Journal of Machine Learning Research 22 (2021) 1-27 Submitted 7/21; Revised 10/21; Published 10/21

Consistency of Gaussian Process Regression in Metric Spaces

Peter Koepernik PETER.KOEPERNIK@STCATZ.OX.AC.UK
Department of Statistics
University of Oxford
24–29 St Giles', Oxford OX1 3LB, United Kingdom

Florian Pfaff PFAFF@KIT.EDU
Intelligent Sensor-Actuator-Systems Laboratory
Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe, Germany

Editor: Marc Peter Deisenroth

Abstract

Gaussian process (GP) regressors are used in a wide variety of regression tasks, and many recent applications feature domains that are non-Euclidean manifolds or other metric spaces. In this paper, we examine formal consistency of GP regression on general metric spaces. Specifically, we consider a GP prior on an unknown real-valued function with a metric domain space and examine consistency of the resulting posterior distribution. If the kernel is continuous and the sequence of sampling points lies sufficiently dense, then the variance of the posterior GP is shown to converge to zero almost surely monotonically and in L^p for all $p > 1$, uniformly on compact sets. Moreover, we prove that if the difference between the observed function and the mean function of the prior lies in the reproducing kernel Hilbert space of the prior's kernel, then the posterior mean converges pointwise in L^2 to the unknown function, and, under an additional assumption on the kernel, uniformly on compacts in L^1. This paper provides an important step towards the theoretical legitimization of GP regression on manifolds and other non-Euclidean metric spaces.

Keywords: Gaussian process, regression, nonparametric inference, Bayesian inference, reproducing kernel Hilbert space

1. Introduction

Gaussian Process (GP) regression (Rasmussen and Williams, 2006, Chapter 2) is an established tool for nonparametric modelling of real-world phenomena. It is a Bayesian approach to regression of functional data that assumes a Gaussian process prior on the unknown function, which is updated based on noisy evaluations thereof. While the technique dates back to the 1940s (Wiener, 1949), it became a major research topic in engineering disciplines only recently with the growing availability of computational resources and the widespread popularity of statistical machine learning.

Fundamental research on GP regression comprises consistency proofs. Formal definitions of consistency vary throughout the literature (see, for example, Ghosh and Ramamoorthi, 2003; Ghosh et al., 2006, and references therein), but it describes the general idea that the posterior distribution converges in some sense to the true function as more observations become available. There are several results available in this direction: Choi and Schervish (2007) established posterior consistency in the case where the unknown function's domain is $\mathbb{R}$ and the kernel used for regression is stationary and sufficiently often differentiable. Shi and Choi (2011) proved consistency of GP regression for a bounded subset of $\mathbb{R}$ and a squared exponential kernel. Ghosal et al. (2006) established consistency results in the context of binary classification with GPs with multidimensional real covariates, under some assumptions on the kernel, including differentiability up to a certain order. Further results have been achieved by Tokdar and Ghosh (2007), who considered non-parametric density estimation using GPs.

All of these results are concerned with the case where the unknown function's domain is Euclidean, and GPs on more general spaces such as manifolds are an active field of research. Lin et al. (2019) mention several applications of GPs on manifolds and achieve some general results by constructing embeddings of the manifolds considered there into Euclidean spaces. Calandra et al. (2016) considered learning a transformation along with the GP regression.

A key challenge when using GPs on manifolds is the definition of a kernel since one has to ensure it is symmetric and positive definite. One idea is to take the formula of a stationary Euclidean kernel and replace the difference between two vectors with the geodesic distance on the manifold. However, the resulting kernel is not necessarily positive definite. For example, Feragen et al. (2015) proved that the Gaussian kernel based on the geodesic distance is only positive definite when the manifold is flat in the sense of Alexandrov (Bridson and Haefliger, 1999). Nonlinear manifolds, such as the unit sphere, are generally not flat. Feragen et al. (2015) further proved that the Laplacian kernel involving the geodesic distance is only positive definite when the metric is conditionally negative definite. Borovitskiy et al. (2020) considered ways to provide valid kernels for manifolds based on the Matérn kernel.

An alternative approach is to define a kernel directly on the considered manifold. Hitczenko and Stein (2012) provided a covariance function for anisotropic GPs based on spherical harmonics (Kennedy and Sadeghi, 2013, Section 7.3.3). A practically motivated GP regression on $[0, 2\pi)$ with a corresponding kernel function was presented by Wahlström and Özkan (2015) for extended object tracking. In that work, all objects were modelled as two-dimensional objects on a plane. A GP regression for three-dimensional objects was examined by Kumru and Özkan (2018), who considered the manifold of the unit sphere. A more complicated manifold was regarded for tracking the position and orientation of a rigid body in a three-dimensional Euclidean space (Lang et al., 2014). The authors used quaternions to describe rotations, which reside on the three-dimensional unit hypersphere. A unique property of this representation is that q and $-q$ describe the same orientation, which was considered in the kernel function.

The consistency of GP regression on manifolds cannot be concluded directly from available results for Euclidean spaces, even if the manifold can be embedded into a Euclidean space in a way that the induced Euclidean distance is topologically equivalent to the geodesic distance on the manifold (such as the unit sphere $\left\{x \in \mathbb{R}^3 \colon \|x\| = 1\right\}$ with the arc-length distance). This is because a positive definite kernel on the manifold can, in general, not be trivially extended to a positive definite kernel on the Euclidean embedding space.

Instead of proving consistency for individual spaces, we provide a proof that covers all manifolds. Specifically, we consider the more general case of GP regression on metric spaces, which includes domain spaces that are not manifolds such as GP regression on probability density functions (Dolgov and Hanebeck, 2018) or function spaces. More precisely, we provide a proof for the consistency of GP regression of an unknown, real-valued function f_0, whose domain is a separable metric space (T, ρ). We show that for any non-degenerate GP prior with continuous kernel K, the variance of the posterior GP converges to zero almost surely monotonically and in L^p for all $p > 1$, uniformly on compact sets. Furthermore, if the difference of f_0 to the prior's mean function lies in the reproducing kernel Hilbert space (RKHS) of K (see van der Vaart and van Zanten, 2008), then the mean of the posterior GP converges in L^2 to the unknown function f_0. We assume that evaluations of f_0 happen at a sequence of known sampling points and are corrupted by Gaussian measurement noise whose variance is allowed to depend continuously on the sampling point. The sampling procedure may be random, provided it satisfies a denseness assumption introduced in this work.

The paper is structured as follows. In the next section, we describe the model and the prior. In Section 3, we state our main results, which we relate to existing definitions of consistency and compare with established results when specialized to the Euclidean case in Section 4. Sections 5 and 6 contain the proofs, and Section 7 comprises a discussion of the result and possible future work. In Appendices A and B, we present brief introductions to reproducing kernel Hilbert spaces and Minkowski dimension.

2. Setup and Model

The task is to estimate an unknown function $f_0 \colon T \to \mathbb{R}$, where T is a metric space, based on noisy evaluations of f_0 at random sampling points $(t_i)_{i\in\mathbb{N}}$. The measurement noise $(\varepsilon_i)_{i\in\mathbb{N}}$ shall be centred Gaussian and independent in subsequent measurements. Its variance may depend on the sampling point, according to a function $\sigma^2 \colon T \to (0, \infty)$. Formally, $(\varepsilon_i)_{i\in\mathbb{N}}$ shall be a sequence of real random variables, conditionally independent given (t_i), such that, for $i \in \mathbb{N}$, the conditional distribution of ε_i, given t_i, is $\mathcal{N}(0, \sigma^2(t_i))$. More precisely,

$$\mathbb{P}^{((\varepsilon_i),(t_i))}(\,\cdot\,) = \int\limits_{T^{\mathbb{N}}}\int\limits_{\mathbb{R}^{\mathbb{N}}} \mathbb{1}_{\{((x_i),(s_i))\in\cdot\}} \left(\bigotimes_{j=1}^{\infty} \mathcal{N}(0, \sigma^2(s_j))\right) \left(\mathrm{d}(x_1, x_2, \ldots)\right) \mathbb{P}^{(t_i)} \left(\mathrm{d}(s_1, s_2, \ldots)\right),$$

where $(\Omega, \mathcal{A}, \mathbb{P})$ denotes the underlying probability space and $\mathbb{P}^X := \mathbb{P}\left(X^{-1}(\,\cdot\,)\right)$ denotes the distribution of a random variable X defined on $(\Omega, \mathcal{A})$. We may achieve this in the following way. Let $\varepsilon := \left(\varepsilon^{(i)}\right)_{i\in\mathbb{N}}$ be an i.i.d. sequence independent of (t_i), with $\varepsilon^{(1)} = \left(\varepsilon^{(1)}(t)\right)_{t\in T}$ an independent family of real random variables such that $\varepsilon^{(1)}(t) \sim \mathcal{N}(0, \sigma^2(t))$ for $t \in T$. Then,

$$\varepsilon_i := \varepsilon^{(i)}(t_i), \quad i \in \mathbb{N}, \tag{1}$$

satisfies the above assumptions. The advantage of this definition is that $\left(\varepsilon^{(i)}\right)$ is defined separately from (t_i), and can thus be used to define a measurement noise via Equation (1) for any given sequence of sampling points. Furthermore, noisy evaluations of f_0 can now be treated as ordinary evaluations of the (random) functions $f_0 + \varepsilon^{(i)}$, $i \in \mathbb{N}$.

We shall now specify the GP estimator in this setting. If $m \colon T \to \mathbb{R}$ is a function and $K \colon T \times T \to \mathbb{R}$ is symmetric and positive definite, then

$$f \sim \mathcal{GP}(m, K),$$

independent of (t_i) and $(\varepsilon^{(i)})$, describes a prior on the unknown observed function. Here $\mathcal{GP}(m, K)$ denotes the law of a Gaussian process on T with mean and covariance functions m and K. Formally, $(f(t))_{t\in T}$ is a family of real random variables, independent of (t_i) and $(\varepsilon^{(i)})$, such that $\boldsymbol{f}(\boldsymbol{s}) \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{s}), \boldsymbol{K}(\boldsymbol{s}, \boldsymbol{s}))$ for any $n \in \mathbb{N}$ and $\boldsymbol{s} \in T^n$. Here we used the notation $\boldsymbol{m}(\boldsymbol{s}) \coloneqq (m(s_i))_{i=1}^n \in \mathbb{R}^n$ and $\boldsymbol{K}(\boldsymbol{s}, \boldsymbol{s}) \coloneqq (K(s_i, s_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$, mutatis mutandis in similar contexts. We also set $\boldsymbol{\varepsilon}(\boldsymbol{s}) \coloneqq \left(\varepsilon^{(i)}(s_i)\right)_{i=1}^n$.

Now if $n \in \mathbb{N}$, $\boldsymbol{s} \in T^n$ and $\boldsymbol{y} \in \mathbb{R}^n$ are given, the distribution of f, conditioned on the events $f(s_i) + \varepsilon^{(i)}(s_i) = y_i$ for $i = 1, \dots, n$, is that of a GP with mean and covariance functions given by

$$f_n(t; \boldsymbol{s}, \boldsymbol{y}) \coloneqq m(t) + \boldsymbol{K}(t, \boldsymbol{s})\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{s})^{-1} (\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{s})), \quad t \in T, \tag{2}$$

$$k_n(t, s; \boldsymbol{s}) \coloneqq K(t, s) - \boldsymbol{K}(t, \boldsymbol{s})\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{s})^{-1}\boldsymbol{K}(\boldsymbol{s}, s), \quad t, s \in T, \tag{3}$$

where

$$\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{s}) \coloneqq \boldsymbol{K}(\boldsymbol{s}, \boldsymbol{s}) + \left(\delta_{ij}\sigma^2(s_i)\right)_{i,j=1}^n .$$

We omit the dependence of f_n and k_n on m, K, and σ^2. In particular,

$$f_n(t; \boldsymbol{s}, \boldsymbol{y}) = \mathbb{E}\left[f(t) \mid \boldsymbol{f}(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s}) = \boldsymbol{y}\right], \tag{4}$$

$$k_n(t, s; \boldsymbol{s}) = \mathrm{Cov}\left(f(t), f(s) \mid \boldsymbol{f}(\boldsymbol{s}) + \boldsymbol{\varepsilon}(\boldsymbol{s}) = \boldsymbol{y}\right), \tag{5}$$

for all $n \in \mathbb{N}$, $\boldsymbol{s} \in T^n$, and $\mathbb{P}^{\boldsymbol{f}(\boldsymbol{s})+\boldsymbol{\varepsilon}(\boldsymbol{s})}$-almost every (a.e.) $\boldsymbol{y} \in \mathbb{R}^n$.

