, see *e.g.* Wahba [19], unbiased risk estimation, see *e.g.* Cavalier *et al.* [6], penalized empirical risk minimisation, see *e.g.* Cavalier and Golubev [7], or Lepski's balancing principle for inverse problems, see *e.g.* Mathé and Pereverzev [15]. They all share the drawback that the estimators $\widehat{\mu}^{(m)}$ have first to be computed for all values of $0 \leq m \leq D$, and then be compared to each other in some way.

In this work, we are motivated by constraints due to the possible obstructive computational complexity of calculating the full SVD in high dimensions. We stress that the initial discretisation of the observation is generally based on a fixed scheme which does not deliver a representation of the observation vector Y in an SVD basis; nor is the full SVD basis of the discretized operator A a *priori* available in general, so that it has to be computed on the fly. Since the calculation of the largest singular value and its corresponding subspace is much less costly, efficient numerical algorithms rely on *deflation* or *locking* methods, which achieve the desired accuracy for the larger singular values first and then iteratively achieve the accuracy also for the next smaller singular values. As a basic example the popular power method can be considered, which usually finds after a few vector-matrix multiplications the top eigenvalue-eigenvector pair (with exponentially small error in the iteration number), so that by iterative application the largest m singular values and vectors are computed with roughly $\mathcal{O}(mD^2)$ multiplications compared to $\mathcal{O}(D^3)$ multiplications for a full SVD in a

worst case scenario. We refer to the monograph by Saad [17] for a comprehensive exposition of the numerical methods.

We investigate the possibility of an approach which is both statistically efficient and sequential along the SVD in the following sense: we aim at *early stopping* methods, in which the truncated SVD estimators $\widehat{\mu}^{(m)}$ for $m = 0, 1, \ldots,$ are computed iteratively, a stopping rule decides to stop at some step $\widehat{m}$ and then $\widehat{\mu}^{(\widehat{m})}$ is used as the estimator.

More generally, we envision our setting as a simple and prototypical model to study the scope of statistical adaptivity using iterative methods, which are widely used in computational statistics and learning. A notable feature of these methods is that not only the numerical, but also the statistical complexity (*e.g.*, measured by the variance) increases with the number of iterations, so that early stopping is essential from both points of view. It is common to use stopping rules based on monitoring the residuals because the user has access without substantial additional cost to the residual norm. Observe that the computation of the residual norm $\|Y - A\widehat{\mu}^{(m)}\|^2 = \|Y\|^2 - \sum_{i=1}^{m} Y_i^2$ does not require the full SVD, but only the knowledge of the m first coefficient and of the full norm $\|Y\|^2$, which is readily available. The properties of such rules have been well studied for deterministic inverse problems (e.g. the discrepancy principle, see Engl *et al.* [10]). In a statistical setting, minimax optimal solutions along the iteration path have been identified in different settings, see *e.g.* Yao *et al.* [20] for gradient descent learning, Blanchard and Mathé [3] for conjugate gradients, Raskutti, Wainwright and Yu [16] for (reproducing) kernel learning and Bühlmann and Hothorn [4] for the application to L^2-boosting. All these methods stop at a fixed iteration step[1], depending on the prior knowledge of the smoothness of the unknown solution.

By contrast, our goal is to analyse an *a posteriori* early stopping rule based on monitoring the residual, which corresponds to proposals by practitioners in the absence of prior smoothness information. An analysis of such a stopping rule for quite general spectral estimators like Landweber iteration is provided in the companion paper [2]. Although the general results can also be adapted to the truncated SVD method, we exhibit a more transparent analysis for this prototypical method which gives more satisfactory results: we establish coherent lower bounds and we obtain adaptivity in strong norm via the oracle property, while for more general spectral estimators only rate results over Sobolev-type classes can be achieved. Moreover, a hybrid two-step procedure enjoys full adaptivity for the truncated SVD-method.

1.2. Non-asymptotic oracle approach

Our approach is a priori non-asymptotic and concentrates on oracle optimality analysis for individual signals. The oracle approach compares the error of $\widehat{\mu}^{(\widehat{m})}$

---

[1]The stopping rule in [16] for a random design setting is data-dependent because the distribution of the design is unknown. The stopping iteration becomes fixed when the design distribution is known, the target function belonging to the unit ball of the kernel space.

to the minimal error among $(\widehat{\mu}^{(m)})_m$ for any signal μ individually, which entails optimal adaptation in minimax settings, see e.g. Cavalier [5].

The risk (mean integrated squared error) for a fixed truncated SVD estimator $\widehat{\mu}^{(m)}$ obeys a standard squared bias-variance decomposition

$$\mathbb{E}\left[\|\widehat{\mu}^{(m)} - \mu\|^2\right] = B_m^2(\mu) + V_m \,,$$

where $\|\bullet\|$ denotes the Euclidean norm in $\mathbb{R}^D$, and

$$B_m^2(\mu) := \mathbb{E}\left[\|\mathbb{E}[\widehat{\mu}^{(m)}] - \mu\|^2\right] = \sum_{i=m+1}^D \mu_i^2, \qquad (1.4)$$

$$V_m := \mathbb{E}\left[\|\widehat{\mu}^{(m)} - \mathbb{E}[\widehat{\mu}^{(m)}]\|^2\right] = \delta^2 \sum_{i=1}^m \lambda_i^{-2}. \qquad (1.5)$$

In distinction with the weak norm quantities defined below, we call $B_m(\mu)$ *strong bias* of μ and V_m *strong variance*.

If we have access to the residual squared norm

$$R_m^2 := \|Y - A\widehat{\mu}^{(m)}\|^2 = \|Y\|^2 - \|A\widehat{\mu}^{(m)}\|^2 = \sum_{i=1}^D (Y_i - \lambda_i\widehat{\mu}_i^{(m)})^2 = \sum_{i=m+1}^D Y_i^2, \qquad (1.6)$$

then $R_m^2 - (D - m)\delta^2$ gives some bias information due to

$$\mathbb{E}[R_m^2 - (D - m)\delta^2] = B_{m,\lambda}^2(\mu), \quad \text{with} \quad B_{m,\lambda}^2(\mu) := \sum_{i=m+1}^D \lambda_i^2\mu_i^2.$$

We call $B_{m,\lambda}^2(\mu)$ the *weak bias* and similarly $V_{m,\lambda} = m\delta^2$ the *weak variance*. They correspond to measuring the error in the *weak norm* (or prediction norm) $\|v\|_\lambda^2 := \|Av\|^2 = \sum_{i=1}^D \lambda_i^2 v_i^2$, which usually (always if $\lambda_1 < 1$) is smaller than the *strong* Euclidean norm $\|\bullet\|$. The squared bias-variance decomposition for the weak risk then reads $\mathbb{E}\left[\|\widehat{\mu}^{(m)} - \mu\|_\lambda^2\right] = B_{m,\lambda}^2(\mu) + V_{m,\lambda}$. Our setting is thus a particular instance of the question raised by Lepski [13] whether adaptation in one loss (here: weak norm) leads to adaptation in another loss (here: strong norm). Our positive answer for truncated SVD or spectral cut-off estimation will also extend the results by Chernousova *et al.* [8].

Intrinsic to the sequential analysis is the fact that at truncation index m we cannot say anything about the way the bias decreases for larger indices: it may drop to zero at $m + 1$ or even stay constant until $D - 1$. Even if we knew the exact value of the bias until index m, we could not minimise the sum of squared bias and variance sequentially. Instead, we should wait until the squared bias is sufficiently small to equal (approximately) the variance. This leads to the notion of the *strongly balanced oracle*

$$m_{\mathfrak{s}} = m_{\mathfrak{s}}(\mu) := \min\{m \in \{0, \ldots, D\} \mid V_m \geq B_m^2(\mu)\}, \qquad (1.7)$$

whose risk is always upper bounded by twice the classical oracle risk, see (3.5) below.

1.3. Setting for asymptotic considerations

Risk estimates over classes of signals and asymptotics for vanishing noise level $\delta \to 0$ often help to reveal main features. This way, we can also provide lower

bounds for sequential estimation procedures and compare them directly to classical minimax convergence rates. In our setting, the magnitude of the discretisation dimension D plays a central role, so that it is sensible to assume in an asymptotic view that $D = D_\delta \to \infty$ as $\delta \to 0$. As classes of signals, we will consider the Sobolev-type ellipsoids

$$H^\beta(R, D) := \{\mu \in \mathbb{R}^D \mid \sum_{i=1}^D i^{2\beta} \mu_i^2 \le R^2\}, \quad \beta \ge 0, R > 0, \qquad (1.8)$$

and we shall use the following polynomial spectral decay assumption

$$C_A^{-1} i^{-p} \le \lambda_i \le C_A i^{-p}, \qquad 1 \le i \le D, \qquad\qquad (\mathbf{PSD}(p, C_A))$$

for $p \ge 0, C_A \ge 1$. The spectrum is allowed to change with D and δ, but p, C_A are considered as fixed constants. Under these assumptions, standard computations yield for $\mu \in H^\beta(R, D)$, $1 \le m \le D$:

$$B_m^2(\mu) \le R^2 m^{-2\beta}; \qquad V_m \le C_A^{-2} \delta^2 m^{2p+1}.$$

Balance between these squared bias and variance bounds is obtained for m of the order of the minimax truncation "time"

$$t_{\beta,p,R}(\delta) := (R^{-1}\delta)^{-2/(2\beta+2p+1)}, \qquad\qquad (1.9)$$

provided the condition $D \ge t_{\beta,p,R}(\delta)$ holds. This gives rise to the risk rate

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

which agrees with the optimal minimax rate in the standard Gaussian sequence model (i.e. $D = \infty$). On the other hand, for $D \lesssim t_{\beta,p,R}(\delta)$ the choice $m = D$ is optimal on $H^\beta(R, D)$ and the rate degenerates to $\mathcal{O}(D^{2p+1}\delta^2)$. This situation is indicative of an insufficient discretisation and will be excluded from the asymptotic considerations.

1.4. Overview of results

Our results consist of lower and upper bounds for sequentially adaptive stopping rules. The stopping rules permitted are most conveniently described in terms of stopping times with respect to an appropriate filtration. Introduce the *frequency filtration*

$$\mathcal{F}_m := \sigma(\widehat{\mu}^{(0)}, \dots, \widehat{\mu}^{(m)}) = \sigma(Y_1, \dots, Y_m), \qquad\qquad (1.10)$$

$\mathcal{F}_0$ being the trivial sigma-field. Stopping rules with respect to the filtration $\mathcal{F} = (\mathcal{F}_m)_{0 \le m \le D}$ must decide whether to halt and output $\widehat{\mu}^{(m)}$ based only on the information of the first m estimators. Statistical adaptation will turn out to be essentially impossible for such stopping rules (Section 2.1). If the residual (1.6) is available at no substantial computational cost, taking this information into account, we define the *residual filtration*

$$\mathcal{G}_m := \mathcal{F}_m \vee \sigma(R_0^2, \dots, R_m^2) = \mathcal{F}_m \vee \sigma(\|Y\|^2), \qquad\qquad (1.11)$$

which is the filtration $\mathcal{F}_m$ enlarged by the residuals up to index m.

Pushing some technical details aside, the main message conveyed by our lower bounds is that oracle statistical adaptation with respect to the residual filtration is *impossible* for signals μ such that the strongly balanced oracle $m_{\mathfrak{s}}(\mu)$ is $o(\sqrt{D})$. (Section 2.2). On the other hand, we establish in Section 3 that this statement is sharp, in the sense that the simple residual-based stopping rule

$$\tau = \min\left\{m \geq m_0 \mid R_m^2 \leq \kappa\right\}, \tag{1.12}$$

with a proper choice of κ and m_0 is statistically adaptive for signals μ such that $m_{\mathfrak{s}}(\mu) \gtrsim \sqrt{D}$. Let us stress that by minimax adaptive we always mean that the procedure attains the optimal rate even among all methods with access to the entire data, that is without information constraints.

Finally, in Section 4 we introduce a hybrid two-step approach consisting of the above stopping rule with $m_0 \sim \sqrt{D}\log D$, followed by a traditional (non-sequential) model selection procedure over $m \leq m_0$, in the case where $\tau = m_0$ (immediate stop hinting at an optimal index smaller than m_0). This procedure enjoys full oracle adaptivity at a computational cost of calculating on average the first $\mathcal{O}(\max(\sqrt{D}\log D, m_{\mathfrak{s}}(\mu)))$ singular values, to be compared to the full SVD with D singular values in non-sequential adaptation. Some numerical simulations illustrate the theoretical analysis. Technical proofs are gathered in an appendix.

