Sample Path Regularity of Gaussian Processes from the Covariance Kernel

Nathaël Da Costa[1] nathael.da-costa@uni-tuebingen.de
Marvin Pförtner[1] marvin.pfoertner@uni-tuebingen.de
Lancelot Da Costa[2,3,4] lance.dacosta@verses.ai
Philipp Hennig[1] philipp.hennig@uni-tuebingen.de

[1] *Tübingen AI Center, University of Tübingen, Tübingen, Germany*
[2] *VERSES AI Research Lab, Los Angeles, USA*
[3] *Department of Mathematics, Imperial College, London, UK*
[4] *Wellcome Centre for Human Neuroimaging, UCL, London, UK*

Abstract

Gaussian processes (GPs) are the most common formalism for defining probability distributions over spaces of functions. While applications of GPs are myriad, a comprehensive understanding of GP sample paths, i.e. the function spaces over which they define a probability measure, is lacking. In practice, GPs are not constructed through a probability measure, but instead through a mean function and a covariance kernel. In this paper we provide necessary and sufficient conditions on the covariance kernel for the sample paths of the corresponding GP to attain a given regularity. We use the framework of Hölder regularity as it grants particularly straightforward conditions, which simplify further in the cases of stationary and isotropic GPs. We then demonstrate that our results allow for novel and unusually tight characterisations of the sample path regularities of the GPs commonly used in machine learning applications, such as the Matérn GPs.

Keywords: Gaussian processes, Gaussian random fields, differentiability, Hölder regularity, Sobolev regularity, stationary kernels, isotropic kernels, Matérn kernels

1 Introduction

Gaussian processes (GPs) provide a formalism to assign probability distributions over spaces of functions. That distribution is principally characterised and controlled by the process' covariance function, which is a positive definite kernel. Such kernels are also associated with reproducing kernel Hilbert spaces (RKHSs). However, it is relatively widely known that the *sample paths* of the associated GP are not generally elements of the RKHS, but form a "larger" space of typically less regular functions (Kanagawa et al., 2018, Section 4). This sample path space is *much harder to characterise* than the RKHS.

GP regression is widely applied in statistics and machine learning for inference from physical observations. In such settings, practitioners mostly concern themselves with the posterior mean function (which *is* an element of the RKHS) and the marginal variance (which is *not* in the RKHS, but inherits its regularity) while largely ignoring the regularity of the sample paths and other properties of the support of the prior probability measure. But recently relevant use cases for GPs in computational tasks urgently require a more careful, and ideally tight analysis of the sample path regularity, which we provide in this work. For instance, Pförtner et al. (2022) recently showed that a large class of classic solution methods for linear partial differential equations (PDEs) can be interpreted as GP inference, i.e. as base instances of probabilistic numerical methods. To infer the solution to a PDE using a probabilistic numerical method, we want to construct a GP prior for it. It should be tailored to the problem as tightly as possible, for the following reasons:

1. We need to ensure that the PDE's differential operator is well-defined on all samples of the GP. Hence, the sample paths must be **regular enough**. Here, regularity typically refers to the existence of a number of strong or weak partial derivatives, which is encapsulated in the frameworks of Hölder and Sobolev regularity respectively.

2. But we also want the GP posterior to provide a useful uncertainty estimate over the solution of the PDE. Hence, we do not want to needlessly impose *additional* regularity constraints on the sample paths. The sample paths should be **as irregular as possible**, to avoid overconfidence. Such cautious models may also be advisable on numerical grounds, to avoid instabilities such as Gibbs or Runge phenomena.

The above is just one example of a setting in which one would like to characterise the regularity of GP sample paths as tightly as possible. Further examples include generalised coordinates, a method that underlies the analysis and filtering of stochastic differential equations driven by noise admitting a high number of derivatives (Friston et al., 2023; Heins et al., 2023). These noise signals are usually modelled with GPs, and hence characterising GP sample path regularity is crucial for understanding in which situations these methods can be applied.

1.1 Summary of Contributions

A consequence of our main result, Theorem 7, may be written as follows:

Corollary 1 *Let* $k\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *be a symmetric positive definite kernel. If either*

- *all partial derivatives of the form* $\frac{\partial^{2n} k}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}} \partial \boldsymbol{y}^{\boldsymbol{\alpha}}}$ *for multi-indices* $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ *with* $|\boldsymbol{\alpha}| = n$ *exist and are (locally) Lipschitz,*

- *for stationary* $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y})$, $\frac{\partial^{2n} k_\delta}{\partial \boldsymbol{x}^{\boldsymbol{\alpha}}}$ *for* $|\boldsymbol{\alpha}| = 2n$ *exist and are (locally) Lipschitz,*

- *for isotropic* $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$, $k_r^{(2n)}$ *exists and is (locally) Lipschitz,*

then the sample paths of $f \sim \mathcal{GP}(0, k)$ *are* n *times continuously differentiable.*

Note that, by the mean value theorem, the existence of one additional continuous derivative of k is sufficient for local Lipschitz continuity of the lower derivatives.

The actual Theorem 7 is sharper than this. It provides necessary and sufficient conditions on the regularity of the kernel for the sample paths to attain a given Hölder regularity. Just as the above corollary, the theorem encompasses statements for stationary and isotropic GPs.

We apply Theorem 7 in particular to the Matérn GPs to obtain Proposition 10, a consequence of which is the following:

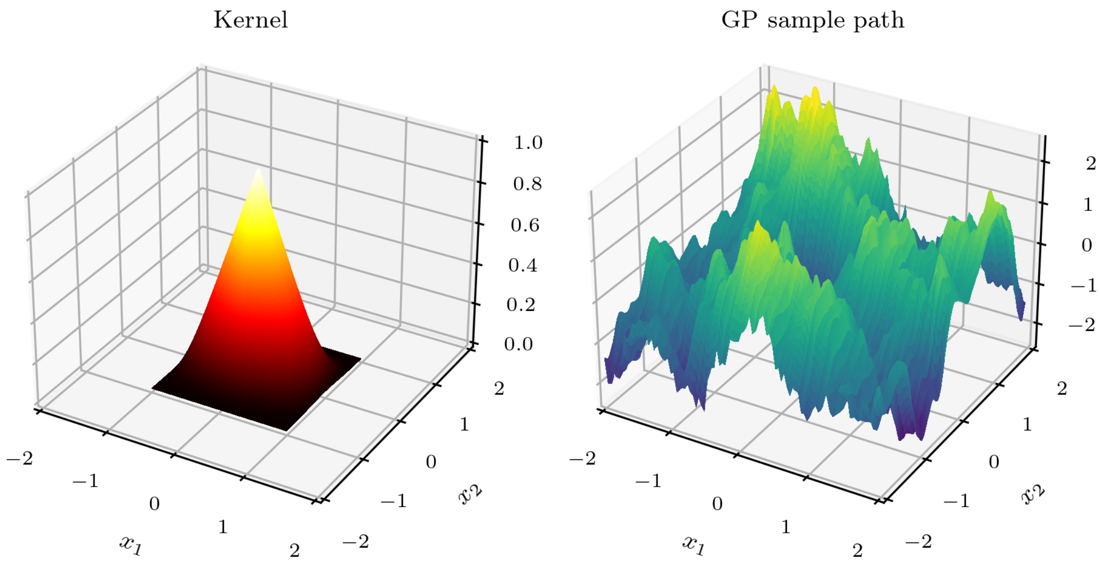


Figure 1: *Left:* surface plot of a kernel defined as the tensor product of two Wendland kernels (with $d = 1$, $n = 0$ on the x_1-axis and $d = 1$, $n = 1$ on the x_2-axis). *Right:* surface plot of a sample path from the corresponding centered GP. The sample path displays different regularity along each axis (non-differentiable along the x_1-axis; once differentiable along the x_2-axis), reflecting the rotational asymmetry in the regularity of the kernel.

Corollary 2 *The sample paths of a centered Matérn GP with smoothness parameter* $\nu \notin \mathbb{N}_0$ *are* $\lfloor \nu \rfloor$ *times continuously differentiable and no more. In particular, when* $\nu = n + 1/2$ *for some* $n \in \mathbb{N}_0$*, the sample paths are* n *times continuously differentiable and no more.*

We then obtain similar results for other GPs, such as the Wendland (Proposition 12), squared exponential and rational quadratic GPs (Remark 14).

In Section 4.6 we describe how to apply Theorem 7 to covariance kernels constructed through algebraic transformations of other kernels, including conic combinations (Proposition 17), products (Proposition 18), tensor products (Proposition 21) and coordinate transformations (Proposition 23). We note in Section 4.7 how the results in this work may be generalised to manifold GPs.

Theorem 25 summarises results about GP sample path Sobolev regularity. Comparing these with Theorem 7, we argue that one should not expect the existence of a greater number of weak (Sobolev) derivatives than of strong derivatives for GP sample paths. This provides further evidence that Theorem 7 contains all that is needed in practice.

We then plot a number of stationary kernel functions as well as sample paths from the corresponding centered GPs, as in Figure 1.

All regularity properties considered in this work are *local*, in the sense that they capture the infinitesimal behaviour of functions around each point in their domain. Note that our proofs can be extended and our assumptions strengthened to consider global regularity properties, such as global Hölder continuity.

1.2 Related Work

Some of the earliest work on sample path regularity can be attributed to Kolmogorov and Chentsov (1956). Later, Fernique (1975) proved the first necessary and sufficient conditions for sample path continuity of stationary GPs. Since then, many authors have extended these continuity results. Potthoff (2009) provides an approachable overview of how such results can be obtained. It also investigates conditions for uniform continuity, which is a *global* regularity property. Azmoodeh et al. (2014), which we rely on in this work, proves a converse to the Kolmogorov continuity theorem for GPs, providing necessary and sufficient conditions for Hölder continuity.

Differentiability has also been investigated by many authors, including Scheuerer (2010b) for general random fields, Adler and Taylor (2007), Potthoff (2010) and Henderson (2024) for Sobolev regularity. The sample path regularity of Matérn GPs was studied in Scheuerer (2010a) (see specifically Scheuerer (2010a, Examples 5.3.19, 5.5.6 & 5.5.12)) and of tensor product of one dimensional Matérn GPs in Wang et al. (2021). Pierce (1970, Proposition I.3) combines differentiability results with the Kolmogorov continuity theorem, and can be seen as the global version of the forward implication of our Theorem 7 (1).

Compared to the existing literature, the present work provides *general*, *versatile*, *easily applicable* and *unusually tight* results for the study of GP sample path continuity and differentiability.

2 Preliminaries

Definition 3 *A Gaussian process on a set O is a map $f\colon O\times\Omega\to\mathbb{R}$, where Ω is a probability space, such that for all $x\in O$, $f(x,\cdot)\colon O\to\mathbb{R}$ is measurable, and such that for each $X := (x_1,\dots,x_N)\in O^N$, the map $f(X,\cdot)\colon\Omega\to\mathbb{R}^N$ given by $f(X,\omega)=(f(x_1,\omega),\dots,f(x_N,\omega))$ is a multivariate Gaussian random variable.*

The maps $f(\cdot,\omega)\colon O\to\mathbb{R}$ for $\omega\in\Omega$ are the sample paths of the GP.

We would like to study the regularity of the sample paths of a GP. However, Definition 3 is not convenient to work with in practice: one rarely constructs the probability space Ω of a GP. Instead, one characterises a GP by its *mean* and its *covariance kernel*.

Definition 4 *The mean μ of a GP $f\colon O\times\Omega\to\mathbb{R}$ is the map*

$$\mu\colon O\to\mathbb{R},\ x\mapsto\mathbb{E}(f(x,\cdot)).$$

The covariance kernel k is the map

$$k\colon O\times O\to\mathbb{R},\ (x,y)\mapsto\operatorname{cov}(f(x,\cdot),f(y,\cdot)).$$

The GP f is said to be centered if $\mu=0$.

Conversely, given maps $\mu\colon O\to\mathbb{R}$ and $k\colon O\times O\to\mathbb{R}$ with k symmetric positive definite – meaning that for any finite set of points $\{x_1,\dots,x_N\}\subset\Omega$, the matrix $(k(x_i,x_j))_{i,j}$ is symmetric positive semi-definite – one can construct a probability space Ω, and a GP $f\colon O\times\Omega\to\mathbb{R}$ with mean μ and covariance kernel k (Klenke, 2014, Theorem 14.36). We then write $f\sim\mathcal{GP}(\mu,k)$.

