JMLR: Workshop and Conference Proceedings 1:1–9, 2018. Symposium on Advances in Approximate Bayesian Inference

Explicit Rates of Convergence for Sparse Variational Inference in Gaussian Process Regression

David Burt DRB62@CAM.AC.UK and **Carl Edward Rasmussen** CER54@CAM.AC.UK
University of Cambridge, Cambridge, UK

Mark van der Wilk MARK@PROWLER.IO
PROWLER.io, Cambridge, UK

Abstract

Sparse variational inference in Gaussian process regression has theoretical guarantees that make it robust to over-fitting. Additionally, it is well-known that in the non-sparse regime, with $M \geq N$, variational inducing point methods can be equivalent to full inference. In this paper, we derive bounds on the KL-divergence from the sparse variational approximation to the full posterior and show convergence for M on the order of $\log(N)$ inducing features when using the squared exponential kernel.

1. Introduction

Gaussian processes can model complex functions, are robust to over-fitting and provide uncertainty estimates. In the case of regression, the marginal likelihood and predictive posterior can be calculated in closed form. However, this becomes computationally intractable for large data sets. The variational framework by Titsias (2009) approximates the model posterior with a simpler Gaussian process. This approximation retains many of the desirable properties of non-parametric models, in contrast to parametric model approximations of Quiñonero Candela and Rasmussen (2005). A precise treatment of the KL-divergence between stochastic processes is given in Matthews et al. (2016). While the variational approximation provides an objective function (ELBO) for selecting the hyperparameters, it can introduce bias into the selection process (Turner and Sahani, 2011). However, if the gap between the ELBO and the log marginal likelihood at the optimum choice of hyperparameters is small, the optimization will result in a choice of hyperparameters similarly good to those learned using the marginal likelihood.

In this work, we introduce *eigenfunction inducing features*, an inter-domain feature in the style of Lázaro-Gredilla and Figueiras-Vidal (2009), based on the eigendecomposition of the kernel. We obtain bounds on the KL-divergence for sparse inference with these features and give theoretical insight into the number of features needed to approximate the posterior process. These features also result in a diagonal covariance that can lead to computational gains (Burt et al., 2018).

2. Bounds on the Marginal Likelihood

The log marginal likelihood for regression with a mean-centered Gaussian process prior with covariance function $k(\cdot, \cdot)$, used to regress $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, is given by (Rasmussen and

Williams, 2006):

$$\mathcal{L} = \log\left(p(\mathbf{y})\right), \tag{1}$$

where $\mathbf{y}$ is distributed according to $\mathcal{N}\left(\mathbf{y}; 0, \mathbf{K}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)$, σ^2_{noise} is the likelihood variance and $\mathbf{K}_{f,f}$ is the data covariance matrix, with $(\mathbf{K}_{f,f})_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$. Titsias (2009) provided an ELBO for (1) based on using a subset of "inducing features," $\{\mathbf{u}_m\}_{m=0}^{M-1}$.

$$\mathcal{L} \geq \mathcal{L}_{lower} = \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}\right)\right) - \frac{t}{2\sigma^2_{noise}}. \tag{2}$$

where $t = \mathrm{tr}\left(\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}\right)$, $\mathbf{Q}_{f,f} = \mathbf{K}_{u,f}^T\mathbf{K}_{u,u}^{-1}\mathbf{K}_{u,f}$, $(\mathbf{K}_{u,f})_{m,i} = \mathrm{cov}(\mathbf{u}_m, f(\mathbf{x}_i))$ and $(\mathbf{K}_{u,u})_{m,n} = \mathrm{cov}(\mathbf{u}_m, \mathbf{u}_n)$.

This ELBO is commonly used as a computationally tractable objective function for learning hyperparameters. Moreover, Matthews et al. (2016) showed that the gap between the log marginal likelihood and the evidence lower bound is the KL divergence between the posterior Gaussian process and the process used in approximate inference.

