Appendix
Reducing training time by efficient localized kernel regression

A Preliminaries

We let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ denote the sample space, where the input space $\mathcal{X}$ is a standard Borel space endowed with a fixed unknown probability measure ν. The kernel space $\mathcal{H}$ is assumed to be separable, equipped with a measurable positive semi-definite kernel K, bounded by κ, implying continuity of the inclusion map $I : \mathcal{H} \longrightarrow L^2(\nu)$. Moreover, we consider the covariance operator $T = \kappa^{-2} I^* I = \kappa^{-2} \mathbb{E}[K_X \otimes K_X]$, which can be shown to be positive self-adjoint trace class (and hence is compact). Given a sample $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, we define the sampling operator $S_{\mathbf{x}} : \mathcal{H} \longrightarrow \mathbb{R}^n$ by $(S_{\mathbf{x}} f)_i = \langle f, K_{x_i} \rangle_{\mathcal{H}}$. The empirical covariance operator is given by $T_{\mathbf{x}} = \kappa^{-2} S_{\mathbf{x}}^* S_{\mathbf{x}} : \mathcal{H} \longrightarrow \mathcal{H}$.

For a partition $\{\mathcal{X}_1, ..., \mathcal{X}_m\}$ of $\mathcal{X}$, we denote by $\hat{\mathcal{H}}_j$ the local RKHS with extended bounded kernel $\hat{K}_j$, supported on $\mathcal{X}_j$, with associated covariance operator $T_j = \kappa_j^{-2} \mathbb{E}_{\nu_j}[\hat{K}_j(X, \cdot) \otimes \hat{K}_j(X, \cdot)]$. Given a sample $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,n_j}) \in \mathcal{X}_j^{n_j}$, we define the sampling operator $S_{\mathbf{x}_j} : \hat{\mathcal{H}}_j \longrightarrow \mathbb{R}^{n_j}$ similarly by $(S_{\mathbf{x}_j} f)_i = \langle f, \hat{K}_j(x_i, \cdot) \rangle_{\hat{\mathcal{H}}_j}$.

The global covariance operator acts as an operator on the direct sum $\mathcal{H} = \hat{\mathcal{H}}_1 \oplus ... \oplus \hat{\mathcal{H}}_m$. According to (8), it decomposes as

$$T = \sum_{j=1}^{m} p_j^{-1} T_j \ ,$$

which can be used to prove that the global effective dimension can be expressed as the sum of the (rescaled) local ones.

Lemma 1 (Effective Dimension)**.** *For any* $\lambda \in [0, 1]$

$$\sum_{j=1}^{m} \mathcal{N}(T_j, p_j \lambda) = \mathcal{N}(T, \lambda) \ .$$

Finally, our error decomposition relies on the the following standard decomposition

Lemma 2. *Given* $j \in [m]$ *let* $p_j = \nu(\mathcal{X}_j)$ *and* $\nu_j(A) = \nu(A|\mathcal{X}_j)$*, for a measurable* $A \subset \mathcal{X}$*. One has*

$$L^2(\mathcal{X}, \nu) = \bigoplus_{j=1}^{m} p_j L^2(\mathcal{X}_j, \nu_j)$$

with

$$||f||^2_{L^2(\nu)} = \sum_{j=1}^{m} p_j ||f_j||^2_{L^2(\nu_j)} \ ,$$

where $f = f_1 + ... + f_m$ *.*

For proving our results we additionally need an appropriate Bernstein condition on the noise.

Assumption 1 (Distributions). *1. The sampling is random i.i.d., where each observation point (X_i, Y_i) follows the model $Y = f_\rho(X) + \epsilon$, and the noise satisfies the following Bernstein-type assumption: For any integer $k \geq 2$ and some $\sigma > 0$ and $M > 0$:*

$$\mathbb{E}[\ |Y - f_\rho(X)|^k \mid X\ ] \leq \frac{1}{2} k!\ \sigma^2 M^{k-2} \quad \nu - \text{a.s.}\ . \qquad (\text{Bern}(M,\sigma))$$

2. Given $\theta = (M, \sigma, R) \in \mathbb{R}^3_+$, the class $\mathcal{M} := \mathcal{M}(\theta, r, b)$ consists of all distributions ρ with X-marginal ν and conditional distribution of Y given X satisfying (Bern(M,σ)) *for the deviations and* (10) *for the mean.*

We remark that point 1 implies for any $j \in [m]$

$$\mathbb{E}[\ |Y - f_j(X)|^k \mid X\ ] \leq \frac{1}{2} k!\ \sigma^2 M^{k-2} \quad \nu_j - \text{a.s.}, \tag{1}$$

where σ and M are uniform with respect to m and k. This is what we actually need in our proofs.

For ease of reading we make use of the following conventions:

- we are interested in a precise dependence of multiplicative constants on the parameter $\sigma, M, R,\ m,\ n$
- the dependence of multiplicative constants on various other parameters, including the kernel parameter κ, the parameters arising from the regularization method, $b > 1$, $r > 0$, etc. will (generally) be omitted
- the value of C might change from line to line
- the expression "for n sufficiently large" means that the statement holds for $n \geq n_0$, with n_0 potentially depending on all model parameters (including σ, M and R).

