Efficient and principled score estimation with Nyström kernel exponential families: Supplementary material

We now prove Theorems 1 and 2, as well as providing a finite-sample bound with explicit constants (Theorem 3).

In Appendix A, we begin with a review of necessary notation and definitions of all necessary objects, as well as an overview of relevant theory for the full kernel exponential family estimator by Sriperumbudur et al. (2017). In Appendix B, we establish a representer theorem for our Nyström estimator and prove Theorem 1. We address consistency and convergence in Appendix C, by first decomposing and bounding the error in Appendix C.1, then developing probabilistic inequalities in Appendix C.2, and finally collecting everything into a final bound to prove Theorem 2 in Appendix C.3. Appendix D establishes auxiliary results used in the proofs, including tools for dimension subsampling, and in particular a concentration inequality for sums of correlated random operators in Appendix D.2.

A Preliminaries

We will first establish some definitions that will be useful throughout, as well as overviewing some relevant results from Sriperumbudur et al. (2017).

A.1 Notation

Our notation is mostly standard: $\mathcal{H}$ is a reproducing kernel Hilbert space of functions $\Omega \subseteq \mathbb{R}^d \to \mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$, with a kernel $k : \Omega \times \Omega \to \mathbb{R}$ given by the reproducing property, $k(x, y) = \langle k(x, \cdot), k(y, \cdot) \rangle_{\mathcal{H}}$. The reproducing property for kernel derivatives (Steinwart and Christmann 2008, Lemma 4.34) will also be important: $\langle \partial_i k(x, \cdot), f \rangle_{\mathcal{H}} = \partial_i f(x)$ as long as k is differentiable; the same holds for higher-order derivatives.

We use $\|\cdot\|$ to denote the operator norm $\|A\| = \sup_{f:\|f\|_{\mathcal{H}} \leq 1} |\langle f, Af \rangle_{\mathcal{H}}|$, and A^* for the adjoint of an operator $A : \mathcal{H}_1 \to \mathcal{H}_2$, $\langle Af, g \rangle_{\mathcal{H}_2} = \langle f, A^* g \rangle_{\mathcal{H}_1}$. $\lambda_{\max}(A)$ denotes the algebraically largest eigenvalue of A. For elements $f \in \mathcal{H}_1$, $g \in \mathcal{H}_2$ we define $f \otimes g$ to be the tensor product, viewed as an operator from $\mathcal{H}_2$ to $\mathcal{H}_1$ with $(f \otimes g)h = f \langle g, h \rangle_{\mathcal{H}_2}$; note that $(f \otimes g)^* = g \otimes f$ and that $A(f \otimes g)B = (Af) \otimes (B^* g)$.

$C^1(\Omega)$ denotes the space of continuously differentiable functions on Ω, and $L^r(\Omega)$ the space of r-power Lebesgue-integrable functions.

As in the main text, $x_{(a,i)}$ will denote $x_{(a-1)d+i}$.

A.2 Operator definitions

The following objects will be useful in our study: C, ξ, and their estimators were defined by Sriperumbudur et al. (2017). C is similar to the standard covariance operator in similar analyses (Caponnetto and De Vito 2007; Rudi et al. 2015).

Definition 1. *Suppose we have a sample set $X = \{X_a\}_{a \in [n]} \subset \mathbb{R}^d$. For any $\lambda > 0$, define the following:*

$$C = \mathbb{E}_{x \sim p_0} \left[\sum_{i=1}^{d} \partial_i k(x, \cdot) \otimes \partial_i k(x, \cdot) \right] : \mathcal{H} \to \mathcal{H}; \qquad C_\lambda = C + \lambda I \tag{10}$$

$$\xi = -C f_0 = \mathbb{E}_{x \sim p_0} \left[\sum_{i=1}^{d} \partial_i k(x, \cdot) \partial_i \log q_0(x) + \partial_i^2 k(x, \cdot) \right] \in \mathcal{H} \tag{11}$$

$$Z_X = \sum_{b=1}^{n} \sum_{i=1}^{d} e_{(b,i)} \otimes \partial_i k(X_b, \cdot) : \mathcal{H} \to \mathbb{R}^{nd};$$

here $e_{(b,i)} \in \mathbb{R}^{nd}$ has component $(b-1)d + i$ equal to 1 and all others 0.

Define estimators of (10) and (11) by

$$\hat{C} = \frac{1}{n} Z_X^* Z_X = \frac{1}{n} \sum_{a=1}^{n} \sum_{i=1}^{d} \partial_i k(X_a, \cdot) \otimes \partial_i k(X_a, \cdot) : \mathcal{H} \to \mathcal{H}; \qquad \hat{C}_\lambda = \hat{C} + \lambda I \tag{12}$$

$$\hat{\xi} = \frac{1}{n} \sum_{a=1}^{n} \sum_{i=1}^{d} \partial_i k(X_a, \cdot) \partial_i \log q_0(X_a) + \partial_i^2 k(X_a, \cdot) \in \mathcal{H}. \tag{13}$$

Further define:

$$\mathcal{N}_\infty(\lambda) := \sup_{x \in \Omega} \sum_{i=1}^{d} \left\| C_\lambda^{-\frac{1}{2}} \partial_i k(x, \cdot) \right\|_{\mathcal{H}}^2$$

$$\mathcal{N}'_\infty(\lambda) := \sup_{\substack{x \in \Omega \\ i \in [d]}} \left\| C_\lambda^{-\frac{1}{2}} \partial_i k(x, \cdot) \right\|_{\mathcal{H}}^2.$$

Here, Z_X evaluates derivatives of its input at the points of X, $(Z_X f)_{(b,i)} = \partial_i f(X_b)$, whereas Z_X^* constructs linear combinations: for $\alpha \in \mathbb{R}^{nd}$, $Z_X^* \alpha = \sum_{b=1}^{n} \sum_{i=1}^{d} \alpha_{(b,i)} \partial_i k(X_b, \cdot)$.

A.3 Assumptions

We will need the following assumptions on p_0, q_0, and $\mathcal{H}$:

(**A**) (Well-specified) The true density is $p_0 = p_{f_0} \in \mathcal{P}$, for some $f_0 \in \mathcal{F}$.

(**B**) $\operatorname{supp} p_0 = \Omega$ is a non-empty open subset of $\mathbb{R}^d$, with a piecewise smooth boundary $\partial\Omega := \bar{\Omega} \setminus \Omega$, where $\bar{\Omega}$ denotes the closure of Ω.

(**C**) p_0 is continuously extensible to $\bar{\Omega}$. k is twice continuously differentiable on $\Omega \times \Omega$, with $\partial^{\alpha,\alpha} k$ continuously extensible to $\bar{\Omega} \times \bar{\Omega}$ for $|\alpha| \leq 2$.

(**D**) $\partial_i \partial_{i+d} k(x, x')|_{x'=x} p_0(x) = 0$ for $x \in \partial\Omega$, and for all sequences of $x \in \Omega$ with $\|x\|_2 \to \infty$ we have have $p_0(x) \sqrt{\partial_i \partial_{i+d} k(x, x')}\Big|_{x'=x} = o\left(\|x\|^{1-d}\right)$ for each $i \in [d]$.

(**E**) (Integrability) For all $i \in [d]$, each of

$$\partial_i \partial_{i+d} k(x, x')|_{x'=x}, \sqrt{\partial_i^2 \partial_{i+d}^2 k(x, x')}\Big|_{x'=x}, \partial_i \log q_0(x) \sqrt{\partial_i^2 \partial_{i+d}^2 k(x, x')}\Big|_{x'=x}$$

are in $L^1(\Omega, p_0)$. Moreover, $q_0 \in C^1(\Omega)$.

(**F**) (Range space) $f_0 \in \operatorname{range}(C^\beta)$ for some $\beta \geq 0$, and $\left\|C^{-\beta} f_0\right\|_{\mathcal{H}} < R$ for some $R < \infty$. The operator C is defined by (10).

(**G**) (Bounded derivatives) $\operatorname{supp}(q_0) = \mathcal{H}$, and the following quantities are finite:

$$\kappa_1^2 := \sup_{\substack{x \in \Omega \\ i \in [d]}} \partial_i \partial_{i+d} k(x, x')|_{x'=x}, \quad \kappa_2^2 := \sup_{\substack{x \in \Omega \\ i \in [d]}} \partial_i^2 \partial_{i+d}^2 k(x, x')\big|_{x'=x}, \quad Q := \sup_{\substack{x \in \Omega \\ i \in [d]}} |\partial_i \log q_0(x)|.$$

(**H**) (Bounded kernel) $\kappa^2 := \sup_{x \in \Omega} k(x, x)$ is finite.

These assumptions, or closely related ones, were all used by Sriperumbudur et al. (2017) for various parts of their analysis. Assumptions (**B**) to (**D**) ensure that the form for $J(p_0 \| p)$ in (2) is valid. Assumption (**E**) implies $J(p_0 \| p_f)$ is finite for any $p_f \in \mathcal{P}$. Assumption (**G**) is used to get probabilistic bounds on the convergence of the estimators, and implies Assumption (**E**). Note that $\kappa_2^2 < \infty$ and $Q < \infty$ can be replaced by $L^2(\Omega, p_0)$ integrability assumptions as in Sriperumbudur et al. (2017) without affecting the asymptotic rates, but $\kappa_1^2 < \infty$ is used to get Nyström-like rates. Assumption (**H**) is additionally needed for the convergence in L^r, Hellinger, and KL distances.

Note that under (**G**), $\mathcal{N}_\infty(\lambda) \leq d \mathcal{N}'_\infty(\lambda) \leq \frac{d\kappa_1^2}{\lambda}$, and $\|C\| \leq d\kappa_1^2$.

A.4 Full-data result

This result is essentially Theorem 3 of Sriperumbudur et al. (2017).

Lemma 1. *Under Assumptions (**A**) to (**E**),*

$$J(f) = J(p_0\|p_f) = \frac{1}{2}\langle f - f_0, C(f - f_0)\rangle_{\mathcal{H}} = \frac{1}{2}\langle f, Cf\rangle_{\mathcal{H}} + \langle f, \xi\rangle_{\mathcal{H}} + J(p_0\|q_0).$$

Thus for $\lambda > 0$, the unique minimizer of the regularized loss function $J_\lambda(f) = J(f) + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}}$ is

$$f_\lambda = \operatorname*{argmin}_{f\in\mathcal{H}} J_\lambda(f) = -C_\lambda^{-1}\xi = C_\lambda^{-1}Cf_0.$$

Using the estimators (12) and (13), define an empirical estimator of the loss function (3), up to the additive constant $J(p_0\|q_0)$, as

$$\hat{J}(f) = \frac{1}{2}\langle f, \hat{C}f\rangle_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}}.$$

There is a unique minimizer of $\hat{J}_\lambda(f) = \hat{J}(f) + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}}$:

$$f^m_{\lambda,n} = \operatorname*{argmin}_{f\in\mathcal{H}} \hat{J}_\lambda(f) = -\hat{C}_\lambda^{-1}\hat{\xi}.$$

$f^m_{\lambda,n}$ can be computed according to Theorem 4 of Sriperumbudur et al. (2017), using (4) and (5).

A.5 Subsampling

In our Nyström projections, we will consider a more general $\mathcal{H}_Y$ than (6), allowing any finite-dimensional subspace of $\mathcal{H}$.