Instead of bounding $\mathcal{L} - \mathcal{L}_{lower}$ directly, we bound the gap between the ELBO and an upper bound on the marginal likelihood (Titsias, 2014):

$$\begin{aligned}\mathcal{L} \leq \mathcal{L}_{upper} = \log\left(\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_{f,f} + t\mathbf{I} + \sigma^2_{noise}\mathbf{I})\right) + \frac{1}{2}\log\left(|\mathbf{Q}_{f,f} + t\mathbf{I} + \sigma^2_{noise}\mathbf{I}|\right) \\ - \frac{1}{2}\log\left(|\mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}|\right).\end{aligned} \tag{3}$$

If $t = 0$, $\mathcal{L}_{lower} = \mathcal{L} = \mathcal{L}_{upper}$, so in order to obtain a bound on the KL-divergence, it suffices to bound t. Explicitly,

Lemma 1 *With the same notation as in (3),*

$$KL(Q\|\hat{P}) \leq \frac{t}{2\sigma^2_{noise}}\left(1 + \frac{\|\mathbf{y}\|_2^2}{\sigma^2_{noise} + t}\right). \tag{4}$$

where $\hat{P}$ is the full posterior process and Q is the variational approximation. Note that under mild conditions, $\|\mathbf{y}\|_2^2 = O(N)$.

The proof is provided in Appendix A. Explicitly writing out the diagonal entries in $\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}$ in order to bound t is difficult due to inverse matrix $\mathbf{K}_{u,u}^{-1}$ appearing in the definition of $\mathbf{Q}_{f,f}$.

3. Eigenvalues and Optimal Rate of Convergence

Before proving upper bounds on t, we consider a lower bound. As observed in Titsias (2014), t is the error in trace norm of a low-rank approximation to the covariance matrix, so

$$t \geq \sum_{m=M+1}^{N} \gamma_m \tag{5}$$

where γ_m denotes the m^{th} largest eigenvalue of $\mathbf{K}_{f,f}$.

Generally, computing eigenvalues and eigenvectors of a large matrix is costly. However, as observed in Williams and Seeger (2001), as $N \to \infty$ if the training points are i.i.d random variables drawn according to the probability measure ρ the eigenvalues of $\mathbf{K}_{f,f}$ converge to those of the operator

$$\mathcal{K}_\rho f \to \int_{\mathcal{X}} f(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') d\rho(\mathbf{x}). \tag{6}$$

3.1. Connection to the Nyström Approximation

In the case of inducing points subsampled from the data, the matrix $\mathbf{Q}_{f,f}$ is a Nyström approximation to the $\mathbf{K}_{f,f}$ (Williams and Seeger, 2001). Therefore, any bound on the rate of convergence of the Nyström approximation, for example (Gittens and Mahoney, 2013, Lemma 2), can be used in conjunction with Lemma 1 to yield a bound on the KL-divergence. While Gittens and Mahoney (2013, Lemma 2) yields similar asymptotic (in M) behavior to the bounds we prove, due to large constants in their proof, direct application of their result to this problem appears to require tens of thousands of points. Additionally, a proper analysis when M is allowed to grow as a function of N involves bounding rates of convergence of the kernel matrix spectrum to the operator spectrum.

Instead, we introduce *eigenfunction features*, defined with respect to $\mathcal{K}$ which yield elegant analytic upper bounds on t. These bounds asymptotically (in N) match the lower bound (5). Eigenfunction features are defined so as to be orthogonal, greatly simplifying the computation of t because $\mathbf{K}_{u,u} = \mathbf{I}$.

3.2. Eigenfunction Inducing Features

Mercer's theorem tells us that

$$k(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{\infty} \lambda_m \phi_m(\mathbf{x}) \phi_m(\mathbf{x}'), \tag{7}$$

where $(\lambda_m, \phi_m)_{m=0}^{\infty}$ are the eigenvalue-eigenfunction pairs of $\mathcal{K}_\rho$, which we assume are sorted so that $\lambda_m > \lambda_{m+1} > 0$ for all m. We define *eigenfunction inducing features* by,

$$\mathbf{u}_m = \int_{\mathcal{X}} f(\mathbf{x}) \frac{1}{\sqrt{\lambda_m}} \phi_m(\mathbf{x}) d\mu(\mathbf{x}).$$

where μ is a probability measure that can be treated as a variational parameter.
Using the orthogonality of eigenfunctions and the eigenfunction property it can be shown that,

$$\operatorname{cov}(\mathbf{u}_m, f(\mathbf{x})) = \sqrt{\lambda_m} \phi_m(\mathbf{x}) \quad \text{and} \quad \operatorname{cov}(\mathbf{u}_m, \mathbf{u}'_m) = \delta_{m,m'}. \tag{8}$$