Recall that we assume a random sequence (t_i) in T to be given.

Definition 1 *Let $f_0 \colon T \to \mathbb{R}$ be a function. Then, for $n \in \mathbb{N}$, set $\boldsymbol{t}_n \coloneqq (t_i)_{i=1}^n$, and let $\widehat{f}_n(\,\cdot\,; f_0) \colon T \to \mathbb{R}$ and $\widehat{v}_n(\,\cdot\,) \colon T \to \mathbb{R}$ be defined by*

$$\widehat{f}_n(t; f_0) \coloneqq f_n\left(t; \boldsymbol{t}_n, \boldsymbol{f}_0(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right), \quad t \in T,$$
$$\widehat{v}_n(t) \coloneqq k_n\left(t, t; \boldsymbol{t}_n\right), \qquad t \in T,$$

the pointwise mean and variance of the posterior GP, given the first n noisy observations of f_0.

Note that we omit the dependence of $\widehat{f}_n$ and $\widehat{v}_n$ on m, K, σ^2, and (t_i), though we will at one point explicitly denote the dependence of $\widehat{f}_n$ on $\varepsilon = (\varepsilon^{(i)})$ by writing $\widehat{f}_n(\,\cdot\,; f_0, \varepsilon)$.

Remark 2 *Definition 1 also gives directions on how to implement GP regression on a general metric space: Given a kernel K, a prior mean function m and an estimate for the noise function σ^2 (in the simplest case one may choose $m \equiv 0$ and $\sigma^2(\cdot) \equiv \sigma^2$ for a fixed, estimated noise variance $\sigma^2 > 0$), as well as n sampling points $\boldsymbol{t}_n \in T^n$ and observations $\boldsymbol{y}_n \in \mathbb{R}^n$, the estimated function and uncertainty are*

$$\widehat{f}_n(t) = f_n\left(t; \boldsymbol{t}_n, \boldsymbol{y}_n\right), \quad t \in T,$$
$$\widehat{v}_n(t) = k_n\left(t, t; \boldsymbol{t}_n\right), \quad t \in T.$$

The calculation—see Equations (2) and (3)—requires $O(n^2)$ evaluations of K, m, and σ^2, and multiplications and an inversion of matrices in $\mathbb{R}^n$.

3. Statement of Main Results

We first make a brief remark on the assumptions required of the sequence of sampling points (t_i). Intuitively speaking, averaging out the corrupting noise requires infinitely many measurements at, or close to, every point $t \in T$. In particular, (t_i) must be dense, but this is not sufficient in general. Indeed, if T contains an isolated point, that is, a point $t \in T$ such that $\{t\}$ is open, then this point would have to be measured infinitely often, but for denseness, it is sufficient for (t_i) to contain t only once. In this spirit, we call a sequence $(s_i) \in T^{\mathbb{N}}$ *recurrently dense* if it is dense and, additionally, contains every isolated point infinitely often. This is equivalent to the less intuitive but technically more handy condition that (s_i) has infinitely many points in every open set. Note the obvious but important fact that such a sequence can only exist if T is separable.

Theorem 3 *Let $f_0 \colon T \to \mathbb{R}$ be a function and suppose that T is separable, K and σ^2 are continuous, and (t_i) is almost surely recurrently dense in T. Then, for every compact set $C \subset T$,*

$$\sup_{t \in C} \widehat{v}_n(t) \longrightarrow 0, \quad n \to \infty,$$

almost surely monotonically, as well as in L^p for every $p \in [1, \infty)$. Furthermore, if $f_0 - m$ lies in the RKHS of K, then

$$\widehat{f}_n(t; f_0) \xrightarrow{L^2} f_0(t), \quad n \to \infty,$$

for all $t \in T$.

The following proposition should convince the reader that the distinction between recurrently dense and dense is only relevant in theoretical edge cases, and denseness is sufficient in many common cases. Moreover, i.i.d. sampling according to a distribution whose support is the full the domain will always yield a recurrently dense sequence. Recall that T is (topologically) connected if it cannot be written as a disjoint union of two non-empty open sets.

Proposition 4 *Suppose T is separable.*

(i) If T is connected or has no isolated points, then any dense sequence is recurrently dense,

(ii) If (t_i) is i.i.d. with a distribution that has full support (that is, assigns positive probability to every open set), then it is almost surely recurrently dense.

We can replace the pointwise L^2-convergence of the posterior mean in Theorem 3 by uniform L^1-convergence on compacts under additional assumptions on T and K. The idea is to apply a technique first developed by Dudley (1967) to bound the expected supremum of a Gaussian process. We briefly introduce the relevant terminology.

Definition 5 *Given a symmetric and positive definite function $k \colon T \times T \to \mathbb{R}$, the* Dudley metric *associated with k is*

$$d_k(t, s) := \sqrt{k(t,t) + k(s,s) - 2k(t,s)}, \quad t, s \in T. \tag{6}$$

For $C \subset T$ compact, denote by $N(C, \varepsilon, d_k) \in \mathbb{N} \cup \{\infty\}$ for $\varepsilon > 0$ the minimal number of points in an ε-net with respect to (w.r.t.) d_k of C (that is, the minimal number of balls of d_k-radius ε needed to cover C). Then the Dudley integral *associated with k and C is*

$$J(C, d_k) := \int_0^\infty \sqrt{\log N(C, \varepsilon, d_k)}\, \mathrm{d}\varepsilon.$$

Remark 6 *The fact that d_k is a metric can be seen by taking a GP $\xi \sim \mathcal{GP}(0, k)$ and noting that $d_k(t, s) = \mathbb{E}\big[\, (\xi_t - \xi_s)^2 \,\big]^{1/2}$, which implies the triangle inequality. Definiteness follows from positive definiteness of k, and symmetry is obvious.*

Our key additional assumption will be that $J(C, d_K) < \infty$ on compact sets $C \subset T$. As the following proposition shows, this is easy to satisfy at least in finite-dimensional applications. Recall the definition of Minkowski dimension from Appendix B and that a function on T is called locally Lipschitz continuous if, for any point, there is a neighbourhood on which it is Lipschitz continuous.

Proposition 7 *Suppose K is locally Lipschitz continuous, and that the Minkowski dimension of T is finite. Then, $J(C, d_K) < \infty$ for all compact sets $C \subset T$.*

In reasonably nice spaces, the Minkowski dimension takes the value one would expect. In fact, if T is a connected, Riemannian n-manifold with the geodesic metric, then its Minkowski dimension is n. For general manifolds, this depends on the metric, but it is certainly always possible to choose one for which the Minkowski dimension is at least finite (see Lemmas B.2 and B.3).

We will further assume that f_0 is continuous, and it is easy to see heuristically why this should be necessary: If we take $m = 0$ for the moment, then the posterior mean $\widehat{f}_n(\cdot; f_0)$ is continuous by continuity of K (regardless of f_0), and the uniform limit of a sequence of continuous functions must be continuous.

Our final additional assumption will be that T is σ-compact, which means it is the countable union of compact sets. This assumption is not as restrictive as it may at first seem. First, every manifold is σ-compact (Lee, 2013, Lem. 1.10). Second, if T is complete and separable, and all samples (t_i) are drawn according to the same probability distribution Q on T, then the support of Q—the subspace that will eventually be explored by (t_i)—is σ-compact, even if T is not (Parthasarathy, 2014, Theorem 3.2). In that case, one might just take T to be that σ-compact subspace to begin with. Denote by $C(T, \mathbb{R})$ the space of continuous functions $T \to \mathbb{R}$, equipped with the topology of uniform convergence on compacts.

Theorem 8 *If the assumptions of Theorem 3 hold and T is σ-compact, f_0 and m are continuous, and $J(C, d_K) < \infty$ for all compact sets $C \subset T$, then*

$$\sup_{t \in C} \left| f_0(t) - \widehat{f}_n(t; f_0) \right| \xrightarrow{L^1} 0,$$

for every compact $C \subset T$. Moreover, the posterior GP is continuous for all $n \in \mathbb{N}$, and if $\Pi_n = \mathcal{GP}(\widehat{f}_n(\cdot; f_0), k_n(\cdot, \cdot; \boldsymbol{t}_n))$ denotes its distribution on $C(T, \mathbb{R})$, then

$$\Pi_n\,(U) \xrightarrow{L^1} 1$$

for every open neighbourhood U of f_0.

4. Comparison with Prior Work

We briefly discuss how our results relate to existing definitions of consistency, and show that they are of interest even if specialized to the well-studied Euclidean case, by comparing with established results.

The usual notion of consistency in Bayesian analysis is that the posterior distribution Π_n on (the Borel σ-algebra of) a topological parameter space Θ satisfies

$$\forall U \subset \Theta \text{ open with } \theta_0 \in U \colon \Pi_n(U) \longrightarrow 1, \tag{7}$$

where $\theta_0 \in \Theta$ is the true parameter (cf. Section 1.3 in Ghosh and Ramamoorthi, 2003). This specializes to GP regression if $\Theta \subset \mathbb{R}^T$ is a space of functions, $\theta_0 = f_0$, and the prior $\Pi_0 = \mathcal{GP}(m, K)$ on Θ is a GP. There are two main variables in this definition. Since Π_n is random (it depends on the measurement noise and the sampling points), the sense in which $\Pi_n(U) \to 1$ has to be specified, which is usually convergence almost surely or in probability. Note that the latter is the same as L^1 convergence here because the two notions coincide for bounded random variables. Secondly, one has to specify $\Theta \subset \mathbb{R}^T$ and its topology. In the present literature, T is usually a compact subset of $\mathbb{R}^d$, and common choices include (cf. Ghosal et al., 2006; Choi and Schervish, 2007)

(i) $\Theta \subset \mathbb{R}^T$ is (a subset of) the space of Borel measurable functions with the topology of convergence in measure w.r.t. a given finite Borel measure Q on T. That is, $g_n \to g$ iff $Q\left(t\colon\ |g_n(t) - g(t)| \geq \varepsilon\right) \to 0$ for all $\varepsilon > 0$,

(ii) Θ and Q as in (i) with the topology of $L^p(Q)$ convergence for some $p > 0$. That is, $g_n \to g$ iff $\int_T |g_n(t) - g(t)|^p \, Q(\mathrm{d}t) \to 0$,

(iii) $\Theta \subset \mathbb{R}^T$ is the space of bounded functions (or a subset such as $C(T, \mathbb{R})$) with the topology of uniform convergence.

The topologies become increasingly fine in the order we have listed them (if defined on a suitable common Θ) in the sense that the notion of convergence becomes stronger. This means that Equation (7) holds for more and finer neighbourhoods of θ_0 for topologies further down the list, hence the associated consistency result becomes stronger.

For general T, (i) and (ii) work without modification, and the natural extension of (iii) is to consider $C(T, \mathbb{R})$ with the topology of uniform convergence on compact sets, in which case (7) is precisely the second part of Theorem 8. If we specialize this to $T = [0, 1]$, it closely resembles a result of Choi and Schervish (2007, Theorem 3). The assertions only differ in that the topology on $C(T, \mathbb{R})$ they consider is of the type (i) and hence weaker than ours; on the other hand, the convergence $\Pi_n(U) \to 1$ they show is almost sure rather than L^1.

Their assumptions are noticeably stronger: m must be continuously differentiable, K must have continuous partial derivatives up to fourth order (which implies our assumptions on K by Proposition 7), (t_i) must be i.i.d. with full support (which is sufficient for us by Proposition 4), and f_0 must be continuously differentiable. Only the latter assumption is not easy to compare with our assumption that $f_0 - m$ is in the RKHS of K, except for certain kernels: The Laplacian kernel on $[0, 1]$, for example, contains continuously differentiable functions (Berlinet and Thomas-Agnan, 2011); in this case at least, our assumption on f_0 is hence weaker.

5. Convergence of Posterior Variance

We turn to the proof of Theorem 3. For the first part, we have to show that if (t_i) is fixed (that is, non-random) and recurrently dense, then $\widehat{v}_n(\cdot)$, which is then also non-random (see Definition 1), converges to zero monotonically and uniformly on compact sets. The asserted L^p-convergence then follows with the monotone convergence theorem, which can be applied since $\mathbb{E}\left[\sup_{t\in C}\widehat{v}_1(t)^p\right] \leq \sup_{t\in C} K(t,t) < \infty$ for $C \subset T$ compact and $p > 1$ by Equation (3) and continuity of K. We first establish monotonicity.