Definition 2 (Subsampling operators)**.** *Let $Y = \{y_a\}_{a\in[m]} \subset \mathcal{H}$ be some basis set, and let its span be $\mathcal{H}_Y = \operatorname{span}(Y)$; note that (6) uses $y_{(a,i)} = \partial_i k(Y_a, \cdot)$. Then define*

$$Z_Y = \sum_{a=1}^{m} e_a \otimes y_a : \mathcal{H} \to \mathbb{R}^m;$$

let Z_Y have singular value decomposition $Z_Y = U\Sigma V^*$, where $\Sigma \in \mathbb{R}^{t\times t}$ for some $t \le M$. Note that $VV^* = P_Y$ is the orthogonal projection operator onto $\mathcal{H}_Y$, while V^*V is the identity on $\mathbb{R}^t$.

For an operator $A : \mathcal{H} \to \mathcal{H}$, let

$$g_Y(A) = V(V^*AV)^{-1}V^*. \tag{14}$$

The projected inverse function g_Y, defined by Rudi et al. (2015), will be crucial in our study, and so we first establish some useful properties of it.

Lemma 2 (Properties of g_Y)**.** *Let $A : \mathcal{H} \to \mathcal{H}$ be a positive operator, and define $A_\lambda = A + \lambda I$ for any $\lambda > 0$. The operator g_Y of (14) satisfies the following:*

(i) $g_Y(A)P_Y = g_Y(A)$,

(ii) $P_Y g_Y(A) = g_Y(A)$,

(iii) $g_Y(A_\lambda)A_\lambda P_Y = P_Y$,

(iv) $g_Y(A_\lambda) = (P_Y A P_Y + \lambda I)^{-1}P_Y$, *and*

(v) $\|A_\lambda^{\frac{1}{2}} g_Y(A_\lambda) A_\lambda^{\frac{1}{2}}\| \le 1$.

Proof. (i) and (ii) follow from $V^*P_Y = V^*VV^* = V^*$ and $P_YV = VV^*V = V$, respectively. (iii) is similar: $g_Y(A_\lambda)A_\lambda P_Y = V(V^*A_\lambda V)^{-1}V^*A_\lambda VV^* = VV^*$. For (iv),

$$P_Y = VV^* = V(V^*A_\lambda V)(V^*A_\lambda V)^{-1}V^* = V(V^*A_\lambda V)V^*V(V^*A_\lambda V)^{-1}V^*.$$

But $V(V^*A_\lambda V)V^* = V(V^*AV + \lambda V^*V)V^* = (P_YAP_Y + \lambda I)P_Y$, so we have

$$P_Y = (P_YAP_Y + \lambda I)P_Y g_Y(A_\lambda);$$

left-multiplying both sides by $(P_YAP_Y + \lambda I)^{-1}$ and using (ii) yields the desired result. Finally,

$$\begin{aligned}\left(A_\lambda^{\frac{1}{2}} g_Y(A_\lambda) A_\lambda^{\frac{1}{2}}\right)^2 &= A_\lambda^{\frac{1}{2}} g_Y(A_\lambda) A_\lambda g_Y(A_\lambda) A_\lambda^{\frac{1}{2}} \\ &= A_\lambda^{\frac{1}{2}} V(V^*A_\lambda V)^{-1}V^*A_\lambda V(V^*A_\lambda V)^{-1}V^*A_\lambda^{\frac{1}{2}} \\ &= A_\lambda^{\frac{1}{2}} V(V^*A_\lambda V)^{-1}V^*A_\lambda^{\frac{1}{2}} \\ &= A_\lambda^{\frac{1}{2}} g_Y(A_\lambda) A_\lambda^{\frac{1}{2}},\end{aligned}$$

so that $A_\lambda^{\frac{1}{2}} g_Y(A_\lambda) A_\lambda^{\frac{1}{2}}$ is a projection. Thus its operator norm is either 0 or 1, and (v) follows. □

B Representer theorem for Nyström optimization problem (Theorem 1)

We will first establish some representations for $f^m_{\lambda,n}$ in terms of operators on $\mathcal{H}$ (in Lemma 3), and then show Lemma 4, which generalizes Theorem 1. This parallels Appendix C of Rudi et al. (2015).

Lemma 3. *Under Assumptions (A) to (E), the unique minimizer of $\hat{J}(f) + \lambda\|f\|^2_{\mathcal{H}}$ in $\mathcal{H}_Y$ is*

$$f^m_{\lambda,n} = -(P_Y\hat{C}P_Y + \lambda I)^{-1}P_Y\hat{\xi} = -g_Y(\hat{C}_\lambda)\hat{\xi}. \tag{15}$$

Proof. We begin by rewriting the minimization using Lemma 1 as

$$\begin{aligned}f^m_{\lambda,n} &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \hat{J}_\lambda(f) \\ &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\langle f, \hat{C}f\rangle_{\mathcal{H}} + \langle f, \hat{\xi}\rangle_{\mathcal{H}} + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} \\ &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\langle P_Yf, \hat{C}P_Yf\rangle_{\mathcal{H}} + \langle P_Yf, \hat{\xi}\rangle_{\mathcal{H}} + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} \\ &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\left\langle \frac{1}{\sqrt{n}}Z_XP_Yf, \frac{1}{\sqrt{n}}Z_XP_Yf\right\rangle_{\mathcal{H}} + \langle f, P_Y\hat{\xi}\rangle_{\mathcal{H}} + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} \\ &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\left\|\frac{1}{\sqrt{n}}Z_XP_Yf\right\|^2_{\mathcal{H}} + \lambda\left\langle f, \frac{1}{\lambda}P_Y\hat{\xi}\right\rangle_{\mathcal{H}} + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} + \frac{1}{2}\lambda\left\|\frac{1}{\lambda}P_Y\hat{\xi}\right\|^2_{\mathcal{H}} \\ &= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\left\|\frac{1}{\sqrt{n}}Z_XP_Yf\right\|^2_{\mathcal{H}} + \frac{1}{2}\lambda\left\|f + \frac{1}{\lambda}P_Y\hat{\xi}\right\|^2_{\mathcal{H}}.\end{aligned}$$

This problem is strictly convex and coercive, thus a unique $f^m_{\lambda,n}$ exists. Now, for any $f \in \mathcal{H}$, we have

$$\left\|f + \frac{1}{\lambda}P_Y\hat{\xi}\right\|^2_{\mathcal{H}} = \left\|P_Yf + \frac{1}{\lambda}P_Y\hat{\xi}\right\|^2_{\mathcal{H}} + \|(I - P_Y)f\|^2_{\mathcal{H}},$$

so that the problem

$$\operatorname*{argmin}_{f\in\mathcal{H}} \frac{1}{2}\left\|\frac{1}{\sqrt{n}}Z_XP_Yf\right\|^2_{\mathcal{H}} + \frac{1}{2}\lambda\left\|f + \frac{1}{\lambda}P_Y\hat{\xi}\right\|^2_{\mathcal{H}}$$

will yield a solution in $\mathcal{H}_Y$. This problem is also strictly convex and coercive, so its unique solution must be $f^m_{\lambda,n}$. By differentiating the objective, we can then see that

$$\begin{aligned}\tfrac{1}{n}P_YZ_X^*Z_Xf^m_{\lambda,n} + \lambda f^m_{\lambda,n} + P_Y\hat{\xi} &= 0 \\ \left(P_Y\hat{C}P_Y + \lambda I\right)f^m_{\lambda,n} &= -P_Y\hat{\xi},\end{aligned}$$

which since $\hat{C}$ is positive yields the first equality of (15). The second follows from Lemma 2 (iv). □

Lemma 4 (Generalization of Theorem 1). *Under Assumptions (A) to (E), $f^m_{\lambda,n}$ can be computed as*

$$f^m_{\lambda,n} = Z^*_Y \beta_Y = \sum_{a=1}^{m} (\beta_Y)_a y_a$$
$$\beta_Y = -(\tfrac{1}{n} B_{XY}^\mathsf{T} B_{XY} + \lambda G_{YY})^\dagger h_Y, \tag{16}$$

where $B_{XY} \in \mathbb{R}^{nd \times m}, G_{YY} \in \mathbb{R}^{m \times m}$, $h_Y \in \mathbb{R}^m$ are given by

$$(B_{XY})_{(b,i),a} = \langle \partial_i k(X_b, \cdot), y_a \rangle_{\mathcal{H}} \tag{17}$$
$$(G_{YY})_{a,a'} = \langle y_a, y_{a'} \rangle_{\mathcal{H}}$$
$$(h_Y)_a = \langle \hat{\xi}, y_a \rangle_{\mathcal{H}}.$$

Proof. First, $B_{XY} = Z_X Z_Y^*$, $G_{YY} = Z_Y Z_Y^*$, and $h_Y = Z_Y \hat{\xi}$. For example, (17) agrees with

$$\begin{aligned} Z_X Z_Y^* &= \left[\sum_{b=1}^{n} \sum_{i=1}^{d} e_{(b,i)} \otimes \partial_i k(X_b, \cdot) \right] \left[\sum_{a=1}^{m} y_a \otimes e_a \right] \\ &= \sum_{b=1}^{n} \sum_{i=1}^{d} \sum_{a=1}^{m} \langle \partial_i k(X_b, \cdot), y_a \rangle_{\mathcal{H}} \left[e_{(b,i)} \otimes e_a \right]. \end{aligned}$$

Recall the full-rank factorization of pseudo-inverses: if a matrix A of rank r can be written as $A = FG$ for F, G each of rank r, then $A^\dagger = G^\dagger F^\dagger$ (Ben-Israel and Greville 2003, chap. 1, sec. 6, ex. 17).

Now we can show that the claimed form (16) matches $f^m_{\lambda,n}$ from (15):

$$\begin{aligned} -Z_Y^* \left(\tfrac{1}{n} B_{XY}^\mathsf{T} B_{XY} + \lambda G_{YY} \right)^\dagger h_Y &= -Z_Y^* \left(\tfrac{1}{n} Z_Y Z_X^* Z_X Z_Y^* + \lambda Z_Y Z_Y^* \right)^\dagger Z_Y \hat{\xi} \\ &= -Z_Y^* \left(Z_Y \hat{C}_\lambda Z_Y^* \right)^\dagger Z_Y \hat{\xi} \\ &= -V \Sigma U^* \left((U\Sigma)(V^* \hat{C}_\lambda V) \Sigma U^* \right)^\dagger U \Sigma V^* \hat{\xi} \\ &= -V \Sigma U^* (\Sigma U^*)^\dagger (V^* \hat{C}_\lambda V)^\dagger (U\Sigma)^\dagger U \Sigma V^* \hat{\xi} \\ &= -V \Sigma U^* U \Sigma^{-1} (V^* \hat{C}_\lambda V)^{-1} \Sigma^{-1} U^* U \Sigma V^* \hat{\xi} \\ &= -V (V^* \hat{C}_\lambda V)^{-1} V^* \hat{\xi} \\ &= -g_Y(\hat{C}_\lambda) \hat{\xi} = f^m_{\lambda,n}. \end{aligned}$$

□

Theorem 1 is the specialization of Lemma 4 to $y_{(a,i)} = \partial_i k(Y_a, \cdot)$.