3.3. An Example: Squared Exponential Kernel

The squared exponential kernel, defined for $\mathcal{X} = \mathbb{R}$, by $k_{se}(x, x') = v_k \left(-(x - x')^2/(2\ell^2)\right)$, is among the most popular choices of kernels. Suppose that the measure used in defining the $\mathcal{K}$ has normal density $\mathcal{N}(0, s^2)$. In this case the eigenvalues of $\mathcal{K}$ are given by (Rasmussen and Williams, 2006):

$$\lambda_m = v_k \sqrt{\frac{2a}{A}} B^m, \tag{9}$$

with $a = 1/(4s^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$, $B = b/A$.

From the above expression, when B is near one, we should expect to need more inducing features in order to bound the KL-divergence. This occurs when the observed data spans many kernel lengthscales. For any lengthscale, the eigenvalues decay exponentially. This asymptotic decay is closely connected to the smoothness of sample functions from the prior, see Rasmussen and Williams (2006, Chapters 4,7). It has previously been shown that truncating the kernel with M such that $\lambda_M \approx \sigma^2_{noise}/N$ eigenbasis functions leads to almost no loss in model performance (Ferrari-Trecate et al., 1999). We prove the following:

Theorem 2 *If the $x_i \sim \mathcal{N}(0, s'^2)$ i.i.d. For sparse inference with eigenfunction inducing features defined with respect to $q(x) \sim \mathcal{N}(0, s^2)$ with $2s'^2 < s^2$ There exists an $N_{s,s'}$ such that for all $N > N_{s,s'}$ inducing points inference with a set of $M = c_{s',s} \log(N)$ features results in:*

$$Pr(KL(Q\|\hat{P}) > \epsilon) < \delta.$$

While the full proof is in Appendix B, we suggest why this should be true in the following section.

4. Bounds on the Trace using Eigenfunction Features

Using the eigenfunction features, deriving a bound on t is straightforward. From (8),

$$\left(\mathbf{K}_{u,f}^T \mathbf{K}_{u,f}\right)_{i,j} = \sum_{m=0}^{M-1} \lambda_m \phi_m(\mathbf{x}_i)\phi_m(\mathbf{x}_j).$$

Combining this with Mercer's theorem yields the following lemma:

Lemma 3 *Suppose the training data consists of N i.i.d. random variables drawn according to probability density $p(x)$ and we use M eigenfunction inducing features defined with respect to a density $q(x)$ such that $k = \max_x \frac{p(x)}{q(x)} < \infty$, then*

$$\lim_{N\to\infty} \frac{1}{N} tr(\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}) = \left(\sum_{m=M}^{\infty} \lambda_m \mathbb{E}_p[\phi_m(\mathbf{x}_i)^2]\right) \leq k \sum_{m=M}^{\infty} \lambda_m. \tag{10}$$

For the squared exponential kernel, (10) is a geometric series, and has a closed form (see Appendix B). An example of the bound on the expected value of the trace and the resulting bound on the KL-divergence is shown in Figure 1. Under the assumed input distribution in Lemma 3 as N tends to infinity for fixed M, the empirical eigenvalues in (5) approach the λ_m so the expected value of $\frac{1}{N}t$ is asymptotically tight. In order to rigorously prove Theorem 2 we need to understand the trace, not the expected value of its entries. In order to achieve this, as well as to obtain bounds that are effective as M grows as a function of N, we show that the square of the eigenfunctions has bounded second moment under $q(x)$. The details given in Appendix B.