At this point it is important to note that $f \sim \mathcal{GP}(\mu, k)$ does not uniquely specify f or the probability space Ω. However it uniquely specifies the finite dimensional distributions of f – since multivariate Gaussians are entirely characterised by their first two moments. In machine learning applications all that will ever be observed are evaluations of f at finitely many points. Therefore we would like to study sample path regularity of f *up to modification*. By this we mean that $f \sim \mathcal{GP}(\mu, k)$ will be said to have samples in $\mathcal{F}(O)$, where $\mathcal{F}(O)$ is some space of functions $O \to \mathbb{R}$, if there exists a construction of the GP f, say $f \colon O \times \Omega \to \mathbb{R}$, such that $f(\cdot, \omega) \in \mathcal{F}(O)$ for all $\omega \in \Omega$.

Finally, note that $f \sim \mathcal{GP}(\mu, k)$ having samples in $\mathcal{F}(O)$ is equivalent to $\tilde{f} + \mu$ having samples in $\mathcal{F}(O)$, where $\tilde{f} \sim \mathcal{GP}(0, k)$. For this it suffices that $\mu \in \mathcal{F}(O)$ and $\tilde{f}$ has samples in $\mathcal{F}(O)$. Therefore, in what follows, we will consider only centered GPs.

Thus we are studying sample path regularity of GPs *from the covariance kernel*, and our goal is to link the regularity of the kernel to the regularity of the GP sample paths.

2.1 Setup

We study GPs defined on open subsets of Euclidean spaces: $O \subset \mathbb{R}^d$ is open, for some $d \in \mathbb{N}$, and $f \sim \mathcal{GP}(0, k)$ is a centered GP on O. An important special case is when k is stationary, i.e. $O = \mathbb{R}^d$ and there is an even function $k_\delta \colon \mathbb{R}^d \to \mathbb{R}$ such that

$$k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y}) \tag{1}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. A further special case of k stationary is when k is isotropic, i.e. $O = \mathbb{R}^d$ and there is an even function $k_r \colon \mathbb{R} \to \mathbb{R}$ such that

$$k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|) \tag{2}$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where $\|\cdot\|$ denotes the standard Euclidean norm on $\mathbb{R}^d$.

3 Hölder Regularity

The main type of sample path regularity we would like to investigate is continuity and the order of continuous differentiability. These are encapsulated in the framework of *Hölder regularity.*

We choose O to be an open set in $\mathbb{R}^d$ in order to capture differentiability of the sample paths. However this implies that Hölder spaces on O are inconvenient for our purposes, as they do not solely characterise local regularity properties of functions on O. Indeed, the Hölder spaces not only constrain the local behaviour of functions, but also their behaviour at infinity, or near the boundary of O. To retain solely local constraints we employ local Hölder spaces.

Definition 5 (Local Hölder and almost-Hölder spaces) *Let $n \in \mathbb{N}_0$ and $\gamma \in [0, 1]$.*

(1) The local Hölder space $C^{n,\gamma}_{loc}(O)$ is the space of functions f on O for which $\partial^{\boldsymbol{\alpha}} f$ exists for all multi-indices $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_d \leq n$, and such that the highest order partial derivatives satisfy a Hölder condition of the form: for all compact subsets $K \subset O$ there is a constant $C_K > 0$ such that

$$|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})| \leq C_K \|\boldsymbol{x} - \boldsymbol{y}\|^\gamma$$

for all $\boldsymbol{x}, \boldsymbol{y} \in K$ *and* $|\boldsymbol{\alpha}| = n$.

(2) *The local almost-Hölder space* $C^{(n+\gamma)^-}_{loc}(O)$ *is defined as* $\bigcap_{n'+\gamma'<n+\gamma} C^{n',\gamma'}_{loc}(O)$.

Remark 6 *For* $n, n' \in \mathbb{N}_0$ *and* $\gamma, \gamma' \in [0,1]$, $n'+\gamma' < n+\gamma$ *implies* $C^{n',\gamma'}_{loc}(O) \supsetneq C^{(n+\gamma)^-}_{loc}(O) \supsetneq C^{n,\gamma}_{loc}(O)$. *Moreover* $C^n(O) = C^{n,0}_{loc}(O)$.

The key point in the definition of $C^{(n+\gamma)^-}_{loc}(O)$ *is the strict inequality* $n' + \gamma' < n + \gamma$, *since allowing for equality we have* $\bigcap_{n'+\gamma'\leq n+\gamma} C^{n',\gamma'}_{loc}(O) = C^{n,\gamma}_{loc}(O)$.

As will become clear, the local almost-Hölder spaces are natural for characterising GP sample path regularity.

To describe the regularity of the covariance kernel k, we will require derivatives of the form $\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k$. Here $\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}$ stands for the application of $\partial^{\boldsymbol{\alpha}}$ with respect to the first variable, followed by the application of $\partial^{\boldsymbol{\beta}}$ with respect to the second variable. In fact the order in which they are applied will not matter, by continuity of these partial derivatives. We moreover write $C^{n\otimes n}(O \times O)$ for the space of functions k on $O \times O$ for which $\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k$ exists and is continuous for all $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq n$.

Also recall the "big O" notation: the functions $f, g\colon \mathbb{R}^d\backslash\{\boldsymbol{0}\} \to \mathbb{R}$ satisfy $f(\boldsymbol{h}) = \mathcal{O}(g(\boldsymbol{h}))$ as $\boldsymbol{h} \to \boldsymbol{0}$ if and only if there is $C > 0$ such that $\limsup_{\boldsymbol{h}\to\boldsymbol{0}} |f(\boldsymbol{h})/g(\boldsymbol{h})| \leq C$. Moreover, for a family of functions $(f_{\boldsymbol{x}})_{\boldsymbol{x}\in O}$, we say $f_{\boldsymbol{x}}(\boldsymbol{h}) = \mathcal{O}(g(\boldsymbol{h}))$ as $\boldsymbol{h} \to \boldsymbol{0}$ locally uniformly in $\boldsymbol{x} \in O$ if, for every compact subset $K \subset O$, $C > 0$ may be chosen independently of $\boldsymbol{x} \in K$.

We are now in a position to state the main result of this paper.

Theorem 7 (Sample path Hölder regularity) *Let* $n \in \mathbb{N}_0$ *and* $\gamma \in (0,1]$. *The process* $f \sim \mathcal{GP}(0,k)$ *has samples in* $C^{(n+\gamma)^-}_{loc}(O)$ *if and only if,*

(1) *for general* k,

- $k \in C^{n\otimes n}(O \times O)$,
- $|\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}}k(\boldsymbol{x}, \boldsymbol{x})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as* $\boldsymbol{h} \to \boldsymbol{0}$, *locally uniformly in* $\boldsymbol{x} \in O$, *for all* $\epsilon \in (0,\gamma)$ *and* $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$.

(2) *for stationary* $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y})$,

- $k_\delta \in C^{2n}(\mathbb{R}^d)$,
- $|\partial^{\boldsymbol{\alpha}}k_\delta(\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}}k_\delta(\boldsymbol{0})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as* $\boldsymbol{h} \to \boldsymbol{0}$ *for all* $\epsilon \in (0,\gamma)$ *and* $|\boldsymbol{\alpha}| = 2n$.

(3) *for isotropic* $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$,

- $k_r \in C^{2n}(\mathbb{R})$,
- $|k_r^{(2n)}(h) - k_r^{(2n)}(0)| = \mathcal{O}(|h|^{2\epsilon})$ *as* $h \to 0$ *for all* $\epsilon \in (0,\gamma)$.

In each case, differentiating sample-wise we have $\partial^{\boldsymbol{\alpha}}f \sim \mathcal{GP}(0, \partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}}k)$ *for all* $|\boldsymbol{\alpha}| \leq n$.

The proof can be found in Appendix A.

Remark 8 *The positive definiteness of k allows us to extrapolate regularity: k, k_δ or k_r have no more derivatives around the diagonal $\{(\boldsymbol{x}, \boldsymbol{x}) \in O \times O : \boldsymbol{x} \in O\}$, $\mathbf{0}$ or 0 respectively as they do on the rest of the domain (see Lemma A.3). Therefore, when applying Theorem 7, it suffices to check the existence of the respective derivatives around the diagonal, $\mathbf{0}$ or 0 respectively.*

Another way to see that information about k in a neighbourhood of the diagonal must be sufficient to characterise any local *sample path regularity property, as it must suffice to consider the covariances at arbitrarily close by points.*

Theorem 7 combines results about sample Hölder continuity of GPs (Azmoodeh et al., 2014) to results about sample differentiability of GPs (Potthoff, 2010), applied inductively on the partial derivatives. It provides necessary and sufficient conditions for f to have almost-Hölder continuous samples of a certain degree. Let us note however that it does not quite give us necessary conditions for f to have samples in $C^n(O)$, as the latter is not an almost-Hölder space. To achieve this, we can adapt the proof of Theorem 7 combining necessary and sufficient conditions for sample continuity of GPs with the results about sample differentiability from Potthoff (2010) (see Scheuerer (2010a, Theorem 5.3.16) for a result of this flavour). Characterising necessary and sufficient conditions on the kernel for the sample continuity of the GP is much more involved than almost-Hölder continuity, and therefore we do not expand our results to this setting (see Fernique (1975) for the stationary case, Talagrand (1987) for the general case). Moreover we demonstrate in the examples below that Theorem 7 is all we need.

4 Examples

In this section we demonstrate how Theorem 7 can be applied in practice. We investigate the sample path regularity of the most widely used GP families, recovering known results as well as proving novel ones.

The reverse implications in Theorem 7 allow us to obtain sharp sample path regularity characterisations in the following sense: f is said to have samples in $C_{loc}^{(n+\gamma)^-}(O)$ *and no more*, if f is not sample $C_{loc}^{(n'+\gamma')^-}(O)$ for any $n' + \gamma' > n + \gamma$.

4.1 Wiener Kernel

The Wiener process is a centered GP on $O = \mathbb{R}_{>0}$ with covariance kernel

$$k(x, y) = \min(x, y)$$

for $x, y \in \mathbb{R}_{>0}$. $\min(\cdot, \cdot)$ is Lipschitz but non-differentiable in both its arguments. So by Theorem 7 (1) the Wiener process has samples in $C_{loc}^{1/2^-}(\mathbb{R}_{>0})$ and no more.

Remark 9 *By the fundamental theorem of calculus, the n-times integrated Wiener process (Schober et al., 2014, Section B) has samples in $C_{loc}^{(n+1/2)^-}(\mathbb{R}_{>0})$ and no more.*

4.2 Matérn Kernels

The Matérn kernels are isotropic kernels on $\mathbb{R}^d$ given by

$$k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{y}\|\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\|\boldsymbol{x} - \boldsymbol{y}\|\right) \tag{3}$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where K_ν is the modified Bessel function of the second kind and $\nu > 0$ is the smoothness parameter. The cases where $\nu = n + 1/2$ for some $n \in \mathbb{N}_0$ are particularly interesting because the expression in Equation (3) simplifies to a product of a polynomial and an exponential of the radial distance (Rasmussen and Williams, 2005, Equation 4.16).

Proposition 10 *A centered Matérn GP with smoothness parameter $\nu > 0$ has samples in $C^{\nu^-}_{loc}(O)$ and no more. In particular, if $\nu = n + 1/2$ for some $n \in \mathbb{N}_0$, the GP has samples in $C^{(n+1/2)^-}_{loc}(\mathbb{R}^d)$ and no more.*

This follows from Theorem 7 (3) and the proof can be found in Appendix B.1.