B.1 Relationship to "lite" kernel exponential families

The lite kernel exponential family of Strathmann et al. (2015) obtains a solution in $\mathcal{H}'_Y = \operatorname{span}\{k(y, \cdot)\}_{y \in Y}$, where in that paper it was assumed that $Y = X$, $k(x, y) = \exp\left(-\tau^{-1} \|x - y\|^2\right)$, and q_0 was uniform. Their estimator, given by their Proposition 1, is

$$\alpha = -\frac{\tau}{2} (A + \lambda I)^{-1} b \tag{18}$$
$$A = \sum_{i=1}^{d} -[D_{x_i} K - K D_{x_i}]^2 \qquad b = \sum_{i=1}^{d} \left(\frac{2}{\tau} (K s_i + D_{s_i} K \mathbf{1} - 2 D_{x_i} K x_i) - K \mathbf{1} \right)$$

where $x_i = \begin{bmatrix} X_{1i} & \dots & X_{ni} \end{bmatrix}^\mathsf{T}$, $s_i = x_i \odot x_i$ with $\odot$ the elementwise product, $D_x = \operatorname{diag}(x)$, and $K \in \mathbb{R}^{m \times m}$ has entries $K_{aa'} = k(X_a, X_{a'})$.

Lemma 4 allows us to optimize over $\mathcal{H}'_Y$; we need not restrict ourselves to $Y = X$, uniform q_0, or a Gaussian kernel. Here $y_a = k(Y_a, \cdot)$, and we obtain

$$\beta'_Y = -\left(\frac{1}{n}(B'_{XY})^\intercal B'_{XY} + \lambda G'_{YY}\right)^\dagger h'_Y.$$

Using that for the Gaussian kernel k

$$\partial_i k(x,y) = -\frac{2}{\tau}(x_i - y_i)k(x,y) \qquad \partial^2_{i+d} k(x,y) = \frac{2}{\tau}\left[\frac{2}{\tau}(x_i - y_i)^2 - 1\right]k(x,y),$$

we can obtain with some algebra similar to the proof of Strathmann et al. (2015)'s Proposition 1 that when $Y = X$ and q_0 is uniform,

$$h'_X = \frac{2}{n\tau}b \qquad (B'_{XX})^\intercal B'_{XX} = \frac{4}{\tau^2}A \qquad G'_{XX} = K.$$

Thus

$$\beta'_X = -\left(\frac{4}{n\tau^2}A + \lambda K\right)^\dagger \frac{2}{n\tau}b = -\frac{\tau}{2}\left(A + \frac{1}{4}n\tau^2\lambda K\right)^\dagger b. \tag{19}$$

(19) resembles (18), except that our approach regularizes A with $\frac{1}{4}n\tau^2\lambda K$ rather than λI. This is because, despite claims by Strathmann et al. (2015) in both the statement and the proof of their Proposition 1 that they minimize $\hat{J}(f) + \lambda\|f\|^2_{\mathcal{H}}$, they in fact minimize $\hat{J}(f) + \frac{1}{2}n\tau^2\lambda\|\alpha\|^2_2$. Our solutions otherwise agree.

C Consistency and convergence rate of the estimator (Theorem 2)

To prove the consistency and convergence of $f^m_{\lambda,n}$, we will first bound the difference between $f^m_{\lambda,n}$ in terms of various quantities (Appendix C.1), which we will then study individually in Appendix C.2 to yield the final result in Appendix C.3. Appendix D gives auxiliary results used along the way.

C.1 Decomposition

We care both about the parameter convergence $\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}}$ and the convergence of $p^m_{\lambda,n} = p_{f^m_{\lambda,n}}$ to p_0 in various distances. But by Lemma 1, we know that $J(p_0\|p^m_{\lambda,n}) = \frac{1}{2}\left\|C^{\frac{1}{2}}(f^m_{\lambda,n} - f_0)\right\|^2_{\mathcal{H}}$. Lemma 20 additionally shows that the L^r, KL, and Hellinger distances between the distributions can be bounded in terms of $\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}}$. Thus it suffices to bound $\|C^\alpha(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}}$ for $\alpha \geq 0$.

Lemma 5. *Under Assumptions (A) to (F), let $\alpha \geq 0$ and define*

$$c(a) := \lambda^{\min\left(0,\ a-\frac{1}{2}\right)}\|C\|^{\max\left(0,\ a-\frac{1}{2}\right)}, \qquad \mathcal{C}_Y := \|C^{\frac{1}{2}}_\lambda(I - VV^*)\|^2.$$

Then

$$\|C^\alpha(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}} \leq R\,(2\mathcal{C}_Y + \lambda)\,c(\alpha)c(\beta) + \frac{1}{\sqrt{\lambda}}\left\|C^\alpha\hat{C}^{-\frac{1}{2}}_\lambda\right\|\left(\|\hat{\xi} - \xi\|_{\mathcal{H}} + \|\hat{C} - C\|R\left(\left(\frac{2\mathcal{C}_Y}{\sqrt{\lambda}} + \sqrt{\lambda}\right)c(\beta) + \|C\|^\beta\right)\right).$$

Proof. We will decompose the error with respect to the best estimator for a fixed basis:

$$\begin{aligned} f^m_\lambda &:= \operatorname*{argmin}_{f\in\mathcal{H}_Y} \frac{1}{2}\langle f, P_Y C P_Y f\rangle_{\mathcal{H}} + \langle f, P_Y\xi\rangle_{\mathcal{H}} + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} \\ &= -(P_Y C P_Y + \lambda I)^{-1}P_y\xi = -g_Y(C_\lambda)\xi = g_Y(C_\lambda)Cf_0. \end{aligned}$$

Then we have

$$\|C^\alpha(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}} \leq \|C^\alpha(f^m_{\lambda,n} - f^m_\lambda)\|_{\mathcal{H}} + \|C^\alpha(f^m_\lambda - f_0)\|_{\mathcal{H}}. \tag{20}$$

We'll tackle the second term first.

Approximation error This term covers both approximation due to the basis $\mathcal{H}_Y$ and the bias due to regularization. We'll break it down using some ideas from the proof of Rudi et al. (2015)'s Theorem 2:

$$\begin{aligned} f_0 - f_\lambda^m &= (I - g_Y(C_\lambda)C) f_0 \\ &= (I - g_Y(C_\lambda)C_\lambda + \lambda g_Y(C_\lambda))\, f_0 \\ &= (I - g_Y(C_\lambda)C_\lambda(VV^*) - g_Y(C_\lambda)C_\lambda(I - VV^*) + \lambda g_Y(C_\lambda))\, f_0 \\ &= ((I - VV^*) - g_Y(C_\lambda)C_\lambda(I - VV^*) + \lambda g_Y(C_\lambda))\, f_0, \end{aligned}$$

where in the last line we used Lemma 2 (iii). Thus, using Assumption (**F**) and Lemma 2 (v),

$$\begin{aligned} \|C^\alpha(f_\lambda^m - f_0)\|_{\mathcal{H}} &\le \|C^\alpha(I - VV^*)f_0\|_{\mathcal{H}} + \|C^\alpha g_Y(C_\lambda)C_\lambda(I - VV^*)f_0\|_{\mathcal{H}} + \lambda \|C^\alpha g_Y(C_\lambda)f_0\|_{\mathcal{H}} \\ &\le \underbrace{\left\|C^\alpha C_\lambda^{-\frac{1}{2}}\right\|}_{\mathcal{S}_\alpha} \left\|C_\lambda^{\frac{1}{2}}(I - VV^*)C^\beta\right\| \underbrace{\left\|C^{-\beta} f_0\right\|_{\mathcal{H}}}_{\le R} \\ &\quad + \underbrace{\left\|C^\alpha C_\lambda^{-\frac{1}{2}}\right\|}_{\mathcal{S}_\alpha} \underbrace{\left\|C_\lambda^{\frac{1}{2}} g_Y(C_\lambda) C_\lambda^{\frac{1}{2}}\right\|}_{\le 1} \left\|C_\lambda^{\frac{1}{2}}(I - VV^*)C^\beta\right\| \underbrace{\left\|C^{-\beta} f_0\right\|_{\mathcal{H}}}_{\le R} \\ &\quad + \lambda \underbrace{\left\|C^\alpha C_\lambda^{-\frac{1}{2}}\right\|}_{\mathcal{S}_\alpha} \underbrace{\left\|C_\lambda^{\frac{1}{2}} g_Y(C_\lambda) C_\lambda^{\frac{1}{2}}\right\|}_{\le 1} \underbrace{\left\|C_\lambda^{-\frac{1}{2}} C^\beta\right\|_{\mathcal{H}}}_{\mathcal{S}_\beta} \underbrace{\left\|C^{-\beta} f_0\right\|_{\mathcal{H}}}_{\le R}. \end{aligned}$$

Because $(I - VV^*)$ is a projection, we have

$$\left\|C_\lambda^{\frac{1}{2}}(I - VV^*)C^\beta\right\| \le \left\|C_\lambda^{\frac{1}{2}}(I - VV^*)^2 C_\lambda^{\frac{1}{2}}\right\| \left\|C_\lambda^{-\frac{1}{2}} C^\beta\right\| \le \left\|C_\lambda^{\frac{1}{2}}(I - VV^*)\right\|^2 \mathcal{S}_\beta.$$

We can also bound the terms $\mathcal{S}_a$ as follows. When $a \ge \frac{1}{2}$, the function $x \mapsto x^a/\sqrt{x+\lambda}$ is increasing on $[0, \infty)$, so that

$$\mathcal{S}_a = \left\|C_\lambda^{-\frac{1}{2}} C^a\right\|_{\mathcal{H}} = \frac{\|C\|^a}{\sqrt{\|C\| + \lambda}} \le \|C\|^{a - \frac{1}{2}}.$$

When instead $0 \le a < \frac{1}{2}$, we have that

$$\mathcal{S}_a = \left\|C_\lambda^{-\frac{1}{2}} C^a\right\|_{\mathcal{H}} \le \max_{x \ge 0} \frac{x^a}{\sqrt{x+\lambda}} = \sqrt{2} a^a \left(\tfrac{1}{2} - a\right)^{\frac{1}{2} - a} \lambda^{a - \frac{1}{2}} \le \lambda^{a - \frac{1}{2}}.$$

Combining the two yields

$$\mathcal{S}_a \le \lambda^{\min\left(0,\ a - \frac{1}{2}\right)} \|C\|^{\max\left(0,\ a - \frac{1}{2}\right)} = c(a),$$

and so

$$\|C^\alpha(f_\lambda^m - f_0)\|_{\mathcal{H}} \le R \left(2 \left\|C_\lambda^{\frac{1}{2}}(I - VV^*)\right\|^2 + \lambda\right) c(\alpha) c(\beta). \tag{21}$$

Estimation error Let $D = P_Y C P_Y$, $\hat{D} = P_Y \hat{C} P_Y$. Then

$$f_\lambda^m = -(D + \lambda I)^{-1} P_Y \xi = -\frac{1}{\lambda}(D + \lambda I - D)(D + \lambda I)^{-1} P_Y \xi = -\frac{1}{\lambda}(P_Y \xi + D f_\lambda^m),$$

and so the error due to finite n is

$$\begin{aligned} f_\lambda^m - f_{\lambda,n}^m &= (\hat{D} + \lambda I)^{-1} P_Y \hat{\xi} + f_\lambda^m \\ &= (\hat{D} + \lambda I)^{-1} \left(P_Y \hat{\xi} + (\hat{D} + \lambda I) f_\lambda^m\right) \\ &= (\hat{D} + \lambda I)^{-1} \left(P_Y \hat{\xi} + \hat{D} f_\lambda^m + \lambda f_\lambda^m\right) \\ &= (\hat{D} + \lambda I)^{-1} \left(P_Y \hat{\xi} + \hat{D} f_\lambda^m - P_Y \xi - D f_\lambda^m\right) \\ &= (\hat{D} + \lambda I)^{-1} \left(P_Y (\hat{\xi} - \xi) + (\hat{D} - D) f_\lambda^m\right) \\ &= (\hat{D} + \lambda I)^{-1} \left(P_Y (\hat{\xi} - \xi) + (\hat{D} - D)(f_\lambda^m - f_0) + (\hat{D} - D) f_0\right). \end{aligned}$$

