Supplementary material for " Optimal learning rates for Kernel Conjugate Gradient regression"

Gilles Blanchard
Mathematics Institute, University of Potsdam
Am neuen Palais 10, 14469 Potsdam
blanchard@math.uni-potsdam.de

Nicole Krämer
Weierstrass Institute
Mohrenstr. 39, 10117 Berlin, Germany
nicole.kraemer@wias-berlin.de

Abstract

This supplementary material contains the proof for Theorem 2.2 and a sketch of the proof of Theorem 2.3 of the main paper.

A Notation

We follow the notation used in the main part, in particular the operators T_n, T_n^*, T, T^* defined in Section 4.1, and we recall that $S_n := T_n^* T_n$; $S = T^* T$; $K_n = T_n T_n^*$; and $K = TT^*$. We define the auxiliary notation $\mathcal{N}(\lambda) := \mathrm{Tr}(K(K+\lambda I)^{-1})$, which is the function entering on the LHS of the **ED** condition.

We denote by $(\xi_i)_{i\geq 1}$ the possibly finite sequence in $[0,\kappa]$ of nonzero eigenvalues of S and K, and by $(\xi_{j,n})_{1\leq j\leq n}$ the n-sequence of eigenvalues of S_n and K_n respectively (in each case in decreasing order and with multiplicity). Finally, $(F_u)_{u\geq 0}$ denotes the spectral family of the operator S_n, i.e. F_u is the orthogonal projector on the subspace of $\mathcal{H}$ spanned by eigenvectors of S_n corresponding to eigenvalues strictly less than u.

It is useful to consider the spectral integral representation: If $(e_{i,n})_{1\leq i\leq n}$ denotes the orthogonal eigensystem of S_n associated to the non-zero eigenvalues $(\lambda_{i,n})_{1\leq i\leq n}$, for any integrable function h on $[0,\kappa]$, we set

$$\int_0^\kappa h(u) d\left\|F_{u,n} T_n^* \mathbf{Y}\right\|^2 := \langle T_n^* \mathbf{Y}, h(S_n) T_n^* \mathbf{Y}\rangle = \sum_{i=1}^n h(\lambda_{i,n}) \left\langle T_n^* \mathbf{Y}, e_{i,n}\right\rangle^2 .$$

By its definition, the output of the m-th iteration of the CG algorithm can be put under the form $f_m = q_m(S_n) T_n^* \mathbf{Y}$, where $q_m \in \mathcal{P}_{m-1}$, the set of real polynomials of degree less than $m-1$. A crucial role is played by the *residual polynomial*

$$p_m(x) = 1 - x q_m(x) \in \mathcal{P}_m^0 ,$$

where $\mathcal{P}_m^0$ is the set of real polynomials of degree less than m and having constant term equal to 1. In particular $T_n^* \mathbf{Y} - S_n f_m = p_m(S_n) T_n^* \mathbf{Y}$. Furthermore, the definition of the CG algorithm implies that the sequence $(p_m)_{m\geq 0}$ are orthogonal polynomials for the scalar product $[.,.]_{(1)}$, where for $i \geq 0$ we define

$$[p,q]_{(i)} := \left\langle p(S_n) T_n^* \mathbf{Y}, S_n^i q(S_n) T_n^* \mathbf{Y}\right\rangle = \int_0^\kappa p(u) q(u) u^i d\left\|F_{u,n} T_n^* \mathbf{Y}\right\|^2 .$$

This can be shown as follows: p_m is the orthogonal projection, of the origin onto the affine space $\mathcal{P}_m^0 = 1 + x\mathcal{P}_{m-1}$ with the scalar product $[.,.]_{(0)},$, where $x\mathcal{P}_{m-1}$ denotes (with some abuse of notation) the set of polynomials of degree less than m with constant coefficient equal to zero. Thus $0 = [p_m, xq]_{(0)} = [p_m, q]_{(1)}$ for any $q \in \mathcal{P}_{m-1}$. From the theory of orthogonal polynomials, it results that for any $m \leq m_{final} := \#\{i : 1 \leq i \leq n, \xi_{i,n} \langle T_n^* \mathbf{Y}, e_{i,n} \rangle \neq 0\}$, the polynomial p_m has exactly m distinct roots belonging to $[0, \kappa]$, which we denote by $(x_{k,m})_{1 \leq k \leq m}$ (in increasing order). Finally, we use the notation $c(a, b)$ to denote a function depending on the stated parameters only, and whose exact value can change from line to line.

B Preparation of the proof

We follow the general architecture of Nemirovskii's proof to establish rates. We recall that since we assume $r \geq 1/2$, the representation $f^* = T f_{\mathcal{H}}^*$ holds. The main difference to Nemirovskii's original result is that (similar to the approach of [2, 3]) we use deviation bounds in a "warped" norm rather than in the standard norm. More precisely, we consider the following type of assumptions:

B1(λ) $\left\| (S + \lambda I)^{-\frac{1}{2}} (T_n^* \mathbf{Y} - S_n f_{\mathcal{H}}^*) \right\| \leq \delta(\lambda)$,

B2(λ) $\left\| (S + \lambda I)(S_n + \lambda I)^{-1} \right\| \leq \Lambda^2$, with $\Lambda \geq 1$
(this implies in particular $\left\| (S + \lambda I)^{\frac{1}{2}} (S_n + \lambda I)^{-\frac{1}{2}} \right\| \leq \Lambda$ via (20) below),

B3 $\|S - S_n\| \leq \kappa \Delta$.