Remark 11 *Proposition 10 also covers the (multivariate) Ornstein-Uhlenbeck process, the centered GP with covariance kernel given by $k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|)$ which corresponds to the Matérn GP with $\nu = 1/2$. So the (multivariate) Ornstein-Uhlenbeck process has samples in $C^{1/2^-}_{loc}(\mathbb{R}^d)$ and no more.*

4.3 Wendland Kernels

The Wendland kernels are isotropic kernels which are compactly supported and piecewise polynomial in the radial distance. On $\mathbb{R}^d$, they are defined as follows (Wendland, 2004, Definition 9.11):

$$k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|) = \mathcal{I}^n \phi_{\lfloor d/2 \rfloor + n + 1}(\|\boldsymbol{x} - \boldsymbol{y}\|)$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where $\phi_j(\rho) := \max((1-\rho)^j, 0)$ and $\mathcal{I}\phi(\rho) := \int_\rho^\infty t\phi(t)\,dt / \int_0^\infty t\phi(t)\,dt$ for $\rho \geq 0$. $n \in \mathbb{N}_0$ is a parameter controlling the degree of the polynomial (precisely, $k_r(x)$ has degree $\lfloor d/2 \rfloor + 3n + 1$ in ρ around 0).

Proposition 12 *A centered Wendland GP with degree parameter n has samples in $C^{(n+1/2)^-}_{loc}(O)$ and no more.*

This follows from Theorem 7 (3) and the proof can be found in Appendix B.2.

4.4 Smooth Kernels

From Theorem 7 immediately follows the following important corollary:

Corollary 13 *The process $f \sim \mathcal{GP}(0, k)$ has samples in $C^\infty(O)$ if and only if,*

(1) for general k, $k \in C^\infty(O \times O)$.

(2) for stationary $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y})$, $k_\delta \in C^\infty(\mathbb{R}^d)$.

(3) for isotropic $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$, $k_r \in C^\infty(\mathbb{R})$.

Remark 14 *The centered GPs with squared exponential ($k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\|\boldsymbol{x} - \boldsymbol{y}\|^2)$), rational quadratic ($k(\boldsymbol{x}, \boldsymbol{y}) = (1 + \|\boldsymbol{x} - \boldsymbol{y}\|^2)^{-a}$ for some $a > 0$) and periodic ($k(x, y) = \exp(-\sin(\pi(x-y))^2)$ for $x, y \in \mathbb{R}$) covariance kernels are therefore sample $C^\infty(\mathbb{R}^d)$.*

4.5 Feature Kernels

For any map $\boldsymbol{\phi}\colon O \to \mathbb{R}^m$ is associated a feature kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{y})$$

for $\boldsymbol{x}, \boldsymbol{y} \in O$. From Theorem 7 (1) we obtain the following result:

Proposition 15 *For $n \in \mathbb{N}_0$ and $\gamma \in (0,1]$, if $\phi_i \in C_{loc}^{(n+\gamma)^-}(O)$ for all $1 \le i \le m$ then $f \sim \mathcal{GP}(0,k)$ has samples in $C_{loc}^{(n+\gamma)^-}(O)$.*

The proof can be found in Appendix B.3.

Remark 16 *As a special case, linear kernels ($k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$) and polynomial kernels ($k(\boldsymbol{x}, \boldsymbol{y}) = (1 + \langle \boldsymbol{x}, \boldsymbol{y} \rangle)^m$ for some $m \in \mathbb{N}$) are feature kernels, with the feature maps being the coordinate maps or polynomials of the coordinate maps respectively. Therefore the associated centered GPs have samples in $C^\infty(\mathbb{R}^d)$.*

4.6 Kernel Algebra

The versatility of the GP framework in applications is largely due to the fact that covariance kernels can be recursively combined using algebraic operations like pointwise sums or products which allows for constructing highly flexible and complex prior models in a systematic and interpretable fashion. It is therefore important to understand how the sample path regularity of a GP depends on the recursive structure of the covariance kernel.

In the following the k_i are always assumed to be symmetric positive definite kernels. Note that in this section we work solely with general kernels, leveraging Theorem 7 (1). It would then be straightforward to similarly leverage Theorem 7 (2) and Theorem 7 (3) for stationary and isotropic kernels respectively.

4.6.1 Conic Combinations

Conic combinations of kernels are pointwise linear combinations of the kernel functions with positive weights, i.e.

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} a_i k_i(\boldsymbol{x}, \boldsymbol{y})$$

for $\boldsymbol{x}, \boldsymbol{y} \in O$, where $a_i > 0$ for all $1 \le i \le m$. Sums of kernels and positive scaling of kernels are special cases of conic combinations with $a_i = 1$ and $m = 1$, respectively.

Proposition 17 *For $n \in \mathbb{N}_0$ and $\gamma \in (0,1]$, if the k_i satisfy the condition in Theorem 7 (1) for all $1 \le i \le m$ then $f \sim \mathcal{GP}(0,k)$ has samples in $C_{loc}^{(n+\gamma)^-}(O)$.*

This follows from the fact that for $f \sim \mathcal{GP}(0,k)$, we can express the sample paths as $f(\cdot,\omega) = \sum_{i=1}^m \sqrt{a_i} f_i(\cdot,\omega)$ where $f_i \sim \mathcal{GP}(0,k_i)$, and that $C_{loc}^{(n+\gamma)^-}(O)$ is a vector space.

4.6.2 Products

Define the product kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{m} k_i(\boldsymbol{x}, \boldsymbol{y})$$

for $\boldsymbol{x}, \boldsymbol{y} \in O$. k is also positive definite (Berg et al., 1984, Chapter 3 Theorem 1.12).

Proposition 18 *For $n \in \mathbb{N}_0$ and $\gamma \in (0, 1]$, if the k_i satisfy the condition in Theorem 7(1) for all $1 \leq i \leq m$ then $f \sim \mathcal{GP}(0, k)$ has samples in $C_{loc}^{(n+\gamma)^-}(O)$.*

The proof can be found in Appendix B.4.

Remark 19 *Combining Proposition 18 with Proposition 15 we get that if $\tilde{k}$ satisfies the condition in Theorem 7(1) for some $n \in \mathbb{N}_0$ and $\gamma \in (0, 1]$, and $\phi \in C_{loc}^{(n+\gamma)^-}(O)$, then defining $k(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})\tilde{k}(\boldsymbol{x}, \boldsymbol{y})\phi(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in O$, $f \sim \mathcal{GP}(0, k)$ has samples in $C_{loc}^{(n+\gamma)^-}(O)$. We could equivalently prove this by noting that we can express the sample paths as $f(\cdot, \omega) = \phi(\cdot)\tilde{f}(\cdot, \omega)$ where $\tilde{f} \sim \mathcal{GP}(0, \tilde{k})$.*

4.6.3 Tensor Products

Let $O_i \subset \mathbb{R}^{d_i}$ be open sets for $1 \leq i \leq m$ and $O_1 \times \cdots \times O_m =: O \subset \mathbb{R}^d = \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_m}$. For $k_i \colon O_i \times O_i \to \mathbb{R}$ for $1 \leq i \leq m$, the tensor product kernel $k \colon O \times O \to \mathbb{R}$, is given by

$$k(\boldsymbol{x}, \boldsymbol{y}) = (k_1 \otimes \cdots \otimes k_m)(\boldsymbol{x}, \boldsymbol{y}) := \prod_{i=1}^{m} k_i(\boldsymbol{x}_i, \boldsymbol{y}_i)$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, where $\boldsymbol{x}_i, \boldsymbol{y}_i \in \mathbb{R}^{d_i}$ are the orthogonal projections of $\boldsymbol{x}$, $\boldsymbol{y}$ respectively onto $\mathbb{R}^{d_i}$.

GPs with tensor product covariance kernels are interesting as they allow *spherically asymmetric* sample path regularity. We need the following definition to formalise this.

Definition 20 *Let $n_1, \ldots, n_m \in \mathbb{N}_0$ and $\gamma_1, \ldots, \gamma_m \in [0, 1]$.*

(1) $C_{loc}^{(n_1,\gamma_1)\otimes\cdots\otimes(n_m,\gamma_m)}(O_1 \times \cdots \times O_m)$ is the space of functions f on O for which $\partial^{\boldsymbol{\alpha}} f$ exists for all multi-indices $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}_i| \leq n_i$, where $\boldsymbol{\alpha}_i \in \mathbb{N}_0^{d_i}$ is the projection of $\boldsymbol{\alpha}$ onto $\mathbb{N}_0^{d_i}$, and such that the highest order partial derivatives satisfy a Hölder condition of the form: for all compact subsets $K \subset O$ there is a constant $C_K > 0$ such that

$$|\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \partial^{\boldsymbol{\alpha}} f(\boldsymbol{y})| \leq C_K \|\boldsymbol{x}_i - \boldsymbol{y}_i\|^{\gamma_i}$$

for all $1 \leq i \leq d$, $\boldsymbol{x}, \boldsymbol{y} \in K$ with $\boldsymbol{x}_j = \boldsymbol{y}_j$ for $j \neq i$, and $|\boldsymbol{\alpha}_i| = n_i$.

(2) $C_{loc}^{(n_1+\gamma_1)^-\otimes\cdots\otimes(n_m+\gamma_m)^-}(O_1 \times \cdots \times O_m)$ $:= \bigcap_{n_i'+\gamma_i'<n_i+\gamma_i} C_{loc}^{(n_1',\gamma_1')\otimes\cdots\otimes(n_m',\gamma_m')}(O_1 \times \cdots \times O_m)$.

The following proposition generalises Theorem 7.

Proposition 21 *For $n_1, \ldots, n_m \in \mathbb{N}_0$ and $\gamma_1, \ldots, \gamma_m \in (0, 1]$, $f \sim \mathcal{GP}(0, k)$ has samples in $C_{loc}^{(n_1+\gamma_1)^- \otimes \cdots \otimes (n_m+\gamma_m)^-}(O_1 \times \cdots \times O_m)$ if and only if the k_i satisfy the condition in Theorem 7(1) with $n = n_i$ and $\gamma = \gamma_i$ for all $1 \le i \le m$.*

The proof of Proposition 21 follows the same steps as the one of Theorem 7 in Appendix A.

Remark 22 *We could generalise Theorem 7 to allow for even more flexibility in the asymmetrical derivatives in Theorem 7. This can be done by defining Hölder spaces extending the spaces $C^A(O)$, where $A \subset \mathbb{N}_0^d$ is a downward closed set of multi-indices which specifies which partial derivatives exist (Pförtner et al., 2022, Definition B.9).*

4.6.4 Coordinate Transformations

For a map $\boldsymbol{\varphi}\colon O \to \mathbb{R}^m$ and a symmetric positive definite kernel $\tilde{k}$ defined on $\mathrm{Im}(\boldsymbol{\varphi}) \times \mathrm{Im}(\boldsymbol{\varphi})$, we have a kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \tilde{k}(\boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{y})).$$

Proposition 23 *For $n \in \mathbb{N}_0$ and $\gamma, \delta \in (0, 1]$, if $\tilde{k}$ satisfies the condition in Theorem 7(1) and $\varphi_i \in C_{loc}^{(n+\delta)^-}(O)$ for all $1 \le i \le m$ then $f \sim \mathcal{GP}(0, k)$ has samples in $C_{loc}^{(n+\gamma\delta)^-}(O)$.*

This follows from the fact that for $\tilde{f} \sim GP(0, \tilde{k})$ we have $\tilde{f} \circ \boldsymbol{\varphi} \sim \mathcal{GP}(0, k)$, and that the composition of a $C_{loc}^{n,\zeta}(O)$ with a $C_{loc}^{n,\epsilon}(O)$ function is $C_{loc}^{n,\epsilon\zeta}(O)$.

4.7 Manifold Kernels

We can generalise the results in the present work to manifold domains; instead of considering GPs on open subsets $O \subset \mathbb{R}^d$, we consider GPs on manifolds M. There exists two ways of constructing positive definite kernels, and hence GPs, on manifolds M.

Extrinsic kernels are defined by viewing $M \subset \mathbb{R}^d$, taking a positive definite kernel k on $\mathbb{R}^d$, and restricting it to M. This restriction is analogous to a coordinate transformation (Section 4.6.4), since $k|_{M\times M}(\boldsymbol{x}, \boldsymbol{y}) = k(\boldsymbol{\iota}(\boldsymbol{x}), \boldsymbol{\iota}(\boldsymbol{y}))$ for $\boldsymbol{x}, \boldsymbol{y} \in M$, where $\boldsymbol{\iota} : M \to \mathbb{R}^d$ is the inclusion map. Therefore, if the manifold is smooth, $k|_{M\times M}$ has the same sample regularity as k.

