Comparing Scale Parameter Estimators for Gaussian Process Interpolation with the Brownian Motion Prior: Leave-One-Out Cross Validation and Maximum Likelihood*

Masha Naslidnyk[†], Motonobu Kanagawa[‡], Toni Karvonen[§], and Maren Mahsereci[¶]

Abstract.
Gaussian process (GP) regression is a Bayesian nonparametric method for regression and interpolation, offering a principled way of quantifying the uncertainties of predicted function values. For the quantified uncertainties to be well-calibrated, however, the kernel of the GP prior has to be carefully selected. In this paper, we theoretically compare two methods for choosing the kernel in GP regression: cross-validation and maximum likelihood estimation. Focusing on the scale-parameter estimation of a Brownian motion kernel in the noiseless setting, we prove that cross-validation can yield asymptotically well-calibrated credible intervals for a broader class of ground-truth functions than maximum likelihood estimation, suggesting an advantage of the former over the latter. Finally, motivated by the findings, we propose *interior cross validation*, a procedure that adapts to an even broader class of ground-truth functions.

Key words. gaussian processes, credible sets, cross-validation, model misspecification

MSC codes. 60G15, 62G20, 68T37, 65D05

1. Introduction.

Gaussian process (GP) regression (or kriging) is a Bayesian nonparametric method for regression and interpolation that has been extensively studied in statistics and machine learning (O'Hagan, 1978; Stein, 1999; Rasmussen and Williams, 2006). Its key property is that it enables uncertainty quantification of estimated function values in a principled manner, which is crucial for applications involving decision-making, safety concerns, and scientific discovery. As such, GP regression has been a core building block of more applied algorithms, including Bayesian optimisation (Jones et al., 1998; Shahriari et al., 2015; Garnett, 2023), probabilistic numerical computation (Hennig et al., 2015; Cockayne et al., 2019; Hennig et al., 2022), and calibration and emulation of computer models (Sacks et al., 1989; Kennedy and O'Hagan, 2001; O'Hagan, 2006; Beck and Guillas, 2016; Gu et al., 2018), to name just a few.

*Received by the editors July 14, 2023; accepted for publication (in revised form) July 23, 2024.

Funding: The work of the first author was supported by the U.K. Research and Innovation under grant number EP/S021566/1. The work of the second author was supported by the French government, through the 3IA Cote d'Azur Investment in the Future Project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The work of the third author was supported by the Academy of Finland postdoctoral researcher grant #338567 "Scalable, adaptive and reliable probabilistic integration". Part of this research was carried out during a visit by TK to EURECOM in May 2023 that was funded by the Institut français de Finlande, the Embassy of France to Finland, and the Finnish Society of Sciences and Letters. The work of the fourth author was supported 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.

[†]Department of Statistical Science, University College London, UK (masha.naslidnyk.21@ucl.ac.uk).

[‡]Data Science Department, EURECOM, France (motonobu.kanagawa@eurecom.fr).

[§]Department of Mathematics and Statistics, University of Helsinki, Finland (toni.karvonen@helsinki.fi).

[¶]Yahoo Research. Work performed while at the University of Tübingen, Germany (maren.mahsereci@yahooinc.com).

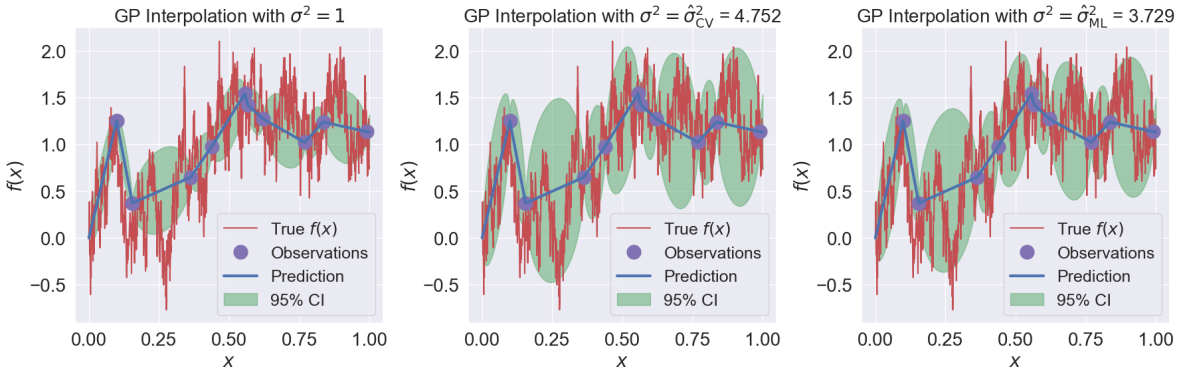

Figure 1. *GP interpolation of a fractional Brownian motion with the Hurst parameter $H = 0.2$ (smoothness $l + \alpha = 0.2$) using the Brownian motion kernel (1.2) with three different scale parameters: $\sigma^2 = 1$ (left), $\sigma^2 = \hat{\sigma}^2_{\mathrm{CV}} = 4.752$ given by the LOO-CV estimator (middle) and $\sigma^2 = \hat{\sigma}^2_{\mathrm{ML}} = 3.729$ obtained with the ML estimator (right). In each figure, the red trajectory represents the path of the fractional Brownian motion, the purple circles the training data, the blue curve the posterior mean $m_N(x)$ and the green shade the 95 % credible interval $[m_N(x) - 1.96\sigma\sqrt{k_N(x)}, m_N(x) + 1.96\sigma\sqrt{k_N(x)}]$.*

GP regression estimates an unknown function f from its observations as follows. One first defines a *prior distribution* for f as a GP by specifying its *kernel* (and mean function). Provided N observations about f, one then derives the *posterior distribution* of f, which is another GP with mean function m_N and kernel (or covariance function) k_N. One can then predict the function value $f(x)$ at any input x by the posterior mean $m_N(x)$ and quantify its uncertainty using the posterior standard deviation $\sqrt{k_N(x)} := \sqrt{k_N(x,x)}$. Specifically, one can construct a *credible interval* of $f(x)$ as the interval $[m_N(x) - \alpha\sqrt{k_N(x)}, m_N(x) + \alpha\sqrt{k_N(x)}]$ for a constant $\alpha > 0$ (for example, $\alpha \approx 1.96$ leads to the 95% credible interval). Such uncertainty estimates constitute key ingredients in the above applications of GP regression.

For GP uncertainty estimates to be reliable, the posterior standard deviation $\sqrt{k_N(x)}$ should, ideally, decay at the *same* rate as the prediction error $|m_N(x) - f(x)|$ decreases, with the increase of sample size N. Otherwise, GP uncertainty estimates are either asymptotically *overconfident* or *underconfident*. For example, if $\sqrt{k_N(x)}$ goes to 0 faster than the error $|m_N(x) - f(x)|$, then the credible interval $[m_N(x) - \alpha\sqrt{k_N(x)}, m_N(x) + \alpha\sqrt{k_N(x)}]$ will *not* contain the true value $f(x)$ as N increases for *any* fixed constant $\alpha > 0$ (asymptotically overconfident). If $\sqrt{k_N(x)}$ goes to 0 slower than the error $|m_N(x) - f(x)|$, then the confidence interval $[m_N(x) - \alpha\sqrt{k_N(x)}, m_N(x) + \alpha\sqrt{k_N(x)}]$ will get larger than the error $|m_N(x) - f(x)|$ as N increases (asymptotically underconfident). Both of these cases are not desirable in practice, as GP credible intervals will not be accurate estimates of prediction errors.

Unfortunately, in general, the posterior standard deviation $\sqrt{k_N(x)}$ does *not* decay at the same rate as the prediction error $|f(x) - m_N(x)|$, because, as is well-known, $\sqrt{k_N(x)}$ does *not* depend on the true function f; see (2.1b) in Section 2.1. Exceptionally, if the function f is a sample path of the GP prior (the well-specified case), GP uncertainty estimates can be well-calibrated. However, in general, the unknown f is not exactly a sample path of the GP prior (the misspecified case), and the posterior standard deviation $\sqrt{k_N(x)}$ does not scale with

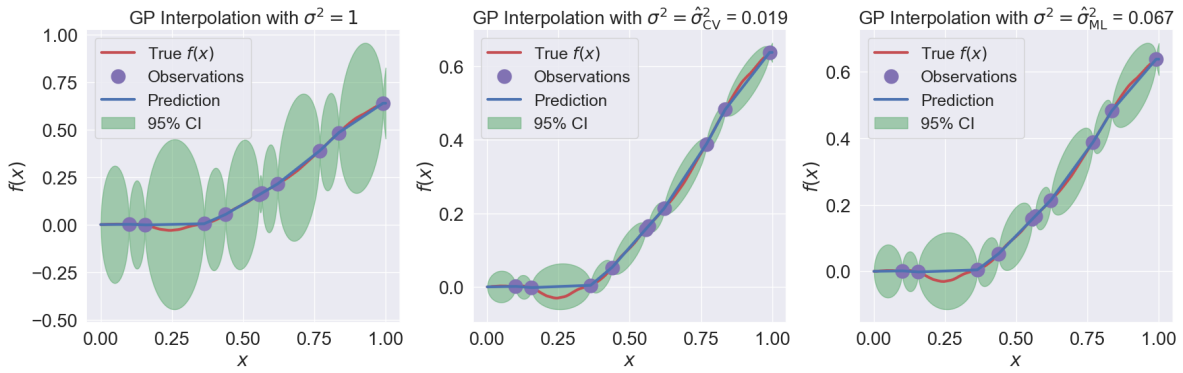

Figure 2. *GP interpolation of an integrated fractional Brownian motion with the Hurst parameter $H = 0.5$ (smoothness $l + \alpha = 1.5$) using the Brownian motion kernel (1.2) with three different scale parameters: $\sigma^2 = 1$ (left), $\sigma^2 = \hat{\sigma}^2_{\mathrm{CV}} = 0.019$ given by the LOO-CV estimator (middle) and $\sigma^2 = \hat{\sigma}^2_{\mathrm{ML}} = 0.067$ obtained with the ML estimator (right). For the explanation of the figures, see the caption of Figure 1.*

the prediction error $|f(x) - m_N(x)|$. Figures 1 and 2 (the left panels) show examples where the true function f is not a sample of the GP prior and where the GP uncertainty estimates are not well-calibrated.

1.1. Scale Parameter Estimation. To obtain sensible uncertainty estimates, one thus needs to adapt the posterior standard deviation $\sqrt{k_N(x)}$ to the function f. One simple way to achieve this is to introduce the *scale parameter* $\sigma^2 > 0$ and parametrize the kernel as

$$k_\sigma(x, x') := \sigma^2 k(x, x'),$$

where k is the original kernel. GP regression with this kernel k_σ yields the posterior mean function m_N, which is not influenced by σ^2, and the posterior covariance function $\sigma^2 k$, which is scaled by σ^2. If one estimates σ^2 from observed data of f, the estimate $\hat{\sigma}^2$ depends on f, and so does the resulting posterior standard deviation $\hat{\sigma}\sqrt{k_N(x)}$.

One approach to scale-parameter estimation is the method of *maximum likelihood (ML)*, which optimizes σ^2 to maximize the marginal likelihood of the GP (Rasmussen and Williams, 2006, Section 5.4). The ML approach is popular for general hyperparameter optimization in GP regression. Another less common way in the GP literature is *cross-validation (CV)*, which optimizes σ^2 to maximize the average predictive likelihood with held out data (Sundararajan and Keerthi, 2001). For either approach, the optimized scale parameter can be obtained analytically in computational complexity $\mathcal{O}(N^3)$. Figures 1 and 2 (middle and right panels) demonstrate that both approaches yield uncertainty estimates better calibrated than the original estimates without the scale parameter.

Do these scale parameter estimators lead to asymptotically well-calibrated uncertainty estimates? To answer this question, one needs to understand their convergence properties as the sample size N increases. Most existing theoretical works focus on the well-specified case where there is a "true" scale parameter σ_0^2 such that the unknown f is a GP with the kernel $\sigma_0^2 k$. In this case, both the ML and CV estimators have been shown to be consistent in estimating

the true σ_0^2 (e.g., Ying, 1991; Zhang, 2004; Bachoc et al., 2017, 2020). However, in general, no "true" scale parameter σ_0^2 exists such that the unknown f is a GP with the covariance $\sigma_0^2 k$. In such misspecified cases, not much is known about the convergence properties of both estimators. Karvonen et al. (2020) analyze the ML estimator for the scale parameter, assuming that f is a deterministic function. They derive upper bounds (and lower bounds in some cases) for the ML estimator; see Wang (2021) for closely related work. To our knowledge, no theoretical work exists for the CV estimator for the scale parameter in the misspecified case. Bachoc (2013) and Petit et al. (2022) empirically compare the ML and CV estimators under different model misspecification settings. We will review other related works in Section 1.3.

1.2. Contributions. This work studies the convergence properties of the ML and CV estimators, $\hat{\sigma}^2_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{CV}}$, of the scale parameter σ^2 in GP regression, to understand whether they lead to asymptotically well-calibrated uncertainty estimates. In particular, we provide the first theoretical analysis of the CV estimator $\hat{\sigma}^2_{\mathrm{CV}}$ when the GP prior is misspecified, and also establish novel results for the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$.

To facilitate the analysis, we focus on the following simplified setting. For a constant $T > 0$, let $[0, T] \subset \mathbb{R}$ be the input domain. Let k in (1.1) be the Brownian motion kernel

$$k(x, x') = \min(x, x') \quad \text{for} \quad x, x' \in [0, T].$$

With this choice, a sample path of the GP prior has roughly a smoothness of $1/2$ (in terms of the differentiability; we will be more rigorous in later sections).

We assume that the true unknown function f has the smoothness $l + \alpha$, where $l \in \{0\} \cup \mathbb{N}$ and $0 < \alpha \leq 1$. The GP prior has well-specified smoothness if $l = 0$ and $\alpha = 1/2$. Other settings of l and α represent misspecified cases. If $l = 0$ and $\alpha < 1/2$, the true function f is rougher than the GP prior (Figure 1); if $l = 0$ and $\alpha > 1/2$ or $l \geq 1$, the function f is smoother than the GP prior. We focus on the noise-free setting where one observes the function values $f(x_1), \ldots, f(x_N)$ at input points $x_1, \ldots, x_N \in [0, T]$.

Our main results are new upper and lower bounds for the asymptotic rates of the CV estimator $\hat{\sigma}^2_{\mathrm{CV}}$ and the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ as $N \to \infty$ (Section 4). The results suggest that the CV estimator can yield asymptotically well-calibrated uncertainty estimates for a broader class of functions f than the ML estimator; thus, the former has an advantage over the latter (Section 5). More specifically, asymptotically well-calibrated uncertainty estimates may be obtained with the CV estimator for the range $0 < l + \alpha \leq 3/2$ of smoothness of the true function, while this range becomes $0 < l + \alpha \leq 1$ with the ML estimator and is narrower. This finding is consistent with the example in Figure 2, where the true function has smoothness $l + \alpha = 3/2$ and is thus smoother than the GP prior. The uncertainty estimates of the CV estimator appear to be well-calibrated, while those of the ML estimator are unnecessarily wide, failing to adapt to the smoothness. Motivated by these insights, we propose a method called *interior cross-validation*, and show it accommodates an even wider range of smoothness of the true function than the CV estimator.

This paper is structured as follows. After reviewing related works in Subsection 1.3, we introduce the necessary background on the ML and CV approaches to scale parameter estimation for GP regression in Section 2. We describe the setting of the theoretical analysis in Section 3, present our main results in Section 4, and discuss its consequences on uncertainty quantification

in Section 5. We report simulation experiments in Section 6, conclude in Section 7, and present proofs in Section 8.

1.3. Related work. We review here related theoretical works on hyper-parameter selection in GP regression in the noiseless setting. We categorize them into two groups based on how the true unknown function f is modelled: random and deterministic.

Random setting. One group of works models the ground truth f as a random function, specifically as a GP. Most of these works model f as a GP with a Matérn-type kernel and analyze the ML estimator. Under the assumption that the GP prior is correctly specified, asymptotic properties of the ML estimator for the scale parameter and other parameters have been studied (Stein, 1990; Ying, 1991, 1993; Loh and Kam, 2000; Zhang, 2004; Loh, 2005; Du et al., 2009; Anderes, 2010; Wang and Loh, 2011; Kaufman and Shaby, 2013; Bevilacqua et al., 2019). Recently Loh et al. (2021) and Loh and Sun (2023) have constructed consistent estimators of various parameters for many commonly used kernels, including Matérns. Chen et al. (2021) and Petit (2023) consider a periodic version of Matérn GPs, and show the consistency of the ML estimator for its smoothness parameter. To our knowledge, the only existing theoretical result for ML estimation of the scale parameter in the misspecified random setting considers oversmoothing (Karvonen, 2021, Theorem 4.2). Oversmoothing refers to the situation where the chosen kernel is smoother than the true function. In Subsection 4.2 (Theorem 4.6), we provide a result for the undersmoothing case, which occurs when the chosen kernel is less smooth than the true function.

In contrast, few theoretical works exist for the CV estimator. Bachoc et al. (2017) study the leave-one-out (LOO) CV estimator for the Matérn-1/2 model (or the Laplace kernel) with one-dimensional inputs, in which case the GP prior is an Ornstein–Uhlenbeck (OU) process. Assuming the well-specified case where the true function is also an OU process, they prove the consistency and asymptotic normality of the CV estimator for the microergodic parameter in the fixed-domain asymptotic setting. Bachoc (2018) and Bachoc et al. (2020) discuss another CV estimator that uses the mean square prediction error as the scoring criterion of CV (thus different from the one discussed here) in the increasing-domain asymptotics. Bachoc (2013) and Petit et al. (2022) perform empirical comparisons of the ML and CV estimators under different model misspecification settings. Thus, to our knowledge, no theoretical result exists for the CV estimator of the scale parameter in the random misspecified setting, which we provide in Subsection 4.2 (Theorem 4.5).