2. Lower bounds

2.1. The frequency filtration

Let τ be an $\mathcal{F}$-stopping time, where $\mathcal{F}$ is the frequency filtration defined in (1.10) and let[2]

$$\mathcal{R}(\mu, \tau)^2 := \mathbb{E}_\mu[\|\widehat{\mu}^{(\tau)} - \mu\|^2].$$

By Wald's identity, we obtain the simple formula

$$\mathcal{R}(\mu, \tau)^2 = \mathbb{E}_\mu\left[\sum_{i=\tau+1}^{D}\mu_i^2 + \sum_{i=1}^{\tau}\lambda_i^{-2}\delta^2\varepsilon_i^2\right] = \mathbb{E}_\mu\left[B_\tau^2(\mu) + V_\tau\right], \tag{2.1}$$

with $B_m^2(\mu)$ and V_m from (1.4), (1.5). This implies in particular that an oracle stopping time, *i.e.*, an optimal $\mathcal{F}$-stopping time constructed using the knowledge of μ, coincides with the deterministic oracle $\operatorname{argmin}_m\left(B_m^2(\mu) + V_m\right)$ almost surely. The next proposition encapsulates the main argument for the lower bound and merely relies on a two-point analysis. It clarifies that if the stopping time τ yields a squared risk comparable to the optimally balanced risk for a given signal μ, then this signal can be changed arbitrarily to $\bar{\mu}$ after the index $\lfloor 3Cm_{\mathfrak{s}}(\mu)\rfloor$, while the risk for the rule τ always stays larger than the squared bias of that part – which can be made arbitrarily large by "hiding" signal in large-index coefficients.

---

[2]We emphasise in the notation the dependence on μ in the distribution of τ and the Y_i by adding the subscript μ when writing the expectation $\mathbb{E} = \mathbb{E}_\mu$ or probability $P = P_\mu$.

2.1 Proposition. *Let $\mu, \bar{\mu} \in \mathbb{R}^D$ with $\mu_i = \bar{\mu}_i$ for all $i \leq i_0$ and $i_0 \in \{1, \ldots, D-1\}$. Then any $\mathcal{F}$-stopping rule τ satisfies*

$$\mathcal{R}(\bar{\mu}, \tau)^2 \geq B_{i_0}^2(\bar{\mu})\Big(1 - \frac{\mathcal{R}(\mu, \tau)^2}{V_{i_0+1}}\Big).$$

Suppose $\mathcal{R}(\mu, \tau)^2 \leq C\mathcal{R}(\mu, m_{\mathfrak{s}})^2$ for the balanced oracle $m_{\mathfrak{s}}$ in (1.7) and some $C \geq 1$. Then for any $\bar{\mu} \in \mathbb{R}^D$ with $\bar{\mu}_i = \mu_i$ for $i \leq 3Cm_{\mathfrak{s}}$ we obtain

$$\mathcal{R}(\bar{\mu}, \tau)^2 \geq \tfrac{1}{3}B_{\lfloor 3Cm_{\mathfrak{s}}\rfloor}^2(\bar{\mu}).$$

Proof. We use the fact that $(Y_i)_{1 \leq i \leq i_0}$ has the same law under P_μ and $P_{\bar{\mu}}$ and so has $\mathbf{1}(\tau \leq i_0)$ by the stopping time property of τ. Moreover, thanks to the monotonicity of $m \mapsto V_m$ and $m \mapsto B_m^2(\bar{\mu})$, Markov's inequality and identity (2.1):

$$\begin{aligned}
\mathcal{R}(\bar{\mu}, \tau)^2 &\geq \mathbb{E}_{\bar{\mu}}[B_\tau^2(\bar{\mu})\mathbf{1}(\tau \leq i_0)] \\
&= \mathbb{E}_\mu[B_\tau^2(\bar{\mu})\mathbf{1}(\tau \leq i_0)] \\
&\geq B_{i_0}^2(\bar{\mu})P_\mu(\tau \leq i_0) \\
&\geq B_{i_0}^2(\bar{\mu})(1 - P_\mu(V_\tau \geq V_{i_0+1})) \\
&\geq B_{i_0}^2(\bar{\mu})\Big(1 - \frac{\mathbb{E}_\mu[V_\tau]}{V_{i_0+1}}\Big) \\
&\geq B_{i_0}^2(\bar{\mu})\Big(1 - \frac{\mathcal{R}(\mu, \tau)^2}{V_{i_0+1}}\Big).
\end{aligned}$$

The second assertion follows by inserting $i_0 = \lfloor 3Cm_{\mathfrak{s}}\rfloor$ and $\mathcal{R}(\mu, \tau)^2 \leq 2CV_{m_{\mathfrak{s}}}$ together with $V_{m_{\mathfrak{s}}}/V_{i_0+1} \leq m_{\mathfrak{s}}/(i_0+1)$ since the singular values λ_i are non-increasing. $\qquad\square$

In Appendix 5.1 we use this proposition to provide a result suitable for asymptotic interpretation (we use the notation from Section 1.3):

2.2 Corollary. *Assume (**PSD**(p, C_A)) and let τ be any $\mathcal{F}$-stopping rule. If there exists $\mu \in H^\beta(R, D)$ with $\mathcal{R}(\mu, \tau) \leq C_\mu \mathcal{R}_{\beta,p,R}^*(\delta)$, then for any $\alpha \in [0, \beta]$, $\bar{R} \geq 2R$, there exists $\bar{\mu} \in H^\alpha(\bar{R}, D)$ such that*

$$\mathcal{R}(\bar{\mu}, \tau) \geq c_1\bar{R}(R^{-1}\delta)^{2\alpha/(2\beta+2p+1)},$$

provided $D \geq c_2 t_{\beta,p,R}(\delta)$. The constants $c_1, c_2 > 0$ only depend on C_μ and C_A.

The conclusion for impossible rate-optimal adaptation is a direct consequence of Corollary 2.2: since for any $\alpha < \beta$ the rate $\delta^{2\alpha/(2\beta+2p+1)}$ is suboptimal, no $\mathcal{F}$-stopping rule can adapt over Sobolev classes with different regularities. Finally, the rate $\bar{R}(R^{-1}\delta)^{2\alpha/(2\beta+2p+1)}$ is attained by a deterministic stopping rule that stops at the oracle frequency for $H^\beta(R, D)$, so that the lower bound is in fact a sharp *no adaptation* result.

2.2. Residual filtration

We start with a key lemma, similar in spirit to the first step in the proof of Proposition 2.1, but valid for an arbitrary random τ. Here and in the sequel the numerical values are not optimised, but give rise to more transparent proofs and convey some intuition for the worst case order of magnitude. The proof is delayed until Appendix 5.2.

2.3 Lemma. *Let* $\tau = \tau\big((Y_i)_{1 \leq i \leq D}\big) \in \{0, \ldots, D\}$ *be an arbitrary (measurable) data-dependent index. Then for any* $m \in \{1, \ldots, D\}$ *the following implication holds true:*

$$V_m \geq 200 \, \mathcal{R}(\mu, \tau)^2 \Rightarrow P_\mu(\tau \geq m) \leq 0.9.$$

For $\mathcal{G}$-stopping rules, where $\mathcal{G}$ is the residual filtration defined in (1.11), we deduce the following lower bound, again based on a two-point argument:

2.4 Proposition. *Let* τ *be an arbitrary* $\mathcal{G}$-*stopping rule. Consider* $\mu \in \mathbb{R}^D$ *and* $i_0 \in \{1, \ldots, D\}$ *such that* $V_{i_0+1} \geq 200\mathcal{R}(\mu, \tau)^2$. *Then*

$$\mathcal{R}(\bar\mu, \tau)^2 \geq 0.05B_{i_0}^2(\bar\mu)$$

holds for any $\bar\mu \in \mathbb{R}^D$ *that satisfies*

(a) $\mu_i = \bar\mu_i$ *for all* $i \leq i_0$,
(b) the weak bias bound $|B_{i_0,\lambda}^2(\bar\mu) - B_{i_0,\lambda}^2(\mu)| \leq 0.05\frac{\sqrt{D-i_0}}{2}\delta^2$ *and*
(c) $B_{i_0,\lambda}(\mu) + B_{i_0,\lambda}(\bar\mu) \geq 5.25\delta$.

Suppose that $\mathcal{R}(\mu, \tau)^2 \leq C_\mu \mathcal{R}(\mu, m_{\mathfrak{s}})^2$ *holds with some* $C_\mu \geq 1$. *Then any* $i_0 \geq 400C_\mu m_{\mathfrak{s}}$ *will satisfy the initial requirement.*

Proof. First, we lower bound the risk of $\bar\mu$ by its bias on $\{\tau \leq i_0\}$ and then transfer to the law of τ under P_μ, using the total variation distance on $\mathcal{G}_{i_0}$:

$$\begin{aligned}
\mathcal{R}(\bar\mu, \tau)^2 &\geq \mathbb{E}_{\bar\mu}[B_\tau^2(\bar\mu)\mathbf{1}(\tau \leq i_0)] \\
&\geq B_{i_0}^2(\bar\mu)P_{\bar\mu}(\tau \leq i_0) \\
&\geq B_{i_0}^2(\bar\mu)\big(P_\mu(\tau \leq i_0) - \|P_\mu - P_{\bar\mu}\|_{TV(\mathcal{G}_{i_0})}\big).
\end{aligned}$$

By Lemma 2.3 we infer $P_\mu(\tau \leq i_0) \geq 0.1$. Denote $W_{i_0} = (Y_1, \ldots, Y_{i_0})$. Since the law of W_{i_0} is identical under P_μ and $P_{\bar\mu}$, and W_{i_0} is independent of $R_{i_0}^2$ for both measures, the total variation distance between P_μ and $P_{\bar\mu}$ on $\mathcal{G}_{i_0}$ equals the total variation distance between the respective laws of the scaled residual $\delta^{-2}R_{i_0}^2$. For $\vartheta \in \mathbb{R}^D$, let $\mathbb{P}_K^\vartheta$ be the non-central χ^2-law of $X_\vartheta = \sum_{k=1}^K (\vartheta_k + Z_k)^2$ with Z_k independent and standard Gaussian. With $K = D - i_0$, $\vartheta_k = \delta^{-1}\lambda_{i_0+k}\mu_{i_0+k}$, $\bar\vartheta_k = \delta^{-1}\lambda_{i_0+k}\bar\mu_{i_0+k}$, the total variation distance between the respective laws of the scaled residual $\delta^{-2}R_{i_0}^2$ exactly equals $\|\mathbb{P}_K^\vartheta - \mathbb{P}_K^{\bar\vartheta}\|_{TV}$. By Lemma 5.1 in the Appendix, taking account of $\|\vartheta\| = \delta^{-1}B_{i_0,\lambda}(\mu)$ and similarly for $\|\bar\vartheta\|$, we infer from (c) the simplified bound

$$\|P_\mu - P_{\bar\mu}\|_{TV(\mathcal{G}_{i_0})} \leq \frac{2|B_{i_0,\lambda}^2(\bar\mu) - B_{i_0,\lambda}^2(\mu)|}{\delta^2\sqrt{D - i_0}}.$$

Under our assumption on $\bar{\mu}$, this is at most 0.05, and the inequality follows. From $\mathcal{R}(\mu, \tau)^2 \leq 2C_\mu V_{m_\mathfrak{s}}$ and $V_{i_0+1}/V_{m_\mathfrak{s}} \geq (i_0 + 1)/m_\mathfrak{s}$, the last statement follows.   □

In comparison with the frequency filtration, the main new hypothesis is that at i_0 the weak bias of $\bar{\mu}$ is sufficiently close to that of μ, while the lower bound is still expressed in terms of the strong bias. This is natural since the bias only appears in weak form in the residuals, while the risk involves the strong bias. Condition (c) is just assumed to simplify the bound. To obtain valuable counterexamples, $\bar{\mu}$ is usually chosen at maximal weak bias distance of μ allowed by (b), so that (c) is always satisfied in the interesting cases where $\sqrt{D - i_0}$ is not small.

Considering the behaviour over Sobolev-type ellipsoids, we obtain in Appendix 5.4 a lower bound result comparable to Corollary 2.2 for the frequency filtration.

2.5 Corollary. *Assume* $(\mathbf{PSD}(p, C_A))$ *and let τ be any $\mathcal{G}$-stopping time. If there exists $\mu \in H^\beta(R, D)$ with $\mathcal{R}(\mu, \tau) \leq C_\mu \mathcal{R}^*_{\beta, p, R}(\delta)$, then for any $\alpha \in [0, \beta]$ and $\bar{R} \geq 2R$, there exists $\bar{\mu} \in H^\alpha(\bar{R}, D)$ such that*

$$\mathcal{R}(\bar{\mu}, \tau) \geq c_1 \bar{R} \min\left(\left(\bar{R}^{-1} \delta D^{1/4}\right)^{2\alpha/(2\alpha+2p)}, (R^{-1}\delta)^{2\alpha/(2\beta+2p+1)} \right),$$

provided $R^{-1}\delta \leq c_2$ and $D \geq c_3 t_{\alpha - \frac{1}{4}, p, \bar{R}}(\delta)$. The constants $c_1, c_3 > 0$, and $c_2 \in (0, 1]$ depend only on C_μ, C_A, α, p.

