Appendix: Rates of Convergence for Sparse Variational Gaussian Process Regression

A. Proof Of Lemma 1

Titsias [2014] proves the tighter upper bound,

$$\begin{aligned}\mathcal{L} \leq \mathcal{L}'_{\text{upper}} := &-\frac{N}{2}\log(2\pi)\\ &-\frac{1}{2}\log(|\mathbf{Q}_n|) - \frac{1}{2}\mathbf{y}^{\intercal}\Big(\mathbf{Q}_n + \widetilde{\lambda}_{max}\mathbf{I}\Big)^{-1}\mathbf{y}.\end{aligned}$$

Subtracting,

$$\begin{aligned}&\mathcal{L}'_{\text{upper}} - \mathcal{L}_{\text{lower}}\\ &= \frac{t}{2\sigma_n^2} + \frac{1}{2}\Big(\mathbf{y}^{\intercal}\Big(\mathbf{Q}_n^{-1} - (\mathbf{Q}_n + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{y}\Big). \qquad (1)\end{aligned}$$

Since $\mathbf{Q_{ff}}$ is symmetric positive semidefinite, $\mathbf{Q}_n$ is positive definite with eigenvalues bounded below by σ_n^2. Write, $\mathbf{Q}_n = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\intercal}$, where $\mathbf{U}$ is unitary and $\mathbf{\Lambda}$ is a diagonal matrix with non-increasing diagonal entries $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_N \geq \sigma_n^2$.

We can rewrite the second term (ignoring the factor of one half) in Equation 1 as,

$$(\mathbf{U}^{\intercal}\mathbf{y})^{\intercal}\Big(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda} + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)(\mathbf{U}^{\intercal}\mathbf{y}).$$

Define, $\mathbf{z} = (\mathbf{U}^{\intercal}\mathbf{y})$. Since $\mathbf{U}$ is unitary, $\|\mathbf{z}\| = \|\mathbf{y}\|$.

$$\begin{aligned}(\mathbf{U}^{\intercal}\mathbf{y})^{\intercal}\big(\mathbf{\Lambda}^{-1} &- (\mathbf{\Lambda} + t\mathbf{I})^{-1}\big)(\mathbf{U}^{\intercal}\mathbf{y})\\ &= \mathbf{z}^{\intercal}\Big(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda} + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{z}\\ &= \sum_i z_i^2 \frac{\widetilde{\lambda}_{max}}{\gamma_i^2 + \gamma_i\widetilde{\lambda}_{max}}\\ &\leq \|\mathbf{y}\|^2 \frac{\widetilde{\lambda}_{max}}{\gamma_N^2 + \gamma_N\widetilde{\lambda}_{max}}.\end{aligned}$$

The last inequality comes from noting that the fraction in the sum attains a maximum when γ_i is minimized. Since σ_n^2 is a lower bound on the smallest eigenvalue of $\mathbf{Q}_n$, we have,

$$\mathbf{y}^{T}\Big(\mathbf{Q}_n^{-1} - (\mathbf{Q}_n + \widetilde{\lambda}_{max}\mathbf{I})^{-1}\Big)\mathbf{y} \leq \frac{\widetilde{\lambda}_{max}\|\mathbf{y}\|^2}{\sigma_n^4 + \sigma_n^2\widetilde{\lambda}_{max}}.$$

Lemma 1 follows.

B. KL Divergence Gaussian Distributions

B.1. KL divergence between multivariate Gaussian distributions

We make use of the formula for KL divergences between multivariate Gaussian distributions in our proof of Lemma 2, and the univariate case in Proposition 1.