For matrices, we use $\geq$ to denote the Löwner partial ordering. That is, if $k \in \mathbb{N}$ and $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{k\times k}$ are symmetric, write $\boldsymbol{A} \geq 0$ if $\boldsymbol{A}$ is positive semi-definite, and $\boldsymbol{A} \geq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B} \geq 0$. Note that this is consistent with the ordinary order on $\mathbb{R}$ if $k = 1$.

Lemma 9 *Let $\boldsymbol{s} \in \mathbb{R}^n$ for some $n \in \mathbb{N}$, $0 \leq m \leq n$, and, if $m \neq 0$, $\boldsymbol{u} \in \mathbb{R}^m$ a subsequence of $\boldsymbol{s}$. Then, for any $k \in \mathbb{N}$ and $\boldsymbol{t} \in T^k$,*

$$\boldsymbol{k}_n(\boldsymbol{t}, \boldsymbol{t}; \boldsymbol{s}) \leq \boldsymbol{k}_m(\boldsymbol{t}, \boldsymbol{t}; \boldsymbol{u}),$$

where the right-hand side is $\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{t})$ if $m = 0$. In particular, $(\widehat{v}_n(t))_{n\in\mathbb{N}}$ is decreasing for every $t \in T$.

Proof Let $t \in T$. By using induction, we may assume that $m = n - 1$ and, WLOG, $\boldsymbol{u} = (s_1, \ldots, s_{n-1})$. We will not treat the case $n = 1$ (so $m = 0$) separately, but if the reader is uncomfortable with this, he will find it easy to translate the following arguments to that case explicitly. By Equation (3), we have to show that

$$\boldsymbol{K}(\boldsymbol{t}, \boldsymbol{s})\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{s})^{-1}\boldsymbol{K}(\boldsymbol{s}, \boldsymbol{t}) \geq \boldsymbol{K}(\boldsymbol{t}, \boldsymbol{u})\boldsymbol{B}(\boldsymbol{u}, \boldsymbol{u})^{-1}\boldsymbol{K}(\boldsymbol{u}, \boldsymbol{t}).$$

Then,

$$\boldsymbol{B}(\boldsymbol{s}, \boldsymbol{s}) = \begin{bmatrix} \boldsymbol{B}(\boldsymbol{u}, \boldsymbol{u}) & \boldsymbol{K}(\boldsymbol{u}, s_n) \\ \boldsymbol{K}(\boldsymbol{u}, s_n)^\top & K(s_n, s_n) \end{bmatrix} =: \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{bmatrix} =: \boldsymbol{A}_\star,$$

$$\boldsymbol{K}(\boldsymbol{s}, \boldsymbol{t}) = \begin{pmatrix} \boldsymbol{K}(\boldsymbol{u}, \boldsymbol{t}) \\ \boldsymbol{K}(s_n, \boldsymbol{t}) \end{pmatrix} =: \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{w}^\top \end{pmatrix} =: \boldsymbol{V}_\star.$$

We now have to show that $\boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V} \leq \boldsymbol{V}_\star^\top\boldsymbol{A}_\star^{-1}\boldsymbol{V}_\star$. Put $\chi := c - \boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{b}$, which is well-defined since $\boldsymbol{A} = \boldsymbol{B}(\boldsymbol{u}, \boldsymbol{u})$ is invertible. Furthermore, χ is the Schur complement of $\boldsymbol{A}$ in $\boldsymbol{A}_\star$, and since $\boldsymbol{A}$ is positive definite, Theorem 1.12(a) of Zhang (2006) implies that $\chi > 0$. Hence, by the block matrix inversion formula, we have

$$\boldsymbol{A}_\star^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{bmatrix} + \chi^{-1}\begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{b} \\ -\boldsymbol{b}^\top\boldsymbol{A}^{-1} & 1 \end{bmatrix}.$$

Thus,

$$\begin{aligned}\boldsymbol{V}_\star^\top\boldsymbol{A}_\star^{-1}\boldsymbol{V}_\star = \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{V} \\ &+ \chi^{-1}\underbrace{\left[\boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{V} - \boldsymbol{V}^\top\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{w}^\top - \boldsymbol{w}\boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{V} + \boldsymbol{w}\boldsymbol{w}^\top\right]}_{=:\boldsymbol{X}}.\end{aligned}$$

Since $\chi > 0$, it remains to show that $\boldsymbol{X} \geq 0$. Set $\boldsymbol{a} \coloneqq \boldsymbol{V}^\top \boldsymbol{A}^{-1}\boldsymbol{b} \in \mathbb{R}$ and note that $\boldsymbol{b}^\top \boldsymbol{A}^{-1}\boldsymbol{V} = \boldsymbol{a}^\top$, so

$$\boldsymbol{X} = (\boldsymbol{a} - \boldsymbol{w})(\boldsymbol{a} - \boldsymbol{w})^\top \eqqcolon \boldsymbol{x}\boldsymbol{x}^\top.$$

This implies that $\boldsymbol{X} \geq 0$: For any $\boldsymbol{y} \in \mathbb{R}^k$, $\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{y} = \left(\boldsymbol{y}^\top \boldsymbol{x}\right)\left(\boldsymbol{x}^\top \boldsymbol{y}\right) = \left(\boldsymbol{x}^\top \boldsymbol{y}\right)^2 \geq 0.$ ■

Proposition 10 *Suppose that T is separable, K and σ^2 are continuous, and (t_i) is fixed and recurrently dense in T. Then, for every compact $C \subset T$,*

$$\sup_{t\in C} \widehat{v}_n(t) \longrightarrow 0$$

monotonically as $n \to \infty$.

Proof Let $C \subset T$ be compact. Since $\left(\sup_{t\in C} \widehat{v}_n(t)\right)_{n\in\mathbb{N}}$ is a decreasing sequence by Lemma 9, it suffices to show convergence of a subsequence. We choose that subsequence in the following way: For $n \in \mathbb{N}$, let $\delta_n > 0$ be such that

$$\left|K(s_1, s_2) - K(s_1', s_2')\right| < \frac{1}{n^3} \tag{8}$$

for all $s_1, s_2 \in C$, $s_1' \in B(s_1, \delta_n)$, and $s_2' \in B(s_2, \delta_n)$, which exists by uniform continuity of K on $C \times C$. We may arrange for $\delta_n \downarrow 0$. Since C is compact, there exists a finite δ_n-net of C for each $n \in \mathbb{N}$, that is, $l_n \in \mathbb{N}$ open sets $O_{(1,n)}, \ldots, O_{(l_n,n)}$ with diameter at most δ_n that cover C. For $t \in C$, let $i_n(t) \in \{1, \ldots, l_n\}$ be such that $t \in O_{(i_n(t),n)}$. Then, for any $n \in \mathbb{N}$, we know (since (t_i) is recurrently dense) that there exists $k(n) \in \mathbb{N}$ large enough such that for each $i = 1, \ldots, l_n$, at least n of the points $t_1, \ldots, t_{k(n)}$ lie in $O_{(i,n)}$, and such that $(k(n))$ is increasing. For each $i = 1, \ldots, l_n$, let $\boldsymbol{t}_{(i,n)} \in T^n$ comprise n of those points and put the corresponding indices into the set $I_{(i,n)}$.

Let $t \in C$. We have to find an upper bound on $\widehat{v}_n(t)$ that vanishes as $n \to \infty$ and does not depend on t. Set $\boldsymbol{t}_{(n)} \coloneqq \boldsymbol{t}_{(i_n(t),n)}$ and $I_n \coloneqq I_{(i_n(t),t)}$ for $n \in \mathbb{N}$. Then, by construction, $\boldsymbol{t}_{(n)}$ is a subsequence of $(t_1, \ldots, t_{k(n)})$ of length n such that every point in $\boldsymbol{t}_{(n)}$ has distance at most δ_n from t, and the set $I_n \subset \{1, \ldots, k(n)\}$ contains the corresponding indices. For $n \in \mathbb{N}$, set

$$\begin{aligned}
\boldsymbol{B}_n &\coloneqq \boldsymbol{B}(\boldsymbol{t}_{(n)}, \boldsymbol{t}_{(n)}) \in \mathbb{R}^{n\times n},\\
\boldsymbol{k}_n &\coloneqq \boldsymbol{K}(\boldsymbol{t}_{(n)}, t) \in \mathbb{R}^n,\\
\boldsymbol{\kappa}_n &\coloneqq K(t,t)\,(1, \ldots, 1) \in \mathbb{R}^n,
\end{aligned}$$

as well as

$$\boldsymbol{C}_n \coloneqq K(t,t) \begin{bmatrix} 1 + \frac{\sigma^2(t)}{K(t,t)} & 1 & \cdots & 1 \\ 1 & 1 + \frac{\sigma^2(t)}{K(t,t)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \cdots & 1 & 1 + \frac{\sigma^2(t)}{K(t,t)} \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Intuitively, we expect $\boldsymbol{C}_n$ and $\boldsymbol{\kappa}_n$ to be close in some sense to $\boldsymbol{B}_n$ and $\boldsymbol{k}_n$, respectively, as n grows large.

Definition 1 and Lemma 9 imply

$$\begin{aligned}\widehat{v}_{k(n)}(t) = k_{k(n)}(t,t;(t_1,\dots,t_{k(n)})) &\leq k_n(t,t;\boldsymbol{t}_{(n)}) = \left|K(t,t) - \boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{k}_n\right| \\ &\leq \underbrace{\left|K(t,t) - \boldsymbol{\kappa}_n^\top \boldsymbol{C}_n^{-1}\boldsymbol{\kappa}_n\right|}_{\text{(i)}} + \underbrace{\left|\boldsymbol{\kappa}_n^\top \boldsymbol{C}_n^{-1}\boldsymbol{\kappa}_n - \boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{k}_n\right|}_{\text{(ii)}}.\end{aligned}$$

We will prove convergence to zero for each of the summands separately. Before we do that, however, we compute $\boldsymbol{C}_n^{-1}$ and, subsequently, its spectral norm $\left\|\boldsymbol{C}_n^{-1}\right\|_2$. The inverse of $\boldsymbol{C}_n$ is given by

$$\boldsymbol{C}_n^{-1} = \frac{1}{\sigma^2(t)}\frac{1}{n+\frac{\sigma^2(t)}{K(t,t)}}\begin{bmatrix} n-1+\frac{\sigma^2(t)}{K(t,t)} & -1 & \cdots & -1 \\ -1 & n-1+\frac{\sigma^2(t)}{K(t,t)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \cdots & -1 & n-1+\frac{\sigma^2(t)}{K(t,t)} \end{bmatrix}, \tag{9}$$

as can easily be verified by multiplication of the above matrix with $\boldsymbol{C}_n$. Before we compute the norm of $\boldsymbol{C}_n$, recall the definitions of the row-sum norm $\|\cdot\|_\infty$ and the column-sum norm $\|\cdot\|_1$ and the fact that $\|\cdot\|_2 \leq \sqrt{\|\cdot\|_1\|\cdot\|_\infty}$. In particular, $\|A\|_2 \leq \|A\|_\infty$ for symmetric matrices A. Since $\boldsymbol{C}_n^{-1}$ is symmetric, we obtain

$$\left\|\boldsymbol{C}_n^{-1}\right\|_2 \leq \left\|\boldsymbol{C}_n^{-1}\right\|_\infty \leq \frac{1}{\sigma^2(t)}\frac{1}{n+\frac{\sigma^2(t)}{K(t,t)}}\left(2n+\frac{\sigma^2(t)}{K(t,t)}\right) \leq \frac{2}{\sigma^2(t)} \leq \frac{2}{m_\sigma}, \tag{10}$$

where $m_\sigma := \min_{s\in C}\sigma^2(s) > 0$ by compactness of C and since $\sigma^2(\cdot) > 0$ by assumption.

Let us now consider (i). Using Equation (9), we obtain

$$\boldsymbol{\kappa}_n^\top \boldsymbol{C}_n^{-1}\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n^\top \frac{1}{n+\frac{\sigma^2(t)}{K(t,t)}}\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \frac{n}{n+\frac{\sigma^2(t)}{K(t,t)}}K(t,t)$$

and thus

$$\text{(i)} = K(t,t)\left|1-\frac{n}{n+\frac{\sigma^2(t)}{K(t,t)}}\right| = K(t,t)\frac{\frac{\sigma^2(t)}{K(t,t)}}{n+\frac{\sigma^2(t)}{K(t,t)}} = \frac{\sigma^2(t)}{n+\frac{\sigma^2(t)}{K(t,t)}} \leq \frac{\max_{s\in C}\sigma^2(s)}{n},$$

which is a vanishing bound independent of t.

