Journal of Machine Learning Research 26 (2025) 1-70 Submitted 7/22; Published 8/25

Relaxed Gaussian Process Interpolation: a Goal-Oriented Approach to Bayesian Optimization

Sébastien J. Petit SEBASTIEN.PETIT@LNE.FR
Laboratoire National de Métrologie et d'Essais,
78197, Trappes Cedex, France[*]

Julien Bect JULIEN.BECT@CENTRALESUPELEC.FR
Université Paris-Saclay,
CNRS, CentraleSupélec,
Laboratoire des signaux et systèmes,
91190, Gif-sur-Yvette, France

Emmanuel Vazquez EMMANUEL.VAZQUEZ@CENTRALESUPELEC.FR
Université Paris-Saclay,
CNRS, CentraleSupélec,
Laboratoire des signaux et systèmes,
91190, Gif-sur-Yvette, France

Editor: Ryan Adams

Abstract

This work presents a new procedure for obtaining predictive distributions in the context of Gaussian process (GP) modeling, with a relaxation of the interpolation constraints outside ranges of interest: the mean of the predictive distribution no longer necessarily interpolates the observed values when they are outside ranges of interest, but is simply constrained to remain outside. This method called relaxed Gaussian process (reGP) interpolation provides better predictive distributions in ranges of interest, especially in cases where a stationarity assumption for the GP model is not appropriate. It can be viewed as a goal-oriented method and becomes particularly interesting in Bayesian optimization, for example, for the minimization of an objective function, where good predictive distributions for low function values are important. When the expected improvement criterion and reGP are used for sequentially choosing evaluation points, the convergence of the resulting optimization algorithm is theoretically guaranteed (provided that the function to be optimized lies in the reproducing kernel Hilbert space attached to the known covariance of the underlying Gaussian process). Experiments indicate that using reGP instead of stationary GP models in Bayesian optimization is beneficial.

Keywords: Gaussian processes; Bayesian optimization; Expected improvement; Goal-oriented modeling; Reproducing kernel Hilbert spaces

∗. Part of this work was funded by the French Agence Nationale de la Recherche et de la Technologie (ANRT) under a CIFRE grant, and was carried out while S. J. Petit was affiliated with Safran Aircraft Engines (Moissy-Cramayel, France) and L2S.

1. Introduction

1.1 Context and Motivation

Gaussian process (GP) interpolation and regression (see, e.g., Stein, 1999; Rasmussen and Williams, 2006) is a classical method for predicting an unknown function from data. It has found applications in active learning techniques, and notably in Bayesian optimization, a popular derivative-free global optimization technique for functions whose evaluations are time-consuming.

A GP model is defined by a mean function and a covariance function, which are generally selected from data within parametric families. The most popular models assume stationarity and rely on standard covariance functions such as the Matérn covariance. The assumption of stationarity yields models with relatively low-dimensional parameters. However, such a hypothesis can sometimes result in poor models when the function to be predicted has different scales of variation or different local regularities across the domain.

This is the case for instance in the motivating example given by Gramacy and Lee (2008), or in the even simpler toy minimization problem shown in Figure 1. The objective function in this example, which we shall call the Steep function, is smooth with an obvious global minimum around the point $x = 8$. However, the variations around the minimum are overshadowed by some steep variations on the left. Figure 2 shows a stationary GP fit with $n = 8$ points, where the parameters of the covariance function have been selected using maximum likelihood. Observe that the confidence bands are too large and that the conditional mean varies too much in the neighborhood of the global minimum, consistently with the stationary GP model that reflects the prior that our function oscillates around a mean value with a constant scale of variations. In this case, even if GP interpolation is consistent (Vazquez and Bect, 2010a), stationarity seems an unsatisfactory assumption for the Steep function. One expects Bayesian optimization techniques to be somehow inefficient on this problem with such a stationary model, whose posterior distributions are too pessimistic in the region of the minimum.

Nevertheless, the Steep function has the characteristics of an easy optimization problem: it has only two local minima, with the global minimum lying in a valley of significant volume. Consequently, a Bayesian optimization technique could be competitive if it relied on a model giving good predictions in regions where the function takes low values. In this work, we propose to explore goal-oriented GP modeling, where we want predictive models in regions of interest, even if it means being less predictive elsewhere.

1.2 Related Works

1.2.1 LOCAL MODELS

Going beyond the stationary hypothesis has been an active direction of research. Local models are one popular solution, which still uses stationary Gaussian processes as a core building block.

A first class of local models is obtained by considering partitions of the input domain with different GP models on each subset. Partitions can be built by splitting the domain along the coordinate axes. This is the case of the treed Gaussian process models proposed by Gramacy and Lee (2008), which combines a fully Bayesian framework and the use of RJ-MCMC techniques for the inference, or the trust-region method by Eriksson et al. (2019). Park and Apley (2018) also propose partition-based local models built by splitting the domain along principal component directions.

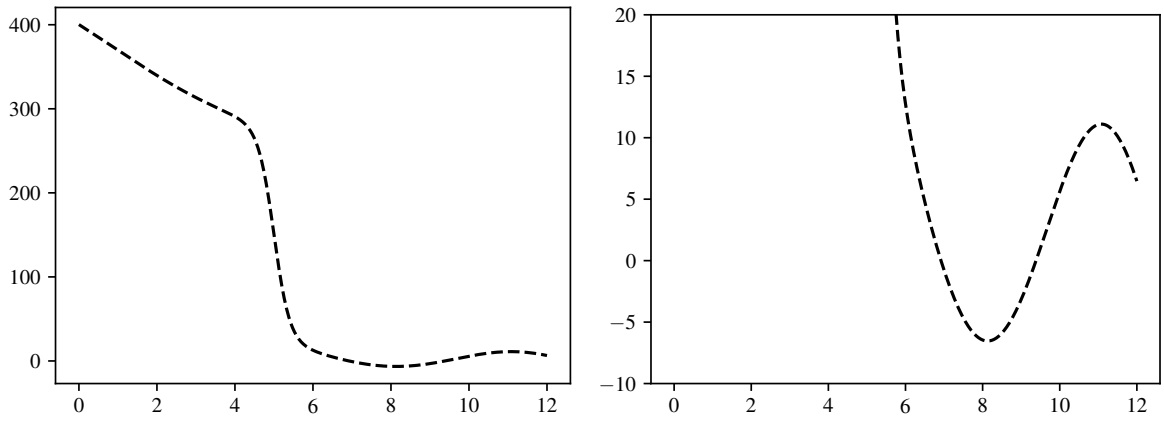

Figure 1: Left: the Steep function. Right: same illustration with a restrained range on the y-axis. The variations on the left tend to overshadow the global minimum on the right.

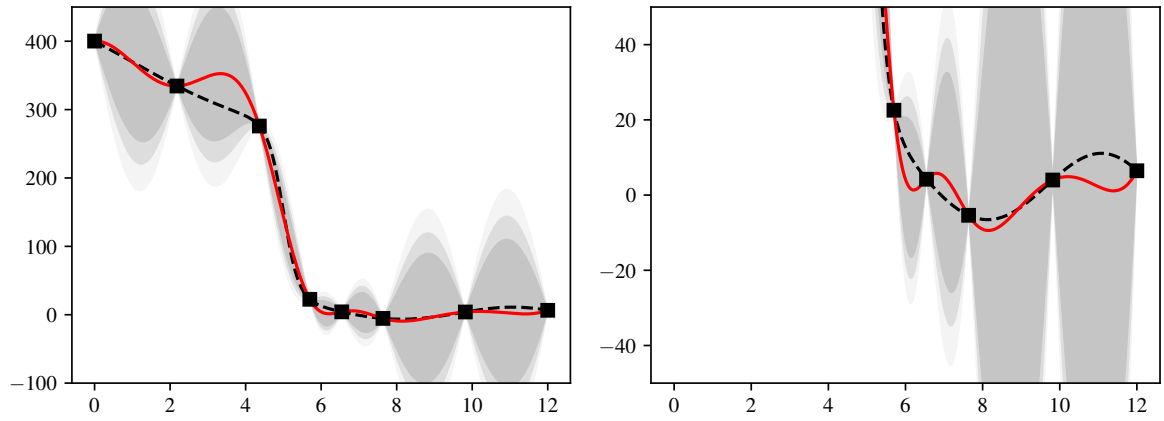

Figure 2: Left: GP fit on the Steep function. Right: same illustration with a restrained range on the y-axis. The squares represent the data. The red line represents the posterior mean μ_n given by the model and the gray envelopes represent the associated uncertainties.

In such techniques, there are parameters related to, e.g., the way the partitions evolve with the data, the size of the partitions, or how local Gaussian processes interact with each other.

A second class of local models is obtained by spatially weighting one or several GP models. Many schemes have been proposed, including methods based on partition of unity (Nott and Dunsmuir, 2002), weightings of covariance functions (Pronzato and Rendas, 2017; Rivoirard and Romary, 2011), and convolution techniques (see, e.g., Higdon, 1998; Gibbs, 1998; Higdon, 2002; Ver Hoef et al., 2004; Stein, 2005). Let us also mention data-driven aggregation techniques: composite Gaussian process models (Ba and Joseph, 2012), and mixture of experts techniques (see, e.g., Tresp, 2001; Rasmussen and Ghahramani, 2002; Meeds and Osindero, 2006; Yuan and Neubauer, 2009; Yang and Ma, 2011; Yuksel et al., 2012). In the latter framework, the weights are called gating functions and the estimation of the parameters and the inference are usually performed using EM, MCMC, or variational techniques. Weighting methods generally have parameters specifying weighting functions, with an increased need to watch for overfitting phenomena.

1.2.2 TRANSFORMATION AND COMPOSITION OF MODELS

Transformation and composition of models is another popular alternative to stationary GP modeling, which also uses stationary Gaussian processes as a core building block.

A first technique for composition of models consists in using a parametric transformation of a GP (Rychlik et al., 1997; Snelson et al., 2004).

Another route is to transform the input domain, using for instance a parametric density (Xiong et al., 2007), or other parametric transformations involving possible dimension reduction (Marmin et al., 2018). Bodin et al. (2020) proposed a framework that uses additional input variables, serving as nuisance parameters, to smooth out some badly behaved data. The practitioner has to specify a prior over the variance of the nuisance parameter and inference is based on MCMC.

Lázaro-Gredilla (2012) takes the step of choosing a GP prior on the output transform and resorts to variational inference techniques. This type of idea can be viewed as an ancestor of deep Gaussian processes (see, e.g., Damianou and Lawrence, 2013; Dunlop et al., 2018; Hebbal et al., 2021; Jakkala, 2021; Bachoc and Lagnoux, 2025), which stack layers of linear combinations of GPs. The practitioner has to specify a network structure among other parameters and resort to variational inference.

Picheny et al. (2019) proposed another approach where predictions are made only from pairwise comparisons between data points, relying on the variational framework of ordinal GP regression proposed by Chu and Ghahramani (2005) for the inference.

1.2.3 GOAL-ORIENTED APPROXIMATE INFERENCE TECHNIQUES

Another related area of research is goal-oriented approximate inference, where the task to be performed with the predictive model is taken into account in the approximations that are used in the inference (a.k.a. training) step.

For instance, in the context of scalable Bayesian optimization, McIntire et al. (2016) and Moss et al. (2023) have proposed goal-oriented criteria to choose inducing points. Another example is Yang et al. (2021), who propose a goal-oriented criterion to select the frequencies of a sparse spectrum approximation. Loss-calibrated approximate Bayesian computation (Lacoste–Julien et al., 2011; Morais and Pillow, 2022), which takes into account the decision-making task when approximating the posterior distribution, also falls in this category.

These works focus on fine-tuning approximations of classical GP models, by taking into account the task to be performed. This is in contrast with the method introduced in this article, which aims to *replace* the classical GP model with a new, goal-oriented GP-based predictive distribution, which no longer stems from conditioning a GP prior. (The development of a massively scalable version of the proposed method is beyond the scope of this article, but could be based on approximate goal-directed inference techniques.)

1.3 Contributions and Outline

The main contribution of this article is a method called *relaxed Gaussian processes* (reGP) for building goal-oriented GP-based models targeting regions of interest specified through function values. The objective is to obtain global models that exhibit good predictive distributions on a range of interest—in the case of a minimization problem, the range of interest would be the values below a threshold—while possibly being less predictive outside the range of interest. This is achieved by relaxing interpolation constraints outside this range. Such a model is presented in Figure 3: compared to the situation in Figure 2, the model is more predictive in the region where the Steep function takes low values, with expected benefits for the efficiency of Bayesian optimization.

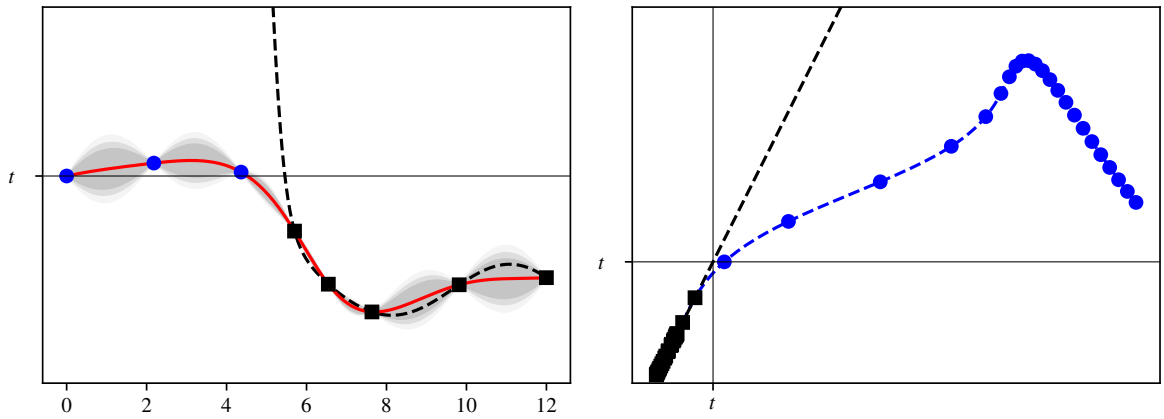

Figure 3: Left: prediction of the Steep function with the proposed methodology (black line: relaxation threshold t; blue points: relaxed observations). Right: μ_n versus f (with more observations for illustration purposes). The model interpolates the data below t. The blue points are relaxed observations.