Recall the KL divergence from $p_1 \sim \mathcal{N}(\mathbf{m_1}, \mathbf{S_1})$ to $p_2 \sim \mathcal{N}(\mathbf{m_2}, \mathbf{S_2})$ both of dimension N is given by

$$\begin{aligned}\mathrm{KL}(p_1\|p_2) = \frac{1}{2}\bigg(&\mathrm{Tr}\big(\mathbf{S_2}^{-1}\mathbf{S_1}\big) + \log\bigg(\frac{|\mathbf{S_2}|}{|\mathbf{S_1}|}\bigg)\\ &+(\mathbf{m_1} - \mathbf{m_2})^{\intercal}\mathbf{S_2}^{-1}(\mathbf{m_1} - \mathbf{m_2}) - N\bigg) \geq 0. \qquad (2)\end{aligned}$$

The inequality is a special case of Jensen's inequality.

B.2. Proof of Upper Bound in Lemma 2

In the main text we showed,

$$\begin{aligned}\mathbb{E}_y\Big[\mathrm{KL}\Big(Q\|\hat{P}\Big)\Big] &= \frac{t}{2\sigma_n^2} + \int \mathcal{N}(\mathbf{y}; 0, \mathbf{K}_n)\\ &\times \log\bigg(\frac{\mathcal{N}(\mathbf{y}; 0, \mathbf{K}_n)}{\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_n)}\bigg)\mathrm{d}\mathbf{y}\end{aligned}$$

In order to complete the proof, we need to show that the second term on the right hand side is bounded above by $t/(2\sigma_n^2)$. Using Equation 2:

$$\begin{aligned}\mathbb{E}_y\Big[KL\Big(Q\|\hat{P}\Big)\Big] &= \frac{t}{2\sigma_n^2} - \frac{N}{2} + \frac{1}{2}\log\bigg(\frac{|\mathbf{Q}_n|}{|\mathbf{K}_n|}\bigg)\\ &\quad + \frac{1}{2}\mathrm{Tr}\big(\mathbf{Q}_n^{-1}(\mathbf{K}_n)\big)\\ &\leq \frac{t}{2\sigma_n^2} - \frac{N}{2} + \frac{1}{2}\mathrm{Tr}\Big(\mathbf{Q}_n^{-1}(\mathbf{Q}_n + \widetilde{\mathbf{K}}_{\mathbf{ff}})\Big). \qquad (3)\end{aligned}$$

The inequality follows from noting the log determinant term is negative, as $\mathbf{K}_n \succ \mathbf{Q}_n$ (i.e. $\mathbf{K}_n - \mathbf{Q}_n$ is positive definite). Simplifying the last term,

$$\begin{aligned}\frac{1}{2}\mathrm{Tr}(\mathbf{I}) + \frac{1}{2}\mathrm{Tr}\Big(\mathbf{Q}_n^{-1}\widetilde{\mathbf{K}}_{\mathbf{ff}})\Big) &\leq N/2 + t\lambda_1\big(\mathbf{Q}_n^{-1}\big)/2\\ &\leq N/2 + t/(2\sigma_n^2).\end{aligned}$$

The first inequality uses that for positive semi-definite symmetric matrices $\mathrm{Tr}(AB) \leq \mathrm{Tr}(A)\lambda_1(B)$ which is a special case of Hölder's inequality. The final line uses that the largest eigenvalue of $\mathbf{Q}_n^{-1}$ is bounded above by σ_n^{-2}. Using this in Equation 3 finishes the proof.

B.3. Proof of Proposition 1

Defining $\epsilon = 2KL(q\|p)$,

$$\begin{aligned}\epsilon &= \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{\sigma_2^2} - \log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - 1 \qquad (4)\\ &\geq x - \log(x) - 1\end{aligned}$$

where we have defined $x = \frac{\sigma_1^2}{\sigma_2^2}$.

