# Addendum to "Information Consistency of Nonparametric Gaussian Process Methods"

Matthias Seeger Max Planck Institute for Biological Cybernetics P.O. Box: 21 69 72012 Tübingen, Germany seeger@tuebingen.mpg.de

August 17, 2019

#### Abstract

In this note, we present additional material for the IEEE Transactions on Information Theory correspondence [1]. The appendix of this paper is technically very dense. With this note, we hope to achieve a better understanding of the concepts used there.

### 1 Widom's Theorem. Application to the Matérn Class

It is argued in [1] how information consistency statements and corresponding rates can be obtained by first proving a regret bound for the cumulative log loss of the Bayesian strategy versus RKHS experts, then bounding the expected regret term

$$E[R], \quad R = \log |I + cK|$$

Now, E[R] depends on the covariance function K for the GP prior (which is also the RKHS kernel for the expert space) and on the covariate distribution  $d\mu(\boldsymbol{x})$ , where  $\boldsymbol{x} \in \mathbb{R}^d$ . For a "linear kernel", when the expert space is a finite-dimensional linear space, then  $E[R] = O(\log n)$  if  $\mu$  has bounded support. For nonlinear kernels<sup>1</sup> however, the RKHS is typically of infinite dimension and contains very irregular functions. Among different kernels, it is now the RKHS norm  $\|f\|_K$  assigned to a function f which correctly captures the complexity of the kernel, rather than the "size" of the RKHS (however this is defined).

A very useful way of studying the kernel K in relation to some probability measure  $\mu$  is to look at the spectrum of the positive semi-definite operator on  $\mathcal{L}_2(\mu)$  induced by K. If Kis continuous and Hilbert-Schmidt in  $L_2(\mu)$ , then the spectrum of this operator is discrete and non-negative, with

$$K(\boldsymbol{x}, \boldsymbol{x}') = \sum_{s \ge 0} \lambda_s \phi_s(\boldsymbol{x}) \phi_s(\boldsymbol{x}').$$
(1)

Here,  $\{(\lambda_s, \phi_s) \mid s \ge 0\}$  is a complete orthonormal eigensystem of K in  $L_2(\mu)$ ,

$$\lambda_s \phi_s(\boldsymbol{x}) = \mathbf{E}_{\boldsymbol{x}' \sim \mu} \left[ K(\boldsymbol{x}, \boldsymbol{x}') \phi_s(\boldsymbol{x}') \right], \qquad (2)$$

<sup>&</sup>lt;sup>1</sup>The so-called "polynomial kernels" are linear kernels from this viewpoint.

with  $\lambda_0 \geq \lambda_1 \geq \ldots \geq 0$ , and  $E[\phi_s(\boldsymbol{x})\phi_t(\boldsymbol{x})] = \delta_{s,t}$ . This is Mercer's theorem. Note that the original proof was for  $\mu$  being an indicator of a compact set, but it has been extended to the case required here and beyond [3]. The Hilbert-Schmidt assumption implies that  $\sum_s \lambda_s^2 < \infty$ , so  $\lambda_s$  decays rapidly to 0, and the series expansion of K converges uniformly. It is shown in [1] how the orthonormality of the  $\phi_s$  can be used in order to bound E[R] in terms of the eigenvalues  $\{\lambda_s\}$  only:

$$\mathbf{E}[R] \le \sum_{s \ge 0} \log\left(1 + c\lambda_s n\right) \tag{3}$$

#### 1.1 Widom's Theorem

A theorem due to Widom [4] gives asymptotic expressions for the eigenvalues  $\lambda_s$  of the operator induced by the kernel on  $\mathcal{L}_2(\mu)$ . It holds for stationary kernels, and the associated computations seem most manageable for *isotropic* kernels:  $K(\boldsymbol{x}, \boldsymbol{x}') = K(\boldsymbol{r}) = K(r)$ , where  $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{x}', \ r = \|\boldsymbol{r}\|$ . In the remainder of this section, we assume that K is isotropic. Let  $\lambda(\boldsymbol{\omega})$  be the spectral density of K, *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's theorem does not hold for all isotropic kernels, but comes with a set of requirements on K.  $\lambda(\eta) \ge 0$  holds exactly for positive semi-definite kernels (Bochner's theorem). The other requirements concern limit behaviour as  $\eta \to \infty$ . Recall that  $X \sim Y$  iff  $X/Y \to 1$ . We require that