We thus have, using $\|P_Y\| \le 1$,

$$\left\|C^\alpha(f_\lambda^m - f_{\lambda,n}^m)\right\|_{\mathcal{H}} \le \left\|C^\alpha(P_Y\hat{C}P_Y + \lambda I)^{-1}P_Y\right\| \left(\|\hat{\xi}-\xi\|_{\mathcal{H}} + \|\hat{C}-C\|\|f_\lambda^m - f_0\|_{\mathcal{H}} + \|\hat{C}-C\|\,\|f_0\|_{\mathcal{H}}\right).$$

We have already bounded $\|f_\lambda^m - f_0\|_{\mathcal{H}}$, and have $\|f_0\|_{\mathcal{H}} \le \|C^\beta\|\|C^{-\beta}f_0\|_{\mathcal{H}} \le R\|C\|^\beta$. Using Lemma 2 (iv) and (v), we have

$$\begin{aligned}\left\|C^\alpha(P_Y\hat{C}P_Y + \lambda I)^{-1}P_Y\right\| = \left\|C^\alpha g_Y(\hat{C}_\lambda)\right\| &\le \left\|C^\alpha\hat{C}_\lambda^{-\frac{1}{2}}\right\| \left\|\hat{C}_\lambda^{\frac{1}{2}} g_Y(\hat{C}_\lambda)\hat{C}_\lambda^{\frac{1}{2}}\right\| \left\|\hat{C}_\lambda^{-\frac{1}{2}}\right\| \\ &\le \frac{1}{\sqrt{\lambda}}\left\|C^\alpha\hat{C}_\lambda^{-\frac{1}{2}}\right\|,\end{aligned}$$

and so

$$\left\|C^\alpha(f_\lambda^m - f_{\lambda,n}^m)\right\|_{\mathcal{H}} \le \frac{\left\|C^\alpha\hat{C}_\lambda^{-\frac{1}{2}}\right\|}{\sqrt{\lambda}}\left(\|\hat{\xi}-\xi\|_{\mathcal{H}} + \|\hat{C}-C\|\left(\|f_\lambda^m - f_0\|_{\mathcal{H}} + R\|C\|^\beta\right)\right). \tag{22}$$

The claim follows by using (21) and (22) in (20). □

C.1.1 Remark on unimportance of $\partial_i^2 k(x,\cdot)$ terms in the basis

This decomposition gives some intuition about why terms of the form $\partial_i^2 k(x,\cdot)$, which are included in the basis of the full-data solution but missing from our solution even when $Y = X$, appear to be unimportant (as we also observe empirically).

The only term in the error decomposition depending on the specific basis chosen is the projection error term $\|C_\lambda^{\frac{1}{2}}(I - VV^*)\|$. Because the $\partial_i^2 k(x,\cdot)$ directions are not particularly aligned with C, unlike the $\partial_i k(x,\cdot)$ terms, whether they are included or not should not have a major effect on this term and therefore does not strongly affect the bound.

Moreover, the primary places where Lemma 5 discards dependence on the basis are that in the estimation error term, we bounded each of $\|P_Y(\hat{\xi}-\xi)\|$, $\|P_Y(\hat{C}-C)P_Y\|$, and $\|C^\alpha(P_Y\hat{C}P_Y+\lambda I)^{-1}P_Y\|$ terms by simply dropping the P_Y. For the C-based terms, we again expect that the $\partial_i^2 k(x,\cdot)$ terms do not have a strong effect on the given norms. Thus the only term that should be very directly affected is $\|P_Y(\hat{\xi}-\xi)\|$; but since we expect that $\hat{\xi}\to\xi$ relatively quickly compared to the convergence of $\hat{C}\to C$, this term should not be especially important to the overall error.

C.2 Probabilistic inequalities

We only need Lemma 5 for $\alpha = 0$ and $\alpha = \frac{1}{2}$; in the former case, we use $\left\|\hat{C}_\lambda^{-\frac{1}{2}}\right\| \le 1/\sqrt{\lambda}$. Thus we are left with four quantities to control: $\|C^{\frac{1}{2}}\hat{C}_\lambda^{-\frac{1}{2}}\|$, $\mathcal{C}_Y = \|C_\lambda^{\frac{1}{2}}(I-VV^*)\|^2$, $\|\hat{\xi}-\xi\|_{\mathcal{H}}$, and $\|\hat{C}-C\|$.

Lemma 6. *Let $\rho,\delta \in (0,1)$. Under Assumptions (B) to (E) and (G), for any $0 < \lambda \le \frac{1}{3}\|C\|$, we have with probability at least $1-\delta$ that*

$$\|C^{\frac{1}{2}}\hat{C}_\lambda^{-\frac{1}{2}}\| \le \frac{1}{\sqrt{1-\rho}}$$

as long as

$$n \ge \max\left(\frac{4}{3\rho}, \frac{40d\mathcal{N}'_\infty(\lambda)}{\rho^2}\right)\log\frac{40\operatorname{Tr}C}{\lambda\delta}.$$

Proof. Let $\gamma := \lambda_{\max}\left(C_\lambda^{-\frac{1}{2}}(C-\hat{C})C_\lambda^{-\frac{1}{2}}\right)$. Lemma 19 gives that $\|C^{\frac{1}{2}}\hat{C}_\lambda^{-\frac{1}{2}}\| \le \frac{1}{\sqrt{1-\gamma}}$. We bound γ with Lemma 17, using $Y_i^a = \partial_i k(X_a,\cdot)$ so that $\mathbb{E}\sum_{i=1}^d Y_i^a \otimes Y_i^a = C$. This gives us that $\gamma \le \rho$ with probability at least $1-\delta$ as long as

$$\rho \le \frac{2w}{3n} + \sqrt{\frac{10d\mathcal{N}'_\infty(\lambda)w}{n}},$$

which is satisfied by the condition on n. □

Lemma 7. *Sample m points $\{Y_a\}_{a\in[m]}$ iid from p_0, and construct a subspace $\mathcal{H}_Y$ from those points in a way determined below; let VV^* be the orthogonal projection onto $\mathcal{H}_Y$. Choose $\rho, \delta \in (0,1)$, and assume that $\lambda \leq \frac{1}{3}\|C\|$. Then, under Assumptions (**B**) to (**E**) and (**G**)*

$$\mathcal{C}_Y = \|C_\lambda^{\frac{1}{2}}(I - VV^*)\|^2 \leq \frac{\lambda}{1-\rho}$$

with probability at least $1-\delta$ in each of the following cases:

(i) We put all components of the m points in our basis: $Y = \{\partial_i k(Y_a, \cdot)\}_{a\in[m]}^{i\in[d]}$, so that we have md components. We require

$$m \geq \max\left(\frac{4}{3\rho}, \frac{40d\mathcal{N}'_\infty(\lambda)}{\rho^2}\right)\log\left(\frac{40}{\lambda\delta}\operatorname{Tr}(C)\right).$$

(ii) Include each of the md components $\partial_i k(Y_a, \cdot)$ with probability p, so that the total number of components is distributed randomly as $\operatorname{Binomial}(md, p)$. The statement holds as long as

$$m \geq \max\left(\frac{4}{3\rho}, \frac{40\left(d+\frac{1}{p}-1\right)\mathcal{N}'_\infty(\lambda)}{\rho^2}\right)\log\left(\frac{40}{\lambda\delta}\operatorname{Tr}(C)\frac{d+\frac{1}{p}-1}{d+15\left(\frac{1}{p}-1\right)}\right).$$

(iii) For each of the m data points, we choose $\ell \in [1, d]$ components uniformly at random without replacement, so that we have $m\ell$ components. Assume here that $d > 1$; otherwise we necessarily have $\ell = d = 1$, covered by case (i). The statement holds as long as

$$m \geq \max\left(\frac{4}{3\rho}, \frac{40d\mathcal{N}'_\infty(\lambda)}{\rho^2}\right)\log\left(\frac{40}{\lambda\delta}\operatorname{Tr}(C)\left(1+14\frac{d-\ell}{\ell(d-1)}\right)\right).$$

Proof. Define the random operator $R_Y : \mathcal{H} \to \mathbb{R}^{md}$ by $R_Y := \frac{1}{\sqrt{m}}\sum_{a=1}^m\sum_{i=1}^d \frac{1}{p_{ai}} e_{ai} \otimes \partial_i k(Y_a, \cdot)$, where p_{ai} is the probability that the corresponding component is included in the basis. Since $p_{ai} > 0$ for each (a,i) in these setups, the operator R_Y is bounded. Note that $\overline{\operatorname{range} Z^*} = \operatorname{range} P_Y = \mathcal{H}_Y$ and that $\|C_\lambda^{\frac{1}{2}}(I-VV^*)\|^2 = \|(I-VV^*)C_\lambda^{\frac{1}{2}}\|^2$ as $C_\lambda^{\frac{1}{2}}$ is symmetric. Thus we can apply Lemmas 18 and 19 to observe that

$$\|C_\lambda^{\frac{1}{2}}(I-VV^*)\|^2 \leq \lambda\left\|(R_Y^*R_Y + \lambda I)^{-\frac{1}{2}}C_\lambda^{\frac{1}{2}}\right\|^2 \leq \frac{\lambda}{1-\lambda_{\max}\left(C_\lambda^{-\frac{1}{2}}\left(C - R_Y^*R_Y\right)C_\lambda^{-\frac{1}{2}}\right)}.$$

It remains to bound the relevant eigenvalue by ρ. We do so with the results of Appendix D.2: Lemma 17 for (i), Lemma 15 for (ii), and Lemma 16 for (iii). □

For the remaining two quantities, we use simple Hoeffding bounds:[2]

Lemma 8 (Concentration of $\hat{\xi}$)**.** *Under Assumption (**G**), with probability at least $1-\delta$ we have*

$$\|\hat{\xi} - \xi\|_{\mathcal{H}} \leq \frac{2d(Q\kappa_1 + \kappa_2)}{\sqrt{n}}\left(1 + \sqrt{2\log\tfrac{1}{\delta}}\right).$$

Proof. Let

$$\nu_a := \sum_{i=1}^d \left(\partial_i \log q_0(X_a)\partial_i k(X_a, \cdot) + \partial_i^2 k(X_a, \cdot)\right) - \xi,$$

so that $\hat{\xi} - \xi = \frac{1}{n}\sum_{a=1}^n \nu_a$, and for each a we have that $\mathbb{E}\,\nu_a = 0$ and

$$\|\nu_a\|_{\mathcal{H}} \leq 2\sup_{x\in\Omega}\left\|\sum_{i=1}^d \partial_i \log q_0(x)\partial_i k(x, \cdot) + \partial_i^2 k(x, \cdot)\right\|_{\mathcal{H}} \leq 2d\left(Q\kappa_1 + \kappa_2\right).$$

Applying Lemma 10 to the vectors ν_a gives the result. □

[2] A Bernstein bound would allow for a slightly better result when κ_1 and κ_2 are large, at the cost of a more complex form.