For (ii), we repeatedly apply the triangle inequality to obtain

$$\begin{aligned}\left\|\boldsymbol{\kappa}_n^\top \boldsymbol{C}_n^{-1}\boldsymbol{\kappa}_n - \boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{k}_n\right\| \leq{}& \underbrace{\left\|\boldsymbol{\kappa}_n^\top \boldsymbol{C}_n^{-1}\boldsymbol{\kappa}_n - \boldsymbol{\kappa}_n \boldsymbol{B}_n^{-1}\boldsymbol{\kappa}_n\right\|}_{\text{(ii.1)}} \\ &+ \underbrace{\left\|\boldsymbol{\kappa}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{\kappa}_n - \boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{\kappa}_n\right\|}_{\text{(ii.2)}} \\ &+ \underbrace{\left\|\boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{\kappa}_n - \boldsymbol{k}_n^\top \boldsymbol{B}_n^{-1}\boldsymbol{k}_n\right\|}_{\text{(ii.3)}}.\end{aligned}$$

For the first summand, we must find a vanishing bound on

$$\begin{aligned}\left\|\boldsymbol{C}_n^{-1} - \boldsymbol{B}_n^{-1}\right\|_2 &= \left\|\boldsymbol{B}_n^{-1}\boldsymbol{B}_n\boldsymbol{C}_n^{-1} - \boldsymbol{B}_n^{-1}\boldsymbol{C}_n\boldsymbol{C}_n^{-1}\right\|_2 = \left\|\boldsymbol{B}_n^{-1}\left(\boldsymbol{B}_n - \boldsymbol{C}_n\right)\boldsymbol{C}_n^{-1}\right\|_2 \\ &\leq \left\|\boldsymbol{B}_n^{-1}\right\|_2 \left\|\boldsymbol{B}_n - \boldsymbol{C}_n\right\|_2 \left\|\boldsymbol{C}_n^{-1}\right\|_2.\end{aligned} \tag{11}$$

Recall that since $\boldsymbol{B}_n$ is symmetric and positive definite (as it is the covariance matrix of a non-degenerate Gaussian distribution), we have

$$\left\|\boldsymbol{B}_n^{-1}\right\|_2 = \frac{1}{\lambda_n},$$

where $\lambda_n > 0$ is the smallest eigenvalue of $\boldsymbol{B}_n$. We will now establish a lower bound on λ_n that is independent of n. For that purpose, we consider

$$\boldsymbol{Z} \coloneqq \boldsymbol{f}(\boldsymbol{t}_{(n)}) + \boldsymbol{\varepsilon}(\boldsymbol{t}_{(n)}) \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{t}_{(n)}), \boldsymbol{B}_n).$$

Now let $\boldsymbol{u} \in \mathbb{R}^n$ be an eigenvector of unit norm of $\boldsymbol{B}_n$ with eigenvalue λ_n. Then

$$\begin{aligned}\lambda_n = \boldsymbol{u}^\top \boldsymbol{B}_n \boldsymbol{u} = \mathbb{V}\left(\boldsymbol{u}^\top \boldsymbol{Z}\right) &= \mathbb{V}\left(\sum_{i\in I_n} u_i f(t_i) + \sum_{i\in I_n} u_i \varepsilon^{(i)}(t_i)\right) \\ &= \mathbb{V}\left(\sum_{i\in I_n} u_i f(t_i)\right) + \sum_{i\in I_n} u_i^2 \mathbb{V}(\varepsilon^{(i)}(t_i)) \\ &\geq 0 + \underbrace{\left(\sum_{i\in I_n} u_i^2\right)}_{=\|\boldsymbol{u}\|_2 = 1} \min_{i\in I_n} \mathbb{V}(\varepsilon^{(i)}(t_i)) \\ &= \min_{i\in I_n} \sigma^2(t_i).\end{aligned}$$

By uniform continuity of σ^2 on C, there is some $\delta > 0$ such that $\sigma^2(s) \geq m_\sigma/2$ whenever $\rho(s, C) < \delta$ (recall that $m_\sigma = \min_{s\in C} \sigma^2(s)$). Hence, if $n \in \mathbb{N}$ is such that $\delta_n < \delta$, then for all $i \in I_n$ we have $\rho(t_i, C) \leq \rho(t_i, t) < \delta_n < \delta$, and thus $\sigma^2(t_i) \geq m_\sigma/2$. Therefore, we have $\lambda_n \geq m_\sigma/2$ and thereby

$$\left\|\boldsymbol{B}_n^{-1}\right\|_2 \leq \frac{2}{m_\sigma}$$

for all $n \geq n_0$, for some $n_0 \in \mathbb{N}$ that depends only on σ^2 and C.

Let us now consider $\|\boldsymbol{C}_n - \boldsymbol{B}_n\|$. Since $\boldsymbol{C}_n - \boldsymbol{B}_n = (K(t,t) - K(t_i,t_j))_{i,j \in I_n}$ is symmetric by symmetry of K, we have

$$\|\boldsymbol{C}_n - \boldsymbol{B}_n\|_2 \leq \|\boldsymbol{C}_n - \boldsymbol{B}_n\|_\infty = \sup_{j \in I_n} \left| \sum_{i \in I_n} |K(t,t) - K(t_i,t_j)| \right|.$$

By Equation (8) we have $|K(t,t) - K(t_i,t_j)| \leq 1/n^3$ for all $i, j \in I_n$ and thus

$$\|\boldsymbol{C}_n - \boldsymbol{B}_n\|_2 \leq \frac{1}{n^2}, \quad n \in \mathbb{N}.$$

Combining these results and recalling Equations (10) and (11), we find

$$\left\|\boldsymbol{C}_n^{-1} - \boldsymbol{B}_n^{-1}\right\| \leq \frac{2}{m_\sigma} \frac{1}{n^2} \frac{2}{m_\sigma} = \frac{4}{n^2 m_\sigma^2}$$

for all $n \geq n_0$. Applying this to (ii.1) now yields

$$\text{(ii.1)} \leq \|\boldsymbol{\kappa}_n\|^2 \left\|\boldsymbol{C}_n^{-1} - \boldsymbol{B}_n^{-1}\right\| \leq \left(\sum_{i \in I_n} K(t,t)^2 \right) \frac{4}{n^2 m_\sigma^2} \leq \frac{1}{n} \frac{4M_K^2}{m_\sigma^2} \longrightarrow 0,$$

where we have put $M_K \coloneqq \max_{s \in C} K(s,s) < \infty$. For the remaining two summands (ii.2) and (ii.3), we first observe that

$$\|\boldsymbol{\kappa}_n - \boldsymbol{k}_n\| = \sqrt{\sum_{i \in I_n} \underbrace{(K(t,t) - K(t,t_i))^2}_{\leq (1/n^3)^2}} \leq \sqrt{n} \frac{1}{n^3} \leq \frac{1}{n^2},$$

for all $n \in \mathbb{N}$, which gives

$$\text{(ii.2)} \leq \|\boldsymbol{\kappa}_n - \boldsymbol{k}_n\| \left\|\boldsymbol{B}_n^{-1}\right\| \|\boldsymbol{\kappa}_n\| \leq \frac{1}{n^2} \frac{2}{m_\sigma} \sqrt{n} M_K \longrightarrow 0,$$

as well as

$$\text{(ii.3)} \leq \|\boldsymbol{k}_n\| \left\|\boldsymbol{B}_n^{-1}\right\| \|\boldsymbol{\kappa}_n - \boldsymbol{k}_n\| \leq 2M_K \sqrt{n} \frac{2}{m_\sigma} \frac{1}{n^2} \longrightarrow 0,$$

where we have used that $|(\boldsymbol{\kappa}_n)_i| = K(t,t) \leq M_K$ for all $n \in \mathbb{N}$ and

$$|(\boldsymbol{k}_n)_i| = |K(t,t_i)| \leq 2M_K$$

for all $i \in I_n$ and all but finitely many $n \in \mathbb{N}$ by continuity of K and compactness of C. ■

This finishes the proof of the first part of Theorem 3.

6. Convergence of Posterior Mean

We consider the second part of Theorem 3. What we have to show is that, under the given assumptions, we have

$$\mathbb{E}\left[\left|f_0(t) - \widehat{f}_n(t; f_0)\right|^2\right] \longrightarrow 0, \quad n \to \infty,$$

for all $t \in T$. Note that the expectation averages over the noise and the sequence of sampling points. As an intermediate step, we show the above in the case where the fixed function f_0 is replaced by the GP prior f.

Lemma 11 *If $n \in \mathbb{N}$, $t \in T$, and the sampling points (t_i) are fixed, then*

$$\mathbb{E}\left[\left|f(t) - \widehat{f}_n(t; f)\right|^2\right] = \widehat{v}_n(t).$$

Proof We claim that

$$\widehat{f}_n(t; f) = \mathbb{E}\left[f(t) \mid \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right], \tag{12}$$

where $\boldsymbol{t}_n = (t_1, \ldots, t_n)$. Indeed, this is a special case of the general fact that if X is a real random variable and Y is a random variable with values in a measurable space E, then, by definition, $\mathbb{E}\left[X \mid Y = \,\cdot\,\right] := h(\,\cdot\,)$, where $h\colon E \to \mathbb{R}$ is measurable such that $\mathbb{E}\left[X \mid Y\right] = h(Y)$ (such a function is uniquely determined up to equality $\mathbb{P}^Y$-almost everywhere and exists since $\mathbb{E}\left[X \mid Y\right]$ is $\sigma(Y)$-measurable). In this case, $X = f(t)$, $Y = \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)$, $E = \mathbb{R}^n$, and $h(\,\cdot\,) = f_n(t; \boldsymbol{t}_n, \,\cdot\,)$ (see Equation (4)), and thus,

$$\begin{aligned}\widehat{f}_n(t; f) &= f_n(t; \boldsymbol{t}_n, \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)) = h(Y) = \mathbb{E}\left[X \mid Y\right] \\ &= \mathbb{E}\left[f(t) \mid \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right],\end{aligned}$$

where the first equality holds by Definition 1. We conclude that

$$\begin{aligned}\mathbb{E}\left[\left|f(t) - \widehat{f}_n(t; f)\right|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left|f(t) - \widehat{f}_n(t; f)\right|^2 \,\middle|\, \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left|f(t) - \mathbb{E}\left[f(t) \mid \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right]\right|^2 \,\middle|\, \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right]\right] \\ &= \mathbb{E}\left[\mathbb{V}\left(f(t) \mid \boldsymbol{f}(\boldsymbol{t}_n) + \boldsymbol{\varepsilon}(\boldsymbol{t}_n)\right)\right], \\ &= \mathbb{E}\left[\widehat{v}_n(t)\right],\end{aligned}$$

where we used the tower property of conditional expectation in the first step. Now by Definition 1 and since we assumed (t_i) to be fixed, $\widehat{v}_n(t)$ is non-random, and the expectation can be omitted. ∎

Proposition 12 *Suppose that T is separable, K and σ^2 are continuous, and (t_i) is almost surely (a.s.) recurrently dense in T. Then*

$$\mathbb{E}\left[\left|f(t) - \widehat{f}_n(t; f)\right|^2\right] \longrightarrow 0, \quad t \in T,$$

monotonically as $n \to \infty$.

Proof By independence of (t_i), ε, and f, we have

$$\mathbb{E}\left[\left|f(t)-\widehat{f}_n(t;f)\right|^2\right]=\int_{T^{\mathbb{N}}}\underbrace{\mathbb{E}\left[\left|f(t)-\widehat{f}_n(t;f)\right|^2\,\middle|\,t_1=t_1',\ldots,t_n=t_n'\right]}_{=:g_n(t;(t_i'))}\mathbb{P}^{(t_i)_{i\in\mathbb{N}}}\left(\mathrm{d}(t_1',\ldots)\right).$$

Then $g_n(t;(t_i'))\downarrow 0$ for all recurrently dense sequences (t_i') by Lemma 11 and Proposition 10, hence for $\mathbb{P}^{(t_i)_{i\in\mathbb{N}}}$-almost-all (t_i'). Thus, monotone convergence yields

$$\mathbb{E}\left[\left|f(t)-\widehat{f}_n(t)\right|^2\right]=\int_{T^{\mathbb{N}}}g_n(t;(t_i))\,\mathbb{P}^{(t_i)_{i\in\mathbb{N}}}(\mathrm{d}(t_1,\ldots))\longrightarrow 0$$

monotonically as $n\to\infty$. ∎

We are now equipped to prove the second part of Theorem 3.