In the rest of this section we set $\mu = r - 1/2$. Under the source condition assumption **SC**(r), for $r \geq \frac{1}{2}$ the representation $f^* = K^r u$ can be rewritten

$$f^* = (TT^*)^r u = T(T^*T)^{r-\frac{1}{2}} (T^*T)^{-\frac{1}{2}} T^* u = T S^\mu (T^*T)^{-\frac{1}{2}} T^* u,$$

by identification we therefore have the source condition for $f_{\mathcal{H}}$ given by $f_{\mathcal{H}} = S^\mu w$ with $w = (T^*T)^{-\frac{1}{2}} T^* u$, and $\|w\|_{\mathcal{H}} \leq \|u\|$, since $(T^*T)^{-\frac{1}{2}} T^*$ is a restricted isometry from $L_2(P_X)$ into $\mathcal{H}$.

We define the shortcut notation

$$Z_\mu(\lambda) = \begin{cases} \lambda^\mu & \text{for } \mu \leq 1, \\ \kappa^\mu \Delta & \text{for } \mu > 1. \end{cases} \tag{1}$$

We start with preliminary technical lemmas, before turning to the proof of Theorem 2.2.

Lemma B.1. *For any $\lambda > 0$, if assumptions* **SC**(r)*,* **B1**(λ)*,* **B2**(λ) *and* **B3** *hold, then for any iteration step $1 \leq m \leq m_{final}$*

$$\begin{aligned} \|T_n^*(T_n f_m - \mathbf{Y})\| \leq & c(\mu)\Lambda^2 \left(|p_m'(0)|^{-(\mu+1)} + Z_\mu(\lambda) \, |p_m'(0)|^{-1} \right) \kappa^{-\mu-\frac{1}{2}} \rho \\ & + \left(|p_m'(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}} \right) \Lambda \delta(\lambda). \end{aligned} \tag{2}$$

Proof. Recall that $(x_{k,m})_{1 \leq k \leq m}$ denote the m roots of the polynomial p_m; define further the function φ_m on the interval $[0, x_{1,m}]$ as

$$\varphi_m(x) = p_m(x) \left(\frac{x_{1,m}}{x_{1,m} - x} \right)^{\frac{1}{2}}$$

Following the idea introduced by Nemirovski, it can be shown that

$$\begin{aligned} \|T_n^*(T_n f_m - \mathbf{Y})\| &= \|p_m(S_n) T_n^* \mathbf{Y}\| \\ &\leq \left\| F_{x_{1,m}} \varphi_m(S_n) T_n^* \mathbf{Y} \right\| \\ &\leq \left\| F_{x_{1,m}} \varphi_m(S_n) S_n f_{\mathcal{H}}^* \right\| + \left\| F_{x_{1,m}} \varphi_m(S_n) (T_n^* \mathbf{Y} - S_n f_{\mathcal{H}}^*) \right\| := (I) + (II). \end{aligned}$$

Above, the first inequality (lemma 3.7. in [4]) is the crucial point, and relies fundamentally on the fact that (p_m) is an orthogonal polynomial sequence.

We start with controlling the second term:

$$\begin{aligned}(II) = \left\|F_{x_{1,m}}\varphi_m(S_n)(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\right\| &= \left\|F_{x_{1,m}}\varphi_m(S_n)(S+\lambda I)^{\frac{1}{2}}(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\right\| \\ &\leq \left\|F_{x_{1,m}}\varphi_m(S_n)(S_n+\lambda I)^{\frac{1}{2}}\right\| \Lambda\delta(\lambda) \\ &\leq \left(\sup_{x\in[0,x_{1,m}]} x^{\frac{1}{2}}\varphi_m(x) + \lambda^{\frac{1}{2}} \sup_{x\in[0,x_{1,m}]} \varphi_m(x)\right)\Lambda\delta(\lambda) \\ &\leq \left(|p_m'(0)|^{-\frac{1}{2}} + \lambda^{\frac{1}{2}}\right)\Lambda\delta(\lambda)\,,\end{aligned}$$

where the last line used the inequality (see (3.10) in [4])

$$\sup_{x\in[0,x_{1,m}]} x^\nu \varphi_m^2(x) \leq \nu^\nu \left|p_m'(0)\right|^{-\nu}\,, \tag{3}$$

for any $\nu \geq 0$ (using the convention $0^0 = 1$), which we applied above for $\nu = 0, 1$. For the first term, we use assumption **SC**(r); first consider the case $\mu > 1$:

$$\begin{aligned}(I) = \left\|F_{x_{1,m}}\varphi_m(S_n)S_n f_{\mathcal{H}}^*\right\| &= \left\|F_{x_{1,m}}\varphi_m(S_n)S_n S^\mu w\right\| \\ &\leq \left(\left\|F_{x_{1,m}}\varphi_m(S_n)S_n^{\mu+1}\right\| + \left\|F_{x_{1,m}}\varphi_m(S_n)S_n\right\| \left\|S^\mu - S_n^\mu\right\|\right)\kappa^{-\mu-\frac{1}{2}}\rho \\ &\leq c(\mu)\left(\left|p_m'(0)\right|^{-(\mu+1)} + \kappa^\mu\Delta\left|p_m'(0)\right|^{-1}\right)\kappa^{-\mu-\frac{1}{2}}\rho\,,\end{aligned}$$