Applying the lower bound $x - \log(x) - 1 \geq (x-1)^2/2 - (x-1)^3/3$,

$$\epsilon \geq (x-1)^2/2 - (x-1)^3/3.$$

A bound on $|x-1|$ that holds for all ϵ can then be found with the cubic formula. Under the assumption that $\epsilon < \frac{1}{5}$, we have $x - \log(x) < 1.2$ which implies $x \in [0.493, 1.77]$. For x in this range, we have

$$x - \log(x) - 1 \geq (x-1)^2/3$$

So,

$$|x-1| \leq \sqrt{3\epsilon}$$

This proves that,

$$1 - \sqrt{3\epsilon} < \frac{\sigma_1^2}{\sigma_2^2} < 1 + \sqrt{3\epsilon}.$$

From Equation 4 and $x - \log x > 1$,

$$|\mu_1 - \mu_2| \leq \sigma_2\sqrt{3\epsilon}.$$

Using our bound on the ratio of the variances completes the proof of Proposition 1.

C. Covariances for Interdomain Features

We compute the covariances for eigenvector and eigenfunction inducing features.

C.1. Eigenvector inducing features

Recall we have defined eigenvector inducing features by,

$$u_m = \sum_{i=1}^N w_i^{(m)} f(\mathbf{x}_i).$$

Then,

$$\begin{aligned}\mathrm{cov}(u_m, u_k) &= \mathbb{E}\left[\sum_{i=1}^N w_i^{(m)} f(\mathbf{x}_i) \sum_{j=1}^N w_j^{(k)} f(\mathbf{x}_j)\right]\\ &= \sum_{i=1}^N w_i^{(m)} \sum_{j=1}^N w_j^{(k)} \mathbb{E}[f(\mathbf{x}_i)f(\mathbf{x}_j)]\\ &= \sum_{i=1}^N w_i^{(m)} \sum_{j=1}^N w_j^{(k)} k(\mathbf{x}_i, \mathbf{x}_j).\end{aligned}$$

We now recognize this expression as $\mathbf{w}^{(m)\intercal}\mathbf{K_{ff}}\mathbf{w}^{(k)}$. Using the defining property of eigenvectors as well as orthonormality,

$$\mathrm{cov}(u_m, u_k) = \lambda_k(\mathbf{K_{ff}})\delta_{m,k}.$$

Similarly,

$$\begin{aligned}\mathrm{cov}(u_m, f(\mathbf{x}_i)) &= \mathbb{E}\left[\sum_{j=1}^N w_j^{(m)} f(\mathbf{x}_j) f(\mathbf{x}_i)\right]\\ &= \sum_{j=1}^N w_j^{(m)} \mathbb{E}[f(\mathbf{x}_j)f(\mathbf{x}_i)]\\ &= \sum_{j=1}^N w_j^{(m)} k(\mathbf{x}_j, \mathbf{x}_i).\end{aligned}$$

This is the i^{th} entry of the matrix vector product $\mathbf{K_{ff}}\mathbf{w}^{(m)} = \lambda_m(\mathbf{K_{ff}})\mathbf{w}_i^{(m)}$.

C.2. Eigenfunction inducing features

Recall we have defined eigenfunction inducing features by,

$$u_m = \int \phi_m(\mathbf{x}) f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

Then,

$$\begin{aligned}&\mathrm{cov}(u_m, u_k)\\ &= \mathbb{E}\left[\int \phi_m(\mathbf{x}) f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \int \phi_k(\mathbf{x}') f(\mathbf{x}') p(\mathbf{x}') d\mathbf{x}'\right]\\ &= \int \phi_m(\mathbf{x}) p(\mathbf{x}) \int \phi_k(\mathbf{x}') \mathbb{E}[f(\mathbf{x})f(\mathbf{x}')] p(\mathbf{x}') d\mathbf{x}' d\mathbf{x}\\ &= \int \phi_m(\mathbf{x}) p(\mathbf{x}) \int \phi_k(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') p(\mathbf{x}') d\mathbf{x}' d\mathbf{x}.\end{aligned}$$

The expectation and integration may be interchanged by Fubini's theorem, as both integrals converge absolutely since $p(\mathbf{x})$ is a probability density, the $\phi_m(\mathbf{x})$ are in $L^2(\mathcal{X})_p \subset L^1(\mathcal{X})_p$ and k is bounded.