Proposition 13 *Let $f_0\colon T\to\mathbb{R}$ be a function such that f_0-m lies in the RKHS of K. If T is separable, K and σ^2 are continuous, and (t_i) is a.s. recurrently dense in T, then*

$$\widehat{f}_n(t;f_0)\xrightarrow{L^2}f_0(t)$$

for all $t\in T$.

Proof Let us first consider the case where $m\equiv 0$. For the scope of this proof, we will explicitly denote the dependence of $\widehat{f}_n$ (see Definition 1) on the measurement noise $\varepsilon=(\varepsilon^{(i)})_{i\in\mathbb{N}}$ by writing $\widehat{f}_n(\,\cdot\,;f,\varepsilon)$.

By Tonelli's theorem, independence of f, ε, and (t_i), and Proposition 12, we have

$$\mathbb{E}_f\left[\mathbb{E}_{\varepsilon,(t_i)}\left[\left|f(t)-\widehat{f}_n(t;f,\varepsilon)\right|^2\right]\right]=\mathbb{E}\left[\left|f(t)-\widehat{f}_n(t;f,\varepsilon)\right|^2\right]\longrightarrow 0,\tag{13}$$

where we used $\mathbb{E}_X[\,\cdot\,]$ as shorthand for $\int\cdot\,\mathrm{d}\mathbb{P}^X$ for a random variable X. In other words,

$$\mathbb{E}_{\varepsilon,(t_i)}\left[\left|f(t)-\widehat{f}_n(t;f,\varepsilon)\right|^2\right]\xrightarrow{L^1}0.$$

In particular, every subsequence of $(n)_{n\in\mathbb{N}}$ contains an almost surely convergent subsequence.

We now use the subsequence criterion to prove that $\mathbb{E}\left[\left|f_0(t)-\widehat{f}_n(t;f_0,\varepsilon)\right|^2\right]\longrightarrow 0$. Let $(l(n))_{n\in\mathbb{N}}$ be a subsequence of $(n)_{n\in\mathbb{N}}$. We choose a subsequence $(k(n))$ of $(l(n))$ such that

$$\mathbb{E}_{\varepsilon,(t_i)}\left[\left|f(t)-\widehat{f}_{k(n)}(t;f,\varepsilon)\right|^2\right]\xrightarrow{\text{a.s.}}0,$$

that is,

$$\mathbb{E}\left[\left|f_1(t)-\widehat{f}_{k(n)}(t;f_1,\varepsilon)\right|^2\right]\longrightarrow 0\quad\text{for }\mathcal{GP}(0,K)\text{-a.e. }f_1\in\mathbb{R}^T,\tag{14}$$

say for all $f_1 \in \Omega_0 \subset \mathbb{R}^T$. We now wish to reduce the consistency of the estimator for f_0 to consistency of suitable functions in Ω_0. Abbreviating $\|\cdot\| := \|\cdot\|_{L^2} = \sqrt{\mathbb{E}\left[(\cdot)^2\right]}$, we find, for any $f_1 \in \Omega_0$,

$$\begin{aligned}\left\|f_0(t) - \widehat{f}_{k(n)}(t; f_0, \varepsilon)\right\| \leq &\left\|f_0(t) - \widehat{f}_{k(n)}(t; f_0, \varepsilon) + f_1(t) - \widehat{f}_{k(n)}(t; f_1, \varepsilon')\right\| \\ &+ \left\|f_1(t) - \widehat{f}_{k(n)}(t; f_1, \varepsilon')\right\|,\end{aligned} \tag{15}$$

where we introduced $\varepsilon' = (\varepsilon'^{(i)})_{i\in\mathbb{N}}$, a noise independent of f and (t_i), such that ε, ε', and $\varepsilon + \varepsilon'$ follow the same distribution. This can be achieved, for example, by choosing independent vectors

$$\begin{pmatrix} \varepsilon^{(i)}(t) \\ \varepsilon'^{(i)}(t) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma^2(t) & -\sigma^2(t)/2 \\ -\sigma^2(t)/2 & \sigma^2(t) \end{pmatrix}\right), \quad t \in T, i \in \mathbb{N}.$$

Then, for all $t \in T$ and $i \in \mathbb{N}$, $\varepsilon^{(i)}(t), \varepsilon'^{(i)}(t), \varepsilon^{(i)}(t) + \varepsilon'^{(i)}(t) \sim \mathcal{N}(0, \sigma^2(t))$.

By Equation (14), the latter expression in Equation (15) vanishes as $n \to \infty$. For the former, we note that by Definition 1, Equation (2), and $m \equiv 0$, the function $\widehat{f}_n(t; \cdot\,, \cdot)$ is linear in the sum of its arguments. Hence, we can write

$$\widehat{f}_{k(n)}(t; f_0, \varepsilon) + \widehat{f}_{k(n)}(t; f_1, \varepsilon') = \widehat{f}_{k(n)}(t; f_0 + f_1, \varepsilon + \varepsilon'), \quad t \in T, n \in \mathbb{N},$$

leading to

$$\begin{aligned}&\left\|f_0(t) - \widehat{f}_{k(n)}(t; f_0, \varepsilon) + f_1(t) - \widehat{f}_{k(n)}(t; f_1, \varepsilon')\right\| \\ &\qquad = \left\|(f_0 + f_1)(t) - \widehat{f}_{k(n)}(t; f_0 + f_1, \varepsilon + \varepsilon')\right\| \\ &\qquad = \left\|(f_0 + f_1)(t) - \widehat{f}_{k(n)}(t; f_0 + f_1, \varepsilon)\right\|\end{aligned} \tag{16}$$

for $t \in T$ and $n \in \mathbb{N}$, where we used that $\varepsilon + \varepsilon'$ has the same distribution as ε and is independent of (t_i) in the last step. Plugging Equation (16) back into Equation (15) gives

$$\begin{aligned}\left\|f_0(t) - \widehat{f}_{k(n)}(t; f_0, \varepsilon)\right\| \leq &\left\|(f_0 + f_1)(t) - \widehat{f}_{k(n)}(t; f_0 + f_1, \varepsilon)\right\| \\ &+ \left\|f_1(t) - \widehat{f}_{k(n)}(t; f_1, \varepsilon)\right\|.\end{aligned} \tag{17}$$

Glancing back at Equation (14), we conclude that the expressions on the right-hand side of the above inequality vanish if $f_0 + f_1 \in \Omega_0$ and $f_1 \in \Omega_0$, respectively. In order to argue the existence of such an f_1, we need to show that

$$\Omega_0 \cap (\Omega_0 - f_0) \neq \varnothing,$$

a sufficient condition for which is that both Ω_0 and $\Omega_0 - f_0$ are one-sets w.r.t. $\mathcal{GP}(0, K)$. The former holds by assumption on Ω_0, the latter follows from Proposition A.2, by which the distributions of f and $f + f_0$ are equivalent and thus have the same one-sets, so

$$\mathbb{P}(f \in \Omega - f_0) = \mathbb{P}(f + f_0 \in \Omega_0) = \mathbb{P}^{f+f_0}(\Omega_0) = 1.$$

Now we regard the case of general m. If we denote the estimator we had obtained if $m \equiv 0$ by $\widehat{g}_n$, then

$$\widehat{f}_n(t; f_0, \varepsilon) = m(t) + \widehat{g}_n(t; f_0 - m, \varepsilon), \tag{18}$$

see Definition 1 and Equation (2). Thus,

$$\begin{aligned} \left\| f_0(t) - \widehat{f}_n(t; f_0, \varepsilon) \right\| &= \| f_0(t) - (m(t) + \widehat{g}_n(t; f_0 - m, \varepsilon)) \| \\ &= \| (f_0 - m)(t) - \widehat{g}_n(t; f_0 - m, \varepsilon) \| , \end{aligned} \tag{19}$$

which vanishes by what we have already shown, since $f_0 - m$ is in the RKHS of K. ■

This concludes the proof of the second assertion in Theorem 3, and we turn to Theorem 8. For the remainder of this section, assume that T is separable and σ-compact, K and σ^2 are continuous, (t_i) is a.s. recurrently dense, f_0 and m are continuous and $f_0 - m$ is in the RKHS of K, and $J(C, d_K) < \infty$ for every compact $C \subset T$.

Theorem 14 *Suppose $(\xi_t)_{t \in T}$ is a centred GP with kernel k and $J(C, d_k) < \infty$ for every compact $C \subset T$.*

(i) If k is continuous, then there exists a continuous modification $\widetilde{\xi}$ of ξ, that is, a continuous process $(\widetilde{\xi}_t)_{t \in T}$ such that $\xi_t = \widetilde{\xi}_t$ a.s. for all $t \in T$.

(ii) There is a universal constant $c > 0$, such that if ξ is continuous, then for any compact $C \subset T$ and any $t_0 \in C$,

$$\mathbb{E}\left[\sup_{t \in C} |\xi_t - \xi_{t_0}|\right] \leq c J(C, d_k).$$

Proof

(i) Let C_i, $i \in \mathbb{N}$, be compact sets such that $C_1 \subset C_2 \subset \ldots$ and $\bigcup_{i=1}^{\infty} C_i = T$. Then, for any $i \in \mathbb{N}$, a result of Bogachev (1998, Corollary 7.1.4) implies that there exists a modification $\widetilde{\xi}_i = (\widetilde{\xi}_i(t))_{t \in C_i}$ of $\xi\big|_{C_i}$ which is continuous w.r.t. d_k, hence also w.r.t. ρ. Indeed, if $\rho(s_n, s) \to 0$, then $d_k(s_n, s) \to 0$ by definition of d_k (Equation (6)) and continuity of k, and thus, $\xi_i(s_n) \to \xi(s)$.

It remains to be shown that the modifications $\widetilde{\xi}_i$ can be “glued” to a single continuous modification on all of T. For $i \in \mathbb{N}$, $\widetilde{\xi}_i$ and $\widetilde{\xi}_{i+1}\big|_{C_i}$ are both continuous and modifications of each other, so $\widetilde{\xi}_i$ and $\widetilde{\xi}_{i+1}$ almost surely coincide on C_i. By intersecting countably many sets of full probability, there is a single set of probability one on which, for every $i, j \in \mathbb{N}$, $i < j$, $\widetilde{\xi}_i$ and $\widetilde{\xi}_j$ coincide on C_i. It is thus well-defined to put, for $t \in T$, $\widetilde{\xi}(t) := \widetilde{\xi}_i(t)$ for any $i \in \mathbb{N}$ with $t \in C_i$, and this yields a continuous modification $\widetilde{\xi} = (\widetilde{\xi}(t))_{t \in T}$ of ξ.

(ii) This is a variant of a result due to Dudley (1967), stated and proved in this form by Zhou (2020, Theorem 1.2).

■

By Theorem 14 (i), we may choose and fix a modification of f (which we again denote by f) such that $f - m$ and hence f is continuous. Recall that the kernel of the posterior GP is $k_n(\cdot,\cdot;\boldsymbol{t}_n)$ (see Equations (3) and (5)), which depends only on the sampling points, and denote by $d_n \coloneqq d_{k_n(\cdot,\cdot;\boldsymbol{t}_n)}$ for $n \in \mathbb{N}$ the associated Dudley metric and $d_0 \coloneqq d_K$.

Lemma 15 *If (t_i) is fixed and recurrently dense, and $C \subset T$ is compact, then*

$$J(C, d_{n+1}) \le J(C, d_n) \le J(C, d_0)$$

for all $n \in \mathbb{N}$, and $J(C, d_n) \longrightarrow 0$ as $n \to \infty$.