where we applied (3) with $\nu = 2(\mu+1)$, $\nu = 2$ and (19). □

For the case $\mu \leq 1$, using (20) and arguments similar to the previous case:

$$\begin{aligned}(I) &= \left\|F_{x_{1,m}}\varphi_m(S_n)S_n f_{\mathcal{H}}^*\right\| \\ &= \left\|F_{x_{1,m}}\varphi_m(S_n)S_n S^\mu w\right\| \\ &\leq \left\|F_{x_{1,m}}\varphi_m(S_n)S_n(S_n+\lambda I)^\mu\right\| \left\|(S_n+\lambda I)^{-\mu}(S+\lambda I)^\mu\right\| \left\|(S+\lambda)^{-\mu}S^\mu\right\| \kappa^{-\mu-\frac{1}{2}}\rho \\ &\leq c(\mu)\Lambda^2\left(\left|p_m'(0)\right|^{-(\mu+1)} + \lambda^\mu\left|p_m'(0)\right|^{-1}\right)\kappa^{-\mu-\frac{1}{2}}\rho\,.\end{aligned}$$

Lemma B.2. *For any $\lambda > 0$, if assumptions* **SC**(r)*,* **B1**(λ)*,* **B2**(λ) *and* **B3** *hold, then for any iteration step $1 \leq m \leq m_{final}$, for any $\varepsilon \in (0, x_{1,m})$:*

$$\begin{aligned}\|T(f_m - f_{\mathcal{H}}^*)\| \leq \Lambda\Bigg(&3\left(1+\lambda\left(|p_m'(0)| + \varepsilon^{-1}\right)\right)\Lambda\delta(\lambda) + c(\mu)\Lambda^2\left(\varepsilon^{\frac{1}{2}} + \lambda^{\frac{1}{2}}\right)\left(\varepsilon^\mu + Z_\mu(\lambda)\right)\kappa^{-\mu-\frac{1}{2}}\rho \\ &+ \sqrt{2}\left(1 + \frac{\lambda^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}}\right)\varepsilon^{-\frac{1}{2}}\left\|T_n^*(T_n f_m - \mathbf{Y})\right\|\Bigg)\end{aligned}$$

If $m = 0$, the above inequality is valid for any $\varepsilon > 0$.

Proof. Set $\bar{f}_m = q_m(S_n)S_n f_{\mathcal{H}}^*$. This is the element in $\mathcal{H}$ that we obtain by applying the mth-iteration CG polynomial q_m to the *noiseless* data. We have

$$\begin{aligned}\|T(f_m - f_{\mathcal{H}}^*)\| = \left\|S^{\frac{1}{2}}(f_m - f_{\mathcal{H}}^*)\right\| &\leq \Lambda\left\|(S_n+\lambda I)^{\frac{1}{2}}(f_m - f_{\mathcal{H}}^*)\right\| \\ &\leq \Lambda\left(\left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}}(f_m - \bar{f}_m)\right\| + \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}}(\bar{f}_m - f_{\mathcal{H}}^*)\right\|\right. \\ &\qquad \left. + \left\|F_\varepsilon^\perp(S_n+\lambda I)^{\frac{1}{2}}(f_m - f_{\mathcal{H}}^*)\right\|\right) \\ &:= \Lambda((I) + (II) + (III))\,,\end{aligned}$$

where we denote $F_\varepsilon^\perp := (I - F_\varepsilon)$. *First summand:*

$$\begin{aligned}(I) = \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}}(f_m - \bar{f}_m)\right\| &= \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} q_m(S_n)(S+\lambda I)^{\frac{1}{2}}(S+\lambda I)^{-\frac{1}{2}}\left(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*\right)\right\| \\ &\le \Lambda \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} q_m(S_n)(S_n+\lambda I)^{\frac{1}{2}}\right\| \delta(\lambda) \\ &\le \Lambda\delta(\lambda)\left(\sup_{x\in[0,\varepsilon]} x q_m(x) + \lambda \sup_{x\in[0,\varepsilon]} q_m(x)\right) \\ &\le \Lambda\delta(\lambda)\left(1 + \lambda\left|p_m'(0)\right|\right).\end{aligned}$$

The last inequality is obtained by the following argument: if $m \ge 1$, since $\varepsilon \le x_{1,m}$, p_m is convex in $[0,\varepsilon]$, we have

$$q_m(x) = \frac{1-p_m(x)}{x} \le |p_m'(0)| \qquad \text{for } x \in [0,\varepsilon];$$

and also $x q_m(x) = 1 - p_m(x) \le 1$ for $x \in [0,\varepsilon]$. If $m = 0$, we have $p_0 \equiv 1$ and $q_m \equiv 0$, so that the above is also trivially satisfied for any x.

