Information Consistency of Nonparametric Gaussian Process Methods

Matthias W. Seeger, Sham M. Kakade, and Dean P. Foster

***Abstract*—Bayesian nonparametric models are widely and successfully used for statistical prediction. While posterior consistency properties are well studied in quite general settings, results have been proved using abstract concepts such as metric entropy, and they come with subtle conditions which are hard to validate and not intuitive when applied to concrete models. Furthermore, convergence rates are difficult to obtain.**

By focussing on the concept of information consistency for Bayesian Gaussian process (GP)models, consistency results and convergence rates are obtained via a regret bound on cumulative log loss. These results depend strongly on the covariance function of the prior process, thereby giving a novel interpretation to penalization with reproducing kernel Hilbert space norms and to commonly used covariance function classes and their parameters. The proof of the main result employs elementary convexity arguments only. A theorem of Widom is used in order to obtain precise convergence rates for several covariance functions widely used in practice.

***Index Terms*—Bayesian prediction, eigenvalue asymptotics, Gaussian process, information consistency, nonparametric statistics, online learning, posterior consistency, regret bound.**

I. Introduction

In this correspondence, we are interested in methods predicting a response $y \in \mathcal{Y}$ from a covariate $\boldsymbol{x} \in \mathcal{X}$. Given some class of functions $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ and a likelihood conditional distribution $P(y|f(\boldsymbol{x}))$ over $\mathcal{Y}$, we assume that data $y_1, \ldots, y_n$, given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, is generated by nature picking f, then sampling $y_i \sim P(\cdot|f(\boldsymbol{x}_i))$ independently.[1] Note that covariates are by definition always given at prediction time, and in the sequel all distributions are implicitly conditional on all necessary covariate instances. We assume that the covariates are independently drawn from a distribution $d\mu(\boldsymbol{x})$, which will not be modeled.

The prediction task may be of batch nature, i.e., given some training data $\{(\boldsymbol{x}_i, y_i)|i = 1, \ldots, n\}$, predict y_{n+1} for unseen $\boldsymbol{x}_{n+1}$, or of sequential nature, i.e., predict y_i, given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i$ and $y_1, \ldots, y_{i-1}$, respectively, for $i = 1, \ldots, n$. The Bayesian prediction strategy is the same in both situations. Initial assumptions about nature's choice are encoded in a prior distribution $P_{bs}(f)$ over $\mathcal{F}$. This distribution is conditioned on observed data in order to obtain the posterior distribution

$$dP_{bs}(f|y_1, \ldots, y_n) = \frac{\left[\prod_{i=1}^n P(y_i|f(\boldsymbol{x}_i))\right] dP_{bs}(f)}{\int \left[\prod_{i=1}^n P(y_i|f'(\boldsymbol{x}_i))\right] dP_{bs}(f')}$$

Manuscript received August 6, 2006; revised November 17, 2007. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

M. W. Seeger is with the Max Planck Institute for Biological Cybernetics, P. O. Box 21 69, D-72012 Tübingen, Germany (e-mail: seeger@tuebingen.mpg.de).

S. M. Kakade is with the Toyota Technological Institute, TTI-C, University Press Building, Chicago, IL 60637 USA (e-mail: sham@tti-c.org).

D. P. Foster is with the Department of Statistics, Wharton School of Economics, University of Pennsylvania, Philadelphia, PA 19104-6340 USA (e-mail: dean@foster.net).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2007.915707

[1]In some settings, it is advisable to parameterize nature's choice by more than one real-valued function. While our results can be extended to this case straightforwardly, we focus on single-function models for simplicity.

from which the *predictive distribution* is obtained as

$$P_{bs}(y_{n+1}|\boldsymbol{y}_{\leq n}) = \int P(y_{n+1}|f(\boldsymbol{x}_{n+1}))dP_{bs}(f|\boldsymbol{y}_{\leq n})$$

thus, as expectation of the likelihood with respect to (w.r.t.) the posterior. Note that this strategy has strong practical and theoretical merits, even if nature does *not* choose f according to P_{bs}. Barron's work [3] can be understood as trying to characterize Bayesian prediction performance depending on the prior specification, assuming that the true likelihood is known, but making minimal or no assumptions about nature's true choice of f.

An intuitive way to information consistency goes via sequential prediction. Let $\boldsymbol{y}_{\leq i} = \{y_1, \ldots, y_i\}$, and $\boldsymbol{x}_{\leq i}$ accordingly. An expert prediction strategy parameterized by $f \in \mathcal{F}$ is

$$P(\boldsymbol{y}_{\leq n}|f) = \prod_{i=1}^n P(y_i|f(\boldsymbol{x}_i)).$$

An expert predicts $P(y_i|f(\boldsymbol{x}_i))$ independently for each unseen point, using a fixed function. The Bayesian prediction strategy is *mixing* over experts, in the sequential case by using the predictive distributions $P_{bs}(y_i|\boldsymbol{y}_{<i})$, so the mixing distribution is always given by the posterior for all observed data. Now, suppose that a prediction strategy, outputting $Q(\cdot)$ in order to predict y_i, incurs the log loss $-\log Q(y_i)$ for each single prediction, and the cumulative log loss overall

$$L_Q(\boldsymbol{y}_{\leq n}) = \sum_{i=1}^n -\log Q(y_i|\boldsymbol{y}_{<i}).$$

For an expert f, the cumulative log loss is

$$L_f(\boldsymbol{y}_{\leq n}) = -\log P(\boldsymbol{y}_{\leq n}|f)$$

while for the Bayesian strategy we have that

$$L_{bs}(\boldsymbol{y}_{\leq n}) = -\log P_{bs}(\boldsymbol{y}_{\leq n})$$

by the chain rule. The Bayesian strategy has been analyzed under the log loss setting by several researchers [7], [8].

Let Q be a prediction strategy, and let $\mathcal{F}_{\text{comp}} \subset \mathcal{F}$ be a *competitor space*. Barron [3] calls Q *information consistent* over $\mathcal{F}_{\text{comp}}$ iff

$$\frac{\mathrm{E}_{\boldsymbol{x}_{\leq n}}\left[\mathrm{D}\left[P(\boldsymbol{y}_{\leq n}|f, \boldsymbol{x}_{\leq n}) \parallel Q(\boldsymbol{y}_{\leq n}|\boldsymbol{x}_{\leq n})\right]\right]}{n} \to 0 \quad (n \to \infty) \quad (1)$$

for all $f \in \mathcal{F}_{\text{comp}}$. Here, the expectation is over $\boldsymbol{x}_{\leq n} \sim \mu^n$, and $\mathrm{D}[P_1 \parallel P_2] = \int (\log P_1 - \log P_2) dP_1$ is the *relative entropy* (or *Kullback–Leibler divergence*). Note that

$$n^{-1}\mathrm{E}[\mathrm{D}[P(\boldsymbol{y}_{\leq n}|f) \parallel Q(\boldsymbol{y}_{\leq n})]] = n^{-1}\sum_{i=1}^n \mathrm{E}[\mathrm{D}[P(y_i|f(\boldsymbol{x}_i)) \parallel Q(y_i|\boldsymbol{y}_{<i})]]$$

which is a type of Cesaro average risk. Information consistency seems a fairly weak mode of consistency, but Barron [3] argues that some stronger notions do have shortcomings which are unintuitive at the least. For example, while the average of the left-hand side of (1) over $f \sim P_{bs}$ is nonincreasing, the individual Kullback risk $\mathrm{D}[P(y_i|f(\boldsymbol{x}_i)) \parallel P_{bs}(y_i|\boldsymbol{y}_{<i})]$ can increase for some i, even if $P_{bs}(f)$ is large. And in order to ensure that for any f, the posterior $P(f|\boldsymbol{y}_{\leq n})$ concentrates on arbitrary small neighborhoods of f (w.r.t. Hellinger, Kullback, or some other metric) [4], unintuitive global conditions on P_{bs} are required (Barron [3] gives an example of posterior inconsistency, where $P_{bs}(\mathcal{F}_{bad}) = 1/2$, but $P_{bs}(\mathcal{F}_{bad}|\boldsymbol{y}_{\leq n}) \to 1$ almost

0018-9448/$25.00 © 2008 IEEE

surely, and $d(f, f') = 1$ for all $f' \in \mathcal{F}_{bad}$, data coming from f). We focus on information consistency in what follows.