Deterministic setting. Another line of research assumes that the ground truth f is a fixed function belonging to a specific function space (Stein, 1993). Xu and Stein (2017) assumed that the ground truth f is a monomial on $[0, 1]$ and proved some asymptotic results for the ML estimator when the kernel k is Gaussian. As mentioned earlier, Karvonen et al. (2020) proved asymptotic upper (and, in certain cases, also lower) bounds on the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$ of the scale parameter σ^2; see Wang (2021) for a closely related work. Karvonen (2023) has studied the ML and LOO-CV estimators for the smoothness parameter in the Matérn model; see also Petit (2023). Ben Salem et al. (2019) and Karvonen and Oates (2023) proved non-asymptotic results on the length-scale parameter in the Matérn and related models. Thus, there has been no work for the CV estimator of the scale parameter σ^2 in the deterministic setting, which we provide in Section 4.1 (Theorem 4.1); we also prove a corresponding result for the ML estimator

(Theorem 4.2).

2. Background.

This section briefly reviews GP regression and the ML and LOO-CV estimators of kernel parameters.

2.1. Gaussian process regression.

We first explain GP regression (or interpolation). Let Ω be a set, and $f\colon \Omega \to \mathbb{R}$ be an unknown function of interest. Suppose one observes N function values $f(x_1), \dots, f(x_N)$ at pairwise distinct input points $x_1, \dots, x_N \in \Omega$. The task here is to estimate f based on the data $(\mathbf{x}, f(\mathbf{x}))$, where $f(\mathbf{x}) \coloneqq [f(x_1), \dots, f(x_N)]^\top \in \mathbb{R}^N$ and $\mathbf{x} \coloneqq [x_1, \dots, x_N]^\top \in \Omega^N$.

In GP regression, one first defines a prior distribution of the unknown f as a GP by specifying its mean function $m\colon \Omega \to \mathbb{R}$ and covariance function (kernel) $k\colon \Omega \times \Omega \to \mathbb{R}$; we may write $f \sim \mathcal{GP}(m, k)$ to indicate this. Conditioned on the data $(\mathbf{x}, f(\mathbf{x}))$, the posterior distribution of f is again a GP whose mean function $m_N : \Omega \to \mathbb{R}$ and covariance function $k_N : \Omega \times \Omega \to \mathbb{R}$ are given by

$$m_N(x) \coloneqq m(x) + k(x, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} \left(f(\mathbf{x}) - m(\mathbf{x})\right), \quad x \in \Omega, \tag{2.1a}$$

$$k_N(x, x') \coloneqq k(x, x') - k(x, \mathbf{x})^\top k(\mathbf{x}, \mathbf{x})^{-1} k(x', \mathbf{x}), \quad x, x' \in \Omega, \tag{2.1b}$$

where $m(\mathbf{x}) \coloneqq [m(x_1), \dots, m(x_N)]^\top \in \mathbb{R}^N$ and $k(x, \mathbf{x}) \coloneqq [k(x, x_1), \dots, k(x, x_N)]^\top \in \mathbb{R}^N$, and

$$k(\mathbf{x}, \mathbf{x}) \coloneqq \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \dots & k(x_N, x_N) \end{bmatrix} \in \mathbb{R}^{N \times N}$$

is the Gram matrix. Throughout this paper, we assume that the points $\mathbf{x}$ are such that the Gram matrix is non-singular. For notational simplicity, we may write the posterior variance as

$$k_N(x) \coloneqq k_N(x, x), \quad x \in \Omega.$$

For simplicity and as commonly done, we henceforth assume that the prior mean function m is the zero function, $m(\cdot) \equiv 0$.

While the GP prior assumes that the unknown function f is a sample path of the GP with the specified kernel k, this assumption does not hold in general, i.e., model misspecification occurs. In this case, as described in Figures 1 and 2 (left), the posterior standard deviation $\sqrt{k_N(x)}$, which is supposed to quantify the uncertainty of the unknown function value $f(x)$, may not be well calibrated with the prediction error $|m_N(x) - f(x)|$. One could address this issue by selecting the kernel k or its parameters from the data $(\mathbf{x}, f(\mathbf{x}))$; we will explain this topic next.

2.2. Kernel parameter estimation.

The selection of the kernel k is typically performed by defining a parametric family of kernels $\{k_\theta\}_{\theta \in \Theta}$ and selecting the parameter θ based on an appropriate criterion. Here Θ is a parameter set, and $k_\theta : \Omega \times \Omega \to \mathbb{R}$ for each $\theta \in \Theta$ is a kernel.

Maximum likelihood (ML) estimation. The ML estimator maximises the log-likelihood of the GP f with kernel k_θ under the data $(\mathbf{x}, f(\mathbf{x}))$:

$$\log p(f(\mathbf{x}) \,|\, \mathbf{x}, \theta) = -\frac{1}{2}\left(f(\mathbf{x})^\top k_\theta(\mathbf{x}, \mathbf{x})^{-1} f(\mathbf{x}) + \log\det k_\theta(\mathbf{x}, \mathbf{x}) + n \log(2\pi) \right),$$

where $\det k_\theta(\mathbf{x}, \mathbf{x})$ is the determinant of the Gram matrix $k_\theta(\mathbf{x}, \mathbf{x})$ (see, e.g., Rasmussen and Williams 2006, Section 5.4.1). With the additive terms that do not depend on θ removed from $\log p(f(\mathbf{x}) \mid \mathbf{x}, \theta)$, this is equivalent to minimising the loss function

$$\mathcal{L}_{\mathrm{ML}}(\theta) := f(\mathbf{x})^\top k_\theta(\mathbf{x}, \mathbf{x})^{-1} f(\mathbf{x}) + \log \det k_\theta(\mathbf{x}, \mathbf{x}).$$

In general, $\mathcal{L}_{\mathrm{ML}}(\theta)$ may not have a unique minimiser, so that any ML estimator satisfies

$$\hat{\theta}_{\mathrm{ML}} \in \arg\min_{\theta \in \Theta} \mathcal{L}_{\mathrm{ML}}(\theta).$$

Leave-one-out cross-validation (LOO-CV). The LOO-CV estimator (e.g., Rasmussen and Williams, 2006, Section 5.4.2), which we may simply call the CV estimator, is an alternative to the ML estimator. It maximizes the average log-predictive likelihood

$$\sum_{n=1}^{N} \log p(f(x_n) \mid x_n, \mathbf{x}_{\setminus n}, f(\mathbf{x}_{\setminus n}), \theta)$$

with held-out data $(x_n, f(x_n))$, where $n = 1, \ldots, N$, based on the data $(\mathbf{x}_{\setminus n}, f(\mathbf{x}_{\setminus n}))$, where $\mathbf{x}_{\setminus n}$ denotes the input points with x_n removed:

$$\mathbf{x}_{\setminus n} = [x_1, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N]^\top \in \Omega^{N-1}.$$

Let $m_{\theta,\setminus n}$ and $k_{\theta,\setminus n}$ denote the posterior mean and covariance functions of GP regression with the kernel k_θ and the data $(\mathbf{x}_{\setminus n}, f(\mathbf{x}_{\setminus n}))$. Because each $p(f(x_n) \mid x_n, \mathbf{x}_{\setminus n}, f(\mathbf{x}_{\setminus n}), \theta)$ is the Gaussian density of $f(x_n)$ with mean $m_{\theta,\setminus n}(x_n)$ and variance $k_{\theta,\setminus n}(x_n) := k_{\theta,\setminus n}(x_n, x_n)$, removing additive terms that do not depend on θ and reversing the sign in (2.2) yields the following CV objective function:

$$\mathcal{L}_{\mathrm{CV}}(\theta) = \sum_{n=1}^{N} \frac{\left[f(x_n) - m_{\theta,\setminus n}(x_n)\right]^2}{k_{\theta,\setminus n}(x_n)} + \log k_{\theta,\setminus n}(x_n).$$

The CV estimator is then defined as its minimizer:

$$\hat{\theta}_{\mathrm{CV}} \in \arg\min_{\theta \in \Theta} \mathcal{L}_{\mathrm{CV}}(\theta).$$

As for the ML estimator, the CV objective function and its first-order gradients can be computed in closed form in $\mathcal{O}(N^3)$ time (Sundararajan and Keerthi, 2001).

Scale parameter estimation. As explained in Section 1, we consider the family of kernels $k_\sigma(x, x') := \sigma^2 k(x, x')$ parametrized with the scale parameter $\sigma^2 > 0$, where k is a fixed kernel, and study the estimation of σ^2 using the CV and ML estimators, denoted as $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$, respectively. In this case, both $\hat{\sigma}^2_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{CV}}$ can be derived in closed form by differentiating (2.2) and (2.2).

Let m_{n-1} and k_{n-1} be the posterior mean and variance functions of GP regression using the kernel k and the first $n-1$ training observations $(x_1, f(x_1)), \ldots, (x_{n-1}, f(x_{n-1}))$. Let

$m_0(\cdot) \coloneqq 0$ and $k_0(x,x) \coloneqq k(x,x)$. Then the ML estimator is given by

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{f(\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} f(\mathbf{x})}{N} = \frac{1}{N}\sum_{n=1}^{N} \frac{[f(x_n) - m_{n-1}(x_n)]^2}{k_{n-1}(x_n)}.$$

This expression of the ML estimator is relatively well known; see e.g. Section 4.2.2 in Xu and Stein (2017) or Proposition 7.5 in Karvonen and Oates (2023).

On the other hand, the CV estimator $\hat{\sigma}^2_{\mathrm{CV}}$ is given by

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N}\sum_{n=1}^{N} \frac{\left[f(x_n) - m_{\backslash n}(x_n)\right]^2}{k_{\backslash n}(x_n)}, \tag{2.2}$$

where $m_{\backslash n}$ and $k_{\backslash n}$ are the posterior mean and covariance functions of GP regression using the kernel k and data $(\mathbf{x}_{\backslash n}, f(\mathbf{x}_{\backslash n}))$ with $(x_n, f(x_n))$ removed:

$$\begin{aligned} m_{\backslash n}(x) &= k(\mathbf{x}_{\backslash n}, x)^\top k(\mathbf{x}_{\backslash n}, \mathbf{x}_{\backslash n})^{-1} f(\mathbf{x}_{\backslash n}), \\ k_{\backslash n}(x,x') &= k(x,x') - k(\mathbf{x}_{\backslash n}, x)^\top k(\mathbf{x}_{\backslash n}, \mathbf{x}_{\backslash n})^{-1} k(\mathbf{x}_{\backslash n}, x'). \end{aligned}$$

Notice the similarity between the two expressions (2.2) and (2.2). The difference is that the ML estimator uses k_{n-1} and m_{n-1}, which are based on the first $n-1$ training observations, while the CV estimator uses $k_{\backslash n}$ and $m_{\backslash n}$ obtained with $N-1$ observations, for each $n = 1, \dots, N$. Therefore, the CV estimator uses all the data points more evenly than the ML estimator. This difference may be the source of the difference in their asymptotic properties established later.

Remark 2.1. As suggested by the similarity between (2.2) and (2.2), there is a deeper connection between ML and CV estimators in general. For instance, Fong and Holmes (2020, Proposition 2) have shown that the Bayesian marginal likelihood equals the average of leave-p-out CV scores. We prove this result for the special case of scale parameter estimation in GP regression in Appendix A. Another notable example is the work in Ginsbourger and Schärer (2021), where the authors showed that, when corrected for the covariance of residuals, the CV estimator of the scale parameter reverts to MLE.

3. Setting.

This section describes the settings and tools for our theoretical analysis: the Brownian motion kernel in Section 3.1; sequences of partitions in Section 3.2; the Hölder class of functions in Section 3.3; fractional Brownian motion in Section 3.4; and functions of finite quadratic variation in Section 3.5.

3.1. Brownian motion kernel.

As explained in Section 1, for the kernel k we focus on the Brownian motion kernel on the domain $\Omega = [0, T]$ for some $T > 0$:

$$k(x,x') = \min(x,x').$$

The resulting kernel $k_\sigma(x,x') = \sigma^2 k(x,x')$ induces a Brownian motion prior for GP regression.

We assume that the input points $\mathbf{x} = [x_1, \dots x_N]^\top$ for GP regression are positive and ordered:

$$0 < x_1 < x_2 < \cdots < x_N \leq T.$$

The positivity ensures that the Gram matrix (2.1) is non-singular; the proof is given in Subsection 8.1.

As is well known (see, for instance, Diaconis, 1988, Example 1) and can be seen in Figures 1 and 2, the posterior mean function m_N in (2.1) using the Brownian motion kernel becomes the *piecewise linear interpolant* of the observations $(\mathbf{x}, f(\mathbf{x}))$. See (8.1) and (8.1) in Section 8.1 for the proof and explicit expressions of the posterior mean and covariance functions.

3.2. Sequences of partitions. For our asymptotic analysis, we assume that the input points $x_1, \ldots, x_N \in [0, T]$ cover the domain $[0, T]$ more densely as the sample size N increases. To make the dependence on the size N explicit, we write $\mathcal{P}_N \coloneqq (x_{N,n})_{n=1}^N \subset [0, T]$ as a point set of size N, and assume that they are ordered as

$$0 =: x_{N,0} < x_{N,1} < x_{N,2} < \cdots < x_{N,N} = T$$

Then $\mathcal{P}_N$ defines a partition of $[0, T]$ into N subintervals $[x_{N,n}, x_{N,n+1}]$. When there is no risk of confusion, we may write x_n instead of $x_{N,n}$ for simplicity. Note that we do *not* require the nesting $\mathcal{P}_N \subset \mathcal{P}_{N+1}$ of partitions.

We define the *mesh size* of partition $\mathcal{P}_N$ as the longest subinterval in the partition:

$$\|\mathcal{P}_N\| \coloneqq \max_{n \in \{0,1,\ldots,N-1\}} (x_{N,n+1} - x_{N,n})$$

The decay rate of the mesh size $\|\mathcal{P}_N\|$ quantifies how quickly the points in $\mathcal{P}_N$ cover the interval $[0, T]$. In particular, the decay rate $\mathcal{P}_N = \mathcal{O}(N^{-1})$ implies that the length of every subinterval is asymptotically upper bounded by $1/N$. At the same time, if each subinterval is asymptotically lower bounded by $1/N$, we call the sequence of partitions $(\mathcal{P}_N)_{N \in \mathbb{N}}$ *quasi-uniform*, more formally defined in Wendland (2005, Definition 4.6) as follows.

Definition 3.1. *For each $N \in \mathbb{N}$, let $\mathcal{P}_N \coloneqq (x_{N,n})_{n=1}^N \subset [0, T]$. Define $\Delta x_{N,n} \coloneqq x_{N,n+1} - x_{N,n}$. Then the sequence of partitions $(\mathcal{P}_N)_{N \in \mathbb{N}}$ is called* quasi-uniform *if there exists a constant $1 \leq C_{\mathrm{qu}} < \infty$ such that*

$$\sup_{N \in \mathbb{N}} \frac{\max_n \Delta x_{N,n}}{\min_n \Delta x_{N,n}} = C_{\mathrm{qu}}.$$

Quasi-uniformity, as defined here, requires that the ratio of the longest subinterval, $\max_n \Delta x_{N,n}$, to the shortest one, $\min_n \Delta x_{N,n}$, is upper-bounded by C_{qu} for all $N \in \mathbb{N}$. Since $\min_n \Delta x_{N,n} \leq TN^{-1}$ and $\max_n \Delta x_{N,n} \geq TN^{-1}$ for any partition of $[0, T]$, quasi-uniformity implies that all subintervals are asymptotically upper and lower bounded by $1/N$, as we have, for all $N \in \mathbb{N}$ and $n_0 \in \{0, \ldots, N-1\}$,

$$\frac{TN^{-1}}{C_{\mathrm{qu}}} \leq \min_n \Delta x_{N,n} \leq \Delta x_{N,n_0} \leq \max_n \Delta x_{N,n} \leq TC_{\mathrm{qu}} N^{-1}.$$

Therefore, quasi-uniform sequences of partitions are *space-filling designs* that cover the space "almost" uniformly. Trivially, equally-spaced points (or uniform grids) satisfy the quasi-uniformity with $C_{\mathrm{qu}} = 1$. Wenzel et al. (2021) showed that points chosen sequentially to minimise GP posterior variance for a Sobolev kernel are quasi-uniform. We refer to Wynne et al. (2021, p. 6) for further examples and a discussion on quasi-uniformity.

3.3. Hölder spaces. Subsection 4.1 studies the deterministic setting where the true unknown function f is assumed to belong to a Hölder space of functions. To define this space, we first need the following definition.

Definition 3.2. *For $0 < \alpha \leq 1$, a function $f : [0,T] \to \mathbb{R}$ is α-Hölder continuous if there exists a constant $L \geq 0$ such that, for all $x, x' \in [0,T]$,*

$$|f(x) - f(x')| \leq L|x - x'|^{\alpha}.$$

Any such constant L is called a Hölder constant *of f.*

For $l \in \mathbb{N} \cup \{0\}$, denote by $C^l([0,T])$ the space of functions $f\colon [0,T] \to \mathbb{R}$ such that the l^{th} derivative $f^{(l)}$ exists and is continuous. For $l = 0$, this is the space of continuous functions. Hölder spaces are now defined as follows.

Definition 3.3. *Let $l \in \mathbb{N} \cup \{0\}$ and $0 < \alpha \leq 1$. The Hölder space $C^{l,\alpha}([0,T])$ consists of functions $f \in C^l([0,T])$ whose l^{th} derivative $f^{(l)}$ is α-Hölder continuous.*

Intuitively, $l + \alpha$ represents the smoothness of least-smooth functions in $C^{l,\alpha}([0,T])$. It is well known that a sample path of Brownian motion is almost surely α-Hölder continuous if and only if $\alpha < 1/2$ (e.g., Mörters and Peres, 2010, Corollary 1.20), and thus it belongs to the Hölder space $C^{l,\alpha}([0,T])$ with $l = 0$ and $\alpha = 1/2 - \varepsilon$ almost surely for arbitrarily small $\varepsilon > 0$; in this sense, the smoothness of a Brownian motion is $1/2$. As such, as is well known (e.g., Mörters and Peres, 2010, Theorem 1.27), a Brownian motion is almost nowhere differentiable almost surely.

Note that we have the following strict inclusions:[1]

- $C^{l_1,\alpha_1}([0,T]) \subsetneq C^{l_2,\alpha_2}([0,T])$ if (a) $l_1 > l_2$ or (b) $l_1 = l_2$ and $\alpha_1 > \alpha_2$,
- $C^{l+1}([0,T]) \subsetneq C^{l,1}([0,T])$.