Second summand: first subcase, $\mu > 1$, using (19), and the fact that $|p_m|(x) \le 1$ for $x \in [0,\varepsilon]$:

$$\begin{aligned}(II) = \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}}(\bar{f}_m - f_{\mathcal{H}}^*)\right\| &= \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} p_m(S_n) S^\mu w\right\| \\ &\le \left(\left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} p_m(S_n) S_n^\mu\right\| + \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} p_m(S_n)\right\| c(\mu)\kappa^\mu\Delta\right)\kappa^{-\mu-\frac{1}{2}}\rho \\ &\le \left(\varepsilon^{\mu+\frac{1}{2}} + \lambda^{\frac{1}{2}}\varepsilon^\mu + c(\mu)\kappa^\mu\left(\varepsilon^{\frac{1}{2}} + \lambda^{\frac{1}{2}}\right)\Delta\right)\kappa^{-\mu-\frac{1}{2}}\rho \\ &\le c(\mu)\left(\varepsilon^{\frac{1}{2}} + \lambda^{\frac{1}{2}}\right)\left(\varepsilon^\mu + \kappa^\mu\Delta\right)\kappa^{-\mu-\frac{1}{2}}\rho\,.\end{aligned}$$

Bounding the second summand: second subcase, $\mu \le 1$:

$$\begin{aligned}(II) = \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}} p_m(S_n) S^\mu w\right\| &\le \left\|F_\varepsilon(S_n+\lambda I)^{\mu+\frac{1}{2}} p_m(S_n)\right\| \Lambda^2\kappa^{-\mu-\frac{1}{2}}\rho \\ &\le c(\mu)(\varepsilon+\lambda)^{\mu+\frac{1}{2}}\Lambda^2\kappa^{-\mu-\frac{1}{2}}\rho\,.\end{aligned}$$

Third summand:

$$\begin{aligned}(III) = \left\|F_\varepsilon^\perp(S_n+\lambda I)^{\frac{1}{2}}(f_m - f_{\mathcal{H}}^*)\right\| &\le \left\|F_\varepsilon^\perp S_n^{\frac{1}{2}}(f_m - f_{\mathcal{H}}^*)\right\| + \lambda^{\frac{1}{2}}\left\|F_\varepsilon^\perp(f_m - f_{\mathcal{H}}^*)\right\| \\ &\le \left(\frac{(\varepsilon+\lambda)^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}} + \lambda^{\frac{1}{2}}\frac{(\varepsilon+\lambda)^{\frac{1}{2}}}{\varepsilon}\right)\left\|F_\varepsilon^\perp(S_n+\lambda I)^{-\frac{1}{2}} S_n(f_m - f_{\mathcal{H}}^*)\right\| \\ &\le \left(1 + \frac{\lambda^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}}\right)\left(1+\frac{\lambda}{\varepsilon}\right)^{\frac{1}{2}}\left\|F_\varepsilon^\perp(S_n+\lambda I)^{-\frac{1}{2}} S_n(f_m - f_{\mathcal{H}}^*)\right\| \\ &\le \left(1 + \frac{\lambda^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}}\right)\left(1+\frac{\lambda}{\varepsilon}\right)^{\frac{1}{2}}\left(\left\|F_\varepsilon^\perp(S_n+\lambda I)^{-\frac{1}{2}} T_n^*(T_n f_m - \mathbf{Y})\right\|\right. \\ &\qquad \left. + \left\|(S_n+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\right\|\right) \\ &\le \left(1 + \frac{\lambda^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}}\right)\varepsilon^{-\frac{1}{2}}\left\|T_n^*(T_n f_m - \mathbf{Y})\right\| + \sqrt{2}\Lambda\left(1+\frac{\lambda}{\varepsilon}\right)\left\|(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\right\| \\ &\le \left(1 + \frac{\lambda^{\frac{1}{2}}}{\varepsilon^{\frac{1}{2}}}\right)\varepsilon^{-\frac{1}{2}}\left\|T_n^*(T_n f_m - \mathbf{Y})\right\| + \sqrt{2}\Lambda\left(1+\frac{\lambda}{\varepsilon}\right)\delta(\lambda)\,.\end{aligned}$$

□

We now consider the sequence of polynomials that are orthogonal with respect to the scalar product $[.,.]_{(2)}$, which we denote by $p_m^{(2)}$, and its roots by $x_m^{(2)}$.

Lemma B.3. *For any $\lambda > 0$, if assumptions* **SC**(r)*,* **B1**(λ)*,* **B2**(λ) *and* **B3** *hold, then for any iteration step $1 \leq m \leq m_{final}$, for any $\varepsilon \in (0, x_{1,m-1})$:*

$$[p_{m-1}, p_{m-1}]_{(0)}^{\frac{1}{2}} = \|p_{m-1}(S_n)T_n^*\mathbf{Y}\|$$
$$\leq \Lambda(\varepsilon+\lambda)^{\frac{1}{2}}\delta(\lambda) + c(\mu)\Lambda^2\varepsilon\left(\varepsilon^\mu + Z_\mu(\lambda)\right)\kappa^{-\mu-\frac{1}{2}}\rho + \varepsilon^{-\frac{1}{2}}\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}^{\frac{1}{2}} . \quad (4)$$

Proof. By the optimality property defining our CG algorithm,

$$\|p_{m-1}(S_n)T_n^*\mathbf{Y}\| \leq \left\|p_{m-1}^{(2)}(S_n)T_n^*\mathbf{Y}\right\| \leq \left\|F_\varepsilon p_{m-1}^{(2)}(S_n)T_n^*\mathbf{Y}\right\| + \left\|F_\varepsilon^\perp p_{m-1}^{(2)}(S_n)T_n^*\mathbf{Y}\right\|$$
$$\leq \|F_\varepsilon T_n^*\mathbf{Y}\| + \varepsilon^{-\frac{1}{2}}\left\|p_{m-1}^{(2)}(S_n)S_n^{\frac{1}{2}}T_n^*\mathbf{Y}\right\| = \|F_\varepsilon T_n^*\mathbf{Y}\| + \varepsilon^{-\frac{1}{2}}\left[p_{m-1}^{(2)}, p_{m-1}^{(2)}\right]_{(1)}^{\frac{1}{2}}$$