In order to relate information consistency to sequential prediction under cumulative log loss, note that

$$\mathrm{D}\left[P(\boldsymbol{y}_{\le n}|f) \parallel P_{bs}(\boldsymbol{y}_{\le n})\right] = \mathrm{E}\left[L_{bs}(\boldsymbol{y}_{\le n}) - L_f(\boldsymbol{y}_{\le n})\right]$$

where the expectation is over $\boldsymbol{y}_{\le n} \sim P(\cdot|f)$. If we can bound $L_{bs}(\boldsymbol{y}_{\le n}) - L_f(\boldsymbol{y}_{\le n})$ uniformly over all $\boldsymbol{y}_{\le n}$, and for all $f \in \mathcal{F}_{\mathrm{comp}}$, this implies information consistency and convergence rate bounds.

Our main result can be stated as follows. Consider a Bayesian Gaussian process (GP) prediction strategy P_{bs}, where the prior distribution $P_{bs}(f)$ is a zero-mean GP with covariance function $K(\boldsymbol{x}, \boldsymbol{x}')$, and let $\mathcal{H}$ be the reproducing kernel Hilbert space determined by K, having norm $\|f\|_K$. Furthermore, let the curvature of $-\log P(y|f(\boldsymbol{x}))$ w.r.t. $f(\boldsymbol{x})$ be bounded by $c > 0$ for any $y \in \mathcal{Y}$. We show that

$$\mathrm{D}\left[P(\boldsymbol{y}_{\le n}|f) \parallel P_{bs}(\boldsymbol{y}_{\le n})\right] \le \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\log|\boldsymbol{I} + c\boldsymbol{K}|$$

for any $f \in \mathcal{H}$, where $\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j} \in \mathbb{R}^{n,n}$ is the covariance matrix depending on K and $\boldsymbol{x}_{\le n}$. Therefore, Bayesian GP prediction is information consistent w.r.t. $\mathcal{H}$ if $n^{-1}\mathrm{E}[\log|\boldsymbol{I}+c\boldsymbol{K}|] \to 0$ $(n \to \infty)$, where the latter criterion depends on the covariance function K and the covariate distribution μ only. We give a range of examples for practically relevant covariance functions and restrictions on μ, for which information consistency and convergence rates can be established along this path, namely, by analyzing the term $\mathrm{E}[\log|\boldsymbol{I} + c\boldsymbol{K}|]$ asymptotically as $n \to \infty$. To this end, we utilize the Mercer eigenexpansion of the covariance function K w.r.t. the measure $d\mu$, and a powerful theorem by Widom [6] in order to obtain asymptotic expressions for the eigenvalues. To the best of our knowledge, our approach to obtain sharp information convergence rates for GP nonparametric prediction methods is novel. The regret term $n^{-1}\mathrm{E}[\log|\boldsymbol{I} + c\boldsymbol{K}|]$ and also our bounds for common kernel classes depend explicitly on parameters of K and μ, thereby giving new characterizations of these regularization parameters in terms of convergence rates.

In Section II, we state our main result, a regret bound for cumulative log loss of Bayesian GP prediction. In Section III, we develop tools in order to bound the expected regret featuring in our result. These tools are applied to several classes of covariance functions frequently used in practice in Section IV. Conclusions are given in Section V, and the Appendices contain details of proofs.

II. Main Result

A *Gaussian process* (GP) model is defined on the space $\mathcal{F}$ of continuous functions $\mathcal{X} \to \mathbb{R}$. A zero mean GP is a random function $f \in \mathcal{F}$ with $\mathrm{E}[f(\boldsymbol{x})] = 0$ and $\mathrm{E}[f(\boldsymbol{x})f(\boldsymbol{x}')] = K(\boldsymbol{x}, \boldsymbol{x}')$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. GPs have the property that all associated finite-dimensional distributions are Gaussian again. Namely, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be arbitrary, and consider the random vector $\boldsymbol{f} = (f(\boldsymbol{x}_i))_i \in \mathbb{R}^k$. Then, $\boldsymbol{f}$ has a multivariate Gaussian distribution with mean $\boldsymbol{0}$ and covariance matrix $(K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j} \in \mathbb{R}^{k,k}$. For details on GPs in Machine Learning, see [9], [10]. GP models form a major class of nonparametric methods which are routinely used for spatial statistics applications in geostatistics and remote sensing [11]. Bayesian GP prediction has been pioneered by O'Hagan [12], and has been applied to many problems in Machine Learning. We note that while Bayesian GP prediction is analytically tractable only for a Gaussian likelihood, Markov chain Monte Carlo techniques may be used to sample from the posterior, or one of several variational approximation techniques proposed in Machine Learning may be applied.

The covariance function K of a GP is a positive semi-definite form, in that all induced covariance matrices $\boldsymbol{K}$ are always positive semi-definite: $\boldsymbol{v}^T\boldsymbol{K}\boldsymbol{v} \ge 0$ for all vectors $\boldsymbol{v}$. A reproducing kernel Hilbert space (RKHS) [13], [14] of functions $\mathcal{X} \to \mathbb{R}$ is associated with K as follows. Consider the linear space of all finite kernel expansions (over any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$) of the form $f(\cdot) = \sum_{i=1}^n \alpha_i K(\cdot, \boldsymbol{x}_i)$, with the inner product

$$\left(\sum_i \alpha_i K(\cdot, \boldsymbol{x}_i), \sum_j \beta_j K(\cdot, \boldsymbol{x}'_j)\right)_K = \sum_{i,j} \alpha_i \beta_j K(\boldsymbol{x}_i, \boldsymbol{x}'_j).$$

The RKHS $\mathcal{H}$ is the completion of this space. By construction, $\mathcal{H}$ contains all finite kernel expansions $f(\cdot) = \sum_{i=1}^n \alpha_i K(\cdot, \boldsymbol{x}_i)$ with

$$\|f\|_K^2 = \boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}, \qquad \boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{2}$$

The characteristic property of $\mathcal{H}$ is that all (Dirac) evaluation functionals are *represented* in $\mathcal{H}$ itself by the functions $K(\cdot, \boldsymbol{x}_i)$, meaning that $(f, K(\cdot, \boldsymbol{x}_i))_K = f(\boldsymbol{x}_i)$. This *reproducing property* means that convergence in norm in $\mathcal{H}$ implies pointwise convergence, so all $f \in \mathcal{H}$ are pointwise defined. Intuitively, $\mathcal{H}$ is a space within $L_2(\mathcal{X})$ of reasonably well-behaved functions. In general, it is the case that functions of larger RKHS norm show a rougher and more irregular behavior, and $\|f\|_K^2$ is commonly used as smoothness penalty. The RKHS $\mathcal{H}$ turns out to be the largest competitor space of experts for which our results are meaningful. We note that for most kernels used in practice, and in fact for all infinite-dimensional kernels mentioned here, $\mathcal{H}$ is dense in the space of continuous functions restricted to a compact domain in $\mathcal{X}$. Also note that the "complexity" $\|f\|_K$ assigned to a function f depends on characteristics of K, and our results render a new interpretation for this dependency.

Theorem 1 (Main Result): Let P_{bs} be the Bayesian GP prediction method, configured by a zero-mean Gaussian process prior with covariance function K. Let $(\boldsymbol{x}_{\le n}, \boldsymbol{y}_{\le n})$ be a sequence from $(\mathcal{X} \times \mathcal{Y})^n$ and f be a function from the RKHS $\mathcal{H}$ associated with K. Then

$$-\log P_{bs}(\boldsymbol{y}_{\le n}) \le -\log P(\boldsymbol{y}_{\le n}|f) + \frac{1}{2}\|f\|_K^2 + \frac{1}{2}\log|\boldsymbol{I} + c\boldsymbol{K}| \tag{3}$$

where $\|f\|_K$ is the RKHS norm of f, $\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j} \in \mathbb{R}^{n,n}$ is the covariance matrix over the input sequence $\boldsymbol{x}_{\le n}$, and $c > 0$ is a constant such that for all $y \in \mathcal{Y}$, $f(\boldsymbol{x}) \in \mathbb{R}$:

$$\frac{d^2}{df(\boldsymbol{x})^2} - \log P(y|f(\boldsymbol{x})) \le c.$$

For the Gaussian likelihood $P(y|f(\boldsymbol{x})) = N(y|f(\boldsymbol{x}), \sigma^2)$, (3) is attained as equality with $c = \sigma^{-2}$ for a function of the form $f(\cdot) = \sum_{i=1}^n \alpha_i K(\cdot, \boldsymbol{x}_i)$.