Intrinsic kernels are constructed directly on the manifold. Examples of these include the intrinsic Matérn and heat kernels (Borovitskiy et al., 2020), and the hyperbolic secant kernel (Da Costa et al., 2023). In this case, since the sample path regularity results in this work are all local, we may study sample path regularity in each coordinate patch $O \subset M$, treating them as open subsets of $\mathbb{R}^d$.

5 Sobolev Regularity

In this section we investigate how the regularity of the kernel affects the weak differentiability of the sample paths. Specifically, we consider the important L^2-Sobolev regularity. The more general L^p-Sobolev regularity is studied in Henderson (2024). To characterise only the local regularity of the sample paths, as for Hölder spaces, we define the local pre-Sobolev spaces.

Definition 24 (Local pre-Sobolev spaces) *Let $n \in \mathbb{N}_0$. The local pre-Sobolev space $\mathcal{H}^n_{loc}(O)$ is the space of functions*[1] *f on O for which for every compact $K \subset O$ the L^2 weak derivative $\partial^{\boldsymbol{\alpha}}_w f$ exists on K for all multi-index $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| \leq n$, i.e. for all such $\boldsymbol{\alpha}$ there is a function $\partial^{\boldsymbol{\alpha}}_w f \in L^2(K)$ such that*

$$\int_K \partial^{\boldsymbol{\alpha}}_w f(\boldsymbol{x})\varphi(\boldsymbol{x})\, d\boldsymbol{x} = (-1)^{|\boldsymbol{\alpha}|} \int_K f(\boldsymbol{x})\partial^{\boldsymbol{\alpha}}\varphi(\boldsymbol{x})\, d\boldsymbol{x}$$

for all $\varphi \in C^\infty(K)$.

Theorem 25 (Sample path Sobolev regularity) *Let $n \in \mathbb{N}$. The process $f \sim \mathcal{GP}(0, k)$ has samples in $\mathcal{H}^n_{loc}(O)$ if,*

(1) for general k, $k \in C^{n \otimes n}(O \times O)$.

(2) for stationary $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y})$, $\partial^{\boldsymbol{\alpha}} k_\delta$ exists at $\mathbf{0}$ for all $|\boldsymbol{\alpha}| \leq 2n$.

(3) for isotropic $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$, $k_r^{(j)}$ exists at 0 for all $j \leq 2n$.

If we assume that k is continuous, or more generally that f has a measurable modification, then the converse holds in the scenarios (2) and (3).

Proof (1) follows from the general theorem for second order measurable random fields (Scheuerer, 2010b, Theorem 1). (2) and (3) follow from its corollary (Scheuerer, 2010b, Corollary 1). The converse in (2) follows from (Scheuerer, 2010b, Proposition 1). Finally, if (3) holds then, by the propagation of regularity from Gneiting (1999), we have $k_r \in C^{2n}(\mathbb{R})$. So by Lemma A.1 (3) $\Rightarrow$ (2) (to be precise an adapted version without the Hölder conditions) we deduce (2), and hence we obtain the converse statement in this case too. ■

As observed in Scheuerer (2010b), Sobolev differentiability is a natural regularity notion for sample paths of general random fields as it can be deduced from continuous mean square differentiability. In the case of GPs however, if, in addition to the condition in Theorem 25 (1), we assume Hölder control on the highest order partial derivatives of the kernel at the diagonal as in Theorem 7 (1), we deduce strong differentiability of the sample paths. The Hölder condition in Theorem 7 (1) being a weak one, one can usually not obtain a greater number of weak derivatives using Theorem 25 than of strong derivatives using Theorem 7. The reverse implications in Theorem 25 (2) and Theorem 25 (3) show that, in the stationary and isotropic cases, this is not a limitation of the theorem, but an inherent property of stationary and isotropic GPs. For example, the proof of Proposition 10 Appendix B.1 shows that, by the converse to Theorem 25 (3), even in the edge case $\nu \in \mathbb{N}$, the samples of the Matérn GPs have as many Sobolev derivatives as of strong derivatives.

Note however that one way to construct non-stationary GPs with more weak derivatives than strong derivatives is through feature kernels, with features that admit more weak derivatives than strong derivatives. This also shows that it is not possible to obtain a converse for Theorem 25 (1) as we have for Theorem 25 (2) and Theorem 25 (3).

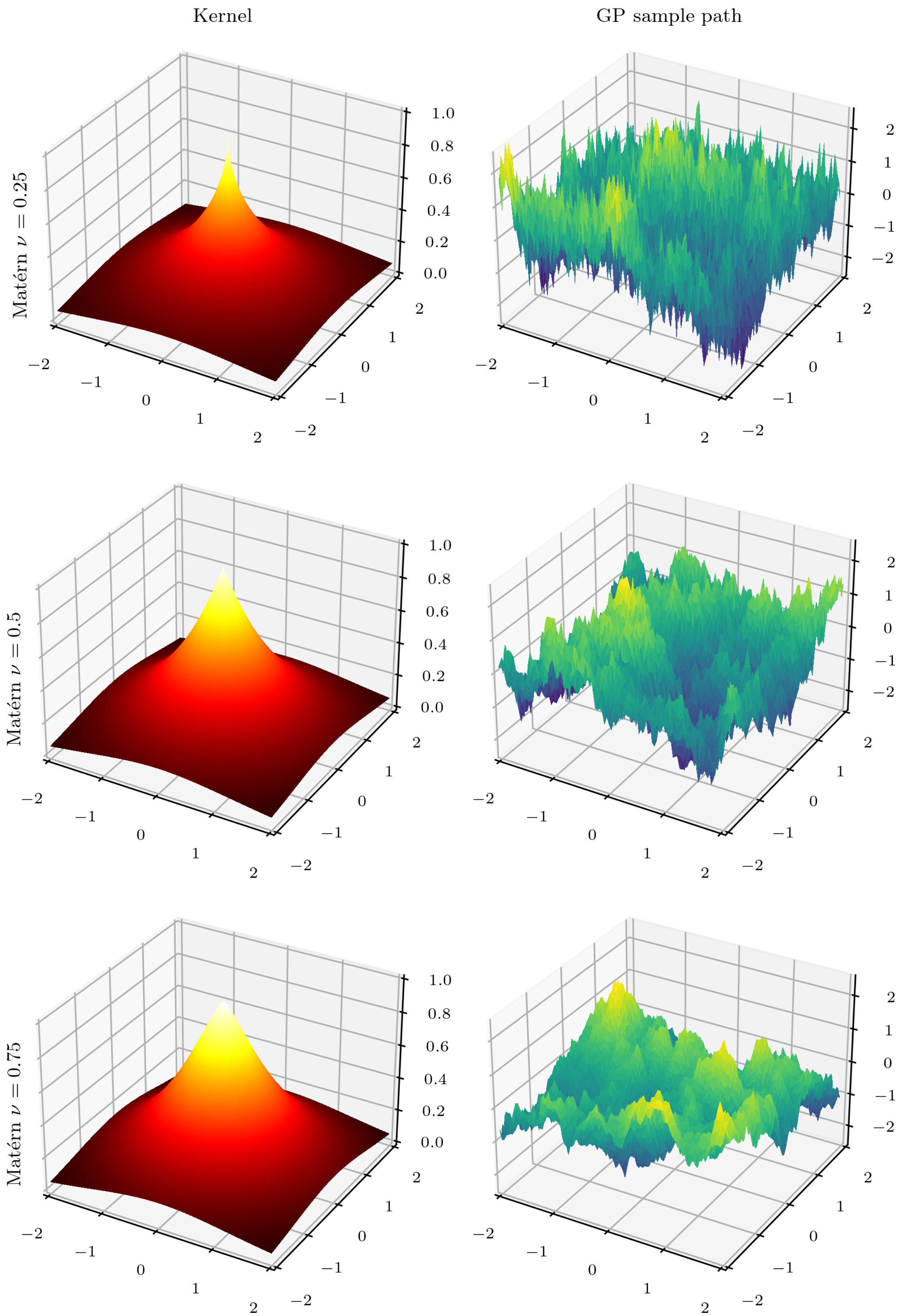
1. A (local) Sobolev space is an L^2 quotient of a (local) pre-Sobolev space. Since we are interested in *function* spaces, pre-Sobolev spaces are the right framework for us.

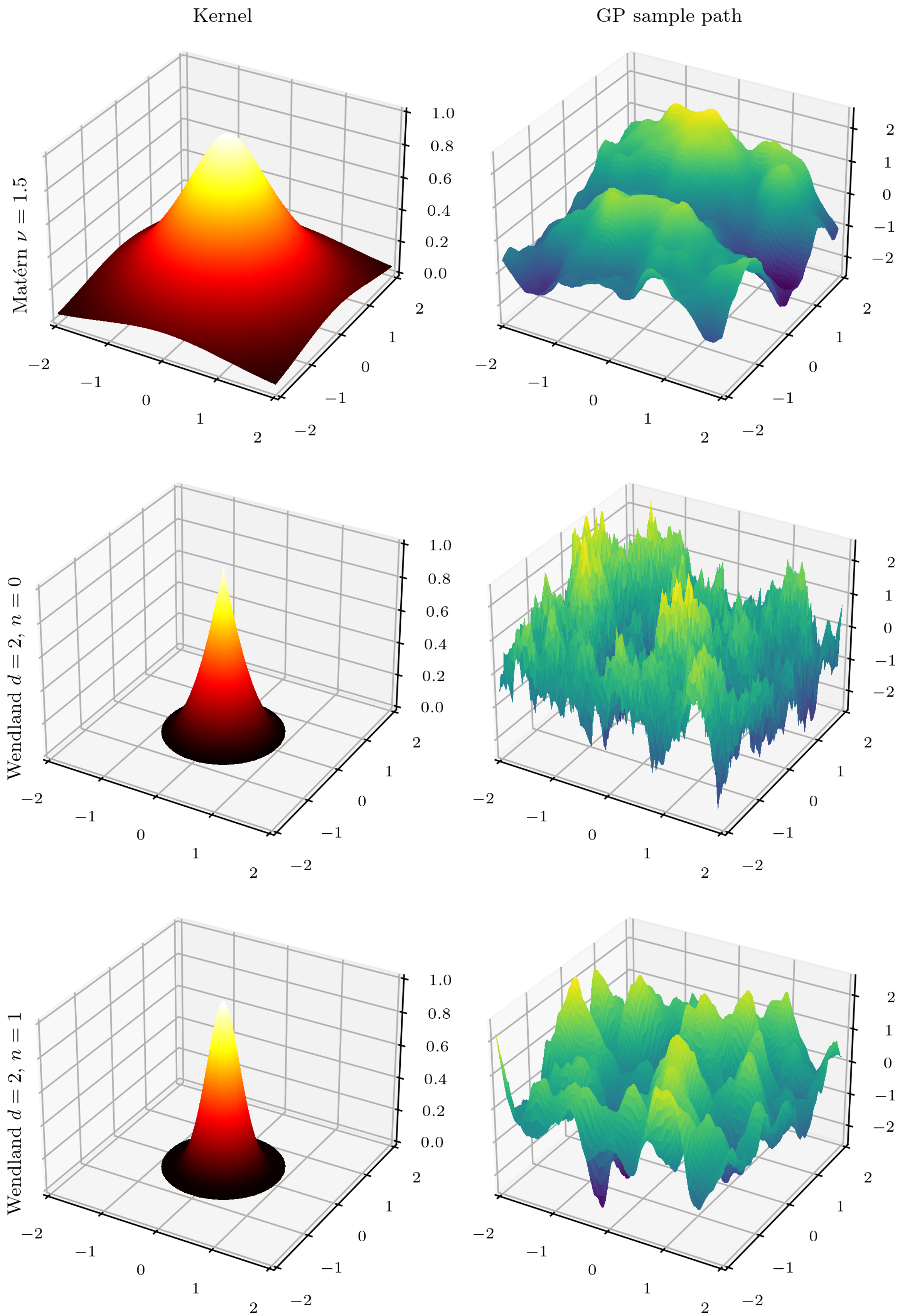
Also note that deducing sample path Hölder regularity from applying Theorem 25 and Sobolev embedding theorems introduces a superfluous dependence on dimension, and yields weaker results than the dimension independent Theorem 7.