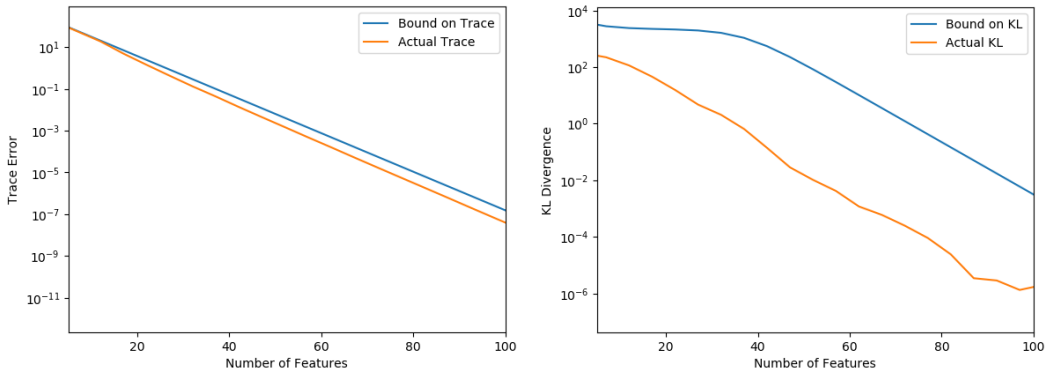

Figure 1: Log-linear plot of bounds (blue) on the trace (left) and KL-divergence (right) plotted against the actual values (yellow) for a synthetic data set with $N = 200$, $x \sim \mathcal{N}(0, 5^2), v_k = 1, \ell^2 = 1$.

References

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r$^{\text{th}}$ Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1): 299–303, 1965. ISSN 00034851.

David Burt, Carl Edward Rasmussen, and Mark van der Wilk. Sparse Variational Gaussian Process Inference with Eigenfunction Inducing Features. Submitted, October 2018.

Pafnutii Lvovich Chebyshev. Des valeurs moyennes, Liouville's. *J. Math. Pures Appl.*, 12: 177–184, 1867.

Giancarlo Ferrari-Trecate, Christopher KI Williams, and Manfred Opper. Finite-dimensional approximation of Gaussian processes. In *Advances in neural information processing systems*, pages 218–224, 1999.

Alex Gittens and Michael Mahoney. Revisiting the Nyström method for improved large-scale machine learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA, June 2013. PMLR.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

Miguel Lázaro-Gredilla and Anbal Figueiras-Vidal. Inter-domain Gaussian Processes for Sparse Inference using Inducing Features. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1087–1095. Curran Associates, Inc., 2009.

Alexander G de G Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec): 1939–1959, 2005.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

Michalis K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009.

Michalis K. Titsias. *Variational Inference for Gaussian and Determinantal Point Processes*. December 2014. Published: Workshop on Advances in Variational Inference (NIPS 2014).

Richard Eric Turner and Maneesh Sahani. *Two problems with variational expectation maximisation for time series models*, pages 104–124. Cambridge University Press, 2011.

Christopher K. I. Williams and Matthias Seeger. Using the Nyström Method to Speed Up Kernel Machines. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

Appendix A. Proof of Lemma 1

Proof The proof has a similar spirit to that of (3) provided in Titsias (2014). Let $\mathbf{R} = \mathbf{Q}_{f,f} + \sigma^2_{noise}\mathbf{I}$.

$$\begin{aligned}\mathcal{L}_{upper} - \mathcal{L}_{lower} &= \frac{t}{2\sigma^2_{noise}} + \frac{1}{2}\left(\mathbf{y}^T\mathbf{R}^{-1}\mathbf{y} - \mathbf{y}^T(\mathbf{R}+t\mathbf{I})^{-1}\mathbf{y}\right) \\ &= \frac{t}{2\sigma^2_{noise}} + \frac{1}{2}\left(\mathbf{y}^T\left(\mathbf{R}^{-1} - (\mathbf{R}+t\mathbf{I})^{-1}\right)\mathbf{y}\right). \qquad (11)\end{aligned}$$

Since $\mathbf{Q}_{f,f}$ is symmetric positive semidefinite, $\mathbf{R}$ is positive definite with eigenvalues bounded below by σ^2_{noise}. Write, $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is unitary and $\mathbf{\Lambda}$ is a diagonal matrix with non-increasing diagonal entries $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_N \geq \sigma^2_{noise}$.