B Proofs of Section 3

This section is devoted to proving the results of Section 3. Recall that by Assumption 2 the regression function belongs to $\mathcal{H}$, i.e. admits an unique representation $f = f_1 + ... + f_m$, with $f_j \in \hat{\mathcal{H}}_j$. For proving our error bounds we shall use a classical bias-variance decomposition

$$f_\rho - \hat{f}^\lambda_{\mathcal{D}} = \sum_{j=1}^m f_j - \hat{f}^\lambda_{\mathcal{D}_j} = \sum_{j=1}^m r_\lambda(T_{\mathbf{x}_j}) f_j + \sum_{j=1}^m g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j} f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\ ,$$

where $\hat{f}^\lambda_{\mathcal{D}}$ is given in (9), with $r_\lambda(t) = 1 - g_\lambda(t)t$ and with $g_\lambda(t) = (t+\lambda)^{-1}$. The final error bound follows then from

$$\begin{aligned}\mathbb{E}\Big[\ \mathcal{E}(f_\rho) - \mathcal{E}(\hat{f}^\lambda_{\mathcal{D}})\ \Big] &= \mathbb{E}\Big[\ ||f_\rho - \hat{f}^\lambda_{\mathcal{D}}||^2_{L^2(\nu)}\ \Big] \\ &\leq \mathbb{E}\left[\ ||\sum_{j=1}^m r_\lambda(T_{\mathbf{x}_j}) f_j||^2_{L^2(\nu)}\ \right] + \mathbb{E}\left[\ ||\sum_{j=1}^m g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j} f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)||^2_{L^2(\nu)}\ \right]\ .\end{aligned} \tag{2}$$

We proceed by bounding each term in the above decomposition separately.

Proposition 1 (Approximation Error)**.** *For any $\lambda \in (0,1]$, one has*

$$\mathbb{E}\Big[\|\sum_{j=1}^m r_\lambda(T_{\mathbf{x}_j})f_j\|^2_{L^2(\nu)}\Big] \le CR^2\sum_{j=1}^m p_j \mathcal{B}^2_{n_j}(T_j,\lambda)\lambda^{2(r_j+\frac{1}{2})}\ ,$$

where $\mathcal{B}^2_{n_j}(T_j,\lambda)$ is defined in Proposition 9 and where C does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$.

Proof of Proposition 1. Recall[1] that $||\sqrt{T_j}f||_{\hat{\mathcal{H}}_j} = ||f||_{L^2(\nu_j)}$ for any $f \in \hat{\mathcal{H}}_j$. According to Lemma 2, by Assumption 10 we have

$$\begin{aligned}
\mathbb{E}\Big[\|\sum_{j=1}^m r_\lambda(T_{\mathbf{x}_j})f_j\|^2_{L^2(\nu)}\Big] &= \sum_{j=1}^m p_j \mathbb{E}\Big[\|r_\lambda(T_{\mathbf{x}_j})f_j\|^2_{L^2(\nu_j)}\Big]\\
&= \sum_{j=1}^m p_j \mathbb{E}\Big[\|\sqrt{T_j}r_\lambda(T_{\mathbf{x}_j})f_j\|^2_{\hat{\mathcal{H}}_j}\Big]\\
&\le CR^2\sum_{j=1}^m p_j \mathbb{E}\Big[\|\sqrt{T_j}r_\lambda(T_{\mathbf{x}_j})T_j^{r_j}\|^2\Big]\ . \qquad (3)
\end{aligned}$$

We bound for any $j \in [m]$ the expectation by first deriving a probabilistic estimate. For any $\eta \in (0,1]$, with probability at least $1-\eta$

$$\begin{aligned}
||\sqrt{T_j}r_\lambda(T_{\mathbf{x}_j})T_j^{r_j}|| &\le C\log^2(2\eta^{-1})\mathcal{B}_{n_j}(T_j,\lambda)\ ||T_j^{\frac{1}{2}}(T_j+\lambda)^{\frac{1}{2}}||\ ||(T_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}r_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}+\lambda)^{r_j}||\ ||(T_j+\lambda)^{r_j}T_j^{r_j}||\\
&\le C\log^2(2\eta^{-1})\mathcal{B}_{n_j}(T_j,\lambda)\lambda^{r_j+\frac{1}{2}}\ .
\end{aligned}$$

Here we have used that

$$||(T_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}r_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}+\lambda)^{r}|| \le C\lambda^{r_j+\frac{1}{2}}$$

and that for $s \in [0,\frac{1}{2}]$

$$||(T_j+\lambda)^sT_j^s|| \le ||(T_j+\lambda)T_j||^s \le 1$$

by Proposition 10 and the spectral theorem. Also, from Proposition 10 and Proposition 9

$$||(T_{\mathbf{x}_j}+\lambda)^{-\frac{1}{2}}(T_j+\lambda)^{\frac{1}{2}}|| \le ||(T_{\mathbf{x}_j}+\lambda)^{-1}(T_j+\lambda)||^{\frac{1}{2}} \le \sqrt{8}\log(2\eta^{-1})\mathcal{B}^{\frac{1}{2}}_{n_j}(T_j,\lambda)\ .$$

From Lemma 7, by integration

$$\mathbb{E}\Big[\|\sqrt{T_j}r_\lambda(T_{\mathbf{x}_j})T_j^{r_j}\|^2\Big] \le C\mathcal{B}^2_{n_j}(T_j,\lambda)\lambda^{2(r_j+\frac{1}{2})}\ .$$

Combining this with (3) finishes the proof. □

Proposition 2 (Sample Error)**.** *For any $\lambda \in (0,1]$, one has*

$$\mathbb{E}\Big[\|\sum_{j=1}^m g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu)}\Big] \le C\sum_{j=1}^m p_j\ \mathcal{B}^2_{n_j}(T_j,\lambda)\lambda\ \left(\frac{M}{n_j\lambda}+\sigma\sqrt{\frac{\mathcal{N}(T_j,\lambda)}{n_j\lambda}}\right)^2\ ,$$

where $\mathcal{B}^2_{n_j}(T_j,\lambda)$ is defined in Proposition 9 and C does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$.