We may then apply the eigenfunction property to the inner integral and orthonormality of eigenfunctions to the result yielding,

$$\mathrm{cov}(u_m, u_k) = \lambda_k \int \phi_k(\mathbf{x})\phi_m(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \lambda_k \delta_{m,k}.$$

With similar considerations,

$$\begin{aligned}\mathrm{cov}(u_m, f(\mathbf{x}_i)) &= \mathbb{E}\left[\int \phi_m(\mathbf{x}) f(\mathbf{x}) f(\mathbf{x}_i) p(\mathbf{x}) d\mathbf{x}\right]\\ &= \int \phi_m(\mathbf{x}) \mathbb{E}[f(\mathbf{x})f(\mathbf{x}_i)] p(\mathbf{x}) d\mathbf{x}\\ &= \lambda_m \phi_m(\mathbf{x}_i).\end{aligned}$$

Algorithm 1 Initialization of Inducing Points

```
**Input:** Training inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, number of points to choose, $M$, kernel $k$.
**Returns:** $Z$, a sample of $M$ inducing points drawn proportional to the determinant of $\mathbf{K}_{Z,Z}$
Initialize $Z = \{\}$
**while** $|Z| < M$ **do**
  **for** $\mathbf{x}_i \in \mathbf{X} \setminus Z$ **do**
    $[\mathbf{k}_{Z,i}]_m :=$ cov$(\mathbf{z}_m, \mathbf{x}_i)$.
    $V_i = k(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k}_{i,Z}\mathbf{K}_{Z,Z}^{-1}\mathbf{k}_{Z,i}$,
  **end for**
  Sample $\mathbf{x}_i$ with probability proportional to $V_i$
  Add $\mathbf{x}_i$ to $Z$
**end while**
```

D. Sampling from a Discrete k-DPP

In this section, we give the algorithm, Algorithm 1, described in Hennig & Garnett [2016] adapted to the discrete setting which is relevant to our application. We additionally show that it can be implemented with complexity $\mathcal{O}(NM^2)$.

D.1. Efficient Implementation of Algorithm 1

Let $\mathbf{K_Z}$ be the matrix with entries $[\mathbf{K_Z}]_{m',m''} = k(\mathbf{z}_{m'}, \mathbf{z}_{m''})$.

$$\mathbf{K_Z} = \begin{bmatrix} \mathbf{K_{Z-1}} & \mathbf{k_{Z-1,m}} \\ \mathbf{k}^\intercal_{\mathbf{Z-1,m}} & k(\mathbf{z}_m, \mathbf{z}_m) \end{bmatrix}$$

where $\mathbf{k_m}$ is an $(m-1)\times 1$ column vector with $[\mathbf{k_{Z-1,m}}]_i = k(\mathbf{z}_i, \mathbf{z}_m)$, for $i \leq m-1$. Using block matrix inversion,

$$\mathbf{K_Z}^{-1} = \begin{bmatrix} \mathbf{K}^{-1}_{\mathbf{Z-1}} + r\mathbf{a}\mathbf{a}^\intercal & \mathbf{a} \\ \mathbf{a}^\intercal & \frac{1}{r} \end{bmatrix}$$

with $r = k(\mathbf{z}_m, \mathbf{z}_m) - \mathbf{k_{Z-1,m}}^\intercal \mathbf{K}^{-1}_{\mathbf{Z-1}} \mathbf{k_{Z-1,m}}$ and $\mathbf{a} = -\frac{1}{r}\mathbf{K}^{-1}_{\mathbf{Z-1}}\mathbf{k_{Z-1,m}}$. Define,

$$\mathbf{t_{Z,i}} = \mathbf{K_Z}^{-1}\mathbf{k_{Z-1,m}} \quad \text{and} \quad \mathbf{t}'_{\mathbf{Z,i}} = \mathbf{K_Z}^{-1}\mathbf{k}_{\mathbf{Z-1,x_i}},$$