We can rewrite the second term (ignoring the factor of one half) in (11) as,

$$\mathbf{y}^T\left(\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T - \mathbf{U}(\mathbf{\Lambda}+t\mathbf{I})^{-1}\mathbf{U}^T\right)\mathbf{y} = (\mathbf{U}^T\mathbf{y})^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)(\mathbf{U}^T\mathbf{y}).$$

Now define, $\mathbf{z} = (\mathbf{U}^T\mathbf{y})$. Since $\mathbf{U}$ is unitary, $\|\mathbf{z}\| = \|\mathbf{y}\|$.

$$\begin{aligned}(\mathbf{U}^T\mathbf{y})^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)(\mathbf{U}^T\mathbf{y}) &= \mathbf{z}^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)\mathbf{z} \\ &= \sum_i z_i^2 \frac{t}{\gamma_i^2 + \gamma_i t} \\ &\leq \|\mathbf{y}\|^2 \frac{t}{\gamma_N^2 + \gamma_N t}. \qquad (12)\end{aligned}$$

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

$$\mathbf{y}^T\left(\mathbf{R}^{-1} - (\mathbf{R}+t\mathbf{I})^{-1}\right)\mathbf{y} \leq \frac{t\|\mathbf{y}\|^2}{\sigma^4_{noise} + \sigma^2_{noise}t},$$

from which Lemma 1 follows. ■

Appendix B. Proof of Theorem

For inference with a squared exponential kernel and M eigenfunction inducing features defined with respect to $\mu \sim \mathcal{N}(0, s^2)$ a stronger version of Lemma 3 is the following:

Theorem 4 *Suppose $x_i \sim \mathcal{N}(0, s'^2)$ with $s' \leq s$ then for all $M \geq 0$,*

$$\lim_{n\to\infty} \frac{1}{N} t \leq k\lambda_0 \frac{B^M}{1-B} \tag{13}$$

for $k = \frac{s}{s'}$ holds almost surely for all $M \geq 0$. For $2s'^2 < s^2$, for any $\alpha > 0$

$$P\left(t > KNB^M\right) \leq \frac{1}{\alpha^2 N}\left(\sqrt{\frac{s^2}{s^2-2s'^2}} - \frac{s^2}{s^2-s'^2}\right) \tag{14}$$

where

$$K := 1.19(4cs^2)^{1/2}\frac{v_k}{1-B}\sqrt{\frac{2a}{A}}\left(\alpha + \sqrt{\frac{s^2}{s^2-s'^2}}\right).$$

Remark 5 *This proof could likely be generalized for all $s' \leq s$ via using a concentration inequality based on the r^{th} moment for $1 < r \leq 2$, for example those in Bahr and Esseen (1965), in place of Chebyshev's inequality.*

The proof of the first part of the statement, (13), comes from the observation that if q is the density associated to μ and p is the density the x_i are drawn from $\max_x \frac{p(x)}{q(x)} = \frac{p(0)}{q(0)} = k$. Note that here it is essential $s' \leq s$ or else this ratio would not be bounded as $x \to \infty$. Then,

$$\begin{aligned}\frac{1}{N}t &= \frac{1}{N}\sum_{i=1}^{N}\sum_{m=M}^{\infty}\lambda_m\phi_m^2(x_i) \\ &= \sum_{m=M}^{\infty}\lambda_m\frac{1}{N}\sum_{i=1}^{N}\phi_m^2(x_i).\end{aligned} \tag{15}$$

The inner sum is an expectation (with respect to p) of i.i.d. random variables. By the strong law of large numbers,

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\phi_m^2(x_i) &\xrightarrow{a.s.} E_p[\phi_m^2(x_i)] = \int |\phi_m(x)^2|p(x)dx \\ &\leq k\int |\phi_m(x)^2|q(x)dx \\ &= k.\end{aligned} \tag{16}$$

This shows (13) holds as N tends to ∞ for any fixed M. As the countable intersection of events that occur with probability one also occurs with probability one, this holds for all M (simultaneously) almost surely.

For the probabilistic bound, we begin by establishing a term-wise bound on elements of $\mathbf{K}_{f,f} - \mathbf{Q}_{f,f}$ that takes into account the locations of the training inputs.