This article provides two other main contributions. On the one hand, we give theoretical and empirical results justifying the method and its use for Bayesian optimization. On the other hand, to assess the predictivity of reGP models, we adopt the formalism of scoring rules (Gneiting and Raftery, 2007) and propose the use of a goal-oriented scoring rule that we call *truncated continuous ranked probability score* (tCRPS), which is designed to assess the predictivity of a model in a range of interest. The tCRPS is used to construct a leave-one-out (LOO) goodness-of-fit criterion to auto-

matically select the range of function values (outside the range of interest) where the interpolation constraints are relaxed.

An important feature of the proposed approach is that, with respect to classical GP modeling, it only requires the choice of one additional key parameter—namely, the range of interest where good predictions are expected. We propose several heuristics to set this range of interest in Bayesian optimization and level-set estimation problems.

The organization of this article is as follows. Section 2 briefly recalls the formalism of Gaussian processes and Bayesian optimization. Section 3 presents reGP and its theoretical properties. The tCRPS and its use for selecting the relaxation range are then presented in Section 4. Section 5 presents a reGP-based Bayesian optimization algorithm called EGO-R, together with the convergence analysis of this algorithm and a numerical benchmark. Finally, Section 6 presents our conclusions and perspectives for future work.

An open source implementation of the reGP method and the numerical experiments is available online at `https://github.com/relaxedGP/regp_paper_experiments`.

2. Background and Notations

2.1 Gaussian Process Modeling

Consider a real-valued function $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X} \subseteq \mathbb{R}^d$, and suppose we want to infer f at a given $x \in \mathbb{X}$ from evaluations of f on a finite set of points $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$, $n \geq 1$. A standard Bayesian approach to this problem consists in using a GP model $\xi \sim \mathrm{GP}(\mu, k)$ as a prior about f, where $\mu : \mathbb{X} \to \mathbb{R}$ is a mean function and $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a covariance function, which is supposed to be strictly positive-definite in this article.

The posterior distribution of ξ given $\underline{Z}_n = (\xi(x_1), \ldots, \xi(x_n))^\mathsf{T}$ is still a Gaussian process, whose mean and covariance functions are given by the standard kriging equations (Matheron, 1971). More precisely:

$$\xi \,|\, \underline{Z}_n \sim \mathrm{GP}(\mu_n, k_n), \tag{1}$$

with

$$\mu_n(x) = \mu(x) + k(x, \underline{x}_n) K_n^{-1} (\underline{Z}_n - \mu(\underline{x}_n)) \tag{2}$$

and

$$k_n(x, y) = k(x, y) - k(x, \underline{x}_n) K_n^{-1} k(y, \underline{x}_n)^\mathsf{T},$$

and where $\mu(\underline{x}_n) = (\mu(x_1), \ldots, \mu(x_n))^\mathsf{T}$, $k(x, \underline{x}_n) = (k(x, x_1), \ldots, k(x, x_n))$, and K_n is the $n \times n$ matrix with entries $k(x_i, x_j)$. We shall also use the notation $\sigma_n^2(x) = k_n(x, x)$ for the posterior variance, a.k.a. the kriging variance, a.k.a. the squared power function, so that $\xi(x) \,|\, \underline{Z}_n \sim \mathcal{N}\left(\mu_n(x), \sigma_n^2(x)\right)$.

The functions μ and k control the posterior distribution (1) and must be chosen carefully. The standard practice is to select them from data within a parametric family $\{(\mu_\theta, k_\theta),\ \theta \in \Theta\}$. A common approach is to suppose stationarity for the GP, which means choosing a constant mean function $\mu \equiv c \in \mathbb{R}$ and a stationary covariance function $k(x, y) = \tau^2 r(x - y)$, where $r : \mathbb{R}^d \to \mathbb{R}$ is a stationary correlation function.

A correlation function often recommended in the literature (Stein, 1999) is the (geometrically anisotropic) Matérn correlation function

$$r(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\|h\|_\rho\right)^\nu \mathcal{K}_\nu \left(\sqrt{2\nu}\|h\|_\rho\right), \quad \|h\|_\rho^2 = \sum_{j=1}^d \frac{h_{[j]}^2}{\rho_j^2}, \tag{3}$$

for $h = (h_{[1]}, \ldots, h_{[d]}) \in \mathbb{R}^d$, and where Γ is the Gamma function and $\mathscr{K}_\nu$ is the modified Bessel function of the second kind. The covariance parameters to be selected in this case are $(\tau^2, \rho_1, \ldots, \rho_d, \nu) \in (0,\infty)^{d+2}$ with τ^2 the process variance, ρ_i the range parameter along the i-th dimension, and ν a regularity parameter controlling the smoothness of the process. Two other standard covariance functions can be recovered for specific values of ν: the exponential covariance function for $\nu = 1/2$ and the squared-exponential covariance function for $\nu \to \infty$.

A variety of techniques for selecting the parameter θ have been proposed in the literature, but we can safely say that maximum likelihood estimation is the most popular and can be recommended in the case of interpolation (Petit et al., 2023). It simply consists in minimizing the negative log-likelihood

$$\mathscr{L}(\theta; \underline{Z}_n) = -\log\left(p\left(\underline{Z}_n \,|\, \theta\right)\right) \propto \log\left(\det\left(K_n\right)\right) + \left(\underline{Z}_n - \mu(\underline{x}_n)\right)^{\mathsf{T}} K_n^{-1} \left(\underline{Z}_n - \mu(\underline{x}_n)\right) + \text{constant}, \quad (4)$$

where p stands for the probability density of $\underline{Z}_n$. Other methods for selecting the parameters include the restricted maximum likelihood method and leave-one-out strategies (see, e.g., Stein, 1999; Rasmussen and Williams, 2006).

2.2 Bayesian Optimization

The framework of GPs is well suited to the problem of sequential design of experiments, or active learning. In particular, for minimizing a real-valued function f defined on a compact domain $\mathbb{X}$, the Bayesian approach consists in sequentially choosing evaluation points $x_1, x_2, \ldots \in \mathbb{X}$ using a GP model ξ for f, which makes it to possible to build a sampling criterion that represents an expected information gain on the minimum of f when an evaluation is made at a new point. One of the most popular sampling criteria (also called acquisition function) is the *Expected Improvement* (EI) (Mockus et al., 1978; Jones et al., 1998), which can be expressed as

$$\rho_n(x) = \mathsf{E}\left(\left(m_n - \xi(x)\right)_+ \,|\, \underline{Z}_n\right), \quad (5)$$

where $m_n = \min(\xi(x_1), \ldots, \xi(x_n))$. The EI criterion corresponds to the expectation of the excursion of ξ below the minimum given n observations, and can be written in closed form:

Proposition 1 *(Jones et al., 1998; Vazquez and Bect, 2010b) The EI criterion may be written as* $\rho_n(x) = \gamma\left(m_n - \mu_n(x), \sigma_n^2(x)\right)$ *with*

$$\gamma : (z, s) \in \mathbb{R} \times \mathbb{R}_+ \mapsto \begin{cases} \sqrt{s}\phi\left(\frac{z}{\sqrt{s}}\right) + z\Phi\left(\frac{z}{\sqrt{s}}\right) & \text{if } s > 0, \\ \max(z, 0) & \text{if } s = 0, \end{cases}$$

where ϕ and Φ stand for the probability density and cumulative distribution functions of the standard Gaussian distribution. Moreover, the function γ is continuous, satisfies $\gamma(z, s) > 0$ if $s > 0$ and is non-decreasing with respect to z and s on $\mathbb{R} \times \mathbb{R}_+$.

When the EI criterion is used for optimization, that is, when the sequence of evaluation points $(x_n)_{n>0}$ of f is chosen using the rule

$$x_{n+1} = \operatorname*{arg\,max}_{x \in \mathbb{X}} \rho_n(x),$$

the resulting algorithm is generally called the Efficient Global Optimization (EGO) algorithm, as proposed by Jones et al. (1998). The EGO algorithm has known convergence properties (Vazquez and Bect, 2010b; Bull, 2011).

A variety of other sampling criteria for the minimization problem can be found in the literature (see, e.g., Frazier et al., 2008; Villemonteix et al., 2009; Srinivas et al., 2010; Vazquez and Bect, 2014), but we shall focus on the EI criterion in this article. (Appendix E also considers the Upper Confidence Bound (UCB) criterion by Srinivas et al., 2010, and Appendix F deals with the task of estimating excursion sets using the straddle heuristic by Bryan et al., 2005.)

2.3 Reproducing Kernel Hilbert Spaces

Reproducing kernel Hilbert spaces (RKHS, see, e.g., Aronszajn, 1950; Berlinet and Thomas-Agnan, 2004) are Hilbert spaces of functions commonly used in the field of approximation theory (see, e.g., Wahba, 1990; Wendland, 2004). A Hilbert space $\mathscr{H}(\mathbb{X})$ of real-valued functions on $\mathbb{X}$ with an inner product $(\cdot,\cdot)_{\mathscr{H}(\mathbb{X})}$ is called an RKHS if it has a reproducing kernel, that is, a function $k\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ such that $k(x,\cdot)\in\mathscr{H}(\mathbb{X})$, and

$$(f, k(x,\cdot))_{\mathscr{H}(\mathbb{X})} = f(x) \tag{6}$$

(the reproducing property), for all $x\in\mathbb{X}$ and $f\in\mathscr{H}(\mathbb{X})$. Furthermore, given a (strictly) positive definite covariance function k, there exists a unique RKHS admitting k as reproducing kernel.

Given locations $\underline{x}_n = (x_1,\ldots,x_n)\in\mathbb{X}^n$, and corresponding values $\underline{z}_n\in\mathbb{R}^n$, suppose we want to find a function $h\in\mathscr{H}(\mathbb{X})$ with minimal norm, such that $h(\underline{x}_n) = (h(x_1),\ldots,h(x_n))^\mathsf{T} = \underline{z}_n$. Then the solution is given by the following classical result, which can be derived from the generalized representer theorem (Schölkopf et al., 2001) and appears in early work on optimal interpolation in RKHS (Kimeldorf and Wahba, 1970).

Proposition 2 *(Minimum-norm interpolant) The problem*

$$\begin{cases} \textit{minimize} & \|h\|_{\mathscr{H}(\mathbb{X})} \\ \textit{subject to} & h\in\mathscr{H}(\mathbb{X}) \\ & h(\underline{x}_n) = \underline{z}_n \end{cases} \tag{7}$$

has a unique solution given by $s_{\underline{z}_n} = k(\cdot,\underline{x}_n)K_n^{-1}\underline{z}_n$.

Observe that the solution $s_{\underline{z}_n}$ is equal to the posterior mean (2) when $\mu = 0$.

Moreover, for any $f\in\mathscr{H}(\mathbb{X})$ and $x\in\mathbb{X}$, the reproducing property (6) yields the upper bound

$$\left|f(x) - s_{\underline{z}_n}(x)\right| \le \sigma_n(x)\,\|f\|_{\mathscr{H}(\mathbb{X})}, \tag{8}$$

with $\sigma_n(x) = \sqrt{k_n(x,x)}$ and $\underline{z}_n = f(\underline{x}_n)$ (see, e.g., Scheurer et al., 2013, p. 13). Note that $\sigma_n(x)$ is the worst-case error at x for the interpolation of functions in the unit ball of $\mathscr{H}(\mathbb{X})$.

3. Relaxed Gaussian Process Interpolation

3.1 Relaxed Interpolation

The example in the introduction (see Figures 1–3) suggests that, in order to gain accuracy over a range of values of interest, it can be beneficial to relax interpolation constraints outside this range.

More precisely, the probabilistic model in Figure 3 interpolates data lying below a selected threshold t, and when data are above t, the model only keeps the information that the data exceeds t.

In the following, we consider the general setting where relaxation is carried out on a set of function values of the form $R = \bigcup_{j=1}^{J} R_j$, where $R_1, \ldots, R_J \subset \mathbb{R}$ are disjoint closed intervals with non-zero lengths. (The set $R = [t, +\infty)$ was used in the example of Figure 3.)

As above, we shall write $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$ for a sequence of locations with corresponding function values $\underline{z}_n = (z_1, \ldots, z_n)^\mathsf{T} \in \mathbb{R}^n$. Then, we introduce the set $C_{R,n} = C_1 \times \cdots \times C_n \subset \mathbb{R}^n$ of relaxed constraints, where

$$\begin{cases} C_i = R_j & \text{if } z_i \in R_j \text{ for some } j, \\ C_i = \{z_i\} & \text{otherwise.} \end{cases} \tag{9}$$

Let also

$$\mathscr{H}_{R,n} = \{h \in \mathscr{H}(\mathbb{X}) \mid h(\underline{x}_n) \in C_{R,n}\}$$

be the set of relaxed-interpolating functions. The following proposition gives the definition of the minimum-norm relaxed predictor.

Proposition 3 *The problem*

$$\begin{cases} \textit{minimize} & \|h\|_{\mathscr{H}(\mathbb{X})} \\ \textit{subject to} & h \in \mathscr{H}_{R,n} \end{cases} \tag{10}$$

has a unique solution given by $s_{\underline{z}_n^\star}$, where $\underline{z}_n^\star$ is the unique solution of the quadratic problem

$$\operatorname*{arg\,min}_{\underline{z} \in C_{R,n}} \underline{z}^\mathsf{T} K_n^{-1} \underline{z}. \tag{11}$$

An extension of this relaxation scheme to the case of noisy observations, where the function no longer exactly interpolates the data, is given in Proposition 41.

3.2 Relaxed Gaussian Process Interpolation

The main advantage of Gaussian processes is the possibility to obtain not only point predictions but also predictive distributions. However, Proposition 3 only defines a function approximation. We now turn relaxed interpolation into a probabilistic model providing predictive distributions whose mean is not constrained to interpolate data on a given relaxation range R. The following proposition makes a step in this direction.