with, $[\mathbf{k}_{\mathbf{Z,x_i}}]_j = k(\mathbf{z}_j, \mathbf{x}_i)$. With these definition,

$$\mathbf{K_Z}^{-1} = \begin{bmatrix} \mathbf{K}^{-1}_{\mathbf{Z-1}} + \frac{1}{r}\mathbf{t_{Z-1,m}}\mathbf{t}^\intercal_{\mathbf{Z-1,m}} & -\frac{1}{r}\mathbf{t_{Z-1,m}} \\ -\frac{1}{r}\mathbf{t}^\intercal_{\mathbf{Z-1,m}} & \frac{1}{r} \end{bmatrix}$$

and

$$r = k(\mathbf{z}_m, \mathbf{z}_m) - \mathbf{k_{Z-1,m}}^\intercal \mathbf{t_{Z-1,m}}.$$

Additionally,

$$V_i = k(\mathbf{x}_i, \mathbf{x}_i) - \mathbf{k_{Z,i}}^\intercal \mathbf{t'_{Z,i}}.$$

We assume the kernel can be evaluated in constant time. The second term is an inner product between vectors of length m, and therefore has computational cost $\mathcal{O}(m)$.

We need to show that given $\mathbf{t'_{Z-1,j}}$ for all $j \leq N$, $\mathbf{t'_{Z,i}}$ can be computed in $\mathcal{O}(m)$. Using the formula for $\mathbf{K_Z}^{-1}$, and

$$\mathbf{k}_{\mathbf{Z,x_i}} = \begin{bmatrix} \mathbf{k}_{\mathbf{Z-1,x_i}} \\ k(\mathbf{z}_m, \mathbf{x}_i) \end{bmatrix},$$

We have

$$\mathbf{t'_{Z,i}} = \begin{bmatrix} \mathbf{t'_{Z-1,i}} + \frac{1}{r}\left(\mathbf{t_{Z-1,m}}\mathbf{t}^\intercal_{\mathbf{Z-1,m}}\mathbf{k}_{\mathbf{Z,x_i}} - k(\mathbf{z}_m, \mathbf{x}_i)\mathbf{t_{Z-1,m}}\right) \\ \frac{1}{r}\left(-\mathbf{t}^\intercal_{\mathbf{Z-1,m}}\mathbf{k}_{\mathbf{Z,x_i}} + k(\mathbf{z}_m, \mathbf{x}_i)\right) \end{bmatrix}$$

Therefore, in iteration m of the outer loop in Algorithm 1 updating each $\mathbf{t'_{Z,i}}$ can be done in $\mathcal{O}(m)$ by computing inner products of length $m-1$ vectors, there are N such vectors, and we must iterate through this loop M times, proving the complexity is $\mathcal{O}(NM^2)$.

E. Proof of Corollaries

E.1. Corollary 1

From Theorem 3,

$$\mathrm{KL}\left(Q\|\hat{P}\right) \leq \frac{C(M+1)}{2\sigma_n^2\delta}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma_n^2}\right) \tag{5}$$

Take $M = \frac{(3+\epsilon)\log(N) + \log D}{\log(B^{-1})}$. If $M \geq N$ the KL-divergence is zero and we are done. Otherwise, $C(M+1) < N^2\sum_{i=M+1}^\infty \lambda_i$. By the geometric series formula,

$$\sum_{i=M+1}^{\infty} \lambda_i = v\frac{\sqrt{2a}}{\sqrt{A}}\frac{B^M}{1-B}$$

Now, $B^M = N^{-3-\epsilon}D^{-1}$, so

$$\sum_{i=M+1}^{\infty} \lambda_i = \delta N^{-3-\epsilon},$$

and $\frac{C(M+1)}{2\delta\sigma_n^2} < N^{-1-\epsilon}$. Using this in Equation 5 completes the proof.