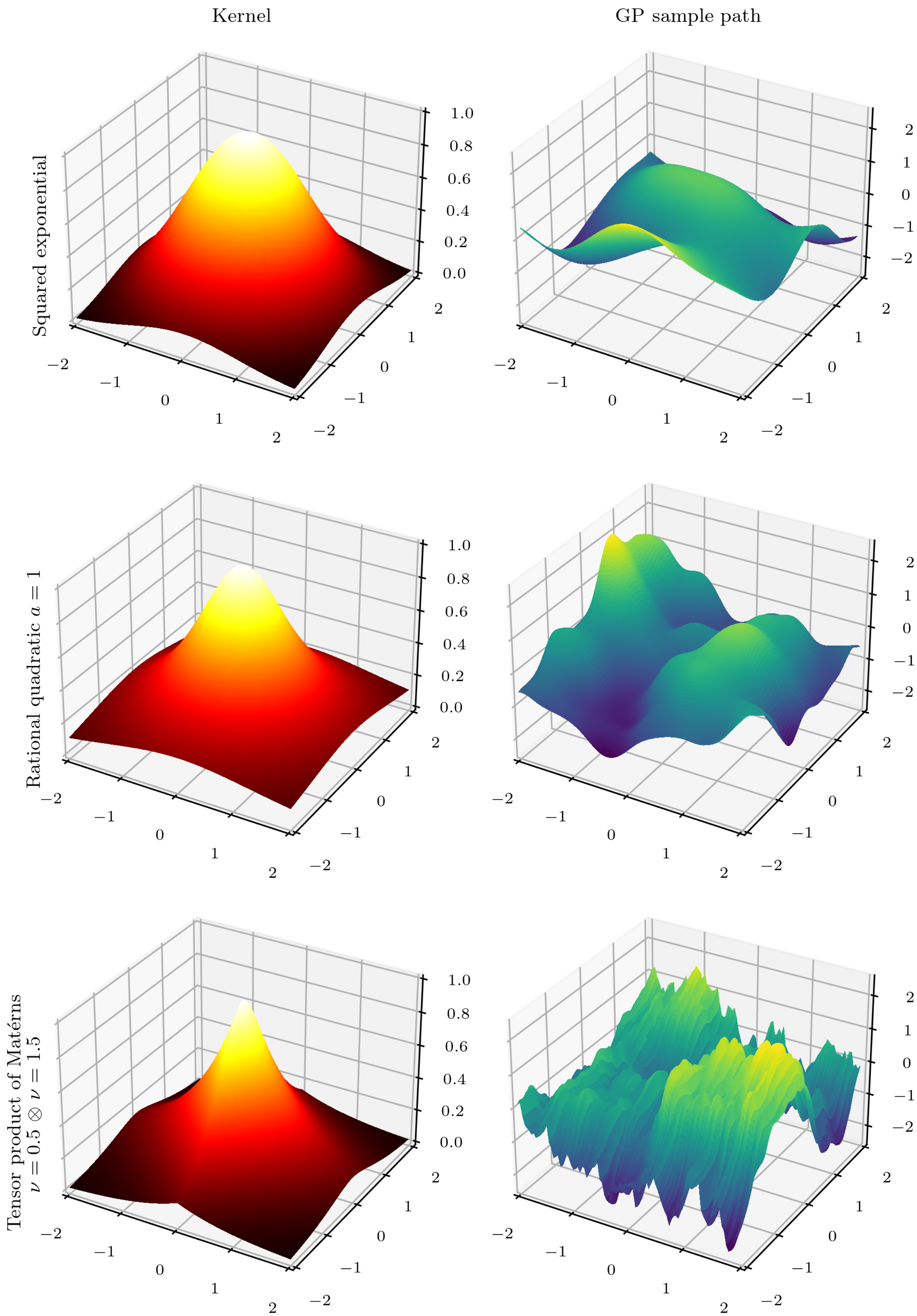
We therefore expect Theorem 25 to have less practical value for GPs as its general version (Scheuerer, 2010b, Theorem 1) for second order measurable random fields. For GPs we encourage the use of Theorem 7.

Acknowledgments and Disclosure of Funding

NDC, MP and PH gratefully acknowledge financial support by the European Research Council through ERC StG Action 757275 / PANAMA; the DFG Cluster of Excellence "Machine Learning - New Perspectives for Science", EXC 2064/1, project number 390727645; the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A); and funds from the Ministry of Science, Research and Arts of the State of Baden-Württemberg. NDC acknowledges the support of the Fonds National de la Recherche, Luxembourg. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting NDC and MP.


Kernel
GP sample path
Matérn ν = 0.25
Matérn ν = 0.5
Matérn ν = 0.75
0.0
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2


Kernel
GP sample path
Matérn $\nu = 1.5$
Wendland $d = 2$, $n = 0$
Wendland $d = 2$, $n = 1$
1.0
0.8
0.6
0.4
0.2
0.0
−2
−1
0
1
2


Kernel
GP sample path
Squared exponential
Rational quadratic $a = 1$
Tensor product of Matérns
$\nu = 0.5 \otimes \nu = 1.5$

Appendix A. Proof of Theorem 7

We start by showing the equivalences of the various kernel conditions in Theorem 7, namely (1) $\Leftrightarrow$ (2) in the stationary case, and (1) $\Leftrightarrow$ (2) $\Leftrightarrow$ (3) in the isotropic case.

Lemma A.1 *Let $n \in \mathbb{N}_0$ and $\epsilon \in (0, 1]$. Then the condition*

(1)
- $k \in C^{n \otimes n}(O \times O)$,
- $|\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}, \boldsymbol{x})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as $\boldsymbol{h} \to \boldsymbol{0}$, locally uniformly in $\boldsymbol{x} \in O$, for all $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$.*

is equivalent to the following:

(2) for stationary $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x}, \boldsymbol{y})$,

- $k_\delta \in C^{2n}(\mathbb{R}^d)$,
- $|\partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{0})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as $\boldsymbol{h} \to \boldsymbol{0}$ for all $|\boldsymbol{\alpha}| = 2n$.*

(3) for isotropic $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$,

- $k_r \in C^{2n}(\mathbb{R})$,
- $|k_r^{(2n)}(h) - k_r^{(2n)}(0)| = \mathcal{O}(|h|^{2\epsilon})$ *as $h \to 0$.*

Proof of Lemma A.1 (1) $\Leftrightarrow$ (2) Let k be stationary, i.e. $k(\boldsymbol{x}, \boldsymbol{y}) = k_\delta(\boldsymbol{x} - \boldsymbol{y})$.

(2) $\Rightarrow$ (1): For $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq n$,

$$\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}, \boldsymbol{y}) = (-1)^{|\boldsymbol{\beta}|} \partial^{\boldsymbol{\alpha}+\boldsymbol{\beta}} k_\delta(\boldsymbol{x} - \boldsymbol{y})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, by the chain rule. So the existence and the continuity of $\partial^{\boldsymbol{\alpha}+\boldsymbol{\beta}} k_\delta$ implies the existence and the continuity of $\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k$. Moreover when $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$,

$$\begin{aligned} &|\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}, \boldsymbol{x})| \\ &= 2|\partial^{\boldsymbol{\alpha}+\boldsymbol{\beta}} k_\delta(\boldsymbol{0}) - \partial^{\boldsymbol{\alpha}+\boldsymbol{\beta}} k_\delta(\boldsymbol{h})| \end{aligned} \tag{4}$$

for all $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^d$, so the Hölder conditions on the highest order partial derivatives of k at the diagonal and of k_δ at $\boldsymbol{0}$ correspond.

(1) $\Rightarrow$ (2): For $|\boldsymbol{\alpha}| \leq 2n$, say $\boldsymbol{\alpha} = \boldsymbol{\beta} + \boldsymbol{\gamma}$ with $|\boldsymbol{\beta}|, |\boldsymbol{\gamma}| \leq n$, $\partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{x})$ exists and is equal to $(-1)^{|\boldsymbol{\gamma}|} \partial^{\boldsymbol{\beta},\boldsymbol{\gamma}} k(\boldsymbol{x}, \boldsymbol{0})$, for all $\boldsymbol{x} \in \mathbb{R}^d$. Moreover the Hölder conditions on the partial derivatives of k and k_δ correspond by Equation (4). ■

To prove the second equivalence in Lemma A.1, we need the following result:

Lemma A.2 *Let $m \in \mathbb{N}_0$, $g \in C^m(\mathbb{R} \setminus \{0\})$, and $\nu \in \mathbb{R}$ such that $g^{(j)}(h) = \mathcal{O}(|h|^{\nu-j})$ as $h \to 0$, for all $0 \leq j \leq m$. Define $f := g \circ \|\cdot\| \colon \mathbb{R}^d \setminus \{\boldsymbol{0}\} \to \mathbb{R}$. Then $f \in C^m(\mathbb{R}^d \setminus \{\boldsymbol{0}\})$ and $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) = \mathcal{O}(\|\boldsymbol{h}\|^{\nu-j})$ as $\boldsymbol{h} \to \boldsymbol{0}$, for $0 \leq j \leq m$ and $|\boldsymbol{\alpha}| = j$.*

Proof By induction on m. $m = 0$ is clear; the growth/decay of f at $\boldsymbol{0}$ is the same as that of g at 0. For $m > 0$, $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, and $1 \leq i \leq d$, note that

$$\partial^{\boldsymbol{e}_i} f(\boldsymbol{x}) = g'(\|\boldsymbol{x}\|) \frac{x_i}{\|\boldsymbol{x}\|} = \tilde{g}(\|\boldsymbol{x}\|) x_i \tag{5}$$

where $\boldsymbol{e}_i$ is the i^{th} unit vector and $\tilde{g}(x) := \frac{g'(x)}{x}$ for $x \in \mathbb{R} \setminus \{0\}$. Clearly $\tilde{g} \in C^{m-1}(\mathbb{R} \setminus \{0\})$ and

$$\tilde{g}^{(j)}(h) = \sum_{l=0}^{j} \binom{j}{l} g^{(l+1)}(h)(-1)^{j-l}(j-l)!h^{-1-j+l} = \mathcal{O}(|h|^{(\nu-2)-j})$$

as $h \to 0$, for all $0 \le j \le m-1$. Let $\tilde{f} = \tilde{g} \circ \|\cdot\|$. By the induction hypothesis, $\partial^{\boldsymbol{\alpha}} \tilde{f}(\boldsymbol{h}) = \mathcal{O}(\|\boldsymbol{h}\|^{(\nu-2)-j})$ as $\boldsymbol{h} \to \boldsymbol{0}$, for $|\boldsymbol{\alpha}| = j$ and $0 \le j \le m-1$. Thus, by (5),

$$\partial^{\boldsymbol{\alpha}} \partial^{\boldsymbol{e}_i} f(\boldsymbol{h}) = \begin{Bmatrix} h_i \partial^{\boldsymbol{\alpha}} \tilde{f}(\boldsymbol{h}) & \text{if } \alpha_i = 0 \\ h_i \partial^{\boldsymbol{\alpha}} \tilde{f}(\boldsymbol{h}) + \partial^{\boldsymbol{\alpha}-\boldsymbol{e}_i} \tilde{f}(\boldsymbol{h}) & \text{if } \alpha_i > 0 \end{Bmatrix} = \mathcal{O}(\|\boldsymbol{h}\|^{\nu-(j+1)})$$

as $\boldsymbol{h} \to \boldsymbol{0}$. Consequently, $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) = \mathcal{O}(\|\boldsymbol{h}\|^{\nu-j})$ as $\boldsymbol{h} \to \boldsymbol{0}$ for $|\boldsymbol{\alpha}| = j$ and $0 \le j \le m$. This concludes the induction step. ■

Proof of Lemma A.1 (1) ⇔ (3) Let k be isotropic, i.e. $k(\boldsymbol{x}, \boldsymbol{y}) = k_r(\|\boldsymbol{x} - \boldsymbol{y}\|)$. By Lemma A.1 (1) ⇔ (2), it suffices to show that the condition on k_δ in (2) is equivalent to the condition on k_r in (3).

