Efficient and principled score estimation

Dougal J. Sutherland* **Heiko Strathmann*** **Michael Arbel** **Arthur Gretton**
Gatsby Computational Neuroscience Unit, University College London
{dougal, heiko.strathmann, michael.n.arbel, arthur.gretton}@gmail.com

Abstract

We propose a fast method with statistical guarantees for learning an exponential family density model where the natural parameter is in a reproducing kernel Hilbert space, and may be infinite dimensional. The model is learned by fitting the derivative of the log density, the *score*, thus avoiding the need to compute a normalization constant. We improved the computational efficiency of an earlier solution with a low-rank, Nyström-like solution. The new solution retains the consistency and convergence rates of the full-rank solution (exactly in Fisher distance, and nearly in other distances), with guarantees on the degree of cost and storage reduction. We evaluate the method in experiments on density estimation and in the construction of an adaptive Hamiltonian Monte Carlo sampler. Compared to an existing score learning approach using a denoising autoencoder, our estimator is empirically more data-efficient when estimating the score, runs faster, and has fewer parameters (which can be tuned in a principled and interpretable way), in addition to providing statistical guarantees.

1 Introduction

We address the problem of efficiently estimating the natural parameter of a density in the exponential family, where this parameter may be infinite dimensional (a member of a function space). While finite dimensional exponential families are a keystone of parametric statistics [4], their generalization to the fully non-parametric setting has proved challenging, despite the benefits and applications envisaged for such models [5]: it is difficult to construct a practical, consistent maximum likelihood solution for infinite dimensional natural parameters [2, 11, 12]. In the absence of a tractable estimation procedure, the infinite exponential family has not seen the widespread adoption and practical successes of other nonparametric generalizations of parametric models, for instance the Gaussian and Dirichlet processes.

Recently, a procedure has been developed to fit infinite exponential family models to sample points drawn i.i.d. from a probability density, where the natural parameter is a member of a reproducing kernel Hilbert space (RKHS) [24]. The approach employs a score matching procedure [13], which minimizes the *Fisher distance*: the expected squared distance between the model *score* (i.e., the derivative of the log model density) and the score of the (unknown) true density. The Fisher distance can be recast to yield a quadratic loss using integration by parts. Unlike the maximum likelihood case, a Tikhonov-regularized solution can be formulated to obtain a well-posed and straightforward solution, which is a linear system defined in terms of the first and second derivatives of the RKHS kernels at the sample points. Details of the model and its empirical fit are given in Section 2. Consistency (in Fisher distance, as well as in L^r, Hellinger, and KL distances) has been established, with rates depending on the smoothness of the density [24].

The infinite dimensional exponential family has been applied in adaptive Hamiltonian Markov chain Monte Carlo [26]. In this setting, the score of the stationary distribution of the Markov chain is

*These authors contributed equally.

learned from the chain history, and used in formulating new, more efficient proposals for a Metropolis-Hastings algorithm. For a practical implementation, it was necessary to approximate the full solution from [24], which has memory cost $\mathcal{O}(n^2d^2)$ and computational cost $\mathcal{O}(n^3d^3)$, where n is the number of training samples and d is the dimension of the problem. Two heuristics were employed: one using random Fourier features [18, 25, 27], and the second using a finite, random set of basis points. While these heuristics greatly improved the runtime, no convergence guarantees are known, nor any means of determining how quickly to increase the complexity of these solutions with increasing n.

In the present work, we present an efficient learning scheme for the infinite dimensional exponential family, using a Nyström approximation to the solution. Our main theoretical contribution, in Theorem 2, is to prove guarantees on the convergence of this algorithm for an increasing number m of Nyström points and n training samples. Depending on the problem difficulty, convergence is attained in the regime $m \sim n^{1/3}$ to $m \sim n^{1/2}$ (up to log terms), thus yielding guaranteed cost savings. The overall Fisher distance between our solution and the true density decreases as $m, n \to \infty$ with rates that match those of the full solution in [24, Theorem 6]; convergence in other distances (e.g., KL and Hellinger) either matches or is slightly worse, depending on the problem smoothness. These tight generalization bounds draw on recent state-of-the-art techniques developed for least-squares regression in [20], which efficiently and directly control the generalization error as a function of the Nyström basis, rather than relying on indirect proofs via the reconstruction error of the Gram matrix, as in e.g. [7]. Details may be found in Sections 3 and 4.

In our experiments (Section 5), we compare our approach against the full solution of [24], the heuristics in [5], and the autoencoder score estimator of [1] (discussed in Section 2.3). We address two problem settings: score function estimation for known, multimodal densities in high dimensions, and adaptive Hamiltonian Monte Carlo, where the score is used in proposing Metropolis-Hastings moves, which will be accepted more often as the quality of the learned score improves. Our approach is significantly more accurate than the autoencoder score estimate, as well as being faster, and easier to tune. Moreover, our method performs as well as the full kernel exponential family solution at a much lower computational cost, and on par with previous heuristic approximations.

2 Unnormalized density and score estimation

Suppose we are given a set of points $X = \{X_b\}_{b\in[n]} \subset \mathbb{R}^d$ sampled i.i.d. from an unknown distribution with density p_0. Our setting is that of *unnormalized density estimation*, where we wish to fit a model p such that $p(x)/Z(p) \approx p_0(x)$ in some divergence measure, and we do not concern ourselves with the normalization factor $Z(p)$ (also called the partition function). In many powerful classes of probabilistic models, computing $Z(p)$ is intractable, but several interesting applications do not require it, including mode finding and sampling via Markov Chain Monte Carlo (MCMC). This setting is closely related to that of *energy-based learning* [14].

The exponential family models with infinite dimensional natural parameters are a particular case for which the partition function is problematic. Here fitting by maximum likelihood is difficult, and becomes completely impractical in high dimensions [2, 11, 12].

Hyvärinen [13] proposed an elegant approach to estimate an *unnormalized* density, by minimizing the Fisher divergence, the expected squared distance between *score functions*[1] $\nabla_x \log p(x)$:

$$J(p_0\|p) = \frac{1}{2}\int p_0(x)\,\|\nabla_x \log p(x) - \nabla_x \log p_0(x)\|_2^2\,\mathrm{d}x \tag{1}$$

$$= \int p_0(x) \sum_{i=1}^{d} \left[\partial_i^2 \log p(x) + \frac{1}{2}\left(\partial_i \log p(x)\right)^2\right] \mathrm{d}x + \text{const}, \tag{2}$$

where (2) assumes some mild regularity conditions and contains a constant depending only on p_0. We use $\partial_i f(x)$ to mean $\frac{\partial}{\partial x_i} f(x)$. Crucially, (2) is independent of $Z(p)$ and, other than the constant, depends on p_0 only through an expectation, so it can be estimated by a simple Monte Carlo average.