[1]If $I_j : \hat{\mathcal{H}}_j \hookrightarrow L^2(\nu_j)$, then $T_j = I_j^*I_j$ and $||\sqrt{T_j}f||^2_{\hat{\mathcal{H}}_j} = \langle T_jf, f\rangle_{\hat{\mathcal{H}}_j} = \langle I_jf, I_jf\rangle_{L^2(\nu_j)} = ||f||^2_{L^2(\nu_j)}$. Here, we identify $I_jf = f$.

Proof of Proposition 2. Using again $||\sqrt{T_j}f||_{\hat{\mathcal{H}}_j} = ||f||_{L^2(\nu_j)}$ we find with Lemma 2

$$\begin{aligned}
\mathbb{E}\Big[\|\sum_{j=1}^{m} g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\big\|^2_{L^2(\nu)}\Big] &= \sum_{j=1}^{m} p_j \mathbb{E}\Big[\big\|g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\big\|^2_{L^2(\nu_j)}\Big] \\
&= \sum_{j=1}^{m} p_j \mathbb{E}\Big[\big\|\sqrt{T_j}g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\big\|^2_{\hat{\mathcal{H}}_j}\Big] . \qquad (4)
\end{aligned}$$

We bound the expectation for each separate subsample of size n_j by first deriving a probabilistic estimate and then by integration. For this reason, we use (16) and Proposition 10 and write for any $f_j \in \hat{\mathcal{H}}_j$, $j \in [m]$

$$\begin{aligned}
||\sqrt{T_j}f_j||_{\hat{\mathcal{H}}_j} &\leq ||\sqrt{T_j}(T_j+\lambda)^{-1/2}||\ ||(T_j+\lambda)^{1/2}(T_{\mathbf{x}_j}+\lambda)^{-1/2}||\ ||(T_{\mathbf{x}_j}+\lambda)^{1/2}f_j||_{\hat{\mathcal{H}}_j} \\
&\leq ||T_j(T_j+\lambda)^{-1}||^{1/2}\ ||(T_j+\lambda)(T_{\mathbf{x}_j}+\lambda)^{-1}||^{1/2}\ ||(T_{\mathbf{x}_j}+\lambda)^{1/2}f_j||_{\hat{\mathcal{H}}_j} \\
&\leq C\log(4\eta^{-1})\mathcal{B}^{1/2}_{n_j}(T_j,\lambda)\ ||(T_{\mathbf{x}_j}+\lambda)^{1/2}f_j||_{\hat{\mathcal{H}}_j} \ , \qquad (5)
\end{aligned}$$

holding with probability at least $1-\frac{\eta}{2}$.

We proceed by splitting

$$(T_{\mathbf{x}_j}+\lambda)^s g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_\rho - S^*_{\mathbf{x}_j}\mathbf{y}_j) = H^{(1)}_{\mathbf{x}_j}\cdot H^{(2)}_{\mathbf{x}_j}\cdot h^\lambda_{\mathbf{z}_j} \ , \qquad (6)$$

with

$$\begin{aligned}
H^{(1)}_{\mathbf{x}_j} &:= (T_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}+\lambda)^{\frac{1}{2}}, \\
H^{(2)}_{\mathbf{x}_j} &:= (T_{\mathbf{x}_j}+\lambda)^{-\frac{1}{2}}(T+\lambda)^{\frac{1}{2}}, \\
h^\lambda_{\mathbf{z}_j} &:= (T+\lambda)^{-\frac{1}{2}}(T_{\mathbf{x}_j}f_\rho - S^*_{\mathbf{x}_j}\mathbf{y}_j) \ .
\end{aligned}$$

The first term is bounded. The second term is now estimated using (16) once more. One has with probability at least $1-\frac{\eta}{4}$

$$H^{(2)}_{\mathbf{x}_j} \leq \sqrt{8}\log(8\eta^{-1})\mathcal{B}_{\frac{n}{m}}(T_j,\lambda)^{\frac{1}{2}} \ .$$

Finally, $h^\lambda_{\mathbf{z}_j}$ is estimated using Proposition 8:

$$h^\lambda_{\mathbf{z}_j} \leq 2\log(8\eta^{-1})\left(\frac{M}{n_j\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}(T_j,\lambda)}{n_j}}\right) ,$$

holding with probability at least $1-\frac{\eta}{4}$. Thus, combining the estimates following (6) with (5) gives for any $j \in [m]$

$$||\sqrt{T_j}g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_\rho - S^*_{\mathbf{x}_j}\mathbf{y}_j)||_{\hat{\mathcal{H}}_j} \leq C\log^3(8\eta^{-1})\mathcal{B}_{n_j}(T_j,\lambda)\sqrt{\lambda}\ \left(\frac{M}{n_j\lambda} + \sigma\sqrt{\frac{\mathcal{N}(T_j,\lambda)}{n_j\lambda}}\right) ,$$

with probability at least $1-\eta$. By integration using Lemma 7 one obtains

$$\mathbb{E}\Big[\big\|\sqrt{T_j}g_\lambda(B_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_\rho - S^*_{\mathbf{x}_j}\mathbf{y}_j)\big\|^2_{\hat{\mathcal{H}}_j}\Big]^{\frac{1}{2}} \leq C\mathcal{B}_{n_j}(T_j,\lambda)\sqrt{\lambda}\ \left(\frac{M}{n_j\lambda} + \sigma\sqrt{\frac{\mathcal{N}(T_j,\lambda)}{n_j\lambda}}\right) .$$