(2) ⇒ (3): This follows from the fact that $x \mapsto k_\delta(x\boldsymbol{e}_1)$ is exactly k_r, and analogously $x \mapsto \partial^{j\boldsymbol{e}_1} k_\delta(x\boldsymbol{e}_1)$ is equal to $k_r^{(j)}$ for $0 \le j \le 2n$.

(3) ⇒ (2): $k_r \colon \mathbb{R} \to \mathbb{R}$ is even, so the odd derivatives of k_r which exist at 0 vanish there. Let

$$g(x) := k_r(x) - \sum_{j=0}^{n} \frac{x^{2j}}{j!} k_r^{(2j)}(0),$$

for all $x \in \mathbb{R}$. The Lagrange form of the remainder in an order $2n-j-1$ Taylor expansion of $g^{(j)}$ then reveals that

$$g^{(j)}(h) = \frac{h^{2n-j}}{(2n-j)!}(k_r^{(2n)}(\xi_j(h)) - k_r^{(2n)}(0)) = \mathcal{O}(|h|^{2n-j+2\epsilon})$$

as $h \to 0$, for all $0 \le j \le 2n$ and where $\xi_j(h) \in (0, h)$ (or $(h, 0)$ if $h < 0$). So $g|_{\mathbb{R} \setminus \{0\}}$ satisfies the conditions of Lemma A.2 with $m = 2n$ and $\nu = 2n + 2\epsilon$. Let $f := g \circ \|\cdot\| \colon \mathbb{R}^d \to \mathbb{R}$. Then, by Lemma A.2, $f|_{\mathbb{R}^d \setminus \{\boldsymbol{0}\}} \in C^{2n}(\mathbb{R}^d \setminus \{\boldsymbol{0}\})$, and $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) = \mathcal{O}(\|\boldsymbol{h}\|^{2n+2\epsilon-j})$ as $\boldsymbol{h} \to \boldsymbol{0}$ for $|\boldsymbol{\alpha}| = j$ and $0 \le j \le 2n$. In particular, $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to \boldsymbol{0}$, and by the mean value theorem this is sufficient to deduce that $\partial^{\boldsymbol{\alpha}} f(\boldsymbol{0})$ exists and is 0. Now by noting that $\|\boldsymbol{x}\|^{2j} = (x_1^2 + \cdots + x_d^2)^j$ is smooth in $\boldsymbol{x}$ for all j, we have that

$$k_\delta(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{j=0}^{n} \frac{\|\boldsymbol{x}\|^{2j}}{j!} k_r^{(2j)}(0)$$

is in $C^{2n}(\mathbb{R})$. Moreover, for $\boldsymbol{h} \in \mathbb{R}^d$ and $|\boldsymbol{\alpha}| = 2n$ there is a constant $C_{\boldsymbol{\alpha}} \in \mathbb{R}$ such that $\partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{h}) = \partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) + C_{\boldsymbol{\alpha}}$, implying that

$$|\partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{0})| = |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{h}) - \underbrace{\partial^{\boldsymbol{\alpha}} f(\boldsymbol{0})}_{=0}| = \mathcal{O}(\|\boldsymbol{h}\|^{2n+2\epsilon-2n}) = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$$

as $\boldsymbol{h} \to \mathbf{0}$. Thus k_δ satisfies the Hölder condition in (2). ■

Note that Lemma A.1 is purely a result about stationary and isotropic functions; the proof does make use of the positive definiteness of k. In fact, the positive definiteness of k gives us additional information about its regularity, which can be seen through Lemma A.3 below.

For a GP f, the mean-square partial derivative $\partial_{ms}^{\boldsymbol{e}_i} f$, when it exists, is a stochastic process defined on the same probability space such that $\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x})$ is the L^2 limit of $\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}$ as $h \to 0$. The existence of a mean-square partial derivative is neither sufficient nor necessary for the existence of sample derivatives. However such derivatives are relevant for our work as they are directly related to the derivatives of the covariance kernel:

Lemma A.3 *Let $n \in \mathbb{N}_0$. For $f \sim \mathcal{GP}(0, k)$, the following conditions are equivalent:*

(1) $k \in C^{n \otimes n}(O \times O)$,

(2) $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k$ exists in a neighbourhood the diagonal and is continuous there for all $|\boldsymbol{\alpha}| \leq n$,

(3) the mean-square partial derivative $\partial_{ms}^{\boldsymbol{\alpha}} f$ exists and is mean-square continuous for all $|\boldsymbol{\alpha}| \leq n$.

Moreover, if either condition holds then we have $\partial_{ms}^{\boldsymbol{\alpha}} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k)$ for all $|\boldsymbol{\alpha}| \leq n$.

Proof We prove this for $n = 1$; for general n it then suffices to apply the same argument inductively on the partial derivatives.

(1) ⇒ (2): This implication is clear.

(2) ⇒ (3): Let $1 \leq i \leq d$. Note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}\right)\left(\frac{f(\boldsymbol{y}+h'\boldsymbol{e}_i)-f(\boldsymbol{y})}{h'}\right)\right] \\
&= \frac{\mathbb{E}[f(\boldsymbol{x}+h\boldsymbol{e}_i)f(\boldsymbol{y}+h'\boldsymbol{e}_i)] - \mathbb{E}[f(\boldsymbol{x}+h\boldsymbol{e}_i)f(\boldsymbol{y})] - \mathbb{E}[f(\boldsymbol{x})f(\boldsymbol{y}+h'\boldsymbol{e}_i)] + \mathbb{E}[f(\boldsymbol{x})f(\boldsymbol{y})]}{hh'} \\
&= \frac{k(\boldsymbol{x}+h\boldsymbol{e}_i, \boldsymbol{y}+h'\boldsymbol{e}_i) - k(\boldsymbol{x}+h\boldsymbol{e}_i, \boldsymbol{y}) - k(\boldsymbol{x}, \boldsymbol{y}+h'\boldsymbol{e}_i) + k(\boldsymbol{x}, \boldsymbol{y})}{hh'} \\
&\to \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}, \boldsymbol{y})
\end{aligned} \tag{6}$$

as $h, h' \to 0$, for $\boldsymbol{x}, \boldsymbol{y} \in O$ close enough, since $\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k$ is assumed to be continuous in a neighbourhood of the diagonal. Thus we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h} - \frac{f(\boldsymbol{x}+h'\boldsymbol{e}_i)-f(\boldsymbol{x})}{h'}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}\right)^2\right] - 2\mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}\right)\left(\frac{f(\boldsymbol{x}+h'\boldsymbol{e}_i)-f(\boldsymbol{x})}{h'}\right)\right] \\
&\quad + \mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h'\boldsymbol{e}_i)-f(\boldsymbol{x})}{h'}\right)^2\right] \\
&\to \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}, \boldsymbol{x}) - 2\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}, \boldsymbol{x}) + \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}, \boldsymbol{x}) = 0
\end{aligned} \tag{7}$$

where the limit is taken as $h, h' \to 0$. Equation (7) shows that $\lim_{h\to 0} \frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}$ exists in L^2, i.e. that $\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x})$ exists. Moreover Equation (6) then implies that $\mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}) \partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{y})\right] = \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x},\boldsymbol{y})$ for $\boldsymbol{x},\boldsymbol{y} \in O$ close enough. Hence

$$
\begin{aligned}
\mathbb{E}\left[(\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}+\boldsymbol{h}) - \partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}))^2\right] &= \mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}+\boldsymbol{h})^2\right] - 2\mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}+\boldsymbol{h}) \partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x})\right] + \mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x})^2\right] \\
&= \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) - 2\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) + \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x},\boldsymbol{x}) \\
&\to 0
\end{aligned}
$$

as $\boldsymbol{h} \to \boldsymbol{0}$, so $\partial_{ms}^{\boldsymbol{e}_i} f$ is mean-square continuous.

(3) ⇒ (1): For $\boldsymbol{x},\boldsymbol{y} \in O$ and $1 \le i \le d$ we have

$$
\frac{k(\boldsymbol{x}+h\boldsymbol{e}_i,\boldsymbol{y}) - k(\boldsymbol{x},\boldsymbol{y})}{h} = \mathbb{E}\left[\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i) - f(\boldsymbol{x})}{h} f(\boldsymbol{y})\right] \to \mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}) f(\boldsymbol{y})\right]
$$

as $h \to 0$. Hence $\partial^{\boldsymbol{e}_i,\boldsymbol{0}} k(\boldsymbol{x},\boldsymbol{y})$ exists. Furthermore, for $1 \le j \le d$,

$$
\begin{aligned}
\frac{\partial^{\boldsymbol{e}_i,\boldsymbol{0}} k(\boldsymbol{x},\boldsymbol{y}+h\boldsymbol{e}_j) - \partial^{\boldsymbol{e}_i,\boldsymbol{0}} k(\boldsymbol{x},\boldsymbol{y})}{h} &= \mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}) \frac{f(\boldsymbol{y}+h\boldsymbol{e}_j) - f(\boldsymbol{y})}{h}\right] \\
&\to \mathbb{E}\left[\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}) \partial_{ms}^{\boldsymbol{e}_j} f(\boldsymbol{y})\right]
\end{aligned} \tag{8}
$$

as $h \to 0$. Hence $\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j} k$ exists, and is continuous by Equation (8), since $\partial_{ms}^{\boldsymbol{e}_i} f$ and $\partial_{ms}^{\boldsymbol{e}_j} f$ are mean-square continuous.

Now it remains to show that in all cases $\partial_{ms}^{\boldsymbol{e}_i} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k)$. Let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset O$ a finite set of points. Then we need to show

$$
(\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}_1), \dots, \partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}_N)) \sim \mathcal{N}\left(\boldsymbol{0}, (\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k(\boldsymbol{x}_p,\boldsymbol{x}_q))_{p,q=1}^N\right). \tag{9}
$$

$(\partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}_1), \dots, \partial_{ms}^{\boldsymbol{e}_i} f(\boldsymbol{x}_N))$ is the limit in $L^2(\Omega; \mathbb{R}^N)$ as $h \to 0$ of

$$
\begin{aligned}
&\left(\frac{f(\boldsymbol{x}_p+h\boldsymbol{e}_i) - f(\boldsymbol{x}_p)}{h}\right)_{p=1}^N \\
\sim \mathcal{N}\Bigg(\boldsymbol{0}, &\left(\frac{k(\boldsymbol{x}_p+h\boldsymbol{e}_i,\boldsymbol{x}_q+h\boldsymbol{e}_i) - k(\boldsymbol{x}_p+h\boldsymbol{e}_i,\boldsymbol{x}_q) - k(\boldsymbol{x}_p,\boldsymbol{x}_q+h\boldsymbol{e}_i) + k(\boldsymbol{x}_p,\boldsymbol{x}_q)}{h^2}\right)_{p,q=1}^N\Bigg).
\end{aligned} \tag{10}
$$

Now convergence in L^2 implies convergence in distribution, and we see that the multivariate normal distribution in Equation (10) converges to the multivariate normal distribution in Equation (9) as $h \to 0$. So this shows $\partial_{ms}^{\boldsymbol{e}_i} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k)$. ■

We can now prove our main result, Theorem 7, which we restate here for convenience:

Theorem 7 (Sample path Hölder regularity) *Let $n \in \mathbb{N}_0$ and $\gamma \in (0,1]$. The process $f \sim \mathcal{GP}(0,k)$ has samples in $C_{loc}^{(n+\gamma)^-}(O)$ if and only if,*

(1) for general k,

- $k \in C^{n\otimes n}(O \times O)$,

- $|\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as* $\boldsymbol{h} \to \boldsymbol{0}$*, locally uniformly in* $\boldsymbol{x} \in O$*, for all* $\epsilon \in (0,\gamma)$ *and* $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$.

(2) for stationary $k(\boldsymbol{x},\boldsymbol{y}) = k_\delta(\boldsymbol{x}-\boldsymbol{y})$,

- $k_\delta \in C^{2n}(\mathbb{R}^d)$,
- $|\partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}} k_\delta(\boldsymbol{0})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$ *as* $\boldsymbol{h} \to \boldsymbol{0}$ *for all* $\epsilon \in (0,\gamma)$ *and* $|\boldsymbol{\alpha}| = 2n$.