For the last inequality, we have used the fact that $|p_{m-1}^{(2)}|(x) \leq 1$ for $x \in [0, x_{m-1}^{(2)}]$, along with the assumption $0 < \varepsilon < x_{1,m-1} \leq x_{1,m-1}^{(2)}$; the latter inequality is due to interlacing properties of the roots of orthogonal polynomials for $[.,.]_{(i)}$ and $[.,.]_{(i+1)}$ (see [4], Cor 2.7). We now bound

$$\|F_\varepsilon T_n^*\mathbf{Y}\| \leq \|F_\varepsilon(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\| + \|F_\varepsilon S_n S^\mu w\|$$
$$\leq \left\|F_\varepsilon(S_n+\lambda I)^{\frac{1}{2}}\right\| \left\|(S_n+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - S_n f_{\mathcal{H}}^*)\right\| + \|F_\varepsilon S_n S^\mu w\|$$
$$\leq \Lambda(\varepsilon+\lambda)^{\frac{1}{2}}\delta(\lambda) + \|F_\varepsilon S_n S^\mu w\| ;$$

for the second term, we divide as usual into two cases: for $\mu > 1$:

$$\|F_\varepsilon S_n S^\mu w\| \leq \left\|F_\varepsilon S_n^{\mu+1} w\right\| + \|F_\varepsilon S_n(S_n^\mu - S^\mu)w\| \leq \varepsilon c(\mu)\left(\varepsilon^\mu + \kappa^\mu\Delta\right)\kappa^{-\mu-\frac{1}{2}}\rho ,$$

and for $\mu \leq 1$:

$$\|F_\varepsilon S_n S^\mu w\| \leq \|F_\varepsilon S_n(S_n+\lambda I)^\mu\|\,\Lambda^2\kappa^{-\mu-\frac{1}{2}}\rho \leq \varepsilon(\varepsilon^\mu + \lambda^\mu)\Lambda^2\kappa^{-\mu-\frac{1}{2}}\rho .$$

□

C Proof of Theorem 2.2

We fix

$$\lambda_* = \left(\left(4D/\sqrt{n}\right)\log\left(6/\gamma\right)\right)^{\frac{2}{2\mu+s+1}}\kappa . \quad (5)$$

and assume n is big enough to ensure $\lambda_* \leq \kappa$. Furthermore we denote $\widetilde{\lambda}_* = \kappa^{-1}\lambda_*$ (this normalization was introduced in [2]).

We rewrite equivalently the discrepancy stopping rule as follows: for some fixed $\tau > 0$,

$$\widehat{m} := \min\left\{0 \geq m : \|T_n^*(T_n f_m - \mathbf{Y})\| \leq (2+\tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\right\} , \quad (6)$$

where

$$\delta(\lambda_*) := \frac{3}{4}M\widetilde{\lambda}_*^{\mu+\frac{1}{2}} . \quad (7)$$

(Observe that the above $\tau > 0$ is deduced from the constant $\tau' > 3/2$ considered in the main part of the paper via $\tau = \frac{4}{3}(\tau - \frac{3}{2})$.)

We first check **B1**(λ_*), **B2**(λ_*) and **B3** are satisfied simultaneously with large probability, using for this concentration results which are recalled in Section E. Concerning **B1**(λ_*), inequality (17) ensures that with probability $1-\gamma$, we have

$$\begin{aligned}\left\|(S+\lambda_* I)^{-\frac{1}{2}}(T_n^*\mathbf{Y}-S_n f_{\mathcal{H}}^*)\right\| &\leq 2M\left(\sqrt{\frac{\mathcal{N}(\lambda_*)}{n}}+\frac{2\sqrt{\kappa}}{\sqrt{\lambda_* n}}\right)\log\frac{6}{\gamma}\\ &\leq \frac{2M}{\sqrt{n}}D\widetilde{\lambda}_*^{-\frac{s}{2}}\left(1+\frac{1}{2D^2}\left(\frac{4D}{\sqrt{n}}\log\frac{6}{\gamma}\right)\widetilde{\lambda}_*^{\frac{s-1}{2}}\right)\log\frac{6}{\gamma}\\ &\leq \frac{M}{2}\widetilde{\lambda}_*^{\mu+\frac{1}{2}}\left(1+\frac{1}{2D^2}\widetilde{\lambda}_*^{\mu+s}\right)\\ &\leq \frac{3}{4}M\widetilde{\lambda}_*^{\mu+\frac{1}{2}}=\delta(\lambda_*)\,, \end{aligned} \tag{8}$$

where we have used **SC**(r), (5) and the assumptions $D \geq 1$ and $\widetilde{\lambda}_* \leq 1$. We now turn to **B2**($\lambda_*$). Inequality (18) along with a repetition of the above reasoning yields that with probability $1-\gamma$:

$$\left\|(S+\lambda_* I)^{-\frac{1}{2}}(S_n-S)\right\|_{HS} \leq \frac{\sqrt{\kappa}}{M}\delta(\lambda_*)\,,$$

so that

$$\left\|(S+\lambda_* I)^{-\frac{1}{2}}(S_n-S)(S+\lambda_* I)^{-\frac{1}{2}}\right\| \leq \frac{\sqrt{\kappa}}{M}\lambda_*^{-\frac{1}{2}}\delta(\lambda_*)\,.$$