This theorem has appeared in [15], using earlier work on parametric models [16]. A proof is given in Appendix I. The bound depends on $\|f\|_K^2$, which states the intuitive fact that a meaningful bound can only be obtained under smoothness assumptions on the set of experts (note that the bound is nonasymptotic and holds for any finite n). The constant c, which bounds the curvature of the log likelihood, exists for most commonly used exponential family likelihoods. For logistic regression, we have $c = 1/4$, and for Gaussian regression, we have $c = \sigma^{-2}$, where σ^2 is the noise variance.

Returning to our introduction to information consistency in Section I, we see that we have to analyze the term $\mathrm{E}[\log|\boldsymbol{I} + c\boldsymbol{K}|]$, which depends on K and the covariate distribution μ. We call $R = \log|\boldsymbol{I} + c\boldsymbol{K}|$ the *regret term*. In Section III, we will provide a thorough analysis of the (expected) regret term, obtaining tight information consistency results for several practically relevant settings.

Note that Theorem 1 is a statement which holds for every fixed $f \in \mathcal{H}$, and the right-hand side depends on f through $\|f\|_K^2$. This is

different from learning curve analyses, where f is assumed random according to a fixed prior, typically just the one that the Bayesian method is using. For example, if the likelihood is $N(y_i|f(\boldsymbol{x}_i),\sigma^2)$ and f is a zero mean GP with kernel K, a simple direct calculation shows that

$$\mathrm{E}_f\left[\mathrm{D}[P(\boldsymbol{y}_{\le n}|f)\parallel P(\boldsymbol{y}_{\le n})]\right]=\frac{1}{2}\log|\boldsymbol{I}+\sigma^{-2}\boldsymbol{K}|$$

so that R controls the learning curve directly [17]. Our analysis is more general, in that we do not assume that f is drawn from a simple, known prior distribution. On the other hand, our result restricts f to lie in $\mathcal{H}$, which in fact is a null set under the GP prior [14]. If $\|\cdot\|_K$ is formally defined over all functions in $L_2(\mathcal{X})$ (with $\|f\|_K=\infty$ for $f\notin\mathcal{H}$), then $\mathrm{E}_f[\|f\|_K]=\infty$ for GP sample paths.

III. Analysis of the Regret Term

Theorem 1 provides a regret bound for Bayesian GP prediction, competing against experts from the RKHS associated with the covariance function K of the GP. The bound depends on the squared RKHS norm $\|f\|_K^2$, where f is the competitor function, and on the regret term $R=\log|\boldsymbol{I}+c\boldsymbol{K}|$, the latter depending on K and the covariates $\boldsymbol{x}_{\le n}$. In this section, we collect some tools from spectral analysis which will be used to obtain bounds on $\mathrm{E}[R]$ under assumptions on K and the covariate distribution μ, thereby obtaining information consistency results via Theorem 1.

It is clear that with no further assumption, the regret term can always be made as large as $\Omega(n)$, rendering our result trivial. For example, for an isotropic covariance function $K(\boldsymbol{x},\boldsymbol{x}')=K(\|\boldsymbol{x}-\boldsymbol{x}'\|)$ and $K(r)\to 0$ $(r\to\infty)$, we can choose all $\boldsymbol{x}_i$ to be very far from each other, equivalently μ to have very heavy tails, so that $\boldsymbol{K}\propto\boldsymbol{I}$ for all n. In such extreme cases, the smoothness constraint on f through the requirement of a small $\|f\|_K^2$ term does not imply any strong constraints on the function values $f(\boldsymbol{x}_i)$, so that even a set of smooth competitors can represent any $\boldsymbol{y}_{\le n}$ very well. Our main result implies that R has to be large in such cases. In the remainder of this correspondence, we are interested in more reasonable cases, in which useful instances of our main result can be obtained.

Suppose K is continuous and Hilbert–Schmidt in $L_2(\mu)$. Note that we choose the covariate distribution μ as base measure in what follows. The spectrum of the linear operator with kernel K is discrete and nonnegative [14]

$$K(\boldsymbol{x},\boldsymbol{x}')=\sum_{s\ge 0}\lambda_s\phi_s(\boldsymbol{x})\phi_s(\boldsymbol{x}'). \tag{4}$$

Here, $\{(\lambda_s,\phi_s)|s\ge 0\}$ is a complete orthonormal eigensystem of K in $L_2(\mu)$ with $\lambda_0\ge\lambda_1\ge\cdots\ge 0$, and $\mathrm{E}[\phi_s(\boldsymbol{x})\phi_t(\boldsymbol{x})]=\delta_{s,t}$. The Hilbert–Schmidt assumption implies that $\sum_s\lambda_s^2<\infty$, so λ_s decays rapidly to 0, and the series expansion of K converges uniformly.

Lemma 1: Suppose K has an eigenexpansion (4). Then

$$R=\log|\boldsymbol{I}+c\boldsymbol{K}|\le\sum_{s\ge 0}\log\left(1+c\lambda_s\sum_{i=1}^{n}\phi_s(\boldsymbol{x}_i)^2\right).$$

Moreover, suppose that $\boldsymbol{x}_{\le n}$ are drawn from a distribution such that the marginal distribution of each component $\boldsymbol{x}_i$ is μ. Then, the *expected* regret is bounded as follows:

$$\mathrm{E}[R]\le\sum_{s\ge 0}\log\left(1+c\lambda_s n\right).$$

Proof: Let $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_s)_s$, $\boldsymbol{\Phi}=(\phi_s(\boldsymbol{x}_i))_{i,s}$, so that $\boldsymbol{K}=\lim_{S\to\infty}\boldsymbol{\Phi}_{\cdot,\le S}\boldsymbol{\Lambda}_{\le S}\boldsymbol{\Phi}_{\cdot,\le S}^T$ uniformly over $\boldsymbol{x}_{\le n}$, where "$\le S$" is short for $\{1,\ldots,S\}$ (and $S\ge n$). By continuity of $\log|\cdot|$, we have that

$$\log|\boldsymbol{I}+c\boldsymbol{K}|=\lim_{S\to\infty}\log\left|\boldsymbol{I}+c\boldsymbol{\Lambda}_{\le S}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}\right|. \tag{5}$$

The last term is equal to $\log\left|\boldsymbol{I}+c\boldsymbol{\Lambda}_{\le S}^{1/2}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}\boldsymbol{\Lambda}_{\le S}^{1/2}\right|$. The first statement follows by Hadamard's inequality (which states that $\log|\boldsymbol{M}|\le\log|\mathrm{diag}\boldsymbol{M}|$ for positive semi-definite $\boldsymbol{M}$).

Consider the eigenexpansion (4) of K with respect to μ. We have $\mathrm{E}\left[n^{-1}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}\right]=\boldsymbol{I}$ by the orthonormality of the eigenfunctions. Using (5) and the concavity of $\boldsymbol{A}\mapsto\log|\boldsymbol{I}+\boldsymbol{A}|$, we have

$$\begin{aligned}\mathrm{E}[\log|\boldsymbol{I}+c\boldsymbol{K}|]&=\lim_{S\to\infty}\mathrm{E}\left[\log\left|\boldsymbol{I}+c\boldsymbol{\Lambda}_{\le S}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}\right|\right]\\&\le\lim_{S\to\infty}\log\left|\boldsymbol{I}+cn\boldsymbol{\Lambda}_{\le S}\mathrm{E}\left[n^{-1}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}\right]\right|\\&=\sum_{s\ge 0}\log\left(1+c\lambda_s n\right)\end{aligned} \tag{6}$$

by Jensen's inequality. In the first equality, we use Lebesgue's monotone convergence theorem, noting that $|\boldsymbol{I}+c\boldsymbol{\Lambda}_{\le S}\boldsymbol{\Phi}_{\cdot,\le S}^T\boldsymbol{\Phi}_{\cdot,\le S}|\ge 1$ is nondecreasing in S. This completes the proof.