The score function is in itself a quantity of interest, and is employed directly in several algorithms. Perhaps best known is Hamiltonian Monte Carlo (HMC, e.g. [16]), where the score is used in

[1] Here we use the term *score* in the sense of Hyvärinen [13]; in traditional statistical parlance, this is the score with respect to a hypothetical location parameter of the model.

constructing Hamiltonian dynamics that yield fast mixing chains. Thus, if the score can be learned from the chain history, it can be used in constructing an approximate HMC sampler with mixing properties close to those attainable using the population score [26]. Another application area is in constructing control functionals for Monte Carlo integration [17]: again, learned score functions could be used where closed-form expressions do not exist.

Computing unnormalized densities from a nonparametric learned score function can be a more challenging task. A direct approach would involve numerical integration of the score estimate, where errors can accumulate; moreover, as discussed by [1, Section 3.6], a given score estimate might not correspond to a valid gradient function, or might not yield a normalizable density. The exponential family model does not suffer these drawbacks, as we will see next.

2.1 Kernel parametrization of exponential families

We now describe the procedure for unnormalized density estimation in the *kernel exponential family* $\mathcal{P}$ [5, 11]. This is an infinite dimensional exponential family model, defined as

$$\mathcal{P} = \{p_f(x) := \exp\left(f(x) - A(f)\right) q_0(x) \mid f \in \mathcal{F}\},$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space [3], $\mathcal{F} \subseteq \mathcal{H}$ is the set of functions for which the normalizer $Z(p_f) = \int \exp(f(x))\, q_0(x)\mathrm{d}x$ is finite, $A(f) = \log Z(p_f)$, and q_0 is a base measure with appropriately vanishing tails. That this is a member of the exponential family becomes apparent when we recall the reproducing property $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}}$: the feature map $x \mapsto k(x,\cdot)$ is the sufficient statistic, and f is the natural parameter. An overview of various finite dimensional members of the exponential family (such as Gamma, Poisson, Binomial, etc) and their corresponding kernel functions may be found in [24, Example 1]. When $\mathcal{H}$ is infinite dimensional, $\mathcal{P}$ can be very rich: for instance, when the kernel on $\mathbb{R}^d$ is a continuous function vanishing at infinity and integrally strictly positive definite, then $\mathcal{P}$ is dense in the family of continuous densities vanishing at infinity for which $\|p/q_0\|_\infty$ is bounded, with respect to the KL, TV, and Hellinger divergences [24, Corollary 2].

As discussed earlier, maximum likelihood estimation is difficult due to the intractability of $A(f)$. Instead, Sriperumbudur et al. [24] propose to use a score-matching approach to find an f such that p_f approximates p_0. Their empirical estimator of (2) is

$$\hat{J}(f) = \hat{J}(p_0\|p_f) - \text{const} = \frac{1}{n}\sum_{b=1}^{n}\sum_{i=1}^{d} \partial_i^2 f(X_b) + \frac{1}{2}\left(\partial_i f(X_b)\right)^2, \tag{3}$$

where the constant depends on p_0 and q_0 but not f. Minimizing a regularized version of (3) gives

$$f_{\lambda,n} = \operatorname*{argmin}_{f\in\mathcal{H}} \hat{J}(f) + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2 = -\frac{\hat{\xi}}{\lambda} + \sum_{a=1}^{n}\sum_{i=1}^{d} \beta_{(a,i)}\partial_i k(X_a,\cdot), \tag{4}$$

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

where $\beta_{(a,i)}$ denotes the $(a-1)d+i$th entry of a vector $\beta \in \mathbb{R}^{nd}$; we use $\partial_i k(x,y)$ to mean $\frac{\partial}{\partial x_i}k(x,y)$, and $\partial_{i+d}k(x,y)$ for $\frac{\partial}{\partial y_i}k(x,y)$. To evaluate the estimated unnormalized log-density $f_{\lambda,n}$ at a point x, we take a linear combination of $\partial_i k(X_a,x)$ and $\partial_i^2 k(X_a,x)$ for each sample X_a. The weights β of this linear combination are obtained by solving the nd-dimensional linear system

$$(G + n\lambda I)\beta = h/\lambda, \tag{5}$$

where $G \in \mathbb{R}^{nd\times nd}$ is the matrix collecting partial derivatives of the kernel at the training points, $G_{(a,i),(b,j)} = \partial_i\partial_{j+d}k(X_a,X_b)$, and $h \in \mathbb{R}^{nd}$ evaluates derivatives of $\hat{\xi}$, $h_{(b,i)} = \partial_i\hat{\xi}(X_b)$.

Solving (5) takes $\mathcal{O}(n^3d^3)$ time and $\mathcal{O}(n^2d^2)$ memory, which quickly becomes infeasible as n grows, especially for large d. We will propose a more scalable approximation in Section 3.

2.2 Related work on fast approximate kernel regression

The system of (5) is related to the problem of kernel ridge regression, which suffers from similar $\mathcal{O}(n^3)$ computational cost. Thus we will briefly review methods for speeding up kernel regression.

Nyström methods We refer here to a class of broadly related Nyström-type methods [20, 23, 28]. The representer theorem [21] guarantees that the minimizer of the empirical regression loss for a training set $X = \{X_b\}_{b\in[n]}$ over the RKHS $\mathcal{H}$ with kernel k will lie in the subspace $\mathcal{H}_X = \operatorname{span}\{k(X_b, \cdot)\}_{b\in[n]}$. Nyström methods find an approximate solution by optimizing over a smaller subspace $\mathcal{H}_Y$, usually given by $\mathcal{H}_Y = \operatorname{span}\{k(y, \cdot)\}_{y\in Y}$ for a set of m points $Y \subseteq X$ chosen uniformly at random. This decreases the computational burden both of training ($\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$ time, $\mathcal{O}(n^2)$ to $\mathcal{O}(nm)$ memory) and testing ($\mathcal{O}(n)$ to $\mathcal{O}(m)$ time and memory).

Guarantees on the performance of Nyström methods have been the topic of considerable study. Earlier approaches have worked by first bounding the error in a Nyström approximation of the kernel matrix on the sample [8], and then separately evaluating the impact of regression with an approximate kernel matrix [7]. This approach, however, results in suboptimal rates; better rates can be obtained by considering the whole problem at once [9], including its direct impact on generalization error [20].

Random feature approximations Another popular method for scaling up kernel methods is to use random Fourier features [18, 25, 27] and their variants. Rather than finding the best solution in a subspace of $\mathcal{H}$, these methods choose a set of parametric features, often independent of the data, such that expected inner products between the features coincide with the kernel. These methods have some attractive computational properties but generally also require the number of features to increase with the data size in a way that can be difficult to analyze: see [19] for such an analysis in regression.

Sketching Another scheme for improving the speed of kernel ridge regression, sketching [29, 30] compresses the kernel matrix and the labels by multiplying with a sketching matrix. These methods have some overlap with Nyström-type approaches, and our method will encompass certain classes of sketches [20, Appendix C.1].