Observe that

$$\frac{\sqrt{\kappa}}{M}\lambda_*^{-\frac{1}{2}}\delta(\lambda_*)=\frac{3}{4}\widetilde{\lambda}_*^{\mu}\leq\frac{3}{4}\,, \tag{9}$$

so that with Lemma E.2, we obtain that **B2**(λ_*) is satisfied with $\Lambda := 2$ (with probability $1-\gamma$). Finally, equation (11) in the main paper implies that **(B3)** is also satified with probability $1-\gamma$, with

$$\Delta := \frac{2}{\sqrt{n}}\log\frac{1}{\gamma}\,. \tag{10}$$

To conclude, by the union bound, the event that **B1**(λ_*), **B2**(λ_*) and **B3** satisfied simultaneously has probability larger than $1-3\gamma$, and we assume for the rest of the proof that we are on this event.

We will assume $\widehat{m} \geq 1$ for the remainder of the proof and postpone to the end the (simpler) case $\widehat{m}=0$.

First step: upper bound on $\left|p'_{\widehat{m}-1}(0)\right|$. By definition of the stopping rule we have $\|T_n^*(T_n f_{\widehat{m}-1}-\mathbf{Y})\| > (2+\tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)$. Now applying this together with the upper bound of Lemma B.1 we get

$$\begin{aligned}\tau\lambda_*^{\frac{1}{2}}\delta(\lambda_*) &\leq c(\mu)\left(\left|p'_{\widehat{m}-1}(0)\right|^{-(\mu+1)}+Z_\mu(\lambda_*)\left|p'_{\widehat{m}-1}(0)\right|^{-1}\right)\kappa^{-\mu-\frac{1}{2}}\rho+2\left|p'_{\widehat{m}-1}(0)\right|^{-\frac{1}{2}}\delta(\lambda_*)\\ &\leq 3\max\Big(2\left|p'_{\widehat{m}-1}(0)\right|^{-\frac{1}{2}}\delta(\lambda_*), c(\mu)\rho\kappa^{-\mu-\frac{1}{2}}\left|p'_{\widehat{m}-1}(0)\right|^{-(\mu+1)},\\ &\qquad c(\mu)\rho\kappa^{-\mu-\frac{1}{2}}Z_\mu(\lambda_*)\left|p'_{\widehat{m}-1}(0)\right|^{-1}\Big)\,.\end{aligned}$$

We examine in succession the possibility that the maximum in the above expression is attained for each of the terms which comprise it. If the first term attains the maximum, this implies $|p'_{\widehat{m}-1}(0)| \leq (9/\tau^2)\lambda_*^{-1}$. If the second term attains the maximum, this entails

$$c(\mu)\rho\kappa^{-\mu-\frac{1}{2}}\left|p'_{\widehat{m}-1}(0)\right|^{-(\mu+1)} \geq \tau\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\,,$$

which using (7) yields:

$$\left|p'_{\widehat{m}-1}(0)\right| \leq c(\mu,\tau)\left(\frac{\rho}{M}\right)^{\frac{1}{\mu+1}}\lambda_*^{-1}\,.$$

Finally, if the third term attains the maximum, we have

$$c(\mu)\rho Z_\mu(\lambda_*)\kappa^{-\mu-\frac{1}{2}}\left|p'_{\widehat{m}-1}(0)\right|^{-1} \geq \tau\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\,,$$

which using (7) yields:

$$\left|p'_{\widehat{m}-1}(0)\right| \leq c(\mu,\tau)\frac{\rho}{M}\lambda_*^{-\mu-1}Z_\mu(\lambda_*)\,.$$

We now establish the inequality

$$Z_\mu(\lambda_*)\lambda_*^{-\mu} \leq 1\,. \tag{11}$$

The inequality is trivial if $\mu \leq 1$ given the definition of $Z_\mu(\lambda_*)$ in (1). If $\mu > 1$ holds, from the definition (10), it holds that $\Delta \leq \frac{1}{2}\widetilde{\lambda}_*^{\frac{2\mu+s+1}{2}}$, hence

$$Z_\mu(\lambda_*)\lambda_*^{-\mu} = \Delta\widetilde{\lambda}_*^{-\mu} \leq \frac{1}{2}\widetilde{\lambda}_*^{\frac{s+1}{2}} \leq \frac{1}{2}\,.$$

Gathering all three cases, we obtain that it always holds that

$$\left|p'_{\widehat{m}-1}(0)\right| \leq c(\mu,\tau)\max\left(\frac{\rho}{M},1\right)\lambda_*^{-1}\,. \tag{12}$$

Second step: upper bound on $|p'_{\widehat{m}}(0)|$. For this we use the result of the first step and relate $\left|p'_{\widehat{m}-1}(0)\right|$ to $|p'_{\widehat{m}}(0)|$. It is a property of orthogonal polynomials (see Hanke, Corollary 2.6) that for any $m \geq 1$

$$p_{m-1}{}'(0) - p_m{}'(0) = \frac{[p_{m-1},p_{m-1}]_{(0)} - [p_m,p_m]_{(0)}}{\left[p_{m-1}^{(2)},p_{m-1}^{(2)}\right]_{(1)}} \leq \frac{[p_{m-1},p_{m-1}]_{(0)}}{\left[p_{m-1}^{(2)},p_{m-1}^{(2)}\right]_{(1)}}\,. \tag{13}$$