3.4. Fractional Brownian motion. Subsection 4.2 considers the random setting where f is a *fractional (or integrated fractional) Brownian motion* (see Mandelbrot (e.g., 1982, Chapter IX)). Examples of these processes can be seen in Figures 1, 2, 5 and 6.

A fractional Brownian motion on $[0,T]$ with Hurst parameter $0 < H < 1$ is a Gaussian process whose kernel is given by

$$k_{0,H}(x,x') = \left(|x|^{2H} + |x'|^{2H} - |x - x'|^{2H} \right)/2.$$

Note that if $H = 1/2$, this is the Brownian motion kernel: $k_{0,1/2}(x,x') = \min(x,x')$. The Hurst parameter H quantifies the smoothness of the fractional Brownian motion. If $f_{\text{FBM}} \sim \mathcal{GP}(0, k_{0,H})$ for $H \in (0,1)$, then $f_{\text{FBM}} \in C^{0,H-\varepsilon}([0,T])$ almost surely for arbitrarily small $\varepsilon > 0$ (e.g., Nourdin, 2012, Proposition 1.6).[2]

[1] These inclusions follow from the following facts: By the definition of Hölder continuity, an α_1-Hölder continuous function is α_2-Hölder continuous if $\alpha_1 > \alpha_2$; continuously differentiable functions are α-Hölder continuous for any $0 < \alpha \leq 1$; not all Lipschitz functions are differentiable.

[2] That $f_{\text{FBM}} \notin C^{0,H}([0,T])$ almost surely for $f_{\text{FBM}} \sim \mathcal{GP}(0, k_{0,H})$ with $H \in (0,1)$ is a straightforward corollary of, for example, Theorem 3.2 in Wang (2007).

An integrated Brownian motion with Hurst parameter H is defined via the integration of a fractional Brownian motion with the same Hurst parameter: if $f_{\mathrm{FBM}} \sim \mathcal{GP}(0, k_{0,H})$, then

$$f_{\mathrm{iFBM}}(x) = \int_0^x f_{\mathrm{FBM}}(z)\,\mathrm{d}z, \quad x \in [0,T]$$

is an integrated Brownian motion with Hurst parameter H. It is a zero-mean GP with the kernel

$$\begin{aligned} k_{1,H}(x,x') &= \int_0^x \int_0^{x'} \left(|z|^{2H} + |z'|^{2H} - |z-z'|^{2H}\right)/2\,\mathrm{d}z\,\mathrm{d}z' \\ &= \frac{1}{2(2H+1)}\Bigg(x'x^{2H+1} + x(x')^{2H+1} \\ &\qquad - \frac{1}{2(H+1)}\left[x^{2H+2} + (x')^{2H+2} - |x-x'|^{2H+2}\right]\Bigg). \end{aligned} \tag{3.1}$$

Because differentiating an integrated fractional Brownian motion $f_{\mathrm{iFBM}} \sim \mathcal{GP}(0, k_{1,H})$ yields a fractional Brownian motion $f_{\mathrm{FBM}} \sim \mathcal{GP}(0, k_{0,H})$, a sample path of the former satisfies $f_{\mathrm{iFBM}} \in C^{1,H-\varepsilon}([0,T])$ almost surely for arbitrarily small $\varepsilon > 0$; therefore the smoothness of f_{iFBM} is $1+H$.

3.5. Functions of finite quadratic variation. Some of our asymptotic results use the notion of functions of *finite quadratic variation*, defined below.

Definition 3.4. *For each $N \in \mathbb{N}$, let $\mathcal{P}_N \coloneqq (x_{N,n})_{n=1}^N \subset [0,T]$, and suppose that $\|\mathcal{P}_N\| \to 0$ as $N \to \infty$. Then a function $f : [0,T] \to \mathbb{R}$ is defined to have* finite quadratic variation *with respect to $\mathcal{P} \coloneqq (\mathcal{P}_N)_{N\in\mathbb{N}}$, if the limit*

$$V^2(f) \coloneqq \lim_{N\to\infty} \sum_{n=0}^{N-1} \left[f(x_{N,n+1}) - f(x_{N,n})\right]^2$$

exists and is finite. We write $V^2(f,\mathcal{P})$ when it is necessary to indicate the sequence of partitions.

Quadratic variation is defined for a specific sequence of partitions $(\mathcal{P}_N)_{N\in\mathbb{N}}$ and may take different values for different sequences of partitions (Mörters and Peres, 2010, Remark 1.36). For conditions that guarantee the invariance of quadratic variation on the sequence of partitions, see, for instance, Cont and Bas (2023). Note also that the notion of quadratic variation differs from that of p-variation for $p = 2$, which is defined as the supremum over all possible sequences of partitions whose mesh sizes tend to zero.

If $f \in C^{0,\alpha}([0,T])$ with $\alpha > 1/2$ and $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$ as $N \to \infty$, then we have $V^2(f) = 0$, because in this case

$$\sum_{n=0}^{N-1} \left[f(x_{N,n+1}) - f(x_{N,n})\right]^2 \le NL^2 \max_n (\Delta x_{N,n})^{2\alpha} = \mathcal{O}(N^{1-2\alpha}) \to 0$$

as $N \to \infty$. Therefore, given the inclusion properties of Hölder spaces (see Section 3.3), we arrive at the following standard proposition.

Proposition 3.5. *Suppose that the partitions* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *are such that* $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$. *If* $f \in C^{l,\alpha}([0,T])$ *for* $l+\alpha > 1/2$, *then* $V^2(f) = 0$.

If the mesh size tends to zero faster than $1/\log N$, in that $\|\mathcal{P}_N\| = o(1/\log N)$, then the quadratic variation of almost every sample path of the Brownian motion on the interval $[0,T]$ equals T (Dudley, 1973). This is of course true for partitions which have the faster decay $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$.

4. Main results.

This section presents our main results on the asymptotic properties of the CV and ML estimators, $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$, for the scale parameter. Subsection 4.1 considers the deterministic setting where the true function f is fixed and assumed to belong to a Hölder space. Subsection 4.2 studies the random setting where f is an (integrated) fractional Brownian motion. In Subsection 4.3, we use the insights obtained in the proofs for the deterministic and random settings to propose a *interior cross-validation* (ICV) estimator, and show its asymptotic properties are an improvement on those of CV and ML estimators.

4.1. Deterministic setting.

We present our main results for the deterministic case where the true function f is fixed and assumed to be in a Hölder space $C^{l,\alpha}([0,T])$. Theorem 4.1 below provides asymptotic upper bounds on the CV estimator $\hat{\sigma}^2_{\mathrm{CV}}$ for different values of the smoothness parameters l and α of the Hölder space.

Theorem 4.1 (Rate of CV decay in Hölder spaces). *Suppose that* f *is an element of* $C^{l,\alpha}([0,T])$, *with* $l \geq 0$ *and* $0 < \alpha \leq 1$, *such that* $f(0) = 0$, *and the interval partitions* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *have bounded mesh sizes* $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$ *as* $N \to \infty$. *Then*

$$\hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}\big(N^{1-\min\{2(l+\alpha),3\}}\big) = \begin{cases} \mathcal{O}\left(N^{1-2\alpha}\right) & \text{if} \quad l = 0, \\ \mathcal{O}\left(N^{-1-2\alpha}\right) & \text{if} \quad l = 1 \text{ and } \alpha < 1/2, \\ \mathcal{O}\left(N^{-2}\right) & \text{if} \quad l = 1 \text{ and } \alpha \geq 1/2, \\ \mathcal{O}\left(N^{-2}\right) & \text{if} \quad l \geq 2. \end{cases}$$

Proof. See Section 8.2. ∎

Theorem 4.2 below is a corresponding result for the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$. Note that a similar result has been obtained by Karvonen et al. (2020, Proposition 4.5), where the function f is assumed to belong to a Sobolev space and the kernel is a Matérn-type kernel. Theorem 4.2 is a version of this result where f is in a Hölder space and the kernel is the Brownian motion kernel; we provide it for completeness and ease of comparison.

Theorem 4.2 (Rate of ML decay in Hölder spaces). *Suppose that* f *is a non-zero element of* $C^{l,\alpha}([0,T])$, *with* $l \geq 0$ *and* $0 < \alpha \leq 1$, *such that* $f(0) = 0$, *and the interval partitions* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *have bounded mesh sizes* $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$ *as* $N \to \infty$. *Then*

$$\hat{\sigma}^2_{\mathrm{ML}} = \mathcal{O}\big(N^{1-\min\{2(l+\alpha),2\}}\big) = \begin{cases} \mathcal{O}\left(N^{1-2\alpha}\right) & \text{if} \quad l = 0, \\ \Theta\left(N^{-1}\right) & \text{if} \quad l \geq 1. \end{cases}$$

Proof. See Section 8.2. The proof is similar to that of Theorem 4.1. ∎

Figure 3 summarises the rates of Theorems 4.1 and 4.2. When $l + \alpha \leq 1$ (or $l = 0$ and $\alpha \leq 1$), the rates of $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ are $\mathcal{O}(N^{1-2\alpha})$, so both of them may decay (or grow, for

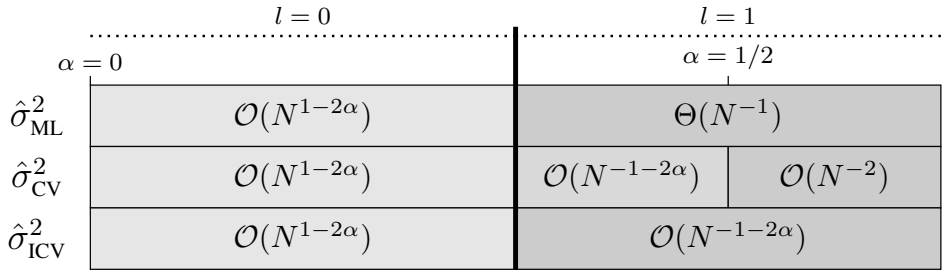

Figure 3. *Rates of decay for the ML, CV and ICV estimators from Theorems 4.1, 4.2, and 4.8. Observe that the CV estimator's range of adaptation to the smoothness $l + \alpha$ is wider than the ML estimator's, and the ICV estimator's range of adaptation is wider than that for both the CV and ML estimators.*

$l + \alpha < 1/2$) adaptively to the smoothness $l + \alpha$ of the function f. However, when $l + \alpha > 1$, the situation is different: the decay rate of $\hat{\sigma}^2_{\mathrm{ML}}$ is always $\Theta(N^{-1})$ and thus insensitive to α, while that of $\hat{\sigma}^2_{\mathrm{CV}}$ is $\left(N^{-1-2\alpha}\right)$ for $l = 1$ and $\alpha \in (0, 1/2]$. Therefore the CV estimator may be adaptive to a broader range of the smoothness $0 < l + \alpha \leq 3/2$ of the function f than the ML estimator (whose range of adaptation is $0 < l + \alpha \leq 1$).

Note that Theorems 4.1 and 4.2 provide asymptotic upper bounds (except for the case $l \geq 1$ of Theorem 4.2) and may not be tight if the function f is smoother than "typical" functions in $C^{l,\alpha}([0, T])$.[3] In Subsection 4.2, we show that the bounds are indeed tight in expectation if f is a fractional (or integrated fractional) Brownian motion with smoothness $l + \alpha$.

In the deterministic setting, a potential approach for obtaining a matching lower bound could use the rate of decay of the Fourier coefficients as a notion of smoothness, instead of the Hölder smoothness condition on the function f. Certain self-similarity conditions based on the decay rate and behaviour of Fourier coefficients are routinely used to study coverage of Bayesian credible sets (e.g., Szabó et al., 2015; Hadji and Szabó, 2021) as they define classes of functions that cannot "deceive" parameter estimators. Motivated by this, we attempted to adapt the argument in Sniekers and van der Vaart (2015, Section 4.2) and Sniekers and van der Vaart (2020, Section 10) to derive a matching lower bound under a self-similarity assumption on the Fourier coefficients. However, the bounds obtained through this approach proved sub-optimal in our setting. A different technique may therefore be required.

Remark 4.3. The proof of Theorem 4.2 shows that for $l = 1$ we have $\hat{\sigma}^2_{\mathrm{ML}} = \Theta(N^{-1})$ whenever $\|\mathcal{P}_N\| \to 0$ as $N \to \infty$. More precisely, it establishes that

$$N\hat{\sigma}^2_{\mathrm{ML}} \to \|f'\|_{\mathcal{L}^2([0,T])} := \int_0^T f'(x)^2 \, \mathrm{d}x \quad \text{as} \quad N \to \infty.$$

Note that the $\mathcal{L}^2([0, T])$ norm of f' in the right hand side equals the norm of f in the reproducing kernel Hilbert space of the Brownian motion kernel (e.g., van der Vaart and van Zanten, 2008, Section 10) Therefore, this fact is consistent with a similar more general statement in Karvonen

[3]For example, if $f(x) = |x - 1/2|$ with $T = 1$, we have $f \in C^{0,1}([0, T])$, as f is Lipschitz continuous in this case. However, f is almost everywhere infinitely differentiable except at one point $x = 1/2$, so it is, in this sense, much smoother than "typical" functions in $C^{0,1}([0, T])$.

et al. (2020, Proposition 3.1).

In addition to the above results, Theorem 4.4 below shows the limit of the CV estimator $\hat{\sigma}^2_{\mathrm{CV}}$ if the true function f is of finite quadratic variation.

Theorem 4.4. *For each $N \in \mathbb{N}$, let $\mathcal{P}_N \subset [0,T]$ be the equally-spaced partition of size N. Suppose that $f : [0,T] \to \mathbb{R}$ has finite quadratic variation $V^2(f)$ with respect to $(\mathcal{P}_N)_{N\in\mathbb{N}}$, $f(0) = 0$, and f is continuous on the boundary, i.e., $\lim_{x\to 0^+} f(x) = f(0)$ and $\lim_{x\to T^-} f(x) = f(T)$. Moreover, suppose that the quadratic variation $V^2(f)$ remains the same for all sequences of quasi-uniform partitions with constant $C_{\mathrm{qu}} = 2$.*[4] *Then*

$$\lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}} = \frac{V^2(f)}{T}.$$

Proof. See Subsection 8.2. ∎

For the ML estimator $\hat{\sigma}^2_{\mathrm{ML}}$, it is straightforward to obtain a similar result by using (3.2) and (8.1) in Section 8.1: Under the same conditions as Theorem 4.4, we have

$$\lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{ML}} = \frac{V^2(f)}{T}.$$

Theorem 4.4 and (4.1) are consistent with Theorems 4.1 and 4.2, which assume $f \in C^{l,\alpha}([0,T])$ with $l + \alpha > 1/2$ and imply $\hat{\sigma}^2_{\mathrm{CV}} \to 0$ and $\hat{\sigma}^2_{\mathrm{ML}} \to 0$ as $N \to \infty$. As summarized in Proposition 3.5, we have $V(f) = 0$ for $f \in C^{l,\alpha}([0,T])$ with $l + \alpha > 1/2$, so Theorem 4.4 and (4.1) imply that $\hat{\sigma}^2_{\mathrm{CV}} \to 0$ and $\hat{\sigma}^2_{\mathrm{ML}} \to 0$ as $N \to \infty$.

When f is a Brownian motion, in which case the Brownian motion prior is well-specified, the smoothness of f is $l + \alpha = 1/2$, and the quadratic variation $V(f)$ becomes a positive constant (Dudley, 1973). Proposition 4.7 in the next subsection shows that this fact, Theorem 4.4, and (4.1) lead to the consistency of the ML and CV estimators in the well-specified setting.

4.2. Random setting. In Subsection 4.1, we obtained asymptotic upper bounds on the CV and ML scale estimators when the true function f is a fixed function in a Hölder space. This section shows that these asymptotic bounds are tight in expectation when f is a fractional (or integrated fractional) Brownian motion.

That is, we consider the asymptotics of the expectations $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}}$ and $\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}}$ under the assumption that $f \sim \mathcal{GP}(0, k_{l,H})$, where $k_{l,H}$ is the kernel of a fractional Brownian motion (3.4) for $l = 0$ or that of an integrated fractional Brownian motion (3.1) for $l = 1$, with $0 < H < 1$ being the Hurst parameter. Recall that $f \sim \mathcal{GP}(0, k_{l,H})$ belongs to the Hölder space $C^{l,H-\varepsilon}([0,T])$ almost surely for arbitrarily small $\varepsilon > 0$, so its smoothness is $l + H$. Figure 4 summarises the obtained upper and lower rates, corroborating the upper rates in Figure 3.

Theorems 4.5 and 4.6 below establish the asymptotic upper and lower bounds for the CV and ML estimators, respectively.

Theorem 4.5 (Expected CV rate for fractional Brownian motion). *Suppose that $(\mathcal{P}_N)_{N\in\mathbb{N}}$ are*

[4] In Appendix B, we discuss the relaxation of this requirement.

quasi-uniform and $f \sim \mathcal{GP}(0, k_{l,H})$ *with* $l \in \{0, 1\}$ *and* $0 < H < 1$. *Then*

$$\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} = \Theta(N^{1-\min\{2(l+H),3\}}) = \begin{cases} \Theta\left(N^{1-2H}\right) & \text{if} \quad l = 0 \text{ and } H \in (0,1), \\ \Theta\left(N^{-1-2H}\right) & \text{if} \quad l = 1 \text{ and } H < 1/2, \\ \Theta\left(N^{-2}\right) & \text{if} \quad l = 1 \text{ and } H \geq 1/2. \end{cases}$$

Proof. See Section 8.3. ∎

Theorem 4.6 (Expected ML rate for fractional Brownian motion). *Suppose that* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *are quasi-uniform and* $f \sim \mathcal{GP}(0, k_{l,H})$ *with* $l \in \{0, 1\}$ *and* $0 < H < 1$. *Then*

$$\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} = \Theta(N^{1-\min\{2(l+H),2\}}) = \begin{cases} \Theta\left(N^{1-2H}\right) & \text{if} \quad l = 0 \text{ and } H \in (0,1), \\ \Theta\left(N^{-1}\right) & \text{if} \quad l = 1 \text{ and } H \in (0,1). \end{cases}$$

Proof. See Section 8.3. The proof is similar to that of Theorem 4.5. ∎

Theorems 4.5 and 4.6 show that the CV estimator is adaptive to the unknown smoothness $l+H$ of the function f for a broader range $0 < l+H \leq 3/2$ than the ML estimator, whose range of adaptation is $0 < l+H \leq 1$. These results imply that the CV estimator can be asymptotically well-calibrated for a broader range of unknown smoothness than the ML estimator, as discussed in Section 5.