Combining this with (4) implies

$$\mathbb{E}\Big[\|\sum_{j=1}^{m} g_\lambda(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\big\|^2_{L^2(\nu)}\Big] \leq C\sum_{j=1}^{m} p_j\ \mathcal{B}^2_{n_j}(T_j,\lambda)\lambda\ \left(\frac{M}{n_j\lambda} + \sigma\sqrt{\frac{\mathcal{N}(T_j,\lambda)}{n_j\lambda}}\right)^2 ,$$

where C does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$. $\square$

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let the regularization parameter λ_n be chosen as

$$\lambda_n = \min\left(1, \left(\frac{\sigma^2}{R^2 n}\right)^{\frac{1}{2r+1+\gamma}}\right) , \tag{7}$$

with $r = \min(r_1, ..., r_m)$ and assume that $n_j = \lfloor \frac{n}{m} \rfloor$. Note that by Lemma 5 we have $\mathcal{B}_{\frac{n}{m}}(T_j, \lambda_n) \leq 2$ for any $j \in [m]$, provided $n > n_0$, with n_0 given by (17). Since $\lambda_n^{r_j} \leq \lambda_n^r$ for any $j \in [m]$, the approximation error bound becomes by Proposition 1

$$\begin{aligned} \mathbb{E}\Big[\Big\| \sum_{j=1}^{m} r_{\lambda_n}(T_{\mathbf{x}_j}) f_j \Big\|^2_{L^2(\nu)}\Big] &\leq CR^2 \sum_{j=1}^{m} p_j \lambda_n^{2(r_j+\frac{1}{2})} \\ &\leq CR^2 \, \lambda_n^{2(r+\frac{1}{2})} , \end{aligned} \tag{8}$$

where we also used that $\sum_j p_j = 1$.

For estimating the sample error firstly observe that

$$\frac{Mm}{n\lambda_n} \leq R\lambda_n^r$$

if

$$n > \left(m\,\frac{M}{R}\right)^{\frac{2r+1+\gamma}{r+\gamma}} \left(\frac{R}{\sigma}\right)^{\frac{2(r+1)}{r+\gamma}} =: n_1 .$$

Thus, from Proposition 2 we obtain (recalling again that $\mathcal{B}_{\frac{n}{m}}(T_j, \lambda_n) \leq 2$)

$$\mathbb{E}\Big[\Big\| \sum_{j=1}^{m} g_{\lambda_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j} f_j - S^*_{\mathbf{x}_j}\mathbf{y}_j) \Big\|^2_{L^2(\nu)}\Big] \leq C\lambda_n \sum_{j=1}^{m} p_j \left(R\lambda_n^r + \sigma\sqrt{\frac{m\mathcal{N}(T_j, \lambda_n)}{n\lambda_n}}\right)^2 . \tag{9}$$

We proceed by applying $(a+b)^2 \leq 2(a^2+b^2)$. Observe that by our Assumption 3 , 2.

$$\begin{aligned} \sum_{j=1}^{m} p_j \sigma^2 \frac{m\mathcal{N}(T_j, \lambda_n)}{n\lambda_n} &= \sigma^2 \frac{m}{n\lambda_n} \sum_{j=1}^{m} p_j \; \mathcal{N}(T_j, \lambda_n) \\ &\leq C\frac{\sigma^2}{n\lambda_n} \mathcal{N}(T, m\lambda_n) \\ &\leq Cm^{-\gamma} \frac{\sigma^2}{n\lambda_n} \lambda_n^{-\gamma} \\ &\leq CR\lambda_n^r , \end{aligned} \tag{10}$$

by definition of λ_n. Finally, combining (2) with (10), (9) and (8) proves the theorem, provided

$$n > \max(n_0, n_1) \geq C_{M,\sigma,R,\gamma,r} \; m^{1+\frac{\gamma+1}{2r}} , \tag{11}$$

for some (explicitly given) $C_{M,\sigma,R,\gamma,r} < \infty$. □

Proof of Theorem 2. Assume that $n_j = \lfloor \frac{n}{m} \rfloor$. Let the regularization parameter λ_n be given by (15) . As above, Lemma 5 yields $\mathcal{B}_{\frac{n}{m}}(T_j, \lambda_n) \leq 2$ provided $n > n_0$, with n_0 satisfying (17) (with r replaced by r_h). From Proposition 1 we immediately obtain for the approximation error

$$\mathbb{E}\Big[\|\sum_{j=1}^{m} r_{\lambda_n}(T_{\mathbf{x}_j}) f_j\|^2_{L^2(\nu)}\Big] \leq C \left(R_l^2 \left(\sum_{j\in E} p_j \right) \lambda_n^{2(r_l+\frac{1}{2})} + R_h^2 \left(\sum_{j\in E^c} p_j \right) \lambda_n^{2(r_h+\frac{1}{2})} \right)$$

$$\leq C R_h^2 \lambda_n^{2(r_h+\frac{1}{2})} \ .$$

Here we have used that by Assumption 4

$$\left(\sum_{j\in E} p_j \right) \leq \left(\frac{R_h}{R_l} \right)^2 \lambda_n^{2(r_h - r_l)} \quad \text{and} \quad \left(\sum_{j\in E^c} p_j \right) \leq 1 \ .$$

The bound for the sample error follows exactly as in the proof of Theorem 1. Finally, the error bound (17) is obtained by using again (2). □