2.3 Prior methods for direct score estimation

Alain and Bengio [1] proposed a deep learning-based approach to directly learn a score function from samples. Denoising autoencoders are networks trained to recover the original inputs from versions with noise added. A denoising autoencoder trained with L_2 loss and noise $\mathcal{N}(0, \sigma^2 I)$ can be used to construct a score estimator: $(r_\sigma(x) - x)/\sigma^2 \approx \nabla_x \log p_0(x)$, where r_σ is the autoencoder's reconstruction function. When the autoencoder has infinite capacity and is trained to its global optimum, Alain and Bengio [1] show that this estimator is consistent as $\sigma \to 0$. For realistic autoencoders with finite representation capacity, however, the consistency of this approach remains an open question. Moreover, this technique has many hyperparameters to choose, both in the architecture of the network and in how it is trained, with no theory yet available to guide those choices.

3 Nyström methods for estimation in kernel exponential families

To alleviate the computational costs of the linear system in (5), we apply the Nyström idea to the estimator of the full kernel exponential family model in (4). More precisely, we select a set of m "basis" points $Y = \{Y_a\}_{a\in[m]}$, and restrict the optimization in (4) to

$$\mathcal{H}_Y := \operatorname{span}\left\{\partial_i k(Y_a, \cdot)\right\}_{\substack{a\in[m]}}^{i\in[d]}, \tag{6}$$

which is a subspace of $\mathcal{H}$ with elements that can be represented using md coefficients, similar to (4). Typically $Y \subset X$; in particular, Y is usually chosen as a uniformly random subset of X. We could, however, use any set of points Y, or even a different set of spanning vectors than $\partial_i k(Y_a, \cdot)$.

Theorem 1. *The regularized minimizer of the empirical Fisher divergence (3) over $\mathcal{H}_Y$ (6) is*

$$f^m_{\lambda,n} = \operatorname*{argmin}_{f\in\mathcal{H}_Y} \hat{J}(f) + \frac{1}{2}\lambda\|f\|^2_{\mathcal{H}} = \sum_{a=1}^{m}\sum_{i=1}^{d}(\beta_Y)_{(a,i)}\partial_i k(Y_b, \cdot),$$
$$\beta_Y = -(\tfrac{1}{n}B_{XY}^{\mathsf{T}}B_{XY} + \lambda G_{YY})^{\dagger}h_Y. \tag{7}$$

Here † denotes the pseudo-inverse, and $B_{XY} \in \mathbb{R}^{nd\times md}$, $G_{YY} \in \mathbb{R}^{md\times md}$, $h_Y \in \mathbb{R}^{md}$ are given by

$$(B_{XY})_{(b,i),(a,j)} = \partial_i\partial_{j+d}k(X_b, Y_a) \qquad (G_{YY})_{(a,i),(a',j)} = \partial_i\partial_{j+d}k(Y_a, Y_{a'})$$

$$(h_Y)_{(a,i)} = \frac{1}{n}\sum_{b=1}^{n}\sum_{j=1}^{d}\partial_i\partial^2_{j+d}k(Y_a, X_b) + \partial_i\partial_{j+d}k(Y_a, X_b)\partial_j \log q_0(X_b).$$

The proof, which is similar to the kernel ridge regression analogue [20], is given in Appendix B. In fact, we show a slight generalization (Lemma 4), which also applies to more general subspaces $\mathcal{H}_Y$.

It is worth emphasizing that in order to evaluate an estimate $f^m_{\lambda,n}$, we need only evaluate derivatives of the kernel between the basis points Y and the test point x. We no longer need X at all: its full contribution is summarized in β_Y.

When $Y \subseteq X$, the above quantities are simply block sub-sampled versions of the terms in the full solution (5). Note, however, that when $Y = X$ we do not exactly recover the solution (5), because $\hat{\xi}$ contains components of the form $\partial_i^2 k(X_b, \cdot) \notin \mathcal{H}_X$.

Computing the the $md \times md$ matrix in (7) takes $\mathcal{O}(nmd^2)$ memory and $\mathcal{O}(nm^2d^3)$ time, both linear in n. Computing the pseudo-inverse takes $\mathcal{O}(m^3d^3)$ computation, independent of n. Moreover, evaluating $f^m_{\lambda,n}$ takes $\mathcal{O}(md)$ time, as opposed to the $\mathcal{O}(nd)$ time for $f_{\lambda,n}$.

Finite and lite kernel exponential families Strathmann et al. [26] proposed two alternative heuristic approximations to the full model of Section 2, used for efficient score learning in adaptive HMC. These approaches currently lack convergence guarantees.

The *finite* form uses an m-dimensional $\mathcal{H}$, defined e.g. by random Fourier features [18], where (4) can be computed directly in $\mathcal{H}$ in time linear in n. Such parametric features limit the expressiveness of the model: Strathmann et al. [26] observed that the score estimate oscillates in regions where little or no data has been observed, leading to poor HMC behavior when the sampler enters those regions. Therefore, we do not further pursue this approach in the present work.

The *lite* approximation instead finds an estimator $f \in \operatorname{span}\{k(x, \cdot)\}_{x \in X}$. This has a similar spirit to Nyström approaches, but note the differing basis from (4), which is based on kernel *derivatives*, and that it uses the entirety of X, so the dependence on n is improved only by simple subsampling. Strathmann et al. [26] derived an estimator for the specific case of Gaussian kernels.

The generalized version of Theorem 1 (Lemma 4 in the appendix) covers the basis used by the lite approximation, allowing us to generalize this method to basis sets $Y \neq X$ and to kernels other than the Gaussian; Appendix B.1 discusses this in more detail.

4 Theory

We analyze the performance of our estimator in the well-specified case: assuming that the true density p_0 is in $\mathcal{P}$ (and thus corresponds to some $f_0 \in \mathcal{H}$), we obtain both the parameter convergence of $f^m_{\lambda,n}$ to f_0 and the convergence of the corresponding density $p_{f^m_{\lambda,n}}$ to the true density p_0.

Theorem 2. *Assume the conditions listed in Appendix A.2 (similar to those of Sriperumbudur et al. [24] for the well-specified case), and use the $\mathcal{H}_Y$ of (6) with the basis set Y chosen uniformly at random from the size-m subsets of the training set X. Let $\beta \geq 0$ be the range-space smoothness parameter of the true density f_0, and define $b = \min\left(\beta, \frac{1}{2}\right)$, $\theta = \frac{1}{2(b+1)} \in [\frac{1}{3}, \frac{1}{2}]$. As long as $m = \Omega\left(n^\theta \log n\right)$, then with $\lambda = n^{-\theta}$ we obtain*

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

The first statement implies that $p_{f^m_{\lambda,n}}$ also converges to p_0 in L_r ($1 \leq r \leq \infty$) and Hellinger distances at a rate $\mathcal{O}_{p_0}\left(n^{-\frac{b}{2(b+1)}}\right)$, and that $\mathrm{KL}(p_0 \| p_{f^m_{\lambda,n}}), \mathrm{KL}(p_{f^m_{\lambda,n}} \| p_0)$ are each $\mathcal{O}_{p_0}\left(n^{-\frac{b}{b+1}}\right)$.

The rate of convergence in J exactly matches the rate for the full-data estimator $f_{\lambda,n}$ shown by [24] in J; the rates in other divergences essentially match, except that ours saturate slightly sooner as β increases. Thus, for any problem satisfying the assumptions, we can achieve the same statistical properties as the full-data setting with $m = \Omega\left(\sqrt{n} \log n\right)$, while in the smoothest problems we need only $m = \Omega\left(n^{1/3} \log n\right)$.