The form of the lower bound is transparent: as in the case of the frequency filtration, the sub-obtimal rate $\bar{R}(R^{-1}\delta)^{2\alpha/(2\beta+2p+1)}$ is the one attained by a deterministic rule that stops at the oracle frequency for $H^\beta(R, D)$, whereas $\bar{R}\left(\bar{R}^{-1}\delta D^{1/4}\right)^{2\alpha/(2\alpha+2p)}$ is the size of a signal that may be hidden in the noise of the residual, *i.e.*, is not detected with positive probability by any test, thus also leading to erroneous early stopping. Note that for the direct problem $(p = 0)$, the latter quantity is just $\delta D^{1/4}$, which is exactly the critical signal strength in nonparametric testing, see Ingster and Suslina [11], while for $p > 0$, it reflects the interplay between the weak bias part in the residual and the strong bias part in the risk within the Sobolev ellipsoid.

Corollary 2.5 implies in turn explicit constraints for the maximal Sobolev regularity to which a $\mathcal{G}$-stopping rule can possibly adapt. Here, we argue asymptotically and let explicitly $D = D_\delta$ tend to infinity as the noise level δ tends to zero. In this setting, a stopping rule τ is to be understood as a family of stopping rules that depend on the knowledge of D and δ.

2.6 Corollary. *Assume* $(\mathbf{PSD}(p, C_A))$. *Let $\beta_+ > \beta_- \geq 0$, $R_+, R_- > 0$. Suppose that there exists a $\mathcal{G}$-stopping rule τ such that $\mathcal{R}(\mu, \tau) \leq C\mathcal{R}^*_{\beta, p, R}(\delta)$ holds for some $C > 0$, all $\delta > 0$ small enough, and for every $\mu \in H^\beta(R, D_\delta)$, simultaneously for $(\beta, R) \in \{(\beta_-, R_-), (\beta_+, R_+)\}$. Then the rate-optimal truncation time for $H^{\beta_-}(R_-, D_\delta)$ must satisfy $\sqrt{D_\delta} = \mathcal{O}(t_{\beta_-, p, R_-}(\delta))$ as $\delta \to 0$ (all other parameters being fixed).*

In particular, if a $\mathcal{G}$-stopping rule τ is rate-optimal over $H^\beta(R, D_\delta)$ for $\beta \in [\beta_{min}, \beta_{max}]$, $\beta_{max} > \beta_{min} \geq 0$, and some $R > 0$, then we necessarily must have $\beta_{max} \leq \liminf_{\delta \to 0} \frac{\log \delta^{-2}}{\log D_\delta} - p - 1/2$.

Proof. In this proof we denote by '$\lesssim$', '$\gtrsim$' inequalities holding up to factors depending on $C_A, p, \beta_+, \beta_-, R_-, R_+$. We apply Corollary 2.5 with $\beta = \beta_+$, $\alpha = \beta_-$ and $\bar{R} = R_-$, $R = \min(R_+, \bar{R}/2)$. Because of $\delta^{-2/(2\beta_- + 2p + 1/2)} \leq \delta^{-4/(2\beta_- + 2p + 1)} = o(D_\delta)$, the conditions are fulfilled for sufficiently small $\delta > 0$ and we conclude (R_+, R_- are fixed)

$$\exists \bar{\mu} \in H^{\beta_-}(R_-, D) : \mathcal{R}(\bar{\mu}, \tau) \gtrsim \min\left(\left(\delta D_\delta^{1/4}\right)^{2\beta_- / (2\beta_- + 2p)}, \delta^{2\beta_- / (2\beta_+ + 2p + 1)} \right).$$

By assumption, that rate must be $\mathcal{O}(\delta^{2\beta_- / (2\beta_- + 2p + 1)})$. Since the second term in the above minimum is of larger order than this, this must imply $(\delta D_\delta^{1/4})^{2\beta_- / (2\beta_- + 2p)} \lesssim \delta^{2\beta_- / (2\beta_- + 2p + 1)}$, and further $\sqrt{D_\delta} \lesssim \delta^{-2/(2\beta_- + 2p + 1)} \lesssim t_{\beta_-, p, R_-}(\delta)$. The first statement is proved.

For the second assertion, we proceed by contradiction and assume $\beta_{max} > \beta_{lim} := \liminf_{\delta \to 0} \frac{\log \delta^{-2}}{\log D_\delta} - p - 1/2$. Choose $\beta_+ = \beta_{max}$ and $\beta_- \in (\beta_{lim}, \beta_{max})$. Then $\beta_- > \beta_{lim}$ implies $t_{\beta_-, p, R_-}(\delta_k) = o(\sqrt{D_{\delta_k}})$ for some sequence $\delta_k \to 0$, contradicting the first part of the corollary. $\square$

For statistical inverse problems with singular values satisfying the polynomial decay (**PSD**(p, C_A)) we may choose the maximal dimension $D_\delta \sim \delta^{-2/(2p+1)}$ without losing in the convergence rate for a Sobolev ellipsoid of any regularity $\beta \geq 0$, see e.g. Cohen *el al.* [9]. In fact, we then have the variance

$$V_{D_\delta} = \delta^2 \sum_{i=1}^{D_\delta} \lambda_i^{-2} \sim \delta^2 D_\delta^{2p+1} \sim 1, \tag{2.2}$$

and the estimator with truncation at the order of D_δ will not be consistent anyway; the oracle index is always of order $o(D_\delta)$ whatever the signal regularity. For this choice of D_δ, optimal adaptation is only possible if the squared minimax rate is within the interval $[\delta, 1]$, faster adaptive rates up to δ^2 cannot be attained.

Usually, D_δ will be chosen much smaller, assuming some minimal a priori regularity β_{min}. The choice $D_\delta \sim \delta^{-2/(2\beta_{min} + 2p + 1)}$ ensures that rate optimality is possible for all (sequence space) Sobolev regularities $\beta \geq \beta_{min}$, when using either oracle (non-adaptive) rules, or adaptive rules that are not stopping times. In contrast, any $\mathcal{G}$-stopping rule can at best adapt over the regularity interval $[\beta_{min}, \beta_{max}]$ with $\beta_{max} = 2\beta_{min} + p + 1/2$ (keeping the radius R of the Sobolev ball fixed). These adaptation intervals, however, are fundamentally understood only when inspecting the corresponding rate-optimal truncation indices $t_{\beta, p, R}(\delta)$, which must at least be of order $\sqrt{D_\delta} \sim \delta^{-1/(2\beta_{min} + 2p + 1)}$ in order to distinguish a signal in the residual from the pure noise case.

3. Upper bounds

Consider the residual-based stopping rule $\tau = \min\{m \geq m_0 \,|\, R_m^2 \leq \kappa\}$ from (1.12). Since R_m^2 is decreasing with $R_D^2 = 0$, the minimum is attained and we have $R_\tau^2 \leq \kappa$.

In order to have clearer oracle inequalities, we work with continuous oracle-type truncation indices in $[0, D]$. Interpolating such that $V_{t,\lambda} = t\delta^2$ continues to hold for real $t \in [0, D]$, we set

$$\widehat{\mu}_i^{(t)} := \left(\mathbf{1}(i \leq \lfloor t \rfloor) + \sqrt{t - \lfloor t \rfloor}\mathbf{1}(i = \lfloor t \rfloor + 1)\right)\lambda_i^{-1}Y_i, \quad i = 1, \ldots, D,$$

and define further

$$R_t^2 = \left(1 - \sqrt{t - \lfloor t \rfloor}\right)^2 Y_{\lfloor t \rfloor + 1}^2 + \sum_{i=\lfloor t \rfloor + 2}^{D} Y_i^2,$$
$$B_t^2(\mu) = \left(1 - \sqrt{t - \lfloor t \rfloor}\right)^2 \mu_{\lfloor t \rfloor + 1}^2 + \sum_{i=\lfloor t \rfloor + 2}^{D} \mu_i^2,$$
$$V_t = (t - \lfloor t \rfloor)\delta^2\lambda_{\lfloor t \rfloor + 1}^{-2} + \delta^2 \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{-2},$$
$$S_t = (t - \lfloor t \rfloor)\delta^2\lambda_{\lfloor t \rfloor + 1}^{-2}\varepsilon_{\lfloor t \rfloor + 1}^2 + \delta^2 \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{-2}\varepsilon_i^2.$$

We thus obtain the following decompositions in a bias and a stochastic error term:

$$\|\widehat{\mu}^{(t)} - \mu\|^2 = B_t^2(\mu) + S_t + 2\delta\left(t - \lfloor t \rfloor - \sqrt{t - \lfloor t \rfloor}\right)\lambda_{\lfloor t \rfloor + 1}^{-1}\mu_{\lfloor t \rfloor + 1}\varepsilon_{\lfloor t \rfloor + 1}, \quad (3.1)$$
$$\mathbb{E}\left[\|\widehat{\mu}^{(t)} - \mu\|^2\right] = B_t^2(\mu) + V_t, \quad \mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] = \mathbb{E}\left[B_\tau^2(\mu) + S_\tau\right], \quad (3.2)$$

noting that the last term in (3.1) has expectation zero for deterministic t and vanishes for the integer-valued random time τ. Analogously, the linear interpolations for bias and variance in weak norm are defined. Thus, the continuously interpolated residual has expectation

$$\mathbb{E}[R_t^2] = B_{t,\lambda}^2(\mu) + \left(\left(1 - \sqrt{t - \lfloor t \rfloor}\right)^2 + (D - \lfloor t \rfloor - 1)\right)\delta^2$$
$$= B_{t,\lambda}^2(\mu) - V_{t,\lambda} + D\delta^2 - 2\left(\sqrt{t - \lfloor t \rfloor} - (t - \lfloor t \rfloor)\right)\delta^2. \quad (3.3)$$

Integrating the last interpolation error term into the definition, we define the *oracle-proxy index* $t^* \in [m_0, D]$ as

$$t^* = \inf\left\{t \geq m_0 \,\Big|\, \mathbb{E}_\mu[R_t^2] \leq \kappa - 2\left(\sqrt{t - \lfloor t \rfloor} - (t - \lfloor t \rfloor)\right)\delta^2\right\}.$$

Then by continuity $\mathbb{E}[R_{t^*}^2] = \kappa - 2(\sqrt{t_* - \lfloor t_* \rfloor} - (t_* - \lfloor t_* \rfloor))\delta^2 \in [\kappa - \frac{1}{2}\delta^2, \kappa]$ holds in the case $t^* > m_0$, implying

$$\kappa = D\delta^2 + B_{t^*,\lambda}^2(\mu) - V_{t^*,\lambda}. \quad (3.4)$$

For $t^* = m_0$ we still have $D\delta^2 + B_{t^*,\lambda}^2(\mu) - V_{t^*,\lambda} \leq \kappa$.

Let us finally define the *weakly and strongly balanced oracles* $t_{\mathfrak{w}}$ and $t_{\mathfrak{s}}$ in a continuous manner:

$$t_{\mathfrak{w}} = t_{\mathfrak{w}}(\mu) = \inf\{t \geq m_0 \mid B_{t,\lambda}^2(\mu) \leq V_{t,\lambda}\} \in [m_0, D],$$
$$t_{\mathfrak{s}} = t_{\mathfrak{s}}(\mu) = \inf\{t \geq m_0 \mid B_t^2(\mu) \leq V_t\} \in [m_0, D].$$

While the balanced oracles are the natural oracle quantities we try to mimic by early stopping, they should be compared to the classical oracles. Since $t \mapsto B_t^2(\mu)$ is decreasing and $t \mapsto V_t$ is increasing, we derive

$$\inf_{t\in[m_0,D]} \mathbb{E}\left[\|\widehat{\mu}^{(t)} - \mu\|^2\right] \geq \inf_{t\in[m_0,D]} \max(B_t^2(\mu), V_t) \geq V_{t_{\mathfrak{s}}} \geq \tfrac{1}{2}\,\mathbb{E}\left[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2\right],$$
(3.5)

noting $B_{t_{\mathfrak{s}}}^2(\mu) \leq V_{t_{\mathfrak{s}}}$.

3.1. Upper bounds in weak norm

The following is an analogue of Proposition 2.1 in [2], but includes a discretisation error for the discrete time stopping rule τ.