This result allows us to bound $\mathrm{E}[R]$, given that we know the asymptotic behavior of λ_s as $s\to\infty$. However, the eigenvalues of the Mercer expansion of K w.r.t. μ are known explicitly only for a few special cases. Widom [6] gives a powerful theorem which characterizes λ_s $(s\to\infty)$ in a useful way, under some conditions on K and μ. In the sequel, $A\sim B$ means that $A/B\to 1$ in the limit which is given by the context.

A kernel K is called *stationary* if $K(\boldsymbol{x},\boldsymbol{x}')=K(\boldsymbol{x}-\boldsymbol{x}')$, and *isotropic* if $K(\boldsymbol{x},\boldsymbol{x}')=K(\|\boldsymbol{x}-\boldsymbol{x}'\|)$. For example, the Gaussian kernel (7) is isotropic. Bochner's theorem [11] asserts that the class of stationary covariance functions with $K(\boldsymbol{0})=1$ (also called stationary correlation functions) is identical to the class of characteristic functions of probability distributions: $K(\boldsymbol{r})=\mathrm{E}[\exp(i\boldsymbol{\omega}^T\boldsymbol{r})]$, where $\boldsymbol{\omega}\in\mathbb{R}^d$ is a random variable. If the distribution of $\boldsymbol{\omega}$ has a density, this is called the *spectral density*[2] $\lambda(\boldsymbol{\omega})$ of $K(\boldsymbol{x})$. For isotropic covariance functions, we have $K(\boldsymbol{r})=K(r)$, $r=\|\boldsymbol{r}\|$, and therefore $\lambda(\boldsymbol{\omega})=\lambda(\eta)$, $\eta=\|\boldsymbol{\omega}\|$.

Widom's theorem applies to isotropic covariance functions with a spectral density $\lambda(\eta)$ which does not decay too fast as $\eta\to\infty$. Moreover, $d\mu$ needs to have a density $\mu(\boldsymbol{x})$ w.r.t. $d\boldsymbol{x}$ which is bounded and has bounded support.[3] The theorem and its requirements are detailed in Appendix III. It is interesting to note that the Gaussian kernel (7) does *not* fulfil Widom's requirements, since the tails of its spectral density decay exponentially fast. We have the following theorem.

Theorem 2: Let $K(r)$ be an isotropic covariance function in $\mathbb{R}^d$ with strictly decreasing spectral density $\lambda(\eta)$, fulfilling the requirements for Widom's theorem (Appendix III). Suppose that the covariate distribution μ has bounded support and a bounded density, in that $\mu(\boldsymbol{x})\le D$, and $\mu(\boldsymbol{x})=0$ for $\|\boldsymbol{x}\|>T$. Then

$$\lambda_s\le D(2\pi)^d\lambda\left(\frac{2\Gamma(d/2+1)^{2/d}}{T}s^{1/d}\right)(1+o(1))$$

asymptotically as $s\to\infty$.

A proof is given in Appendix III. In the sequel, we apply this result in order to bound $\mathrm{E}[R]$ for a class of kernels which is frequently used in practice.

IV. Applications to Consistency and Convergence Rates

In this section, we apply the spectral techniques introduced in Section III in order to bound the expected regret term $\mathrm{E}[R]$ for several practically important settings K,μ, thereby obtaining information consistency rates via our main result.

[2]We have that $\lambda(\boldsymbol{\omega})=(2\pi)^{-d}\int\exp(-i\boldsymbol{\omega}^T\boldsymbol{r})K(\boldsymbol{r})d\boldsymbol{r}$.

[3]It is conjectured in [6] that this requirement may not be necessary, but the proof given there uses the bounded support of μ.

A. Finite-Dimensional Covariance Functions

If $K(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}',\ \boldsymbol{x},\boldsymbol{x}' \in \mathbb{R}^d$, we obtain the parametric linear model: $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}$, $\boldsymbol{w} \sim N(\boldsymbol{0},\boldsymbol{I})$ *a priori*. The RKHS $\mathcal{H}$ is $\{\boldsymbol{x} \mapsto \boldsymbol{w}^T\boldsymbol{x}\}$. Let $\boldsymbol{X} = (\boldsymbol{x}_1 \ldots \boldsymbol{x}_n)^T \in \mathbb{R}^{n,d}$, then $R = \log|\boldsymbol{I} + c\boldsymbol{X}^T\boldsymbol{X}|$. It is shown in [16] that if $\|\boldsymbol{x}\| \leq 1$ for all $\boldsymbol{x}$, then $R \leq d\log(1 + cn/d)$. If the covariate distribution μ has bounded support, we have $\mathrm{E}[R] = O(\log n)$, therefore, Bayesian prediction with the parametric linear model is information consistent with rate $O(n^{-1}\log n)$. Note that there is a linear dependence on the covariate dimensionality d. A more general result, covering other parametric models, is given in [2].

B. Gaussian Kernel, Gaussian Covariates

The Gaussian (or Radial Basis Function) kernel is

$$K(r) = \exp\left(-br^2\right), \qquad r = \|\boldsymbol{x} - \boldsymbol{x}'\| \tag{7}$$

for input points $\boldsymbol{x} \in \mathbb{R}^d$. $b > 0$ is a scale parameter, in that $b^{-1/2}$ is the typical length scale in $\mathcal{X}$. The Gaussian kernel is frequently used in Machine Learning for tasks where d can be quite large. For small input dimensions common in geostatistical applications, the Gaussian kernel is not suitable, because it enforces an unreasonably high degree of smoothness [11]. If we choose the covariate distribution to be Gaussian, namely $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0},(4a)^{-1}\boldsymbol{I})$, the kernel eigenvalues are known [18], and by using Lemma 1 we obtain a tight bound on $\mathrm{E}[R]$

$$\mathrm{E}\left[\log|\boldsymbol{I} + c\boldsymbol{K}|\right] = O\left((\log n)^{d+1}\right).$$

Here, the leading constant is $[\log(1 + 2a/b)]^{-d}$, which decreases in a/b, being the squared ratio of the length scale of the kernel and the standard deviation of μ. This makes sense: if a/b is small, typical functions (with RKHS norm of $O(1)$) change on average rapidly and significantly within the typical range of μ. In other words, the penalization of such rapid variations is weaker under the RKHS norm, and therefore the expected regret term has to be larger. If a/b is large, typical functions do not change much in the typical range of μ, which justifies a small expected regret term.

A proof is provided in Appendix II. The result matches our intuition in that the regularization imposed by the RKHS norm becomes weaker with a higher input dimensionality (the RKHS for dimension d is actually the tensor product of d copies of the RKHS for dimension 1). To conclude, even though the RKHS $\mathcal{H}$ for this kernel is a space dense in the continuous functions, the expected regret is very small. This can be explained by the strong smoothness constraint enforced via $\|\cdot\|_K$, which grows quickly with irregularities in f. With a view on Section I and Theorem 1, we see that Bayesian GP prediction with the Gaussian kernel is information consistent in any dimension d, and for all $a, b > 0$, and we have an information rate bound of

$$\frac{1}{2n}\|f\|_K^2 + O\left(n^{-1}(\log n)^{d+1}\right).$$

C. Matérn Kernels, Bounded Support Covariates

Recall that isotropic correlation functions are characteristic functions of probability distributions. An important class of isotropic kernels is obtained this way from Student-t distributions, it is referred to as *Matérn* class (see [11] for details; we use [10, Sec. 4.2.1.] here)

$$K(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}(r/\alpha)^\nu K_\nu(r/\alpha), \qquad \alpha > 0,\ \ \nu > 0 \tag{8}$$

where K_ν is a modified Bessel function. The spectral density is

$$\lambda(\eta) = f_{\alpha,\nu}(\eta) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)}\alpha^d\left(1 + (\alpha\eta)^2\right)^{-\nu-d/2} \tag{9}$$

which is the multivariate t-density in $\mathbb{R}^d$ with $2\nu+d$ degrees of freedom and scale matrix $\alpha^{-2}\boldsymbol{I}$. While α is a scale parameter, the parameter ν directly controls the smoothness of sample paths of the process: they are l times differentiable for some version of the process iff $l < \nu$. For $\nu = 1/2$, $K(r) \propto e^{-r/\alpha}$ is the Ornstein–Uhlenbeck kernel, corresponding GPs are Markov processes, and therefore very irregular. On the other hand, if $\nu \to \infty$ and $\alpha = l(2\nu)^{-1/2}$ for fixed l, then $K(r)$ becomes the Gaussian kernel $e^{-(r/l)^2}$, whose sample paths are analytic functions.