When the smoothness of f is less than $1/2$, i.e., when $l + H < 1/2$, the Brownian motion prior, whose smoothness is $1/2$, is smoother than f. In this case, the expected rates of $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ are $\Theta\left(N^{1-2H}\right)$ and increase as N increases. The increase of $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ can be interpreted as compensating the overconfidence of the posterior standard deviation $\sqrt{k_N(x)}$, which decays too fast to be asymptotically well-calibrated. This interpretation agrees with the illustration in Figure 1.

On the other hand, when $l+H > 1/2$, the function f is smoother than the Brownian motion prior. In this case, $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ decrease as N increases, compensating the under-confidence of the posterior standard deviation $\sqrt{k_N(x)}$. See Figure 2 for an illustration.

When $l + H = 1/2$, this is the well-specified case in that the smoothness of f matches the Brownian motion prior. In this case, Theorems 4.5 and 4.6 yield $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} = \Theta(1)$ and $\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} = \Theta(1)$, i.e., when the CV and ML estimators converge, they converge to a positive constant. The following proposition, which follows from Theorem 4.4 and (4.1), shows that this limiting constant is the true value of the scale parameter σ_0^2 in the well-specified setting $f \sim \mathcal{GP}(0, \sigma_0^2 k)$, recovering similar results in the literature (e.g., Bachoc et al., 2017, Theorem 2).

Proposition 4.7. *Suppose that* $f \sim \mathcal{GP}(0, \sigma_0^2 k)$ *for* $\sigma_0 > 0$ *and that partitions* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *are equally-spaced. Then*

$$\lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}} = \lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{ML}} = \sigma_0^2 \quad \textit{almost surely.}$$

Proof. Since the quadratic variation of almost all sample paths of the unscaled (i.e., $\sigma_0 = 1$) Brownian motion on $[0, T]$ equals T (Dudley, 1973), the claim follows from (4.4) and (4.1). ∎

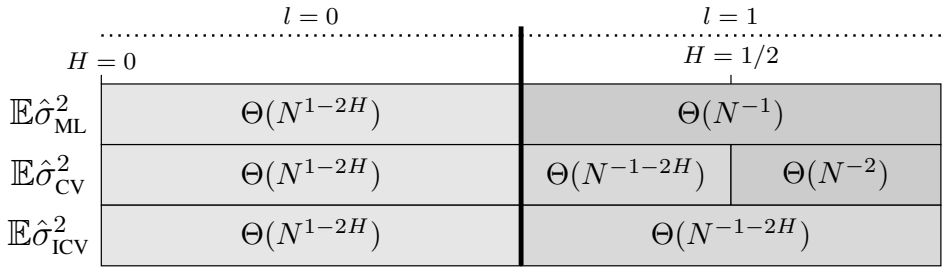

Figure 4. *Expected decay rates for the ML, CV and ICV estimators from Theorems 4.5, 4.6, and 4.9. Observe that the CV estimator's range of adaptation to the smoothness $l + H$ is wider than the ML estimator's, and the ICV estimator's range of adaptation is wider than that for both the CV and ML estimators.*

In Section 5, we discuss the implications of the obtained asymptotic rates of $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ on the reliability of the resulting GP uncertainty estimates. But first, motivated by the results in Theorem 4.1 and Theorem 4.5, we propose a modification to the cross-validation procedure that may have better asymptotic properties than the CV estimator.

4.3. Interior cross-validation estimators. The proofs of Theorems 4.1 and 4.5 show that when $l = 1$ and $\alpha \in (1/2, 1]$, the bound on $\hat{\sigma}^2_{\mathrm{CV}}$ is dominated by the bound on what we call the boundary terms. These are the terms corresponding to $n = 1$ and $n = N$ in (2.2); see also (10). That the boundary terms dominate is unsurprising since prediction at boundary points is a more challenging task than prediction at the interior. Motivated by this observation, we propose an alternative estimation method called *interior cross validation* (ICV) that maximises

$$\sum_{n=2}^{N-1} \log p(f(x_n) \,|\, x_n, \mathbf{x}_{\setminus n}, f(\mathbf{x}_{\setminus n}), \theta).$$

The corresponding scale parameter estimator is

$$\hat{\sigma}^2_{\mathrm{ICV}} = \frac{1}{N} \sum_{n=2}^{N-1} \frac{\left[f(x_n) - m_{\setminus n}(x_n)\right]^2}{k_{\setminus n}(x_n)}.$$

With the boundary points removed, the estimator's range of adaptation to the smoothness of the true function is greater than that for the CV estimator, as illustrated in Figure 3 for the deterministic setting and Figure 4 for the random setting. We present formal results for the deterministic and the random settings in the following theorems.

Theorem 4.8 (Rate of ICV decay in Hölder spaces). *Suppose that f is an element of $C^{l,\alpha}([0, T])$, with $l \geq 0$ and $0 < \alpha \leq 1$, such that $f(0) = 0$, and the interval partitions $(\mathcal{P}_N)_{N \in \mathbb{N}}$ have bounded mesh sizes $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$ as $N \to \infty$. Then*

$$\hat{\sigma}^2_{\mathrm{ICV}} = \mathcal{O}\big(N^{1-\min\{2(l+\alpha),4\}}\big) = \begin{cases} \mathcal{O}\left(N^{1-2\alpha}\right) & \text{if} \quad l = 0, \\ \mathcal{O}\left(N^{-1-2\alpha}\right) & \text{if} \quad l = 1, \\ \mathcal{O}\left(N^{-3}\right) & \text{if} \quad l \geq 2. \end{cases}$$

Proof. See Section 8.4. ∎

Theorem 4.9 (Expected ICV rate for fractional Brownian motion). *Suppose that* $(\mathcal{P}_N)_{N\in\mathbb{N}}$ *are quasi-uniform and* $f \sim \mathcal{GP}(0, k_{l,H})$ *with* $l \in \{0,1\}$ *and* $0 < H < 1$. *Then*

$$\mathbb{E}\hat{\sigma}^2_{\mathrm{ICV}} = \Theta(N^{1-\min\{2(l+H),4\}}) = \begin{cases} \Theta\left(N^{1-2H}\right) & \text{if} \quad l = 0, \\ \Theta\left(N^{-1-2H}\right) & \text{if} \quad l = 1. \end{cases}$$

Proof. See Section 8.4. ∎

One may take this idea further. For the Brownian motion kernel, an estimator that does not attempt to predict on points "close enough" to the boundary,

$$\hat{\sigma}^2_{\mathrm{ICV}}[N_0] = \frac{1}{N}\sum_{n=N_0}^{N-N_0} \frac{\left[f(x_n) - m_{\setminus n}(x_n)\right]^2}{k_{\setminus n}(x_n)}$$

for some fixed N_0, has the same range of adaptation as $\hat{\sigma}^2_{\mathrm{ICV}} = \hat{\sigma}^2_{\mathrm{ICV}}[1]$, the estimator that only ignores the points on the boundary. However, for smoother kernels like fractional Brownian motion (iFBM) and the Matérn family, $\hat{\sigma}^2_{\mathrm{ICV}}[N_0]$ may exhibit adaptation beyond the level $l = 2$. The number of boundary points N_0 to remove would likely depend on the smoothness of the kernel. Investigating model-dependent cross-validation estimators that discard a proportion of boundary points would be an interesting direction for future work.

5. Consequences for credible intervals. This section discusses whether the estimated scale parameter, given by the CV or ML estimator, leads to asymptotically well-calibrated credible intervals. With the kernel $\hat{\sigma}^2 k(x, x')$, where $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{CV}}$ or $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$, a GP credible interval at $x \in [0, T]$ is given by

$$[m_N(x) - \alpha\hat{\sigma}\sqrt{k_N(x)}, \quad m_N(x) + \alpha\hat{\sigma}\sqrt{k_N(x)}]$$

where $\alpha > 0$ is a constant (e.g., $\alpha \approx 1.96$ leads to the 95% credible interval).

As discussed in Section 1, this credible interval (5) is asymptotically well-calibrated, if it shrinks to 0 at the same speed as the decay of the error $|m_N(x) - f(x)|$ as N increases, i.e., the ratio

$$\frac{|f(x) - m_N(x)|}{\hat{\sigma}\sqrt{k_N(x)}}$$

should neither diverge to infinity nor converge to 0. If this ratio diverges to infinity, the credible interval (5) is asymptotically overconfident, in that (5) shrinks to 0 faster than the actual error $|f(x) - m_N(x)|$. If the ratio converges to 0, the credible interval is asymptotically underconfident, as it increasingly overestimates the actual error. Therefore, the ratio (5) should ideally converge to a positive constant for the credible interval (5) to be reliable.

For ease of analysis, we focus on the random setting in Section 4.2 where f is a fractional (or integrated fractional) Brownian motion and where we obtained asymptotic upper and lower bounds for $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}}$ and $\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}}$. We study how the expectation of the posterior variance $\mathbb{E}\hat{\sigma}^2 k_N(x)$ scales with the expected squared error $\mathbb{E}[f(x) - m_N(x)]^2$. Specifically, we analyze their ratio for $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\mathrm{ML}}$:

$$R^{\mathbb{E}}_{\mathrm{CV}}(x, N) := \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} k_N(x)} \quad \text{and} \quad R^{\mathbb{E}}_{\mathrm{ML}}(x, N) := \frac{\mathbb{E}[f(x) - m_N(x)]^2}{\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} k_N(x)}.$$

The ratio diverging to infinity (resp. converging to 0) as $N \to \infty$ suggests that the credible interval (5) is asymptotically overconfident (resp. underconfident) for a non-zero probability of the realisation of f. Thus ideally, the ratio should converge to a positive constant.

Theorem 5.1 below establishes the asymptotic rates of the ratios in (5).

Theorem 5.1. *Suppose that $(\mathcal{P}_N)_{N\in\mathbb{N}}$ are quasi-uniform and $f \sim \mathcal{GP}(0, k_{l,H})$ for $l \in \{0,1\}$ and $0 < H < 1$. Then,*

$$\sup_{x\in[0,T]} R^{\mathbb{E}}_{\mathrm{CV}}(x,N) = \begin{cases} \Theta(1) & \text{if} \quad l=0 \text{ and } H\in(0,1), \\ \Theta(1) & \text{if} \quad l=1 \text{ and } H\in(0,1/2), \\ \Theta\left(N^{1-2H}\right) & \text{if} \quad l=1 \text{ and } H\in(1/2,1), \end{cases}$$

and

$$\sup_{x\in[0,T]} R^{\mathbb{E}}_{\mathrm{ML}}(x,N) = \begin{cases} \Theta(1) & \text{if} \quad l=0 \text{ and } H\in(0,1), \\ \Theta\left(N^{-2H}\right) & \text{if} \quad l=1 \text{ and } H\in(0,1). \end{cases}$$

Proof. See Subsection 8.5. ∎

We have the following observations from Theorem 5.1, which suggest an advantage of the CV estimator over the ML estimator for uncertainty quantification:

- The ratio for the CV estimator neither diverges to infinity nor decays to 0 in the range $0 < l + H < 3/2$, which is broader than that of the ML estimator, $0 < l + H < 1$. This observation suggests that the CV estimator can yield asymptotically well-calibrated credible intervals for a broader range of the unknown smoothness $l + H$ of the function f than the ML estimator.
- The ratio decays to 0 for the CV estimator in the range $3/2 < l + H < 2$ and for the ML estimator in the range $1 < l + H < 2$. Therefore, the ML estimator may yield asymptotically underconfident credible intervals for a broader range of the smoothness $l + H$ than the CV estimator.

Moreover, for the interior CV estimator introduced in Subsection 4.3, it immediately follows from the proof in Subsection 8.5 that

$$\sup_{x\in[0,T]} R^{\mathbb{E}}_{\mathrm{ICV}}(x,N) = \begin{cases} \Theta(1) & \text{if} \quad l=0 \text{ and } H\in(0,1), \\ \Theta(1) & \text{if} \quad l=1 \text{ and } H\in(0,1), \end{cases}$$

which implies the ICV estimator can yield asymptotically well-calibrated credible intervals for a broader range of the smoothness than either the CV or the ML estimator.

6. Experiments.

This section describes numerical experiments to substantiate the theoretical results in Section 4. We define test functions in Subsection 6.1, show empirical asymptotic results for the CV estimator in Subsection 6.2, and report comparisons between the CV and ML estimators in Subsection 6.3.

To this end, for a continuous function f, define $l[f] \in \mathbb{N} \cup \{0\}$ and $\alpha \in (0,1]$ as

$$l[f] := \sup\{l \in \mathbb{N}\cup\{0\} : f \in C^l([0,T])\}, \quad \alpha[f] := \sup\{\alpha \in (0,1] : f \in C^{l[f],\alpha}([0,T])\}.$$

Then, for arbitrarily small $\varepsilon_1 \in \mathbb{N}$ and $\varepsilon_2 > 0$, we have

$$f \in C^{\max(l[f]-\varepsilon_1,0),\alpha[f]-\varepsilon_2}([0,T]) \quad \text{and} \quad f \notin C^{l[f]+\varepsilon_1,\alpha[f]+\varepsilon_2}([0,T]).$$

In this sense, $l[f]$ and $\alpha[f]$ characterize the smoothness of f.

6.1. Test functions. We generate test functions $f : [0,1] \to \mathbb{R}$ as sample paths of stochastic processes with varying degrees of smoothness, as defined below. The left columns of Figures 5 and 6 show samples of these functions.

- To generate nowhere differentiable test functions, we use the Brownian motion (BM), the Ornstein–Uhlenbeck process (OU), and the fractional Brownian motion (FBM[5]) which are zero-mean GPs with kernels

$$k_{\mathrm{BM}}(x,x') = \min(x,x'), \quad k_{\mathrm{OU}}(x,x') = \big(e^{-\lambda|x-x'|} - e^{-\lambda(x+x')}\big)/4,$$
$$k_{\mathrm{FBM}}(x,x') = \big(|x|^{2H} + |x'|^{2H} - |x-x'|^{2H}\big)/2,$$

  where $\lambda > 0$ and $0 < H < 1$ is the Hurst parameter (recall that the FBM = BM if $H = 1/2$). We set $\lambda = 0.2$ in the experiments below. Almost all samples f from these processes satisfy $l[f] = 0$. For BM and OU we have $\alpha[f] = 1/2$ and for FBM $\alpha[f] = H$ (see Subsection 3.4). It is well known that the OU process with the kernel k_{OU} above satisfies the stochastic differential equation

$$\mathrm{d}f(t) = -\lambda f(t)\mathrm{d}t + \sqrt{\frac{\lambda}{2}}\,\mathrm{d}B(t),$$

  where B is the standard Brownian motion whose kernel is k_{BM}.
- To generate differentiable test functions, we use once (iFBM) and twice (iiFBM) integrated fractional Brownian motions

$$f_{\mathrm{iFBM}}(x) = \int_0^x f_{\mathrm{FBM}}(z)\,\mathrm{d}z \quad \text{and} \quad f_{\mathrm{iiFBM}}(x) = \int_0^x f_{\mathrm{iFBM}}(z)\,\mathrm{d}z,$$

  where $f_{\mathrm{FBM}} \sim \mathcal{GP}(0, k_{\mathrm{FBM}})$. See (3.1) for the iFBM kernel. With H the Hurst parameter of the original FBM, almost all samples f from the above processes satisfy $l[f] = 1$ and $\alpha[f] = H$ (iFBM) or $l[f] = 2$ and $\alpha[f] = H$ (iiFBM).
- We also consider a piecewise infinitely differentiable function $f(x) = \sin 10x + [x > x_0]$, where x_0 is randomly sampled from the uniform distribution on $[0,1]$ and $[x > x_0]$ is 1 if $x > x_0$ and 0 otherwise. This function is of finite quadratic variation with $V^2(f) = 1$.

Denote $\hat{\sigma}^2 = \lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}}$. For the above test functions, with equally-spaced partitions, we expect the following asymptotic behaviours for the CV estimator from Theorems 4.1, 4.4,

[5]We use https://github.com/crflynn/fbm to sample from FBM.

and 4.5, Proposition 4.7, the definition of quadratic variation, and Equation (6.1):

$$\begin{aligned}
\text{BM } (l[f]=0,\,\alpha[f]=1/2)\text{:} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(1) &&\text{and} \quad \hat{\sigma}^2 = 1,\\
\text{OU } (l[f]=0,\,\alpha[f]=1/2)\text{:} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(1) &&\text{and} \quad \hat{\sigma}^2 = \lambda/2,\\
\text{FBM } (l[f]=0,\,\alpha[f]=H)\text{:} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(N^{1-2H}) &&\text{and} \quad \hat{\sigma}^2 = 0,\\
\text{iFBM } (l[f]=1,\,\alpha[f]=H)\text{:} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(N^{-1-2H}) &&\text{and} \quad \hat{\sigma}^2 = 0,\\
\text{iiFBM } (l[f]=2,\,\alpha[f]=H)\text{:} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(N^{-2}) &&\text{and} \quad \hat{\sigma}^2 = 0,\\
\sin 10x + [x > x_0]\text{ :} &\quad \hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(1) &&\text{and} \quad \hat{\sigma}^2 = 1.
\end{aligned}$$

Note that the above rate for the iFBM holds for $0 < H \leq 1/2$. The chosen functions allow us to cover a range of $\alpha[f]$ and $l[f]$ relevant to the varying rate of convergence in Theorems 4.1 and 4.5, as well as a range of $V^2(f)$ relevant to the limit in Theorem 4.4, $\lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}} = V^2(f)/T$.

6.2. Asymptotics of the CV estimator. Figure 5 shows the asymptotics of $\hat{\sigma}^2_{\mathrm{CV}}$, where each row corresponds to one stochastic process generating test functions f; the rows are displayed in the increasing order of smoothness as quantified by $l[f] + \alpha[f]$. The estimates are obtained for equally-spaced partitions of sizes $N = 10, 10^2, \ldots, 10^5$. In each row, the left panel plots a single sample of generated test functions f. The middle panel shows the mean and confidence intervals (of two standard deviations) of $\hat{\sigma}^2_{\mathrm{CV}}$ for 100 sample realisations of f for each sample size N. The right panel describes the convergence rate of $\hat{\sigma}^2_{\mathrm{CV}}$ to its limit point $\hat{\sigma}^2 = \lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}}$ on the log scale.