3.1 Proposition. *The balanced oracle inequality in weak norm*

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \widehat{\mu}^{(t^*)}\|_\lambda^2\right] \leq \sqrt{2D}\delta^2 + 2\delta B_{t^*,\lambda}(\mu) + \Delta_\tau(\mu)^2 \qquad (3.6)$$

holds with the discretisation error

$$\Delta_\tau(\mu) = \max_{i \geq \lfloor t^*\rfloor+1}|\lambda_i\mu_i| + 4\delta\Big(\big(\log(\sqrt{2}D)\big)^{1/2} + 1\Big).$$

Proof. The main argument is completely deterministic. For $\tau > t^* \geq m_0$ we obtain by $R_\tau^2 + Y_\tau^2 = R_{\tau-1}^2 > \kappa \geq \mathbb{E}[R_{t^*}^2]$:

$$\|\widehat{\mu}^{(t^*)} - \widehat{\mu}^{(\tau)}\|_\lambda^2 = (1 - \sqrt{t^* - \lfloor t^*\rfloor})^2 Y_{\lfloor t^*\rfloor+1}^2 + \sum_{i=\lfloor t^*\rfloor+2}^{\tau} Y_i^2$$
$$= R_{t^*}^2 - R_\tau^2 < R_{t^*}^2 - \mathbb{E}[R_{t^*}^2] + Y_\tau^2.$$

For $t^* > \tau \geq m_0$ we use $t^* - \lfloor t^*\rfloor \leq 1 - (1 - \sqrt{t^* - \lfloor t^*\rfloor})^2$ and $R_\tau^2 \leq \kappa \leq \mathbb{E}[R_{t^*}^2] + \tfrac{1}{2}\delta^2$:

$$\|\widehat{\mu}^{(t^*)} - \widehat{\mu}^{(\tau)}\|_\lambda^2 = (t^* - \lfloor t^*\rfloor) Y_{\lfloor t^*\rfloor+1}^2 + \sum_{i=\tau+1}^{\lfloor t^*\rfloor} Y_i^2$$
$$\leq R_\tau^2 - R_{t^*}^2 \leq \mathbb{E}[R_{t^*}^2] - R_{t^*}^2 + \tfrac{1}{2}\delta^2.$$

Consequently, we find

$$\mathbb{E}\left[\|\widehat{\mu}^{(t^*)} - \widehat{\mu}^{(\tau)}\|_\lambda^2\right]$$
$$\leq \mathbb{E}\left[\Big|\sum_{i=\lfloor t^*\rfloor+1}^{D} \gamma_i\big(\delta^2(\varepsilon_i^2 - 1) + 2\lambda_i\mu_i\delta\varepsilon_i\big)\Big|\right] + \mathbb{E}\left[\max_{i \geq \lfloor t^*\rfloor+1} Y_i^2\right] + \tfrac{1}{2}\delta^2,$$

with $\gamma_i = 1$ for $i > \lfloor t^* \rfloor + 1$ and $\gamma_i = (1 - \sqrt{t^* - \lfloor t^* \rfloor})^2$ for $i = \lfloor t^* \rfloor + 1$. The maximal inequality in Corollary 1.3 of [18] implies

$$\mathbb{E}\left[\max_{i \geq \lfloor t^* \rfloor + 1} Y_i^2\right] \leq \left(\max_{i \geq \lfloor t^* \rfloor + 1} |\lambda_i \mu_i| + 4\delta \left(\log\left(\sqrt{2}(D - \lfloor t^* \rfloor)\right)\right)^{1/2}\right)^2,$$

which is smaller than $\Delta_\tau(\mu)^2 - \frac{1}{2}\delta^2$. By bounding the main term via Jensen's inequality, using $\mathrm{Var}(\varepsilon_i^2) = 2$, $\mathrm{Cov}(\varepsilon_i^2, \varepsilon_i) = 0$, this gives

$$\mathbb{E}_\mu\left[\|\widehat{\mu}^{(t^*)} - \widehat{\mu}^{(\tau)}\|_\lambda^2\right] \leq \left(2(D - t^*)\delta^4 + 4\delta^2 B_{t^*,\lambda}^2(\mu)\right)^{1/2} + \Delta_\tau(\mu)^2,$$

and thus by $\sqrt{A + B} \leq \sqrt{A} + \sqrt{B}$, $A, B \geq 0$, the asserted inequality. $\qquad\square$

Remark that the proof only relies on the moments of (ε_i) up to fourth order and a maximal deviation inequality, so that an extension to sub-Gaussian distributions is straightforward. More heavy-tailed distributions can be treated at the cost of a looser bound on $\Delta_\tau(\mu)$.

So far, the choice of κ has not been addressed. The identity (3.4) shows that the choice $\kappa = D\delta^2$ balances weak squared bias and variance exactly such that $t^* = t_\mathfrak{w}$. In practice, however, we might have to estimate the noise level δ^2, or we prefer a larger threshold κ to reduce numerical complexity. Therefore, precise bounds for general κ between the oracle-proxy and the weakly balanced errors in weak norm are useful.

3.2 Lemma. *We have*

$$(B_{t^*,\lambda}^2(\mu) - B_{t_\mathfrak{w},\lambda}^2(\mu))_+ \leq (\kappa - D\delta^2)_+, \quad (V_{t^*,\lambda} - V_{t_\mathfrak{w},\lambda})_+ \leq (D\delta^2 - \kappa)_+,$$

so that

$$\mathbb{E}[\|\widehat{\mu}^{(t^*)} - \mu\|_\lambda^2] \leq \mathbb{E}[\|\widehat{\mu}^{(t_\mathfrak{w})} - \mu\|_\lambda^2] + |\kappa - D\delta^2|.$$

Proof. Suppose $t_\mathfrak{w} > t^* \geq m_0$. Then $V_{t^*,\lambda} < V_{t_\mathfrak{w},\lambda}$ and from $\kappa \geq B_{t^*,\lambda}^2(\mu) + D\delta^2 - V_{t^*,\lambda}$ (see (3.4) and afterwards), $V_{t_\mathfrak{w},\lambda} = B_{t_\mathfrak{w},\lambda}^2(\mu)$ we deduce

$$B_{t^*,\lambda}^2(\mu) \leq V_{t^*,\lambda} + \kappa - D\delta^2 < V_{t_\mathfrak{w},\lambda} + \kappa - D\delta^2 = B_{t_\mathfrak{w},\lambda}^2(\mu) + \kappa - D\delta^2.$$

Conversely, for $t^* > t_\mathfrak{w} \geq m_0$ we have $B_{t^*,\lambda}^2(\mu) \leq B_{t_\mathfrak{w},\lambda}^2(\mu)$ as well as (3.4) and $V_{t_\mathfrak{w},\lambda} \geq B_{t_\mathfrak{w},\lambda}^2(\mu)$, so that

$$V_{t^*,\lambda} = B_{t^*,\lambda}^2(\mu) - \kappa + D\delta^2 \leq B_{t_\mathfrak{w},\lambda}^2(\mu) - \kappa + D\delta^2 \leq V_{t_\mathfrak{w},\lambda} - \kappa + D\delta^2.$$

This gives the result. $\qquad\square$

Remark that the weak variance control of Lemma 3.2 implies directly $(t^* - t_\mathfrak{w})_+ \leq (D - \kappa\delta^{-2})_+$. From the inequalities $B_t^2(\mu) \geq \lambda_{\lfloor t \rfloor}^{-2} B_{t,\lambda}^2(\mu)$ and $V_t \leq \lambda_{\lfloor t \rfloor}^{-2} V_{t,\lambda}$ we infer further $t_\mathfrak{w} \leq t_\mathfrak{s}$, and thus it always holds

$$t^* - (D - \kappa\delta^{-2})_+ \leq t_\mathfrak{w} \leq t_\mathfrak{s}. \tag{3.7}$$

As a consequence of the preceding two results, we obtain directly a weakly balanced oracle inequality with error terms of order $\sqrt{D}\delta^2$, provided $|\kappa - D\delta|^2$ is at most of that order:

3.3 Theorem. *We have*

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|_\lambda^2\right] \le C\Big(\mathbb{E}\left[\|\widehat{\mu}^{(t_{\mathfrak{w}})} - \mu\|_\lambda^2\right] + \sqrt{D}\delta^2 + |\kappa - D\delta^2|\Big)$$

$$\le C\Big(2\min_{t\in[m_0,D]}\mathbb{E}\left[\|\widehat{\mu}^{(t)} - \mu\|_\lambda^2\right] + \sqrt{D}\delta^2 + |\kappa - D\delta^2|\Big)$$

with a numerical constant $C > 0$.

Proof. For the first bound use

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|_\lambda^2\right] \le 2\,\mathbb{E}\left[\|\widehat{\mu}^{(t_{\mathfrak{w}})} - \mu\|_\lambda^2\right] + 2\,\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \widehat{\mu}^{(t_{\mathfrak{w}})}\|_\lambda^2\right]$$

and apply Proposition 3.1 and Lemma 3.2 with the estimates $2\delta B_{t^*,\lambda}(\mu) \le \delta^2 + B_{t^*,\lambda}^2(\mu)$, $\Delta_\tau(\mu)^2 \lesssim B_{t^*,\lambda}^2(\mu) + \sqrt{D}\delta^2$, $B_{t^*,\lambda}^2(\mu) \le B_{t_{\mathfrak{w}},\lambda}^2(\mu) + |\kappa - D\delta^2|$ ('$\lesssim$' denotes an inequality up to a numerical factor). The second bound follows exactly as (3.5). □

In weak norm, we have thus obtained a completely general oracle inequality for our early stopping rule. In view of the lower bounds, the "residual term" of order $\sqrt{D}\delta^2$, which is much larger than the usual parametric order δ^2, is unavoidable. This will be developed further in the strong norm error analysis.

3.2. Upper bounds in strong norm

In Appendix 5.5 we derive exponential bounds for $P(R_m^2 \le \kappa)$, $m < t^*$, in terms of the weak bias and deduce by partial summation the following weak bias deviation inequality:

3.4 Proposition. *We have*

$$\mathbb{E}\left[(B_{\tau,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu))_+\right] \le \left(17\sqrt{D} + 64\right)\delta^2 + B_{t^*,\lambda}^2(\mu)D^{-1/2}.$$

This is the probabilistic basis for the main bias oracle inequality.

3.5 Proposition. *We have the balanced oracle inequality for the strong bias*

$$\mathbb{E}[(B_\tau^2(\mu) - B_{t_{\mathfrak{s}}}^2(\mu))_+] \le 81\lambda_{\lfloor t_{\mathfrak{s}}\rfloor+1}^{-2}\delta^2\Big(t_{\mathfrak{s}} + \sqrt{D} + (\kappa\delta^{-2} - D)_+\Big).$$

Proof. On the event $\{\tau \ge t_{\mathfrak{s}}\}$ we have $B_\tau^2(\mu) \le B_{t_{\mathfrak{s}}}^2(\mu)$. On $\{\tau < t_{\mathfrak{s}}\}$ we have

$$B_\tau^2(\mu) - B_{t_{\mathfrak{s}}}^2(\mu) \le \lambda_{\lfloor t_{\mathfrak{s}}\rfloor+1}^{-2}(B_{\tau,\lambda}^2(\mu) - B_{t_{\mathfrak{s}},\lambda}^2(\mu)),$$

using the fact that only coefficients up to index $\lfloor t_{\mathfrak{s}}\rfloor + 1$ enter into the bias differences. From the weak bias control given by Proposition 3.4 it follows that

$$\mathbb{E}[(B_\tau^2(\mu) - B_{t_{\mathfrak{s}}}^2(\mu))_+]$$
$$\le \lambda_{\lfloor t_{\mathfrak{s}}\rfloor+1}^{-2}\Big(\mathbb{E}[(B_{\tau,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu))_+] + (B_{t^*,\lambda}^2(\mu) - B_{t_{\mathfrak{s}},\lambda}^2(\mu))_+\Big)$$

$$\leq \lambda_{\lfloor t_{\mathfrak{s}} \rfloor + 1}^{-2} \Big((17\sqrt{D} + 64)\delta^2 + (1 + D^{-1/2})B_{t^*,\lambda}^2(\mu) \Big)$$

$$\leq \lambda_{\lfloor t_{\mathfrak{s}} \rfloor + 1}^{-2} \delta^2 \Big(81\sqrt{D} + 2(t^* + \kappa\delta^{-2} - D) \Big),$$

where in the last line we used $\kappa \geq B_{t^*,\lambda}^2(\mu) + D\delta^2 - V_{t^*,\lambda}$ (see (3.4) and afterwards) and $V_{t^*,\lambda} = t^*\delta^2$. By (3.7) we see $t^* \leq t_{\mathfrak{s}} + (D - \kappa\delta^{-2})_+$ and the result follows. □

To assess the size of the bias bound, let us assume the polynomial decay (**PSD**(p, C_A)). Then a Riemann sum approximation yields for any $t \in [1, D]$

$$\delta^{-2}V_t = \sum_{i=1}^{\lfloor t \rfloor} \lambda_i^{-2} + (t - \lfloor t \rfloor)\lambda_{\lfloor t \rfloor + 1}^{-2} \geq C_A^{-2} \int_0^t x^{2p}dx = C_A^{-2}(2p+1)^{-1}t^{2p+1}.$$

Noting $t\lambda_{\lfloor t \rfloor + 1}^{-2} \leq C_A^2(\lfloor t \rfloor + 1)^{2p+1} \leq C_A^2(2t)^{2p+1}$, we thus obtain

$$\lambda_{\lfloor t \rfloor + 1}^{-2} t\delta^2 \leq (1 + 2p)2^{2p+1}C_A^4 V_t. \tag{3.8}$$

Consequently, we can estimate $\mathbb{E}[(B_\tau^2(\mu) - B_{t_{\mathfrak{s}}}^2(\mu))_+] \lesssim V_{t_{\mathfrak{s}}}$ in the case $t_{\mathfrak{s}} \gtrsim \max(\sqrt{D}, \kappa\delta^{-2} - D)$. This means that the bias bound is upper bounded by the balanced strong oracle risk.