Lemma 9 (Concentration of $\hat{C}$)**.** *Under Assumption (G), with probability at least $1-\delta$ we have*

$$\|\hat{C} - C\| \le \frac{2d\kappa_1^2}{\sqrt{n}} \left(1 + \sqrt{2\log\tfrac{1}{\delta}}\right).$$

Proof. Let

$$C_x := \sum_{i=1}^{d} \partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot),$$

so that $\hat{C} = \frac{1}{n}\sum_{a=1} nC_{X_a}$, $C = \mathbb{E}\, C_x$. We know that

$$\|C_x - C\| \le 2\sum_{i=1}^{d} \|\partial_i k(x,\cdot)\|_{\mathcal{H}}^2 \le 2d\kappa_1^2$$
$$\|C_x - C\|_{\mathrm{HS}} \le 2\sum_{i=1}^{d} \sup_{x\in\Omega} \|\partial_i k(x,\cdot)\|_{\mathcal{H}}^2 \le 2d\kappa_1^2,$$

so applying Lemma 11 shows the result. □

C.3 Final bound

Theorem 3 (Finite-sample convergence of $f^m_{\lambda,n}$)**.** *Under Assumptions (A) to (G), let $\delta \in (0,1)$ and define $S_\delta := 1 + \sqrt{2\log\frac{4}{\delta}}$. Sample basis m points $\{Y_a\}_{a\in[m]}$ iid from p_0, not necessarily independent of X, and choose a basis as:*

(i) All d components $\{\partial_i k(Y_a,\cdot)\}_{a\in[m]}^{i\in[d]}$: set $w := 1$, $r := 0$.

(ii) A random subset, choosing each of the md components $\partial_i k(Y_a,\cdot)$ independently with probability p: set $w := \frac{dp+1-p}{dp+80(1-p)/3}$, $r := \frac{1}{p} - 1$.

(iii) A random subset, choosing ℓ components $\partial_i k(Y_a,\cdot)$ uniformly without replacement for each of the m points: set $w := 1 + 14\frac{d-\ell}{\ell(d-1)}$, $r := 0$. (If $d = 1$, use case (i).)

Assume that $0 < \lambda < \frac{1}{3}\|C\|$. When

$$m \ge \frac{90(d+r)\kappa_1^2}{\lambda}\log\frac{160d\kappa_1^2 w}{\lambda\delta} \quad \text{and} \quad n \ge \frac{90d\kappa_1^2}{\lambda}\log\frac{160d\kappa_1^2}{\lambda\delta},$$

we have with probability at least $1-\delta$ that both of the following hold simultaneously:

$$\begin{aligned}\|f^m_{\lambda,n} - f_0\|_{\mathcal{H}} &\le 7R\lambda^{\min\left(\frac{1}{2},\ \beta\right)}(d\kappa_1^2)^{\max\left(0,\ \beta-\frac{1}{2}\right)} \\ &\quad + \frac{2d}{\lambda\sqrt{n}}\left(Q\kappa_1 + \kappa_2 + R\kappa_1^2\left(7\lambda^{\min\left(\frac{1}{2},\ \beta\right)}(d\kappa_1^2)^{\max\left(0,\ \beta-\frac{1}{2}\right)} + (d\kappa_1^2)^{\beta}\right)\right)S_\delta \\ \|C^{\frac{1}{2}}(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}} &\le 7R\lambda^{\min\left(1,\ \beta+\frac{1}{2}\right)}(d\kappa_1^2)^{\max\left(0,\ \beta-\frac{1}{2}\right)} \\ &\quad + \frac{2d\sqrt{3}}{\sqrt{\lambda n}}\left(Q\kappa_1 + \kappa_2 + R\kappa_1^2\left(7\lambda^{\min\left(\frac{1}{2},\ \beta\right)}(d\kappa_1^2)^{\max\left(0,\ \beta-\frac{1}{2}\right)} + (d\kappa_1^2)^{\beta}\right)\right)S_\delta.\end{aligned}$$

Proof. Recall from Lemma 5 that

$$\begin{aligned}\|C^\alpha(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}} &\le R\left(2\mathcal{C}_Y + \lambda\right)c(\alpha)c(\beta) \\ &\quad + \frac{1}{\sqrt{\lambda}}\left\|C^\alpha\hat{C}_\lambda^{-\frac{1}{2}}\right\|\left(\|\hat{\xi} - \xi\|_{\mathcal{H}} + \|\hat{C} - C\|R\left(\left(\frac{2\mathcal{C}_Y}{\sqrt{\lambda}} + \sqrt{\lambda}\right)c(\beta) + \|C\|^\beta\right)\right),\end{aligned}$$

for $c(\alpha) = \lambda^{\min\left(0,\alpha-\frac{1}{2}\right)}\|C\|^{\max\left(0,\alpha-\frac{1}{2}\right)}$.

We'll use a union bound over the results of Lemmas 6 to 9. Note that under Assumption (G), each of $\|C\|$ and $\operatorname{Tr} C$ are at most $d\kappa_1^2$ and $\mathcal{N}'_\infty(\lambda) \le \kappa_1^2/\lambda$.

We first use $\rho = \frac{2}{3}$ in Lemmas 6 and 7 to get that $\|C^{\frac{1}{2}}\hat{C}_\lambda^{-\frac{1}{2}}\| \le \sqrt{3}$ and $\mathcal{C}_Y \le 3\lambda$ with probability at least $\frac{\delta}{2}$ when n and m are each at least

$$\max\left(2,\ 90(d+r)\mathcal{N}'_\infty(\lambda)\right)\log\frac{40\operatorname{Tr}(C)w}{\lambda\frac{\delta}{4}} \le \frac{90(d+r)\kappa_1^2}{\lambda}\log\frac{160d\kappa_1^2 w}{\lambda\delta},$$

where for m we use r and w as defined in the statement, and for n we use $r = 0$, $w = 1$; we also used that $\lambda < \frac{1}{3}\|C\|$ to resolve the max. The claim follows from applying Lemmas 8 and 9. □

Theorem 2 now follows from considering the asymptotics of Theorem 3, once we additionally make Assumption (H):

Proof of Theorem 2. Let $b := \min\left(\frac{1}{2},\ \beta\right)$. Under Assumptions (A) to (G), as $n \to \infty$ Theorem 3 gives:

$$\|f_{\lambda,n}^m - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}\left(\lambda^b + n^{-\frac{1}{2}}\lambda^{-1} + n^{-\frac{1}{2}}\lambda^{b-1}\right) = \mathcal{O}_{p_0}\left(\lambda^b + n^{-\frac{1}{2}}\lambda^{-1}\right)$$
$$\|C^{\frac{1}{2}}(f_{\lambda,n}^m - f_0)\|_{\mathcal{H}} = \mathcal{O}_{p_0}\left(\lambda^{b+\frac{1}{2}} + n^{-\frac{1}{2}}\lambda^{-\frac{1}{2}} + n^{-\frac{1}{2}}\lambda^{b-\frac{1}{2}}\right) = \mathcal{O}_{p_0}\left(\lambda^{b+\frac{1}{2}} + n^{-\frac{1}{2}}\lambda^{-\frac{1}{2}}\right)$$

as long as $\min(n,m) = \Omega(\lambda^{-1}\log\lambda^{-1})$. Choosing $\lambda = n^{-\theta}$, this requirement is $\min(n,m) = \Omega(n^\theta \log n)$ and the bounds become

$$\|f_{\lambda,n}^m - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}\left(n^{-b\theta} + n^{\theta-\frac{1}{2}}\right)$$
$$\|C^{\frac{1}{2}}(f_{\lambda,n}^m - f_0)\|_{\mathcal{H}} = \mathcal{O}_{p_0}\left(n^{-b\theta-\frac{1}{2}\theta} + n^{\frac{1}{2}\theta-\frac{1}{2}}\right).$$

Both bounds are minimized when $\theta = \frac{1}{2(1+b)}$, which since $0 \le b \le \frac{1}{2}$ leads to $\frac{1}{2} \ge \theta \ge \frac{1}{3}$, and the requirement on n is always satisfied once n is large enough. This shows, as claimed, that

$$\|f_{\lambda,n}^m - f_0\|_{\mathcal{H}} = \mathcal{O}_{p_0}\left(n^{-\frac{b}{2(b+1)}}\right) \qquad J(p_0\|p_{f_{\lambda,n}^m}) = \mathcal{O}_{p_0}\left(n^{-\frac{2b+1}{2(b+1)}}\right)$$

when $m = \Omega\left(n^{\frac{1}{2(1+b)}}\log n\right)$.

The bounds on L^r, Hellinger, and KL convergence follow from Lemma 20 under Assumption (H). □

D Auxiliary results

D.1 Standard concentration inequalities in Hilbert spaces

Lemma 10 (Hoeffding-type inequality for random vectors)**.** *Let $X_1, \dots, X_n$ be iid random variables in a (separable) Hilbert space, where $\mathbb{E}\,X_i = 0$ and $\|X_i\| \le L$ almost surely. Then for any $\varepsilon > L/\sqrt{n}$,*

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| > \varepsilon\right) \le \exp\left(-\frac{1}{2}\left(\frac{\sqrt{n}\varepsilon}{L} - 1\right)^2\right);$$

equivalently, we have with probability at least $1-\delta$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| \le \frac{L}{\sqrt{n}}\left(1 + \sqrt{2\log\tfrac{1}{\delta}}\right).$$

Proof. Following Example 6.3 of Boucheron et al. (2013), we can apply McDiarmid's inequality. The function $f(X_1,\dots,X_n) = \left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|$ satisfies bounded differences:

$$\left|\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| - \left\|\frac{1}{n}\hat{X}_1 + \frac{1}{n}\sum_{i=2}^n X_i\right\|\right| \le \left\|\frac{1}{n}(X_1 - \hat{X}_1)\right\| \le \frac{2L}{n}.$$

Thus for $\varepsilon \geq \mathbb{E}\left\|\frac{1}{n}\sum_i X_i\right\|$,

$$\Pr\left(\left\|\frac{1}{n}\sum_i X_i\right\| > \varepsilon\right) \leq \exp\left(-\frac{n\left(\varepsilon - \mathbb{E}\left\|\frac{1}{n}\sum_i X_i\right\|\right)^2}{2L^2}\right).$$

We also know that

$$\mathbb{E}\left\|\frac{1}{n}\sum_i X_i\right\| \leq \frac{1}{n}\sqrt{\mathbb{E}\left\|\sum_i X_i\right\|^2} = \frac{1}{n}\sqrt{\sum_{i,j}\mathbb{E}\langle X_i, X_j\rangle} = \frac{1}{n}\sqrt{\sum_i \mathbb{E}\|X_i\|^2} \leq \frac{1}{n}\sqrt{nL^2} = \frac{L}{\sqrt{n}},$$

so

$$\Pr\left(\left\|\frac{1}{n}\sum_i X_i\right\| > \varepsilon\right) \leq \exp\left(-\frac{n\left(\varepsilon - \frac{L}{\sqrt{n}}\right)^2}{2L^2}\right) = \exp\left(-\frac{1}{2}\left(\frac{\sqrt{n}\varepsilon}{L} - 1\right)^2\right)$$

as desired. The second statement follows by simple algebra. □

Lemma 11 (Hoeffding-type inequality for random Hilbert-Schmidt operators)**.** *Let $X_1, \ldots, X_n$ be iid random operators in a (separable) Hilbert space, where $\mathbb{E}\, X_i = 0$ and $\|X_i\| \leq L$, $\|X_i\|_{\mathrm{HS}} \leq B$ almost surely. Then for any $\varepsilon > B/\sqrt{n}$,*

$$\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| < \varepsilon\right) \leq \exp\left(-\frac{1}{2}\left(\frac{\sqrt{n}\varepsilon}{L} - \frac{B}{L}\right)^2\right);$$

equivalently, we have with probability at least $1-\delta$ that

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\| \leq \frac{1}{\sqrt{n}}\left(B + L\sqrt{2\log\tfrac{1}{\delta}}\right).$$