We have the following observations:

- The first two rows (the FBM and OU) and the last (the piece-wise infinitely differentiable function) confirm Theorem 4.4, which states the convergence $\hat{\sigma}^2_{\mathrm{CV}} \to V^2(f)/T$ as $N \to \infty$. While Theorem 4.4 does not provide convergence rates, the rates in the first two rows appear to be $N^{-1/2}$. In the last row the rate is N^{-2}.
- The remaining rows show that the observed rates of $\hat{\sigma}^2_{\mathrm{CV}}$ to 0 are in complete agreement with the rates predicted by Theorems 4.1 and 4.5. In particular, the rates are adaptive to the smoothness $l[f] + \alpha[f]$ of the function if $l[f] + \alpha[f] \leq 3/2$, as predicted.

6.3. Comparison of CV and ML estimators. Figure 6 shows the decay rates of $\hat{\sigma}^2_{\mathrm{CV}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$ to 0 for test functions f with $l[f] = 1$, under the same setting as for Figure 5. In this case, Theorems 4.2 and 4.6 predict that $\hat{\sigma}^2_{\mathrm{ML}}$ decays at the rate $\Theta(N^{-1})$ regardless of the smoothness; this is confirmed in the right column. In contrast, the middle column shows again that $\hat{\sigma}^2_{\mathrm{CV}}$ decays with a rate that adapts to $l[f]$ and $\alpha[f]$ as long as $l[f] + \alpha[f] \leq 3/2$, as predicted by Theorems 4.1 and 4.5. These results empirically support our theoretical finding that the CV estimator is adaptive to the unknown smoothness $l[f] + \alpha[f]$ of a function f for a broader range of smoothness than the ML estimator.

Additionally, in Appendix D, we compare the asymptotics of the CV and ML estimators when the underlying kernel is a Matérn kernel and the Sobolev smoothness of the true functions differs from that of the kernel. Similarly to the results presented in this section, we observe that the CV estimator exhibits a larger range of adaptation than the ML estimator.

7. Conclusion and future work. We have analysed the asymptotics of the CV and ML estimators for the scale parameter in GP interpolation with the Brownian motion kernel. As

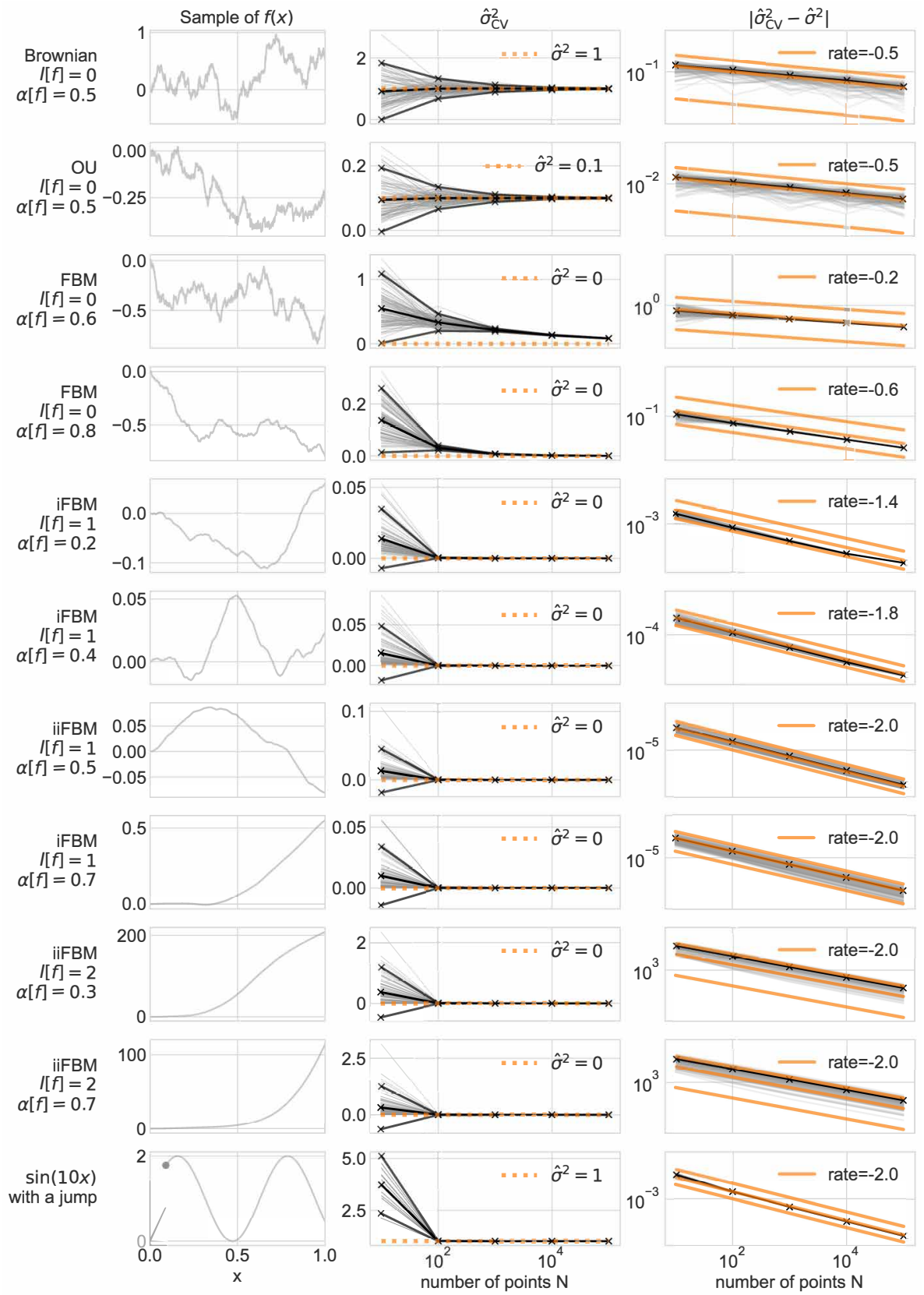

Figure 5. *Asymptotics of CV estimators for functions of varying smoothness as quantified by $l[f]$ and $\alpha[l]$ in (6). Runs on individual 100 samples from f are in gray, means and confidence intervals (of two standard deviations) are in black.*

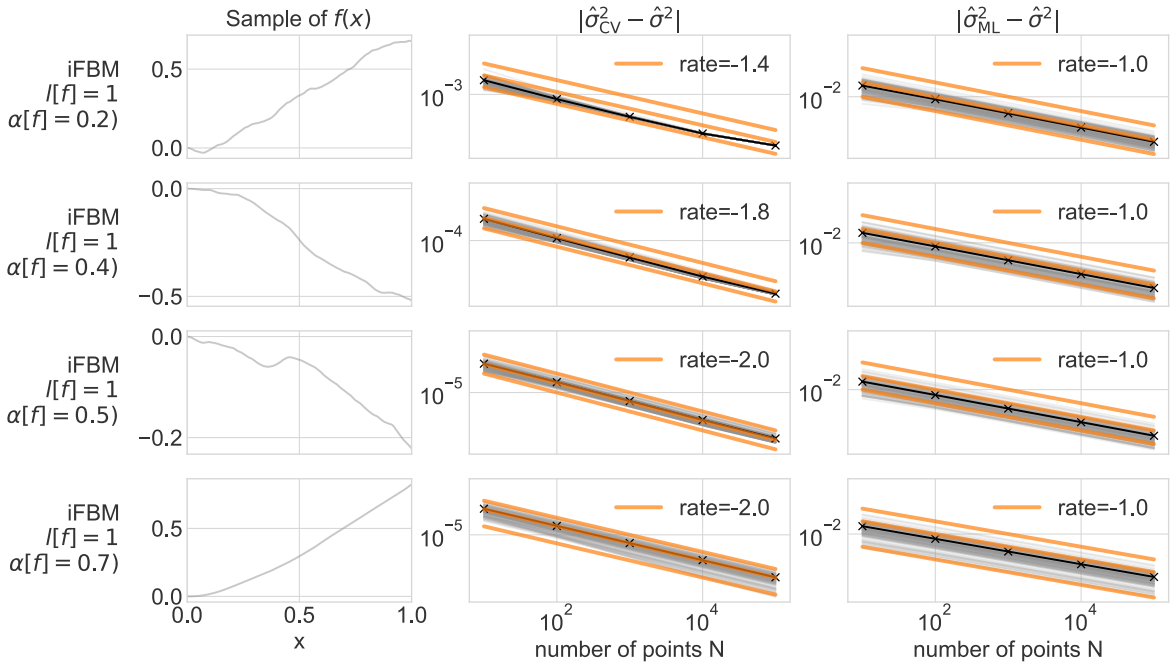

Figure 6. *Asymptotics of CV estimator compared to asymptotics of ML estimators, for once differentiable functions.*

a novel contribution, our analysis covers the misspecified case where the smoothness of the true function f is different from that of the samples from the GP prior. Our main results in Theorems 4.1, 4.2, 4.5, and 4.6 indicate that both CV and ML estimators can adapt to the unknown smoothness of f, but the range of smoothness for which this adaptation happens is broader for the CV estimator. Accordingly, the CV estimator can make GP uncertainty estimates asymptotically well-calibrated for a wider range of smoothness than the ML estimator, as indicated in Theorem 5.1. In this sense, the CV estimator has an advantage over the ML estimator. The experiments provide supporting evidence for the theoretical results.

Natural next steps include the following:

- Supplement the asymptotic upper bounds in Theorems 4.1 and 4.2 of the deterministic setting with matching lower bounds.
- Extend the analyses (of both the deterministic and random settings) to more generic finitely smooth kernels, higher dimensions, and a noisy setting.

The matching lower bounds, if obtained, would allow one to analyse the ratio between the prediction error $|f(x) - m_N(x)|$ and the posterior standard deviation $\hat{\sigma}\sqrt{k_N(x)}$ in the deterministic setting, corresponding to the one in Section 5 for the random setting. Such an analysis would need additional assumptions on the true function f, such as the homogeneity of the smoothness of f across the input space. It also requires a sharp characterisation of the error $|f(x) - m_N(x)|$, which could use super convergence results in Wendland (2005, Section 11.5) and Schaback (2018). Most natural kernel classes for extension are Matérns and other kernels

whose RKHS are norm-equivalent to Sobolev spaces; we conduct initial empirical analysis in Appendix D and observe results consistent with the main results in this paper. To this end, it would be possible to adapt the techniques used in Karvonen et al. (2020) for analyzing the ML estimator to the CV estimator. In any case, one would need much more advanced techniques than those used here. A potentially more straightforward extension could be one to multiple times integrated Brownian motion kernels for which Gaussian process interpolation corresponds to spline interpolation (Wahba, 1990, Chapter 1). In particular, finding analytic expression for the mean and variance of a cubic spline kernel given in, for example, Equation (6.28) of Rasmussen and Williams (2006) can be reduced to the problem of inverting a tridiagonal matrix targeted in Mallik (2001) and Kılıç (2008).

8. Proofs.

This section provides the proofs of the main results and other lengthy computations. For $x_0 = 0$ and $x_1, \dots, x_N \in [0, T]$, we will use the following notation whenever it can improve the readability or highlight a point:

$$\begin{aligned} \Delta x_n &:= x_{n+1} - x_n, \quad n = 0, 1, \dots, N-1, \\ f_n &:= f(x_n), \quad n = 0, 1, \dots, N. \end{aligned} \tag{8.1}$$

8.1. Explicit expressions for the CV and ML estimators.

Let us define $x_0 = 0$ and use the convention $f(x_0) = 0$. By a direct computation it is straightforward to verify that the inverse of the Gram matrix of the Brownian motion kernel $k(x, x') = \min(x, x')$ over the points $0 = x_0 < x_1 < x_2 < \cdots < x_N$ is the band matrix

$$k(\mathbf{x}, \mathbf{x})^{-1} = \begin{bmatrix} x_1 & x_1 & x_1 & \dots & x_1 & x_1 \\ x_1 & x_2 & x_2 & \dots & x_2 & x_2 \\ x_1 & x_2 & x_3 & \dots & x_3 & x_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_1 & x_2 & x_3 & \dots & x_{N-1} & x_{N-1} \\ x_1 & x_2 & x_3 & \dots & x_{N-1} & x_N \end{bmatrix}^{-1} = \begin{bmatrix} b_1 & c_1 & 0 & \dots & 0 & 0 \\ c_1 & b_2 & c_2 & \dots & 0 & 0 \\ 0 & c_2 & b_3 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & b_{N-1} & c_{N-1} \\ 0 & 0 & 0 & \dots & c_{N-1} & b_N \end{bmatrix},$$

where

$$\begin{aligned} b_i &= \frac{x_{i+1} - x_{i-1}}{(x_{i-1} - x_i)(x_i - x_{i+1})} \quad \text{for} \quad i \in \{2, \dots, N-1\}, \qquad b_N = -\frac{1}{x_{N-1} - x_N}, \\ c_i &= \frac{1}{(x_i - x_{i+1})} \quad \text{for} \quad i \in \{1, \dots, N-1\}. \end{aligned}$$

It follows that the posterior mean and covariance functions in (2.1) can be expressed as

$$m_N(x) = \begin{cases} \dfrac{(x_n - x) f(x_{n-1}) + (x - x_{n-1}) f(x_n)}{x_n - x_{n-1}} & \text{if } x \in [x_{n-1}, x_n] \text{ for some } 1 \le n \le N, \\ f(x_N) & \text{if } x \in [x_N, T] \end{cases}$$

and

$$k_N(x,x') = \begin{cases} \dfrac{(x_n - x')(x - x_{n-1})}{x_n - x_{n-1}} & \text{if } x_{n-1} \le x \le x' \le x_n \text{ for some } 1 \le n \le N, \\ x - x_N & \text{if } x_N \le x \le x' \le T, \\ 0 & \text{otherwise.} \end{cases}$$

We omit the case $x' \le x$ for $k_N(x,x')$ as this case is obtained by the symmetry $k_N(x,x') = k_N(x',x)$.

Using these expressions, we have, for each $1 \le n < N$:

$$m_{\backslash n}(x_n) = \frac{(x_n - x_{n+1})f(x_{n-1}) + (x_{n-1} - x_n)f(x_{n+1})}{x_{n-1} - x_{n+1}}$$

and

$$k_{\backslash n}(x_n) = k_{\backslash n}(x_n, x_n) = \frac{(x_n - x_{n+1})(x_n - x_{n-1})}{x_{n-1} - x_{n+1}}$$

For $n = N$, we have $m_{\backslash N}(x_N) = f(x_{N-1})$ and $k_{\backslash N}(x_N) = x_N - x_{N-1}$. Inserting these expressions in (2.2) and using the notation (8.1), the CV estimator can be written as

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N}\left[\frac{(x_2 f_1 - x_1 f_2)^2}{x_1 x_2 \Delta x_1} + \sum_{n=2}^{N-1} \frac{(\Delta x_{n-1}[f_{n+1} - f_n] - \Delta x_n[f_n - f_{n-1}])^2}{(\Delta x_n + \Delta x_{n-1})\Delta x_n \Delta x_{n-1}} + \frac{(f_N - f_{N-1})^2}{\Delta x_{N-1}}\right]. \tag{8.2}$$

For the ML estimator (2.2), we obtain the explicit expression

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \frac{[f(x_n) - f(x_{n-1})]^2}{\Delta x_{n-1}}$$

by observing that $m_{n-1}(x_n) = f(x_n)$ and $k_{n-1}(x_n) = x_n - x_{n-1}$.

Remark 8.1. The leave-p-out estimator $\hat{\sigma}^2_{\mathrm{CV}(p)}$ can be expressed in a form similar (albeit more complicated) to Equation (8.2). We derive this expression in Appendix C. This suggests that the analysis in Section 4 could potentially be generalised to apply to the leave-p-out estimators, a possibility that we leave open for future research to explore.

8.2. Proofs for Section 4.1.

Proof of Theorem 4.1. The estimator $\hat{\sigma}^2_{\mathrm{CV}}$ in (8.2) may be written as

$$\hat{\sigma}^2_{\mathrm{CV}} = B_{1,N} + I_N + B_{2,N}$$

in terms of the boundary terms

$$B_{1,N} = \frac{1}{N} \cdot \frac{(x_2 f_1 - x_1 f_2)^2}{x_1 x_2 \Delta x_1} \quad \text{and} \quad B_{2,N} = \frac{1}{N} \cdot \frac{(f_N - f_{N-1})^2}{\Delta x_{N-1}}$$

and the interior term

$$I_N = \frac{1}{N}\sum_{n=2}^{N-1} \frac{(\Delta x_{n-1}[f_{n+1}-f_n] - \Delta x_n[f_n - f_{n-1}])^2}{(\Delta x_n + \Delta x_{n-1})\Delta x_n \Delta x_{n-1}}.$$

The claimed rate in (4.1) is $\mathcal{O}(N^{-2})$ if $l \geq 2$ or $l = 1$ and $\alpha \geq 1/2$. By the inclusion properties of Hölder spaces in Section 3.3, it is therefore sufficient to consider the cases (a) $l = 0$ and (b) $l = 1$ and $\alpha \in (0, 1/2]$.