It is easy to see that $\lambda = f_{\alpha,\nu}$ fulfils the conditions of Widom's theorem. For large η, $\lambda(\eta) \sim A\eta^{-(2\nu+d)}$, and from Theorem 2 we obtain $\lambda_s = O(s^{-(2\nu+d)/d})$ if μ has bounded support and density. We show in Appendix IV that this implies that

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right). \tag{10}$$

Note that the regret term is much larger than for the Gaussian kernel. It decays the faster, the larger the smoothness parameter ν becomes, or the smaller the dimension d of the input space. Recalling Section I and Theorem 1, we see that Bayesian GP prediction with the Matérn class is information consistent in any dimension d and for any $\nu > 0$, and we have an information rate bound of

$$\frac{1}{2n}\|f\|_K^2 + O\left(n^{-2\nu/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right). \tag{11}$$

Note that the leading constant in the bound on $\mathrm{E}[R]$ just derived depends on the size T of the support of μ. In fact, the dependence is as large as $T^{2\nu+d}$. If μ has unbounded support, we could try to obtain insight into the setup K, μ by defining $\mu_T(\boldsymbol{x}) = \mu(\boldsymbol{x})\mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}$, then studying the behavior of $\mathrm{E}_{\mu_T}[R]$. The result obtained above is not useful in that respect.

D. Matérn Kernels: General Covariates

Let K be the Matérn kernel with spectral density $\lambda = f_{\alpha,\nu}$, and suppose that $\mu(\boldsymbol{x})$ is bounded, but does not necessarily have bounded support. In this case, Theorem 2 is not useful. On the other hand, Widom's theorem we used so far has been proven only for μ of bounded support, so that it cannot be used directly in order to obtain a bound on $\mathrm{E}[R]$. We can still obtain some insight into the pair K, μ through the following theorem.

Theorem 3: Let $K(r)$ be from the Matérn class, with spectral density $\lambda(\eta) = f_{\alpha,\nu}(\eta)$. Suppose that the covariate distribution μ has a bounded density, such that

$$\int \mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\mu(\boldsymbol{x})^{d/(2\nu+d)}d\boldsymbol{x} \leq \tilde{C}$$

where $\tilde{C}$ is a constant independent of $T > 0$. Define the bounded support measure μ_T with density $\mu_T(\boldsymbol{x}) = \mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\mu(\boldsymbol{x})$, and let $\{\lambda_s^{(T)}\}$ be the spectrum of K w.r.t. μ_T. Then, for all $T > 0$ large enough and for all $\delta > 0$, there exists an s_0 such that

$$\lambda_s^{(T)} \leq C(1+\delta)s^{-(2\nu+d)/d}, \qquad \forall s \geq s_0.$$

Here, C is a constant independent of T, δ.

A proof is given in Appendix IV. While the term $s^{-(2\nu+d)/d}$ is the same as that obtained from Theorem 2, the present theorem is stronger (under an additional assumption on μ), in that the leading constant does not grow with T. However, s_0 (defining the speed of convergence) may depend on T, so the theorem does not imply any strong statement on the asymptotics of λ_s. Some examples for μ admissable in Theorem 3 are given in Appendix IV: any Gaussian, or any Student-t with smoothness

parameter $\nu_2 > \nu$ (tails of μ lighter than tails of λ). On the other hand, if $\nu_2 \leq \nu$, μ is not admissable.

We can use Theorem 3 in order to obtain a bound on $\mathrm{E}_{\mu_T}[R]$ of the same form as (10). While the leading constant in this expression does not depend on T, the speed of convergence may do so, and at present we cannot infer a result for $\mathrm{E}_{\mu}[R]$ (if $\mathrm{supp}\mu$ is unbounded). In the same sense, the information rate bound of (11) holds true for any single $T > 0$. Note that for large enough T, μ_T can be renormalized as probability measure, with negligible effect on the constants. Obtaining a rate bound for Matérn K and μ of unbounded support remains an important point for future work.

V. Conclusion

We stated a regret bound for cumulative log loss of Bayesian GP prediction, compared to experts from the RKHS of the prior covariance function, and we gave a fairly elementary proof. We argued how this result can be used to obtain tight information consistency results and rate bounds, namely by bounding the expected regret term $\mathrm{E}[\log|\boldsymbol{I}+c\boldsymbol{K}|]$, where $\boldsymbol{K}$ is the covariance matrix for the covariates $\boldsymbol{x}_{\leq n}$. We gave a number of examples for classes of covariance functions of central importance in practice, bounding the expected regret by way of the covariance operator eigenvalues, which are known in some cases or can be obtained asymptotically in others. Our results depend strongly on parameters of the covariance function and the covariate distribution, and they provide a novel insight into regularization characteristics of these parameters.

Many results about consistency of nonparametric Bayes predictors are known [1]–[4], [19]. A strong notion of consistency is that the posterior has to concentrate on arbitrarily small environments (w.r.t. some metric) of the data-generating function f. Barron *et al.* [3], [4] give such consistency results for general nonparametric methods, but they show that apart from a simple local condition on the prior, namely, that Kullback–Leibler environments of the true f have to be given positive prior mass, additional nonintuitive global conditions are necessary for posterior consistency. The weaker notion of information consistency is used in [2], [3] and is shown to have nicer properties. In contrast to that work, our results here are specific to Bayesian GP prediction, although part of our argument holds for general Bayesian conditional prediction. This has the advantage that our results depend strongly on parameters of the model, such as the prior covariance function or the assumed covariate distribution, which have a clear meaning for practitioners working with these models. Since our results give a novel interpretation of these parameters in terms of regularization properties, they may serve as guidelines for prior choice.

We obtained information convergence rate bounds in a fairly direct manner, and these depend strongly on the specifics of the model. In contrast, rates are very difficult to obtain for stronger notions of consistency [5]. Zhang [19], [20] obtains convergence results and rates using the same convex duality relationship we do here. His results hold for general nonparametric methods, and not surprisingly he requires a global condition on the prior as well. His rate bounds and global condition depend on upper metric entropies, which are very hard to work with in a concrete case such as ours here. Opper and Vivarelli [17] provide bounds on $\mathrm{E}[R]$ for the Gaussian kernel, their motivation is that R controls the learning curve of GP regression with Gaussian noise (see end of Section II).

Appendix I

Proof of Theorem 1: In this appendix, we provide a proof for Theorem 1. We begin with the representer theorem [14], which is proved here for completeness.