C Proofs of Section 4

For proving Theorem 3 we use the non-asymptotic error decomposition given in Theorem 2 of [4], somewhat reformulated and streamlined using our estimate (16). We adopt the notation and idea of [4] and write $\hat{f}^{\lambda}_{n,l} = g_{\lambda,l}(T_{\mathbf{x}}) S^*_{\mathbf{x}} \mathbf{y}$, with $g_{\lambda,l}(T_{\mathbf{x}}) = V(V^* T_{\mathbf{x}} V + \lambda)^{-1} V^*$ and $VV^* = P_l$, the projection operator onto $\mathcal{H}_l$, $l \leq n$. Consider

$$||\sqrt{T}(\hat{f}^{\lambda}_{n,l} - f_\rho)||_{\mathcal{H}} \leq T_1 + T_2$$

with

$$T_1 = ||g_{\lambda,l}(T_{\mathbf{x}})(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}} f_\rho)||_{L^2(\nu)} = ||\sqrt{T} g_{\lambda,l}(T_{\mathbf{x}})(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}}$$

and

$$T_2 = ||\sqrt{T} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} f_\rho - f_\rho)||_{\mathcal{H}} \ ,$$

which we bound in Proposition 3 and Proposition 4 .

Proposition 3 (Expectation Sample Error KRLS-Nyström)**.**

$$\mathbb{E}\Big[\big\|g_{\lambda,l}(T_{\mathbf{x}})(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}} f_\rho)\big\|^2_{L^2(\nu)}\Big]^{\frac{1}{2}} \leq C\, \sqrt{\lambda} \mathcal{B}_n(T, \lambda) \left(\frac{M}{n\lambda} + \sigma \sqrt{\frac{\mathcal{N}(T,\lambda)}{n\lambda}} \right)$$

where C does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$.

Proof of Proposition 3. For estimating T_1 we use Proposition 9 and obtain for any $\lambda \in (0,1]$ with probability at least $1-\eta$

$$\begin{aligned} T_1 &\leq C \log(2\eta^{-1}) \mathcal{B}_n(T,\lambda)\ ||(T_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}} \\ &\leq C \log^2(4\eta^{-1}) \mathcal{B}^2_n(T,\lambda)\ ||(T_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)^{1/2}|| \\ &\quad ||(T+\lambda)^{-1/2}(S^*_{\mathbf{x}}\mathbf{y} - T_{\mathbf{x}} f_\rho)||_{\mathcal{H}} \ . \end{aligned}$$

From Proposition 6 in [4] and from the spectral Theorem we obtain

$$||(T_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)^{1/2}|| \leq 1 \ .$$

Thus, applying Proposition 7 one has with probability at least $1-\eta$

$$T_1 \leq C \log^3(8\eta^{-1})\ \sqrt{\lambda}\ \mathcal{B}^2_n(T,\lambda)\ \left(\frac{M}{n\lambda} + \sigma \sqrt{\frac{\mathcal{N}(T,\lambda)}{n\lambda}} \right) \ ,$$

where C does not depend on $(\sigma, M, R) \in \mathbb{R}^3_+$. Integration using Lemma 7 gives the result. □

Before we proceed we introduce the computational error: For $u \in [0, \frac{1}{2}]$, $\lambda \in (0,1]$ define

$$\mathcal{C}_u(l,\lambda) := ||(Id - VV^*)(T+\lambda)^u|| \; .$$

The proof of the following Lemma can be found in [4], Proof of Theorem 2.

Lemma 3. *For any* $u \in [0, \frac{1}{2}]$

$$\mathcal{C}_u(l,\lambda) \leq \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2u} \; .$$

Lemma 4. *If* λ_n *is defined by* (12) *and if*

$$l_n \geq n^\beta \qquad \beta > \frac{\gamma+1}{2r+1+\gamma}$$

one has with probability at least $1-\eta$

$$\mathcal{C}_{\frac{1}{2}}(l_n,\lambda_n) \leq C \log(2\eta^{-1}) \sqrt{\lambda_n} \; ,$$

provided n *is sufficiently large.*

Proof of Lemma 4. Using Proposition 3 in [4] one has with probability at least $1-\eta$

$$\begin{aligned} \mathcal{C}_{\frac{1}{2}}(l,\lambda_n) &\leq \sqrt{\lambda_n} \; ||(T_{\mathbf{x}_l} + \lambda_n)^{-1}(T+\lambda_n)||^{\frac{1}{2}} \\ &\leq C \log(2\eta^{-1}) \sqrt{\lambda_n} \; \mathcal{B}_l^{\frac{1}{2}}(T,\lambda_n) \; . \end{aligned}$$

Recall that $\mathcal{N}(T,\lambda) \leq C_b \lambda^{-\frac{1}{b}}$, implying

$$\mathcal{B}_l(T,\lambda_n) \leq C \left(1 + \left(\frac{2}{l\lambda_n} + \sqrt{\frac{\lambda_n^{-\gamma}}{l\lambda_n}} \right)^2 \right) \; .$$

Straightforward calculation shows that

$$\frac{2}{l_n\lambda_n} = o(1) \; , \quad \text{if} \;\; l_n \geq n^\beta \; , \; \beta > \frac{1}{2r+1+\gamma}$$

and

$$\sqrt{\frac{\lambda_n^{-\gamma}}{l_n\lambda_n}} = o(1) \; , \quad \text{if} \;\; l_n \geq n^\beta \; , \; \beta > \frac{\gamma+1}{2r+1+\gamma} \; .$$

Thus, $\mathcal{C}_{\frac{1}{2}}(l_n,\lambda_n) \leq C \log(2\eta^{-1})\sqrt{\lambda_n}$, with probability at least $1-\eta$. □