Suppose first that $l = 0$. Let L be a Hölder constant of a function $f \in C^{0,\alpha}([0,T])$. Using the Hölder condition, the bounding assumption on Δx_n, and $f_0 = f(0) = 0$, the boundary terms can be bounded as

$$\begin{aligned} B_{1,N} = \frac{1}{N}\cdot\frac{(x_1(f_1-f_2)+\Delta x_1(f_1-f_0))^2}{x_1 x_2 \Delta x_1} &\leq \frac{1}{N}\cdot\frac{2(x_1^2(f_1-f_2)^2+\Delta x_1^2(f_1-f_0)^2)}{x_1 x_2 \Delta x_1} \\ &\leq \frac{1}{N}\cdot\frac{2L^2(x_1^2\Delta x_1^{2\alpha}+x_1^{2\alpha}\Delta x_1^2)}{x_1 x_2 \Delta x_1} \\ &= \mathcal{O}(N^{-1}\Delta x_1^{2\alpha-1}) \\ &= \mathcal{O}(N^{-2\alpha}) \end{aligned} \tag{8.3}$$

and

$$B_{2,N} = \frac{1}{N}\cdot\frac{(f_N - f_{N-1})^2}{\Delta x_{N-1}} \leq \frac{1}{N}L^2\Delta x_{N-1}^{2\alpha-1} = \mathcal{O}(N^{-2\alpha}).$$

Similarly, the interior term is bounded as

$$\begin{aligned} I_N &\leq \frac{2}{N}\sum_{n=2}^{N-1}\frac{\Delta x_{n-1}^2(f_{n+1}-f_n)^2+\Delta x_n^2(f_n-f_{n-1})^2}{(\Delta x_n+\Delta x_{n-1})\Delta x_n\Delta x_{n-1}} \\ &\leq \frac{2L^2}{N}\sum_{n=2}^{N-1}\frac{\Delta x_{n-1}^2\Delta x_n^{2\alpha}+\Delta x_n^2\Delta x_{n-1}^{2\alpha}}{(\Delta x_n+\Delta x_{n-1})\Delta x_n\Delta x_{n-1}} \\ &= \frac{2L^2}{N}\sum_{n=2}^{N-1}\frac{\Delta x_{n-1}\Delta x_n^{2\alpha-1}+\Delta x_n\Delta x_{n-1}^{2\alpha-1}}{\Delta x_n+\Delta x_{n-1}} \\ &= \frac{2L^2}{N}\sum_{n=2}^{N-1}\left(\frac{\Delta x_{n-1}}{\Delta x_n+\Delta x_{n-1}}\Delta x_n^{2\alpha-1}+\frac{\Delta x_n}{\Delta x_n+\Delta x_{n-1}}\Delta x_{n-1}^{2\alpha-1}\right) \\ &\leq \frac{2L^2}{N}\sum_{n=2}^{N-1}\left(\Delta x_n^{2\alpha-1}+\Delta x_{n-1}^{2\alpha-1}\right) \\ &= \mathcal{O}(N^{1-2\alpha}). \end{aligned}$$

Inserting the above bounds in (10) yields $\hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(N^{-2\alpha}+N^{1-2\alpha}) = \mathcal{O}(N^{1-2\alpha})$, which is the claimed rate when $l = 0$.

Suppose then that $l = 1$ and $\alpha \in (0, 1/2]$, so that the first derivative f' of $f \in C^{1,\alpha}([0,T])$ is α-Hölder and hence continuous. Because a continuously differentiable function is Lipschitz,

we may set $\alpha = 1$ in the estimates (8.3) and (10) for the boundary terms $B_{1,N}$ and $B_{2,N}$ in the preceding case. This shows these terms are $\mathcal{O}(N^{-2})$. Because f is differentiable, we may use the mean value theorem to write the interior term as

$$\begin{aligned} I_N &= \frac{1}{N}\sum_{n=2}^{N-1} \frac{\Delta x_{n-1}\Delta x_n}{\Delta x_{n-1}+\Delta x_n}\left(\frac{f_{n+1}-f_n}{\Delta x_n} - \frac{f_n - f_{n-1}}{\Delta x_{n-1}}\right)^2 \\ &= \frac{1}{N}\sum_{n=2}^{N-1} \frac{\Delta x_{n-1}\Delta x_n}{\Delta x_{n-1}+\Delta x_n}\left[f'(\tilde{x}_n) - f'(\tilde{x}_{n-1})\right]^2, \end{aligned}$$

where $\tilde{x}_n \in (x_n, x_{n+1})$. Let L be a Hölder constant of f'. Then the Hölder continuity of f' and the assumption that $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$ yield

$$\begin{aligned} I_N \le \frac{L^2}{N}\sum_{n=2}^{N-1} \frac{\Delta x_{n-1}\Delta x_n}{\Delta x_{n-1}+\Delta x_n}|\tilde{x}_n - \tilde{x}_{n-1}|^{2\alpha} &\le \frac{L^2}{N}\sum_{n=2}^{N-1} \frac{\Delta x_{n-1}\Delta x_n}{\Delta x_{n-1}+\Delta x_n}(\Delta x_{n-1}+\Delta x_n)^{2\alpha} \\ &\le \frac{L^2}{N}\sum_{n=2}^{N-1} \Delta x_n(\Delta x_{n-1}+\Delta x_n)^{2\alpha} \\ &= \mathcal{O}(N^{-2\alpha-1}). \end{aligned}$$

Using the above bounds in (10) yields $\hat{\sigma}^2_{\mathrm{CV}} = \mathcal{O}(N^{-2} + N^{-2\alpha-1}) = \mathcal{O}(N^{-2\alpha-1})$, which is the claimed rate when $l = 1$. ∎

Proof of Theorem 4.2. From (8.1) we have

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \frac{(f_n - f_{n-1})^2}{\Delta x_{n-1}}.$$

Suppose first that $l = 0$. As in the proof of Theorem 4.1, we get

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \frac{(f_n - f_{n-1})^2}{\Delta x_{n-1}} \le \frac{L^2}{N}\sum_{n=1}^{N} \Delta x_{n-1}^{2\alpha-1} = \mathcal{O}\big(N^{1-2\alpha}\big)$$

when $\|\mathcal{P}_N\| = \mathcal{O}(N^{-1})$. Suppose then that $l = 1$. By the mean value theorem there are $\xi_n \in (x_{n-1}, x_n)$ such that

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \frac{(f_n - f_{n-1})^2}{\Delta x_{n-1}} = \frac{1}{N}\sum_{n=1}^{N} \Delta x_{n-1}\left(\frac{f_n - f_{n-1}}{\Delta x_{n-1}}\right)^2 = \frac{1}{N}\sum_{n=1}^{N} \Delta x_{n-1} f'(\xi_n)^2.$$

Since f' is continuous on $[0, T]$ and hence Riemann integrable, we obtain the asymptotic equivalence

$$N\hat{\sigma}^2_{\mathrm{ML}} \to \int_0^T f'(x)^2\,\mathrm{d}x \quad \text{as} \quad N \to \infty$$

when $\|\mathcal{P}_N\| \to 0$ as $N \to \infty$. The integral is positive because f has been assumed non-constant. ∎

Proof of Theorem 4.4. For equally-spaced partitions, $\Delta x_n = x_1 = T/N$ for all $n \in \{0, \dots, N-1\}$, the estimator $\hat{\sigma}^2_{\mathrm{CV}}$ in (8.2) takes the form

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{T}\left[\frac{(x_2 f_1 - x_1 f_2)^2}{x_1 x_2} + \frac{1}{2}\sum_{n=2}^{N-1}((f_{n+1} - f_n) - (f_n - f_{n-1}))^2 + (f_N - f_{N-1})^2\right].$$

Recall from the proof of Theorem 4.1 the decomposition

$$\hat{\sigma}^2_{\mathrm{CV}} = B_{1,N} + I_N + B_{2,N}$$

in terms of the boundary terms $B_{1,N}$ and $B_{2,N}$ in (10) and the interior term I_N in (10). Because f is assumed continuous on the boundary and equispaced partitions are quasi-uniform, both $B_{1,N}$ and $B_{2,N}$ tend to zero as $N \to \infty$. We may therefore focus on the interior term, which decomposes as

$$\begin{aligned} I_N &= \frac{1}{2}\sum_{n=2}^{N-1}((f_{n+1} - f_n) - (f_n - f_{n-1}))^2 \\ &= \sum_{n=2}^{N-1}(f_{n+1} - f_n)^2 + (f_n - f_{n-1})^2 - \frac{1}{2}(f_{n+1} - f_{n-1})^2 \end{aligned}$$

The sums $\sum_{n=2}^{N-1}(f_{n+1} - f_n)^2$ and $\sum_{n=2}^{N-1}(f_n - f_{n-1})^2$ tend to $V^2(f)$ by definition. To establish the claimed bound we are therefore left to prove that

$$\sum_{n=2}^{N-1}(f_{n+1} - f_{n-1})^2 \to 2V^2(f) \qquad \text{as} \qquad N \to \infty.$$

We may write the sum as

$$\sum_{n=2}^{N-1}(f_{n+1} - f_{n-1})^2 = \sum_{n=1}^{\lfloor \frac{N-1}{2} \rfloor}(f_{2n+1} - f_{2n-1})^2 + \sum_{n=1}^{\lfloor \frac{N-2}{2} \rfloor}(f_{2n+2} - f_{2n})^2.$$

Consider a sub-partition of $\mathcal{P}_N$ that consists of odd-index points $x_1, x_3, \dots x_{2\lfloor \frac{N-1}{2} \rfloor + 1}$ of $\mathcal{P}_N$. The sequence of these sub-partitions is quasi-uniform with constant 2. The assumption that the quadratic variation is $V^2(f)$ for all partitions with quasi-uniformity constant 2 implies that

$$\lim_{N\to\infty} \sum_{n=1}^{\lfloor \frac{N-1}{2} \rfloor}(f_{2n+1} - f_{2n-1})^2 = V^2(f).$$

The same will hold for sub-partitions formed of even-index points of $\mathcal{P}_N$, giving

$$\lim_{N\to\infty} \sum_{n=1}^{\lfloor \frac{N-2}{2} \rfloor}(f_{2n+2} - f_{2n})^2 = V^2(f).$$

Thus, (12) holds. This completes the proof. ∎

8.3. Proofs for Section 4.2.

Proof of Theorem 4.5. Recall the explicit expression of $\hat{\sigma}^2_{\mathrm{CV}}$ in (8.2):

$$\hat{\sigma}^2_{\mathrm{CV}} = \frac{1}{N}\left[\frac{(x_2 f_1 - x_1 f_2)^2}{x_1 x_2 \Delta x_1} + \sum_{n=2}^{N-1} \frac{(\Delta x_{n-1}[f_{n+1} - f_n] - \Delta x_n[f_n - f_{n-1}])^2}{(\Delta x_n + \Delta x_{n-1})\Delta x_n \Delta x_{n-1}} + \frac{(f_N - f_{N-1})^2}{\Delta x_{N-1}}\right]. \tag{8.4}$$

We consider the cases $l = 0$ and $l = 1$ separately. Recall that $f \sim \mathcal{GP}(0, k_{l,H})$ implies that $\mathbb{E}[f(x)f(x')] = k_{l,H}(x, x')$.

Suppose first that $l = 0$, in which case $f \sim \mathcal{GP}(0, k_{0,H})$ for the fractional Brownian motion kernel $k_{0,H}$ in (3.4). In this case the expected values of squared terms in the expression for $\hat{\sigma}^2_{\mathrm{CV}}$ are $\mathbb{E}[x_2 f_1 - x_1 f_2]^2 = x_1 x_2 \Delta x_1 (x_1^{2H-1} + \Delta x_1^{2H-1} - (x_1 + \Delta x_1)^{2H-1})$,

$$\begin{aligned}&\mathbb{E}\big[\Delta x_{n-1}(f_{n+1} - f_n) - \Delta x_n (f_n - f_{n-1})\big]^2 \\ &\quad = \big(\Delta x_n^{2H-1} + \Delta x_{n-1}^{2H-1} - (\Delta x_{n-1} + \Delta x_n)^{2H-1}\big)\Delta x_{n-1}\Delta x_n(\Delta x_n + \Delta x_{n-1}),\end{aligned}$$

and $\mathbb{E}[f_N - f_{N-1}]^2 = \Delta x_{N-1}^{2H}$. Substituting these in the expectation of $\hat{\sigma}^2_{\mathrm{CV}}$ and using the fact that $\Delta x_n = \Theta(N^{-1})$ for all n by quasi-uniformity we get

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} &= \frac{1}{N}\Bigg[(x_1^{2H-1} + \Delta x_1^{2H-1} - (x_1 + \Delta x_1)^{2H-1}) \\ &\qquad + \sum_{n=2}^{N-1}\left(\Delta x_{n-1}^{2H-1} + \Delta x_n^{2H-1} - (\Delta x_{n-1} + \Delta x_n)^{2H-1}\right) + \Delta x_{N-1}^{2H-1}\Bigg] \\ &= \frac{1}{N}\Bigg[\Delta x_1^{2H-1}\left(\left(\frac{x_1}{\Delta x_1}\right)^{2H-1} + 1 - \left(\frac{x_1}{\Delta x_1^{2H-1}} + 1\right)^{2H-1}\right) \\ &\qquad + \Delta x_n^{2H-1}\sum_{n=2}^{N-1}\left(\left(\frac{\Delta x_{n-1}}{\Delta x_n}\right)^{2H-1} + 1 - \left(\frac{\Delta x_{n-1}}{\Delta x_n} + 1\right)^{2H-1}\right) + \Delta x_{N-1}^{2H-1}\Bigg] \\ &=: \frac{1}{N}\left[\Delta x_1^{2H-1} c_1 + \Delta x_n^{2H-1}\sum_{n=2}^{N-1} c_n + \Delta x_{N-1}^{2H-1}\right].\end{aligned}$$

Notice that the function $x \mapsto x^{2H-1} + 1 - (x+1)^{2H-1}$ is positive for $x > 0$ and $H \in (0,1)$, and increasing for $H \in (1/2, 1)$ and non-increasing for $H \in (0, 1/2]$. By quasi-uniformity we have $C_{\mathrm{qu}}^{-1} \le \Delta x_{n-1}/\Delta x_n \le C_{\mathrm{qu}}$, and can bound c_n for any n and N as

$$\begin{aligned}&0 < C_{\mathrm{qu}}^{2H-1} + 1 - (C_{\mathrm{qu}} + 1)^{2H-1} \le c_n \le C_{\mathrm{qu}}^{1-2H} + 1 - (C_{\mathrm{qu}}^{-1} + 1)^{2H-1} \quad \text{if } H \in (0, 1/2], \\ &0 < C_{\mathrm{qu}}^{1-2H} + 1 - (C_{\mathrm{qu}}^{-1} + 1)^{2H-1} \le c_n \le C_{\mathrm{qu}}^{2H-1} + 1 - (C_{\mathrm{qu}} + 1)^{2H-1} \quad \text{if } H \in (1/2, 1).\end{aligned}$$

Finally, by quasi-uniformity $\Delta x_n = \Theta(N^{-1})$, and $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} = \Theta(N^{-2H}) + \Theta(N^{1-2H}) + \Theta(N^{-2H}) = \Theta(N^{1-2H})$.

Suppose then that $l = 1$, in which case $f \sim \mathcal{GP}(0, k_{1,H})$ for the integrated fractional Brownian motion kernel $k_{1,H}$ in (3.1). It is straightforward (though, in the case of the second expectation, somewhat tedious) to compute that the expected values of squared terms in the expression (8.4) for $\hat{\sigma}^2_{\mathrm{CV}}$ are

$$\mathbb{E}[x_2 f_1 - x_1 f_2]^2 = \frac{x_1 x_2 \Delta x_1}{2(H+1)(2H+1)} \left(x_2^{2H+1} - x_1^{2H+1} - \Delta x_1^{2H+1}\right)$$

and

$$\begin{aligned} &\mathbb{E}\left[\Delta x_{n-1}(f_{n+1} - f_n) - \Delta x_n (f_n - f_{n-1})\right]^2 \\ &\quad = \frac{\Delta x_n \Delta x_{n-1}(\Delta x_n + \Delta x_{n-1})}{2(H+1)(2H+1)} \left[(\Delta x_n + \Delta x_{n-1})^{2H+1} - \Delta x_n^{2H+1} - \Delta x_{n-1}^{2H+1}\right] \end{aligned} \tag{8.5}$$

and

$$\mathbb{E}[f_N - f_{N-1}]^2 = \frac{\Delta x_{N-1}}{2H+1} \left(x_N^{2H+1} - x_{N-1}^{2H+1} - \frac{1}{2(H+1)} \Delta x_{N-1}^{2H+1} \right).$$

Therefore, by (8.4),

$$\begin{aligned} \mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} &= \frac{\left(x_2^{2H+1} - x_1^{2H+1} - \Delta x_1^{2H+1}\right)}{2(H+1)(2H+1)N} \\ &\quad + \frac{1}{2(H+1)(2H+1)N} \sum_{n=2}^{N-1} \left[(\Delta x_n + \Delta x_{n-1})^{2H+1} - \Delta x_n^{2H+1} - \Delta x_{n-1}^{2H+1}\right] \\ &\quad + \frac{1}{(2H+1)N} \left(x_N^{2H+1} - x_{N-1}^{2H+1} - \frac{1}{2(H+1)} \Delta x_{N-1}^{2H+1} \right) \\ &=: \frac{1}{2(H+1)(2H+1)} B_{1,N} + \frac{1}{2(H+1)(2H+1)} I_N + \frac{1}{(2H+1)} B_{2,N}. \end{aligned}$$

By quasi-uniformity, $B_{1,N} \leq N^{-1} x_2^{2H+1} = \mathcal{O}(N^{-2-2H})$. Consider then the interior term

$$\begin{aligned} I_N &= \frac{1}{N} \sum_{n=2}^{N-1} \Delta x_n^{2H+1} \left[\left(1 + \frac{\Delta x_{n-1}}{\Delta x_n}\right)^{2H+1} - \left(1 + \left(\frac{\Delta x_{n-1}}{\Delta x_n}\right)^{2H+1}\right) \right] \\ &=: \frac{1}{N} \sum_{n=2}^{N-1} \Delta x_n^{2H+1} c_n'. \end{aligned} \tag{8.6}$$

Because the function $x \mapsto (1+x)^{2H+1} - (1 + x^{2H+1})$ is positive and increasing for $x > 0$ if $H \in (0,1)$ and $C_{\mathrm{qu}}^{-1} \leq \Delta x_{n-1} / \Delta x_n \leq C_{\mathrm{qu}}$ by quasi-uniformity, we have

$$0 < (1 + C_{\mathrm{qu}}^{-1})^{2H+1} - (1 + C_{\mathrm{qu}}^{-(2H+1)}) \leq c_n' \leq \left(1 + \frac{\Delta x_{n-1}}{\Delta x_n}\right)^{2H+1} \leq (1 + C_{\mathrm{qu}})^{2H+1}$$

for every n. Because $N^{-1} \sum_{n=2}^{N-1} \Delta x_n^{2H+1} = \Theta(N^{-1-2H})$ by quasi-uniformity, we conclude from (8.6) that $I_N = \Theta(N^{-1-2H})$. For the last term $B_{2,N}$, recall that we have set $x_N = T$.