Let us see by a counterexample that $(B_{t^*}^2(\mu) - B_{t_{\mathfrak{s}}}^2(\mu))_+$ can be of the same order as the strongly balanced risk itself, meaning that the bound of Proposition 3.5 is not too pessimistic in general. Suppose $\kappa = D\delta^2$ (so that $t^* = t_{\mathfrak{w}}$), $\mu_D \neq 0$ and δ, D such that $t_{\mathfrak{s}} = D - 3/4$. This gives $\mu_D^2/4 = B_{t_{\mathfrak{s}}}^2(\mu) = V_{t_{\mathfrak{s}}}$. In weak norm we have $B_{D-1,\lambda}^2(\mu) = \lambda_D^2\mu_D^2$, $V_{D-1,\lambda} = \delta^2(D-1)$ and consequently $t_{\mathfrak{w}} \leq D - 1$ if $\lambda_D^2\mu_D^2 \leq \delta^2(D-1)$. In that case, $B_{t^*}^2(\mu) \geq \mu_D^2 = 4B_{t_{\mathfrak{s}}}^2(\mu)$ holds and we must indeed pay a positive factor for using the weak oracle in strong norm. We can meet the bound $\lambda_D^2\mu_D^2 \leq \delta^2(D-1)$ under the constraint $\mu_D^2/4 = V_{D-3/4} = \delta^2(\lambda_D^{-2}/4 + \sum_{i=1}^{D-1}\lambda_i^{-2})$ for instance for $\lambda_i = i^{-p}$ with $p > 3/2$ and D sufficiently large.

For the stochastic error we use in Appendix 5.6 exponential inequalities for $P(R_{m-1}^2 > \kappa)$, $m > t^*$, to obtain the following bound:

3.6 Proposition. *We have the oracle-proxy inequality for the strong norm stochastic error*

$$\mathbb{E}[(S_\tau - S_{t^*})_+] \leq r_{V,\tau}\delta^2 \text{ with } r_{V,\tau} := \min\Big(2\sqrt{3}\sum_{m=\lfloor t^* \rfloor + 1}^{D} \lambda_m^{-2}e^{-\frac{(m-1-t^*)_+^2}{16D+32\kappa\delta^{-2}}}, D\Big).$$

*If the polynomial decay condition (**PSD**(p, C_A)) is satisfied, then*

$$r_{V,\tau}\delta^2 \leq C_p C_A^4 \Big(\big((D + \kappa\delta^{-2})/(t^*)^2\big)^{1/2} + \big((D + \kappa\delta^{-2})/(t^*)^2\big)^{p+1/2} \Big)V_{t^*} \tag{3.9}$$

holds with a constant C_p, only depending on p.

3.7 Corollary. *We have the balanced oracle inequality for the stochastic error*

$$\mathbb{E}[(S_\tau - S_{t_s})_+] \leq \Big(r_{V,\tau} + \lambda_{\lfloor t^*\rfloor+1}^{-2}(D - \kappa\delta^{-2})_+\Big)\delta^2.$$

Proof. By the monotonicity of S_t and V_t in t we bound

$$\mathbb{E}[(S_\tau - S_{t_s})_+] \leq \mathbb{E}[(S_\tau - S_{t^*})_+] + \mathbb{E}[(S_{t^*} - S_{t_s})_+] = \mathbb{E}[(S_\tau - S_{t^*})_+] + (V_{t^*} - V_{t_s})_+\,.$$

In view of Proposition 3.6 it suffices to prove $(V_{t^*} - V_{t_s})_+ \leq \lambda_{\lfloor t^*\rfloor+1}^{-2}(D\delta^2 - \kappa)_+$. By definition of the variances, $(V_{t^*} - V_{t_w})_+ \leq \lambda_{\lfloor t^*\rfloor+1}^{-2}(V_{t^*,\lambda} - V_{t_w,\lambda})_+$ holds. We apply Lemma 3.2 and note $V_{t_w} \leq V_{t_s}$ by (3.7) to conclude. □

Everything is prepared to prove our main strong norm result.

3.8 Theorem. *Assume $|\kappa - D\delta^2| \leq C_\kappa\sqrt{D}\delta^2$. Then the following balanced oracle inequality holds in strong norm*

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] \leq \mathbb{E}\left[\|\widehat{\mu}^{(t_s)} - \mu\|^2\right]$$
$$+ \Big(81\lambda_{\lfloor t_s+C_\kappa\sqrt{D}\rfloor+1}^{-2}\big(t_s + (1+C_\kappa)\sqrt{D}\big) + r_{V,\tau}\Big)\delta^2.$$

*If in addition the polynomial decay condition (**PSD**(p, C_A)) is satisfied, then there is a constant $C > 0$, only depending on p, C_A, C_κ, so that*

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] \leq C\,\mathbb{E}\left[\|\widehat{\mu}^{(t_s\vee\sqrt{D})} - \mu\|^2\right]. \tag{3.10}$$

3.9 Remarks.

(a) *The impact of the polynomial decay condition (**PSD**(p, C_A)) is quite transparent here. If the eigenvalues decay exponentially, $\lambda_i = e^{-\alpha i}$ say, then a factor $e^{2\alpha C_\kappa\sqrt{D}}$ appears in the balanced oracle inequality, which we then also lose compared to the optimal minimax rate in strong norm. Polynomial decay ensures $\lambda_{\lfloor t_s+C_\kappa\sqrt{D}\rfloor+1}^{-2} \lesssim \lambda_{t_s}^{-2}$ for $t_s \geq \sqrt{D}$. Intuitively, stopping about $\sqrt{D}$ steps later does not affect the rate under (**PSD**(p, C_A)), but does affect it under exponential singular value decay.*

(b) *Compared with the main Theorem 3.5 in [2] this is a proper oracle inequality since for the truncated SVD method $t_w \leq t_s$ always holds. Note also that the more direct proof here gives much simpler and tighter bounds.*

Proof. By (3.2) we have

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2 - \|\widehat{\mu}^{(t_s)} - \mu\|^2\right] \leq \mathbb{E}\left[\big(B_\tau^2(\mu) - B_{t_s}^2(\mu)\big)_+ + \big(S_\tau - S_{t_s}\big)_+\right].$$

Combining Proposition 3.5 and Corollary 3.7 we thus obtain

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] \leq \mathbb{E}\left[\|\widehat{\mu}^{(t_s)} - \mu\|^2\right]$$
$$+ 81\lambda_{\lfloor t_s\vee t^*\rfloor+1}^{-2}\delta^2\big(t_s + \sqrt{D} + |\kappa\delta^{-2} - D|\big) + r_{V,\tau}\delta^2.$$

By (3.7) we have $t^* \leq t_{\mathfrak{s}} + (D - \kappa\delta^{-2})_+$ and the first inequality follows.

Under $(\mathbf{PSD}(p, C_A))$ we use (3.9), $V_{t^*} \lesssim (t_*)^{2p+1}\delta^2$, $\kappa\delta^{-2} \lesssim D$ and $t^* \leq t_{\mathfrak{s}} + C_\kappa\sqrt{D}$ to further bound

$$r_{V,\tau} \lesssim t_{\mathfrak{s}}^{2p}\sqrt{D} + D^{p+1/2},$$

with a factor depending on p, C_A, C_κ. Finally, note

$$\mathbb{E}[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2] \leq 2V_{t_{\mathfrak{s}}} \leq 2V_{t_{\mathfrak{s}} \vee \sqrt{D}} \leq 2\,\mathbb{E}[\|\widehat{\mu}^{(t_{\mathfrak{s}} \vee \sqrt{D})} - \mu\|^2],$$
$$V_{t_{\mathfrak{s}} \vee \sqrt{D}} \sim (t_{\mathfrak{s}} \vee \sqrt{D})^{2p+1}\delta^2,$$

and apply $\lambda^{-2}_{\lfloor t_{\mathfrak{s}} + C_\kappa\sqrt{D}\rfloor + 1} \lesssim (t_{\mathfrak{s}} \vee \sqrt{D})^{2p}$ to deduce the second bound. $\qquad\square$

Again, the proof only relies on concentration bounds for the residuals R_m^2 and easily extends to sub-Gaussian errors. Let us now derive from Theorem 3.8 an asymptotic minimax upper bound over the Sobolev-type ellipsoids $H^\beta(R, D)$. For $m_0 = \lfloor\sqrt{D}\rfloor + 1$ the bound (3.10) gives

$$\mathbb{E}\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] \leq C\,\mathbb{E}\left[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2\right]$$

because of $t_{\mathfrak{s}} \geq m_0 \geq \sqrt{D}$. Now, $t_{\mathfrak{s}}(\mu) \lesssim t_{\beta,p,R}(\delta)$ holds with the optimal truncation index $t_{\beta,p,R}(\delta)$ from (1.9) for $\mu \in H^\beta(R, D)$ and if $t_{\beta,p,R}(\delta) \in [m_0, D]$; under these conditions $\mathbb{E}[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2] \lesssim \mathcal{R}^*_{\beta,p,R}(\delta)$ is thus true and we obtain the following adaptive upper bound:

3.10 Corollary. *Assume* $(\mathbf{PSD}(p, C_A))$, $|\kappa - D\delta^2| \leq C_\kappa\sqrt{D}\delta^2$ *and choose* $m_0 = \lfloor\sqrt{D}\rfloor + 1$. *Then there is a constant* $C > 0$, *depending only on* p, C_A *and* C_κ, *such that for all* (β, R) *with* $t_{\beta,p,R}(\delta) \in [\sqrt{D}, D]$

$$\sup_{\mu \in H^\beta(R,D)} \mathbb{E}_\mu\left[\|\widehat{\mu}^{(\tau)} - \mu\|^2\right] \leq C\mathcal{R}^*_{\beta,p,R}(\delta).$$

In summary, together with the matching lower bound of Corollary 2.6 this shows that the stopping rule τ is sequentially minimax adaptive.

4. An adaptive two-step procedure

4.1. Construction and results

The lower bounds show that, in general, there is no hope for an early stopping rule attaining the order of the (unconstrained) oracle risk if the strongly balanced oracle $t_{\mathfrak{s}}$ is of smaller order than $\sqrt{D}$. We can therefore always start the stopping rule τ at some $m_0 \gtrsim \sqrt{D}$. If, however, immediate stopping $\tau = m_0$ occurs, we might have stopped too late in the sense that $t_{\mathfrak{s}} \ll m_0$. To avoid this overfitting, we propose to run a second model selection step on $\{\widehat{\mu}^{(0)}, \ldots, \widehat{\mu}^{(m_0)}\}$ in the event $\tau = m_0$.

Below, we shall formalise this procedure and prove that this combined model selection indeed achieves adaptivity, that is, its risk is controlled by an oracle inequality. While violating the initial stopping rule prescription, we still gain substantially in terms of numerical complexity. At the heart of this twofold model selection procedure is a simple observation of independence.

4.1 Lemma. *The stopping rule τ is independent of the estimators $\widehat{\mu}^{(0)}, \ldots \widehat{\mu}^{(m_0)}$.*

Proof. By construction, τ is measurable with respect to the σ-algebra $\sigma(R_{m_0}^2, \ldots, R_D^2) = \sigma(Y_{m_0+1}^2, \ldots, Y_D^2)$ and $\widehat{\mu}^{(m)}$ is $\sigma(Y_1, \ldots, Y_m)$-measurable. By the independence of $(Y_1, \ldots, Y_{m_0})$ and $(Y_{m_0+1}, \ldots, Y_D)$ the claim follows. □

For the second step, we suppose that $\widehat{m} \in \{0, \ldots, m_0\}$ is obtained from any model selection procedure among $\{\widehat{\mu}^{(0)}, \ldots, \widehat{\mu}^{(m_0)}\}$ that satisfies with a constant $C_2 \geq 1$, for any signal μ, the oracle inequality

$$\mathbb{E}[\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2] \leq C_2 \Big(\min_{m \in \{0, \ldots, m_0\}} \mathbb{E}[\|\widehat{\mu}^{(m)} - \mu\|^2] + \delta^2 \Big). \qquad (4.1)$$

Such an oracle inequality holds for standard procedures, for instance the AIC-criterion

$$\widehat{m} \in \operatorname{argmin}_{m \in \{0, \ldots, m_0\}} \Big(-\sum_{i=1}^m \lambda_i^{-2} Y_i^2 + 2\delta^2 \sum_{i=1}^m \lambda_i^{-2} \Big).$$

We refer to Section 2.3 in Cavalier and Golubev [7] for the corresponding result and further discussion. If we are interested in a weak norm oracle inequality, the AIC-criterion takes the weak empirical risk and reduces to the minimisation of $-\sum_{i=1}^m Y_i^2 + 2m\delta^2$, which is classical. Based on the lemma and the tools developed in the previous section, we prove in Appendix 5.7 the following oracle inequality in an asymptotic setting.