Proof Fix $C \subset T$. Lemma 9 implies, for $s, t \in T$, $n \in \mathbb{N}_0$, $\boldsymbol{s} \coloneqq (s,t)$,

$$\begin{aligned} d_{n+1}(s,t)^2 &= (1,-1)\boldsymbol{k}_{n+1}(\boldsymbol{s},\boldsymbol{s};\boldsymbol{t}_{n+1}) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &\le (1,-1)\boldsymbol{k}_{n}(\boldsymbol{s},\boldsymbol{s};\boldsymbol{t}_{n}) \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= d_n(s,t)^2. \end{aligned}$$

In particular, any ε-net w.r.t. d_n is an ε-net w.r.t. d_{n+1}, so

$$N(C,\varepsilon,d_{n+1}) \le N(C,\varepsilon,d_n) \le N(C,\varepsilon,d_0), \quad n \in \mathbb{N}, \varepsilon > 0, \tag{20}$$

and thus

$$J(C, d_{n+1}) \le J(C, d_n) \le J(C, d_0) < \infty, \quad n \in \mathbb{N}.$$

Now take a GP $f_n \sim \mathcal{GP}(0, k_n(\cdot,\cdot;\boldsymbol{t}_n))$, so

$$\begin{aligned} \sup_{s,t\in C} d_n(s,t)^2 &= \sup_{s,t\in C} \mathbb{E}\left[(f_n(t) - f_n(s))^2\right] \le \sup_{s,t\in C} \mathbb{E}\left[2\left(f_n(t)^2 + f_n(s)^2\right)\right] \\ &= 2\sup_{s,t\in C}\left(\widehat{v}_n(t) + \widehat{v}_n(s)\right) \le 4\sup_{t\in C}\widehat{v}_n(t) \eqqcolon D_n, \end{aligned}$$

and recall that $D_n \downarrow 0$ by Proposition 10. Hence, if $\varepsilon > 0$ is fixed, then for $n \in \mathbb{N}$ so large that $D_n \le \varepsilon$, C can be covered in a single ball of d_n-radius ε (centred at an arbitrary point), so $\log N(C,\varepsilon,d_n) = \log 1 = 0$ for all such $n \in \mathbb{N}$. Thus, $\log N(C,\varepsilon,d_n) \longrightarrow 0$ as $n \to \infty$ for every $\varepsilon > 0$, so

$$J(C, d_n) = \int_0^\infty \sqrt{\log N(C,\varepsilon,d_n)}\,\mathrm{d}\varepsilon \longrightarrow 0, \quad n \to \infty,$$

where we could apply dominated convergence with $\sqrt{\log N(C,\varepsilon,d_0)}$ as a dominating integrand by Equation (20) and since $\int_0^\infty \sqrt{\log N(C,\varepsilon,d_0)}\,\mathrm{d}\varepsilon = J(C,d_0) < \infty$. ■

The following result replaces Proposition 12 as the central ingredient in the proof of the first part of Theorem 8.

Proposition 16 *For any compact $C \subset T$,*

$$\mathbb{E}\left[\sup_{t\in C}\left|f(t)-\widehat{f}_n(t;f)\right|\right] \longrightarrow 0.$$

Proof We may assume $m=0$ by the same argument employed in the proof of Proposition 13 (cf. Equations (18) and (19)), so f is a continuous, centred GP. Consider first the case where (t_i) is non-random. Denote, for $n\in\mathbb{N}$, $\boldsymbol{y}_n \coloneqq \boldsymbol{f}(\boldsymbol{t}_n)+\boldsymbol{\varepsilon}(\boldsymbol{t}_n)$, and

$$\overline{f}_n(t) \coloneqq f(t)-\widehat{f}_n(t;f) = f(t)-\mathbb{E}\left[f(t)\,|\,\boldsymbol{y}_n\right], \quad t\in T,$$

where we used Equation (12). Then, $\overline{f}_n$ is a continuous, centred GP (f is continuous, and so is $\widehat{f}_n(\cdot;f)=\boldsymbol{K}(\cdot,\boldsymbol{t}_n)^\top \boldsymbol{B}(\boldsymbol{t}_n,\boldsymbol{t}_n)^{-1}\boldsymbol{y}_n$ by continuity of K), and

$$\begin{aligned}
\operatorname{Cov}\left(\overline{f}_n(t),\overline{f}_n(s)\right) &= \mathbb{E}\left[\overline{f}_n(t)\overline{f}_n(s)\right] = \mathbb{E}\left[\mathbb{E}\left[\overline{f}_n(t)\overline{f}_n(s)\,\middle|\,\boldsymbol{y}_n\right]\right]\\
&= \mathbb{E}\Big[\mathbb{E}\left[f(t)f(s)\,|\,\boldsymbol{y}_n\right]+\mathbb{E}\left[f(t)\,|\,\boldsymbol{y}_n\right]\mathbb{E}\left[f(s)\,|\,\boldsymbol{y}_n\right]\\
&\qquad -\underbrace{\mathbb{E}\Big[f(t)\mathbb{E}\left[f(s)\,|\,\boldsymbol{y}_n\right]\,\Big|\,\boldsymbol{y}_n\Big]}_{\overset{(\star)}{=}\mathbb{E}[f(t)\,|\,\boldsymbol{y}_n]\mathbb{E}[f(s)\,|\,\boldsymbol{y}_n]}-\underbrace{\mathbb{E}\Big[f(s)\mathbb{E}\left[f(t)\,|\,\boldsymbol{y}_n\right]\,\Big|\,\boldsymbol{y}_n\Big]}_{\overset{(\star)}{=}\mathbb{E}[f(s)\,|\,\boldsymbol{y}_n]\mathbb{E}[f(t)\,|\,\boldsymbol{y}_n]}\Big]\\
&= \mathbb{E}\Big[\mathbb{E}\left[f(t)f(s)\,|\,\boldsymbol{y}_n\right]-\mathbb{E}\left[f(t)\,|\,\boldsymbol{y}_n\right]\mathbb{E}\left[f(s)\,|\,\boldsymbol{y}_n\right]\Big]\\
&= \mathbb{E}\left[\operatorname{Cov}\left(f(t),f(s)\,|\,\boldsymbol{y}_n\right)|\right]\\
&= k_n(t,s;\boldsymbol{t}_n),
\end{aligned}$$

for $s,t\in T$, where we used in $(\star)$ that $\mathbb{E}\left[f(q)\,|\,\boldsymbol{y}_n\right]$ for $q=s,t$ is measurable w.r.t. $\boldsymbol{y}_n$ and can hence be pulled out of $\mathbb{E}\left[\cdot\,|\,\boldsymbol{y}_n\right]$; and in the final step we used Equation (5) and that (t_i) is non-random. Hence, $\overline{f}_n=(\overline{f}_n(t))_{t\in T}$ is a centred GP with kernel $k_n(\cdot,\cdot;\boldsymbol{t}_n)$, so we can apply Theorem 14 (ii) for some fixed $t_0\in C$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\sup_{t\in C}\left|\overline{f}_n(t)\right|\right] &\le \overbrace{\mathbb{E}\left[\left|\overline{f}_n(t_0)\right|\right]}^{\le\sqrt{\mathbb{E}[\overline{f}_n(t_0)^2]}=\sqrt{\widehat{v}_n(t_0)}} + \mathbb{E}\left[\sup_{t\in C}\left|\overline{f}_n(t)-\overline{f}_n(t_0)\right|\right]\\
&\le \sqrt{\widehat{v}_n(t_0)}+cJ(C,d_n)\\
&\longrightarrow 0,
\end{aligned} \tag{21}$$

where $\widehat{v}_n(t_0)\longrightarrow 0$ by Proposition 10 and $J(C,d_n)\longrightarrow 0$ by Lemma 15. This finishes the proof for non-random (t_i). If (t_i) is random and a.s. recurrently dense, then

$$\mathbb{E}\left[\sup_{t\in C}\left|\overline{f}_n(t)\right|\right] = \mathbb{E}_{(t_i)}\left[\underbrace{\mathbb{E}_{f,\varepsilon}\left[\sup_{t\in C}\left|\overline{f}_n(t)\right|\right]}_{\longrightarrow 0}\right] \longrightarrow 0,$$

by dominated convergence, for the application of which we use that, by Equation (21),

$$\mathbb{E}_{f,\varepsilon}\left[\sup_{t\in C}\left|\overline{f}_n(t)\right|\right] \le \sqrt{\widehat{v}_n(t_0)}+cJ(C,d_n) \le \sqrt{K(t_0,t_0)}+cJ(C,d_0),$$

where we used that $\widehat{v}_n(t_0) = k_n(t_0, t_0; \boldsymbol{t}_n) \leq K(t_0, t_0)$ by Lemma 9, and $J(C, d_n) \leq J(C, d_0)$ by Lemma 15. This is a finite bound independent of (t_i) and $n \in \mathbb{N}$. ∎

Proof of Theorem 8 For the proof of the first assertion, we may assume $m = 0$ by the same argument employed in the proof of Proposition 13 (cf. Equations (18) and (19)). Then f is continuous, hence defines a random element in $C(T, \mathbb{R})$, and by $\mathbb{P}^f$ we now denote its distribution on the Borel σ-algebra of $C(T, \mathbb{R})$—instead of, as before, the product σ-algebra on the larger space $\mathbb{R}^T$.

We can now essentially copy the proof of Proposition 13, using Proposition 16 in place of Proposition 12 in Equation (13). That is, we start by writing

$$\mathbb{E}_f \left[\mathbb{E}_{\varepsilon,(t_i)} \left[\sup_{t \in C} \left| f(t) - \widehat{f}_n(t; f) \right| \right] \right] = \mathbb{E} \left[\sup_{t \in C} \left| f(t) - \widehat{f}_n(t; f) \right| \right] \longrightarrow 0.$$

In other words, $\mathbb{E}\big[\Delta_n(C, f)\big] \longrightarrow 0$, where $\Delta_n(C, g) := \mathbb{E}\left[\sup_{t\in C} \left| g(t) - \widehat{f}_n(t; g) \right| \right]$ for $g \in C(T, \mathbb{R})$. In particular, for every subsequence $(l(n))$ of $(n)_{n\in\mathbb{N}}$, there is a subsubsequence $(k(n))$ such that $\Delta_{k(n)}(C, f) \xrightarrow{a.s.} 0$, that is (cf. Equation (14)),

$$\Delta_{k(n)}(C, f_1) \longrightarrow 0 \quad \text{for } \mathbb{P}^f\text{-a.e. } f_1 \in C(T, \mathbb{R}), \tag{22}$$

say for all $f_1 \in \Omega_0 \subset C(T, \mathbb{R})$. Here it becomes clear why we must work over $C(T, \mathbb{R})$ instead of $\mathbb{R}^T$: Equation (22) is another way of saying $\mathbb{P}^f\big(g \in C(T, \mathbb{R})\colon \Delta_n(C, g) \longrightarrow 0\big) = 1$; this would not be possible if we worked over $\mathbb{R}^T$, as the event $\big\{g \in \mathbb{R}^T\colon \Delta_n(C, g) \longrightarrow 0\big\}$ is ill-defined. Indeed, the quantity $\Delta_n(C, g)$ is well-defined for continuous g only because the supremum inside the expectation can be reduced to a countable one and is thus measurable, but this is not true for general $g \in \mathbb{R}^T$.

By copying the arguments leading from Equation (14) to Equation (17), we obtain that, for any $f_1 \in \Omega_0$,

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in C} \left| f_0(t) - \widehat{f}_{k(n)}(t; f_0, \varepsilon) \right|\right] \leq{}& \mathbb{E}\left[\sup_{t\in C} \left| (f_0 + f_1)(t) - \widehat{f}_{k(n)}(t; f_0 + f_1, \varepsilon) \right|\right] \\ &+ \mathbb{E}\left[\sup_{t\in C} \left| f_1(t) - \widehat{f}_{k(n)}(t; f_1, \varepsilon) \right|\right].\end{aligned}$$

This converges to zero if we can find $f_1 \in \Omega_0$ that also satisfies $f_1 + f_0 \in \Omega_0$, that is, $f_1 \in \Omega_0 \cap (\Omega_0 - f_0)$. By Proposition A.3 and because f_0 is continuous, Ω_0 is also a one-set w.r.t. $\mathbb{P}^{f+f_0}$, so $\mathbb{P}^f(\Omega_0 - f_0) = \mathbb{P}^{f+f_0}(\Omega_0) = 1$, so both Ω_0 and $\Omega_0 - f_0$ and thus $\Omega_0 \cap (\Omega_0 - f_0)$ are one-sets w.r.t $\mathbb{P}^f$. In particular, the latter set is non-empty and we find the desired f_1, which finishes the proof of the first part.