Thus

$$B_{2,N} = \frac{1}{N}\left(T^{2H+1} - (T - \Delta x_{N-1})^{2H+1} - \frac{1}{2(H+1)}\Delta x_{N-1}^{2H+1}\right).$$

By the generalised binomial theorem,

$$T^{2H+1} - (T - \Delta x_{N-1})^{2H+1} = (2H+1)T^{2H}\Delta x_{N-1} + \mathcal{O}(\Delta x_{N-1}^2)$$

as $\Delta x_{N-1} \to 0$. It follows that under quasi-uniformity we have $B_{2,N} = \Theta(N^{-2})$ for every $H \in (0,1)$. Putting these bounds for $B_{1,N}$, I_N and $B_{2,N}$ together we conclude that

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} &= \frac{1}{2(H+1)(2H+1)}B_{1,N} + \frac{1}{2(H+1)(2H+1)}I_N + \frac{1}{(2H+1)}B_{2,N} \\ &= \mathcal{O}(N^{-2-2H}) + \Theta(N^{-1-2H}) + \Theta(N^{-2}),\end{aligned}$$

which gives $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} = \Theta(N^{-1-2H})$ if $H \in (0, 1/2]$ and $\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} = \Theta(N^{-2})$ if $H \in [1/2, 1)$. ∎

Observe that in the proof of Theorem 4.5 it is the boundary term $B_{2,N}$ that determines the rate when there is sufficient smoothness, in that $l = 1$ and $H \in [1/2, 1)$. Similar phenomenon occurs in the proof of Theorem 4.1. The smoother a process is, the more correlation there is between its values at far-away points. Because the Brownian motion (as well as fractional and integrated Brownian motions) has a zero boundary condition at $x = 0$ but no boundary condition at $x = T$ and no information is available at points beyond T, the importance of $B_{2,N}$ is caused by the fact that around T one has the least information about the process.

Proof of Theorem 4.6. From (8.1) we get

$$\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\mathbb{E}[f_n - f_{n-1}]^2}{\Delta x_{n-1}}.$$

We may then proceed as in the proof of Theorem 4.5 and use quasi-uniformity to show that

$$\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\mathbb{E}[f_n - f_{n-1}]^2}{\Delta x_{n-1}} = \frac{1}{N}\sum_{n=1}^{N}\frac{\Delta x_{n-1}^{2H}}{\Delta x_{n-1}} = \frac{1}{N}\sum_{n=1}^{N}\Delta x_{n-1}^{2H-1} = \Theta(N^{1-2H})$$

when $l = 0$ and

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2_{\mathrm{ML}} &= \sum_{n=1}^{N}\frac{\mathbb{E}[f_n - f_{n-1}]^2}{\Delta x_{n-1}} \\ &= \frac{1}{(2H+1)N}\sum_{n=1}^{N}\left(x_n^{2H+1} - x_{n-1}^{2H+1} - \frac{1}{2(H+1)}\Delta x_{n-1}^{2H+1}\right) \\ &= \frac{1}{(2H+1)N}\sum_{n=1}^{N}\left((2H+1)x_n^{2H}\Delta x_{n-1} + \mathcal{O}(\Delta x_{n-1}^2) - \frac{1}{2(H+1)}\Delta x_{n-1}^{2H+1}\right) \\ &= \Theta(N^{-1})\end{aligned}$$

when $l = 1$. ∎

8.4. Proofs for Subsection 4.3. For the Brownian motion kernel, the ICV estimator defined in (4.3) takes the explicit form

$$\hat{\sigma}^2_{\mathrm{ICV}} = \frac{1}{N}\sum_{n=2}^{N-1} \frac{(\Delta x_{n-1}[f_{n+1}-f_n] - \Delta x_n[f_n - f_{n-1}])^2}{(\Delta x_n + \Delta x_{n-1})\Delta x_n \Delta x_{n-1}}.$$

We analyse this estimator below.

Proof of Theorem 4.8. The proof of Theorem 4.1 shows that when $l = 1$ and $\alpha \in (1/2, 1]$, the bound is dominated by the bound on the boundary terms, $B_{1,N} = \mathcal{O}(N^{-2})$ and $B_{2,N} = \mathcal{O}(N^{-2})$, since

$$\hat{\sigma}^2_{\mathrm{CV}} = B_{1,N} + I_N + B_{2,N} = \mathcal{O}(N^{-2}) + \mathcal{O}(N^{-1-2\alpha}) + \mathcal{O}(N^{-2}) = \mathcal{O}(N^{-2}).$$

As $\hat{\sigma}^2_{\mathrm{ICV}} = I_N$, it follows that $\hat{\sigma}^2_{\mathrm{ICV}} = \mathcal{O}(N^{-1-2\alpha})$ when $l = 1$. ∎

Proof of Theorem 4.9. The proof of Theorem 4.5 shows that when $l = 1$ and $H \in [1/2, 1)$, the bound is dominated by the bound on the right boundary terms, $B_{2,N} = \Theta(N^{-2})$, since

$$\begin{aligned}\mathbb{E}\hat{\sigma}^2_{\mathrm{CV}} &= \frac{1}{2(H+1)(2H+1)} B_{1,N} + \frac{1}{2(H+1)(2H+1)} I_N + \frac{1}{(2H+1)} B_{2,N} \\ &= \mathcal{O}(N^{-2-2H}) + \Theta(N^{-1-2H}) + \Theta(N^{-2})\end{aligned}$$

As $\mathbb{E}\hat{\sigma}^2_{\mathrm{ICV}} = I_N/(2(H+1)(2H+1))$, it follows that $\hat{\sigma}^2_{\mathrm{ICV}} = \Theta(N^{-1-2H})$ when $l = 1$. ∎

8.5. Proofs for Section 5.

Proof of Theorem 5.1. We only provide the proof for the case $l = 1$ and leave the simpler case $l = 0$ to the reader. Let $x \in (x_{n-1}, x_n)$. From the expression for m_N in Subsection 8.1, we get

$$\begin{aligned}\mathbb{E}[f(x) - m_N(x)]^2 &= \mathbb{E}\left[f(x) - \frac{(x_n - x)f(x_{n-1}) + (x - x_{n-1})f(x_n)}{\Delta x_{n-1}}\right]^2 \\ &= \frac{1}{\Delta x^2_{n-1}} \mathbb{E}\big[(x - x_{n-1})(f(x_n) - f(x)) - (x_n - x)(f(x) - f(x_{n-1}))\big]^2.\end{aligned}$$

Then, we can use (8.5) with x_n instead of x_{n+1} and x instead of x_n to get

$$\mathbb{E}[f(x) - m_N(x)]^2 = \frac{(x_n - x)(x - x_{n-1})}{C_H \Delta x_{n-1}}\big[\Delta x^{2H+1}_{n-1} - (x_n - x)^{2H+1} - (x - x_{n-1})^{2H+1}\big],$$

where $C_H = 2(H+1)(2H+1)$. The expression for k_N in Subsection 8.1 gives

$$\frac{\mathbb{E}[f(x) - m_N(x)]^2}{k_N(x)} = \frac{1}{C_H}\big[\Delta x^{2H+1}_{n-1} - (x_n - x)^{2H+1} - (x - x_{n-1})^{2H+1}\big].$$

By removing the negative terms and using the quasi-uniformity (3.2), we obtain

$$\sup_{x \in [0,T]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{k_N(x)} \le \frac{(TC_{\mathrm{qu}})^{2H+1}}{C_H} N^{-1-2H},$$

To see that this bound is tight, observe that for the midpoint $x = (x_n + x_{n-1})/2$ we have $x_n - x = x - x_{n-1} = \Delta x_{n-1}/2$ and

$$\frac{\mathbb{E}[f(x) - m_N(x)]^2}{k_N(x)} = \frac{1}{C_H}\left(1 - \frac{1}{2^{2H}}\right)\Delta x_{n-1}^{2H+1} \geq \frac{T^{2H+1}}{C_H C_{\text{qu}}^{2H+1}}\left(1 - \frac{1}{2^{2H}}\right)N^{-1-2H}$$

by the quasi-uniformity. Therefore

$$\sup_{x\in[0,T]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{k_N(x)} = \Theta(N^{-1-2H})$$

when $l = 1$. One can similarly show that

$$\sup_{x\in[0,T]} \frac{\mathbb{E}[f(x) - m_N(x)]^2}{k_N(x)} = \Theta(N^{1-2H})$$

when $l = 0$. The claims then follow from the rates for $\mathbb{E}\hat{\sigma}^2_{\text{CV}}$ and $\mathbb{E}\hat{\sigma}^2_{\text{ML}}$ in Theorems 4.5 and 4.6. ∎

Acknowledgements. The authors are grateful to the anonymous reviewers and the Associate Editor for their time and effort in reviewing the manuscript, and the comments and suggestions that improved the paper. In particular, one of the reviewers suggested removal of the rightmost boundary point in the CV estimator, which led to the introduction on the interior cross validation estimator.

Appendix A. Connection between the ML and CV estimators.

Here we prove a connection between the ML and CV estimators; see Remark 2.1. Let

$$C(N,p) = \binom{N}{p} = \frac{N!}{p!(N-p)!}$$

denote the binomial coefficient. The leave-p-out cross-validation (LPO-CV) estimator of σ^2 is

$$\hat{\sigma}^2_{\mathrm{CV}(p)} = \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \frac{[f(x_{p,i,n}) - m_{\setminus\{p,i\}}(x_{p,i,n})]^2}{k_{\setminus\{p,i\}}(x_{p,i,n})},$$

where i indexes the N-choose-p possible sets of held-out datapoints, $\mathbf{x}_{\setminus\{p,i\}}$, among $\mathbf{x}$ and $n \leq p$ the data points left out of each of these sets. That is, for each p and i we have

$$\mathbf{x} = \mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,1}, \ldots, x_{p,i,p}\}.$$

The functions $m_{\setminus\{p,i\}}$ and $k_{\setminus\{p,i\}}$ are the GP conditional mean and variance based on the set $\mathbf{x}_{\setminus\{p,i\}}$, which contains $N-p$ points. The purpose of this section is to prove that

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{N} \sum_{p=1}^{N} \hat{\sigma}^2_{\mathrm{CV}(p)}.$$

Denote $\nu(\mathbf{x}) = f(\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} f(\mathbf{x})$. The block matrix inversion formula applied to $g(\mathbf{x}_{\setminus\{p,i\}} \cup \{x\})$ and the equations in Section 2 for the conditional mean and variance yield

$$\frac{[f(x) - m_{\setminus\{p,i\}}(x)]^2}{k_{\setminus\{p,i\}}(x)} = \nu(\mathbf{x}_{\setminus\{p,i\}} \cup \{x\}) - \nu(\mathbf{x}_{\setminus\{p,i\}})$$

for any $1 \leq p \leq N$ and $x \notin \mathbf{x}_{\{p,i\}}$, where we use the convention $\nu(\mathbf{x}_{\setminus\{N,i\}}) = \nu(\emptyset) = 0$. For each $1 \leq p \leq N$, $i \leq C(N,p)$ and $n \leq p$ there is a unique index $j(p,i,n) \leq C(N,p-1)$ such that

$$\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\} = \mathbf{x}_{\setminus\{p-1,j(p,i,n)\}}.$$

Setting $x = x_{p,i,n}$ in (A) gives

$$\frac{[f(x_{p,i,n}) - m_{\setminus\{p,i\}}(x_{p,i,n})]^2}{k_{\setminus\{p,i\}}(x_{p,i,n})} = \nu(\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\}) - \nu(\mathbf{x}_{\setminus\{p,i\}}).$$

Therefore

$$\begin{aligned} \sum_{p=1}^{N} \hat{\sigma}^2_{\mathrm{CV}(p)} &= \frac{1}{N} \sum_{p=1}^{N} \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \frac{[f(x_{p,i,n}) - m_{\setminus\{p,i\}}(x_{p,i,n})]^2}{k_{\setminus\{p,i\}}(x_{p,i,n})} \\ &= \sum_{p=1}^{N} \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \left[\nu(\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\}) - \nu(\mathbf{x}_{\setminus\{p,i\}})\right]. \end{aligned} \tag{A.1}$$

By (A) from each set $\mathbf{x}_{\setminus\{p,i\}}$ on level p (i.e., sets from which p points have been left out) one can obtain p sets on level $p-1$ by adding one of the left-out datapoints. However, there are $C(N,p)$ sets on level p and $C(N,p-1)$ sets on level $p-1$. Hence for each set $\mathbf{x}_{\setminus\{p-1,j\}}$ on level $p-1$ there are

$$p \cdot \frac{C(N,p)}{C(N,p-1)} = p \cdot \frac{N!(p-1)!(N-p+1)!}{N!p!(N-p)!} = N-p+1$$

combinations of sets $\mathbf{x}_{\setminus\{p,i\}}$ on level p and points $x_{p,i,n}$ left out of these sets such that $\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\} = \mathbf{x}_{\setminus\{p-1,j\}}$. Therefore

$$\begin{aligned}
&\sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \left[\nu(\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\}) - \nu(\mathbf{x}_{\setminus\{p,i\}})\right] \\
&\quad = \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \nu(\mathbf{x}_{\setminus\{p,i\}} \cup \{x_{p,i,n}\}) - \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \nu(\mathbf{x}_{\setminus\{p,i\}}) \\
&\quad = \frac{N-p+1}{p} \sum_{j=1}^{C(N,p-1)} \nu(\mathbf{x}_{\setminus\{p-1,j\}}) - \sum_{i=1}^{C(N,p)} \nu(\mathbf{x}_{\setminus\{p,i\}})
\end{aligned}$$

and consequently (A.1) writes

$$\begin{aligned}
\sum_{p=1}^{N} \hat{\sigma}^2_{\mathrm{CV}(p)} &= \sum_{p=1}^{N} \frac{1}{C(N,p)} \left[\frac{N-p+1}{p} \sum_{j=1}^{C(N,p-1)} \nu(\mathbf{x}_{\setminus\{p-1,j\}}) - \sum_{i=1}^{C(N,p)} \nu(\mathbf{x}_{\setminus\{p,i\}})\right] \\
&= \sum_{p=1}^{N} \left[\frac{1}{C(N,p-1)} \sum_{j=1}^{C(N,p-1)} \nu(\mathbf{x}_{\setminus\{p-1,j\}}) - \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \nu(\mathbf{x}_{\setminus\{p,i\}})\right],
\end{aligned}$$

which is a telescoping sum. We are left with

$$\sum_{p=1}^{N} \hat{\sigma}^2_{\mathrm{CV}(p)} = \frac{1}{C(N,0)} \sum_{j=1}^{C(N,0)} \nu(\mathbf{x}_{\setminus\{0,j\}}) - \frac{1}{C(N,N)} \sum_{i=1}^{C(N,N)} \nu(\mathbf{x}_{\setminus\{N,i\}}),$$

where $\nu(\mathbf{x}_{\setminus\{0,j\}}) = f(\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} f(\mathbf{x})$ and $\nu(\mathbf{x}_{\setminus\{N,i\}}) = \nu(\emptyset) = 0$. Thus

$$\frac{1}{N} \sum_{p=1}^{N} \hat{\sigma}^2_{\mathrm{CV}(p)} = \frac{f(\mathbf{x})^\top k(\mathbf{x},\mathbf{x})^{-1} f(\mathbf{x})}{N} = \hat{\sigma}^2_{\mathrm{ML}},$$

which establishes (A).

Appendix B. Further discussion on Theorem 4.4 . The requirement of having the same $V^2(f)$ for all sequences of partitions quasi-uniform with constant 2 can be relaxed somewhat: trivially, it is sufficient that the quadratic variation is $V^2(f)$ specifically with respect to even-points and odd-points sequences of sub-partitions used in the proof in Subsection 8.2.

Furthermore, we may even have different quadratic variations with respect to said sequences. Then the results becomes

$$\lim_{N\to\infty} \hat{\sigma}^2_{\mathrm{CV}} = \frac{\nu}{T} \qquad \text{for} \qquad \nu = \frac{V_0^2(f) + V_1^2(f)}{2},$$

where $V_0^2(f)$ and $V_1^2(f)$ are quadratic variations with respect to the even- and odd-points sub-partitions respectively, meaning that

$$\begin{aligned} V^2(f) &= \lim_{N\to\infty} \sum_{n=1}^{N-1} (f_{n+1} - f_n)^2, \\ V_0^2(f) &= \lim_{N\to\infty} \sum_{n=1}^{\lfloor \frac{N-2}{2} \rfloor} (f_{2n+2} - f_{2n})^2, \\ V_1^2(f) &= \lim_{N\to\infty} \sum_{n=1}^{\lfloor \frac{N-1}{2} \rfloor} (f_{2n+1} - f_{2n-1})^2. \end{aligned}$$

Appendix C. Explicit expression for the leave-p-out estimator.