This substantial reduction in computational expense is in contrast to the comparable analysis for kernel ridge regression [20], which for the hardest problems requires $m = \Omega(n \log n)$, giving no computational savings at all. In the best general case, it also needs $m = \Omega(n^{1/3} \log n)$. This rate was itself a significant advance: a prior analysis based on stability of the kernel approximation [7]

results in a severe additional penalty when using Nyström, matching the worst-case error rates for the full solution, yet still requiring $m = \Omega(n)$ (i.e., according to the earlier reasoning, we would not be guaranteed to benefit from improved rates in easier problems). A finite-sample version of Theorem 2, with explicit constants, is stated and proved in Appendix C.

Proof outline Our proof uses techniques both from the analysis of the full-data estimator [24] and from an analysis of generalization error for Nyström-subsampled kernel ridge regression [20].

Each of the losses considered in Theorem 2 can be bounded in terms of $\|f - f_0\|_{\mathcal{H}}$. We decompose this loss relative to the $f_\lambda^m = \operatorname{argmin}_{f \in \mathcal{H}_Y} J(f) + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2$, the best regularized estimator in population with the particular basis Y. That is,

$$\|f_{\lambda,n}^m - f_0\|_{\mathcal{H}} \leq \|f_{\lambda,n}^m - f_\lambda^m\|_{\mathcal{H}} + \|f_\lambda^m - f_0\|_{\mathcal{H}}. \tag{8}$$

The first term on the right-hand side of (8) is the *estimation error*, which represents our error due to having a finite number of samples n: this term decreases as $n \to \infty$, but it will increase as $\lambda \to 0$. It could conceivably increase as $m \to \infty$ as well, but we show using concentration inequalities in $\mathcal{H}$ that no matter the m, the estimation error is $\mathcal{O}_{p_0}\left(\frac{1}{\lambda\sqrt{n}}\right)$.

The last term of (8) is the *approximation error*, where "approximation" refers both to the regularization by λ and the restriction to the subspace $\mathcal{H}_Y$. This term is independent of n; it decreases as $\mathcal{H}_Y$ grows (i.e. as $m \to \infty$), and also decreases as $\lambda \to 0$, as we allow ourselves to more directly minimize the population risk. The key to bounding this term is to exploit the nature of the space $\mathcal{H}_Y$. This can be done by analogy with the treatment of the "computational error" term of Rudi et al. [20], where we show that any components of f_0 not lying within $\mathcal{H}_Y$ are relatively small in the parts of the space we observe; this is the only step of the proof that depends on the specific basis $\mathcal{H}_Y$. Having handled this contribution, we show that the approximation error term is $\mathcal{O}_{p_0}\left(\lambda^b\right)$ as long as $m = \Omega\left(\frac{1}{\lambda}\log\frac{1}{\lambda}\right)$.

The decay of the two terms is then optimized when $\lambda = n^\theta$, with θ as given in the proof.

The rate in the Fisher divergence J is better because that metric is weighted towards points in the space where we actually see data, as opposed to uniformly across the space as in (8). Our proof technique, similarly to that of Sriperumbudur et al. [24], allows us to account for this with an improved dependence on λ in the evaluation of both estimation and approximation errors.

Remarks Approaches like those of [9, 30], which bound the difference in training error of Nyström-type approximations to kernel ridge regression, are insufficient for our purposes: we need to ensure that the estimated function $f_{\lambda,n}^m$ converges to f_0 everywhere, so that the full distribution matches, not just its values at the training points. In doing so, our work is heavily indebted to [6], as are [20, 24].

We previously noted that using $Y = X$ does not yield an identical estimator, $f_{\lambda,n}^n \neq f_{\lambda,n}$. In fact, we could achieve this by additionally including $\hat{\xi}$ within the space (6); it would also not be too hard to alter our proof to account for this, achieving the same asymptotic rates. Since evaluating $\hat{\xi}$ requires touching all the data points, however, we would lose the test-time improvements in both computation and memory achieved by the estimator of Theorem 1. Moreover, the experiments of Section 5 show that dropping $\hat{\xi}$ from the basis does not seem to be harmful in practice.

5 Experiments

We now validate our estimator empirically. We first consider synthetic densities in Section 5.1, where we know the true densities and can evaluate convergence of the score estimates analytically with (1). In Section 5.2 we evaluate our estimator in the gradient-free Hamiltonian Monte Carlo setting of [26], where (in the absence of a ground truth) we compare the efficiency of the resulting sampler.

For all exponential family variants, we take q_0 to be a uniform distribution with support encompassing the samples, and use a Gaussian kernel $k(x, y) = \exp\left(-\|x - y\|^2/\sigma\right)$, with a tuned bandwidth σ and regularization parameter λ. We compare the following models: (i) The **full** model, (4) and (5), from [24]. (ii) The **lite** model from [26], which subsamples the dataset X to size m, and uses the basis $\{k(X_a, \cdot)\}$, ignoring the remaining datapoints (unlike the Nyström case). Their latest code uses the regularization $\lambda(\|f\|_{\mathcal{H}}^2 + \|\beta\|_2^2)$, which we maintain for our experiments. (iii) The **nyström**

estimator of Theorem 1, choosing m distinct data points uniformly at random for Y. For numerical stability, we add $10^{-5}I$ to the matrix being inverted in (7), corresponding to a small L_2 regularizer on the weights β. (iv) The **dae** model of Alain and Bengio [1], where we train a two-layer denoising autoencoder, with $\tanh$ code activations and linear decoding. We train with decreasing noise levels (100σ, 10σ, σ), using up to 1000 iterations of BFGS each. We tune the number of hidden units and σ, since, while [1] recommend simply choosing some small σ, this plays a similar role to a bandwidth, and its careful choice is essential. We differentiate the score estimate to obtain the second derivative needed to evaluate (2).

5.1 Score convergence on synthetic densities

We first consider two synthetic densities, where the true score is available: The 'ring' dataset takes inspiration from the 'spiral' dataset of [1, Figure 5], being a similarly-shaped distribution but possessing a probability density for evaluation purposes. We sample points uniformly along three circles with radii $(1, 3, 5)$ in $\mathbb{R}^2$ and add Gaussian noise of standard deviation 0.1 in the radial direction. We then add extra dimensions consisting of independent Gaussian noise with standard deviation 0.1. The 'grid' dataset generalizes the 2-component mixture example of Sriperumbudur et al. [24, Figure 1] into a more challenging d-dimensional mixture. We first pick d random (using a fixed seed) vertices of a d-dimensional hypercube, and then construct a mixture of normal distributions, one at each selected vertex.

For each run, we generate $n = 500$ training points and estimate the score on 1500 (grid) or 5000 (ring) newly generated test points. In both cases, we estimate the true score (1) on these test points to ensure a "best case" comparison of the models, though we confirmed that using (2) leads to indistinguishable parameter selections and performance. For **lite** and **nyström**, we independently evaluated the parameters for each subsampling level. We report performances for the best parameters found for each method. All experiments were conducted in a single CPU thread for timing comparisons, although multi-core parallelization is straightforward for all models.