We now turn to the second assertion, for which we drop our assumption $m = 0$. It will be convenient to define a metric on $C(T, \mathbb{R})$ that induces the correct topology, since it will give us a concrete description of open neighbourhoods of f_0. One such metric can be obtained by taking compact sets $C_1 \subset C_2 \subset \ldots$ with $\bigcup_{i=1}^\infty C_i = T$ (recall that T is σ-finite) and putting

$$d(x, y) := \sum_{i=1}^\infty 2^{-i} \left(1 \wedge \sup_{t \in C_i} |x(t) - y(t)|\right), \quad x, y \in C(T, \mathbb{R}),$$

where $a \wedge b = \min(a, b)$. Indeed, one easily checks that d is a metric, and that $d(x_n, x) \to 0$ iff $x_n \to x$ uniformly on C_i for all $i \in \mathbb{N}$, iff $x_n \to x$ uniformly on every compact set. Denote by $\Pi_n = \mathcal{GP}(\widehat{f}_n(\cdot; f_0), k_n(\cdot, \cdot; \boldsymbol{t}_n))$ the distribution of the posterior GP, and let $g_n \sim \Pi_n$, defined on some probability space $(F, \mathcal{F}, Q)$. It is important to note here that we have two nested probability spaces: First there are random realizations of the sampling noise ε and points (t_i), governed by the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then, for fixed $\omega \in \Omega$ (that is, for fixed ε and (t_i)), the posterior GP is itself a random distribution Π_n on $\mathbb{R}^T$ that depends on ω. To clarify the distinction, we will use Q and $\mathbb{E}_Q$ to denote probabilities and expectation w.r.t. the posterior GP for fixed ε and (t_i).

Put $\overline{g}_n \coloneqq g_n - \widehat{f}_n(\cdot; t_0) \overset{Q}{\sim} \mathcal{GP}(0, k_n(\cdot, \cdot; \boldsymbol{t}_n))$, so that by Theorem 14, and since $J(C, d_n) \leq J(C, d_0) < \infty$ by Lemma 15, we may choose g_n such that $\overline{g}_n$ takes values in $C(T, \mathbb{R})$, and hence also g_n, because $\widehat{f}_n(\cdot; t_0)$ is continuous by definition and continuity of m and K.

It now suffices to show that, for fixed but arbitrary $\delta > 0$, $\Pi_n(B(f_0, \delta)) = Q(d(g_n, f_0) < \delta) \xrightarrow{L^1} 1$. Let $i_0 \in \mathbb{N}$ such that $\sum_{i > i_0} 2^{-i} < \delta/2$, so that $d(g_n, f_0) < \delta$ is implied by $\sum_{i=1}^{i_0} 2^{-i} \left(1 \wedge \sup_{t \in C_i} |g_n(t) - f_0(t)|\right) < \delta/2$, which is implied by $\sup_{t \in C_{i_0}} |g_n(t) - f_0(t)| < \delta/2$. Put $C \coloneqq C_{i_0}$. Then,

$$\sup_{t \in C} |g_n(t) - f_0(t)| \leq \sup_{t \in C} |\overline{g}_n(t)| + \sup_{t \in C} \left|\widehat{f}_n(t; f_0) - f_0(t)\right|,$$

and hence

$$\begin{aligned} Q\left(d(g_n, f_0) \geq \delta\right) &\leq Q\left(\sup_{t \in C} |g_n(t) - f_0(t)| \geq \delta/2\right) \\ &\leq Q\left(\sup_{t \in C} |\overline{g}_n(t)| \geq \delta/4\right) + \underbrace{\mathbb{1}\left\{\sup_{t \in C} \left|\widehat{f}_n(t; f_0) - f_0(t)\right| \geq \delta/4\right\}}_{\leq \frac{4}{\delta} \sup_{t \in C} \left|\widehat{f}_n(t; f_0) - f_0(t)\right|}. \end{aligned} \tag{23}$$

Since $\overline{g}_n = g_n - \widehat{f}_n(\cdot; f_0) \overset{Q}{\sim} \mathcal{GP}(0, k_n(\cdot, \cdot; \boldsymbol{t}_n))$, we may apply Theorem 14 (ii) and Lemma 15 to obtain, for some fixed $t_0 \in C$,

$$\begin{aligned} Q\left(\sup_{t \in C} |\overline{g}_n(t)| \geq \delta/4\right) &\leq \frac{4}{\delta} \mathbb{E}_Q\left[\sup_{t \in C} |\overline{g}_n(t)|\right] \\ &\leq \frac{4}{\delta}\left(\mathbb{E}_Q\left[|\widehat{g}_n(t_0)|\right] + \mathbb{E}_Q\left[\sup_{t \in C} |\overline{g}_n(t) - \overline{g}_n(t_0)|\right]\right) \\ &\leq \frac{4}{\delta}\left(\sqrt{\widehat{v}_n(t_0)} + cJ(C, d_n)\right), \end{aligned} \tag{24}$$

where we used that $\mathbb{E}_Q\left[|\overline{g}_n(t_0)|\right] \leq \sqrt{\mathbb{E}_Q\left[\overline{g}_n(t_0)^2\right]} = \sqrt{\mathbb{V}(\overline{g}_n(t_0))} = \sqrt{\widehat{v}_n(t_0)}$. Combining Equations (23) and (24), we conclude that

$$\begin{aligned} \mathbb{E}\left[\Pi_n(B(f_0, \delta))^c\right] &= \mathbb{E}\left[Q\left(d(g_n, f_0) \geq \delta\right)\right] \\ &\leq \frac{4}{\delta}\left(\mathbb{E}\left[\sqrt{\widehat{v}_n(t_0)}\right] + c\mathbb{E}\left[J(C, d_n)\right] + \mathbb{E}\left[\sup_{t \in C}\left|\widehat{f}_n(t; f_0) - f_0(t)\right|\right]\right) \longrightarrow 0, \end{aligned}$$

where the first term goes to zero by Proposition 10 and monotone convergence, the second by Lemma 15 and dominated convergence (where $J(C, d_n)$ is dominated by $J(C, d_0) < \infty$), and the third by the first part of Theorem 8, which has already been proven. ■

We close with proofs of Propositions 4 and 7.

Proof of Proposition 4

(i) If T has no isolated points, then definitions of recurrently dense and dense obviously coincide. Now suppose T is connected and (s_n) is a dense sequence. If $T = \{t\}$ is a singleton, then t is an isolated point and $s_n = t$ for all $n \in \mathbb{N}$, so (s_n) is recurrently dense. Otherwise, T cannot have isolated points. Indeed, that would imply that $\{t\}$ is open for some $t \in T$, but then $T = \{t\} \cup (T \setminus \{t\})$ would be the disjoint union of two non-empty open sets.

(ii) Since T is separable, its topology admits a countable base $\{O_j : j \in \mathbb{N}\}$. We have to show that almost surely, every O_j contains infinitely many t_i. Let $O \in \{O_j : j \in \mathbb{N}\}$ fix and set

$$A_i := \{t_i \in O\}, \quad i \in \mathbb{N},$$
$$A := \limsup_{i\to\infty} A_i = \{t_i \in O \text{ for infinitely many } i \in \mathbb{N}\}.$$

We have to show $\mathbb{P}(A) = 1$, for which we employ the Borel–Cantelli lemma. Clearly, the events A_i, $i \in \mathbb{N}$, are independent, hence it remains to verify that the sequence $(\mathbb{P}(A_i))$ is not summable. This is obvious, however, since

$$\mathbb{P}(A_i) = \mathbb{P}(t_i \in O) = \mathbb{P}(t_1 \in O) =: p > 0$$

for all $i \in \mathbb{N}$ by assumption and thus $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} p = \infty$. Since there are only countably many O_j, we conclude

$$\mathbb{P}\left(\exists j \in \mathbb{N}\colon\ |\{t_i : i \in \mathbb{N}\} \cap O_j| < \infty\right) \leq \sum_{j=1}^{\infty} \underbrace{\mathbb{P}\left(|\{t_i : i \in \mathbb{N}\} \cap O_j| < \infty\right)}_{=0} = 0.$$

■

Proof of Proposition 7 Recall Definition B.1, and fix any $d \in (\dim C, \infty)$. Then, there is an $\varepsilon_0 > 0$ such that $\log N(C, \varepsilon, \rho)/(-\log \varepsilon) \leq d$ and thus $N(C, \varepsilon, \rho) \leq \varepsilon^{-d}$ for all $\varepsilon < \varepsilon_0$.

By assumption, K is locally Lipschitz and hence Lipschitz on C, so there is a $c > 0$ such that, for $s, t \in C$,

$$d_K(s,t)^2 \leq |K(t,t) - K(t,s)| + |K(t,s) - K(s,s)| \leq c\rho(t,s).$$

This implies that

$$N(C, \varepsilon, d_K) \leq N(C, \sqrt{\varepsilon/c}, \rho) \leq (\sqrt{\varepsilon/c})^{-d} = c'\varepsilon^{-d/2}$$

for any $\varepsilon \in (0, \varepsilon_0)$, so $\log N(C, \varepsilon, d_K) = O\left(\log\left(\varepsilon^{-1}\right)\right)$ for small $\varepsilon > 0$. Since convergence of the Dudley integral is determined by its convergence at zero (Bogachev, 1998, p. 334), this implies $J(C, d_K) < \infty$. ■

7. Conclusion and Outlook

In this paper, we made a crucial first step towards the theoretical legitimization of using GP regression on non-Euclidean manifolds and other metric domain spaces, as has already been done by various authors. However, for all of our results to be applicable, one must verify that the difference between the unknown function and the GP prior's mean is (expected to be) in the RKHS of the prior's kernel. This may be challenging because explicit descriptions of RKHSes are available only for few kernels, such as the Laplacian kernel (Berlinet and Thomas-Agnan, 2011). Hence, further research into RKHSes of common kernels is desirable, especially in the context of metric domain spaces. A first step would be to examine popular stationary kernels with known RKHS on $\mathbb{R}$ when defined over a metric space by replacing the absolute difference with the metric. Another approach to this problem was suggested by Shi and Choi (2011), who used a parameterized exponential kernel and assigned a prior to the parameter. This way, the kernel was not fixed, and they could circumvent the assumption that the unknown function must be contained in a specific RKHS. It appears plausible that a similar approach could be used in more general settings, and further research in this direction may be fruitful. Lastly, it seems reasonable to conjecture that the result presented here can be generalized to apply whenever the unknown function is contained in the support of the GP prior, of which the RKHS is a dense subset. If this is not true, it would be interesting to characterize or bound the error of the posterior mean in that case.

While we considered (possibly random) sampling sequences in this paper, we assumed the positions to be precisely known to the observer. Another potential area of future research is to investigate whether it is possible to adapt the model to account for uncertainties in the sampling points (as has been suggested by Dallaire et al., 2009) and inspect if the presented results remain true in that setting.

Another assumption we made is that the variance of the measurement noise is known. In many practical applications, the variance is not known accurately, motivating the question whether the presented result remains valid in the case where the assumed variance of the noise differs from its true variance.

Finally, this work aimed to establish asymptotic consistency of GP regression under as weak as possible assumptions on the observed function and the GP prior, allowing us to provide results for a large class of domains and kernels. Future work may examine additional assumptions necessary to prove more specific results on error bounds and rates of convergence.

Appendix A. Reproducing Kernel Hilbert Spaces

Here, we introduce the notion of the reproducing kernel Hilbert space (RKHS) of a kernel, which is a special case of the more general concept of a Cameron–Martin space. As is covered in detail by a work of Bogachev (1998), the Cameron–Martin space is a Hilbert space associated with a Gaussian measure, which is a generalization of a Gaussian distribution to locally convex topological vector spaces. As we will see shortly, a GP defines a Gaussian measure on $\mathbb{R}^T$, and the arising Cameron–Martin space is then called the RKHS of the GP's kernel.

We briefly introduce the notion of a Gaussian measure as it is considered by Bogachev (1998). A *topological vector space* is a real vector space, endowed with a topology with respect to which scalar multiplication and addition are continuous. It is called *locally convex* if every neighbourhood of zero contains a convex neighbourhood of zero. If X is a locally convex topological vector space, then $\mathcal{E}(X)$ denotes the σ-algebra generated by *cylinders*, which are sets of the form

$$\{x \in X\colon\ (\ell_1(x), \dots, \ell_n(x)) \in B\} \subset X,$$

where $n \in \mathbb{N}$, $\ell_1, \dots, \ell_n \in X^\star$, and $B \in \mathcal{B}(\mathbb{R}^n)$. Here, $X^\star$ denotes the topological dual space of X, which consists of all continuous linear functionals on X. A *Gaussian measure* on X is a probability measure γ defined on $\mathcal{E}(X)$ such that $\gamma \circ \ell^{-1}$ is a Gaussian distribution in $\mathbb{R}$ for every $\ell \in X^\star$.