$$\lambda(\eta + o(\eta)) \sim \lambda(\eta)$$

for any  $o(\eta)$  such that  $o(\eta)/\eta \to 0$ . Moreover,

$$\lambda(\eta) = o(\lambda(o(\eta)))$$

for any  $o(\eta) \to \infty$  such that  $o(\eta)/\eta \to 0$ . Note that the second assumption is equivalent to

$$\frac{\lambda(\eta)}{\lambda(o(\eta))} \to 0$$

These requirements restrict  $\lambda(\eta)$  (and therefore K) in two different ways. First,  $\lambda(\eta)$  must be rather "regular" as a function. Below, we show an example of  $\lambda(\eta)$  decaying as poly $(1/\eta)$ , which does not fulfil the requirements. This aspect of the conditions may not be of high relevance in practice, since  $\lambda(\eta)$  for common covariance functions tend to be simple functions with regular behaviour.

More importantly, the rate of decay to zero of  $\lambda(\eta)$  is rather heavily restricted by the conditions for Widom's theorem. In essence, this rate of decay must not be too fast. Since this rate of decay controls very directly the amount of smoothing implied by using a particular kernel [2], we can also say that the conditions require these smoothness constraints to be weak enough. It is shown below that the frequently used Gaussian kernel does not fulfil the conditions for Widom's theorem. However, assume that  $\lambda(\eta) \propto p(\eta^{-1})$  for a polynomial p with smallest non-zero monomial  $x^D$ . Then,

$$\frac{\lambda(\eta + o(\eta))}{\lambda(\eta)} \doteq \left(\frac{\eta + o(\eta)}{\eta}\right)^D \to 1, \quad \frac{\lambda(\eta)}{\lambda(o(\eta))} \doteq \left(\frac{o(\eta)}{\eta}\right)^D \to 0.$$

where  $X \doteq Y$  denotes that  $\lim X = \lim Y$ , or that both limits do not exist. The Gaussian kernel  $K(r) = \exp(-br^2)$  has spectral density  $\lambda(\eta) \propto \exp(-\eta^2/(4b))$ . Then,

$$\frac{\lambda(\eta + o(\eta))}{\lambda(\eta)} = e^{-o(\eta)(2\eta + o(\eta))/(4b)},$$

which can be made to diverge or converge to any value in [0, 1] for certain  $o(\eta)$ . In other words, the spectral density decays so fast that  $\lambda(\eta + o(\eta))$  becomes arbitrarily smaller than  $\lambda(\eta)$  for some  $o(\eta)$ , even though  $o(\eta)/\eta \to 0$ . Widom's theorem must not be used for the Gaussian kernel. Nevertheless, E[R] is bounded in [1] for the Gaussian kernel and Gaussian  $\mu$ , since in this case, the eigenspectrum  $\{\lambda_s\}$  is known analytically.

Finally, it is not the case that any spectral density which merely decays no faster than  $\operatorname{poly}(1/\eta)$ , fulfils the conditions for the theorem. For a counter-example, define  $\lambda(\eta) = \eta^{-(N+\sin\eta)}$  with some N > 1. We will consider  $o(\eta) = \pi$  and the sequence  $\eta_k = (2k+1)\pi/2 \to \infty$  as  $k \to \infty$ . Then,

$$\frac{\lambda(\eta_k + o(\eta_k))}{\lambda(\eta_k)} \doteq \eta_{k+1}^{\Delta_k},$$

where  $\Delta_k = \pm 2$ , depending on whether k is odd or even. The ratio oscillates with ever larger amplitude. Note that this example is rather pathetical, and that spectral densities of kernels typically in use do not show such irregular behaviour.