E.2. Corollary 2

It is sufficient to consider the case of isotropic kernels and input distributions.[1] From [Seeger et al., 2008] in the isotropic case (i.e. $B_i = B_j =: B$ for all $i, j \leq D$),

$$\lambda_{s+D-1} \leq \left(\frac{2a}{A}\right)^{D/2} B^{s^{1/D}}.$$

[1] For the general case, the eigenvalues can be bounded above by constant times the eigenvalues of an operator with an isotropic kernel with all lengthscales equal to the shortest kernel lengthscale and the input density standard deviation set to the largest standard deviation of any one-dimensional marginal of $p(\mathbf{x})$.

Define $\tilde{M} = M - D + 1$.

$$\sum_{s=\tilde{M}}^{\infty} \lambda_s \leq \left(\frac{2a}{A}\right)^{D/2} \sum_{s=M+1}^{\infty} B^{s^{1/D}} \tag{6}$$

$$< \left(\frac{2a}{A}\right)^{D/2} \int_{s=M}^{\infty} B^{s^{1/D}} ds =: \mathcal{I}. \tag{7}$$

In the second line, we use that $B < 1$, so $B^{s^{1/D}}$ obtains its minimum on the interval $s \in [s', s'+1]$ at the right endpoint (i.e. monotonicity). We now define $\alpha = -\log(B)$. So,

$$\mathcal{I} = \left(\frac{2a}{A}\right)^{D/2} \int_{s=M}^{\infty} \exp\Big(-\alpha s^{1/D}\Big) ds \tag{8}$$

$$= \left(\frac{2a}{A}\right)^{D/2} \alpha^{-D} D \int_{t=\alpha M^{1/D}}^{\infty} e^{-t} t^{D-1} dt \tag{9}$$

In the second line we made the substitution $t = \alpha s^{1/D}$, so $ds = \alpha^{-D} D t^{D-1}$. We now recognize,

$$\int_{t=\alpha M^{1/D}}^{\infty} e^{-t} t^{D-1} dt$$

as an incomplete gamma function, $\Gamma(D, \alpha M^{1/D})$. From Gradshteyn & Ryzhik [2014, 8.352],

$$\Gamma(D, y) = (D-1)! e^{-y} \sum_{k=0}^{D-1} \frac{y^k}{k!}$$

So,

$$\mathcal{I} = \left(\frac{2a}{A}\right)^{D/2} \alpha^{-D} D! e^{-\alpha M^{1/D}} \sum_{k=0}^{D-1} \frac{\alpha^k M^{k/D}}{k!} \tag{10}$$

As M grows as a function of N and D is fixed, for N large $D \leq \alpha M^{1/D}$. This implies that that the largest term in the sum on the right hand side is the final term, so

$$\mathcal{I} \leq \left(\frac{2a}{A}\right)^{D/2} \alpha^{-1} D^2 e^{-\alpha M^{1/D}} M^{(D-1)/D}. \tag{11}$$

Choose $M = \frac{1}{\alpha} \log\Big(N^{\gamma'} \big(\frac{2a}{A}\big)^{D/2} D^2 \alpha^{-1}\Big)^D$. Then

$$\left(\frac{2a}{A}\right)^{D/2} \alpha^{-1} D^2 e^{-\alpha M^{1/D}} = N^{-\gamma'}$$

and

$$M^{(D-1)/D} < M = \frac{1}{\alpha} \log\left(N^{\gamma'} \left(\frac{2a}{A}\right)^{D/2} D^2 \alpha^{-1}\right)^D,$$

so

$$\mathcal{I} \leq \frac{1}{\alpha} \log\left(N^{\gamma'} \left(\frac{2a}{A}\right)^{D/2} D^2 \alpha^{-1}\right)^D N^{-\gamma'}.$$

For any fixed D, for this choice of M for any $\epsilon > 0$ for N large,

$$\mathcal{I} = \mathcal{O}\Big(N^{-\gamma' + \epsilon}\Big).$$

By choosing $\gamma' > 3 + \epsilon'$, for some fixed $\epsilon' > 0$ the proof is complete, using a similar argument as the one used in the proof of the previous corollary.