Proposition 4 *Let $\xi \sim \mathrm{GP}(0, k)$, $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$, $\underline{z}_n \in \mathbb{R}^n$ and $\underline{x}'_m = (x'_1, \ldots, x'_m) \in \mathbb{X}^m$ be a set of locations of interest where predictions should be made. Write $\underline{Z}_n = (\xi(x_1), \ldots, \xi(x_n))^\mathsf{T}$ and $\underline{Z}'_m = (\xi(x'_1), \ldots, \xi(x'_m))^\mathsf{T}$. Then the mode of the probability density function*

$$p\left(\underline{Z}'_m, \underline{Z}_n \,|\, \underline{Z}_n \in C_{R,n}\right) \tag{12}$$

is given by $\left(s_{\underline{z}_n^\star}(\underline{x}'_m),\ \underline{z}_n^\star\right)$.

In other words, the relaxed interpolation solution of Proposition 3 corresponds to the maximum a posteriori (MAP) estimate under the predictive model (12). Conditioning on events of the form $\underline{Z}_n \in C_{R,n}$ has been used in Bayesian statistics as a principled way to handle outliers and model

misspecification; see, e.g., Lewis et al. (2021) and references therein. Such conditional distributions also appear in other settings: in constrained Gaussian process modeling, where constraints encode expert knowledge (Da Veiga and Marrel, 2012; Maatouk and Bay, 2017; López-Lopera et al., 2018), and in classification, ordinal regression, and preference learning problems, where observations impose inequality constraints (Benavoli et al., 2021).

However, the predictive distribution (12) is non-Gaussian since the support of $\underline{Z}_n$ is truncated. In particular, no closed-form expression is available for any of its moments, and sampling requires advanced techniques (e.g., variational, MCMC; see Section 6 for a discussion). Motivated by this observation, we propose instead to build a goal-oriented probabilistic model using the following definition.

Definition 5 (Relaxed-GP predictive distribution; fixed μ and k) *Given $\underline{x}_n \in \mathbb{X}^n$, $\underline{z}_n \in \mathbb{R}^n$, and a relaxation range R (finite union of closed intervals), the* relaxed-GP *(reGP) predictive distribution with fixed mean function μ and covariance function k is defined as the (Gaussian) conditional distribution of $\xi \sim \mathrm{GP}(\mu, k)$ given $\underline{Z}_n = \underline{z}_n^\star$, where*

$$\underline{z}_n^\star = \operatorname*{arg\,min}_{\underline{z} \in C_{R,n}} \; (\underline{z} - \mu(\underline{x}_n))^\mathsf{T} K_n^{-1} (\underline{z} - \mu(\underline{x}_n)), \tag{13}$$

with $C_{R,n}$ defined by (9)*.*

Observe that (13) reduces to (11) when $\mu = 0$. Consequently, the mean of the distribution is the predictor $s_{\underline{z}_n^\star}$ from Proposition 3 in this particular case. Moreover, the reGP predictive distribution can be seen as an approximation of (12), where $p(\underline{Z}_n \,|\, \underline{Z}_n \in C_{R,n})$ has been replaced by its mode. As discussed earlier, the main advantage of the reGP predictive distribution compared to (12) is its reasonable computational cost since it is a GP. Therefore, it makes it possible to use adaptive strategies for the choice of R, as demonstrated in Section 4. Moreover, it also has appealing theoretical approximation properties, as discussed in Section 3.3.

As discussed in Section 2.1, the standard practice is to select the mean and the covariance functions within a parametric family $\{(\mu_\theta, k_\theta),\ \theta \in \Theta\}$. Leveraging the connection between (13) and (4), we propose to perform the parameter selection and the relaxation jointly. This is formalized by the following definition of relaxed Gaussian process interpolation.

Definition 6 (Relaxed-GP predictive distribution; estimated parameters) *Given $\underline{x}_n \in \mathbb{X}^n$, $\underline{z}_n \in \mathbb{R}^n$, a relaxation range R (finite union of closed intervals), and parametric families (μ_θ) and (k_θ) as in Section 2.1, the* relaxed-GP *(reGP) predictive distribution with estimated parameters is the (Gaussian) conditional distribution of $\xi \sim \mathrm{GP}(\mu_\theta, k_\theta)$ given $\underline{Z}_n = \underline{z}_n^\star$, where $\underline{z}_n^\star$ and $\theta = \widehat{\theta}_n$ are obtained jointly by minimizing the negative log-likelihood:*

$$\left(\widehat{\theta}_n,\ \underline{z}_n^\star\right) \;=\; \operatorname*{arg\,min}_{\theta \in \Theta,\ \underline{z} \in C_{R,n}} \; \mathscr{L}(\theta; \underline{z}), \tag{14}$$

with $C_{R,n}$ defined by (9)*.*

An extension of this definition to the case of noisy observations is presented in Appendix G.

The mean function and the covariance function usually vary continuously with θ on an open subset of $\mathbb{R}^q$. In this case, it can be shown that the mapping $\theta \mapsto \min_{\underline{z} \in C_{R,n}} \mathscr{L}(\theta; \underline{z})$ is continuous (using, e.g., Theorem 2.2 from Best and Ding, 1995). Consequently, the minimum of (14) is reached when θ is restricted to a compact subset $\Theta \subset \mathbb{R}^q$.

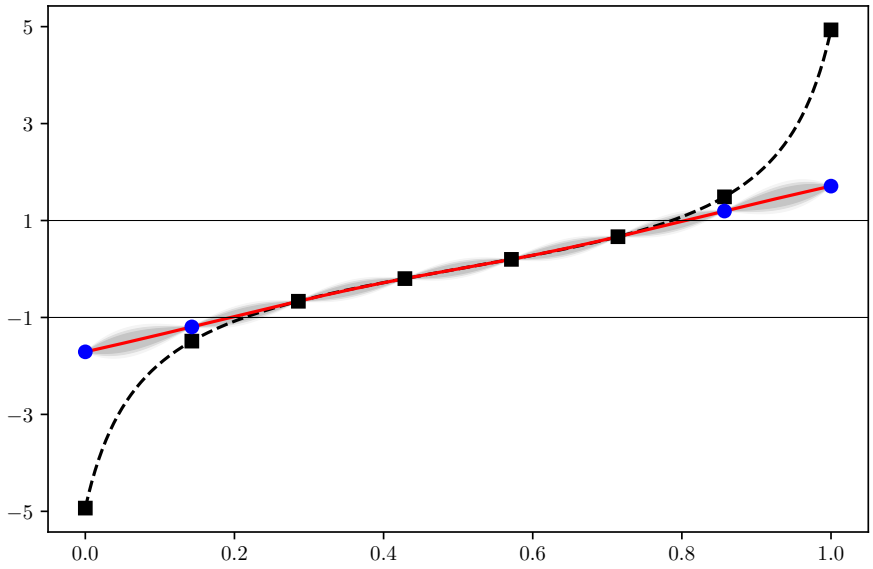

Figure 4: An example of reGP predictive distribution with $R = (-\infty, -1] \cup [1, +\infty)$ on a function f represented in dashed black lines. The solid black lines represent the relaxation thresholds. The problem (14) was solved only in $\underline{z}$ as the parameters of the (constant) mean and ($\nu = 5/2$ Matérn) covariance functions were held fixed for illustration purposes.

Remark 7 (On minimizing (14) jointly) *Let $\underline{Z}_{n,1}$ be the values within the relaxation range R, and $\underline{Z}_{n,0}$ the values in $R^c = \mathbb{R} \setminus R$ that are not relaxed. The negative log-likelihood can be written as*

$$\mathscr{L}(\theta; \underline{Z}_n) = -\ln\left(p\left(\underline{Z}_{n,0} \,|\, \theta\right)\right) - \ln\left(p\left(\underline{Z}_{n,1} \,|\, \theta, \underline{Z}_{n,0}\right)\right),$$

where the first term is a goodness-of-fit criterion based on the values in R^c, and where the second term can mainly be viewed as an imputation term, which "reshapes" the values in R with the information from $\underline{Z}_{n,0}$. (Note also that θ appears in the second term. When this term is minimized with respect to $\underline{Z}_{n,1}$, it becomes a parameter selection term that promotes the θs compatible with the excursions in $C_{R,n}$.)

For illustration, we provide an example of a reGP predictive distribution in Figure 4, with an union of two intervals for the relaxation range R.

Remark 8 (Numerical details) *Minimizing (14) with respect to $\underline{z}$ falls under the scope of quadratic programming (see, e.g., Nocedal and Wright, 2006) and could be solved efficiently using dedicated algorithms. This suggests that specific algorithms could be developed for the problem. In this work, we simply use a standard SLSQP solver (see Schittkowski, 1982) using the gradient of (14). The SciPy implementation was used with default parameters. Good performances were obtained by the following parameter initialization procedure: first, initialize the GP parameter θ using common rules while excluding the most extreme observations, and then optimize the relaxed observations $\underline{z}$ using quadratic programming.*

3.3 Convergence Analysis of reGP

In this section, we provide theoretical results concerning the convergence of the method proposed. This section can be skipped on first reading.

3.3.1 KNOWN CONVERGENCE RESULTS ABOUT INTERPOLATION IN RKHS

Recall that the fractional-order Sobolev space $W_2^\beta(\mathbb{R}^d)$, with regularity $\beta \geq 0$, is the space of functions on $\mathbb{R}^d$ defined by

$$W_2^\beta(\mathbb{R}^d) = \left\{ h \in L^2(\mathbb{R}^d), \quad \|h\|^2_{W_2^\beta(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \left(1 + \|\omega\|^2\right)^\beta |\widehat{h}(\omega)|^2 \mathrm{d}\omega < +\infty \right\},$$

where $\widehat{h} \in L^2(\mathbb{R}^d)$ is the Fourier transform of $h \in L^2(\mathbb{R}^d)$. For $\beta > d/2$, the usual identification with a set of continuous functions given by the Sobolev embedding theorem will be used in this article.

For a given $\mathbb{X} \subset \mathbb{R}^d$, define the Sobolev spaces $W_2^\beta(\mathbb{X}) = \left\{ h_{|\mathbb{X}}, h \in W_2^\beta(\mathbb{R}^d) \right\}$ endowed with the norm

$$\|h\|_{W_2^\beta(\mathbb{X})} = \inf_{g \in W_2^\beta(\mathbb{R}^d), g_{|\mathbb{X}} = h} \|g\|_{W_2^\beta(\mathbb{R}^d)}. \tag{15}$$

The following assumption about $\mathbb{X}$ will sometimes be used in this section.

Assumption 9 *The domain is non-empty, compact, connected, has locally Lipschitz boundary (see, e.g., Adams and Fournier, 2003, Section 4.9), and is equal to the closure of its interior.*

Assumption 9 ensures that the previous definition coincides with other commons definitions, and makes it possible to use well-known results from the field of scattered data approximation, by preventing the existence of cusps. Many common domains—such as hyperrectangles or balls, for instance—satisfy Assumption 9.

A strictly positive-definite reproducing kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is said to have regularity $\alpha > 0$ if the associated RKHS $\mathscr{H}(\mathbb{X})$ coincides with $W_2^{\alpha+d/2}(\mathbb{X})$ as a function space, with equivalent norms. As such, the Matérn stationary kernels (3) have correlation functions r whose Fourier transform satisfies (see, e.g., Wendland, 2004, Theorem 6.13)

$$C_1 \left(1 + \|\cdot\|^2\right)^{-\nu - d/2} \leq \widehat{r} \leq C_2 \left(1 + \|\cdot\|^2\right)^{-\nu - d/2}$$

for some $C_2 \geq C_1 > 0$, and have therefore Sobolev regularity $\alpha = \nu$ on $\mathbb{R}^d$ (see, e.g., Wendland, 2004, Corollary 10.13) and consequently also on $\mathbb{X}$, using (15) and Lemma 31. Other examples are given by Wendland (2004), for instance.

We now recall a classical convergence result about interpolation in RKHS with evaluation points in a bounded domain. Consider a kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, and let $(x_n)_{n \geq 1} \in \mathbb{X}^{\mathbb{N}}$ be a sequence of distinct points. The following property (a minor reformulation of Theorem 4.1 of Arcangéli et al., 2007) gives error bounds that depend on the Sobolev regularity of k and the so-called fill distance of $\underline{x}_n \in \mathbb{X}^n$, defined by

$$h_n = \sup_{x \in \mathbb{X}} \min_{1 \leq i \leq n} \|x - x_i\|.$$

Proposition 10 *Let k be a reproducing kernel with regularity $\alpha > 0$. If $\mathbb{X}$ satisfies Assumption 9, then*

$$\sup_{x\in\mathbb{X}} \sigma_n(x) \lesssim h_n^\alpha, \quad n \geq 1\,, \tag{16}$$

where $\lesssim$ denotes inequality up to a constant, that does not depend on $(x_n)_{n\geq 1}$.

Using (8) and Proposition 10, this yields the following uniform bound.

Corollary 11 *Let k be a reproducing kernel with regularity $\alpha > 0$, $\mathscr{H}(\mathbb{X})$ the RKHS generated by k, and let $f \in \mathscr{H}(\mathbb{X})$. As above, let $s_{\underline{z}_n}$ be the solution of (7) for $\underline{z}_n = (f(x_1),\dots,f(x_n))^\mathsf{T}$, $n \geq 1$. If $\mathbb{X}$ satisfies Assumption 9, then*

$$\|f - s_{\underline{z}_n}\|_{L^\infty(\mathbb{X})} \lesssim h_n^\alpha \|f\|_{\mathscr{H}(\mathbb{X})}\,. \tag{17}$$

3.3.2 CONVERGENCE OF RELAXED INTERPOLATION WITH FIXED COVARIANCE FUNCTION

Let $k\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ be a continuous strictly positive-definite reproducing kernel. In this section, we consider the zero-mean reGP predictive distribution obtained from $\xi \sim \mathrm{GP}(0,k)$, with relaxed interpolation constraints on a union $R = \bigcup_{j=1}^q R_j$ of disjoints closed intervals R_j with non-zero length. Let $\mathscr{H}(\mathbb{X})$ be the RKHS attached to k, $f \in \mathscr{H}(\mathbb{X})$, and consider a sequence $(x_n)_{n\geq 1} \in \mathbb{X}^\mathbb{N}$ of distinct points. Furthermore, define the regions $\mathbb{X}_j = \{x \in \mathbb{X},\ f(x) \in R_j\}$ for $1 \leq j \leq q$ and $\mathbb{X}_0 = \mathbb{X} \setminus \bigcup_{j\geq 1} \mathbb{X}_j$. We give results about the limit of the sequence of reGP predictive distributions that indicate an improved fit in $\mathbb{X}_0$.