Moreover, the proof of Widom's theorem seems to require that  $\mu$  has a density  $\mu(\boldsymbol{x})$  which is bounded and has bounded support. While it is conjectured in [4] that the latter assumption may not be required, the arguments given there are not very convincing, and we will employ the theorem under the bounded support assumption only<sup>2</sup>.

Define

$$\psi(\varepsilon) = (2\pi)^{-d} \int I_{\{\mu(\boldsymbol{x})\lambda(\boldsymbol{\omega}) > (2\pi)^{-d}\varepsilon\}} d\boldsymbol{x} d\boldsymbol{\omega}$$
(4)

and  $s = s(\varepsilon) = \min\{s' \mid \lambda_{s'} > \varepsilon\}$ . Note that both  $\psi(\varepsilon)$  and  $s(\varepsilon)$  are non-increasing. Moreover,  $s(\varepsilon)$  is unbounded as  $\varepsilon \to 0$ . Strictly speaking, both are required only as  $\varepsilon \to 0$ , so we can always restrict ourselves to an arbitrarily small interval  $\varepsilon \in (0, \varepsilon_0]$ . Widom's theorem states that under the conditions above, we have that

$$\psi(\varepsilon) \sim s(\varepsilon), \quad \varepsilon \to 0.$$
 (5)

It is clear that if we have computed some  $\tilde{\psi}(\varepsilon)$  such that  $\tilde{\psi}(\varepsilon) \sim \psi(\varepsilon)$ , the theorem implies that  $\tilde{\psi}(\varepsilon) \sim s(\varepsilon)$  as well, since "~" is transitive. In the sequel, we will not distinguish between  $\psi(\varepsilon)$  and  $\tilde{\psi}(\varepsilon)$  anymore.

In the cases we are interested in, one can show that  $\psi(\varepsilon)$  is strictly decreasing and  $\psi^{-1}(s + o(s)) \sim \psi^{-1}(s)$  for  $s \to \infty$ . Then, Widom's theorem implies that

$$\lambda_s \sim \psi^{-1}(s), \quad s \to \infty.$$
 (6)

<sup>&</sup>lt;sup>2</sup>This will lead to some rather cumbersome formulations below.

Namely, instantiating (5) with the particular sequence  $(\lambda_s)$  for  $\varepsilon \to 0$ , we have that  $\psi(\lambda_s) \sim s - 1$ , so that  $\psi(\lambda_s) = s + o(s)$ , therefore  $\lambda_s = \psi^{-1}(s + o(s)) \sim \psi^{-1}(s)$ . As an aside, note that under these additional assumptions, we must have that  $\lambda_{s+1} \sim \lambda_s$ . Namely,  $\lambda_{s+1} = \psi^{-1}(s+1+o(s+1)) \sim \psi^{-1}(s) \sim \lambda_s$ , since  $(1+o(s+1))/s \to 0$ . This rules out the Gaussian kernel once more, whose eigenspectrum is decaying exponentially, so  $\lambda_s/\lambda_{s+1} \ge B > 1$ .

The expected regret E[R] is bounded for several setups in [1], using Widom's theorem. The underlying technique is always the same. The infinite sum in (3) is split into a finite beginning  $(s = 0, ..., s_0 - 1)$  and the rest  $(s \ge s_0)$ . Here,  $s_0$  is chosen such that  $s_0 \to \infty$  as  $n \to \infty$ . We then use (6), which means that for any  $\delta > 0$ , there exists some S such that  $\lambda_s \le (1 + \delta)\psi^{-1}(s)$  for all  $s \ge S$ . Since  $s_0$  is unbounded w.r.t. n, we have that  $s_0 \ge S$  for almost all n, and the rest term of (3) can be bounded using  $\psi^{-1}(s)$ . The beginning term is a finite sum which can be bounded by other means. It should be clear that the outcome of this procedure is an upper bound on E[R] which holds for almost all n. It does not directly allow a statement about the tightness of this bound, say in order to obtain a statement of the form  $E[R] = \Omega(...)$ .