Figure 1 shows convergence of the score as the dimension increases. On both the ring and grid datasets, **nyström** performs very close to the full solution, in addition to large computational savings. With reasonable performance penalties at $m = 42$, we achieve a major reduction in cost and storage over the original $n = 500$ sample size. The **lite** performance is similar to that of **nyström** at comparable levels of data retention. As expected, the performance of **nyström** gets closer to that of **full** as m increases towards n. The autoencoder performs consistently worse than any of the kernel models, on both datasets. Autoencoder results are also strongly clustered, with only small performance improvements as the number of hidden units increases. For the grid data, we observe that in 20 dimensions, all solutions start to converge to a similar score. This indicates that none of the methods are able to learn the structure for this number of training points and dimensions, and all solutions effectively revert to smooth, uninformative estimates.

In respect of computational cost, the **lite** solution does best, followed by **nyström** for low to moderate m, with significant savings over the full solution even at $m = 167$ on the grid, and across all m on the ring. The additional cost of **nyström** over **lite** arises since it must compute all derivatives at the retained samples. The autoencoder runtimes are longer than the other methods, although we point out that the settings recommended in Alain and Bengio [1] are not optimized for run-time. We observed, however, that replacing BFGS with stochastic gradient descent or avoiding the "decreasing noise" schedule both lead to instabilities in the solution.

5.2 Gradient-free Hamiltonian Monte Carlo

Our final experiment follows methodology and code by Sejdinovic et al. [22] and Strathmann et al. [26] in constructing a gradient-free HMC sampler using score estimates learned on the previous MCMC samples. Our goal is to efficiently sample from the marginal posterior over hyperparameters of a Gaussian process (GP) classifier on the UCI Glass dataset [15]. Closed-form expressions for the score (and therefore HMC itself) are not available, due to the intractability of the marginal data likelihood given the hyperparameters. Using an Expectation Propagation approximation to the GP posterior and importance sampling, however, it is possible to compute unbiased estimates of the likelihood, and to construct a Pseudo-Marginal MCMC method [10]. We compare all score estimators' ability to generate an HMC-like proposal as described in Strathmann et al. [26]. An

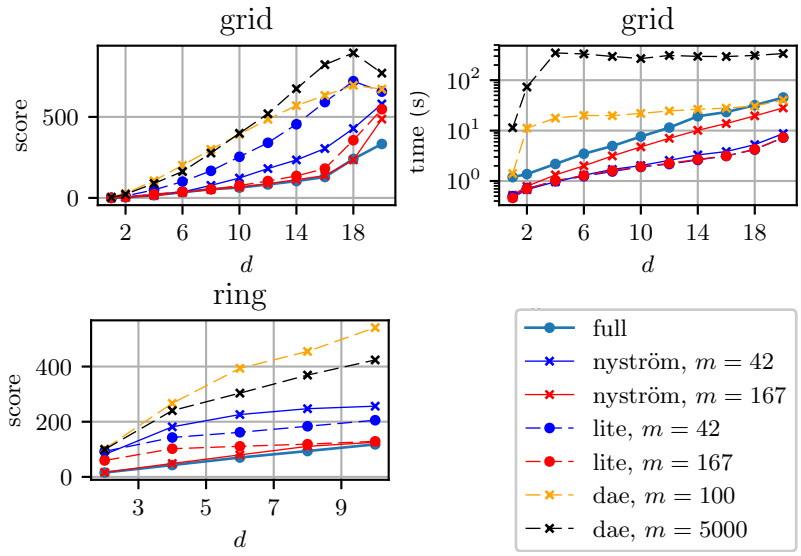

Figure 1: Convergence and timing on synthetic ring and grid data.

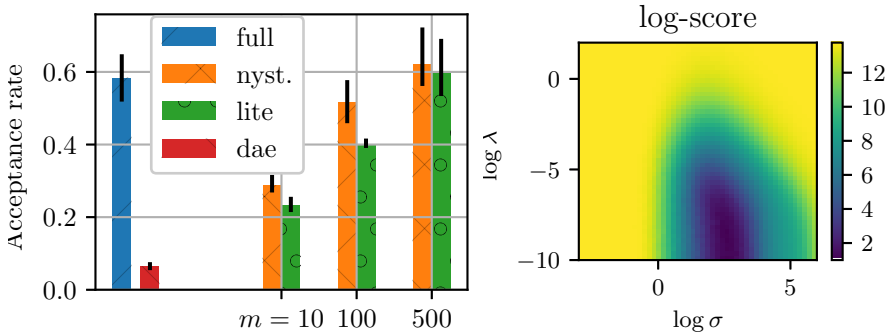

Figure 2: Left: HMC acceptance rate on the GP-glass posterior, with 90% quantiles. Right: Hyperparameter surface of the score for **nyström** with $m = 42$.

accurate score estimate would result in proposals close to an idealized HMC move, which would have a high acceptance probability. Thus, higher acceptance rates indicate better score estimates.

Our experiment assumes the idealized scenario where a burn-in is successfully completed. We run 40 random walk adaptive-Metropolis MCMC samplers for 30 000 iterations, discard the first 10 000 samples, and thin by a factor of 400. Merging these samples results in 2 000 posterior samples. We fit all score estimators on a random subset of $n = 500$ of these samples, and use the remaining 1500 samples to tune the model hyperparameters. The validation surface obtained for **nyström** by the estimated score objective on the held-out set is shown in Figure 2: we note that it is smooth and easily optimised. For **dae** (not shown here), a well-tuned level of corruption noise is essential. Starting from a random point of the initial posterior sketch, we construct trajectories along the surrogate Hamiltonian using 100 steps of size 0.1, and a standard Gaussian momentum. We compute the hypothetical acceptance probability for each step, and average over the trajectory.

Figure 2 shows the results averaged over 200 repetitions. As before, **nyström** matches the performance of **full** for $m = n = 500$, while for $m = 100$ it attains a high acceptance rate at a considerably reduced computational cost. It also reliably outperforms **lite** for lower m, which might occur since **lite** sub-samples the data while **nyström** only sub-samples the basis. **dae** does relatively poorly, despite a large grid-search for its hyperparameters. For any of the models, untuned hyperparameters easily lead to an acceptance rate close to zero.

Acknowledgements

The authors would like to thank Mladen Kolar for productive discussions.

References

[1] G. Alain and Y. Bengio. "What regularized auto-encoders learn from the data-generating distribution." In: *JMLR* 15.1 (2014), pp. 3563–3593. arXiv: 1211.4246.

[2] A. Barron and C.-H. Sheu. "Approximation of density functions by sequences of exponential families." In: *Annals of Statistics* 19.3 (1991), pp. 1347–1369.

[3] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

[4] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Hayward, CA: IMS, 1986.

[5] S. Canu and A. J. Smola. "Kernel methods and the exponential family." In: *Neurocomputing* 69.7 (2006), pp. 714–720.