4.2 Proposition. *Assume $D \geq 3$, (**PSD**(p, C_A)), $|\kappa - D\delta^2| \leq C_\kappa \sqrt{D} \delta^2$ and set $m_0 = \lfloor 128 \log(D) \sqrt{D} \rfloor + 1$. Suppose the model selector $\widehat{m}$ satisfies (4.1) with $C_2 \geq 1$. Then there is a constant $C > 0$, depending only on p, C_A, C_κ and C_2, such that uniformly over all signals μ the estimator*

$$\widehat{\mu}^{(\rho)} = \begin{cases} \widehat{\mu}^{(\widehat{m})}, & \text{if } \tau = m_0, \\ \widehat{\mu}^{(\tau)}, & \text{if } \tau > m_0 \end{cases}$$

satisfies

$$\mathbb{E}\left[\|\widehat{\mu}^{(\rho)} - \mu\|^2\right] \leq C \Big(\min_{m \in \{0, \ldots, D\}} \mathbb{E}\left[\|\widehat{\mu}^{(m)} - \mu\|^2\right] + \delta^2 \Big).$$

In particular, $\widehat{\mu}^{(\rho)}$ is minimax adaptive over all Sobolev-type balls $H^\beta(R, D)$ at a usually much reduced computational complexity compared to standard model selection procedures requiring all $\widehat{\mu}^{(m)}$, $m = 0, \ldots, D$.
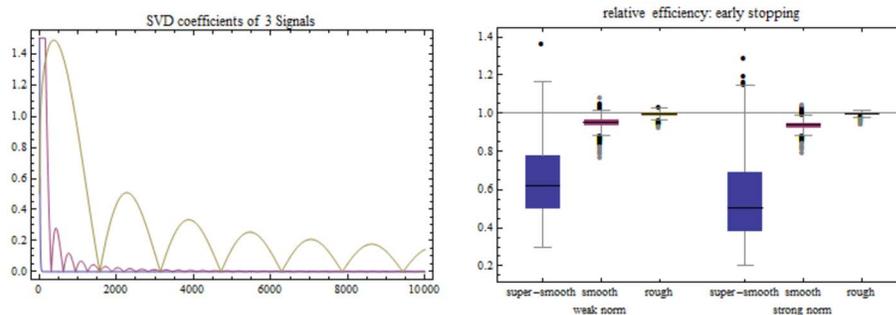
FIG 1. *Left: SVD representation of a super-smooth (blue), a smooth (red) and a rough (olive) signal. Right: Relative efficiency for early stopping with $m_0 = 0$.*

4.2. Numerical illustration

Let us exemplify the procedure by some Monte Carlo results. As a test bed we take the moderately ill-posed case $\lambda_i = i^{-1/2}$ with noise level $\delta = 0.01$ and dimension $D = 10\,000$. We consider early stopping at τ with $\kappa = D\delta^2 = 1$.

In Figure 1 (left), we see the SVD representation of three signals: a very smooth signal $\mu(1)$, a relatively smooth signal $\mu(2)$ and a rough signal $\mu(3)$, the attributes coming from the interpretation via the decay of Fourier coefficients. The corresponding weakly balanced oracle indices $t_\mathfrak{w}$ are $(34, 316, 1356)$. The classical oracle indices in strong norm are $(43, 504, 1331)$. Figure 1 (right) shows box-plots of the relative efficiency of early stopping in 1000 Monte Carlo replications defined as $\min_m \mathbb{E}[\|\widehat{\mu}^{(m)} - \mu\|^2]^{1/2}/\|\widehat{\mu}^{(\tau)} - \mu\|$, both for strong and weak norm. Ideally, the relative efficiency should concentrate around one. This is well achieved for the smooth and rough signals and even better than for the corresponding Landweber results in [2]. The super-smooth case with its very small oracle risk suffers from the variability within the residual and attains on average an efficiency of about 0.5, meaning that its root mean squared error is about twice as large as the oracle error. Let us mention that in unreported situations with higher ill-posedness the *relative* efficiency is similarly good or even better.

We are lead to consider the two-step procedure. According to Proposition 4.2 we have to choose an initial index somewhat larger than $\sqrt{D}$. The factor in the choice there is very conservative due to non-tight concentration bounds. For the implementation we choose m_0 such that for a zero signal $\mu = 0$ the probability of $\{\tau > m_0\} = \{R_{m_0}^2 > \kappa\}$ is about 0.01, when applying a normal approximation, that is $m_0 = \lfloor q_{0.99}\sqrt{2D} \rfloor + 1 = 329$ with the 99%-percentile $q_{0.99}$ of $N(0, 1)$. In Figure 2(left) we see that with this choice for the super-smooth signal, 6 out of $1\,000$ MC realisations lead to $\tau > m_0$, for the others we apply the second model selection step. The truncation for the smooth signal varies around m_0, and the second step is applied to about 50% of the realisations. In the rough case, $\tau > m_0$ was always satisfied and no second model selection step was applied.
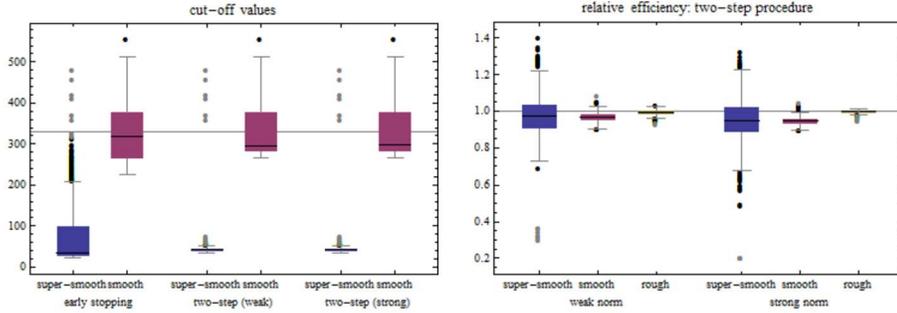
FIG 2. *Left: truncation levels for early stopping with $m_0 = 0$ and the two-step procedures with $m_0 = \lfloor q_{0.99}\sqrt{2D} \rfloor + 1 = 329$, AIC in weak and strong norm. Right: Relative efficiencies for the two-step procedure.*

As model selection procedure we apply the AIC-criterion, based on the weak and strong empirical norm for the weak and strong norm criterion, respectively. The results are shown in Figure 2(right). We see that the efficiency for the super-smooth signal improves significantly (with the 6 outliers not being affected). The variability is still considerably higher than for the other two signals. This phenomenon is well known for unbiased risk estimation. Especially for more strongly ill-posed problems, one should penalise stronger, see the comparison with the risk hull approach in Cavalier and Golubev [7] and the numerical findings in Lucka *et al.* [14]. Here let us rather emphasize that a pure AIC-minimisation for the super-smooth signal gives exactly the same result, apart from the 6 outliers, but requires to calculate the AIC-criterion for $D = 10\,000$ indices in $1\,000$ MC iterations. The two-step procedure, even for known SVD, is about 30 times faster.

5. Appendix

5.1. Proof of Corollary 2.2

Proof. For $i_0 = \lfloor (2C_\mu^2 C_A)^{1/(2p+1)}(R^{-1}\delta)^{-2/(2\beta+2p+1)} \rfloor$, we can choose c_2 (in dependence of C_μ, C_A) big enough so that our assumptions imply $i_0 \leq D$ and

$$1 - \frac{\mathcal{R}(\mu, \tau)^2}{V_{i_0+1}} \geq 1 - C_\mu^2 C_A \left(\frac{(R^{-1}\delta)^{-2/(2\beta+2p+1)}}{i_0 + 1} \right)^{1+2p} \geq \frac{1}{2}.$$

Put $\bar{\mu}_i = \mu_i$ for $i \neq i_0 + 1$ and $\bar{\mu}_{i_0+1} = \frac{1}{2}\bar{R}(i_0 + 1)^{-\alpha}$. Then $\bar{\mu} \in H^\alpha(\bar{R}, D)$ follows from $\mu \in H^\beta(R, D) \subseteq H^\alpha(R, D)$ and $\bar{R} \geq 2R$. The bias bound $B_{i_0}^2(\bar{\mu}) \geq \frac{1}{4}\bar{R}^2(i_0 + 1)^{-2\alpha}$ inserted in Proposition 2.1 yields the result. $\qquad \square$

5.2. Proof of Lemma 2.3

Proof. With $S_m = \delta^2 \sum_{i=1}^{m} \lambda_i^{-2} \varepsilon_i^2$ we obtain

$$\mathcal{R}(\mu, \tau)^2 \geq \mathbb{E}\Big[\delta^2 \sum_{i=1}^{\tau} \lambda_i^{-2} \varepsilon_i^2\Big] \geq \mathbb{E}\Big[\mathbf{1}(\tau \geq m) S_m\Big].$$

By Lemma 1 in Laurent and Massart [12], for nonnegative weights a_i, we have

$$P\Big(\sum_{i=1}^{D} a_i(\varepsilon_i^2 - 1) < -2\|a\|\sqrt{x}\Big) < e^{-x}. \tag{5.1}$$

Picking $a = \delta^2(\lambda_1^{-2}, \ldots, \lambda_m^{-2})$ and $x = \log(5/4)$ so that $2\sqrt{x} \leq 0.95$, with probability larger than $1 - e^{-x} = 0.2$, it follows that

$$S_m \geq \mathbb{E}[S_m] - 2\|a\|\sqrt{x} \geq \frac{\delta^2}{20} \sum_{i=1}^{m} \lambda_i^{-2} = \frac{V_m}{20},$$

where we used $\|a\| \leq \sum_{i=1}^{m} a_i = \mathbb{E}[S_m] = V_m$ (observe that we could tighten the latter inequality significantly under some additional assumptions on the singular value decay). We now have

$$\begin{aligned}
\mathcal{R}(\mu, \tau)^2 &\geq \mathbb{E}\left[\mathbf{1}(\tau \geq m) S_m\right] \\
&\geq V_m P(\{\tau \geq m\} \cap \{S_m \geq V_m/20\})/20 \\
&\geq V_m \big(1 - P(\tau < m) - P(S_m < V_m/20)\big)/20 \\
&\geq V_m \big(0.2 - P(\tau < m)\big)/20.
\end{aligned}$$

We deduce from this that $V_m \geq 200 \mathcal{R}(\mu, \tau)^2$ implies $P(\tau \geq m) \leq 0.9$. $\qquad\square$

5.3. A total variation bound for non-central χ^2-laws

5.1 Lemma. *Let $\vartheta = (\vartheta_1, \ldots, \vartheta_K) \in \mathbb{R}^K$ and $\mathbb{P}_K^\vartheta$ be the non-central χ^2-law of $X_\vartheta = \sum_{k=1}^{K} (\vartheta_k + Z_k)^2$ with Z_k independent and standard Gaussian. Then, for $\vartheta, \bar\vartheta \in \mathbb{R}^K$ we have*

$$\|\mathbb{P}_K^\vartheta - \mathbb{P}_K^{\bar\vartheta}\|_{TV} \leq e \frac{|\|\vartheta\|^2 - \|\bar\vartheta\|^2| + \sqrt{8/\pi}\,|\|\vartheta\| - \|\bar\vartheta\||}{\sqrt{\pi K}},$$

For $\|\vartheta\| + \|\bar\vartheta\| \geq \frac{\sqrt{8}e}{2\pi - \sqrt{\pi}e} \approx 5.248$ this bound simplifies to

$$\|\mathbb{P}_K^\vartheta - \mathbb{P}_K^{\bar\vartheta}\|_{TV} \leq 2 \frac{|\|\vartheta\|^2 - \|\bar\vartheta\|^2|}{\sqrt{K}}.$$

Proof. Writing $\vartheta = (\vartheta_k), Z = (Z_k) \in \mathbb{R}^k$ we see by orthogonal transformation that $X_\vartheta = \|\vartheta\|^2 + 2\langle\vartheta, Z\rangle + \|Z\|^2$ equals in law $X'_\vartheta = \|\vartheta\|^2 + 2\|\vartheta\|Z'_1 + \|Z'\|^2$ with $Z'_1, \ldots, Z'_K \sim N(0,1)$ i.i.d. We can therefore first consider the conditional law $\mathbb{Q}^\vartheta_K(z)$ of $\mathbb{P}^\vartheta_K$ given $\{Z'_1 = z\}$, which is nothing but the $\chi^2(K-1)$-distribution translated by $\|\vartheta\|^2 + 2\|\vartheta\|z + z^2$.