Proposition 4 (Expectation Approximation- and Computational Error KRLS-Nyström)**.** *Assume that*

$$l_n \geq n^\beta \; , \qquad \beta > \frac{\gamma+1}{2r+1+\gamma}$$

and $(\lambda_n)_n$ *is chosen according to* (12)*. If* n *is sufficiently large*

$$\mathbb{E}\Big[\big\| \sqrt{T} g_{\lambda_n, l_n}(T_{\mathbf{x}})(T_{\mathbf{x}} f_\rho - f_\rho) \big\|^2_{L^2(\nu)} \Big]^{\frac{1}{2}} \leq C \; a_n \; ,$$

where C *does not depend on* $(\sigma, M, R) \in \mathbb{R}^3_+$*.*

Proof of Proposition 4. Using that $||T^{-r} f_\rho||_{\mathcal{H}} \leq R$ one has for any $\lambda \in (0,1]$

$$T_2 \leq CR\ (\ (a) + (b) + (c)\ )\ , \tag{12}$$

with

$$(a) = ||\sqrt{T}(Id - VV^*)T^r||\ , \qquad (b) = \lambda||\sqrt{T} g_{\lambda,l}(T_{\mathbf{x}})T^r||$$

and

$$(c) = ||\sqrt{T} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)(Id - VV^*)T^r||\ .$$

Since $(Id - VV^*)^2 = (Id - VV^*)$ we obtain by Lemma 3

$$(a) \leq \mathcal{C}_{\frac{1}{2}}(l,\lambda)\ \mathcal{C}_r(l,\lambda) \leq \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r+1}\ .$$

Furthermore, using (16) , with probability at least $1 - \frac{\eta}{2}$

$$\begin{aligned}(b) &\leq C \log^2(8\eta^{-1})\lambda \mathcal{B}_n^{\frac{1}{2}+r}(T,\lambda)\ ||(T_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)^r|| \\ &\leq C \log^2(8\eta^{-1})\lambda^{\frac{1}{2}+r} \mathcal{B}_n^{\frac{1}{2}+r}(T,\lambda)\ ,\end{aligned}$$

by again using Proposition 6 in [4].

The last term gives with probability at least $1 - \frac{\eta}{2}$

$$\begin{aligned}(c) &\leq C \log(8\eta^{-1})||(T_{\mathbf{x}} + \lambda)^{1/2} g_{\lambda,l}(T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)||\ \mathcal{C}_r(l,\lambda) \\ &\leq C \log(8\eta^{-1})\sqrt{\lambda}\ \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r}\ .\end{aligned}$$

Combining the estimates for (a), (b) and (c) gives

$$T_2 \leq CR \log^2(8\eta^{-1}) \left(\mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r+1} + \lambda^{\frac{1}{2}+r} \mathcal{B}_n^{\frac{1}{2}+r}(T,\lambda) + \sqrt{\lambda}\ \mathcal{C}_{\frac{1}{2}}(l,\lambda)^{2r} \right)\ .$$

We now choose λ_n according to (12) . Notice that by Lemma 6 one has $\mathcal{B}_n(T,\lambda_n) \leq C$ for any n sufficiently large. Applying Lemma 4 we obtain, with probability at least $1 - \eta$

$$T_2 \leq C \log^2(8\eta^{-1}) R \lambda_n^{r+\frac{1}{2}}\ ,$$

provided n is sufficiently large and

$$l_n \geq n^\beta\ , \qquad \beta > \frac{\gamma + 1}{2r + 1 + \gamma}\ .$$

The result follows from integration by applying Lemma 7 and recalling that $a_n = R\lambda_n^{r+\frac{1}{2}}$. □

With these preparations we can now prove the main result of Section 4.

Proof of Theorem 3. The proof easily follows by combining Proposition 3 and Proposition 4 . In particular, the estimate for the sample error by choosing $\lambda = \lambda_n$ follows by recalling that $\mathcal{N}(T,\lambda_n) \leq C_\gamma \lambda_n^{-\gamma}$, by definition of $(a_n)_n$ in Theorem 3 , by Lemma 6 and by

$$\frac{M}{n\lambda_n} = o\left(\sigma\sqrt{\frac{\lambda_n^{-\gamma}}{n\lambda_n}}\right)\ .$$

□

D Proofs of Section 5

Following the lines in the previous sections we divide the error analysis in bounding the Sample error, Approximation error and Computational error.

Proposition 5 (Sample Error). *Let λ_n be defined as in* (12). *We have*

$$\mathbb{E}\Big[\|\sum_{j=1}^{m} g_{\lambda_n,l}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}\hat{f}_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu)}\Big] \leq CR^2\,\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{2(r+\frac{1}{2})}{2r+1+\gamma}}\;,$$

where n has to be chosen sufficiently large, i.e.

$$n > C_{\sigma,R,\gamma,r}\; m^{1+\frac{\gamma+1}{2r+1+\gamma}}\;,$$

for some $C_{\sigma,R,\gamma,r} < \infty$. Moreover, C does not depend on the model parameter $\sigma, M, R \in \mathbb{R}^3_+$.