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

[7] C. Cortes, M. Mohri, and A. Talwalkar. "On the impact of kernel approximation on learning accuracy." In: *AISTATS*. 2010.

[8] P. Drineas and M. W. Mahoney. "On the Nyström method for approximating a Gram matrix for improved kernel-based learning." In: *Journal of Machine Learning Research* 6 (2005), pp. 2153–2175.

[9] A. El Alaoui and M. W. Mahoney. "Fast Randomized Kernel Methods With Statistical Guarantees." In: *NIPS*. 2015. arXiv: 1411.0306.

[10] M. Filippone and M. Girolami. "Pseudo-marginal Bayesian inference for Gaussian Processes." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2014).

[11] K. Fukumizu. "Exponential manifold by reproducing kernel Hilbert spaces." In: *Algebraic and Geometric Methods in Statistics*. Cambridge University Press, 2009, pp. 291–306.

[12] C. Gu and C. Qiu. "Smoothing spline density estimation: Theory." In: *Annals of Statistics* 21.1 (1993), pp. 217–234.

[13] A. Hyvärinen. "Estimation of non-normalized statistical models by score matching." In: *JMLR* 6.Apr (2005), pp. 695–709.

[14] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang. "A Tutorial on Energy-Based Learning." In: *Predicting Structured Data*. MIT Press, 2006, pp. 191–246.

[15] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: http://archive.ics.uci.edu/ml.

[16] R. Neal. "MCMC using Hamiltonian dynamics." In: *Handbook of Markov Chain Monte Carlo* 2 (2011).

[17] C. J. Oates, M. Girolami, and N. Chopin. "Control functionals for Monte Carlo integration." In: *J. R. Statist. Soc. B* 79.3 (2017), pp. 695–718. arXiv: 1410.2392.

[18] A. Rahimi and B. Recht. "Random Features for Large-Scale Kernel Machines." In: *NIPS*. 2007.

[19] A. Rudi, R. Camoriano, and L. Rosasco. *Generalization Properties of Learning with Random Features*. 2016. arXiv: 1602.04474.

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

[21] B. Schölkopf, R. Herbrich, and A. J. Smola. "A Generalized Representer Theorem." In: *COLT*. 2001.

[22] D. Sejdinovic, H. Strathmann, M. Lomeli, C. Andrieu, and A. Gretton. "Kernel Adaptive Metropolis-Hastings." In: *ICML*. 2014.

[23] A. J. Smola and B. Schölkopf. "Sparse Greedy Matrix Approximation for Machine Learning." In: *ICML*. 2000.

[24] B. K. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. *Density Estimation in Infinite Dimensional Exponential Families*. 2013. arXiv: 1312.3516.

[25] B. K. Sriperumbudur and Z. Szábo. “Optimal rates for random Fourier features.” In: *NIPS*. 2015. arXiv: 1506.02155.

[26] H. Strathmann, D. Sejdinovic, S. Livingstone, Z. Szábo, and A. Gretton. “Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families.” In: *NIPS*. 2015. arXiv: 1506.02564.

[27] D. J. Sutherland and J. Schneider. “On the Error of Random Fourier Features.” In: *UAI*. 2015. arXiv: 1506.02785.

[28] C. K. I. Williams and M. Seeger. “Using the Nyström method to speed up kernel machines.” In: *NIPS*. 2000.

[29] D. P. Woodruff. “Sketching as a Tool for Numerical Linear Algebra.” In: *Foundations and Trends in Theoretical Computer Science* 10.1–2 (2014), pp. 1–157. arXiv: 1411.4357.

[30] Y. Yang, M. Pilanci, and M. J. Wainwright. “Randomized Sketches for Kernels: Fast and Optimal Non-Parametric Regression.” In: *Annals of Statistics* (forthcoming). arXiv: 1501.06195.

Appendices

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

A Preliminaries

We will first establish some definitions that will be useful throughout, as well as overviewing some relevant results from [5].

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 [6, 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 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 $L^1(\Omega, p_0)$, and $q_0 \in C^1(\Omega)$.
- **(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 $R < \infty$.
- **(G)** (Bounded derivatives) $\operatorname{supp}(q_0) = \mathcal{H}$, and the following quantities are finite:

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

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

These assumptions, or closely related ones, were all used by [5] 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 [5] 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.

We will also use the following quantities:

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

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

A.3 Operator definitions

The following objects will be useful in our study:

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:*

$$C = \mathbb{E}_{x\sim p_0}\left[\sum_{i=1}^{d} \partial_i k(x,\cdot) \otimes \partial_i k(x,\cdot)\right] : \mathcal{H} \to \mathcal{H}; \qquad C_\lambda = C + \lambda I \tag{9}$$

$$\xi = -Cf_0 = \mathbb{E}_{x\sim p_0}\left[\sum_{i=1}^{d} \partial_i k(x,\cdot)\partial_i \log q_0(x) + \partial_i^2 k(x,\cdot)\right] \in \mathcal{H} \tag{10}$$

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

here $e_{(b,i)}$ has component $(b-1)d+i$ equal to 1 and all others 0. Note Z_X evaluates derivatives, $(Z_X f)_{(b,i)} = \partial_i f(X_b)$, whereas for $\alpha \in \mathbb{R}^{nd}$, $Z_X^*\alpha = \sum_{b=1}^{n}\sum_{i=1}^{d} \alpha_{(b,i)}\partial_i k(X_b,\cdot)$.

Define estimators of (9) and (10) by

$$\hat{C} = \frac{1}{n} Z_X^* Z_X = \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \partial_i k(X_a,\cdot) \otimes \partial_i k(X_a,\cdot) : \mathcal{H} \to \mathcal{H}; \qquad \hat{C}_\lambda = \hat{C} + \lambda I \tag{11}$$

$$\hat{\xi} = \frac{1}{n}\sum_{a=1}^{n}\sum_{i=1}^{d} \partial_i k(X_a,\cdot)\partial_i \log q_0(X_a) + \partial_i^2 k(X_a,\cdot) \in \mathcal{H}. \tag{12}$$

C, ξ, and their estimators were defined by Sriperumbudur et al. [5]. C is similar to the standard covariance operator in similar analyses [3, 4].

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

$$g_Y(A) = V(V^*AV)^{-1}V^*. \tag{13}$$

The projected inverse function g_Y, defined by Rudi et al. [4], will be crucial in our study, and so we first establish some useful properties of it.

Lemma 1 (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 (13) 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) $g_Y(A_\lambda) = (P_Y A P_Y + \lambda I)^{-1} P_Y$

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

Proof. (i) and (ii) follow from $V^* P_Y = V^* V V^* = V^*$ and $P_Y V = V V^* V = V$, respectively. (iii) is similar: $g_Y(A_\lambda)A_\lambda P_Y = V(V^* A_\lambda V)^{-1} V^* A_\lambda V V^* = V V^*$. 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 $V(V^*A_\lambda V)V^* = V(V^*AV + \lambda V^*V)V^* = (P_Y A P_Y + \lambda I)P_Y$, so we have

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

left-multiplying both sides by $(P_Y A P_Y + \lambda I)^{-1}$ 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. □

A.4 Full-data result

This result is essentially Theorem 3 of [5].