Proof. The argument is the same as Lemma 10, except that

$$\mathbb{E}\left\|\frac{1}{n}\sum_i X_i\right\| \leq \frac{1}{n}\sqrt{\mathbb{E}\left\|\sum_i X_i\right\|_{\mathrm{HS}}^2} = \frac{1}{n}\sqrt{\sum_{i,j}\mathbb{E}\langle X_i, X_j\rangle_{\mathrm{HS}}} = \frac{1}{n}\sqrt{\sum_i \mathbb{E}\|X_i\|_{\mathrm{HS}}^2} \leq \frac{B}{\sqrt{n}}$$

using $\|X_i\| \leq \|X_i\|_{\mathrm{HS}}$. □

Lemma 12 (Bernstein's inequality for a sum of random operators; Proposition 12 of Rudi et al. (2015))**.** *Let $\mathcal{H}$ be a separable Hilbert space, and $X_1, \ldots, X_n$ a sequence of iid self-adjoint positive random operators on $\mathcal{H}$, with $\mathbb{E}\, X_1 = 0$, $\lambda_{\max}(X_1) \leq L$ almost surely for some $L > 0$. Let S be a positive operator such that $\mathbb{E}[X_1^2] \preceq S$. Let $\beta = \log\frac{2\operatorname{Tr} S}{\|S\|\delta}$. Then for any $\delta \geq 0$, with probability at least $1-\delta$*

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq \frac{2L\beta}{3n} + \sqrt{\frac{2\|S\|\beta}{n}}.$$

D.2 Concentration of sum of correlated operators

The following result is similar to Proposition 8 of Rudi et al. (2015), but the proof is considerably more complex due to the sum over correlated operators.

We also allow for a random "masking" operation via the U_i^a. Lemma 13 applies to general sampling schemes U_i^a; Lemmas 15 to 17 specialize it to particular sampling schemes.

Lemma 13. *Let $W_a = (Y_i^a)_{i\in[d]}$ be a random d-tuple of vectors in a separable Hilbert space $\mathcal{H}$, with $\{W_a\}_{a\in[n]}$ iid.*

Let $U^a = (U_i^a)_{i\in d}$ be a corresponding d-tuple of random vectors, with $\Pr(U_i^a \in \{0,1\}) = 1$, such that the $\{U^a\}_{a\in[n]}$ are iid, $\mathbb{E}[U_i^a] := \mu_i \in (0,1]$, and U^a is independent of W^a. Define $\nu_{ij} := \mathbb{E}[U_i^a U_j^a]/(\mu_i\mu_j)$, $\nu_i = \sum_{j=1}^d \nu_{ij}$.

Suppose that $Q = \mathbb{E}\sum_{i=1}^d Y_i^1 \otimes Y_i^1$ *exists and is trace class, and that for any* $\lambda > 0$ *there is* $\mathcal{N}'_\infty(\lambda) < \infty$ *such that* $\langle Y_l^a, (Q+\lambda I)^{-1} Y_l^a\rangle_{\mathcal{H}} \le \mathcal{N}'_\infty(\lambda)$ *almost surely. Let* $Q_\lambda = Q + \lambda I$, $V_a = \sum_{i=1}^d \frac{1}{\mu_i} U_i^a (Y_i^a \otimes Y_i^a)$.

Let

$$S := \mathcal{N}'_\infty(\lambda) Q_\lambda^{-\frac{1}{2}} \left(2\,\mathbb{E}\left[\sum_{i,j}^d \nu_{ij}(Y_i \otimes Y_j)\right] + 3\,\mathbb{E}\left[\sum_{i=1}^d \nu_i (Y_i \otimes Y_i)\right]\right) Q_\lambda^{-\frac{1}{2}},$$

and suppose that $\operatorname{Tr} S \le t$, $s_* \le \|S\| \le s^*$. *(These bounds will depend on the distribution of* U^a*.)*

Then with probability at least $1-\delta$ *we have that*

$$\lambda_{\max}\left(Q_\lambda^{-\frac{1}{2}} \left(Q - \frac{1}{n}\sum_{a=1}^n V_a \right) Q_\lambda^{-\frac{1}{2}} \right) \le \frac{2\beta}{3n} + \sqrt{\frac{2s^*\beta}{n}}, \qquad \beta = \log\left(\frac{2t}{\delta s_*}\right).$$

Proof. We will apply the Bernstein inequality for random operators, Lemma 12, to $Z_a := Q_\lambda^{-\frac{1}{2}}(Q - V_a)Q_\lambda^{-\frac{1}{2}}$. For each a,

$$\mathbb{E}\, V_a = \sum_{i=1}^d \frac{\mathbb{E}\, U_i^a}{\mu_i}\, \mathbb{E}[Y_i^a \otimes Y_i^a] = Q$$

so that $\mathbb{E}\, Z_a = 0$, and since V_a is positive and Q_λ is self-adjoint,

$$\sup_{\|f\|_{\mathcal{H}}=1} \langle f, Z_a f\rangle_{\mathcal{H}} = \sup_{\|f\|_{\mathcal{H}}=1} \langle f, Q_\lambda^{-1} Q f\rangle_{\mathcal{H}} - \langle f, Q_\lambda^{-\frac{1}{2}} V_a Q_\lambda^{-\frac{1}{2}} f\rangle_{\mathcal{H}} \le \sup_{\|f\|_{\mathcal{H}}=1} \langle f, Q_\lambda^{-1} Q f\rangle_{\mathcal{H}} \le 1.$$

To apply Lemma 12, we now need to show that the positive operator S upper bounds the second moment of Z_a. Letting $u \in \mathcal{H}$, and dropping the subscript a for brevity, we have that

$$\begin{aligned}
\langle u, \mathbb{E}[Z^2] u\rangle_{\mathcal{H}} &= \left\langle u, \mathbb{E}[Q_\lambda^{-\frac{1}{2}} V Q_\lambda^{-1} V Q_\lambda^{-\frac{1}{2}}] u \right\rangle_{\mathcal{H}} - \left\langle u, Q_\lambda^{-\frac{1}{2}} Q Q_\lambda^{-1} Q Q_\lambda^{-\frac{1}{2}} u \right\rangle_{\mathcal{H}} \\
&\le \left\langle u, Q_\lambda^{-\frac{1}{2}}\, \mathbb{E}[V Q_\lambda^{-1} V] Q_\lambda^{-\frac{1}{2}} u \right\rangle_{\mathcal{H}} \\
&= \left\langle Q_\lambda^{-\frac{1}{2}} u, \mathbb{E}[V Q_\lambda^{-1} V] Q_\lambda^{-\frac{1}{2}} u \right\rangle_{\mathcal{H}} \\
&= \sum_{i,j}^d \left\langle Q_\lambda^{-\frac{1}{2}} u, \mathbb{E}\left[\frac{U_i}{\mu_i}(Y_i \otimes Y_i) Q_\lambda^{-1} (Y_j \otimes Y_j)\frac{U_j}{\mu_j}\right] Q_\lambda^{-\frac{1}{2}} u \right\rangle_{\mathcal{H}} \\
&= \sum_{i,j}^d \frac{\mathbb{E}[U_i U_j]}{\mu_i \mu_j}\, \mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_i\rangle_{\mathcal{H}} \langle Q_\lambda^{-\frac{1}{2}} u, Y_j\rangle_{\mathcal{H}} \langle Y_i, Q_\lambda^{-1} Y_j\rangle_{\mathcal{H}}\right].
\end{aligned}$$

Let $\nu_{ij} = \mathbb{E}[U_i U_j]/(\mu_i \mu_j)$. Using $2\langle x, Ay\rangle = \langle x+y, A(x+y)\rangle - \langle x, Ax\rangle - \langle y, Ay\rangle$, we get:

$$\begin{aligned}
\langle u, \mathbb{E}[Z^2] u\rangle_{\mathcal{H}} \le{}& \frac{1}{2}\sum_{i,j}^d \nu_{ij}\, \mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_i\rangle_{\mathcal{H}} \langle Q_\lambda^{-\frac{1}{2}} u, Y_j\rangle_{\mathcal{H}} \langle Y_i + Y_j, Q_\lambda^{-1}(Y_i + Y_j)\rangle_{\mathcal{H}}\right] \\
&- \sum_{i,j}^d \nu_{ij}\, \mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_i\rangle_{\mathcal{H}} \langle Q_\lambda^{-\frac{1}{2}} u, Y_j\rangle_{\mathcal{H}} \langle Y_i, Q_\lambda^{-1} Y_i\rangle\right].
\end{aligned}$$

Similarly using $2\langle A, x\rangle\langle A, y\rangle = \langle A, x+y\rangle^2 - \langle A, x\rangle^2 - \langle A, y\rangle^2$, we get that the first line is

$$\begin{aligned}
\frac{1}{4}\sum_{i,j}^d \nu_{ij} \Big(&\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_i + Y_j\rangle_{\mathcal{H}}^2 \langle Y_i + Y_j, Q_\lambda^{-1}(Y_i + Y_j)\rangle_{\mathcal{H}}\right] \\
&- \mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_i\rangle_{\mathcal{H}}^2 \langle Y_i + Y_j, Q_\lambda^{-1}(Y_i + Y_j)\rangle_{\mathcal{H}}\right] - \mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}} u, Y_j\rangle_{\mathcal{H}}^2 \langle Y_i + Y_j, Q_\lambda^{-1}(Y_i + Y_j)\rangle_{\mathcal{H}}\right]\Big)
\end{aligned}$$

and the second is

$$\frac{1}{2}\sum_{i,j}^{d}\nu_{ij}\left(-\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i+Y_j\rangle_{\mathcal{H}}^2\langle Y_i, Q_\lambda^{-1}Y_i\rangle\right]\right.$$
$$\left.+\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\langle Y_i, Q_\lambda^{-1}Y_i\rangle\right]+\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}^2\langle Y_i, Q_\lambda^{-1}Y_i\rangle\right]\right).$$

Each of these expectations is nonnegative, so dropping the ones with negative coefficients gives:

$$\langle u, \mathbb{E}[Z^2]u\rangle_{\mathcal{H}} \le \frac{1}{4}\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i+Y_j\rangle_{\mathcal{H}}^2\langle Y_i+Y_j, Q_\lambda^{-1}(Y_i+Y_j)\rangle_{\mathcal{H}}\right]$$
$$+\frac{1}{2}\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\langle Y_i, Q_\lambda^{-1}Y_i\rangle\right]+\frac{1}{2}\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}^2\langle Y_i, Q_\lambda^{-1}Y_i\rangle\right].$$