Lemma 2 (Representer Theorem): Let $\mathcal{H}$ be the RKHS for kernel K, and let $\rho(\boldsymbol{x}_{\leq n}, f)$ be a functional of $\boldsymbol{x}_{\leq n} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ and $f \in \mathcal{H}$. Let $\mathcal{H}_n$ be the span of $\{K(\cdot, \boldsymbol{x}_i)\}$. If $\rho(\boldsymbol{x}_{\leq n}, f) = \rho(\boldsymbol{x}_{\leq n}, (f(\boldsymbol{x}_i))_i)$, then

$$\inf_{f\in\mathcal{H}} \rho(\boldsymbol{x}_{\leq n}, f) + \|f\|_K^2 = \inf_{f\in\mathcal{H}_n} \rho(\boldsymbol{x}_{\leq n}, f) + \|f\|_K^2.$$

Proof: Since $\mathcal{H}_n \subset \mathcal{H}$, one direction is trivial. For the other one, let $f \in \mathcal{H}$, and let $\tilde{f}$ be the orthogonal projection of f onto $\mathcal{H}_n$ w.r.t. $\|\cdot\|_K$. Now, $f(\boldsymbol{x}_i) = \tilde{f}(\boldsymbol{x}_i) + (f - \tilde{f}, K(\cdot, \boldsymbol{x}_i))_K = \tilde{f}(\boldsymbol{x}_i)$, because $f - \tilde{f}$ is orthogonal to $\mathcal{H}_n$. Here, we used the reproducing property of K. Therefore, $\rho(\boldsymbol{x}_{\leq n}, f) = \rho(\boldsymbol{x}_{\leq n}, \tilde{f})$, and $\|\tilde{f}\|_K \leq \|f\|_K$, which proves the reverse direction.

We now prove our main result. Let $\mathcal{H}_n$ be the span of $\{K(\cdot, \boldsymbol{x}_i)\}$. Fix $f(\cdot) = \sum_i \alpha_i K(\cdot, \boldsymbol{x}_i) \in \mathcal{H}_n$. We start with the following inequality:

$$\begin{aligned} -\log P_{bs}(\boldsymbol{y}_{\leq n}) &\leq \mathrm{E}_Q\left[-\log P(\boldsymbol{y}_{\leq n}|u(\cdot))\right] + \mathrm{D}[Q \parallel P_{bs}] \\ &= -\sum_{i=1}^{n} \mathrm{E}_Q\left[\log P(y_i|u(\boldsymbol{x}_i))\right] + \mathrm{D}[Q \parallel P_{bs}] \quad (12) \end{aligned}$$

where Q, P_{bs} are distributions over the function $u(\cdot)$. This inequality is an instance of the following Fenchel–Legendre duality relationship [21], [22]

$$\mathrm{E}_Q[g(\boldsymbol{v})] \leq \log \mathrm{E}_P\left[e^{g(\boldsymbol{v})}\right] + \mathrm{D}[Q \parallel P]$$

where P, Q are distributions over $\boldsymbol{v}$. The inequality is an equality for $dQ \propto e^g dP$. In our case, P is the zero-mean GP prior P_{bs}, and Q is a GP constructed as follows. Let $\tau^2 > 0$ (to be specified below), and let Q be the posterior from a GP model with prior P_{bs} and Gaussian likelihood term $\prod_{i=1}^n N(\hat{y}_i|u(\boldsymbol{x}_i), \tau^2)$, where $\hat{\boldsymbol{y}} = (\boldsymbol{K} + \tau^2\boldsymbol{I})\boldsymbol{\alpha}$. We have $\mathrm{E}_Q[u(\cdot)] = f(\cdot)$. Let $\boldsymbol{u} = (u(\boldsymbol{x}_i))_i$.

Since $dQ(u(\cdot)) \propto N(\boldsymbol{u}|\hat{\boldsymbol{y}}, \tau^2\boldsymbol{I})dP_{bs}(u(\cdot))$, we have that[4]

$$\mathrm{D}[Q(u(\cdot)) \parallel P_{bs}(u(\cdot))] = \mathrm{D}[Q(\boldsymbol{u}) \parallel P_{bs}(\boldsymbol{u})]$$

and if $\boldsymbol{B} = \boldsymbol{I} + \tau^{-2}\boldsymbol{K}$, then

$$\begin{aligned} \mathrm{D}[Q \parallel P_{bs}] &= \mathrm{D}[Q(\boldsymbol{u}) \parallel P_{bs}(\boldsymbol{u})] \\ &= (1/2)\left(\log|\boldsymbol{B}| + \mathrm{tr}\boldsymbol{B}^{-1} - n + \boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha}\right). \end{aligned}$$

By expanding $-\log P(y|u(\boldsymbol{x}))$ to second order around $\mathrm{E}_Q[u(\boldsymbol{x}_i)] = f(\boldsymbol{x}_i)$, we have

$$\mathrm{E}_Q[-\log P(y_i|u(\boldsymbol{x}_i))] \leq -\log P(y_i|f(\boldsymbol{x}_i)) + (c/2)\mathrm{Var}_Q[u(\boldsymbol{x}_i)]$$

so that

$$\mathrm{E}_Q\left[-\log P(\boldsymbol{y}_{\leq n}|u(\cdot))\right] \leq -\log P(\boldsymbol{y}_{\leq n}|f(\cdot)) + \frac{c}{2}\mathrm{tr}\mathrm{Var}_Q[\boldsymbol{u}].$$

Here, $\mathrm{Var}_Q[\boldsymbol{u}] = (\boldsymbol{K}^{-1} + \tau^{-2}\boldsymbol{I})^{-1} = \boldsymbol{K}\boldsymbol{B}^{-1}$. Combining the bounds gives

$$\begin{aligned} -\log P_{bs}(\boldsymbol{y}_{\leq n}) \leq\ & -\log P(\boldsymbol{y}_{\leq n}|f(\cdot)) + \frac{1}{2}\|f\|_K^2 \\ & + \frac{1}{2}\left(c\,\mathrm{tr}\boldsymbol{K}\boldsymbol{B}^{-1} + \log|\boldsymbol{B}| + \mathrm{tr}\boldsymbol{B}^{-1} - n\right) \quad (13) \end{aligned}$$

where we used $\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha} = \|f\|_K^2$(2). Minimizing over τ^2 results in $\tau^2 = c^{-1}$ (using the spectral decomposition of $\boldsymbol{K}$), and plugging this into (13) proves the theorem in the restricted case

$$\inf_{f\in\mathcal{H}_n} -\log P(\boldsymbol{y}_{\leq n}|f(\cdot)) + \frac{1}{2}\|f\|_K^2.$$

[4]The relative entropy is defined as $\mathrm{D}[Q \parallel P] = \mathrm{E}_Q[\log(dQ/dP)]$, dQ/dP the Radon–Nikodym derivative, if $Q \ll P$, and ∞ otherwise [23, Theorem 1.31]. In our case, dQ/dP depends on $\boldsymbol{u}$ only.

Since the first term depends on f only through the $f(\boldsymbol{x}_i)$, Lemma 2 allows us to take the infimum over all of $\mathcal{H}$ instead. This completes the proof of the inequality.

Now, suppose that $P(y|f(\boldsymbol{x})) = N(y|f(\boldsymbol{x}), \sigma^2)$. There are two bounding steps in the proof: the convex duality argument of (12), and the quadratic expansion of $-\log P(y|f(\boldsymbol{x}))$. The latter is an equality in this case. We noted above that the convex duality step is an equality for $dQ \propto e^g dP$, where $g = \log P(\boldsymbol{y}_{\leq n}|f)$. This Q is constructed as above if $\boldsymbol{\alpha} = (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_{\leq n}$, so that equality holds for the corresponding $f = \sum_i \alpha_i K(\cdot, \boldsymbol{x}_i)$.