Lemma 2. *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 $J_\lambda(f) = J(f) + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2$ is

$$f_\lambda = \operatorname*{argmin}_{f \in \mathcal{H}} J_\lambda(f) = -C_\lambda^{-1}\xi = C_\lambda^{-1} C f_0.$$

Using the estimators (11) and (12), 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 $\hat{J}_\lambda(f) = \hat{J}(f) + \frac{1}{2}\lambda\|f\|_{\mathcal{H}}^2$:

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

$f_{\lambda,n}^m$ can be computed according to [5, Theorem 4].

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 [4].

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

$$f^m_{\lambda,n} = -(P_Y \hat{C} P_Y + \lambda I)^{-1} P_Y \hat{\xi} = -g_Y(\hat{C}_\lambda)\hat{\xi}. \tag{14}$$

Proof. We begin by rewriting the minimization using Lemma 2 as

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

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

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

so that the problem

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

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

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

which since $\hat{C}$ is positive yields the first equality of (14). The second follows from Lemma 1 (iv). □

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

$$\begin{aligned}
f^m_{\lambda,n} &= Z^*_Y \beta_Y = \sum_{a=1}^{M} (\beta_Y)_a y_a \\
\beta_Y &= -(\tfrac{1}{n} B^{\mathsf{T}}_{XY} B_{XY} + \lambda G_{YY})^\dagger h_Y,
\end{aligned} \tag{15}$$

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

$$\begin{aligned}
(B_{XY})_{(b,i),a} &= \langle \partial_i k(X_b, \cdot), y_a\rangle_{\mathcal{H}} \\
(G_{YY})_{a,a'} &= \langle y_a, y_{a'}\rangle_{\mathcal{H}} \\
(h_Y)_a &= \langle \hat{\xi}, y_a\rangle_{\mathcal{H}}.
\end{aligned} \tag{16}$$

Proof. First, $B_{XY} = Z_X Z_Y^*$, $G_{YY} = Z_Y Z_Y^*$, and $h_Y = Z_Y \hat{\xi}$. For example, (16) 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 $A^\dagger = G^\dagger F^\dagger$ [1, chap. 1, sec. 6, ex. 17].

Now we can show that the claimed form (15) matches $f^m_{\lambda,n}$ from (14):

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

□

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 [7] 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

$$\alpha = -\frac{\tau}{2}(A + \lambda I)^{-1} b \tag{17}$$

$$A = \sum_{i=1}^{d} -[D_{x_i}K - KD_{x_i}]^2 \qquad b = \sum_{i=1}^{d}\left(\frac{2}{\tau}(Ks_i + D_{s_i}K\mathbf{1} - 2D_{x_i}Kx_i) - K\mathbf{1}\right)$$

where $x_i = [X_{1i} \quad \dots \quad X_{ni}]^\mathsf{T}$, $s_i = x_i \odot x_i$ 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

$$\beta'_Y = -\left(\frac{1}{n}(B'_{XY})^\mathsf{T} B'_{XY} + \lambda G'_{YY}\right)^\dagger h'_Y.$$

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 [7]'s Proposition 1 that when $Y = X$ and q_0 is uniform,

$$h'_X = \frac{2}{n\tau} b \qquad (B'_{XX})^\mathsf{T} B'_{XX} = \frac{4}{\tau^2} A \qquad G'_{XX} = K.$$

Thus

$$\beta'_X = -\left(\frac{4}{n\tau^2}A + \lambda K\right)^\dagger \frac{2}{n\tau} b = -\frac{\tau}{2}\left(A + \frac{1}{4}n\tau^2\lambda K\right)^\dagger b. \tag{18}$$

(18) resembles (17), except that our approach regularizes A with $\frac{1}{4}n\tau^2\lambda K$ rather than λI. This is because, despite claims by [7] 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 2, 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 16 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

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

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

$$\begin{aligned} f^m_\lambda &:= \operatorname*{argmin}_{f \in \mathcal{H}_Y} \frac{1}{2}\langle f, P_Y C P_Y f\rangle_{\mathcal{H}} + \langle f, P_Y \xi\rangle_{\mathcal{H}} + \frac{1}{2}\lambda \|f\|^2_{\mathcal{H}} \\ &= -(P_Y C P_Y + \lambda I)^{-1} P_y \xi = -g_Y(C_\lambda)\xi = g_Y(C_\lambda) C f_0.\end{aligned}$$

Then we have

$$\|C^\alpha(f^m_{\lambda,n} - f_0)\|_{\mathcal{H}} \leq \|C^\alpha(f^m_{\lambda,n} - f^m_\lambda)\|_{\mathcal{H}} + \|C^\alpha(f^m_\lambda - f_0)\|_{\mathcal{H}}. \tag{19}$$

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 [4, Theorem 2]:

$$\begin{aligned} f_0 - f^m_\lambda &= (I - g_m(C_\lambda)C) f_0 \\ &= (I - g_m(C_\lambda)C_\lambda + \lambda g_m(C_\lambda))\, f_0 \\ &= (I - g_m(C_\lambda)C_\lambda(VV^*) - g_m(C_\lambda)C_\lambda(I - VV^*) + \lambda g_m(C_\lambda))\, f_0 \\ &= ((I - VV^*) - g_m(C_\lambda)C_\lambda(I - VV^*) + \lambda g_m(C_\lambda))\, f_0,\end{aligned}$$

where in the last line we used Lemma 1 (iii). Thus, using Assumption (F) and Lemma 1 (v),

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

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

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

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

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

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

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

Combining the two yields

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

and so

$$\|C^\alpha (f_\lambda^m - f_0)\|_{\mathcal{H}} \leq R \left(2 \left\| C_\lambda^{\frac{1}{2}} (I - VV^*) \right\|^2 + \lambda \right) c(\alpha) c(\beta). \tag{20}$$

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 $\|P_Y\| \leq 1$,

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

We have already bounded $\|f_\lambda^m - f_0\|_{\mathcal{H}}$, and have $\|f_0\|_{\mathcal{H}} \leq \|C^\beta\| \|C^{-\beta} f_0\|_{\mathcal{H}} \leq R \|C\|^\beta$. Using Lemma 1 (iv) and (v), we have

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

and so

$$\left\| C^\alpha (f_\lambda^m - f_{\lambda,n}^m) \right\|_{\mathcal{H}} \leq \frac{\left\| C^\alpha \hat{C}_\lambda^{-\frac{1}{2}} \right\|}{\sqrt{\lambda}} \Big(\|\hat{\xi} - \xi\|_{\mathcal{H}} + \|\hat{C} - C\| \left(\|f_\lambda^m - f_0\|_{\mathcal{H}} + R \|C\|^\beta \right) \Big). \tag{21}$$