(3) for isotropic $k(\boldsymbol{x},\boldsymbol{y}) = k_r(\|\boldsymbol{x}-\boldsymbol{y}\|)$,

- $k_r \in C^{2n}(\mathbb{R})$,
- $|k_r^{(2n)}(h) - k_r^{(2n)}(0)| = \mathcal{O}(|h|^{2\epsilon})$ *as* $h \to 0$ *for all* $\epsilon \in (0,\gamma)$.

In each case, differentiating sample-wise we have $\partial^{\boldsymbol{\alpha}} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k)$ *for all* $|\boldsymbol{\alpha}| \le n$.

Proof We only prove the general case (1). Cases (2) and (3) then follow by taking intersections over $\epsilon \in (0,\gamma)$ in Lemma A.1.

$\Longleftarrow$**:** For $n = 0$ the result follows from applying the Kolmogorov continuity theorem to GPs, see Azmoodeh et al. (2014, Theorem 1), or Potthoff (2009, Page 347) for the result in arbitrary dimension d.

For $n = 1$, the existence and continuity of $\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k$ implies by Lemma A.3 (1) $\Rightarrow$ (3) that f is mean-square differentiable in direction i. Moreover this mean-square partial derivative satisfies $\partial^{\boldsymbol{e}_i}_{ms} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k)$. Now applying the case $n = 0$ to $\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k$ we deduce that $\partial^{\boldsymbol{e}_i}_{ms} f$ has samples in[2] $C^{\gamma^-}_{loc}(O)$. In particular it has continuous samples, and this implies by (Potthoff, 2010, Theorem 3.2) that f has differentiable samples in direction i, with $\partial^{\boldsymbol{e}_i} f(\boldsymbol{x}) = \partial^{\boldsymbol{e}_i}_{ms} f(\boldsymbol{x})$ almost surely, for all $\boldsymbol{x} \in O$. Thus $\partial^{\boldsymbol{e}_i} f \sim \mathcal{GP}(0, \partial^{\boldsymbol{e}_i,\boldsymbol{e}_i} k)$ and $\partial^{\boldsymbol{e}_i} f$ has samples in $C^{\gamma^-}_{loc}(O)$. $1 \le i \le d$ was arbitrary, so f has samples in $C^{(1+\gamma)^-}_{loc}(O)$.

For $n > 1$, we apply the same argument inductively on the partial derivatives.

$\Longrightarrow$**:** For $n = 0$, Azmoodeh et al. (2014, Theorem 1) gives the converse to Kolmogorov's theorem for GPs. To be precise, this theorem is only stated for dimension $d = 1$, but we can see that the proof generalises to arbitrary d.

Now suppose $n = 1$. Pick $1 \le i \le d$. For any finite set of points $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\} \subset O$, $(\partial^{\boldsymbol{e}_i} f(\boldsymbol{x}_1), \dots, \partial^{\boldsymbol{e}_i} f(\boldsymbol{x}_N))$ is the almost sure limit of centered multivariate Gaussians with distribution

$$\mathcal{N}\left(\boldsymbol{0}, \left(\frac{k(\boldsymbol{x}_p + h\boldsymbol{e}_i, \boldsymbol{x}_q + h\boldsymbol{e}_i) - k(\boldsymbol{x}_p + h\boldsymbol{e}_i, \boldsymbol{x}_q) - k(\boldsymbol{x}_p, \boldsymbol{x}_q + h\boldsymbol{e}_i) + k(\boldsymbol{x}_p, \boldsymbol{x}_q)}{h^2}\right)_{p,q=1}^{N}\right)$$

(see Equation (10)), so is itself a centered multivariate Gaussian distribution. Hence $\partial^{\boldsymbol{e}_i} f$ is a GP.

Let $\boldsymbol{x}, \boldsymbol{x} + h\boldsymbol{e}_i \in O$. By the mean value theorem, for each $\omega \in \Omega$ there is $\xi_{\boldsymbol{x},\omega}(h)$ between $\boldsymbol{x}$ and $\boldsymbol{x} + h\boldsymbol{e}_i$ such that

$$\frac{f(\boldsymbol{x} + h\boldsymbol{e}_i, \omega) - f(\boldsymbol{x}, \omega)}{h} = \partial^{\boldsymbol{e}_i} f(\xi_{\boldsymbol{x},\omega}(h), \omega).$$

2. The mean-square partial derivative $\partial^{\boldsymbol{e}_i}_{ms} f(\boldsymbol{x})$ at a point $\boldsymbol{x} \in O$ is only well-defined almost surely on Ω. But we still talk of sample paths for the process $\partial^{\boldsymbol{e}_i}_{ms} f$, since we mean this up to modification, as described in Section 2.

Thus

$$\left|\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i,\omega)-f(\boldsymbol{x},\omega)}{h}-\partial^{\boldsymbol{e}_i}f(\boldsymbol{x},\omega)\right|^2=|\partial^{\boldsymbol{e}_i}f(\xi_{\boldsymbol{x},\omega}(h),\omega)-\partial^{\boldsymbol{e}_i}f(\boldsymbol{x},\omega)|^2\leq C_\omega^2|h|^\gamma$$

since $\gamma\in(0,2\gamma)$, for some constant C_ω depending on ω but not on h, assuming h is small enough. Thus we have

$$\mathbb{E}\left[\left(\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}-\partial^{\boldsymbol{e}_i}f(\boldsymbol{x})\right)^2\right]\leq\underbrace{\mathbb{E}\left[C_\omega^2\right]}_{(*)}\cdot\underbrace{|h|^\gamma}_{(**)}$$

for h small enough. $(*)<\infty$ by Azmoodeh et al. (2014, Theorem 1), which we can apply since we showed that $\partial^{\boldsymbol{e}_i}f$ is a GP. Also $(**)\to 0$ as $h\to 0$. Hence

$$\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}\xrightarrow{L^2}\partial^{\boldsymbol{e}_i}f(\boldsymbol{x})$$

as $h\to 0$, i.e. f is mean-square differentiable in direction i. Hence by Lemma A.3 (3) $\Rightarrow$ (1), $\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j}k$ exists and is continuous for all $1\leq i,j\leq d$, and $\partial^{\boldsymbol{e}_i}f\sim\mathcal{GP}(0,\partial^{\boldsymbol{e}_i,\boldsymbol{e}_i}k)$. Now

$$\begin{aligned}&\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j}k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h})-\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j}k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x})-\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j}k(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h})+\partial^{\boldsymbol{e}_i,\boldsymbol{e}_j}k(\boldsymbol{x},\boldsymbol{x})\\&=\mathbb{E}\left[(\partial^{\boldsymbol{e}_i}f(\boldsymbol{x}+\boldsymbol{h})-\partial^{\boldsymbol{e}_i}f(\boldsymbol{x}))(\partial^{\boldsymbol{e}_j}f(\boldsymbol{x}+\boldsymbol{h})-\partial^{\boldsymbol{e}_j}f(\boldsymbol{x}))\right]\\&\leq\mathbb{E}\left[(\partial^{\boldsymbol{e}_i}f(\boldsymbol{x}+\boldsymbol{h})-\partial^{\boldsymbol{e}_i}f(\boldsymbol{x}))^2\right]^{1/2}\mathbb{E}\left[(\partial^{\boldsymbol{e}_j}f(\boldsymbol{x}+\boldsymbol{h})-\partial^{\boldsymbol{e}_j}f(\boldsymbol{x}))^2\right]^{1/2}\\&\leq\underbrace{\mathbb{E}\left[A_\omega^2\right]^{1/2}}_{<\infty}\|\boldsymbol{h}\|^\epsilon\cdot\underbrace{\mathbb{E}\left[B_\omega^2\right]^{1/2}}_{<\infty}\|\boldsymbol{h}\|^\epsilon=\mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})\end{aligned}\tag{11}$$

as $\boldsymbol{h}\to\mathbf{0}$, locally uniformly in $\boldsymbol{x}\in O$, for all $\boldsymbol{\epsilon}\in(0,\gamma)$, where A_ω, B_ω are some constants depending on ω, by Azmoodeh et al. (2014, Theorem 1) applied to $\partial^{\boldsymbol{e}_i}f$ and $\partial^{\boldsymbol{e}_j}f$, for $1\leq i,j\leq d$.

Finally, for $n>1$ we apply the same argument inductively on the partial derivatives. ■

Appendix B. Proofs in Section 4

B.1 Proof of Proposition 10

Proof For $\rho>0$ we can write for $\nu\notin\mathbb{N}$

$$K_\nu(\rho)=\frac{\pi}{2}\frac{I_{-\nu}(\rho)-I_\nu(\rho)}{\sin(\nu\pi)}$$

where

$$I_{\pm\nu}(\rho)=\left(\frac{\rho}{2}\right)^{\pm\nu}\sum_{j=0}^{\infty}\frac{1}{j!\,\Gamma(j\pm\nu+1)}\left(\frac{\rho}{2}\right)^{2j}$$

are modified Bessel functions of the first kind (Abramowitz and Stegun, 1965, Equations 9.6.2 & 9.6.10). So we can write the Matérn kernel (3) as

$$k_r(x) = C_\nu \Bigg(\underbrace{\sum_{j=0}^{\infty} \frac{1}{j!\,\Gamma(j-\nu+1)} \left(\frac{x}{2}\right)^{2j}}_{(*)} - \frac{|x|^{2\nu}}{2^\nu} \underbrace{\sum_{j=0}^{\infty} \frac{1}{j!\,\Gamma(j+\nu+1)} \left(\frac{x}{2}\right)^{2j}}_{(**)} \Bigg)$$

for $x \in \mathbb{R}$, where $C_\nu > 0$ is some constant. The power series $(*)$ and $(**)$ are smooth in x, and $|x|^{2\nu}$ has precisely $\lceil 2\nu - 1 \rceil$ continuous derivatives with $(2\nu - 2\lceil \nu - 1 \rceil)$-Hölder continuous $2\lceil \nu - 1 \rceil^{\text{th}}$ derivative at 0. So, for $\nu \notin \mathbb{N}$, the result follows by Theorem 7 (3).

For $\nu = n \in \mathbb{N}$, we instead use the following formula for the modified Bessel function of the second kind (Abramowitz and Stegun, 1965, Equations 9.6.11): for $\rho > 0$

$$K_n(\rho) = \frac{1}{2}\left(\frac{\rho}{2}\right)^{-n} \sum_{j=0}^{n-1} \frac{(n-j-1)!}{j!} \left(-\frac{\rho^2}{4}\right)^j + (-1)^{n+1} \log\left(\frac{\rho}{2}\right) I_n(\rho)$$
$$+ (-1)^n \frac{1}{2} \left(\frac{\rho}{2}\right)^n \sum_{j=0}^{\infty} \left(\psi(j+1) + \psi(n+j+1)\right) \frac{\left(\frac{\rho^2}{4}\right)^j}{j!(n+j)!}$$

where ψ is the digamma function. So we can write the Matérn kernel (3) as

$$k_r(x) = A_n \underbrace{\sum_{j=0}^{n-1} \frac{(n-j-1)!}{j!} (-1)^j \left(-\frac{x^2}{4}\right)^j}_{(I)} + B_n \, x^{2n} \underbrace{\log\left(\frac{x}{2}\right)}_{(II)} \underbrace{\sum_{j=0}^{\infty} \frac{1}{j!(n+j)!} \left(\frac{x}{2}\right)^{2j}}_{(III)}$$
$$+ C_n \underbrace{x^{2n} \sum_{j=0}^{\infty} \left(\psi(j+1) + \psi(n+j+1)\right) \frac{\left(\frac{x}{2}\right)^{2j}}{j!(n+j)!}}_{(IV)}$$

for some constants $A_n, B_n, C_n \neq 0$. (I) and (IV) are smooth in x. (III) is smooth in x and is non-zero at $x = 0$. (II) has precisely $2n-1$ continuous derivatives, with almost-2-Hölder continuous $(2n-2)^{\text{th}}$ derivative at 0 (precisely, a Hölder decay of $\mathcal{O}(h^2 \log h)$). So the result follows by Theorem 7 (3). ■