If f_p denotes the $\chi^2(p)$-density, then we have for any $t > 0$ that $f_p(x-t) > f_p(x)$ holds iff $x \geq x_t = \frac{t}{1-e^{-t/(p-2)}}$. Thus, we obtain

$$\int_0^\infty |f_p(x-t) - f_p(x)| \, dx$$

$$= 2 \int_0^\infty \left(f_p(x-t) - f_p(x) \right)_+ dx$$

$$= \frac{2^{1-p/2}}{\Gamma(p/2)} \int_{x_t}^\infty \left((1-t/x)^{p/2-1}e^{t/2} - 1 \right) x^{p/2-1}e^{-x/2}dx$$

$$= \frac{2^{1-p/2}}{\Gamma(p/2)} \int_{x_t-t}^{x_t} x^{(p-2)/2}e^{-x/2}dx$$

$$\leq \frac{2^{(2-p)/2}}{\Gamma(p/2)} t(p-2)^{(p-2)/2}e^{-(p-2)/2},$$

knowing that $x = p-2$ is the mode of f_p. Stirling's formula guarantees $\Gamma(x) \geq \sqrt{2\pi/x}(x/e)^x$ for all $x > 0$ such that the last expression is always bounded by $t(\pi p)^{-1/2}e$. This yields

$$\|\mathbb{Q}^\vartheta_K(z) - \mathbb{Q}^{\bar\vartheta}_K(z)\|_{TV} \leq e(\pi K)^{-1/2}\big|\|\vartheta\|^2 - \|\bar\vartheta\|^2 + 2(\|\vartheta\| - \|\bar\vartheta\|)z\big|.$$

Taking expectation with respect to $Z'_1 \sim N(0,1)$ we conclude

$$\|\mathbb{P}^\vartheta_K - \mathbb{P}^{\bar\vartheta}_K\|_{TV} \leq e(\pi K)^{-1/2}\big|\|\vartheta\| - \|\bar\vartheta\|\big| \mathbb{E}\left[\big|\|\vartheta\| + \|\bar\vartheta\| + 2Z'_1\big|\right].$$

Using the triangle inequality and $\mathbb{E}[|Z'_1|] = \sqrt{2/\pi}$, the upper bound follows. $\square$

5.4. Proof of Corollary 2.5

Proof. Set $\bar\mu_i = \mu_i$ for $i \neq i_0 + 1$ and $\bar\mu^2_{i_0+1} = \mu^2_{i_0+1} + \frac{1}{4}\bar R^2(i_0+1)^{-2\alpha}$ for some $i_0 \in \{1, \ldots, D\}$, so that $\bar\mu \in H^\alpha(\bar R, D)$ and condition (a) of Proposition 2.4 is satisfied. If

(b'): $\frac{C_A^2}{4}\bar R^2(i_0+1)^{-2(\alpha+p)} \leq 0.025\delta^2\sqrt{D-i_0}$

holds, then condition (b) of Proposition 2.4 is ensured, whereas

(c'): $\bar R^2(i_0+1)^{-2(\alpha+p)} \geq 2C_A^2 5.25^2\delta^2$

implies condition (c) of Proposition 2.4. Finally, for

(d'): $i_0 \geq \lfloor(200(1+2p)C_A^2 C_\mu^2)^{1/(2p+1)}(R^2\delta^{-2})^{1/(2\beta+2p+1)}\rfloor,$

we have $V_{i_0+1} \geq 200\mathcal{R}(\mu,\tau)^2$. Hence, by Proposition 2.4, (b')-(c')-(d') imply

$$\mathcal{R}(\bar{\mu},\tau)^2 \geq 0.05 B_{i_0}^2(\bar{\mu}) \geq \frac{0.05}{4}\bar{R}^2(i_0+1)^{-2\alpha}.$$

For $i_0 = \lfloor C_0 \max\left((\bar{R}^2\delta^{-2}/\sqrt{D})^{1/(2\alpha+2p)}, (R^2\delta^{-2})^{1/(2\beta+2p+1)}\right)\rfloor$ with some suitably large constant $C_0 > 0$, depending only on C_μ, C_A, p, and for $D \geq 2i_0$, conditions (b') and (d') are satisfied. To check condition (c'), a sufficient condition is $i_0 + 1 \leq (\bar{R}^2\delta^{-2}/(56C_A^2))^{1/(2\alpha+2p)}$. The first term in the maximum defining i_0 satisfies this condition (here again using $D \geq 2i_0$) provided $R^{-1}\delta$ is smaller than a suitable constant c_2' depending on C_A, C_μ, α, p. The second term in the maximum defining i_0 satisfies the sufficient condition

$$C_0(R^2\delta^{-2})^{1/(2\beta+2p+1)} \leq C_0(\bar{R}^2\delta^{-2})^{1/(2\alpha+2p+1)} \leq (\bar{R}^2\delta^{-2}/(56C_A^2))^{1/(2\alpha+2p)},$$

again as soon as $R^{-1}\delta$ is smaller than a suitable constant c_2'' depending on the same parameters as c_2'. Finally, putting $c_2 = \min(c_2', c_2'', 1)$ and unwrapping the condition $D \geq 2i_0$, yields (using $R\delta^{-1} \geq 1$) the sufficient condition $D \geq c_3'(\bar{R}^2\delta^{-2})^{1/(2\alpha+2p+1/2)}$, which is equivalent to the assumption $D \geq c_3 t_{\alpha-\frac{1}{4},p,\bar{R}}(\delta)$ postulated in the statement, for suitable c_3', c_3 depending on C_μ, C_A, α, p. This yields the result. $\qquad\square$

5.5. Proof of Proposition 3.4

Proof. By partial summation, we deduce from $B_{\tau,\lambda}^2(\mu) > B_{t^*,\lambda}^2(\mu)$, which implies $\tau \leq \lfloor t^* \rfloor$:

$$\begin{aligned}
\mathbb{E}[(B_{\tau,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu))_+] &= \sum_{m=m_0}^{\lfloor t^* \rfloor} (B_{m,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu))P(\tau=m) \\
&= \sum_{m=m_0}^{\lfloor t^* \rfloor} \sum_{k=m}^{\lfloor t^* \rfloor} (B_{k,\lambda}^2(\mu) - B_{(k+1)\wedge t^*,\lambda}^2(\mu))P(\tau=m) \\
&= \sum_{m=m_0}^{\lfloor t^* \rfloor} (B_{m,\lambda}^2(\mu) - B_{(m+1)\wedge t^*,\lambda}^2(\mu))P(\tau \leq m).
\end{aligned}$$

In the case $t^* = m_0$ all expressions evaluate to zero because of $\tau \geq t^*$ and we suppose $t^* > m_0$ from now on, so that (3.4) holds. For $m_0 \leq m < t^*$ we have $\{\tau \leq m\} = \{R_m^2 \leq \kappa\}$, $\mathbb{E}[R_m^2] \geq \kappa$ and by Lemma 1 in [12] (recall (5.1) in the proof of Lemma 2.3) together with $P(Z < -x) \leq e^{-x^2/(2\sigma^2)}$ for $Z \sim N(0,\sigma^2)$, $x \geq 0$ we obtain the bound

$$P(R_m^2 \leq \kappa) = P\Big(\sum_{i=m+1}^{D} \left(\delta^2(\varepsilon_i^2-1) + 2\lambda_i\mu_i\delta\varepsilon_i\right) \leq -(\mathbb{E}[R_m^2] - \kappa)\Big)$$

$$\leq P\Big(\sum_{i=m+1}^{D} \delta^2(\varepsilon_i^2-1) \leq -\frac{\mathbb{E}[R_m^2] - \kappa}{2}\Big) + P\Big(\sum_{i=m+1}^{D} \lambda_i\mu_i\delta\varepsilon_i \leq -\frac{\mathbb{E}[R_m^2] - \kappa}{4}\Big)$$

$$\leq \exp\Big(-\frac{(\mathbb{E}[R_m^2]-\kappa)^2}{16\delta^4(D-m)}\Big) + \exp\Big(-\frac{(\mathbb{E}[R_m^2]-\kappa)^2}{32\delta^2 B_{m,\lambda}^2(\mu)}\Big)$$

$$\leq F(B_{m,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu)),$$

where we use $\mathbb{E}[R_m^2] - \kappa \geq B_{m,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu)$ for $m < t^*$ (see (3.3), (3.4)), and put

$$F(z) := \exp\Big(-\frac{z^2}{16\delta^4 D}\Big) + \exp\Big(-\frac{z^2}{32\delta^2(B_{t^*,\lambda}^2(\mu)+z)}\Big), \quad z \geq 0.$$

We conclude by monotonicity of $B_{\bullet,\lambda}^2(\mu)$ and F via a Riemann-Stieltjes sum approximation:

$$\mathbb{E}\left[\big(B_{\tau,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu)\big)_+\right]$$

$$\leq \sum_{m=m_0}^{\lfloor t^*\rfloor}(B_{m,\lambda}^2(\mu) - B_{(m+1)\wedge t^*,\lambda}^2(\mu))F(B_{m,\lambda}^2(\mu) - B_{t^*,\lambda}^2(\mu))$$

$$\leq \int_{B_{t^*,\lambda}^2(\mu)}^{B_{m_0,\lambda}^2(\mu)} F(y - B_{t^*,\lambda}^2(\mu))\,dy$$

$$\leq \int_0^\infty F(z)\,dz$$

$$\leq \sqrt{4\pi\delta^4 D} + \int_0^{B_{t^*,\lambda}^2(\mu)} e^{-z^2/(64\delta^2 B_{t^*,\lambda}^2(\mu))}\,dz + \int_{B_{t^*,\lambda}^2(\mu)}^\infty e^{-z/(64\delta^2)}\,dz$$

$$\leq \sqrt{4\pi\delta^4 D} + \sqrt{16\pi\delta^2 B_{t^*,\lambda}^2(\mu)} + 64\delta^2$$

$$\leq \big(17\sqrt{D} + 64\big)\delta^2 + B_{t^*,\lambda}^2(\mu)D^{-1/2},$$

using $4\sqrt{\pi}\delta B_{t^*,\lambda}(\mu) \leq 4\pi\sqrt{D}\delta^2 + D^{-1/2}B_{t^*,\lambda}^2(\mu)$ by the binomial identity and $\sqrt{4\pi} + 4\pi \leq 17$ in the last line. $\square$

5.6. Proof of Proposition 3.6

Proof. By the Cauchy-Schwarz inequality and $\mathbb{E}[\varepsilon_m^4]^{1/2} = \sqrt{3}$, we have

$$\mathbb{E}\left[(S_\tau - S_{\lfloor t^*\rfloor+1})_+\right] = \delta^2 \sum_{m=\lfloor t^*\rfloor+2}^D \lambda_m^{-2}\,\mathbb{E}[\varepsilon_m^2 \mathbf{1}(\tau \geq m)]$$

$$\leq \sqrt{3}\delta^2 \sum_{m=\lfloor t^*\rfloor+2}^D \lambda_m^{-2} P(\tau \geq m)^{1/2}.$$

For $m \geq t^*+1 \geq m_0+1$ we have $\{\tau \geq m\} = \{R_{m-1}^2 > \kappa\}$, $\mathbb{E}[R_{m-1}^2] \leq \kappa$ and by $P(\sum_{i=1}^D a_i(\varepsilon_i^2 - 1) > x) \leq \exp\big(-x^2/(4\|a\|^2 + 4x\max_i a_i)\big)$ for $a_i \geq 0$ (Lemma 1

in [12]) together with $P(Z > x) \leq e^{-x^2/(2\sigma^2)}$ for $Z \sim N(0, \sigma^2)$, $x \geq 0$, we obtain the bound

$$P(R^2_{m-1} > \kappa) = P\Big(\sum_{i=m}^{D} \big(\delta^2(\varepsilon_i^2 - 1) + 2\lambda_i \mu_i \delta \varepsilon_i \big) > \kappa - \mathbb{E}[R^2_{m-1}] \Big)$$

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

$$\leq \exp\Big(-\frac{(\kappa - \mathbb{E}[R^2_{m-1}])^2}{16\delta^4(D - m + 1) + 8\delta^2(\kappa - \mathbb{E}[R^2_{m-1}])} \Big) + \exp\Big(-\frac{(\kappa - \mathbb{E}[R^2_{m-1}])^2}{32\delta^2 B^2_{m-1,\lambda}(\mu)} \Big).$$