To upper bound the above quantity, we apply Lemma B.3 whithe the choice $\lambda = \lambda_*$ and

$$\varepsilon = \varepsilon_* := a(\mu,\tau)\min\left(\frac{M}{\rho},1\right)\lambda_*\,,$$

where $0 < a(\mu,\tau) \leq 1$ should be chosen small enough in order to satisfy some constraints to be specified below. The first constraint is the requirement $\varepsilon_* \in (0, x_{1,m-1})$ in order to apply Lemma B.3. For this, it can be seen from (12) that $a(\mu,\tau)$ can be chosen small enough to ensure

$$\varepsilon_* \leq \left|p'_{m-1}(0)\right|^{-1} \leq x_{1,m-1}\,,$$

the last inequality is an easy consequence of the fact that p_{m-1} has exactly $(m-1)$ positive real roots and $p_{m-1}(0) = 1$. We now turn to upper bound the following quantity appearing on the RHS of (4):

$$\begin{aligned}\Lambda(\varepsilon_*+\lambda_*)^{\frac{1}{2}}\delta(\lambda_*) + c(\mu)\Lambda^2\varepsilon_*\left(\varepsilon_*^\mu + Z_\mu(\lambda_*)\right)\kappa^{-\mu-\frac{1}{2}}\rho& \\ \leq 2(a(\mu,\tau)+1)\lambda_*^{\frac{1}{2}}\delta(\lambda_*) + c(\mu)a(\mu,\tau)\min(\rho,M)\,\lambda_*\widetilde{\lambda}_*^\mu\kappa^{-\frac{1}{2}}\rho& \\ \leq (c(\mu)a(\mu,\tau)+2)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\,,&\end{aligned} \tag{14}$$

where we have used the definition (7) for $\delta(\lambda_*)$ and inequality $Z_\mu(\lambda_*) \leq \lambda_*^\mu$, see (11). Now, we chose $a(\mu,\tau)$ so that the factor in the last display satisfies $c(\mu)a(\mu,\tau) \leq \frac{\tau}{2}$. Remember that the definition of the stopping rule entails

$$[p_{m-1},p_{m-1}]_{(0)}^{\frac{1}{2}} = \|T_n^*(T_n f_{\widehat{m}-1} - \mathbf{Y})\| > (2+\tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*) > (2+\tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\,, \tag{15}$$

Now combining (4), (15) and (14), we obtain

$$\left(1 - \frac{\tau+\frac{1}{2}}{\tau+2}\right)[p_{m-1},p_{m-1}]_{(0)}^{\frac{1}{2}} \leq \varepsilon_*^{-\frac{1}{2}}\left[p_{m-1}^{(2)},p_{m-1}^{(2)}\right]_{(1)}^{\frac{1}{2}}\,;$$

using this inequality in relation with (13) and (12), we obtain

$$|p'_{\widehat{m}}(0)| \leq \left|p'_{\widehat{m}-1}(0)\right| + c(\tau)\varepsilon_*^{-1} \leq c(\mu,\tau)\max\left(\frac{\rho}{M},1\right)\lambda_*^{-1}\,.$$

Final step. We apply Lemma B.2 (with $\lambda = \lambda_*$ and $\varepsilon = \varepsilon_*$), together with the bound on $|p'_{\widehat{m}}(0)|$ just obtained, and the inequality (by definition of the stopping rule)

$$\|T_n^*(T_n f_{\widehat{m}} - \mathbf{Y})\| \leq (2+\tau)\lambda_*^{\frac{1}{2}}\delta(\lambda_*)\,,$$

obtaining, using again (11):

$$\begin{aligned}\|f_{\widehat{m}} - f^*\|_2 &= \|T(f_{\widehat{m}} - f^*_{\mathcal{H}})\| \\ &\leq c(\mu,\tau)\left(\delta(\lambda_*)\max\left(\frac{\rho}{M},1\right) + \min(\rho,M)\widetilde{\lambda}_*^{\mu+\frac{1}{2}}\right) \leq c(\mu,\tau)(M+\rho)\widetilde{\lambda}_*^{\mu+\frac{1}{2}}\,.\end{aligned}$$

If $\widehat{m} = 0$, we can apply directly Lemma B.2 as above without requiring the two previous steps, since in this case $p'_0(0) = 0$, so that we obtain the same final bound.

D Sketch of the proof of Theorem 2.3

For the proof of Theorem 2.3, the condition **B1(**λ**)** is replaced by

B1'(λ**)** $\left\|(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - T^*f^*)\right\| \leq \delta(\lambda)\,.$

We check that **B1'(**λ_***)**, **B2(**λ_***)** and **B3** are satisfied in the setting of Theorem 2.3. To check **B1'(**λ_***)**, we use (16) instead of (17). Since the easily checked relation $T_n^*\mathbf{Y} = T_{\tilde{n}}^*\widetilde{\mathbf{Y}}$ holds, the upper bound obtained here has the same form as for Theorem 2.2, therefore we can use the same value $\delta(\lambda^*)$ for condition **B1'(**λ_***)** as in the previous section, given by (8). Notice however that we must now use the condition $\mu + s = r + s - \frac{1}{2} \geq 0$ to ensure that the chain of inequalities leading to (8) is valid.

For condition **B2(**λ_***)**, we can apply the deviation inequality (18) but with n replaced by $\widetilde{n}$, since we make use of all the unlabeled data. Using the fact that $\frac{n}{\widetilde{n}} \leq \widetilde{\lambda}_*^{-(1-2r)_+}$ and some elementary algebra leads to **B2(**λ_***)** being satisfied with $\Lambda := 2$.