Proof of Proposition 5. Applying Proposition 3 we obtain

$$\begin{aligned}\mathbb{E}\Big[\|\sum_{j=1}^{m} g_{\lambda,l}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}\hat{f}_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu)}\Big] &= \sum_{j=1}^{m} p_j\mathbb{E}\Big[\|g_{\lambda,l}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}\hat{f}_j - S^*_{\mathbf{x}_j}\mathbf{y}_j)\|^2_{L^2(\nu_j)}\Big]\\ &\leq C\sum_{j=1}^{m} p_j\; \mathcal{B}^2_{\frac{n}{m}}(T_j,\lambda)\lambda\,\left(\frac{Mm}{n\lambda} + \sigma\sqrt{\frac{m\mathcal{N}(T_j,\lambda)}{n\lambda}}\right)^2\;.\end{aligned}$$

Arguing as in the proof of Theorem 1, using Lemma 5, implies the result. □

Proposition 6 (Approximation and Computational Error). *Let λ_n be defined by* (12). *Assume the number of subsampled points satisfies $l_n \geq n^{\beta}$ with*

$$\beta > \frac{\gamma+1}{2r+\gamma+1}\;.$$

Then

$$\mathbb{E}\Big[\|\sum_{j=1}^{m} g_{\lambda_n,l_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - f_j)\|^2_{L^2(\nu)}\Big] \leq CR^2\,\left(\frac{\sigma^2}{R^2 n}\right)^{\frac{2(r+\frac{1}{2})}{2r+\gamma+1}}\;,$$

where C does not depend on the model parameter σ, M, R.

Proof of Proposition 6. For proving this Proposition we combine techniques from both the partitioning and subsampling approach. More precisely:

$$\begin{aligned}\mathbb{E}\Big[\|\sum_{j=1}^{m} g_{\lambda_n,l_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - f_j)\|^2_{L^2(\nu)}\Big] &= \sum_{j=1}^{m} p_j\mathbb{E}\Big[\|g_{\lambda_n,l_l}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - f_j)\|^2_{L^2(\nu_j)}\Big]\\ &= \sum_{j=1}^{m} p_j\mathbb{E}\Big[\|\sqrt{T_j}g_{\lambda_n,l_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}f_j - f_j)\|^2_{\hat{\mathcal{H}}_j}\Big]\;.\end{aligned}$$

We shall decompose as in (12), with T replaced by T_j and $T_{\mathbf{x}}$ replaced by $T_{\mathbf{x}_j}$,

$$||\sqrt{\bar{T}_j}g_{\lambda_n,l_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j}\hat{f}_j - f_j)||_{\hat{\mathcal{H}}_j} \leq CR\;(\;(a)+(b)+(c)\;) = (*)\;.$$

Following the lines of the proof of Proposition 4 leads to an upper bound (with probability at least $1-\eta$) for the rhs of the last inequality, which is

$$\begin{aligned}(*) &\leq CR\log^2(8\eta^{-1})\left(\mathcal{C}_{\frac{1}{2}}(l,\lambda_n)^{2r+1} + \lambda_n^{\frac{1}{2}+r}\mathcal{B}^{\frac{1}{2}+r}_{\frac{n}{m}}(T_j,\lambda_n) + \sqrt{\lambda_n}\;\mathcal{C}_{\frac{1}{2}}(l,\lambda_n)^{2r}\right)\\ &\leq CR\log^2(8\eta^{-1})\lambda_n^{r+\frac{1}{2}}\left(\mathcal{B}^{2r+1}_{l}(T_j,\lambda_n) + \mathcal{B}^{r+\frac{1}{2}}_{\frac{n}{m}}(T_j,\lambda_n) + \mathcal{B}^{2r}_{l}(T_j,\lambda_n)\right)\;.\end{aligned}$$

Thus, by integration and since $r \leq \frac{1}{2}$

$$\mathbb{E}\Big[\|\sum_{j=1}^{m} g_{\lambda_n,l_n}(T_{\mathbf{x}_j})(T_{\mathbf{x}_j} f_j - f_j)\|^2_{L^2(\nu)}\Big] \leq CR^2 \lambda_n^{2(r+\frac{1}{2})} \sum_{j=1}^{m} p_j \left(\mathcal{B}_l^4(T_j,\lambda_n) + \mathcal{B}^2_{\frac{n}{m}}(T_j,\lambda_n) + \mathcal{B}_l^2(T_j,\lambda_n)\right) .$$

Note that by Lemma 5, if

$$n \geq C_{\sigma,R,\gamma,r} m^{1+\frac{\gamma+1}{2r}} \tag{13}$$

we have

$$\begin{aligned}
\mathcal{B}_{\frac{n}{m}}(T_j,\lambda_n) &= \left[1 + \left(\frac{2m}{n\lambda_n} + \sqrt{\frac{m_n \mathcal{N}(T_j,\lambda_n)}{n\lambda}}\right)^2\right] \\
&\leq C\left[1 + \left(\frac{2m}{n\lambda_n}\right) + \left(\frac{m\mathcal{N}(T_j,\lambda_n)}{n\lambda}\right)\right] \\
&\leq C \; .
\end{aligned}$$

Moreover, since $\mathcal{N}(T_j,\lambda_n) \leq \mathcal{N}(T,\lambda_n/p_j)$, by Assumption 3, 2. and since $p_j \leq 1$

$$\mathcal{B}_{l_n}(T_j,\lambda_n) \leq 1 + \left(\frac{2}{l_n\lambda_n} + \sigma\sqrt{\frac{\lambda_n^{-\gamma}}{l_n\lambda_n}}\right)^2 .$$

Straightforward calculation shows that

$$\frac{2}{l_n\lambda_n} = o(1) \; , \quad \text{if } \; l_n \geq n^{\beta'} \; , \; \beta' > \frac{1}{2r+\gamma+1}$$

and

$$\sqrt{\frac{\lambda_n^{-\gamma}}{l_n\lambda_n}} = \mathcal{O}(1) \; , \quad \text{if } \; l_n \geq n^{\beta'} \; , \; \beta' \geq \frac{\gamma+1}{2r+\gamma+1} \; . \tag{14}$$

Thus, (14) ensures $\mathcal{B}_{l_n}(T_j,\lambda_n) = \mathcal{O}(1)$. Finally, on each local set we have the requirement $l_n \lesssim \frac{n}{m_n}$, which is implied by

$$l_n \lesssim n^{1-\alpha} \sim n^{\frac{\gamma+1}{2r+\gamma+1}} \; .$$

Together with (14) we get a sharp bound

$$l_n \sim n^{\frac{\gamma+1}{2r+\gamma+1}} \; .$$

□

E Probabilistic Inequalities

In this section we recall some well-known probabilistic inequalities.