We close by noting that the eigenproblem considered in [4] is stated in a different way from (2). Namely, the eigenequation there is

$$\int \mu(\boldsymbol{x})^{1/2} K(\boldsymbol{x}, \boldsymbol{x}') \mu(\boldsymbol{x}')^{1/2} f(\boldsymbol{x}') \, d\boldsymbol{x}' = \lambda f(\boldsymbol{x}), \tag{7}$$

where  $\mu(\boldsymbol{x})$  is called  $V(\boldsymbol{x})$  there. The two formulations are equivalent. If  $(g, \lambda)$  solves (2), let  $f(\boldsymbol{x}) = \mu(\boldsymbol{x})^{1/2}g(\boldsymbol{x})$ . Then,  $(f, \lambda)$  solves (7). Conversely, if  $(f, \lambda)$  solves (7) with  $\lambda > 0$ , then  $\operatorname{supp} f \subset \operatorname{supp} \mu$ . Define  $g(\boldsymbol{x}) = \lambda^{-1} \int K(\boldsymbol{x}, \boldsymbol{x}') \mu(\boldsymbol{x}')^{1/2} f(\boldsymbol{x}') d\boldsymbol{x}'$ . Then, (7) implies that  $g(\boldsymbol{x}') = \mu(\boldsymbol{x}')^{-1/2} f(\boldsymbol{x}')$  for  $\boldsymbol{x}' \in \operatorname{supp} \mu$ , therefore

$$\lambda g(\boldsymbol{x}) = \int_{\boldsymbol{x}' \in \operatorname{supp} \mu} K(\boldsymbol{x}, \boldsymbol{x}') \mu(\boldsymbol{x}')^{-1/2} f(\boldsymbol{x}') \mu(\boldsymbol{x}') \, d\boldsymbol{x}' = \operatorname{E}_{\boldsymbol{x}' \sim \mu} \left[ K(\boldsymbol{x}, \boldsymbol{x}') g(\boldsymbol{x}') \right],$$

so that  $(g, \lambda)$  solves (2).

#### **1.2** Theorems on Asymptotic Eigenvalue Expressions

Two theorems are given in [1] which employ Widom's theorem. The first holds for any strictly decreasing  $\lambda(\eta)$  fulfilling the Widom requirements, and any bounded  $\mu(\boldsymbol{x})$  with bounded support. However, the leading constant in the final expression bounding  $\lambda_s$  or E[R] asymptotically grows with the size of supp  $\mu$ . The second theorem is concerned with kernels K from the Matérn class and  $\mu$  with potentially unbounded support. However, it is essentially required that the tails of  $\mu(\boldsymbol{x})$  decrease faster than those of  $\lambda(\eta)$ . Also, the second theorem does not make a statement about  $\{\lambda_s\}$  or E[R], but rather about the corresponding expressions based on  $\mu(\boldsymbol{x})I_{\{||\boldsymbol{x}|| \leq T\}}, T > 0$ . Importantly, the leading constants in our bounds do not depend on T.

**Theorem 1** 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 (Section 1.1). Suppose that the covariate distribution  $\mu$  has bounded support and a bounded density, in that  $\mu(\mathbf{x}) \leq 1$ 

D, and  $\mu(\mathbf{x}) = 0$  for  $\|\mathbf{x}\| > T$ . 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  $s \to \infty$ .