For the numerator, we use the lower bound

$$\kappa - \mathbb{E}[R^2_{m-1}] \geq \kappa - \mathbb{E}[R^2_{t^*}] + \delta^2(m - 1 - t^*) \geq \delta^2(m - 1 - t^*).$$

For the denominators, we use $16\delta^4(D-m+1)+8\delta^2(\kappa-\mathbb{E}[R^2_{m-1}]) \leq 16\delta^4 D + 8\delta^2 \kappa$ for the first term and $32\delta^2 B^2_{m-1,\lambda}(\mu) \leq 32\delta^2 B^2_{t^*,\lambda}(\mu) \leq 32\delta^2 \kappa$ for the second term. We arrive at

$$\mathbb{E}\Big[(S_\tau - S_{\lfloor t^* \rfloor + 1})_+ \Big] \leq 2\sqrt{3}\delta^2 \sum_{m=\lfloor t^* \rfloor + 2}^{D} \lambda_m^{-2} \exp\Big(-\frac{(m - 1 - t^*)^2}{16D + 32\kappa\delta^{-2}} \Big).$$

We add $\mathbb{E}[(S_{\lfloor t^* \rfloor + 1} - S_{t^*})_+] \leq 2\sqrt{3}\delta^2 \lambda_{\lfloor t^* \rfloor + 1}^{-2}$ and note $(S_\tau - S_{t^*})_+ \leq S_D$, which gives the trivial bound $\mathbb{E}[S_D] = D\delta^2$. Under the polynomial eigenvalue decay this yields the bound

$$r_{V,\tau} \leq 2\sqrt{3} C_A^2 \Big(1 + \sum_{k \geq 0} (t^* + 1 + k)^{2p} e^{-k^2/(16D + 32\kappa\delta^{-2})} \Big) \wedge D.$$

In the sequel '$\lesssim$', '$\gtrsim$' denote inequalities up to a factor only depending on p. A Riemann sum approximation shows for any $R > 0$

$$\sum_{k \geq 0} e^{-k^2/R} \leq 1 + \int_0^\infty e^{-x^2/R} dx \lesssim \sqrt{R}.$$

Similarly, we obtain

$$\sum_{k \geq 0} k^{2p} e^{-k^2/R} \leq \int_0^\infty (1 + x)^{2p} e^{-x^2/R} dx \lesssim R^{p+1/2}.$$

This yields

$$r_{V_\tau} \lesssim C_A^2 \Big((t^*)^{2p} \sqrt{D + \kappa\delta^{-2}} + (D + \kappa\delta^{-2})^{p+1/2} \Big).$$

On the other hand, we have

$$V_{t^*} = \delta^2 \Big(\sum_{m=1}^{\lfloor t^* \rfloor} \lambda_m^{-2} + (t^* - \lfloor t^* \rfloor)\lambda_{\lfloor t^* \rfloor + 1}^{-2} \Big) \gtrsim \delta^2 C_A^{-2} (t^*)^{2p+1},$$

implying the result with a suitable constant C_p. $\qquad\qquad\square$

5.7. **Proof of Proposition** 4.2

Proof. In this proof '$\lesssim$' denotes an inequality holding up to factors depending only on p, C_A, C_κ and C_2; similarly, '$\sim$' denotes a two-sided inequality holding up to factors depending on these parameters. In the case $t_{\mathfrak{s}} > m_0$ we use the independence of τ from $\widehat{\mu}^{(0)}, \ldots \widehat{\mu}^{(m_0)}$ by Lemma 4.1 to obtain

$$
\begin{aligned}
\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau = m_0)] &= \mathbb{E}[\|\widehat{\mu}^{(\widehat{m})} - \mu\|^2] P(\tau = m_0) \\
&\leq C_2 \Big(\mathbb{E}[\|\widehat{\mu}^{(m_0)} - \mu\|^2] + \delta^2 \Big) P(\tau = m_0) \\
&= C_2 \Big(\mathbb{E}[\|\widehat{\mu}^{(\tau)} - \mu\|^2 \mathbf{1}(\tau = m_0)] + \delta^2 P(\tau = m_0) \Big).
\end{aligned}
$$

On $\{\tau > m_0\}$ we have $\rho = \tau$ and we apply Theorem 3.8 with $t_{\mathfrak{s}} \geq m_0 > \sqrt{D}$ to get

$$
\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau > m_0)] \leq \mathbb{E}[\|\widehat{\mu}^{(\tau)} - \mu\|^2] \lesssim \mathbb{E}[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2].
$$

Because of $t_{\mathfrak{s}} > m_0$ we have $\mathbb{E}[\|\widehat{\mu}^{(t_{\mathfrak{s}})} - \mu\|^2] \leq 2 \min_{t \in [0,D]} \mathbb{E}[\|\widehat{\mu}^{(t)} - \mu\|^2]$. This gives the result in this case.

Next, consider the case $t_{\mathfrak{s}} = m_0$ where $B_{m_0}^2(\mu) \leq V_{m_0}$. Then the estimator $\widehat{\mu}^{(m_{\mathfrak{s}})}$ with $m_{\mathfrak{s}} \in \{0, 1, \ldots, m_0\}$ from (1.7) satisfies

$$
\mathbb{E}\left[\|\widehat{\mu}^{(m_{\mathfrak{s}})} - \mu\|^2\right] \leq 2 \max_i \frac{\lambda_i^2}{\lambda_{i+1}^2} \min_{m \in \{0, \ldots, D\}} \mathbb{E}\left[\|\widehat{\mu}^{(m)} - \mu\|^2\right],
$$

noting that the factor $\max_i \lambda_i^2 / \lambda_{i+1}^2$ comes from the discretisation $m_{\mathfrak{s}}$ of the balanced oracle and is bounded by $C_A^4 2^{2p} \lesssim 1$. Given the independence of τ from $\{\widehat{\mu}^{(0)}, \ldots, \widehat{\mu}^{(m_0)}\}$ by Lemma 4.1 and the properties of the model selector $\widehat{m}$, we have

$$
\begin{aligned}
E[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau = m_0)] &\leq C_2 \Big(\mathbb{E}[\|\widehat{\mu}^{(m_{\mathfrak{s}})} - \mu\|^2] + \delta^2 \Big) \\
&\lesssim \min_{m \in \{0, \ldots, D\}} \mathbb{E}[\|\widehat{\mu}^{(m)} - \mu\|^2] + \delta^2.
\end{aligned}
$$

For $m_{\mathfrak{s}} \in [m_0/2, m_0]$

$$
\mathbb{E}[\|\widehat{\mu}^{(m_0)} - \mu\|^2] \leq B_{m_{\mathfrak{s}}}^2(\mu) + V_{m_0} \lesssim \mathbb{E}[\|\widehat{\mu}^{(m_{\mathfrak{s}})} - \mu\|^2]
$$

follows from $V_{m_0} \sim \delta^2 m_0^{2p+1} \sim V_{m_{\mathfrak{s}}}$. By Theorem 3.8 with $t_{\mathfrak{s}} = m_0$ this gives

$$
\begin{aligned}
\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau > m_0)] &\lesssim \mathbb{E}[\|\widehat{\mu}^{(m_0)} - \mu\|^2] \lesssim \mathbb{E}[\|\widehat{\mu}^{(m_{\mathfrak{s}})} - \mu\|^2] \\
&\lesssim \min_{m \in \{0, \ldots, D\}} \mathbb{E}[\|\widehat{\mu}^{(m)} - \mu\|^2].
\end{aligned}
$$

For $m_{\mathfrak{s}} < m_0/2$ we obtain by $\{\tau > m_0\} = \{R_{m_0}^2 > \kappa\}$, $S_\tau \leq S_D$, $B_\tau^2(\mu) \leq B_{m_0}^2(\mu)$ and the Cauchy-Schwarz inequality:

$$
\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau > m_0)] \leq \mathbb{E}[(B_{m_0}^2(\mu) + S_D) \mathbf{1}(R_{m_0}^2 > \kappa)]
$$

$$\leq B_{m_0}^2(\mu)P(R_{m_0}^2 > \kappa) + \mathbb{E}[S_D^2]^{1/2}P(R_{m_0}^2 > \kappa)^{1/2}.$$

We have $B_{m_0}^2(\mu) \leq B_{m_\mathfrak{s}}^2(\mu) \leq V_D$ and $\mathbb{E}[S_D^2]^{1/2} \lesssim \mathbb{E}[S_D] = V_D$ (by comparison of Gaussian moments), so that

$$\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau > m_0)] \lesssim V_D P(R_{m_0}^2 > \kappa)^{1/2} \sim \delta^2 D P(R_{m_0}^2 > \kappa)^{1/2}.$$

Observing $m_\mathfrak{w} := \min\{m \geq 0 \,|\, B_{m,\lambda}^2(\mu) \leq V_{m,\lambda}\} \leq m_\mathfrak{s} < m_0/2$ we obtain

$$\mathbb{E}[R_{m_0}^2] - \kappa = B_{m_0,\lambda}^2(\mu) - V_{m_0,\lambda} \leq B_{m_\mathfrak{w},\lambda}^2(\mu) - m_0\delta^2 \leq -(m_0/2)\delta^2.$$

As in the proof of Proposition 3.6 we therefore find

$$P(R_{m_0}^2 > \kappa) \leq \exp\Big(-\frac{m_0^2}{64(D-m_0)+16m_0}\Big) + \exp\Big(-\frac{m_0^2\delta^2}{128B_{m_0,\lambda}^2(\mu)}\Big).$$

By the choice of m_0 and $B_{m_0,\lambda}^2(\mu) \leq B_{m_0/2,\lambda}^2(\mu) \leq (m_0/2)\delta^2$, using $D \geq 3 \Rightarrow \log D \geq 1$, we deduce

$$P(R_{m_0}^2 > \kappa) \leq 2\exp\big(-2\log D\big) = 2D^{-2}.$$

Insertion of this bound yields $\mathbb{E}[\|\widehat{\mu}^{(\rho)} - \mu\|^2 \mathbf{1}(\tau > m_0)] \lesssim \delta^2$, which accomplishes the proof for the case $t_\mathfrak{s} = m_0$ and $m_\mathfrak{s} \leq m_0/2$. □

Acknowledgement

References

[1] N. Bissantz, T. Hohage, A. Munk, and F. Ruymgaart. Convergence rates of general regularization methods for statistical inverse problems and applications. *SIAM Journal on Numerical Analysis*, 45:2610–2636, 2007. MR2361904

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

[3] G. Blanchard and P. Mathé. Discrepancy principle for statistical inverse problems with application to conjugate gradient iteration. *Inverse Problems*, 28:pp. 115011, 2012. MR2992966

[4] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007. MR2420454

[5] L. Cavalier. Inverse problems in statistics. In *Inverse problems and high-dimensional estimation*, pages 3–96. Lecture Notes in Statistics 203, Springer, 2011. MR2868199

[6] L. Cavalier, G.K. Golubev D. Picard, and A.B. Tsybakov. Oracle inequalities for inverse problems. *Annals of Statistics*, 30:843–874, 2002. MR1922543

[7] L. Cavalier and Y. Golubev. Risk hull method and regularization by projections of ill-posed inverse problems. *Annals of Statistics*, 34:1653–1677, 2006. MR2283712

[8] E. Chernousova, Y. Golubev, and E. Krymova. Ordered smoothers with exponential weighting. *Electronic Journal of Statistics*, 7:2395–2419, 2013. MR3108818

[9] A. Cohen, M. Hoffmann, and M. Reiß. Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM Journal on Numerical Analysis*, 42(4):1479–1501, 2004. MR2114287

[10] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, London, 1996. MR1408680

[11] Y. Ingster and I. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Lecture Notes in Statistics 169, Springer, 2012. MR1991446

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

[13] O. Lepski. Some new ideas in nonparametric estimation. *arXiv:1603.03934*, 2016.

[14] F. Lucka, K. Proksch, C. Brune, N. Bissantz, M. Burger, H. Dette, and F. Wübbeling. Risk estimators for choosing regularization parameters in ill-posed problems - properties and limitations. e-print, arXiv:1701.04970, 2017. MR3836582

[15] P. Mathé and S. V. Pereverzev. Geometry of linear ill-posed problems in variable hilbert scales. *Inverse problems*, 19(3):789, 2003. MR1984890

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

[17] Y. Saad. *Numerical Methods for Large Eigenvalue Problems: Revised Edition*. Society for Industrial and Applied Mathematics (SIAM), 2011. MR3396212

[18] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. MR2724359

[19] G. Wahba. Practical approximate solutions to linear operator equations when the data are noisy. *SIAM Journal on Numerical Analysis*, 14(4):651–667, 1977. MR0471299

[20] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. MR2327601