Let $s_{R,n} = s_{\underline{z}_n^\star}$ be the relaxed predictor from Proposition 3 based on $(x_1,\dots,x_n)$ and $(f(x_1),\dots,f(x_n))^\mathsf{T}$, $n \geq 1$. The following proposition establishes the limit behavior of $(s_{R,n})_{n\geq 1}$.

Proposition 12 *Let $\mathbb{U} \subset \mathbb{X}$ and let $\mathscr{H}_{R,\mathbb{U}}$ denote the set of functions $h \in \mathscr{H}(\mathbb{X})$ such that, for all $x \in \mathbb{U}$,*

$$\begin{cases} h(x) \in R_j & \text{if } f(x) \in R_j \text{ for some } j, \\ h(x) = f(x) & \text{otherwise.} \end{cases} \tag{18}$$

Then the problem

$$\min_{h\in\mathscr{H}_{R,\mathbb{U}}} \|h\|_{\mathscr{H}(\mathbb{X})} \tag{19}$$

has a unique solution denoted by $s_{R,\mathbb{U}}$. Moreover, when $\mathbb{U}$ is the closure of $\{x_n\}$,

$$s_{R,n} \xrightarrow{\mathscr{H}(\mathbb{X})} s_{R,\mathbb{U}}\,.$$

In particular, when $\{x_n\}$ is dense in $\mathbb{X}$, then $\mathbb{U} = \mathbb{X}$ and $(s_{R,n})_{n\geq 1}$ converges to $s_{R,\mathbb{X}}$, which is the minimal-norm element of the set $\mathscr{H}_{R,\mathbb{X}}$.

The next proposition tells us that the interpolation error on $\mathbb{X}_0$ can be bounded by a term that depends on the norm of $s_{R,\mathbb{X}}$.

Proposition 13 *For any $x \in \mathbb{X}_0$ and $n \geq 1$,*

$$|f(x) - s_{R,n}(x)| \leq 2\sigma_{n,0}(x)\|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})}, \tag{20}$$

where $\sigma_{n,0}$ is the power function obtained using only points in $\mathbb{X}_0$ for predictions and with the convention that $\sigma_{n,0}^2(x) = k(x,x)$ if there are no points.

This yields the following error bounds when the design is dense.

Proposition 14 *Suppose that $\mathbb{X}$ is compact, $\{x_n\}$ is dense, and that k has regularity $\alpha > 0$. Let $B \subset \mathbb{X}_0$ satisfy Assumption 9. Then, for all $n \geq 1$,*

$$\|f - s_{R,n}\|_{L^\infty(B)} \lesssim h_n^\alpha \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})}. \tag{21}$$

Let $d(y, A)$ be the distance of $y \in \mathbb{R}$ to $A \subset \mathbb{R}$. For $j \geq 1$, $x \in \mathbb{X}_j$, and for all $n \geq 1$:

$$d(s_{R,n}(x), R_j) \lesssim h_n^\alpha \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})} \quad \text{if } \alpha < 1,$$

$$d(s_{R,n}(x), R_j) \lesssim (|\ln(h_n)| + 1)^{1/2} \, h_n \, \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})} \quad \text{if } \alpha = 1,$$

and

$$d(s_{R,n}(x), R_j) \lesssim h_n \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})} \quad \text{if } \alpha > 1\,,$$

where $\lesssim$ denotes inequality up to a constant, that does not depend on f, n, x or (x_n).

Finally, we investigate the following question: how large can the norm of f be compared to that of the approximation $\|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})}$?

Proposition 15 *Suppose that k has regularity $\alpha > 0$ and that there exists some $j \geq 1$ such that $\mathbb{X}_j$ has a non-empty interior. We have*

$$\sup_{h \in \mathscr{H}_{R,\mathbb{X}}} \|h\|_{\mathscr{H}(\mathbb{X})} = +\infty, \tag{22}$$

with $\mathscr{H}_{R,\mathbb{X}}$ given by (18) *for $f \in \mathscr{H}(\mathbb{X})$.*

This result shows that the norm reduction obtained by approximating f with relaxed interpolation constraints can therefore be arbitrarily high in the finite-smoothness case. A stronger version of Proposition 15 for the special case where $R = [t, +\infty)$ can be derived, and shows that

$$\sup_{h \in \mathscr{H}_{R,\mathbb{X}}} \|h\|_{L^\infty(\mathbb{X})} = +\infty\,.$$

Overall, no matter the element of $\mathscr{H}_{R,\mathbb{X}}$ at hand, reGP converges to a function $s_{R,\mathbb{X}}$ which: coincides with f on $\mathbb{X}_0$, satisfies $f(x) \in R_j \Leftrightarrow s_{R,\mathbb{X}}(x) \in R_j$ for all $x \in \mathbb{X}$, and is "nicer" than f in the sense of $\|\cdot\|_{\mathscr{H}(\mathbb{X})}$. Furthermore, reGP yields error bounds carrying the norm of $s_{R,\mathbb{X}}$, which can be arbitrarily smaller than the norm of f in the case of a finite-smoothness covariance function.

Remark 16 *Note that $\sigma_n \leq \sigma_{n,0}$ due to the projection residuals interpretation. Empirical and theoretical results about the* screening effect *(see, e.g., Stein, 2011; Bao et al., 2020), suggests that $\sigma_n \simeq \sigma_{n,0}$, if k has smoothness $\alpha > 0$. In this case, observe that—no matter the element of $\mathscr{H}_{R,\mathbb{X}}$ at hand—the bound* (20) *is larger by only a small factor compared to* (8) *with $f = s_{R,\mathbb{X}}$. (However, to the best of our knowledge, no result exists concerning the screening effect for arbitrary designs.)*

Remark 17 (22) *does not hold in general for infinitely smooth covariance functions. For instance, Steinwart et al. (2006, Corollary 3.9) show that $\mathscr{H}_{R,\mathbb{X}} = \{f\}$ if the interior of $\mathbb{X}_0$ is not empty and k is the squared-exponential covariance function, i.e.,* (3)*, with $\nu \to \infty$.*

Appendix B offers insights into the convergence of relaxed interpolation for functions that lie outside the RKHS attached to the covariance. Additionally, we present results on the behavior of the approximation when the covariance parameters are not fixed.

4. Choice of the Relaxation Range

4.1 Towards Goal-Oriented Cross-Validation

The framework of reGP makes it possible to predict a function f from point evaluations of f. Suppose we are specifically interested in obtaining good predictive distributions in a range $Q \subset \mathbb{R}$ of function values, and accept degraded predictions outside this range. To achieve this goal, the idea of reGP is to relax interpolation constraints. Naturally, it makes sense to relax interpolation constraints outside the range of interest Q, but it can happen that relaxing all the interpolation constraints outside Q does not improve predictive distributions on Q. Therefore, the question arises as to how to automatically select a range R in $\mathbb{R} \setminus Q$, on which interpolation constraints should be relaxed.

In the following, we put $R^{(0)} = \mathbb{R} \setminus Q$, and we view the relaxation range R as a parameter of the reGP model, which has to be chosen in $R^{(0)}$ along with the parameters θ of the underlying GP ξ. A first idea for the selection of R is to rely on the standard leave-one-out cross-validation approach to select the parameters of a GP (Dubrule, 1983; Rasmussen and Williams, 2006; Zhang and Wang, 2010). Using the formalism of *scoring rules* (see, e.g., Gneiting and Raftery, 2007; Petit et al., 2023), selecting parameters by a leave-one-out approach amounts to minimizing a selection criterion written as

$$J_n(R) = \frac{1}{n} \sum_{i=1}^{n} S\left(P_{R,n,-i}, f(x_i)\right), \tag{23}$$

where $P_{R,n,-i}$ is the reGP predictive distribution with data $z_{n,-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$ and relaxation range R. The function S in (23) is a scoring rule, that is, a function $S : \mathscr{P} \times \mathbb{R} \to \mathbb{R} \cup \{-\infty, +\infty\}$, acting on a class $\mathscr{P}$ of probability distributions on $\mathbb{R}$, such that $S(P, z)$ assigns a loss for choosing a predictive distribution $P \in \mathscr{P}$, while observing $z \in \mathbb{R}$. Scoring rules make it possible to quantify the quality of probabilistic predictions.

Since the user is not specifically interested in good predictive distributions in $R^{(0)}$, validating the model on $R^{(0)}$ should not be a primary focus. However, simply restricting the sum (23) by removing indices i such that $f(x_i) \in R^{(0)}$ would make it impossible to assess if the model is good at predicting that $f(x) \in R^{(0)}$ for a given $x \in \mathbb{X}$. For instance, in the case of minimization, with $Q = (-\infty, t^{(0)})$ and $R^{(0)} = [t^{(0)}, +\infty)$, it is important to identify the regions corresponding to f being above $t^{(0)}$, even if we are not interested in accurate predictions above $t^{(0)}$, because we expect that an optimization algorithm should avoid the exploration of these regions.

In the next section, we propose instead to keep the whole leave-one-out sum (23), but to choose a scoring rule S that serves our goal-oriented approach.

4.2 Truncated Continuous Ranked Probability Score

An appealing class of scoring rules for goal-oriented predictive distributions is the class of weighted scoring rules for binary predictors (Gneiting and Raftery, 2007; Matheson and Winkler, 1976), which may be written as

$$S(P, z) = \int_{-\infty}^{+\infty} s(F_P(u), \mathbb{1}_{z \le u})\, \mu(\mathrm{d}u), \tag{24}$$

where $s\colon [0, 1] \times \{0, 1\} \to \mathbb{R} \cup \{-\infty, +\infty\}$ is a scoring rule for binary predictors, F_P is the cumulative distribution function of P, and μ is a Borel measure on $\mathbb{R}$. A well-known instance of (24) is the

continuous ranked probability score (Gneiting et al., 2005) written as

$$S^{\mathrm{CRPS}}(P,z)=\int_{-\infty}^{+\infty}(F_P(u)-\mathbb{1}_{z\leq u})^2\,\mathrm{d}u\,,$$

which is obtained by choosing the Brier score for s and the Lebesgue measure for μ.

For the case where we are specifically interested in obtaining good predictive distributions in a range of interest $Q\subset\mathbb{R}$, we propose to use the following scoring rule, which we call *truncated continuous ranked probability score* (tCRPS):

$$S_Q^{\mathrm{tCRPS}}(P,z)\;=\;\int_Q(F_P(u)-\mathbb{1}_{z\leq u})^2\,\mathrm{d}u. \tag{25}$$

This scoring rule, proposed by Lerch and Thorarinsdottir (2013) in a different context, reduces to S^{CRPS} when $Q=\mathbb{R}$. It can be seen as a special case of the weighted CRPS (Matheson and Winkler, 1976; Gneiting and Raftery, 2007; Gneiting and Ranjan, 2011), in which the indicator function $\mathbb{1}_Q$ plays the role of the weight function—in other words, the measure μ in (24) has density $\mathbb{1}_Q$ with respect to Lebesgue's measure.

Consider for instance the case $Q=\left(-\infty,t^{(0)}\right)$:

$$S_Q^{\mathrm{tCRPS}}(P,z)\;=\;\int_{-\infty}^{t^{(0)}}(F_P(u)-\mathbb{1}_{z\leq u})^2\,\mathrm{d}u.$$

The upper endpoint $t^{(0)}$ of the range will be referred to as the validation threshold. Note that, in this case, $S_Q^{\mathrm{tCRPS}}(P,z)$ does not depend on the specific value of z when z is above the validation threshold. This scoring rule is thus well suited to the problem of measuring the performance of a predictive distribution in such a way as to fully assess the goodness-of-fit of the distribution when the true value is below a threshold, and only ask that the support of the predictive distribution is concentrated above the threshold when the true value is above the threshold.

We provide in Appendix A some properties of the scoring rule (25) and closed-form expressions for the case where Q is an interval (or a finite union of intervals) and P is Gaussian. To the best of our knowledge, these expressions are new.

4.3 Choosing the Relaxation Range using the tCRPS Scoring Rule

Given a range of interest Q, the tCRPS scoring rule makes it possible to derive a goal-oriented leave-one-out selection criterion for the relaxation range R, which we call the LOO-tCRPS criterion:

$$J_n(R)=\frac{1}{n}\sum_{i=1}^{n}S_Q^{\mathrm{tCRPS}}\left(P_{R,n,-i},f(x_i)\right). \tag{26}$$

Using (26), we suggest the following procedure to select a reGP model. First, choose a sequence of nested candidate relaxation ranges $R^{(0)}\supset R^{(1)}\supset\cdots\supset R^{(G-1)}=\emptyset$. The next step is the computation of $J_n(R^{(g)})$, $g=0,\ldots,G-1$, which involves the predictive distributions $P_{R^{(g)},n,-i}$.