Lemma 6 *Let $q_{i,j}$ denote the i,j^{th} entry in $\mathbf{Q}_{n,n}$ and $k_{i,j} = k(x_i, x_j)$ denote the i,j^{th} entry in $\mathbf{K}_{n,n}$, then*

$$|k_{i,j} - q_{i,j}| \leq 1.19\sqrt{4cs^2}v_k\sqrt{\frac{2a}{A}}\frac{B^M}{1-B}\exp\left(\frac{x_i^2 + x_j^2}{4s^2}\right). \tag{17}$$

holds for all pairs x_i, x_j.

Proof As in earlier proofs, we have

$$|k_{i,j} - q_{i,j}| = \Big|\sum_{m=M}^{\infty} \lambda_m \phi_m(x_i)\phi_m(x_j)\Big|,$$

We need to now take into account the location of the x_i in this bound. For the squared exponential kernel,

$$\phi_m(x) = \frac{(4cs)^{1/4}}{\sqrt{m!}2^m} H_m(\sqrt{2c}x)\exp(-(c-a)x^2),$$

where $H_m(x)$ is the m^{th} Hermite polynomial, (Rasmussen and Williams, 2006), (we have normalized the basis so $\|\phi_m\|_{L^2(\mu)} = 1$).

We use the following bound on Hermite functions, obtained by squaring the bound in Gradshteyn and Ryzhik (2014),

$$|H_m(x_i)||H_m(x_j)| < 1.19m!2^m e^{(x_i^2+x_j^2)/2}.$$

Expanding into the definition of the ϕ_m we obtain

$$\begin{aligned}
|k_{i,j} - q_{i,j}| &= \sqrt{4cs^2}\exp(a(x_i^2 + x_j^2))\left|\sum_{m=M}^{\infty}\lambda_m \frac{H_m(\sqrt{2c}x_j)e^{-cx_j^2}H_m(\sqrt{2c}x_i)e^{-cx_i^2}}{2^m m!}\right| \\
&\leq \sqrt{4cs^2}\exp(a(x_i^2 + x_j^2))\sum_{m=M}^{\infty}\frac{\lambda_m\left|H_m(\sqrt{2c}x_j)e^{-cx_j^2}\right|\left|H_m(\sqrt{2c}x_i)e^{-cx_i^2}\right|}{2^m m!} \\
&\leq 1.19\sqrt{4cs^2}\exp(a(x_i^2 + x_j^2))\sum_{m=M}^{\infty}\lambda_m \\
&= 1.19\sqrt{4cs^2}\exp\left(\frac{x_i^2 + x_j^2}{4s^2}\right)v_k\sqrt{\frac{2a}{A}}\frac{B^M}{1-B}.
\end{aligned}$$

The first inequality (triangle inequality) is sharp on the terms effecting the trace since when $x_i = x_j$ the sum must be term-wise positive. ■

Let $A_i = \exp\left(\frac{x_i^2}{2s^2}\right)$. It remains to derive a probabilistic bound on $S := \sum_{i=1}^{N} A_i$. We first compute the moments of the A_i under the input distribution.

$$\begin{aligned}\mathbb{E}[A_i^r] &= \frac{1}{\sqrt{2\pi s'^2}} \int_{\mathbb{R}} \exp\left(\frac{rx^2}{2s^2}\right) \exp\left(\frac{-x^2}{2s'^2}\right) dx \\ &= \sqrt{\frac{s^2}{s^2 - s'^2 r}}, \end{aligned} \tag{18}$$

for $rs'^2 < s^2$.

From this we deduce $\mathrm{Var}(A_i) = \mathbb{E}[A_i^2] - \mathbb{E}[A_i]^2 = \sqrt{\frac{s^2}{s^2-2s'^2}} - \frac{s^2}{s^2-s'^2}$. Chebyshev's inequality Chebyshev (1867) tells us that for any $\alpha > 0$,

$$P\left(S > N\left(\alpha + \sqrt{\frac{s^2}{s^2 - s'^2}}\right)\right) \leq \frac{Var(A_1)}{\alpha^2 N}. \tag{19}$$

Theorem 4 follows.

In order to prove Theorem 2 choose $M \gg 3\log(N) = \log(N^3)$ then $KNB^M \ll \frac{1}{N^2}$. For large N, we have that this is an upper bound on the trace with high probability. Using this upper bound in Lemma 1 gives Theorem 2 as a corollary.