Recalling that $\langle Y_i, Q_\lambda^{-1}Y_i\rangle \le \mathcal{N}'_\infty(\lambda)$, the second line is upper-bounded by $\mathcal{N}'_\infty(\lambda)$ times

$$\frac{1}{2}\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\right]+\frac{1}{2}\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}^2\right]=\sum_{i=1}^{d}\nu_i\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\right],$$

where $\nu_i = \sum_{j=1}^{d}\nu_{ij}$. We also have that

$$\langle Y_i+Y_j, Q_\lambda^{-1}(Y_i+Y_j)\rangle_{\mathcal{H}} = \|Q_\lambda^{-\frac{1}{2}}(Y_i+Y_j)\|_{\mathcal{H}}^2 \le 2(\|Q_\lambda^{-\frac{1}{2}}Y_i\|_{\mathcal{H}}^2+\|Q_\lambda^{-\frac{1}{2}}Y_j\|_{\mathcal{H}}^2) \le 4\mathcal{N}'_\infty(\lambda),$$

so the first sum is at most $\mathcal{N}'_\infty(\lambda)$ times

$$\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i+Y_j\rangle_{\mathcal{H}}^2\right]$$
$$=\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2+\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}^2+2\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}\right]$$
$$=2\sum_{i=1}^{d}\nu_i\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\right]+2\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}\right].$$

Thus

$$\langle u, \mathbb{E}[Z^2]u\rangle_{\mathcal{H}} \le \mathcal{N}'_\infty(\lambda)\left(2\sum_{i,j}^{d}\nu_{ij}\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}\langle Q_\lambda^{-\frac{1}{2}}u, Y_j\rangle_{\mathcal{H}}\right]+3\sum_{i=1}^{d}\nu_i\,\mathbb{E}\left[\langle Q_\lambda^{-\frac{1}{2}}u, Y_i\rangle_{\mathcal{H}}^2\right]\right)$$
$$=\left\langle u, \mathcal{N}'_\infty(\lambda)Q_\lambda^{-\frac{1}{2}}\left(2\,\mathbb{E}\left[\sum_{i,j}^{d}\nu_{ij}(Y_i\otimes Y_j)\right]+3\,\mathbb{E}\left[\sum_{i=1}^{d}\nu_i(Y_i\otimes Y_i)\right]\right)Q_\lambda^{-\frac{1}{2}}u\right\rangle_{\mathcal{H}}$$
$$=\langle u, Su\rangle_{\mathcal{H}},$$

recalling that

$$S=\mathcal{N}'_\infty(\lambda)Q_\lambda^{-\frac{1}{2}}\left(2\,\mathbb{E}\left[\sum_{i,j}^{d}\nu_{ij}(Y_i\otimes Y_j)\right]+3\,\mathbb{E}\left[\sum_{i=1}^{d}\nu_i(Y_i\otimes Y_i)\right]\right)Q_\lambda^{-\frac{1}{2}}.$$

Thus we have the desired upper bound $\mathbb{E}[Z^2] \preceq S$.

Recall that $\operatorname{Tr} S \le t$, $s_* \le \|S\| \le s^*$. Then by Lemma 12, with probability at least $1-\delta$ we have that

$$\lambda_{\max}\left(\frac{1}{n}Z_a\right) \le \frac{2\beta'}{3n}+\sqrt{\frac{2\|S\|\beta'}{n}} \le \frac{2\beta}{3n}+\sqrt{\frac{2s^*\beta}{n}},$$

where

$$\beta' := \log \frac{2 \operatorname{Tr} S}{\delta \|S\|} \le \log \frac{2t}{\delta s_*} =: \beta,$$

as desired. □

We will now find t, s_*, s^* for some particular sampling schemes. The following initial lemma will be useful for this purpose:

Lemma 14. *In the setup of Lemma 13, define* $M := \mathbb{E}\left[\left(\sum_{i=1}^d Y_i\right) \otimes \left(\sum_{i=1}^d Y_i\right)\right]$. *We have:*

$$M \preceq dQ, \quad \operatorname{Tr}\left(Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}}\right) \le \frac{d}{\lambda} \operatorname{Tr}(Q), \quad \left\|Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}}\right\| \le d.$$

Proof. We first show $M \preceq dQ$:

$$\begin{aligned}\langle u, Mu\rangle_{\mathcal{H}} &= \left\langle u, \mathbb{E}\left[\left(\sum_{i=1}^d Y_i\right) \otimes \left(\sum_{i=1}^d Y_i\right)\right] u \right\rangle_{\mathcal{H}} = \mathbb{E}\left[\left\langle u, \sum_{i=1}^d Y_i \right\rangle_{\mathcal{H}}^2\right] \\ &\le \mathbb{E}\left[d \sum_{i=1}^d \langle u, Y_i\rangle_{\mathcal{H}}^2\right] = \mathbb{E}\left[d \sum_{i=1}^d \langle u, (Y_i \otimes Y_i) u\rangle_{\mathcal{H}}\right] = \langle u, dQu\rangle_{\mathcal{H}}.\end{aligned}$$

Thus $\operatorname{Tr}(M) \le d \operatorname{Tr}(Q)$, and since $\|Q_\lambda^{-1}\| \le \frac{1}{\lambda}$ we have

$$\operatorname{Tr}\left(Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}}\right) = \operatorname{Tr}\left(Q_\lambda^{-1} M\right) \le \frac{1}{\lambda} \operatorname{Tr}(M) \le \frac{d}{\lambda} \operatorname{Tr}(Q).$$

For any u with $\|u\|_{\mathcal{H}} = 1$:

$$\langle u, Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} u\rangle_{\mathcal{H}} = \langle Q_\lambda^{-\frac{1}{2}} u, M(Q_\lambda^{-\frac{1}{2}} u)\rangle_{\mathcal{H}} \le \langle Q_\lambda^{-\frac{1}{2}} u, dQ(Q_\lambda^{-\frac{1}{2}} u)\rangle_{\mathcal{H}} = d\langle u, QQ_\lambda^{-1} u\rangle_{\mathcal{H}} \le d,$$

and so the norm inequality follows. □

Lemma 15. *Take the setup of Lemma 13 where each* U_i^a *is independently distributed as* Bernoulli(p), *for* $p \in (0, 1]$. *The number of sampled components is random, distributed as* Binomial(nd, p).

For any $\rho \in (0, \frac{1}{2})$, $\lambda \in (0, \rho\|Q\|]$, *and* $\delta \ge 0$, *it holds with probability at least* $1 - \delta$ *that*

$$\lambda_{\max}\left(Q_\lambda^{-\frac{1}{2}}\left(Q - \frac{1}{n}\sum_{a=1}^n V_a\right) Q_\lambda^{-\frac{1}{2}}\right) \le \frac{2\beta}{3n} + \sqrt{\frac{10\,(d + 1/p - 1)\,\mathcal{N}'_\infty(\lambda)\beta}{n}}$$

where

$$\beta := \log \frac{10\,(d + 1/p - 1) \operatorname{Tr} Q}{\lambda\delta\left(\frac{5/p - 5 + 3d}{1+\rho} - 2d\right)}.$$

Proof. Here we have for $i \ne j$

$$\mu_i = p, \qquad \nu_{ii} = \frac{\mathbb{E}[U_i^2]}{\mu_i^2} = \frac{1}{\mu_i} = \frac{1}{p}, \qquad \nu_{ij} = \frac{\mathbb{E}[U_i U_j]}{\mu_i \mu_j} = \frac{\mathbb{E}\, U_i}{\mu_i} \frac{\mathbb{E}\, U_j}{\mu_j} = 1.$$

Define $r := \frac{1}{p} - 1$; then $\nu_i = r + d$. Using Lemma 14, we get that

$$\mathbb{E}\left[\sum_{i=1}^d \nu_i (Y_i \otimes Y_i)\right] = (r + d)Q$$

and

$$\mathbb{E}\left[\sum_{i,j}^d \nu_{ij}(Y_i \otimes Y_j)\right] = \mathbb{E}\left[\sum_{i,j}^d Y_i \otimes Y_j\right] + \left(\frac{1}{p} - 1\right) \mathbb{E}\left[\sum_{i=1}^d Y_i \otimes Y_i\right] = M + rQ,$$

so that

$$\begin{aligned} S &= \mathcal{N}'_\infty(\lambda) Q_\lambda^{-\frac{1}{2}} \left(2(M + rQ) + 3(r+d)Q\right) Q_\lambda^{-\frac{1}{2}} \\ &= \mathcal{N}'_\infty(\lambda) Q_\lambda^{-\frac{1}{2}} \left(2M + (5r+3d)Q\right) Q_\lambda^{-\frac{1}{2}}. \end{aligned}$$

Thus

$$\begin{aligned} \operatorname{Tr} S &= \mathcal{N}'_\infty(\lambda) \left(2 \operatorname{Tr}(Q_\lambda^{-1} M) + (5r+3d) \operatorname{Tr}(Q_\lambda^{-1} Q)\right) \\ &\le \frac{5(r+d)}{\lambda} \mathcal{N}'_\infty(\lambda) \operatorname{Tr}(Q). \end{aligned}$$

Likewise, since $\left\| Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| \le d$,

$$\|S\| \le \mathcal{N}'_\infty(\lambda) \left(2 \left\| Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| + (3d+5r) \|Q Q_\lambda^{-1}\|\right) \le 5(d+r) \mathcal{N}'_\infty(\lambda).$$

Since we have $\lambda \le \rho \|Q\|$, $\|Q Q_\lambda^{-1}\| = \frac{\|Q\|}{\|Q\| + \lambda} \ge \frac{1}{1+\rho}$ and so

$$\begin{aligned} \|S\| &= \mathcal{N}'_\infty(\lambda) \left\| (5r+3d) Q Q_\lambda^{-1} - 2 Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| \\ &\ge \mathcal{N}'_\infty(\lambda) \left((5r+3d) \|Q Q_\lambda^{-1}\| - 2 \left\| Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| \right) \\ &\ge \mathcal{N}'_\infty(\lambda) \left(\frac{5r+3d}{1+\rho} - 2d \right). \end{aligned}$$

This bound is positive when $\frac{5r+3d}{1+\rho} > 2d$, i.e. $\rho < \frac{1}{2}\left(\frac{5r}{d} + 1\right)$; it suffices that $\rho < \frac{1}{2}$.