In principle, (14) should be solved again each time a data point (x_i,z_i) is removed, to obtain a pair $(\widehat{\theta}_{n,-i}^{(g)},\underline{z}_{n,-i}^{(g)})$ and then the corresponding reGP distribution $P_{R^{(g)},n,-i}$. To alleviate computational cost, a simple idea is to rely on the fast leave-one-out formulas (Dubrule, 1983) for Gaussian processes: for each set $R^{(g)}$, solve (14) to obtain $\widehat{\theta}_n^{(g)}$ and $\underline{z}_n^{(g)}=(z_1^{(g)},\ldots,z_n^{(g)})^{\mathsf{T}}$, and then compute the

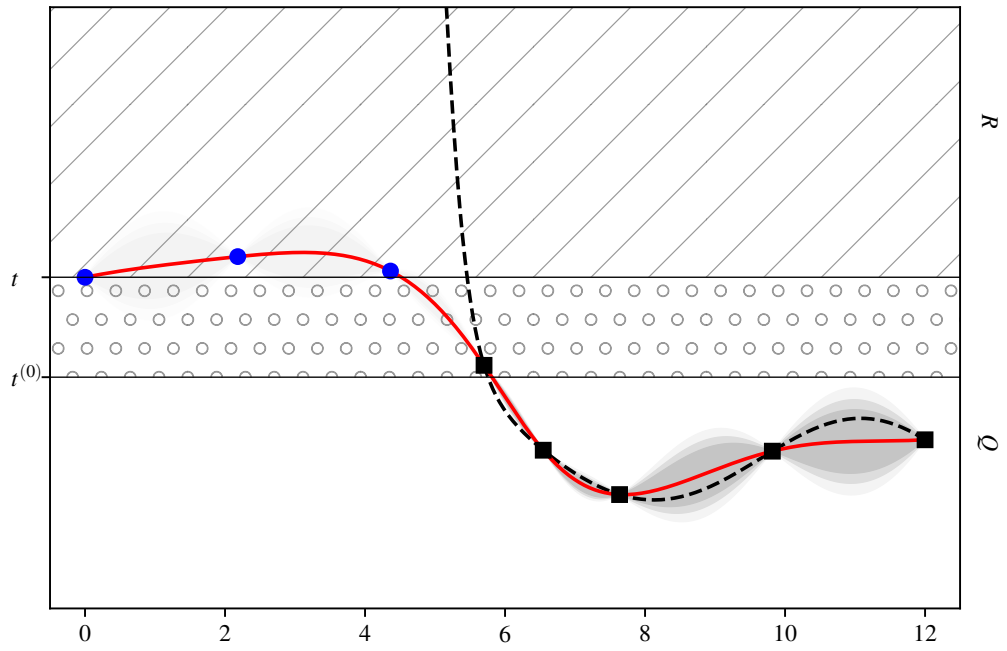

Figure 5: Illustration of the choice of a relaxation range. The range of interest Q is determined by the threshold $t^{(0)}$. The relaxation range R corresponding to the region above t has been obtained by the procedure described in Section 4.3.

conditional distributions $\xi(x_i) \mid \{\xi(x_j) = z_j^{(g)}, j \neq i\}$, where $\xi \sim \mathrm{GP}(\mu, k)$, and where μ and k have parameter $\widehat{\theta}_n^{(g)}$, using the fast leave-one-out formulas. By doing so, we neglect the difference between $\widehat{\theta}_{n,-i}^{(g)}$ and $\widehat{\theta}_n^{(g)}$ and the difference between $\underline{z}_{n,-i}^{(g)}$ and the vector $(z_1^{(g)}, \ldots, z_{i-1}^{(g)}, z_{i+1}^{(g)}, \ldots, z_n^{(g)})^\mathsf{T}$. The procedure ends by choosing the relaxation range $R^{(g)}$ that achieves the best LOO-tCRPS value.

Figure 5 illustrates the selection of the relaxation range used in Figure 3. Algorithm 1 summarizes the general reGP procedure using LOO-tCRPS for selecting relaxation range in a given list of candidates for relaxation range. Section 5.3.1 and Appendix F give specific implementations for Bayesian optimization and the problem of estimation of an excursion set.

Algorithm 1 reGP with automatic selection of the relaxation range.

Input: Data $(\underline{x}_n, \underline{z}_n)$; a range of interest Q; and a list $\mathbb{R} \setminus Q = R^{(0)} \supset \cdots \supset R^{(G-1)} = \emptyset$ of relaxation range candidates.
for $g = 0$ **to** $G - 1$ **do**
  Obtain $\widehat{\theta}_n^{(g)}$ and $\underline{z}_n^{(g)}$ by solving (14) with $R^{(g)}$
  Compute $J_n(R^{(g)})$ with Q, $\widehat{\theta}_n^{(g)}$, and $\underline{z}_n^{(g)}$ using (26)
end for
Output: The pair $\widehat{\theta}_n^{(g)}$, $\underline{z}_n^{(g)}$ that minimizes (26).

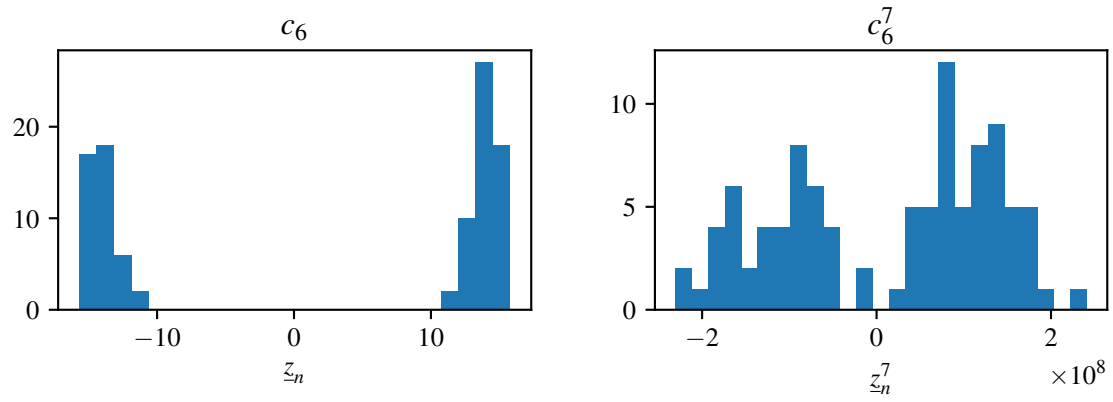

Figure 6: Left: Histogram of the values of the function c_6 from the G10 problem. Right: Same illustration but for the function c_6^α, with $\alpha = 7$. The histograms are obtained from the values of the functions on a space-filling design of size $n = 100$. On the left, the values are very separated and concentrated on two modes, yielding a function close to a piecewise constant function. After transformation, the phenomenon is mitigated.

4.4 An Example for the Estimation of an Excursion Set

We illustrate the method on the problem of estimating an excursion set $\{x \in \mathbb{X}, f(x) \leq 0\}$. We consider the G10 optimization problem used by Regis (2014), and focus on the constraint $c_6 \leq 0$. Finding solutions satisfying the $c_6 \leq 0$ constraint using a GP model is difficult, probably because the values of c_6 are very bi-modal, as illustrated in Figure 6. However, Feliot et al. (2017) found that the difficulty could be overcome by performing an ad-hoc monotonic transformation $z \mapsto z^\alpha$, with $\alpha = 7$, on the constraint.

The estimation of an excursion set $\{f \leq 0\}$ involves capturing precisely the behavior of f around zero. Thus, we define a range of interest $Q = (-t^{(0)}, t^{(0)})$ centered on zero, with $t^{(0)}$ sufficiently small (note that there may be no data in Q). Then, we consider relaxation range candidates $R^{(g)} = (-\infty, -t^{(g)}] \cup [t^{(g)}, +\infty)$ with a sequence of thresholds $t^{(0)} < \cdots < t^{(G-1)} = +\infty$, and we select $t^{(g)}$ by minimizing the LOO-tCRPS as described in the previous section.

Appendix F presents numerical experiments on active learning of the set $\{x \in \mathbb{X}, c_6(x) \leq 0\}$ with reGP. Figure 7 illustrates the relaxation obtained using a value of $t^{(0)}$ such that Q contains 25% of the observations of a design of size $n = 300$ in the experiments presented in Appendix F. Here, the LOO-tCRPS chooses $t^{(g)} = t^{(0)}$. Observe that the transformation after relaxation resembles the transformation $z \mapsto z^\alpha$ proposed by Feliot et al. (2017). Applying the reGP framework on the transformed function c_6^α (details omitted for brevity), we find that the LOO-tCRPS chooses a large $t^{(g)}$ such that the interpolation constraints are relaxed for only a few observations.

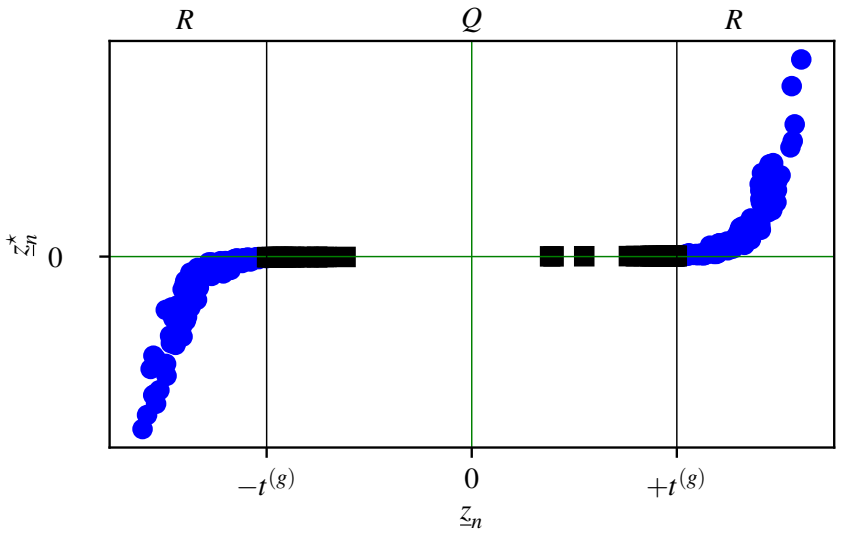

Figure 7: A reGP fit of c_6, where the relaxation thresholds have been selected by LOO-tCRPS. The observations z_n are shown on the x-axis, whereas the "relaxed" observations $z_n^\star$ are represented on the y-axis. The black squares represent interpolated observations and the blue points show relaxed observations. Moreover, the green lines represent the value zero, and the black lines represent $\pm\, t^{(g)}$, with $t^{(g)}$ chosen to be $t^{(0)}$ by the LOO-tCRPS.

5. Application to Bayesian Optimization

5.1 Efficient Global Optimization with Relaxation

The first motivation for introducing reGP models is Bayesian optimization, where obtaining good predictive distributions over ranges corresponding to optimal values is a key issue. In this article, we focus more specifically on the minimization problem

$$\min_{x\in\mathbb{X}} f(x)\,,$$

where f is a real-valued function defined on a compact set $\mathbb{X}\subset\mathbb{R}^d$, but the methodology can be generalized to constrained and/or multi-objective formulations.

Given f, our objective is to construct a sequence of evaluation points $x_1, x_2 \ldots \in \mathbb{X}$ by choosing each point x_{n+1} as the maximizer of the expected improvement criterion (5) computed with respect to the reGP predictive distribution, with a relaxation range $R_n = [t_n, +\infty)$. More precisely, the sequence (x_n) is constructed sequentially using the rule

$$x_{n+1} \;=\; \operatorname*{arg\,max}_{x\in\mathbb{X}} \; \mathsf{E}_n\left((m_n - \xi(x))_+\right)\,, \tag{27}$$

where $m_n = f(x_1)\wedge\cdots\wedge f(x_n)$, and E_n is the expectation under the reGP predictive distribution with relaxation range R_n and data $\underline{z}_n = (f(x_1),\ldots,f(x_n))^{\mathsf{T}}$.

As in Section 4.3, the relaxation threshold t_n at iteration n is chosen using the LOO-tCRPS criterion (26) among candidate values

$$t_n^{(0)} < t_n^{(1)} < \cdots < t_n^{(G-1)}\,, \tag{28}$$

where $t_n^{(0)}$ is the validation threshold, which delimits the range of interest $Q_n = \left(-\infty, t_n^{(0)}\right)$ used at iteration n. In the following, the optimization method just described will be called efficient global optimization with relaxation (EGO-R), in reference to the EGO name proposed by Jones et al. (1998).

Implementation specifics are detailed in Section 5.3. In the next section, we show that using the EI criterion with a reGP model yields a convergent algorithm.

5.2 Convergence of EGO-R with Fixed Parameters and Varying Threshold

In this section, we extend the result of Vazquez and Bect (2010b) and show the convergence of the EGO-R algorithm, in the case where the predictive distributions derive from a zero-mean Gaussian process with fixed covariance function.

We suppose that $\mathbb{X}\subset\mathbb{R}^d$ is a compact domain and that $k:\mathbb{X}\times\mathbb{X}\to\mathbb{R}$ is continuous, strictly positive-definite, and has the NEB (no-empty ball) property (Vazquez and Bect, 2010b), which says that the posterior variance cannot go to zero at a given point if there is no evaluation points in a ball centered on this point. In other words, the NEB property requires that the posterior variance $\sigma_n^2(x)$ at $x\in\mathbb{X}$ remains bounded away from zero for any x not in the closure of the sequence of points (x_n) evaluated by the optimization algorithm. A stationary covariance function with smoothness $\alpha > 0$ satisfies the NEB property (Vazquez and Bect, 2010b), whereas the squared-exponential covariance function does not (Vazquez and Bect, 2010a).

Proposition 18 *Let $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a continuous strictly positive-definite covariance function that satisfies the NEB property, $\mathscr{H}(\mathbb{X})$ the corresponding RKHS and $f \in \mathscr{H}(\mathbb{X})$. Let $n_0 > 0$. Let $(x_n)_{n\geq 1}$ be a sequence in $\mathbb{X}$ such that, for each $n \geq n_0$, x_{n+1} is obtained by* (27) *with $t_n > m_n$. Then the sequence $(x_n)_{n\geq 1}$ is dense in $\mathbb{X}$.*

Proposition 18 implies the convergence of EGO-R with a fixed threshold $t > \min_{i\leq n_0} f(x_i)$. In this case, the theoretical insights from Section 3.3 suggest a faster convergence might be achieved due to the improved error estimates (20) and (21) in a neighborhood of the global minimum.

The convergence of EGO-R also holds in the case of a varying relaxation range $R_n = [t_n, +\infty)$, with $t_n > m_n$, and in particular when t_n is selected at each step using the LOO-tCRPS criterion (26) with a validation threshold $t_n^{(0)} > m_n$. In this case, the norm term in (20) gets smaller if $(t_n)_{n\geq 1}$ is decreasing.