The claim follows by using (20) and (21) in (19). □

C.2 Probabilistic inequalities

We only need Lemma 5 for $\alpha = 0$ and $\alpha = \frac{1}{2}$; in the former case, we use $\left\|\hat{C}_\lambda^{-\frac{1}{2}}\right\| \leq 1/\sqrt{\lambda}$. 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 $0 < \lambda \leq \frac{1}{4}\|C\|$, define $w := \log \frac{25 \operatorname{Tr} C}{\lambda \delta}$. When*

$$n \geq \max\left(\frac{4w}{3\rho}, \frac{40 d \mathcal{N}'_\infty(\lambda) w}{\rho^2}\right),$$

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

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

Proof. Let $\gamma := \lambda_{\max}\left(C_\lambda^{-\frac{1}{2}}(C - \hat{C})C_\lambda^{-\frac{1}{2}}\right)$. Lemma 15 gives that $\|C^{\frac{1}{2}}\hat{C}_\lambda^{-\frac{1}{2}}\| \leq \frac{1}{\sqrt{1-\gamma}}$ as long as $\gamma < 1$. We bound γ with Lemma 13, 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 $\gamma \leq \rho$ with probability at least $1 - \delta$ as long as

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

which is satisfied by the condition on n. □

Lemma 7. *Choose the basis $y_{(a,i)} = \partial_i k(Y_a, \cdot)$, with the points $\{Y_a\}_{a \in [m]}$ sampled iid from p_0. Let $\rho, \delta \in (0,1)$, and define $w := \log \frac{25 \operatorname{Tr}(C)}{\lambda \delta}$. 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 $1 - \delta$ as long as

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

Proof. We have that $\|C_\lambda^{\frac{1}{2}}(I - VV^*)\|^2 \leq \lambda \left\|(\frac{1}{m} Z_Y^* Z_Y + \lambda I)^{-\frac{1}{2}} C_\lambda^{\frac{1}{2}}\right\|^2$ via Lemma 14. Again applying Lemmas 13 and 15 as in the proof of Lemma 6 yields the result. □

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*

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

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

$$\|C_x - C\| \leq 2\sum_{i=1}^{d} \|\partial_i k(x,\cdot)\|_{\mathcal{H}}^2 \leq 2d\kappa_1^2$$

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

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* $S_\delta := 1 + \sqrt{2\log\frac{4}{\delta}}$*. Use the basis* $y_{(a,i)} = \partial_i k(Y_a,\cdot)$*, for* $\{Y_a\}_{a=1}^m$ *an iid sample from* p_0 *not necessarily independent of* X*. Assume that* $0 < \lambda < \frac{1}{4}\|C\|$*. When*

$$\min(n,\, m) \geq \frac{90d\kappa_1^2}{\lambda}\log\frac{100d\kappa_1^2}{\lambda\delta},$$

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

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

Proof. Recall from Lemma 5 that

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

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 $\mathcal{N}'_\infty(\lambda) \leq \kappa_1^2/\lambda$.

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

$$\max\left(2,\ 90d\mathcal{N}'_\infty(\lambda)\right)\log\frac{25\operatorname{Tr} C}{\lambda\frac{\delta}{4}} \leq \frac{90d\kappa_1^2}{\lambda}\log\frac{100d\kappa_1^2}{\lambda\delta},$$

where we used that $\lambda < \frac{1}{4}\|C\|$. 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:

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

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

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

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

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

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 16 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 [2], 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 $\varepsilon \ge \mathbb{E}\left\|\frac{1}{n}\sum_i X_i\right\|$,

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

We also know that

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

so

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

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

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

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

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

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

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

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

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

Lemma 12 (Bernstein's inequality for a sum of random operators; Proposition 12 of [4])**.** *Let $\mathcal{H}$ be a separable Hilbert space, and $X_1, \dots, X_n$ a sequence of iid self-adjoint positive random operators on $\mathcal{H}$, with $\mathbb{E}\, X_1 = 0$, $\lambda_{\max}(X_1) \le L$ 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 $\delta \ge 0$, with probability at least $1-\delta$*

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

D.2 Concentration of sum of correlated operators

The following result is similar to Proposition 8 of [4], but the proof is considerably more complex due to the sum over correlated operators.

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. 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 $Q_\lambda = Q + \lambda I$, $V_a = \sum_{i=1}^d Y_i^a \otimes Y_i^a$.*

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

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

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, clearly $\mathbb{E}\, Z_a = 0$, and since V_a is positive

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

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

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

Using the identity $2\langle x, Ay\rangle = \langle x+y, A(x+y)\rangle - \langle x, Ax\rangle - \langle y, Ay\rangle$, we get:

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

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

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

and the second is

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

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

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

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

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

We also have that

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

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

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

Thus

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

where we defined $M := \mathbb{E}\left[\left(\sum_{i=1}^{d}Y_i\right)\otimes\left(\sum_{i=1}^{d}Y_i\right)\right]$. Thus we have the desired upper bound $E[Z^2] \preceq S := \mathcal{N}'_\infty(\lambda)Q_\lambda^{-\frac{1}{2}}(2M+3dQ)Q_\lambda^{-\frac{1}{2}}$. In order to be able to finally use Lemma 12, we still need to find an upper bound on $\operatorname{Tr} S$, and both upper and lower bounds on $\|S\|$.

To do so, we will first show that $M \preceq dQ$:

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

Thus $\operatorname{Tr}(M) \le d\operatorname{Tr}(Q)$, and so

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

We next wish to bound $\|S\|$. Again because $M \preceq dQ$, we have that

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

and so

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

A lower bound can be obtained as

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

so that when $\lambda \le \rho\|Q\|$ we have that

$$\|S\| \ge d\mathcal{N}'_\infty(\lambda)\left(\frac{3}{1+\rho} - 2\right).$$

At last we can apply Lemma 12 to obtain that with probability at least $1-\delta$,

$$\lambda_{\max}\left(\frac{1}{n}Z_a\right) \le \frac{2\beta'}{3n} + \sqrt{\frac{2\|S\|\beta'}{n}} \le \frac{2\beta}{3n} + \sqrt{\frac{10d\mathcal{N}'_\infty(\lambda)\beta}{n}},$$

where

$$\beta' := \log\frac{2\operatorname{Tr} S}{\delta\|S\|} \le \log\left(\frac{2}{\delta}\frac{5d\mathcal{N}'_\infty(\lambda)\operatorname{Tr} Q}{\lambda d\mathcal{N}'_\infty(\lambda)\left(\frac{3}{1+\rho}-2\right)}\right) = \log\left(\frac{10\operatorname{Tr} Q}{\lambda\delta\left(\frac{3}{1+\rho}-2\right)}\right) =: \beta$$

as required. □

D.3 Results on Hilbert space operators

Lemmas 14 and 15 were proven and used by [4].

Lemma 14 (Proposition 3 of [4]). *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 15 (Proposition 7 of [4]). *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 16 (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 [5] is simply $\mathcal{P}$, and the result applies directly. □

References

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

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

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

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

[5] B. K. Sriperumbudur, K. Fukumizu, R. Kumar, A. Gretton, and A. Hyvärinen. *Density Estimation in Infinite Dimensional Exponential Families*. 2013. arXiv: 1312.3516.

[6] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

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