Proof: The support of  $\mu$  is contained in the ball  $\{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq T\}$ , 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  $\mu(\boldsymbol{x})$  by  $\mu_U(\boldsymbol{x}) = DI_{\{\|\boldsymbol{x}\| \leq T\}} \geq \mu(\boldsymbol{x})$ . We have

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

where  $\gamma = (2\pi)^{-d}D^{-1}$ . Here,  $\varepsilon$  is taken small enough, so that  $\lambda^{-1}(\gamma\varepsilon)$  exists. The right hand side  $R(\varepsilon)$  is invertible, and Widom's theorem gives that  $s - 1 \leq R(\lambda_s)(1 + o(1))$ asymptotically. The statement of the theorem is obtained by inverting R, using that  $\lambda(\eta + o(\eta)) \sim \lambda(\eta)$ .

The next theorem is concerned with kernels K from the Matérn class, whose spectral densities are Student-t densities with  $\alpha > 0$ ,  $\nu > 0$ :

$$\lambda(\eta) = f_{\alpha,\nu}(\eta) = C_t(\alpha,\nu) \left(1 + (\alpha\eta)^2\right)^{-\nu - d/2}, \quad C_t(\alpha,\nu) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)} \alpha^d.$$
(8)

**Theorem 2** Let K(r) be from the Matérn class, with spectral density  $\lambda(\eta) = f_{\alpha,\nu}(\eta)$ . Suppose that the covariate distribution  $\mu$  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  $\mu_T$  with density  $\mu_T(\boldsymbol{x}) = I_{\{\|\boldsymbol{x}\| \leq T\}} \mu(\boldsymbol{x})$ , and let  $\{\lambda_s^{(T)}\}$  be the spectrum of K w.r.t.  $\mu_T$ . Then, for all T > 0 large enough and for all  $\delta > 0$ , there exists a  $s_0$  such that

$$\lambda_s^{(T)} \le C(1+\delta)s^{-(2\nu+d)/d} \quad \forall s \ge s_0$$

Here, C is a constant independent of T,  $\delta$ .

Note that the term  $s^{-(2\nu+d)/d}$  is the same as obtained from Theorem 1 for Matérn K, but the present theorem is stronger (for bounded support  $\mu$ ), in that the leading constant does not grow with T. On the other hand, one has to be careful not to overinterpret Theorem 2 for  $\mu$  of unbounded support. As such, it does not make any statement about the spectrum  $\{\lambda_s\}$  of  $\mu$  without bounded support. This is because  $s_0$  can depend on T, and we cannot prove right now that it stays bounded as  $T \to \infty$ . We conjecture that these difficulties are technical and can probably be overcome. Some more discussion is given below.

*Proof:* It is easily checked that  $\lambda = f_{\alpha,\nu}$  fulfils the conditions for Widom's theorem. We need to upper bound (4) for the measure  $\mu_T$ . We first transform to polar coordinates.

Recall that  $d\boldsymbol{\omega} = A^{d-1}\eta^{d-1}d\eta d\boldsymbol{\sigma}$  with  $d\boldsymbol{\sigma}$  the uniform distribution on the unit sphere, and  $A^{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ . If  $q = \nu + d/2$ ,  $y = 1 + (\alpha \eta)^2$ , then

$$\psi_T(\varepsilon) = C_1 \int_0^\infty \int_{\|\boldsymbol{x}\| \le T} \mathbf{I}_{\{y^{-q}\mu(\boldsymbol{x}) > c_1\varepsilon\}} \eta^{d-1} \, d\boldsymbol{x} \, d\eta,$$

where

$$C_1 = \frac{2^{1-d}\pi^{-d/2}}{\Gamma(d/2)}, \quad c_1 = (2\pi)^{-d} C_t(\alpha,\nu)^{-1}.$$