5.3 Optimization Benchmark

In this section, we run numerical experiments to demonstrate the interest of using EGO-R instead of EGO for minimization problems. Appendix E presents a similar study based on the UCB sampling criterion (Cox and John, 1992; Srinivas et al., 2010). Appendix F considers the task of estimating an excursion set.

5.3.1 METHODOLOGY

In practice, we must choose the sequence of thresholds (28). The validation threshold $t_n^{(0)}$ should be set above m_n to ensure there is enough data to carry out the validation. We propose three different heuristics to this end: a) a *constant* heuristic, where $t_n^{(0)}$ remains constant throughout the iterations and is set to an empirical quantile of an initial data set constructed prior to running EGO-R, b) a *concentration* heuristic, where $t_n^{(0)}$ corresponds to an empirical quantile of $\underline{z}_n$, and c) a *spatial* heuristic, where $t_n^{(0)}$ corresponds to a spatial quantile estimate of the function f. By spatial quantile, we mean a quantile of $f(U)$ with U uniformly distributed over the domain $\mathbb{X}$.

In the case of the constant heuristic, we set $t_n^{(0)}$ to the α-quantile of the function values on an initial design, which is typically built to cover $\mathbb{X}$ as uniformly as possible using, for example, maximin Latin hypercube sampling (McKay et al., 2000). In this article, the numerical experiments were conducted with $\alpha = 0.25$.

For the spatial heuristic, the spatial quantile is estimated by building a one-nearest-neighbor regression model trained on the data $(\underline{x}_n, \underline{z}_n)$, and then taking an α-quantile of its predictions on a uniform sample on $\mathbb{X}$. The numerical value $\alpha = 0.25$ is also used here. This can be viewed as a refinement of the constant heuristic.

For the concentration heuristic, we consider the α-quantile of the values of f at the points visited by the algorithm (again with $\alpha = 0.25$). As the optimization algorithm progresses, the evaluations are expected to cluster around the global minimum. Thus, $t_n^{(0)}$ will approach the minimum value, and the validation range $Q_n = (-\infty, t_n^{(0)})$ will shrink. As a result, since better predictive distributions are expected in this range, a better convergence may be achieved.

All three heuristics are supported by the idealized convergence result from the previous section. Proposing alternative heuristics, or conducting a theoretical comparison of their performances, is out of the scope of this article.

Table 1: Optimization benchmark.

Problem	d
Branin	2
Six-hump Camel	2
Three-hump Camel	2
Hartman	3, 6
Ackley	4, 6, 10
Rosenbrock	4, 6, 10
Shekel	5, 7, 10
Goldstein-Price	2
Log-Goldstein-Price	2
Cross-in-Tray	2
Beale	2
Dixon-Price	4, 6, 10
Perm	4, 6, 10
Michalewicz	4, 6, 10
Zakharov	4, 6, 10

For a given $t_n^{(0)}$, the candidate relaxation thresholds $t_n^{(g)}$, $g = 0, \ldots, G-1$, are chosen such that $t_n^{(g)} - m_n$ ranges logarithmically from $t_n^{(0)} - m_n$ to $\max f(x_i) - m_n$ (with $G = 10$ in the experiments below). We adopt the convention that $t_n^{(G-1)} > \max f(x_i)$, so that the G-th model is a standard GP.

To assess the performances of EGO-R with these three heuristics, we compare them to the standard EGO algorithm. For all four algorithms, we use an initial design of size $n_0 = 10d$, and we consider GPs with a constant mean function and a Matérn covariance function with regularity $\nu = 5/2$. The maximization of the sampling criteria (5) and (27) is performed using a sequential Monte Carlo approach (Benassi et al., 2012; Feliot et al., 2017).

The optimization algorithms are tested against a benchmark of test functions from Surjanovic and Bingham (2013) summarized in Table 1, with $n_{\text{rep}} = 100$ (random) repetitions, and a budget of $n_{\text{tot}} = 300$ evaluations for each repetition. This benchmark is partly inspired by Jones et al. (1998) and Merrill et al. (2021). In particular, we also use a log-version of the Goldstein-Price function as Jones et al. (1998).

To evaluate the algorithms, we construct for each test function a list of target values, defined as spatial quantiles of the function (estimated with a subset simulation algorithm; see, e.g., Au and Beck, 2001). The probability levels associated with this list of spatial quantiles define a non-linear scale that we use instead of raw function values. We plot in this scale the 10%/50%/90% quantiles of the best function value found, as a function of n. We also represent, for each probabilty level, the fraction of runs that reach the corresponding target.

5.3.2 FINDINGS

The full set of results is provided in Appendix D. In Figure 8, we present a representative subset of these results.

First, observe in Figure 8 that the EGO-R methods outperform EGO significantly on functions that are difficult to model with stationary GPs, such as Goldstein-Price, Perm (10), and Beale.

In these cases, relaxing the constraints for the highest observations proves beneficial. Notice also that, among the three EGO-R variants, the one based on the concentration heuristic converges faster in such situations. (Approximately half of the cases of our benchmark fall into this category; see Appendix D for details.)

In the other cases (Log-Goldstein-Price, Ackley (4) and Ackley (10) on Figure 8) the EGO-R methods perform similarly to EGO. This corresponds to functions where classical stationary GP modeling already provides good predictions. In these cases, the LOO-tCRPS criterion for selecting the relaxation range detects that the largest values also contribute to predicting near the minimum, resulting in little to no relaxation being necessary. The concentration heuristic sometimes provides a small advantage on such functions (see Log-Goldstein-Price and Ackley (4)), but occasionally degrades performances compared to EGO and the other EGO-R variants, as seen with the Ackley (10) function. A closer examination shows that the concentration heuristic sometimes gets temporarily trapped in a local minimum. We explain this by the fact that the reGP model with the concentration heuristic can become overly predictive in a small region around the local minimum, but underestimate the function variations elsewhere (the variance of the predictive distributions above $t_n^{(0)}$ is too low, and the optimization algorithm does not explore unknown regions sufficiently). In this sense, the constant and spatial heuristics can be considered more conservative than the concentration heuristic.

Finally, it is useful to compare the performance of the EGO-R algorithms on the Goldstein-Price function with the performance of the EGO algorithm on the Log-Goldstein-Price function. Using reGP modeling allows for performance comparable to that achieved with a logarithmic transformation, but in an automatic manner. This is illustrated in Figure 9, where the (non-parametric) transform learned by reGP resembles a logarithmic transformation.

5.3.3 RUNNING TIME OF REGP

Training a reGP model takes longer than training a standard GP for several reasons. First, the optimization problem in (14) is an extension of (4) and involves a higher dimension of the search space, as the number of relaxed observations increases, which likely increases the time required to solve it. Second, the automatic selection procedure described in Section 4 involves testing G relaxation ranges, making reGP at least G times more costly to compute.

For $G = 10$, Figure 10 provides a quantitative assessment of the increase in computational time as a function of the number of evaluations n, across four representative test cases. With automatic selection of the relaxation range, training a reGP model is, roughly speaking, between $G = 10$ and 1000 times more expensive than training a standard GP. This trend holds for the other test functions in Table 1 (figures omitted for brevity). Among the heuristics, the concentration heuristic is the most computationally expensive version of reGP in this benchmark. This is because the number of observations outside the range of interest—and thus the maximum number of relaxed observations—grows linearly with n. In contrast, for the spatial and constant heuristics, the computational overhead remains around a factor of 10, which suggests that the primary cost increase comes from having to test G relaxation ranges.

Note that the procedure described in Section 4.3 involves an exhaustive search over the set of relaxation range candidates. Assessing the performance of reGP with a smaller set of candidates, or using adaptive strategies such as dichotomies, could help reduce computation time. Note also that a factor G in computation time can be saved if the user already knows how to choose a relaxation

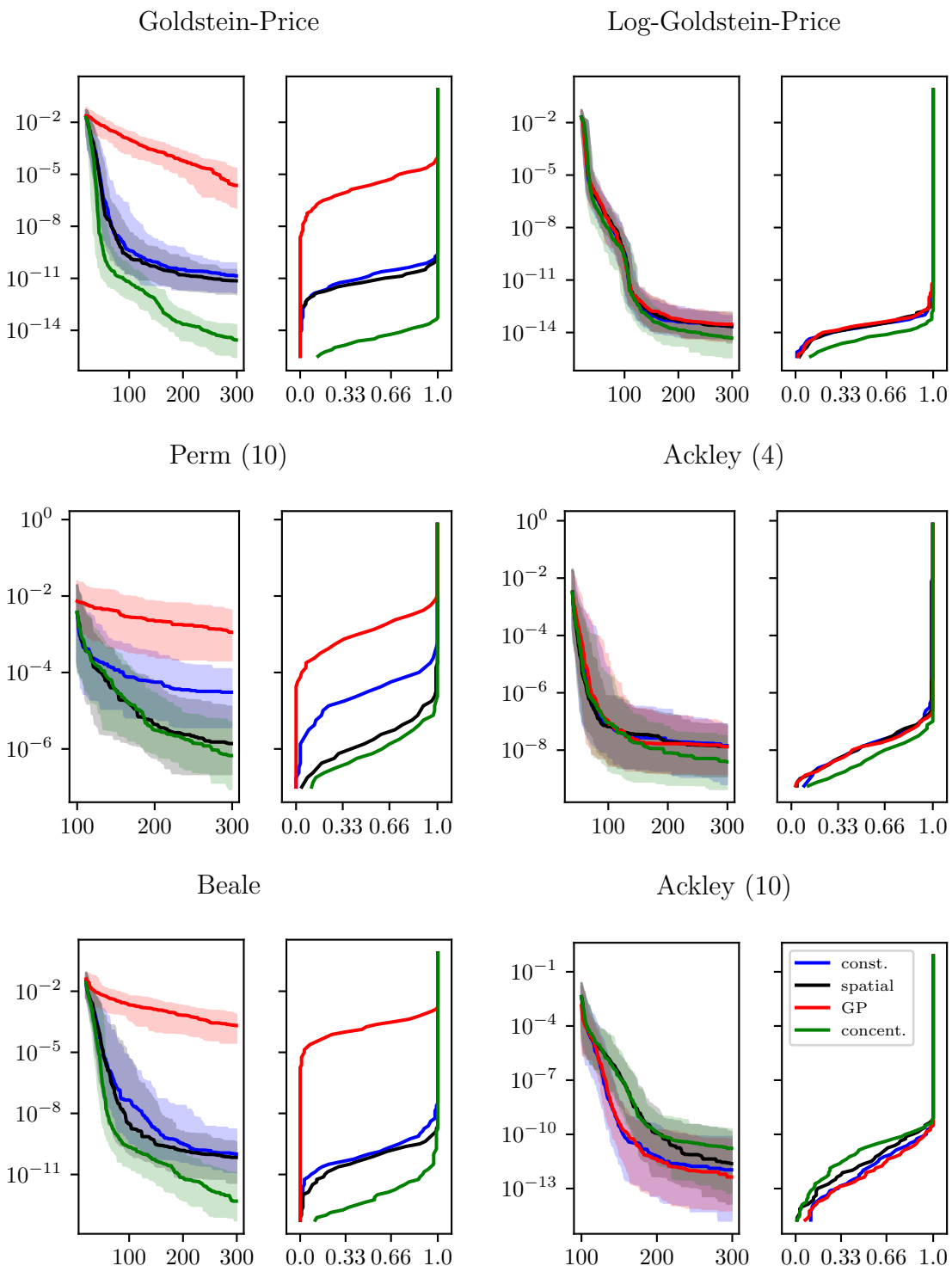

Figure 8: For each case, the left plot shows the evolution of quantiles of the best function value against n, and the right plot represents on the x-axis the fraction of runs that reach each target value. Both plots use the probability level of the target as a scale for the y-axis. Left plot: solid lines represent the median; shaded areas are deliminated by the 10% and 90% quantiles. Both plots: EGO (red), EGO-R + constant heuristic (blue), EGO-R + concentration heuristic (green), EGO-R + spatial heuristic (black).

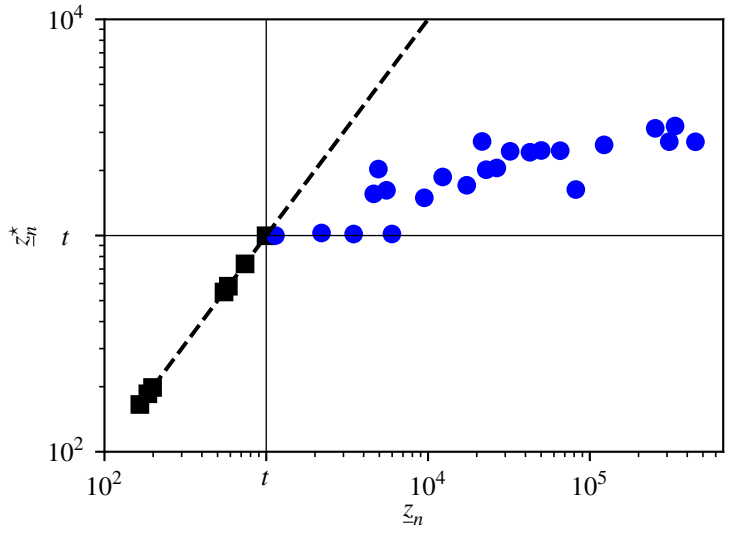

Figure 9: A reGP fit to the Goldstein-Price function with $n = 30$ points, with $R = \left[10^3, +\infty\right)$. The observations z_n are shown on the x-axis, whereas the relaxed observations $z_n^\star$ are represented on the y-axis.

range for his task. Another approach, left for future work, would be to solve (14) with dedicated algorithms (see Remark 8).

Finally, it should be stressed that the computation time is less of a concern for active learning tasks, since evaluating the objective function—which can take several hours—is usually the time bottleneck in this setting.