B.2 Proof of Proposition 12

Proof It is shown in (Wendland, 2004, Theorem 9.12) that the Wendland kernels may be written as

$$k_r(x) = \sum_{j=0}^{\lfloor d/2 \rfloor + 3n + 1} d_{j,n}^{(\lfloor d/2 \rfloor + n + 1)} |x|^j$$

for $x \in \mathbb{R}$, where $d_{j,n}^{(\lfloor d/2 \rfloor + n + 1)} \in \mathbb{R}$ are coefficients. Furthermore the odd degree coefficients satisfy $d_{2j+1,n}^{(\lfloor d/2 \rfloor + n + 1)} = 0$ if and only if $0 \leq j \leq 2n - 1$. Now for all $j \in \mathbb{N}_0$, $|x|^{2j}$ is smooth and $|x|^{2j+1}$ is precisely $2n$ times continuously differentiable with Lipschitz $2n^{\text{th}}$ derivative. So the result follows by Theorem 7 (3). ■

B.3 Proof of Proposition 15

Proof This follows from Theorem 7 (1), by noting that

$$\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{y}) = \partial^{\boldsymbol{\alpha}} \boldsymbol{\phi}(\boldsymbol{x})^T \partial^{\boldsymbol{\beta}} \boldsymbol{\phi}(\boldsymbol{y})$$

for all $\boldsymbol{x},\boldsymbol{y} \in O$ and $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq n$, and moreover when $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$,

$$\begin{aligned}
&|\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x})| \\
&= (\partial^{\boldsymbol{\alpha}} \boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}} \boldsymbol{\phi}(\boldsymbol{x}))^T (\partial^{\boldsymbol{\beta}} \boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\beta}} \boldsymbol{\phi}(\boldsymbol{x})) \\
&\leq \left(\sum_{i=1}^m |\partial^{\boldsymbol{\alpha}} \phi_i(\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha}} \phi_i(\boldsymbol{x})|^2 \right)^{1/2} \left(\sum_{i=1}^m |\partial^{\boldsymbol{\beta}} \phi_i(\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\beta}} \phi_i(\boldsymbol{x})|^2 \right)^{1/2} \\
&= \mathcal{O}(\|\boldsymbol{h}\|^{\epsilon}) \cdot \mathcal{O}(\|\boldsymbol{h}\|^{\epsilon}) = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})
\end{aligned}$$

as $\boldsymbol{h} \to \boldsymbol{0}$, locally uniformly in $\boldsymbol{x} \in O$, for all $\epsilon \in (0,\gamma)$, where we used the Cauchy-Schwarz inequality. ■

B.4 Proof of Proposition 18

Proof We assume $m = 2$; the proof for general m can then be done inductively. Note that by the product rule

$$\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{\gamma}_1+\boldsymbol{\gamma}_2=\boldsymbol{\alpha}} \sum_{\boldsymbol{\delta}_1+\boldsymbol{\delta}_2=\boldsymbol{\beta}} \binom{|\boldsymbol{\alpha}|}{|\boldsymbol{\gamma}_1|} \binom{|\boldsymbol{\beta}|}{|\boldsymbol{\delta}_1|} \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x},\boldsymbol{y}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x},\boldsymbol{y})$$

for all $\boldsymbol{x},\boldsymbol{y} \in O$ and $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \leq n$.

Now for $|\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = n$,

$$\begin{aligned}
&\partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) - \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\alpha},\boldsymbol{\beta}} k(\boldsymbol{x},\boldsymbol{x}) \\
&= \sum_{\boldsymbol{\gamma}_1+\boldsymbol{\gamma}_2=\boldsymbol{\alpha}} \sum_{\boldsymbol{\delta}_1+\boldsymbol{\delta}_2=\boldsymbol{\beta}} \binom{|\boldsymbol{\alpha}|}{|\boldsymbol{\gamma}_1|} \binom{|\boldsymbol{\beta}|}{|\boldsymbol{\delta}_1|} \Big(\partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) \\
&\quad - \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) - \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) \\
&\quad + \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x},\boldsymbol{x}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x},\boldsymbol{x}) \Big).
\end{aligned} \tag{12}$$

To show the expression in Equation (12) is $\mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$, it is therefore sufficient to show that

$$\begin{aligned}
&|\partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}+\boldsymbol{h}) - \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h},\boldsymbol{x}) \\
&\quad - \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x},\boldsymbol{x}+\boldsymbol{h}) + \partial^{\boldsymbol{\gamma}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x},\boldsymbol{x}) \partial^{\boldsymbol{\gamma}_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x},\boldsymbol{x})| = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})
\end{aligned} \tag{13}$$

as $\boldsymbol{h} \to \mathbf{0}$, locally uniformly in $\boldsymbol{x} \in O$, for all $\epsilon \in (0, \gamma)$, $\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_2 = \boldsymbol{\alpha}$ and $\boldsymbol{\delta}_1 + \boldsymbol{\delta}_2 = \boldsymbol{\beta}$. The inside of the absolute value on the left hand side of Equation (13) can be written as

$$\begin{aligned}
&\partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) \\
&\quad - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x}) \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}, \boldsymbol{x}) \\
&= \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) \\
&\quad \cdot \underbrace{\left(\partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x})\right)}_{(*)} \\
&\quad + \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x}) \\
&\quad \cdot \underbrace{\left(\partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}, \boldsymbol{x})\right)}_{(**)} \\
&\quad + \underbrace{\left(\partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x})\right)}_{(I)} \underbrace{\left(\partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x})\right)}_{(II)} \\
&\quad + \underbrace{\left(\partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x})\right)}_{(III)} \underbrace{\left(\partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_2,\boldsymbol{\delta}_2} k_2(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h})\right)}_{(IV)}
\end{aligned} \tag{14}$$

By arguing as in the proof of Theorem 7, Equation (11) we have $(*), (**) = \mathcal{O}(\|\boldsymbol{h}\|^{2\epsilon})$. Moreover, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
|(I)| &= \left|\mathbb{E}\left[\partial^{\gamma_1} f_1(\boldsymbol{x}+\boldsymbol{h}) \partial^{\boldsymbol{\delta}_1} f_1(\boldsymbol{x}) - \partial^{\gamma_1} f_1(\boldsymbol{x}) \partial^{\boldsymbol{\delta}_1} f_1(\boldsymbol{x})\right]\right| \\
&\leq \mathbb{E}\left[\partial^{\boldsymbol{\delta}_1} f_1(\boldsymbol{x})^2\right]^{1/2} \mathbb{E}\left[(\partial^{\gamma_1} f_1(\boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_1} f_1(\boldsymbol{x}))^2\right]^{1/2} \\
&= \partial^{\boldsymbol{\delta}_1,\boldsymbol{\delta}_1} k_1(\boldsymbol{x}, \boldsymbol{x})^{1/2} (\partial^{\gamma_1,\gamma_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}+\boldsymbol{h}) - \partial^{\gamma_1,\gamma_1} k_1(\boldsymbol{x}+\boldsymbol{h}, \boldsymbol{x}) \\
&\quad - \partial^{\gamma_1,\gamma_1} k_1(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}) + \partial^{\gamma_1,\gamma_1} k_1(\boldsymbol{x}, \boldsymbol{x}))^{1/2} \\
&= \mathcal{O}(\|\boldsymbol{h}\|^{\epsilon})
\end{aligned} \tag{15}$$

where $f_1 \sim \mathcal{GP}(0, k_1)$. Similarly $(II), (III), (IV) = \mathcal{O}(\|\boldsymbol{h}\|^{\epsilon})$. By Equation (14) we therefore deduce Equation (13), which concludes the proof. ■

References

Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables.* Courier Corporation, 1965. ISBN 978-0-486-61272-0.

Robert Adler and Jonathan Taylor. *Random Fields and Geometry.* Springer Monographs in Mathematics. Springer, 2007. ISBN 978-0-387-48116-6. doi:10.1007/978-0-387-48116-6.

Ehsan Azmoodeh, Tommi Sottinen, Lauri Viitasaari, and Adil Yazigi. Necessary and sufficient conditions for Hölder continuity of Gaussian processes. *Statistics & Probability Letters*, 94(C):230–235, 2014. doi:10.1016/j.spl.2014.07.030.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*, volume 100 of *Graduate Texts in Mathematics*. Springer, 1984. ISBN 978-1-4612-1128-0. doi:10.1007/978-1-4612-1128-0.

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Matérn Gaussian processes on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, volume 33, pages 12426–12437. Curran Associates, Inc., 2020.

N. N. Chentsov. Weak convergence of stochastic processes whose trajectories have no discontinuities of the second kind and the "heuristic" approach to the Kolmogorov-Smirnov tests. *Theory of Probability & Its Applications*, 1(1):140–144, 1956. ISSN 0040-585X. doi:10.1137/1101013.

Nathaël Da Costa, Cyrus Mostajeran, Juan-Pablo Ortega, and Salem Said. Invariant kernels on Riemannian symmetric spaces: a harmonic-analytic approach. 2023. doi:10.48550/arXiv.2310.19270.

Xavier Fernique. Régularité des trajectoires des fonctions aleatoires gaussiennes. In *Ecole d'Eté de Probabilités de Saint-Flour IV—1974*, Lecture Notes in Mathematics, pages 1–96. Springer, 1975. ISBN 978-3-540-37600-2. doi:10.1007/BFb0080190.

Karl Friston, Lancelot Da Costa, Noor Sajid, Conor Heins, Kai Ueltzhöffer, Grigorios A. Pavliotis, and Thomas Parr. The free energy principle made simpler but not too simple. *Physics Reports*, 1024:1–29, 2023. ISSN 0370-1573. doi:10.1016/j.physrep.2023.07.001.

Tilmann Gneiting. On the derivatives of radial positive definite functions. *Journal of Mathematical Analysis and Applications*, 236(1):86–93, 1999. ISSN 0022-247X. doi:10.1006/jmaa.1999.6434.

Conor Heins, Beren Millidge, Lancelot Da Costa, Richard Mann, Karl Friston, and Iain Couzin. Collective behavior from surprise minimization. 2023. doi:10.48550/arXiv.2307.14804.

Iain Henderson. Sobolev regularity of Gaussian random fields. *Journal of Functional Analysis*, 286(3):110241, 2024. ISSN 0022-1236. doi:10.1016/j.jfa.2023.110241.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K. Sriperumbudur. Gaussian processes and kernel methods: a review on connections and equivalences. 2018. doi:10.48550/arXiv.1807.02582.

Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, 2014. ISBN 978-1-4471-5361-0. doi:10.1007/978-1-4471-5361-0.

Marvin Pförtner, Ingo Steinwart, Philipp Hennig, and Jonathan Wenger. Physics-informed Gaussian process regression generalizes linear PDE solvers. 2022. doi:10.48550/arXiv.2212.12474.

Albert Pierce. *Fundamentals of Nonparametric Statistics*. Dickenson Publishing Company, 1970. ISBN 978-0822143109.

Jürgen Potthoff. Sample properties of random fields. II. Continuity. *Communications on Stochastic Analysis*, 3(3), 2009. ISSN 0973-9599. doi:10.31390/cosa.3.3.02.

Jürgen Potthoff. Sample properties of random fields III: differentiability. *Communications on Stochastic Analysis*, 4(3), 2010. ISSN 2688-6669. doi:10.31390/cosa.4.3.03.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. ISBN 978-0-262-25683-4. doi:10.7551/mitpress/3206.001.0001.

Michael Scheuerer. *A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis*. PhD thesis, 2010a.

Michael Scheuerer. Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120(10):1879–1897, 2010b. ISSN 0304-4149. doi:10.1016/j.spa.2010.05.009.

Michael Schober, David K Duvenaud, and Philipp Hennig. Probabilistic ODE solvers with Runge-Kutta means. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Michel Talagrand. Regularity of gaussian processes. *Acta Mathematica*, 159:99–149, 1987. ISSN 0001-5962, 1871-2509. doi:10.1007/BF02392556.

Junyang Wang, Jon Cockayne, Oksana Chkrebtii, T. J. Sullivan, and Chris. J. Oates. Bayesian numerical methods for nonlinear partial differential equations. *Statistics and Computing*, 31(5):55, 2021. ISSN 1573-1375.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2004. ISBN 978-0-521-84335-5. doi:10.1017/CBO9780511617539.