Let  $\rho = (c_1 \varepsilon)^{-1}$ , a = (d-2)/2 > -1. Note that  $\rho \to \infty$  as  $\varepsilon \to 0$ . Now,

$$\eta^{d-1}(d\eta) = \frac{1}{2}\alpha^{-d}(y-1)^a(dy)$$

so that

$$\psi_T(\varepsilon) = C_2 \int_{\|\boldsymbol{x}\| \le T} \int_1^\infty \mathbf{I}_{\{y^q < \rho\mu(\boldsymbol{x})\}} (y-1)^a \, dy \, d\boldsymbol{x}$$

with  $C_2 = C_1 \alpha^{-d}/2$ . Integrating out y, we have that

$$\psi_T(\varepsilon) \sim C_2(a+1)^{-1} \rho^{(a+1)/q} \int \mathrm{I}_{\{\|\boldsymbol{x}\| \le T\}} \mu(\boldsymbol{x})^{(a+1)/q} d\boldsymbol{x}$$

In fact, the integration leaves us with  $((\rho\mu(\boldsymbol{x}))^{1/q}-1)^{a+1}$ . We can use the binomial theorem in order to write that as polynomial in  $(\rho\mu(\boldsymbol{x}))^{1/q}$ , which is dominated by the highest degree term as  $\varepsilon \to 0$ . Moreover, since  $(y-1)^{a+1} \leq y^{a+1}$  for  $y \geq 1$ , the r.h.s. is also an exact upper bound once  $\rho \geq \rho_0 := \sup\{\mu(\boldsymbol{x})^{-1} | \|\boldsymbol{x}\| \leq T\}$ . Note that  $(a+1)/q = d/(2\nu + d)$ . If  $C_3 = C_2(a+1)^{-1}\tilde{C}c_1^{-d/(2\nu+d)}$ , then  $\psi_T(\varepsilon) \leq C_3(1+o(1))\varepsilon^{-d/(2\nu+d)}$  as  $\varepsilon \to 0$ . Widom's theorem gives that  $s-1 \leq C_3(1+o(1))(\lambda_s^{(T)})^{-d/(2\nu+d)}$ . If  $C = C_3^{(2\nu+d)/d}$ , the statement of the theorem follows by solving for  $\lambda_s^{(T)}$ .

We give some concrete examples for  $\mu$  in order to get an idea about the additional requirement posed in Theorem 2. First, let  $\mu(\boldsymbol{x}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , a multivariate Gaussian with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . Then,

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

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

Second, let  $\mu(\boldsymbol{x}) = f_{\alpha_2,\nu_2}(\|\boldsymbol{x}\|)$  be a Student-*t* density itself. Let  $q_2 = \nu_2 + d/2$ ,  $\sigma = q_2/q$ . Transforming to polar coordinates as above gives

$$\int \mathrm{I}_{\{\|\boldsymbol{x}\| \le T\}} \mu(\boldsymbol{x})^{(a+1)/q} \, d\boldsymbol{x} = \tilde{C}_2 \int_1^{\tilde{T}} z^{-(a+1)\sigma} (z-1)^a \, dz$$

where  $\tilde{T} = 1 + (\alpha_2 T)^2$ , and  $\tilde{C}_2 = A^{d-1}C_t(\alpha_2,\nu_2)^{(a+1)/q}\alpha_2^{-d}/2$ . 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  $z^{\kappa}$  have  $\kappa < -1$ . The final term features different  $\tilde{T}^{\kappa}$  with  $\kappa < 0$ , so can be bounded independently of T > 0. If  $\nu_2 = \nu$ , the integrand features  $z^{-1}$ , so the final term

is  $(a+1)^{-1}(\log \tilde{T}) + O(1)$ , which grows logarithmically in T. If  $\nu_2 < \nu$ , so  $\sigma < 1$ , the final terms contains  $\tilde{T}^{(a+1)(1-\sigma)}$ , growing polynomially in T. In this case, Theorem 2 is applicable only for  $\nu_2 > \nu$ , if the tails of  $\mu(\boldsymbol{x})$  are lighter than the ones of  $\lambda(\boldsymbol{\omega})$ .