Appendix II

Regret for Gaussian Kernel: In this appendix, we bound the expected regret $\mathrm{E}[R]$ for the Gaussian kernel K and Gaussian covariate distribution μ (see Section IV-B). We have $K(r) = \exp(-br^2)$, $\mu(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{0}, (4a)^{-1}\boldsymbol{I})$. In this case, the eigenexpansion of K w.r.t. $\mu(4)$ is known explicitly for $d = 1$ [18]

$$\lambda_l = \sqrt{\frac{2a}{A}} B^l, \quad g = \sqrt{a^2 + 2ab}, \quad A = a + b + g, \quad B = \frac{b}{A}.$$

We first need to obtain a bound on the eigenvalues in the general case of $d \geq 1$. We use the fact that $K(\boldsymbol{x}) = \prod_{j=1}^d K(x_j)$ and $\mu(\boldsymbol{x}) = \prod_{j=1}^d \mu(x_j)$. Therefore, it is clear that the eigenvalues of K in $\mathbb{R}^d$ are d-products of the eigenvalues for the scalar K, thus, $(2a/A)^{d/2} B^l$ appears with multiplicity

$$N = N(l, d) = \binom{l + d - 1}{d - 1}.$$

N is the number of ordered sets $(n_1, \ldots, n_d)$, $n_i \geq 0$ with $\sum_i n_i = l$. This can be seen by noting that $N(l, 2) = l + 1$ and $N(l, d + 1) = \sum_{i=0}^l N(i, d)$. We need the bound $N(l, d) \leq l^d - (l - 1)^d$ for $d \geq 2$, $l \geq 2$. The proof is elementary, using the semantics of $N(l, d)$. Now, consider the sequence of eigenvalues λ_s for d dimensions, consisting of the values $v_l = (2a/A)^{d/2} B^l$ with multiplicity $N(l, d)$. Since $N(1, d) = d > 1$, we alter the sequence λ_s by removing $d - 1$ of the replicas of v_1. For this altered sequence, we have that

$$\lambda_s \leq (2a/A)^{d/2} B^{s^{1/d}}, \qquad s \geq 0.$$

To see this, split the range $s \geq 0$ into blocks of size $N(l, d)$ corresponding to the value of v_l. Now, for $s = (l-1)^d + 1, \ldots, l^d$ we have that $v_l \leq (2a/A)^{d/2} B^{s^{1/d}}$, because $B < 1$. Furthermore, we have $N(l, d) \leq l^d - (l - 1)^d$. We effectively replace the v_l block of size $N(l, d)$ by a block with more elements, whose prefix is a pointwise upper bound.

The modification of the λ_s sequence leads to an additional $O(\log n)$ term in the final result, which is subdominant and will not be mentioned in the sequel. Let $\tilde{c} = c(2a/A)^{d/2}$. Using Lemma 1, we have that

$$\mathrm{E}[R] \leq \sum_{s \geq 0} \log(1 + cn\lambda_s) \leq \sum_{k \geq 0} \log\left(1 + \tilde{c} n B^{k^{1/d}}\right). \qquad (14)$$

B is strictly decreasing in $4a/b$, with $B \to 1$ as $a/b \to 0$. Let $\alpha = -\log B > 0$. We split the right-hand side of (14) into two parts $S_1 + S_2$. For $k_0 = \lceil ((\log n)/\alpha)^d \rceil$, we have

$$\begin{aligned} S_1 &= \sum_{k=0}^{k_0 - 1} \log\left(1 + \tilde{c} n B^{k^{1/d}}\right) \\ &\leq \alpha^{-d} (\log n)^d \log(1 + \tilde{c} n) \\ &= O\left((\log n)^{d+1}\right). \end{aligned}$$

Next, $n B^{k_0^{1/d}} \leq 1$, so that

$$\begin{aligned} S_2 &= \sum_{k \geq k_0} \log\left(1 + \tilde{c} n B^{k^{1/d}}\right) \\ &\leq \tilde{c} n \sum_{k \geq k_0} B^{k^{1/d}} \leq \tilde{c}\left(1 + n \int_{k_0}^\infty \exp\left(-\alpha x^{1/d}\right) dx\right) \\ &\leq \tilde{c}\left(1 + d\alpha^{-d} n \Gamma(d, \log n)\right). \end{aligned}$$

Here, we first use $\log(1 + x) \leq x$, then bound the series by an integral and make use of $k_0^{1/d} \geq (\log n)/\alpha$.

$$\Gamma(d, \beta) = \int_\beta^\infty e^{-t} t^{d-1} dt$$

is the incomplete Gamma function (tail version). We use the substitution $t = \alpha x^{1/d}$ and the fact that $\beta \mapsto \Gamma(d, \beta)$ is nonincreasing. Since $\Gamma(d, \beta) = (d-1)! e^{-\beta} \sum_{k=0}^{d-1} \beta^k / k!$ for $d \in \mathbb{N}$ [24, Eq. 8.352.2], we have

$$S_2 \leq \tilde{c}\left(1 + d!\alpha^{-d} \sum_{k=0}^{d-1} \frac{(\log n)^k}{k!}\right) = O\left((\log n)^{d-1}\right)$$

thus the expected regret for the Gaussian kernel is $O((\log n)^{d+1})$. The leading constant is $\alpha^{-d} \approx [\log(1 + 2a/b)]^{-d}$. While the leading term does not depend on c, there is a term $((\log n)/\alpha)^d (\log c)$, clarifying the dependence on c.

Appendix III
Widom's Theorem

In this appendix, we state a theorem of Widom [6] and show how Theorem 2 is derived from this result. More details can be found in [25].

Let $K(\boldsymbol{r})$ be an isotropic covariance function with spectral density $\lambda(\boldsymbol{\omega})$, i.e.,

$$\lambda(\boldsymbol{\omega}) = (2\pi)^{-d} \int K(\boldsymbol{r}) e^{-i\boldsymbol{\omega}^T \boldsymbol{r}} d\boldsymbol{r}.$$

Note that $\lambda(\boldsymbol{\omega}) = \lambda(\eta)$, $\eta = \|\boldsymbol{\omega}\|$. Widom requires that $\lambda(\eta) \geq 0$, and that its tails do not decay too fast. First, as $\eta \to \infty$: $\lambda(\eta + o(\eta)) \sim \lambda(\eta)$. Second: $\lambda(\eta) = o(\lambda(o(\eta)))$ for any $o(\eta) \to \infty$, $o(\eta)/\eta \to 0$. These are fulfilled for common spectral densities if $\lambda(\eta)$ does not decay faster than $\mathrm{poly}(1/\eta)$. Moreover, the distribution μ has to have a bounded density $\mu(\boldsymbol{x})$ and bounded support. Let

$$\psi(\varepsilon) = (2\pi)^{-d} \int \mathrm{I}_{\{\mu(\boldsymbol{x})\lambda(\boldsymbol{\omega}) > (2\pi)^{-d}\varepsilon\}} d\boldsymbol{x} d\boldsymbol{\omega}$$

and $s = s(\varepsilon) = \min\{s' | \lambda_{s'} > \varepsilon\}$. Widom's theorem states that $\psi(\varepsilon) \sim s(\varepsilon)$ as $\varepsilon \to 0$. Note that if ψ is strictly decreasing, and if $\psi^{-1}(s + o(s)) \sim \psi^{-1}(s)$, then this implies that $\lambda_s \sim \psi^{-1}(s)$.

We now prove Theorem 2. The support of μ is contained in the ball $\{\boldsymbol{x} | \|\boldsymbol{x}\| \leq T\}$, whose volume is $V_T = \pi^{d/2} \Gamma(d/2 + 1)^{-1} T^d$. Furthermore, $\mu(\boldsymbol{x}) \leq D$. We can upper-bound $\psi(\varepsilon)$ by replacing μ by $\mu_U(\boldsymbol{x}) = D \mathrm{I}_{\{\|\boldsymbol{x}\| \leq T\}} \geq \mu(\boldsymbol{x})$. We have

$$\begin{aligned} \psi(\varepsilon) &\leq (2\pi)^{-d} V_T \int \mathrm{I}_{\{\lambda(\boldsymbol{\omega}) \geq (2\pi)^{-d} D^{-1} \varepsilon\}} d\boldsymbol{\omega} \\ &= (2\pi)^{-d} V_T \int \mathrm{I}_{\{\|\boldsymbol{\omega}\| \leq \lambda^{-1}(\gamma\varepsilon)\}} d\boldsymbol{\omega} = (2\pi)^{-d} V_T V_{\lambda^{-1}(\gamma\varepsilon)} \end{aligned}$$

where $\gamma = (2\pi)^{-d} D^{-1}$. Here, ε is taken small enough, so that $\lambda^{-1}(\gamma\varepsilon)$ exists. We equate the right-hand side with s and solve for ε, noting that λ^{-1} is strictly decreasing.