Using the expressions for posterior mean and covariance functions in (8.1) and (8.1), we may derive an explicit expression for the leave-p-out cross-validation (LPO-CV) estimator of the scale parameter, given in (A) by

$$\hat{\sigma}^2_{\mathrm{CV}(p)} = \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \frac{1}{p} \sum_{n=1}^{p} \frac{[f(x_{p,i,n}) - m_{\setminus\{p,i\}}(x_{p,i,n})]^2}{k_{\setminus\{p,i\}}(x_{p,i,n})}.$$

The expression is less straightforward than that for $p = 1$. Denote by $x_{\lfloor p,i,n \rfloor}$ the largest point in the set $\mathbf{x}_{\setminus\{p,i\}} = \mathbf{x} \setminus \{x_{p,i,1}, \ldots, x_{p,i,p}\}$ that does not exceed $x_{p,i,n}$, and by $x_{\lceil p,i,n \rceil}$ the smallest point in the set $\mathbf{x}_{\setminus\{p,i\}}$ that exceeds $x_{p,i,n}$. Through somewhat cumbersome arithmetic derivations one can show that the estimator takes the form

$$\hat{\sigma}^2_{\mathrm{CV}(p)} = \frac{1}{C(N,p)} \sum_{i=1}^{C(N,p)} \left[B_{p,i,1} + \sum_{n=2}^{p-1} I_{p,i,n} + B_{p,i,p} \right]$$

where, for $\Delta x^-_{p,i,n} = (x_{p,i,n} - x_{\lfloor p,i,n \rfloor})$ and $\Delta x^+_{p,i,n} = (x_{\lceil p,i,n \rceil} - x_{p,i,n})$, the inner term is

$$I_{p,i,n} = \frac{\Delta x^-_{p,i,n}(f_{\lceil p,i,n \rceil} - f_{p,i,n}) - \Delta x^+_{p,i,n}(f_{p,i,n} - f_{\lfloor p,i,n \rfloor})}{(\Delta x^+_{p,i,n} + \Delta x^-_{p,i,n}) \Delta x^+_{p,i,n} \Delta x^-_{p,i,n}},$$

and the boundary terms $B_{p,i,1}$ and $B_{p,i,p}$ depend on whether the i^{th} set contains x_1 or x_N,

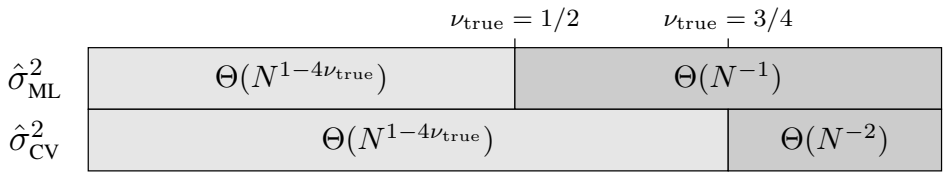

Figure 7. *Rates of decay for the ML and CV estimators for the Matérn kernel of order* 1*, and a true function that is a linear combination of Matérn kernels of order* ν_{true}*. The ML rate is given in Karvonen et al. (2020, Equation 5.2). The CV rate is empirically observed in Figure 8. Observe that the CV estimator's range of adaptation to the smoothness* ν_{true} *is wider than the ML estimator's.*

respectively. Specifically,

$$B_{p,i,1} = \begin{cases} \frac{(x_{\lceil p,i,1 \rceil} f_{p,i,1} - x_{p,i,1} f_{\lceil p,i,1 \rceil})^2}{x_{p,i,1} x_{\lceil p,i,1 \rceil} \Delta x^+_{p,i,1}} & \text{if the } i^{\text{th}} \text{ set contains } x_1, \\ I_{p,i,1} & \text{otherwise,} \end{cases}$$

$$B_{p,i,p} = \begin{cases} \frac{(f_{p,i,p} - f_{\lfloor p,i,p \rfloor})^2}{\Delta x^-_{p,i,p}} & \text{if the } i\text{'th set contains } x_N, \\ I_{p,i,p} & \text{otherwise.} \end{cases}$$

Though more cumbersome, it may be feasible to conduct convergence analysis similar to that in Section 4 for $\hat{\sigma}^2_{\text{CV}(p)}$. We leave this up to future work.

Appendix D. Comparison of CV and ML estimators for Matérn kernels.

A natural next step is extending the analysis to kernels whose reproducing kernel Hilbert spaces (RKHSs) are norm-equivalent to Sobolev spaces, such as the commonly used Matérn kernels. The ML estimator for Matérn kernels was analysed in Karvonen et al. (2020). Their experiments in Section 5.1 suggest that, for $x_+ \coloneqq \max(x, 0)$,

$$\hat{\sigma}^2_{\text{ML}} = \Theta(N^{2(\nu_{\text{model}} - 2\nu_{\text{true}})_+ - 1})$$

when $k_{\nu_{\text{model}}}$ is a Matérn kernel of order ν_{model} and f is a finite linear combination of the form $f = \sum_{i=1}^m \alpha_i k_{\nu_{\text{true}}}(\cdot, x_i)$ for some $m \in \mathbb{N}$, $\alpha_i \in \mathbb{R}$, $x_i \in [0,1]$, and the Matérn kernel $k_{\nu_{\text{true}}}$ of order ν_{true}. Empirically, we compare this to the rate of the CV estimator in Figure 8. The test functions f are posterior means of a GP with the $k_{\nu_{\text{true}}}$ kernel conditioned on points $\{(x_1, y_1), \ldots, (x_{10}, y_{10})\}$, where each x_i and y_i is sampled i.i.d from the uniform distribution on $[0,1]$. Since such f are of the form $f = \sum_{i=1}^{10} \alpha_i k_{\nu_{\text{true}}}(\cdot, x_i)$, we expect the MLE rate in (D) to apply; we use experimental data and the results in Theorems 4.1 and 4.6 to hypothesise what the rate in each individual example is. Similarly to the observations for the Brownian motion kernel, we see that the CV estimator adapts to the smoothness of the true function over a larger ranger of smoothness compared to the ML estimator. For instance, for $\nu_{\text{model}} = 1$, the experimental results suggest that the dependence of rate on ν_{true} is as illustrated in Figure 7. While the CV and the ML estimators adapt to the function smoothness when $\nu_{\text{true}} \leq 1/2$, for $\nu_{\text{true}} \in [1/2, 3/4]$ only the CV estimator continues adapting to the smoothness. This implies the CV estimator is less likely to become asymptotically overconfident in the event of undersmoothing.

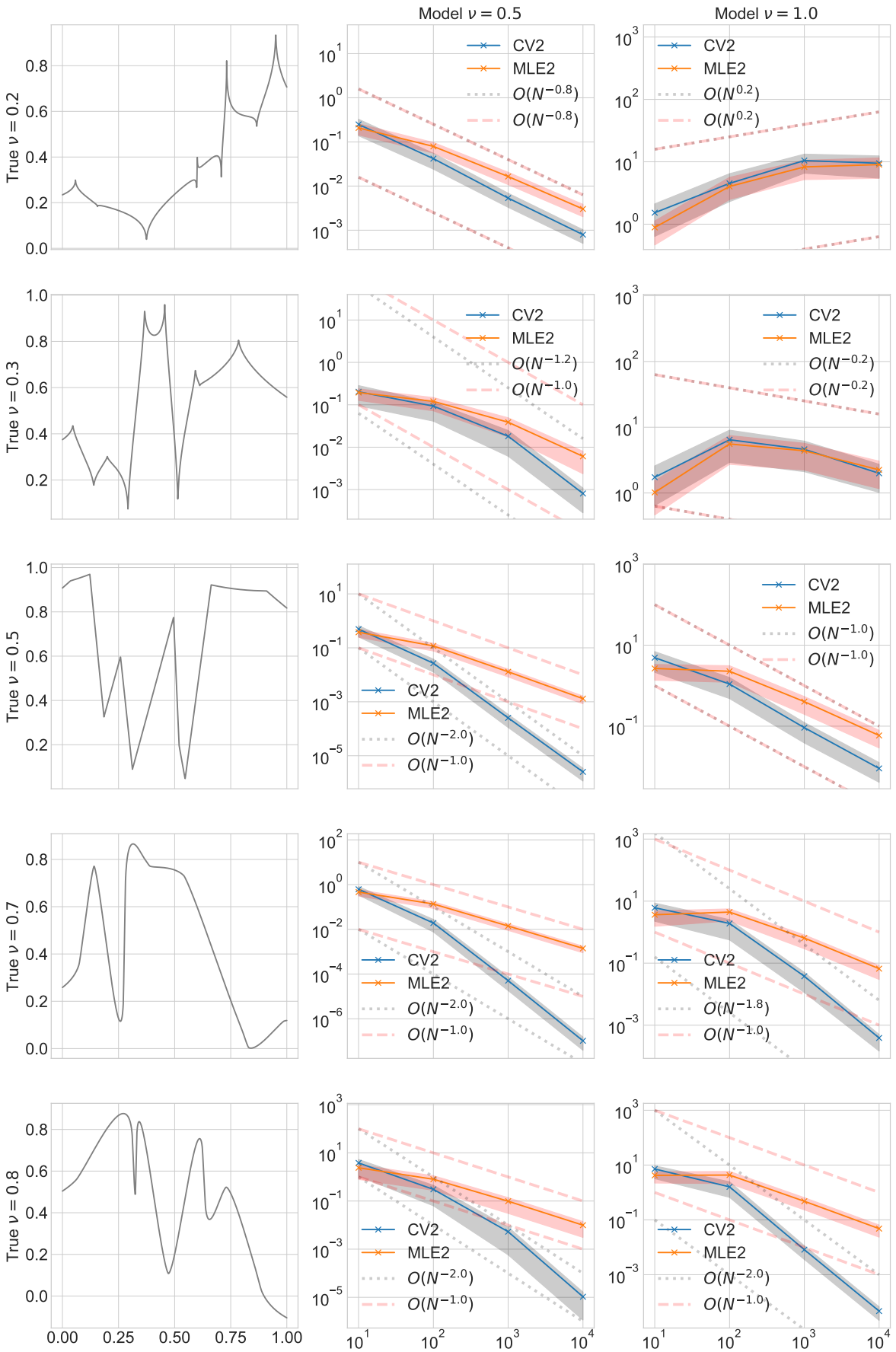

Figure 8. *Asymptotics of CV estimator compared to asymptotics of the ML estimator, for the Matérn kernel* ν_{model}*, and a true function that is a finite linear combination of Matérn kernels* ν_{true}*.*

References.

Anderes, E. (2010): "On the consistent separation of scale and variance for Gaussian random fields," *The Annals of Statistics*, 38, 870–893.

Bachoc, F. (2013): "Cross validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification," *Computational Statistics & Data Analysis*, 66, 55–69.

——— (2018): "Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case," *Bernoulli*, 24, 1531–1575.

Bachoc, F., J. Betancourt, R. Furrer, and T. Klein (2020): "Asymptotic properties of the maximum likelihood and cross validation estimators for transformed Gaussian processes," *Electronic Journal of Statistics*, 14, 1962–2008.

Bachoc, F., A. Lagnoux, and T. M. N. Nguyen (2017): "Cross-validation estimation of covariance parameters under fixed-domain asymptotics," *Journal of Multivariate Analysis*, 160, 42–67.

Beck, J. and S. Guillas (2016): "Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model," *SIAM/ASA Journal on Uncertainty Quantification*, 4, 739–766.

Ben Salem, M., F. Bachoc, O. Roustant, F. Gamboa, and L. Tomaso (2019): "Gaussian process-based dimension reduction for goal-oriented sequential design," *SIAM/ASA Journal on Uncertainty Quantification*, 7, 1369–1397.

Bevilacqua, M., T. Faouzi, R. Furrer, and E. Porcu (2019): "Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics," *The Annals of Statistics*, 47, 828–856.

Chen, Y., H. Owhadi, and A. M. Stuart (2021): "Consistency of empirical Bayes and kernel flow for hierarchical parameter estimation," *Mathematics of Computation*, 90, 2527–2578.

Cockayne, J., C. J. Oates, T. J. Sullivan, and M. Girolami (2019): "Bayesian probabilistic numerical methods," *SIAM Review*, 61, 756–789.

Cont, R. and P. Bas (2023): "Quadratic variation and quadratic roughness," *Bernoulli*, 29, 496–522.

Diaconis, P. (1988): "Bayesian numerical analysis," in *Statistical Decision Theory and Related Topics IV*, Springer-Verlag New York, vol. 1, 163–175.

Du, J., H. Zhang, and V. S. Mandrekar (2009): "Fixed-domain asymptotic properties of tapered maximum likelihood estimators," *The Annals of Statistics*, 37, 3330–3361.

Dudley, R. M. (1973): "Sample Functions of the Gaussian Process," *The Annals of Probability*, 1, 66–103.

Fong, E. and C. C. Holmes (2020): "On the marginal likelihood and cross-validation," *Biometrika*, 107, 489–496.

Garnett, R. (2023): *Bayesian Optimization*, Cambridge University Press.

Ginsbourger, D. and C. Schärer (2021): "Fast calculation of Gaussian process multiple-fold cross-validation residuals and their covariances," *arXiv preprint arXiv:2101.03108*.

Gu, M., X. Wang, and J. O. Berger (2018): "Robust Gaussian stochastic process emulation," *The Annals of Statistics*, 46, 3038–3066.

Hadji, A. and B. Szabó (2021): "Can we trust Bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel?" *SIAM/ASA Journal*

on Uncertainty Quantification, 9, 185–230.

Hennig, P., M. A. Osborne, and M. Girolami (2015): "Probabilistic numerics and uncertainty in computations," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471, 20150142.

Hennig, P., M. A. Osborne, and H. P. Kersting (2022): *Probabilistic Numerics: Computation as Machine Learning*, Cambridge University Press.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, 13, 455.

Karvonen, T. (2021): "Estimation of the scale parameter for a misspecified Gaussian process model," *arXiv preprint arXiv:2110.02810*.

——— (2023): "Asymptotic bounds for smoothness parameter estimates in Gaussian process interpolation," *SIAM/ASA Journal on Uncertainty Quantification*, 11, 1225–1257.

Karvonen, T. and C. J. Oates (2023): "Maximum likelihood estimation in Gaussian process regression is ill-posed," *Journal of Machine Learning Research*, 24, 1–47.

Karvonen, T., G. Wynne, F. Tronarp, C. J. Oates, and S. Särkkä (2020): "Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions," *SIAM/ASA Journal on Uncertainty Quantification*, 8, 926–958.

Kaufman, C. G. and B. A. Shaby (2013): "The role of the range parameter for estimation and prediction in geostatistics," *Biometrika*, 100, 473–484.

Kennedy, M. C. and A. O'Hagan (2001): "Bayesian calibration of computer models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 425–464.

Kiliç, E. (2008): "Explicit formula for the inverse of a tridiagonal matrix by backward continued fractions," *Applied Mathematics and Computation*, 197, 345–357.

Loh, W.-L. (2005): "Fixed-domain asymptotics for a subclass of Matérn-type Gaussian random fields," *The Annals of Statistics*, 33, 2344–2394.

Loh, W.-L. and T.-K. Kam (2000): "Estimating structured correlation matrices in smooth Gaussian random field models," *Annals of Statistics*, 28, 880–904.

Loh, W.-L. and S. Sun (2023): "Estimating the parameters of some common Gaussian random fields with nugget under fixed-domain asymptotics," *Bernoulli*, 29, 2519–2534.

Loh, W.-L., S. Sun, and J. Wen (2021): "On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions," *The Annals of Statistics*, 49, 3127–3152.

Mallik, R. K. (2001): "The inverse of a tridiagonal matrix," *Linear Algebra and its Applications*, 325, 109–139.

Mandelbrot, B. B. (1982): *The Fractal Geometry of Nature*, WH Freeman New York.

Mörters, P. and Y. Peres (2010): *Brownian motion*, vol. 30, Cambridge University Press.

Nourdin, I. (2012): *Selected aspects of fractional Brownian motion*, no. 4 in Bocconi & Springer Series, Springer.

O'Hagan, A. (1978): "Curve fitting and optimal design for prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, 40, 1–42.

O'Hagan, A. (2006): "Bayesian analysis of computer code outputs: A tutorial," *Reliability Engineering & System Safety*, 91, 1290–1300.

Petit, S. (2023): "Maximum likelihood estimation and prediction error for a Matérn model on the circle," *arXiv:2209.07791v5*.

Petit, S., J. Bect, P. Feliot, and E. Vazquez (2022): "Parameter selection in Gaussian process interpolation: An empirical study of selection criteria," *arXiv:2107.06006v4*.

Rasmussen, C. E. and C. K. Williams (2006): *Gaussian processes for machine learning*, vol. 2, MIT press Cambridge, MA.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989): "Design and analysis of computer experiments," *Statistical Science*, 4, 409–423.

Schaback, R. (2018): "Superconvergence of kernel-based interpolation," *Journal of Approximation Theory*, 235, 1–19.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas (2015): "Taking the human out of the loop: A review of Bayesian optimization," *Proceedings of the IEEE*, 104, 148–175.

Sniekers, S. and A. van der Vaart (2015): "Adaptive Bayesian credible sets in regression with a Gaussian process prior," *Electronic Journal of Statistics*, 9, 2475 – 2527.

——— (2020): "Adaptive Bayesian credible bands in regression with a Gaussian process prior," *Sankhya A*, 82, 386–425.

Stein, M. L. (1990): "A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process," *The Annals of Statistics*, 18, 1139–1157.

——— (1993): "Spline smoothing with an estimated order parameter," *The Annals of Statistics*, 21, 1522–1544.

——— (1999): *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Series in Statistics, Springer.

Sundararajan, S. and S. S. Keerthi (2001): "Predictive approaches for choosing hyperparameters in Gaussian processes," *Neural Computation*, 13, 1103–1118.

Szabó, B., A. W. van der Vaart, and J. H. van Zanten (2015): "Frequentist coverage of adaptive nonparametric Bayesian credible sets," *The Annals of Statistics*, 43, 1391–1428.

van der Vaart, A. W. and J. H. van Zanten (2008): *Reproducing Kernel Hilbert spaces of Gaussian Priors*, Institute of Mathematical Statistics, vol. 3 of *IMS Collections*, 200–222.

Wahba, G. (1990): *Spline Models for Observational Data*, no. 59 in CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics.

Wang, D. and W.-L. Loh (2011): "On fixed-domain asymptotics and covariance tapering in Gaussian random field models," *Electronic Journal of Statistics*, 5, 238–269.

Wang, W. (2007): "Almost-sure path properties of fractional Brownian sheet," *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 43, 619–631.

——— (2021): "On the inference of applying Gaussian process modeling to a deterministic function," *Electronic Journal of Statistics*, 15, 5014–5066.

Wendland, H. (2005): *Scattered Data Approximation*, no. 17 in Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.

Wenzel, T., G. Santin, and B. Haasdonk (2021): "A novel class of stabilized greedy kernel approximation algorithms: Convergence, stability and uniform point distribution," *Journal of Approximation Theory*, 262, 105508.

Wynne, G., F.-X. Briol, and M. Girolami (2021): "Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness," *Journal of Machine Learning Research*, 22, 1–40.

Xu, W. and M. L. Stein (2017): "Maximum Likelihood Estimation for a Smooth Gaussian Random Field Model," *SIAM/ASA Journal on Uncertainty Quantification*, 5, 138–175.

Ying, Z. (1991): "Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process," *Journal of Multivariate Analysis*, 36, 280–296.

——— (1993): "Maximum likelihood estimation of parameters under a spatial sampling scheme," *The Annals of Statistics*, 21, 1567–1590.

Zhang, H. (2004): "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, 99, 250–261.