Finally, let us return to the problem regarding the scope of Theorem 2. What we would like to obtain is an asymptotic characterization of  $\{\lambda_s\}$ , the spectrum of K w.r.t.  $\mu$ , the latter of unbounded support. Widom's theorem does not seem to provide that, in particular it does not seem sound to use it in the case of unbounded supp  $\mu$ , since its proof seems to rely on this fact. Therefore, while we obtain a statement of the form  $\lambda_s^{(T)} \sim g(s)$ , where g is essentially independent of T, this need not hold uniformly over all T > 0. More specifically, the ratio  $\lambda_s^{(T)}/q(s)$  converges to 1 for every T > 0, but the rate of convergence may well depend on T. In a worst-case scenario, convergence slows down drastically as T grows. There are two limit processes of relevance here,  $s \to \infty$  and  $T \to \infty$ . If Widom's theorem is used for each T, without any additional arguments, we may do the limit  $T \to \infty$  only at the very end. However, no useful results are obtained this way for  $\mu$  whose tails do not decay faster than  $\lambda$ 's. Even worse, the technique of bounding E[R] by splitting (3) into two parts (see Section 1.3) works by setting the split point  $s_0$  as an expression which grows with n. For a single spectrum, we can argue that the asymptotic expression holds (up to a small constant) for  $s \ge s_0$ , if only  $s_0$  is large enough. In a sense, this couples the limit  $s \to \infty$  to the limit  $n \to \infty$ . However, if supp  $\mu$  is not bounded, the  $s_0$  in Theorem 2 can depend on T, and in the worst-case scenario grows unboundedly as  $T \to \infty$ . No matter how we set  $s_0$ as expression of n then, it could potentially be out-run by this growth.

Additional arguments about the convergence  $\mu_T \to \mu$ , or better  $\mu_T^{1/2} K \mu_T^{1/2} \to \mu^{1/2} K \mu_T^{1/2}$ are probably needed here. Intriguingly, the spectrum  $\{\lambda_s\}$  of  $\mu$  certainly exists, moreover the volume of  $\mu - \mu_T$  converges to zero rapidly. On the other hand, the instantiation of Widom's theorem for  $\mu$  heavier tailed than  $\lambda$  is a clear warning sign.

#### 1.3 Expected Regret Bound for Matérn Class

In this section, let K be the Matérn kernel with Student-t spectral density  $\lambda = f_{\alpha,\nu}$ . In Section 1.2, we showed that  $\lambda_s^{(T)} \sim As^{-(2\nu+d)/d}$  for the eigenspectrum w.r.t.  $\mu_T$  of bounded support. Note that if Theorem 1 is used, then  $\mu = \mu_T$ . We will use the notation  $\mu_T$  and  $\lambda_s^{(T)}$ in this section, even if Theorem 1 is used. The constant A may depend on T > 0 in general (Theorem 1), but is independent of T under an additional assumption on  $\mu$  (Theorem 2). As mentioned above, the speed of convergence (in s) may in general depend on T. At present, this hinders us proving a definite bound on  $E_{\mu}[R]$  for  $\mu$  of unbounded support and K a Matérn kernel. Nevertheless, the results obtained here are useful in interpreting K and  $\mu$ .

The general idea is described in Section 1.1, see (3). For any  $\delta > 0$ , there is a  $\tilde{s}_0$  such that  $\lambda_s^{(T)} \leq A(1+\delta)s^{-(2\nu+d)/d}$  for all  $s \geq \tilde{s}_0$ . If Theorem 2 is used, then A does not depend on T, but  $\tilde{s}_0 = \tilde{s}_0(T)$  in general. The split point for the two parts of (3) is  $s_0 = n^{d/(2\nu+d)}(\log n)^{\tau}$ ,  $\tau$  is chosen below. Now, there exists a  $n_0$  such that  $s_0 \geq \tilde{s}_0$  for all  $n \geq n_0$ , and our final statement will hold only for such n. First,