We now show that a GP f on T defines a Gaussian measure on $\mathbb{R}^T$ in the above sense. It is elementary to confirm that $\mathbb{R}^T$ (endowed with the product topology) is a locally convex topological vector space. Furthermore, it is known that the dual space of $\mathbb{R}^T$ consists of functionals of the form

$$\ell = \sum_{i=1}^n a_i \pi_{t_i}, \quad n \in \mathbb{N}, a_1, \dots, a_n \in \mathbb{R}, t_1, \dots, t_n \in T,$$

where $\pi_t\colon \mathbb{R}^T \to \mathbb{R}$ for $t \in T$ denotes the projection map $f \mapsto f(t)$. Hence, $\mathcal{E}(\mathbb{R}^T)$ is the σ-algebra generated by the projection maps π_t for $t \in T$, with respect to which f is measurable by assumption. The distribution of f thus defines a probability measure on $\mathcal{E}(\mathbb{R}^T)$. Now if $\ell = \sum_{i=1}^n a_i \pi_{t_i} \in (\mathbb{R}^T)^\star$ for some $n \in \mathbb{N}$, $a_1, \dots, a_n \in \mathbb{R}$, and $t_1, \dots, t_n \in T$, then

$$\ell(f) = \sum_{i=1}^n a_i \pi_{t_i}(f) = \sum_{i=1}^n a_i f(t_i),$$

which follows a Gaussian distribution, since $(f(t_1), \dots, f(t_n))$ does by assumption on f.

We have shown that the distribution of a GP defines a Gaussian measure in the sense introduced above. Thus, we may apply the concept of a Cameron–Martin space as introduced by Bogachev (1998, p. 44) to this setting, resulting in the following definition.

Definition A.1 *Let $K\colon T \times T \to \mathbb{R}$ be symmetric and positive definite. Then the* Cameron–Martin space of $\mathcal{GP}(0, K)$*, also called the* reproducing kernel Hilbert space (RKHS) of K*, is given by*

$$\begin{aligned} H &:= \{x\colon T \to \mathbb{R}\colon\ \|x\|_H < \infty\}, \\ \|x\|_H &:= \sup\left\{\langle \boldsymbol{a}, x(\boldsymbol{t})\rangle : n \in \mathbb{N}, \boldsymbol{a} \in \mathbb{R}^n, \boldsymbol{t} \in T^n, \boldsymbol{a}^\top K(\boldsymbol{t}, \boldsymbol{t})\boldsymbol{a} \le 1\right\}, \quad x \in \mathbb{R}^T. \end{aligned}$$

There exists a scalar product $\langle\cdot,\cdot\rangle_H$ that induces $\|\cdot\|_H$, equipped with which H becomes a Hilbert space.

Recall that two probability measures are called *equivalent* if they are absolutely continuous with respect to each other, that is, if they have the same sets of measure zero. Two random variables are called *equivalent (in distribution)* if their distributions are equivalent. One remarkable property of the RKHS is given by the following proposition.

Proposition A.2 *Let H be the RKHS associated with a kernel K. If $g \in H$ and $f \sim \mathcal{GP}(0, K)$, then the distributions of f and $f + g$ on $\mathcal{E}(\mathbb{R}^T)$ are equivalent.*

Proof Follows from a result of Bogachev (1998, Cor. 2.4.3). ■

Proposition A.3 *Suppose that T is σ-compact and equipped with the topology of uniform convergence on compacts, and let H be the RKHS associated with a kernel K. If $g \in H \cap C(T, \mathbb{R})$, and $f \sim \mathcal{GP}(0, K)$ takes values in $C(T, \mathbb{R})$, then the distributions of f and $f + g$ on the Borel σ-algebra $\mathcal{B}(C(T, \mathbb{R}))$ of $C(T, \mathbb{R})$ are equivalent.*

Proof Abbreviate $C \coloneqq C(T, \mathbb{R})$ for this proof, and denote by P and Q the distributions of f and $f + g$ on $\mathcal{E} \coloneqq \mathcal{E}(\mathbb{R}^T)$, respectively, which are equivalent by Proposition A.2. It is a standard fact that $\mathcal{B}(C) = \mathcal{E} \cap C$. Hence, if $A \in \mathcal{B}(C)$ is such that $\mathbb{P}^f(A) = 0$, then there is an $\widetilde{A} \in \mathcal{E}$ with $\widetilde{A} \cap C = A$, so

$$P(\widetilde{A}) = \mathbb{P}(f \in \widetilde{A}) \stackrel{(\star)}{=} \mathbb{P}(f \in \widetilde{A} \cap C) = \mathbb{P}(f \in A) = \mathbb{P}^f(A) = 0, \tag{25}$$

where we used in $(\star)$ that f takes values in C. This implies that $Q(\widetilde{A}) = 0$, and by the same argument (for which we need that $f + g$ takes values in C by continuity of g), $\mathbb{P}^{f+g}(A) = Q(\widetilde{A}) = 0$. This shows that $\mathbb{P}^{f+g}$ is absolutely continuous w.r.t. $\mathbb{P}^f$, and the converse follows the same way. ■

Appendix B. Minkowski Dimension

We give a brief introduction to Minkowski dimension, which is also known as Kolmogorov or upper box counting dimension.

Definition B.1 *Let $C \subset T$ be compact. For $\varepsilon > 0$, denote by $N(C, \varepsilon) \coloneqq N(C, \varepsilon, \rho) \in \mathbb{N}$ the minimal number of balls with radius ε required to cover C. Then the* Minkowski dimension *of C is*

$$\dim C \coloneqq \overline{\lim_{\varepsilon \to 0}} \frac{\log N(C, \varepsilon)}{-\log \varepsilon}. \tag{26}$$

Loosely speaking, $\dim C = d$ means that $N(C, \varepsilon) \approx \varepsilon^{-d}$. Note that this notion only makes sense if $N(C, \varepsilon) < \infty$ for all $\varepsilon > 0$, which is why it is only defined for compact sets. However, this can naturally be extended by putting

$$\dim T \coloneqq \sup \{\dim C \colon C \subset T \text{ is compact}\}, \tag{27}$$

which is consistent with Equation (26) if T itself is compact. It is an easy fact that $\dim \mathbb{R}^d = d$ for all $d \in \mathbb{N}$, and in fact for any reasonably nice space this notion coincides with other definitions of dimension and takes the value one would expect. We make this a bit more explicit in the case of manifolds.

Lemma B.2 *If T is a connected Riemannian n-manifold and ρ the geodesic distance, then* $\dim T = n$.

Proof By Equation (27), we may assume that T is compact. In that case, T is Ahlfor's n-regular (Tholozan, 2021), that is, there exists a Borel measure ν on T and constants $c_1, c_2 > 0$ such that $c_1 r^n \leq \nu(B(x, r)) \leq c_2 r^n$ for all $x \in E$ and $r \in (0, r_0]$, where r_0 is the diameter of T. Let $\varepsilon > 0$, and $(B(x_i, \varepsilon/2))_{i=1}^m$ be a family of minimal size among all collections of disjoint balls with radius ε. Then, $(B(x_i, \varepsilon))_{i=1}^m$ covers T (assuming the opposite would contradict minimality), and

$$c_2 r_0^n \geq \nu(T) \geq \sum_{i=1}^m \nu(B(x_i, \varepsilon/2)) \geq m \cdot c_1 (\varepsilon/2)^n,$$

so $N(T, \varepsilon) \leq m \leq c_3 \varepsilon^{-n}$ with $c_3 = c_2 (2r_0)^n$, and hence

$$\dim T = \varlimsup_{\varepsilon \to 0} \frac{\log N(T, \varepsilon)}{-\log \varepsilon} \leq \varlimsup_{\varepsilon \to 0} \left(n + \frac{\log c_3}{-\log \varepsilon} \right) = n.$$

The lower bound follows by a similar argument and is of less relevance to us, so we omit it. ■

Lemma B.3 *If T is a manifold, then it can be metrized in a way that* $\dim T < \infty$.

Proof Again, we may assume that T is compact. By a theorem due to Hurewicz and Wallman (2015), T is homeomorphic to a compact subset E of $\mathbb{R}^{2n+1}$, where $n \in \mathbb{N}$ denotes the covering dimension of T. If $\phi \colon T \to E$ denotes the homeomorphism, then $\rho(s, t) := |\phi(t) - \phi(s)|$ for $s, t \in T$ defines a metric on T which induces the correct topology (since ϕ is a homeomorphism). Since ϕ is now an isometry between the metric spaces (T, ρ) and $(E, |\cdot - \cdot|)$, the Minkowski dimension of T is the same as that of E, in particular no more than $2n + 1$. ■

References

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science & Business Media, 2011.

Vladimir I. Bogachev. *Gaussian Measures.* Mathematical Surveys and Monographs. American Mathematical Society, 1998. ISBN 9780821810545.

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc P. Deisenroth. Matérn Gaussian processes on Riemannian manifolds. *arXiv preprint arXiv:2006.10160*, 2020.

Martin R. Bridson and André Haefliger. *Metric Spaces of Non-Positive Curvature.* Springer Science & Business Media, 1999.

Roberto Calandra, Jan Peters, Carl E. Rasmussen, and Marc P. Deisenroth. Manifold Gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3338–3345. IEEE, 2016.

Taeryon Choi and Mark J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007.

Patrick Dallaire, Camille Besse, and Brahim Chaib-Draa. Learning Gaussian process models from uncertain data. In *International Conference on Neural Information Processing*, pages 433–440. Springer, 2009.

Maxim Dolgov and Uwe D. Hanebeck. A distance-based framework for Gaussian processes over probability distributions. *arXiv preprint arXiv:1809.09193*, September 2018. URL `https://arxiv.org/abs/1809.09193`.

Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967. ISSN 0022-1236. doi: https://doi.org/10.1016/0022-1236(67)90017-1.

Aasa Feragen, Francois Lauze, and Søren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3042, 2015.

Subhashis Ghosal, Anindya Roy, et al. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006.

Jayanta K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer Science & Business Media, 2003.

Jayanta K. Ghosh, M. Delampady, and T. Samanta. *An Introduction to Bayesian Analysis Theory and Method*, volume 279. Springer, 2006.

Marcin Hitczenko and Michael L. Stein. Some theory for anisotropic processes on the sphere. *Statistical Methodology*, 9(1-2):211–227, 2012.

Witold Hurewicz and Henry Wallman. *Dimension Theory (PMS-4), Volume 4*. Princeton University Press, 2015.

Rodney A. Kennedy and Parastoo Sadeghi. *Hilbert Space Methods in Signal Processing*. Cambridge University Press, 2013. ISBN 978-0-511-84451-5. doi: 10.1017/CBO9780511844515.

Murat Kumru and Emre Özkan. 3-D extended object tracking using recursive Gaussian processes. In *2018 21st International Conference on Information Fusion (FUSION)*, 2018. doi: 10.23919/ICIF.2018.8455480.

Muriel Lang, Oliver Dunkley, and Sandra Hirche. Gaussian process kernels for rotations and 6-D rigid body motions. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5165–5170. IEEE, 2014.

John M. Lee. *Smooth manifolds*. Springer, 2nd edition, 2013.

Lizhen Lin, Niu Mu, Pokman Cheung, and David Dunson. Extrinsic Gaussian processes for regression and classification on manifolds. *Bayesian Analysis*, 14(3):887–906, 2019.

Kalyanapuram R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 2014.

Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, USA, 2006.

Jian Qing Shi and Taeryon Choi. *Gaussian Process Regression Analysis for Functional Data*. CRC Press, 2011.

Nicolas Tholozan. Compact connected Riemannian manifolds are Ahlfors regular metric space. MathOverflow, 2021. URL `https://mathoverflow.net/q/402078`. Accessed: 2021-10-01.

Surya T. Tokdar and Jayanta K. Ghosh. Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference*, 137(1):34–42, 2007.

Aad W. van der Vaart and J. Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K.*, pages 200–222. Institute of Mathematical Statistics, 2008.

Niklas Wahlström and Emre Özkan. Extended target tracking using Gaussian processes. *IEEE Transactions on Signal Processing*, 63(16):4165–4178, 2015.

Norbert Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1949.

Fuzhen Zhang. *The Schur Complement and Its Applications*, volume 4 of *Numerical Methods and Algorithms*. Springer Science & Business Media, 2006.

Wenxin Zhou. Math 281c: Mathematical statistics, lecture 7. `http://www.math.ucsd.edu/~xip024/Teaching/Math281C_Spring2020/Math281C_2020.html`, 2020. Accessed: 2021-10-01.