Applying Lemma 13 proves the result. □

Lemma 16. *Take the setup of Lemma 13 where each U^a is chosen uniformly from the set of binary vectors with $\|U^a\|_1 = \ell \in [1, d]$, i.e. we choose ℓ components of each vector at random without replacement. Assume that $d > 1$; otherwise, we simply have $\ell = d = 1$, which is covered by Lemma 15 with $p = 1$.*

For any $\rho \in (0, \frac{1}{2})$, $\lambda \in (0, \rho\|Q\|)$, and $\delta \ge 0$, it holds with probability at least $1 - \delta$ that

$$\lambda_{\max} \left(Q_\lambda^{-\frac{1}{2}} \left(Q - \frac{1}{n} \sum_{a=1}^n V_a \right) Q_\lambda^{-\frac{1}{2}} \right) \le \frac{2\beta}{3n} + \sqrt{\frac{10 d \mathcal{N}'_\infty(\lambda) \beta}{n}}$$

where

$$\beta := \log \frac{10 \operatorname{Tr}(Q)}{\lambda \delta \left(\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \frac{1}{1+\rho} - 2\frac{d(\ell-1)}{\ell(d-1)} \right)}.$$

Proof. In this case, for $i \ne j$ we have

$$\mu_i = \frac{\ell}{d}, \qquad \nu_{ii} = \frac{\mathbb{E}[U_i^2]}{\mu_i^2} = \frac{1}{\mu_i} = \frac{d}{\ell}, \qquad \nu_{ij} = \frac{\Pr(U_i = U_j = 1)}{\mu_i \mu_j} = \frac{\binom{d-2}{\ell-2}}{\binom{d}{\ell}} \frac{d^2}{\ell^2} = \frac{d(\ell-1)}{\ell(d-1)}.$$

Thus

$$\nu_i = \frac{d}{\ell} + (d-1)\frac{d(\ell-1)}{\ell(d-1)} = \frac{d}{\ell}\left(1 + (\ell - 1)\right) = d,$$

and $\mathbb{E}\left[\sum_{i=1}^d \nu_i (Y_i \otimes Y_i)\right] = dQ$, while

$$\begin{aligned} \mathbb{E}\left[\sum_{i,j}^d \nu_{ij} (Y_i \otimes Y_j)\right] &= \frac{d(\ell-1)}{\ell(d-1)} \mathbb{E}\left[\sum_{i,j}^d Y_i \otimes Y_j\right] + \left(\frac{d}{\ell} - \frac{d(\ell-1)}{\ell(d-1)}\right) \mathbb{E}\left[\sum_{i=1}^d Y_i \otimes Y_i\right] \\ &= \frac{d(\ell-1)}{\ell(d-1)} M + \frac{d(d-\ell)}{\ell(d-1)} Q \end{aligned}$$

using M from Lemma 14, and so

$$S = \mathcal{N}'_\infty(\lambda) Q_\lambda^{-\frac{1}{2}} \left(2\frac{d(\ell-1)}{\ell(d-1)} M + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) Q \right) Q_\lambda^{-\frac{1}{2}}.$$

Thus

$$\begin{aligned}
\operatorname{Tr} S &= \mathcal{N}'_\infty(\lambda) \left(2\frac{d(\ell-1)}{\ell(d-1)} \operatorname{Tr}(Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}}) + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \operatorname{Tr}(QQ_\lambda^{-1}) \right) \\
&\le \frac{1}{\lambda} \mathcal{N}'_\infty(\lambda) \left(2\frac{d(\ell-1)}{\ell(d-1)} d + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \right) \operatorname{Tr}(Q) \\
&= \frac{1}{\lambda} \mathcal{N}'_\infty(\lambda) d \left(2\frac{d(\ell-1)+d-\ell}{\ell(d-1)} + 3 \right) \operatorname{Tr}(Q) \\
&= \frac{5d}{\lambda} \mathcal{N}'_\infty(\lambda) \operatorname{Tr}(Q).
\end{aligned}$$

We similarly have

$$\begin{aligned}
\|S\| &\le \mathcal{N}'_\infty(\lambda) \left(2\frac{d(\ell-1)}{\ell(d-1)} \left\| Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \left\| QQ_\lambda^{-1} \right\| \right) \\
&\le \mathcal{N}'_\infty(\lambda) \left(2\frac{d(\ell-1)}{\ell(d-1)} d + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \right) \\
&= 5d\mathcal{N}'_\infty(\lambda).
\end{aligned}$$

Note also that $1 \le \ell \le d$ implies $2\frac{d(\ell-1)}{\ell(d-1)} \le 3 + 2\frac{d-\ell}{\ell(d-1)}$ for integral ℓ and d. Since $M \le dQ$, and like in Lemma 15 we have that $\|QQ_\lambda^{-1}\| \ge \frac{1}{1+\rho}$, we obtain that

$$\begin{aligned}
\|S\| &= \mathcal{N}'_\infty(\lambda) \left\| -2\frac{d(\ell-1)}{\ell(d-1)} Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) QQ_\lambda^{-1} \right\| \\
&\ge \mathcal{N}'_\infty(\lambda) \left(-2\frac{d(\ell-1)}{\ell(d-1)} \left\| Q_\lambda^{-\frac{1}{2}} M Q_\lambda^{-\frac{1}{2}} \right\| + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \left\| QQ_\lambda^{-1} \right\| \right) \\
&\ge \mathcal{N}'_\infty(\lambda) \left(-2\frac{d(\ell-1)}{\ell(d-1)} d + d\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \frac{1}{1+\rho} \right) \\
&= d\mathcal{N}'_\infty(\lambda) \left(\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \frac{1}{1+\rho} - 2\frac{d(\ell-1)}{\ell(d-1)} \right).
\end{aligned}$$

We then have that

$$\frac{t}{s_*} = \frac{5 \operatorname{Tr}(Q)/\lambda}{\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \frac{1}{1+\rho} - 2\frac{d(\ell-1)}{\ell(d-1)}},$$

which is well-defined and positive as long as either $\ell = 1$ or $\left(3 + 2\frac{d-\ell}{\ell(d-1)}\right) \frac{1}{1+\rho} > 2\frac{d(\ell-1)}{\ell(d-1)}$, i.e. $\rho < \frac{1}{2}\frac{\ell+4-5\frac{\ell}{d}}{\ell-1}$; since $\frac{\ell}{d} \le 1$, it suffices that $\rho < \frac{1}{2}$. The claim follows from Lemma 13. □

An interesting special case of Lemma 16 is $\ell = 1$, where t/s_* reduces to $\frac{1+\rho}{\lambda} \operatorname{Tr}(Q)$.

Lemma 17. *Take the setup of Lemma 13 where each U_i^a is identically 1: we always sample all components of the considered points.*

For any $\rho \in (0, \frac{1}{2})$, $\lambda \in (0, \rho\|Q\|)$, and $\delta \ge 0$, it holds with probability at least $1 - \delta$ that

$$\lambda_{\max}\left(Q_\lambda^{-\frac{1}{2}} \left(Q - \frac{1}{n}\sum_{a=1}^{n} V_a \right) Q_\lambda^{-\frac{1}{2}} \right) \le \frac{2\beta}{3n} + \sqrt{\frac{10d\mathcal{N}'_\infty(\lambda)\beta}{n}}, \qquad \beta := \log \frac{10 \operatorname{Tr} Q}{\lambda\delta\left(\frac{3}{1+\rho} - 2\right)}.$$

Proof. Special case of either Lemma 15 with $p = 1$ or Lemma 16 with $\ell = d$. □

D.3 Results on Hilbert space operators

Lemmas 18 and 19 were proven and used by Rudi et al. (2015).

Lemma 18 (Proposition 3 of Rudi et al. (2015)). *Let $\mathcal{H}_1$, $\mathcal{H}_2$, $\mathcal{H}_3$ be three separable Hilbert spaces, with $Z : \mathcal{H}_1 \to \mathcal{H}_2$ a bounded linear operator and P a projection operator on $\mathcal{H}_1$ with $\operatorname{range} P = \overline{\operatorname{range} Z^*}$. Then for any bounded linear operator $F : \mathcal{H}_3 \to \mathcal{H}_1$ and any $\lambda > 0$,*

$$\|(I - P)F\| \leq \sqrt{\lambda}\|(Z^*Z + \lambda I)^{-\frac{1}{2}}F\|.$$

Lemma 19 (Proposition 7 of Rudi et al. (2015)). *Let $\mathcal{H}$ be a separable Hilbert space, with A, B bounded self-adjoint positive linear operators on $\mathcal{H}$ and $A_\lambda = A + \lambda I$, $B_\lambda = B + \lambda I$. Then for any $\lambda > 0$,*

$$\|A_\lambda^{-\frac{1}{2}} B^{\frac{1}{2}}\| \leq \|A_\lambda^{-\frac{1}{2}} B_\lambda^{\frac{1}{2}}\| \leq (1 - \gamma(\lambda))^{-\frac{1}{2}}$$

when

$$\gamma(\lambda) := \lambda_{\max}\left(B_\lambda^{-\frac{1}{2}}(B - A)B_\lambda^{-\frac{1}{2}}\right) < 1.$$

D.4 Distances between distributions in $\mathcal{P}$

Lemma 20 (Distribution distances from parameter distances). *Let $f_0, f \in \mathcal{F}$ correspond to distributions $p_0 = p_{f_0}, p = p_f \in \mathcal{P}$. Under Assumption **(H)**, we have that for all $r \in [1, \infty]$:*

$$\begin{aligned}
\|p - p_0\|_{L^r(\Omega)} &\leq 2\kappa e^{2\kappa\|f - f_0\|_{\mathcal{H}}} e^{2\kappa \min(\|f\|_{\mathcal{H}}, \|f_0\|_{\mathcal{H}})} \|f - f_0\|_{\mathcal{H}} \|q_0\|_{L^r(\Omega)} \\
\|p - p_0\|_{L^1(\Omega)} &\leq 2\kappa e^{2\kappa\|f - f_0\|_{\mathcal{H}}} \|f - f_0\|_{\mathcal{H}} \\
\mathrm{KL}(f\|f_0) &\leq c\kappa^2 \|f - f_0\|_{\mathcal{H}}^2 e^{\kappa\|f - f_0\|_{\mathcal{H}}}(1 + \kappa\|f - f_0\|_{\mathcal{H}}) \\
\mathrm{KL}(f_0\|f) &\leq c\kappa^2 \|f - f_0\|_{\mathcal{H}}^2 e^{\kappa\|f - f_0\|_{\mathcal{H}}}(1 + \kappa\|f - f_0\|_{\mathcal{H}}) \\
h(f, f_0) &\leq \kappa e^{\frac{1}{2}\|f - f_0\|_{\mathcal{H}}} \|f - f_0\|_{\mathcal{H}}
\end{aligned}$$

where c is a universal constant and h denotes the Hellinger distance $h(p, q) = \|\sqrt{p} - \sqrt{q}\|_{L^2(\Omega)}$.

Proof. First note that

$$\|f - f_0\|_\infty = \sup_{x \in \Omega} |f(x) - f_0(x)| = \sup_{x \in \Omega} |\langle f - f_0, k(x, \cdot)\rangle_{\mathcal{H}}| \leq \kappa\|f - f_0\|_{\mathcal{H}}.$$

Then, since each $f \in \mathcal{H}$ is bounded and measurable, $\mathcal{P}_\infty$ of Lemma A.1 of Sriperumbudur et al. (2017) is simply $\mathcal{P}$, and the result applies directly. □

References

Ben-Israel, A. and T. N. E. Greville (2003). *Generalized inverses: theory and applications*. Second edition. Springer.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, UK: Oxford University Press.

Caponnetto, A. and E. De Vito (2007). "Optimal rates for regularized least-squares algorithm." In: *Foundations of Computational Mathematics* 7.3, pp. 331–368.

Rudi, A., R. Camoriano, and L. Rosasco (2015). "Less is More: Nyström Computational Regularization." In: *NIPS*. arXiv: 1507.04717.

Sriperumbudur, B. K., K. Fukumizu, R. Kumar, A. Gretton, A. Hyvärinen, and R. Kumar (2017). "Density Estimation in Infinite Dimensional Exponential Families." In: *Journal of Machine Learning Research* 18.57, pp. 1–59. arXiv: 1312.3516.

Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. Springer.

Strathmann, H., D. Sejdinovic, S. Livingstone, Z. Szábo, and A. Gretton (2015). "Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families." In: *NIPS*. arXiv: 1506.02564.