APPENDIX IV
THE MATÉRN CLASS

In this appendix, we bound $\mathrm{E}[R]$ for the Matérn class (8) with parameters α, ν. More details are given in [25]. Recall that there exists A, $\tilde{s}_0$ such that $\lambda_s \leq As^{-(2\nu+d)/d}$ for all $s \geq \tilde{s}_0$. We use Lemma 1 and split the sum over s into two parts $S_1 + S_2$, where S_1 runs up to $s_0 = n^{d/(2\nu+d)}(\log n)^\tau$, and τ is chosen below. Since s_0 grows with n, we can assume that $s_0 \geq \tilde{s}_0$. Then

$$S_1 = \sum_{s=0}^{s_0-1} \log(1 + cn\lambda_s) = O\left(n^{d/(2\nu+d)}(\log n)^{1+\tau}\right)$$

since $\lambda_s \leq K(0)^{1/2}$. Furthermore

$$\begin{aligned} S_2 &= \sum_{s\geq s_0} \log(1 + cn\lambda_s) = O\left(n \sum_{s\geq s_0} s^{-(2\nu+d)/d}\right) \\ &= O\left((\log n)^{-\tau(2\nu+d)/d} \sum_{s\geq s_0} (s/s_0)^{-(2\nu+d)/d}\right). \end{aligned}$$

We lower-bound s/s_0 by s_0 repetitions of $1, 2, \ldots$, thus

$$\begin{aligned} S_2 &= O\left((\log n)^{-\tau(2\nu+d)/d} s_0 \sum_{k\geq 1} k^{-(2\nu+d)/d}\right) \\ &= O\left(n^{d/(2\nu+d)}(\log n)^{\tau(1-(2\nu+d)/d)}\right) \end{aligned}$$

because the series converges for $\nu > 0$ (it is a zeta function). Choosing $\tau = -d/(2\nu + d)$, we have

$$\mathrm{E}[R] = O\left(n^{d/(2\nu+d)}(\log n)^{2\nu/(2\nu+d)}\right).$$

Note that the leading constant is an affine function of A, the leading constant in the eigenvalue asymptotics.

Next, we prove Theorem 3. Recall Appendix III, we use $\psi_T(\varepsilon)$ for the clipped measure μ_T. Let $q = \nu + d/2$, $y = 1 + (\alpha\eta)^2$. Transforming to polar coordinates, then to y, gives

$$\psi_T(\varepsilon) \propto \int_{\|\boldsymbol{x}\|\leq T} \int_1^\infty \mathrm{I}_{\{y^q < \rho\mu(\boldsymbol{x})\}}(y-1)^a dy d\boldsymbol{x}$$

where $\rho = (c_1\varepsilon)^{-1}$, c_1 a constant, and $a = (d-2)/2 > -1$. Integrating out y gives

$$\psi_T(\varepsilon) \sim C_1\rho^{(a+1)/q} \int \mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\mu(\boldsymbol{x})^{(a+1)/q} d\boldsymbol{x}$$

where C_1 is a constant. The latter integral is bounded by $\tilde{C}$, so $\psi_T(\varepsilon) \sim C_2\varepsilon^{-d/(2\nu+d)}$, $C_2 = C_1\tilde{C}c_1^{-d/(2\nu+d)}$. If $C = C_2^{(2\nu+d)/d}$, our statement follows from Widom's theorem.

Finally, we give some examples for μ fulfilling the assumptions of Theorem 3. If $\mu(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate Gaussian, then

$$\begin{aligned} &\int \mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\mu(\boldsymbol{x})^{(a+1)/q} d\boldsymbol{x} \\ &\quad = |2\pi\boldsymbol{\Sigma}|^{\nu/(2\nu+d)} \left(\frac{2\nu+d}{d}\right)^{d/2} \mathrm{E}_{N(\boldsymbol{\mu},(2\nu+d)/d\boldsymbol{\Sigma})}\left[\mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\right] \end{aligned}$$

where the latter expectation is bounded above by one, giving a bound independent of T, which is tight as $T \to \infty$.

Next, let $\mu(\boldsymbol{x}) = f_{\alpha_2,\nu_2}(\|\boldsymbol{x}\|)$ be a Student-t density. Let $q_2 = \nu_2 + d/2$, $\sigma = q_2/q$. The same mathematical operation as above gives

$$\int \mathrm{I}_{\{\|\boldsymbol{x}\|\leq T\}}\mu(\boldsymbol{x})^{(a+1)/q} d\boldsymbol{x} \propto \int_1^{\tilde{T}} z^{-(a+1)\sigma}(z-1)^a dz$$

with $\tilde{T} = 1 + (\alpha_2 T)^2$. We employ the binomial theorem to write $(z-1)^a$ as polynomial in z of degree a. If $\nu_2 > \nu$, then $\sigma > 1$, so all terms z^κ in the integrand have $\kappa < -1$, and the final value can be bounded independently of T. If $\nu_2 = \nu$, the integrand features z^{-1}, giving a term $\log\tilde{T}$. If $\nu_2 < \nu$, the final value contains $\tilde{T}^{(a+1)(1-\sigma)}$. We see that μ is admissable for $\nu_2 > \nu$ (lighter tails than λ), but inadmissable for $\nu_2 \leq \nu$.

ACKNOWLEDGMENT

The authors thank Peter Grünwald, Tong Zhang, Chris Williams, Manfred Opper, and Andrew Ng for many discussions.

REFERENCES

[1] P. Diaconis and D. Freedman, "On the consistency of Bayes estimates," *Ann. Statist.*, vol. 14, pp. 1–26, 1986.

[2] B. Clarke and A. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453–471, May 1990.

[3] A. Barron, "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems," in *Bayesian Statistics 6*, J. Bernardo, J. Berger, A. Dawid, and A. Smith, Eds., 1998, pp. 27–52.

[4] A. Barron, M. Schervish, and L. Wasserman, "The consistency of posterior distributions in nonparametric problems," *Ann. Statist.*, vol. 2, no. 27, pp. 536–561, 1999.

[5] X. Shen and L. Wasserman, "Rates of convergence of posterior distributions," *Ann. Statist.*, vol. 29, no. 3, pp. 687–714, 2001.

[6] H. Widom, "Asymptotic behavior of the eigenvalues of certain integral equations," *Trans. Amer. Math. Soc.*, vol. 109, no. 2, pp. 278–295, 1963.

[7] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, no. 6, pp. 2451–2492, 1997.

[8] N. Cesa-Bianchi and G. Lugosi, "Worst case prediction over sequences under log loss," *Mach. Learn.*, vol. 43, 2001.

[9] M. Seeger, "Gaussian processes for machine learning," *Int. J. Neural Syst.*, vol. 14, no. 2, pp. 69–106, 2004.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

[11] M. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer, 1999.

[12] A. O'Hagan, "Curve fitting and optimal design," *J. Roy. Stat. Soc. B*, vol. 40, no. 1, pp. 1–42, 1978.

[13] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[14] G. Wahba, *Spline Models for Observational Data*, ser. CBMS-NSF Regional Conference Series. Philadelphia, PA: SIAM, 1990.

[15] S. Kakade, M. Seeger, and D. Foster, "Worst-case bounds for Gaussian process models," in *Advances in Neural Information Processing Systems 18*, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006.

[16] S. Kakade and A. Ng, "Online bounds for Bayesian algorithms," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[17] M. Opper and F. Vivarelli, "General bounds on Bayes errors for regression with Gaussian processes," in *Advances in Neural Information Processing Systems 11*, M. Kearns, S. Solla, and D. Cohn, Eds. Cambridge, MA: MIT Press, 1999.

[18] H. Zhu, C. K. I. Williams, R. Rohwer, and M. Morciniec, "Gaussian regression and optimal finite dimensional linear models," in *Neural Networks and Machine Learning*, ser. NATO ASI, C. Bishop, Ed. New York: Springer, 1998, vol. 168.

[19] T. Zhang, "Learning bounds for a generalized family of Bayesian posterior distributions," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004.

[20] T. Zhang, "From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation," *Ann. Statist.*, vol. 34, pp. 2180–2210, 2006.

[21] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[22] R. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

[23] M. Shervish, *Theory of Statistics*. New York: Springer, 1995.

[24] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. San Diego, CA: Academic, 2000.

[25] M. Seeger, Addendum to: Information Consistency of Nonparametric Gaussian Process Methods Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2007, Tech. Rep.