Note that using the bound proven in Theorem 2 (the tightest of our bounds) the exponential scaling in dimension is unavoidable. If both k and $p(\mathbf{x})$ are isotropic, then the eigenvalue $(\frac{2a}{A})^D B^m$ appears $\binom{m+D-1}{D-1}$ times. This follows from noting that this is the number of ways to write m as a sum of D non-negative integers. Using the "rising sum" identity, as well as a standard lower bound for binomial coefficients $\sum_{i=1}^{K} \binom{m+D-1}{D-1} = \binom{K+D}{D} \geq \left(\frac{K+D}{D}\right)^D > C(D) K^D$. for some constant $C(D)$. Using Theorem 2 we need to choose M such that $\lambda_M = \mathcal{O}(1/N)$. This means choosing $K \gg \log(N)$ in the sum above, leading to at least $\alpha \log^D(N)$ features being needed for some constant α. The constant in this lower bound decays rapidly with D, while the constant in the upper bound did not exhibit this behavior. Better understanding this gap is important for understanding the performance of sparse Gaussian process approximations in high dimensions.

If the data actually lies on a lower dimensional manifold, we expect the scaling to only depend on the dimensionality of the manifold. In particular, if the manifold is linear, then the kernel matrix only depends on distances along the manifold (not in the space it is embedded in) so the eigenvalues will not be effected by the higher dimensional embedding. We conjecture that similar properties are exhibited when the data manifold is nonlinear.

F. Smoothness and Sacks-Ylivasker Conditions

In many instances the precise eigenvalues of the covariance operator are not available, but the asymptotic properties are well understood. A notable example is when the data is distributed uniformly on the unit interval. If the kernel satisfies the Sacks-Ylivasker condtions of order r:

- $k(x, x')$ is r-times continuously differentiable on $[0,1]^2$ Moreover, $k(x, x')$ has continuous partial derivatives up to order $r+2$ times on $(0,1)^2 \cap (x > x')$ and $(0,1)^2 \cap (x < x')$. These partial derivatives can be continuously extended to the closure of both regions.
- Let L denote $k^{(r,r)}(x, x')$, L_+ denote the restriction of L to the upper triangle and L_- the restriction to the lower triangle, then $L_+^{(1,0)} < L_-^{(1,0)}$ on the diagonal $x = x'$.

- $L_+^{(2,0)}(s,\cdot)$ is an element of the RKHS associated to L and has norm bounded independent of s.

Notably, Matérn half integer kernels of order $r+1/2$ meet the S-Y condition of order r. See [Ritter et al., 1995] for a more detailed explanation of these conditions and extensions to the multivariate case.

References

Gradshteyn, I. S. and Ryzhik, I. M. *Table of Integrals, Series, and Products*. Academic press, 2014.

Hennig, P. and Garnett, R. Exact Sampling from Determinantal Point Processes. *arXiv preprint arXiv:1609.06840*, 2016.

Ritter, K., Wasilkowski, G. W., and Woźniakowski, H. Multivariate Integration and Approximation for Random Fields Satisfying Sacks-Ylvisaker Conditions. In *The Annals of Applied Probability*, volume 5, pp. 518–540. Institute of Mathematical Statistics, 1995.

Seeger, M. W., Kakade, S. M., and Foster, D. P. Information Consistency of Nonparametric Gaussian Process Methods. In *IEEE Transactions on Information Theory*, volume 54, pp. 2376–2382. IEEE, 2008.

Titsias, M. K. Variational Inference for Gaussian and Determinantal Point Processes. In *Workshop on Advances in Variational Inference (NIPS)*, December 2014.