Proposition 7 ([2])**.** *For $n \in \mathbb{N}$, $\lambda \in (0,1]$ and $\eta \in (0,1]$, one has with probability at least $1-\eta$:*

$$\left\|(T+\lambda)^{-\frac{1}{2}} \; (T_{\mathbf{x}} f_\rho - S^*_{\mathbf{x}}\mathbf{y})\right\|_{\mathcal{H}} \; \leq \; 2\log(2\eta^{-1}) \left(\frac{M}{n\sqrt{\lambda}} + \sigma\sqrt{\frac{\mathcal{N}(T,\lambda)}{n}}\right) .$$

Proposition 8 ([2], Proposition 5.3)**.** *For any $\lambda \in (0,1]$ and $\eta \in (0,1)$ one has with probability at least $1-\eta$:*

$$\left\|(T+\lambda)^{-1}(T-T_{\mathbf{x}})\right\|_{HS} \leq 2\log(2\eta^{-1}) \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(T,\lambda)}{n\lambda}}\right) .$$

Proposition 9 ([3])**.** *Define*

$$\mathcal{B}_n(T,\lambda) := \left[1 + \left(\frac{2}{n\lambda} + \sqrt{\frac{\mathcal{N}(T,\lambda)}{n\lambda}} \right)^2 \right] \tag{15}$$

For any $\lambda > 0$, $\eta \in (0,1]$, with probability at least $1-\eta$ one has

$$\left\| (T_{\mathbf{x}} + \lambda)^{-1}(T+\lambda) \right\| \leq 8 \log^2(2\eta^{-1}) \mathcal{B}_n(T,\lambda) \; . \tag{16}$$

Lemma 5. *Let $m \in \mathbb{N}$ and λ_n be defined by* (12)*. Then for any $j \in [m]$ and $n > n_0$*

$$\mathcal{B}_{\frac{n}{m}}(T_j, \lambda_n) \leq 2 \; .$$

Here, n_0 depends on the number m of subsets and the model parameter R, σ, γ, r and is explicitly given in (17)*.*

Proof of Lemma 5. Recall that we assume $\mathcal{N}(T,\lambda) \leq C_\gamma \lambda^{-\gamma}$, for some $b \geq 1$, $C_\gamma < \infty$. Thus, by Lemma 1 we have for any $j \in [m]$

$$\mathcal{N}(T_j,\lambda) \;\leq\; \mathcal{N}(T, \lambda/p_j) \;\leq\; C_\gamma \; p_j^\gamma \; \lambda^{-\gamma}$$

and thus

$$\frac{m\mathcal{N}(T_j,\lambda_n)}{n\lambda_n} \leq C_\gamma p_j \frac{m}{n} \lambda_n^{-(1+\gamma)} < \frac{1}{2} \; ,$$

provided

$$n > (2C_\gamma p_j m)^{\frac{2r+\gamma+1}{2r}} \; \left(\frac{R}{\sigma} \right)^{\frac{2(\gamma+1)}{2r}} \; .$$

Moreover,

$$\frac{2m}{n\lambda_n} < \frac{1}{2} \; ,$$

provided

$$n > (4m)^{\frac{2r+\gamma+1}{2r+1}} \; \left(\frac{R}{\sigma} \right)^{\frac{2}{2r+\gamma}} \; .$$

Finally, setting $p_{max} = \max(p_1, ..., p_m)$, if

$$n > n_0 := (4m)^{\frac{2r+\gamma+1}{2r}} \; \max\Big(\; (R/\sigma)^{\frac{2}{2r+\gamma}} \; , (p_{max} \; C_\gamma)^{\frac{2r+\gamma+1}{2r}} \; (R/\sigma)^{\frac{2(\gamma+1)}{2r}} \; \Big) \tag{17}$$

we have

$$\mathcal{B}_{\frac{n}{m}}(T_j,\lambda_n) \leq 1 + \left(\frac{1}{2} + \frac{1}{2} \right)^2 = 2 \; ,$$

uniformly for any $j \in [m]$. □

Lemma 6. *If λ_n is defined by* (12)

$$\mathcal{B}_n(T,\lambda_n) \leq 2 \; ,$$

provided n is sufficiently large.

Proof of Lemma 6. The proof is a straightforward calculation using Definition (12) and recalling that $\mathcal{N}(T,\lambda) \leq C_\gamma \lambda^{-\gamma}$. □

F Miscellanea

Proposition 10 (Cordes Inequality,[1], Theorem IX.2.1-2)**.** *Let A, B be two bounded, self-adjoint and positive operators on a Hilbert space. Then for any $s \in [0, 1]$:*

$$\|A^s B^s\| \leq \|AB\|^s \ . \tag{18}$$

Lemma 7. *Let X be a non-negative random variable with $\mathbb{P}[X > C \log^u(k\eta^{-1})] < \eta$ for any $\eta \in (0, 1]$. Then $\mathbb{E}[X] \leq \frac{C}{k} u\Gamma(u)$.*

Proof. Apply $\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t]dt$. □

References

[1] R. Bhatia. *Matrix Analysis.* Springer, 1997.

[2] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 2017. doi:10.1007/s10208-017-9359-7.

[3] Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 2017.

[4] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems 28*, 2015.