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

Here,  $\lambda_{\max}$  is a universal upper bound on  $\lambda_s^{(T)}$  for all s, T, for example  $\lambda_{\max} = K(0)^{1/2}$  does the job for isotropic K. Next,

$$S_2 = \sum_{s \ge s_0} \log(1 + cn\lambda_s) = O\left(n \sum_{s \ge s_0} s^{-(2\nu+d)/d}\right)$$
$$= O\left((\log n)^{-\tau(2\nu+d)/d} \sum_{s \ge s_0} (s/s_0)^{-(2\nu+d)/d}\right),$$

where we used that  $\log(1+x) \leq x$ . Note that if Theorem 2 is used, then the leading constant in this bound on  $S_2$  does not depend on T. We lower-bound  $s/s_0$  by  $1, \ldots, 1, 2, \ldots, 2, \ldots$ (each block of length  $s_0$ ), thus

$$S_2 = O\left( (\log n)^{-\tau(2\nu+d)/d} s_0 \sum_{k \ge 1} k^{-(2\nu+d)/d} \right) = O\left( n^{d/(2\nu+d)} (\log n)^{\tau(1-(2\nu+d)/d)} \right),$$

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

$$E_{\mu_T}[R] = O\left(n^{d/(2\nu+d)} (\log n)^{2\nu/(2\nu+d)}\right).$$
(9)

Once more, this must not be interpreted as a statement which holds uniformly over all T > 0. If Theorem 1 has been used in order to bound the eigenspectrum, then the leading constant in (9) depends on T. In fact, it features a term  $T^{2\nu+d}$ . If  $\operatorname{supp} \mu$  is unbounded and Theorem 2 has been used, the leading constant does not depend on T, but the speed of convergence in (9) can in general depend on T. Especially, at present we cannot infer any strong result about  $E_{\mu}[R]$  from (9).

## 2 Information Consistency for Kernels with Summable Spectrum

While we obtain information consistency results and rates for a wide class of kernels and input distributions, these hold only at present if the input distribution has bounded support. The following observation was made after the IEEE IT paper was in print by Ingo Steinwart.

Suppose we have a Mercer kernel for which  $\sum_{s\geq 0} \lambda_s < \infty$ : the spectrum is *summable*. This is stronger than  $\sum_{s\geq 0} \lambda_s^2 < \infty$ , but still true for all explicit kernels and input distributions we address above. As in Appendix IV of [1], we split the r.h.s. of (3) into two parts:

$$S_1 = \sum_{s=0}^{s_0-1} \log(1 + cn\lambda_s) = O(s_0(\log n)), \quad S_2 = \sum_{s \ge s_0} \log(1 + cn\lambda_s) \le cn \sum_{s \ge s_0} \lambda_s.$$

Therefore,

$$\mathbf{E}[R]/n \le C \frac{s_0(\log n)}{n} + c \sum_{s \ge s_0} \lambda_s.$$

For any  $s_0(n) \to \infty$ ,  $s_0(n)(\log n)/n \to 0$ , the r.h.s. converges to zero, showing that we have information consistency. Importantly, summability of eigenvalues can be shown also for input distributions  $\mu(\boldsymbol{x})$  without bounded support. It holds for any stationary kernel with  $K(\mathbf{0}) < \infty$  and any probability measure, because summability of the spectrum is implied by

$$\int K(\boldsymbol{x},\boldsymbol{x})\,d\mu(\boldsymbol{x})<\infty,$$

given the other conditions for Mercer's theorem.

## References

- M. Seeger, S. Kakade, and D. Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.
- [2] M. Stein. Interpolation of Spatial Data: Some Theory for Kriging. Springer, 1999.
- [3] H. Sun. Mercer theorem for RKHS on noncompact sets. Journal of Complexity, 21:337– 349, 2005.
- [4] H. Widom. Asymptotic behavior of the eigenvalues of certain integral equations. Trans. Amer. Math. Soc., 109(2):278–295, 1963.