6. Conclusion

This article presents a new technique called reGP to build predictive distributions for a function observed on a sequence of points. This technique can be applied when a user wants good predictive distributions in a range of function values, for example below a given threshold, and accepts degraded predictions outside this range. The technique relies on Gaussian process interpolation, and operates by relaxing interpolation constraints outside the range of interest. The relaxation range can be selected automatically, using a scoring rule adapted to reGP models. With respect to classical GP modeling, this goal-oriented technique simply involves the choice of one additional key parameter, which is the range of interest—i.e., the user only needs to specify a range of function values where good predictions should be obtained.

Such goal-oriented models can then be used in Bayesian sequential search algorithms. Here we are more specifically interested in the problem of mono-objective optimization and we propose to study the EI / EGO algorithm with such models. We provide theoretical guarantees for the convergence of the reGP-based algorithm on the RKHS attached to the underlying GP covariance,

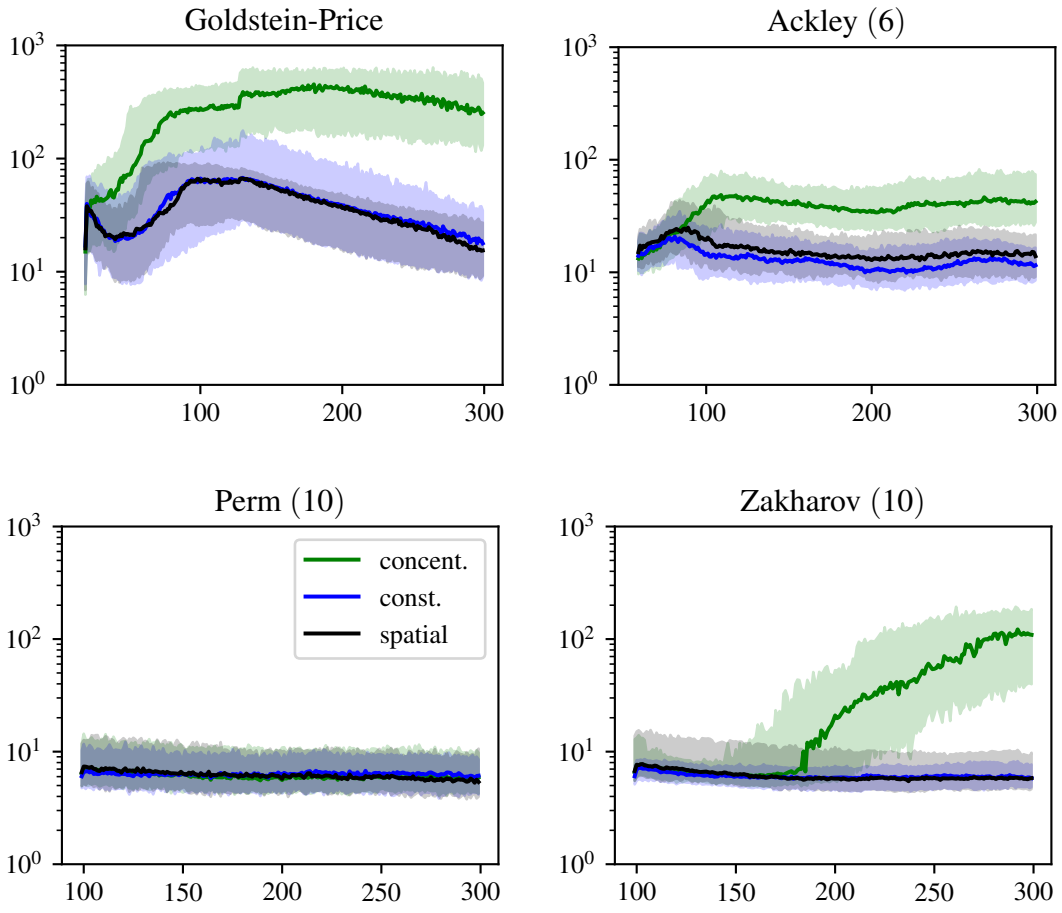

Figure 10: Ratio between the time needed to train a reGP model (including the automatic selection of the relaxation range) and that required to train a GP during Bayesian optimization with the EI criterion, across four representative test cases, as a function of n. The curves show median ratios, while the shaded areas represent intervals bounded by the 10% and 90% quantiles. Blue, green, and black represent reGP using the constant, concentration, and spatial heuristics, respectively.

and a numerical benchmark that shows very clear benefits of using reGP models for the optimization of various functions.

A key element of the reGP approach is the definition of the range of interest, which is for instance of the form $\left(-\infty, t^{(0)}\right)$ in a minimization problem. In some use cases the range will be provided by the user, but in others it is desirable to set it automatically. Three simple heuristics have been proposed to achieve this goal in our optimization benchmark, and it has been observed that the choice of heuristic has an impact on the exploratory behaviour of the resulting Bayesian optimization algorithm. Finding better heuristics, studying their properties, and assessing their impact in Bayesian optimization applications, is an important direction for future research.

More generally, the goal-oriented approach proposed in this article is not limited to single-objective (Bayesian) optimization. The example from Section 4.4 and the benchmark results of Appendix F show that it is also readily applicable, for instance, to level set estimation problems. Appendix G provides a proof of concept for extending the approach to noisy observations. A thorough investigation of the best modeling choices and theoretical guarantees is left for future work. Other extensions are possible but will require more work. Constrained and/or multi-objective optimization is another interesting but challenging direction for future research: in this case the function of interest is multivariate (one objective and several constraints, or several objectives), which requires significant adaptations to the proposed methodology.

Another possible direction for future research would be to infer the unknown function using the full predictive distribution (12), of which reGP is a mode approximation. This distribution defines a Skew GP, for which various inference methods have been proposed in the literature, using for instance quasi Monte-Carlo sampling, linear elliptical slice sampling or Gibbs sampling (see, e.g., Da Veiga and Marrel, 2012; Botev, 2017; Benavoli et al., 2021; Gessner et al., 2020; Takeno et al., 2023). Adapting these ideas to the reGP setting may be relevant in a context where uncertainty quantification is critical.

Acknowledgments

The authors thank the action editor and three anonymous reviewers for their insightful comments, which significantly improved both the content and the presentation of the article. This work was partially supported by the French Agence Nationale de la Recherche et de la Technologie (ANRT) through a CIFRE grant.

Appendix A. Properties of the Truncated CRPS

We shall now write (25) more explicitly for the case where the range of interest is an interval $Q = (a,b)$, $-\infty \leq a < b \leq +\infty$, and provide closed-form expressions for the case where, in addition, the predictive distribution P is Gaussian.

Remark 19 *The value of the tCRPS for an interval $Q = (a,b)$ remains unchanged if the interval is closed at one or both of its endpoints.*

Remark 20 *The value of the tCRPS for a finite (or countable) union of disjoint intervals follows readily from its values on intervals, since $Q \mapsto S_Q^{\mathrm{tCRPS}}(P,z)$ is σ-additive.*

We shall start by defining a quantity that shares similarities with (5).

Definition 21 $\mathrm{EI}_q^{\uparrow}(P,z) \;=\; \mathsf{E}\left((N_1 \vee \cdots \vee N_q - z)_+\right)$ *with* $N_j \overset{\text{iid}}{\sim} P$.

The following expressions hold for a general predictive distribution P.

Proposition 22 *Suppose that P has a first order moment.*

- *Let $a,b \in \mathbb{R}$ with $a \leq b$. Then,*

$$\begin{aligned} S_{a,b}^{\mathrm{tCRPS}}(P,z) \;=\; & (b \wedge z - a)_+ + \mathrm{EI}_2^{\uparrow}(P,b) - \mathrm{EI}_2^{\uparrow}(P,a) \\ & - 2\,\mathbb{1}_{z \leq b}\left(\mathrm{EI}_1^{\uparrow}(P,b) - \mathrm{EI}_1^{\uparrow}(P,a \vee z)\right). \end{aligned}$$

- *Let $b \in \mathbb{R}$ and $N_1, N_2 \overset{\text{iid}}{\sim} P$. Then,*

$$\begin{aligned} S_{-\infty,b}^{\mathrm{tCRPS}}(P,z) \;=\; & b \wedge z + \mathrm{EI}_2^{\uparrow}(P,b) - \mathsf{E}\left(N_1 \vee N_2\right) \\ & - 2\,\mathbb{1}_{z \leq b}\left(\mathrm{EI}_1^{\uparrow}(P,b) - \mathrm{EI}_1^{\uparrow}(P,z)\right). \end{aligned}$$

- *Finally, if $a \in \mathbb{R}$, then*

$$S_{a,+\infty}^{\mathrm{tCRPS}}(P,z) \;=\; S_{-\infty,-a}^{\mathrm{tCRPS}}(\underline{P},-z),$$

where $\underline{P}$ is the distribution of $-U$ if U is P-distributed.

Now, leveraging well-known analytic expressions (see, e.g., Nadarajah and Kotz, 2008; Chevalier and Ginsbourger, 2013), we have the following closed-form expressions in the Gaussian case.

Proposition 23 *(Nadarajah and Kotz, 2008; Chevalier and Ginsbourger, 2013)*
Suppose that $P = \mathcal{N}(\mu,\sigma^2)$ and let ϕ and Φ denote respectively the pdf and the cdf of the standard Gaussian distribution. Then

- $\mathrm{EI}_1^{\uparrow}(P,z) = \sigma\, h_1\left(\frac{\mu - z}{\sigma}\right)$, *with*

$$h_1(t) = t\,\Phi(t) + \phi(t),$$

- *for $q \geq 2$, we have* $\mathrm{EI}_q^{\uparrow}(P,z) = \sigma\, h_q\left(\frac{\mu - z}{\sigma}\right)$, *where*

$$h_q(t) \;=\; qt\,\Phi_q(\delta_q^t; 0, D_q D_q^{\mathsf{T}}) + q\,\Phi(-t)^{q-1}\,\phi(t) + \frac{q(q-1)}{2\sqrt{\pi}}\,\Phi_{q-1}(\delta_{q-1}^t; 0, \tfrac{1}{2}B_q),$$

where $\Phi_q(\cdot; m, \Sigma)$ is the cdf of the multivariate $\mathcal{N}(m,\Sigma)$ distribution,

$$B_q = 2\,\mathrm{diag}(0, \mathbb{1}_{q-2}^{\mathsf{T}}) + \mathbb{1}_{q-1}\mathbb{1}_{q-1}^{\mathsf{T}},$$

D_q is the matrix representing the linear map

$$\mathbb{R}^q \to \mathbb{R}^q, \quad (y_1, \ldots, y_q)^{\mathsf{T}} \mapsto (-y_1, y_2 - y_1, y_3 - y_1, \ldots, y_q - y_1)^{\mathsf{T}},$$

and $\delta_q^t = (t, 0_{q-1}^{\mathsf{T}})^{\mathsf{T}}$,

- *finally for $q = 2$ we have*

$$\mathsf{E}\,(N_1 \vee N_2) = \mu + \frac{\sigma}{\sqrt{\pi}}.$$

The propriety of scoring rules is an important notion that formalizes "well-calibration" in the sense that a generating distribution must be identified to be optimal on average.

Definition 24 *(see, e.g., Gneiting and Raftery, 2007) A scoring rule $S : \mathscr{P} \times \mathbb{R} \to \mathbb{R}$ is said to be (strictly) proper with respect to $\mathscr{P}$ if, for all $P_1, P_2 \in \mathscr{P}$, the mapping $y \in \mathbb{R} \mapsto S(P_1, y)$ is P_2-quasi-integrable and the mapping*

$$P_1 \in \mathscr{P} \mapsto S(P_1, P_2) := \mathsf{E}_{U \sim P_2}\,(S(P_1, U))$$

admits P_2 as a (unique) minimizer.

In the case of the truncated CRPS, simple calculations lead to (Matheson and Winkler, 1976):

$$S_Q^{\mathrm{tCRPS}}(P_1, P_2) \;=\; S_Q^{\mathrm{tCRPS}}(P_2, P_2) + \int_Q \left(F_{P_1}(u) - F_{P_2}(u)\right)^2 \mathrm{d}u.$$

It follows that S_Q^{tCRPS} is proper with respect to the class of all Borel probability measures on $\mathbb{R}$ for any measurable $Q \subset \mathbb{R}$, and is strictly proper with respect to the class of non-degenerate Gaussian measures on $\mathbb{R}$ as soon as Q has non-empty interior.

Appendix B. Theoretical Insights on reGP

This appendix provides further theoretical study of relaxed interpolation. We start by recalling a well-known approximation result for the interpolation of classes of functions outside the RKHS. Then, we give a corresponding adaptation for relaxed interpolation. Finally, Appendix B.3 presents results on approximation and uncertainty quantification for reGP with estimated parameters. The notations are that of Section 3.3.

B.1 Escape Theorem for Gaussian Process Interpolation

In GP interpolation, error bounds are sometimes available for interpolation of functions not belonging to the RKHS. Let $q_n = \min_{i\neq j}\|x_i - x_j\|/2$ be the separation distance and $\rho_n = h_n/q_n$ be the mesh ratio. A sequence $(x_n)_{n\geq 1}$ is quasi-uniform if there exists a positive constant C_{qu} such that $q_n \leq h_n \leq C_{\mathrm{qu}} q_n$. The following proposition is a reminder of the well-known "escape" theorem, which provides error bounds for standard Gaussian process interpolation of classes of functions outside the RKHS attached to a kernel with regularity $\alpha > 0$ (see Theorem 4.2 of Narcowich et al., 2006, and Theorem 4.2 of Karvonen et al., 2020, and Theorem 1 of Wynne et al., 2021, for recent extensions).

Proposition 25 *Let $0 < \beta < \alpha < +\infty$ be such that $f \in W_2^{\beta+d/2}(\mathbb{X})$ and assume that $\mathbb{X}$ satisfies Assumption 9. If k has regularity α, then*

$$\|f - s_{z_n}\|_{L^\infty(\mathbb{X})} \lesssim h_n^\beta (1 + \rho_n^{\alpha-\beta}) \|f\|_{W_2^{\beta+d/2}(\mathbb{X})},$$

where $\lesssim$ denotes inequality up to a constant, that does not depend on f, n or (x_n). In particular, if $(x_n)_{n\geq 1}$ is quasi-uniform, we have

$$\|f - s_{z_n}\|_{L^\infty(\mathbb{X})} \lesssim n^{-\beta/d} \|f\|_{W_2^{\beta+d/2}(\mathbb{X})}.$$

B.2 Convergence of Relaxed Interpolation of Functions Outside the RKHS

Section 3.3.2 only considers functions belonging to the RKHS. There are other situations in which reGP is expected to offer advantages. For the remainder of this appendix, assume $f\colon \mathbb{X} \to \mathbb{R}$ is an arbitrary function. For a relaxation range $R = \bigcup_{j=1}^q R_j$, the subsets $\mathbb{X}_j \subset \mathbb{X}$ are defined as in Section 3.3.2.