Finally condition **B3** is satisfied with Δ given by (10) with n replaced by $\tilde{n}$.

Once these conditions are established, intermediate results similar in structure to Lemmas B.1, B.2 and B.3 can be derived, but where **B1(**λ**)** is replaced by **B1'(**λ**)**. The details are omitted here.

E More technical lemmas

In this section we collect some technical lemmas which underpin the main results. These are taken from previous sources and are recalled here for completeness. The main statistical tool is the following deviation inequality:

Lemma E.1. *Let λ be a positive number. Under assumption* **(Bounded)***, the following holds:*

$$\mathbb{P}\left[\left\|(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - T^*f^*)\right\| \leq 2M\left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda}n}\right)\log\frac{6}{\gamma}\right] \geq 1-\gamma\,. \tag{16}$$

If the representation $f^* = Tf^*_{\mathcal{H}}$ holds and under assumption **(Bernstein)***, we have the following:*

$$\mathbb{P}\left[\left\|(S+\lambda I)^{-\frac{1}{2}}(T_n^*\mathbf{Y} - S_n f^*_{\mathcal{H}})\right\| \leq 2M\left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda}n}\right)\log\frac{6}{\gamma}\right] \geq 1-\gamma\,. \tag{17}$$

Finally, the following holds:

$$\mathbb{P}\left[\left\|(S+\lambda I)^{-\frac{1}{2}}(S_n - S)\right\|_{HS} \leq 2\sqrt{\kappa}\left(\sqrt{\frac{\mathcal{N}(\lambda)}{n}} + \frac{2\sqrt{\kappa}}{\sqrt{\lambda}n}\right)\log\frac{6}{\gamma}\right] \geq 1-\gamma\,, \tag{18}$$

where we recall that $\|.\|_{HS}$ denotes the Hilbert-Schmidt norm.

The proof can be found in [3], and is based on a Bernstein-type inequality for random variables taking values in a Hilbert space, as established in [5, 6].

Inequality (18) can be fruitfully combined with the following:

Lemma E.2. *Assume there exists $\eta > 0$ such that the following inequality holds:*

$$\left\|(S+\lambda)^{-\frac{1}{2}}(S_n - S)(S+\lambda)^{-\frac{1}{2}}\right\| < 1-\eta\,,$$

then

$$\left\|(S+\lambda)^{\frac{1}{2}}(S_n+\lambda)^{-\frac{1}{2}}\right\| \leq \frac{1}{\sqrt{\eta}}.$$

Proof. First we have

$$\left\|(S+\lambda)^{\frac{1}{2}}(S_n+\lambda)^{-\frac{1}{2}}\right\| = \left\|(S+\lambda)^{\frac{1}{2}}(S_n+\lambda)^{-1}(S+\lambda)^{\frac{1}{2}}\right\|^{\frac{1}{2}};$$

then simple algebraic manipulation shows

$$(S+\lambda)^{\frac{1}{2}}(S_n+\lambda)^{-1}(S+\lambda)^{\frac{1}{2}} = \left(I - (S+\lambda)^{\frac{1}{2}}(S-S_n)^{-1}(S+\lambda)^{\frac{1}{2}}\right)^{-1}.$$

Finally, using the inequality $\left\|(I-A)^{-1}\right\| = \left\|\sum_{k\geq 0} A^k\right\| \leq (1-\|A\|)^{-1}$ for $\|A\| < 1$ yields the conclusion. □

We make use of the following operator inequalities:

Lemma E.3. *Let A, B be two positive, self-adjoint operators with $\max(\|A\|, \|B\|) \leq C$. Then for any $r \geq 0$, putting $\zeta = (r-1)_+$, the following inequality holds:*

$$\|A^r - B^r\| \leq (\zeta+1)C^\zeta \|A-B\|^{r-\zeta}\,. \tag{19}$$

Proof. Follows from the fact that the power function $x \mapsto x^r$ is operator monotone for $r \leq 1$ and Lipschitz with constant rC^{r-1} on $[0, C]$ if $r > 1$. □

Lemma E.4 ([1], Theorem IX.2.1-2)**.** *Let A, B be to self-adjoint, positive operators. Then for any $s \in [0,1]$:*

$$\|A^s B^s\| \leq \|AB\|^s\,. \tag{20}$$

Note: this result is stated for positive matrices in [1], but it is easy to check that the proof applies as well to positive operators on a Hilbert space.

References

[1] R. Bathia. *Matrix Analysis*, volume 169 of *Graduate texts in mathematics*. Springer, 1997.

[2] A. Caponnetto. Optimal Rates for Regularization Operators in Learning Theory. Technical Report CBCL Paper 264/ CSAIL-TR 2006-062, Massachusetts Institute of Technology, 2006.

[3] A. Caponnetto and E. De Vito. Optimal Rates for Regularized Least-squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[4] M. Hanke. *Conjugate Gradient Type Methods for Linear Ill-posed Problems*. Pitman Research Notes in Mathematics Series, 327, 1995.

[5] I. F. Pinelis and A. I. Sakhanenko. Remarks on inequalities for probabilities of large deviations. *Theory Probab. Appl.*, 1(30):143–148, 1985.

[6] V. Yurinksi. *Sums and Gaussian vectors*, volume 1617 of *Lecture notes in mathematics*. Springer, 1995.