Definition 26 *Let $\mathscr{F}$ be a set of functions $\mathbb{X} \to \mathbb{R}$. We say that f is an R-element of $\mathscr{F}$ if there exists a function $g \in \mathscr{F}$ such that g and f coincide on $\mathbb{X}_0$ and $f(x) \in R_j \Leftrightarrow g(x) \in R_j$, for all $j \geq 1$ and $x \in \mathbb{X}$.*

Naturally, if $f \in \mathscr{F}$, then f is an R-element of $\mathscr{F}$. Note also that f can be an R-element of a space of continuous functions and be discontinuous in the interior of an $\mathbb{X}_j$, for $j \geq 1$. When $\mathscr{F} = \mathscr{H}(\mathbb{X})$, this condition is equivalent to requiring that the set $\mathscr{H}_{R,\mathbb{X}}$, defined by (18), is non-empty.

Suppose $f \notin \mathscr{H}(\mathbb{X})$ but f is an R-element of $\mathscr{H}(\mathbb{X})$. In this case, the error bounds (8) and (17) are invalid, whereas the conclusions of Propositions 12–14 still hold. (Indeed, observe that the reGP predictor of f is that of a function g given by Definition 26. The propositions apply to g.)

Moreover, even if f is not an R-element of $\mathscr{H}(\mathbb{X})$, error bounds can sometimes be stated on $\mathbb{X}_0$. Recall that the space $W_2^{\beta+d/2}(\mathbb{X})$ is an RKHS with a continuous kernel for $\beta > 0$. If f is

an R-element of $W_2^{\beta+d/2}(\mathbb{X})$, then Proposition 12 ensures the existence of a minimum $W_2^{\beta+d/2}(\mathbb{X})$-norm element $s_{R,\mathbb{X}}^{(\beta)}$ of the set of functions $g \in W_2^{\beta+d/2}(\mathbb{X})$ coinciding with f on $\mathbb{X}_0$ and such that $f(x) \in R_j \Leftrightarrow g(x) \in R_j$, for all $j \geq 1$ and $x \in \mathbb{X}$. The following proposition is an adaptation of Proposition 25 for bounding the error of the reGP predictor on $\mathbb{X}_0$. It relies on membership in the sense of Definition 26.

Proposition 27 *Suppose that $\mathbb{X}$ is bounded and measurable. Let $0 < \beta < \alpha < +\infty$ be such that f is an R-element of $W_2^{\beta+d/2}(\mathbb{X})$ and $B \subset \mathbb{X}_0$ satisfy Assumption 9. If k has regularity α, then it holds that:*

$$\|f - s_{R,n}\|_{L^\infty(B)} \lesssim h_n^\beta (1 + \rho_n^{\alpha-\beta}) \|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})}, \tag{29}$$

where $\lesssim$ denotes inequality up to a constant, that does not depend on f, n or (x_n). In particular, if $(x_n)_{n\geq 1}$ is quasi-uniform, then it holds that:

$$\|f - s_{R,n}\|_{L^\infty(B)} \lesssim n^{-\beta/d} \|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})}. \tag{30}$$

Contrasting Proposition 27 with Proposition 25 suggests an improved fit with reGP in $\mathbb{X}_0$ for two reasons. First, Proposition 27 applies to a larger class of functions since f can be an R-element of $W_2^{\beta+d/2}(\mathbb{X})$ without lying in $W_2^{\beta+d/2}(\mathbb{X})$. Furthermore, Proposition 15 shows that $\|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})}$ can be arbitrarily smaller than $\|f\|_{W_2^{\beta+d/2}(\mathbb{X})}$ when $f \in W_2^{\beta+d/2}(\mathbb{X})$.

B.3 reGP with Estimated Parameters

The reGP predictive distribution is obtained by conditioning the GP on the relaxed observations and the model parameters given by (14). The results from Section 3.3.2 do not take parameter selection into account. This appendix presents adaptations of recent theoretical results about parameter selection of Matérn covariance functions for Gaussian process interpolation of deterministic functions.

We still assume a zero mean GP and use the notations from Section 3.3.2, but in this appendix, we consider the RKHS $\mathscr{H}(\mathbb{X})$ with reproducing kernel r, not k, and the posterior variance will be $\tau^2\sigma_n^2(\cdot)$, not $\sigma_n^2(\cdot)$ (recall from Section 2.1 that the covariance function is $k(\cdot,\cdot) = \tau^2 r(\cdot - \cdot)$).

First, we consider the maximum likelihood estimation of the variance parameter τ^2. Suppose that r is a Matérn correlation function with known regularity parameter ν and range parameters $(\rho_1,\ldots,\rho_d)$. Using the standard Gaussian process model, and writing $z_n = (f(x_1),\ldots,f(x_n))^\mathsf{T}$, it is well-known that the maximum likelihood estimate of τ^2 is given by:

$$\widehat{\tau}_n^2 = \frac{\|s_{z_n}\|_{\mathscr{H}(\mathbb{X})}^2}{n}.$$

The estimated posterior variance at $x \in \mathbb{X}$ is $\widehat{\tau}_n^2\sigma_n^2(x)$. Suppose that $(x_n)_{n\geq 1}$ is dense in $\mathbb{X}$. It holds that $\|s_{z_n}\|_{\mathscr{H}(\mathbb{X})} \to \|f\|_{\mathscr{H}(\mathbb{X})}$ if $f \in \mathscr{H}(\mathbb{X})$ (see, e.g., Iske, 2018, Theorem 8.37). Therefore, if $f \in \mathscr{H}(\mathbb{X})$ and $x \in \mathbb{X}$, then (8) implies that:

$$\frac{|f(x) - s_{z_n}(x)|}{\widehat{\tau}_n \sigma_n(x)} \leq C\sqrt{n}$$

for some $C > 0$ and n large enough and using the convention $0/0 = 1$. In other words, the GP can be overconfident by a factor of magnitude at most $\sqrt{n}$ (Karvonen et al., 2020; Karvonen, 2022).

If $f \notin \mathscr{H}(\mathbb{X})$, then (8) is not valid and one can typically expect larger errors. Karvonen et al. (2020) study how $\|s_{z_n}\|_{\mathscr{H}(\mathbb{X})}$ diverges to compensate in this case.

Returning to relaxed Gaussian process interpolation with relaxation range R, notice that the likelihood can be optimized with respect to the relaxed observation independently of τ^2. Then, given the optimal relaxed observations $\underline{z}_n^\star$, the reGP variance estimate writes

$$\widehat{\tau}^2_{n,R} = \frac{\|s_{R,n}\|^2_{\mathscr{H}(\mathbb{X})}}{n},$$

where $s_{R,n} = s_{\underline{z}_n^\star}$ is the relaxed predictor. Using (7) and (10), it is easy to show that $\widehat{\tau}^2_{n,R} \leq \widehat{\tau}^2_n$. The following proposition shows that the situations where $\widehat{\tau}^2_{n,R} = \mathscr{O}(n^{-1})$ are characterized by membership of f in $\mathscr{H}(\mathbb{X})$ in the sense of Definition 26.

Proposition 28 *Suppose that f is an R-element of the space of continuous functions from $\mathbb{X}$ to $\mathbb{R}$ and that $(x_n)_{n\geq 1}$ is dense in $\mathbb{X}$. Using the notations of Proposition 12, it holds that $\|s_{R,n}\|_{\mathscr{H}(\mathbb{X})} \to \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})}$ if f is an R-element of $\mathscr{H}(\mathbb{X})$ and $\|s_{R,n}\|_{\mathscr{H}(\mathbb{X})} \to +\infty$ otherwise.*

Coordinating Proposition 28 with (20) then shows that, for $x \in \mathbb{X}_0$, the reGP predictor can be overconfident by a factor of magnitude at most

$$\sqrt{n}\frac{\sigma_{n,0}(x)}{\sigma_n(x)},$$

when f is an R-element of $\mathscr{H}(\mathbb{X})$. As mentioned in Remark 16, empirical and theoretical evidence about the screening effect suggests that the ratio $\sigma_{n,0}(x)/\sigma_n(x)$ is asymptotically close to one, for finite-smoothness covariance functions. (Alternatively, matching bounds on power functions, such as Equation 16 and Theorem 4.4 from Karvonen et al., 2020, can be used to prove the existence of a constant C such that $\sigma_{n,0}(x) \leq C\sigma_n(x)$, for n large enough, in the case of a quasi-uniform sequence of points.) In this case, reGP can be overconfident by a factor of magnitude at most $\sqrt{n}$ as for vanilla GPs.

If f is not an R-element of $\mathscr{H}(\mathbb{X})$, then (20) does not hold anymore and Proposition 28 shows that $\|s_{R,n}\|_{\mathscr{H}(\mathbb{X})}$ diverges to compensate. The following proposition is an adaptation of Proposition 4.5 from Karvonen et al. (2020), giving an upper bound of $\widehat{\tau}^2_{n,R}$ for quasi-uniform sequences.

Proposition 29 *Suppose that $\mathbb{X}$ is bounded and measurable, that $(x_n)_{n\geq 1}$ is quasi-uniform, and that there exists a $\beta > 0$ such that f is an R-element of $W_2^{\beta+d/2}(\mathbb{X})$. Then, with a Matérn covariance function with regularity $\nu > \beta$, it holds that:*

$$\widehat{\tau}^2_{n,R} \lesssim n^{2(\nu-\beta)/d-1}\|s^{(\beta)}_{R,\mathbb{X}}\|^2_{W_2^{\beta+d/2}(\mathbb{X})}, \tag{31}$$

with $s^{(\beta)}_{R,\mathbb{X}}$ defined in the paragraph preceding Proposition 27.

Following Wang and Jing (2022)—but using a slightly different convention—we define the smoothness $\beta(f) = \sup\{\beta > 0 : f \in W_2^{\beta+d/2}(\mathbb{X})\}$ of a function f. Given a relaxation range R, we use the definition:

$$\beta_R(f) = \sup\left\{\beta > 0 : f \text{ is an } R\text{−element of } W_2^{\beta+d/2}(\mathbb{X})\right\}, \tag{32}$$

taking the relaxation into account. It holds that $\beta(f) = \beta_\emptyset(f) \leq \beta_R(f)$.

Let $\nu > \beta_R(f) > 0$ and assume the supremum (32) defining $\beta_R(f)$ is reached for simplicity. The rate in the upper bound (31) is optimized by taking $\beta = \beta_R(f)$. Under more restrictive assumptions on the spectrum and the support of f, Karvonen et al. (2020) give matching lower bounds for traditional Gaussian process interpolation and discuss uncertainty quantification for functions outside the RKHS. Obtaining a similar result for reGP is more involved and thus left for future work. Considering (30) and Theorem 4.4 from Karvonen et al. (2020), it can be seen that a class of functions such that the bound (31) with $\beta = \beta_R(f)$ is sharp would also essentially be such that reGP can be overconfident by a factor of magnitude at most $\sqrt{n}$ on $\mathbb{X}_0$ for quasi-uniform observations.

Finally, we consider the selection of the smoothness parameter ν given fixed prior variance and range parameters. For Gaussian process interpolation, Karvonen (2023) gives asymptotic bounds for maximum likelihood estimates of ν with quasi-uniform observations from a deterministic function. The Matérn covariance function depends continuously on ν. Therefore, for $0 < \nu_{\min} < \nu_{\max} < +\infty$ and $\Theta = [\nu_{\min}, \nu_{\max}]$, the minimum of (14) is reached on $\Theta \times C_{R,n}$ according to the paragraph following Definition 6. For each n, let $(\widehat{\nu}_n^R, \underline{z}_n^\star)$ be such a minimizer, with ties broken arbitrarily. The following asymptotic lower bound is an adaptation of the first statement of Theorem 3.12 from Karvonen (2023).

Proposition 30 *Let $0 < \nu_{\min} < \nu_{\max} < +\infty$ and $\Theta = [\nu_{\min}, \nu_{\max}]$ and assume $\beta_R(f) + d/2 \in \Theta$, with $\beta_R(f)$ given by* (32)*. If $\mathbb{X}$ satisfies Assumption 9 and the sequence $(x_n)_{n\geq 1}$ is quasi-uniform, it holds that* $\liminf \widehat{\nu}_n^R \geq \beta_R(f) + d/2$.

A smoothness parameter estimate based on reGP is asymptotically strictly larger than $\beta_R(f)$. Consequently, the function f is not an R-element of the RKHS. This echoes the results obtained for Gaussian process interpolation in the previous reference. These results must be seen considering that parameters are selected using the likelihood of a model whose sample paths are almost surely not contained in the RKHS (see, e.g., Lukić and Beder, 2001). Petit (2025, Section 6) studies asymptotically the effect of a joint estimation of ν and τ^2 in a simplified Gaussian process interpolation framework.

Assume the supremum (32) is reached for simplicity and $0 < \beta_R(f) < +\infty$. With a Matérn covariance function with regularity $\nu > 0$, quasi-uniform observations, and $B \subset \mathbb{X}_0$ satisfying Assumption 9, the upper bounds (21) and (30) yield:

$$\begin{cases} \|f - s_{R,n}\|_{L^\infty(B)} \lesssim n^{-\nu/d} & \text{if } \nu \leq \beta_R(f), \\ \|f - s_{R,n}\|_{L^\infty(B)} \lesssim n^{-\beta_R(f)/d} & \text{otherwise.} \end{cases} \tag{33}$$

As with standard Gaussian processes, this suggests that, for quasi-uniform observations, the most accurate reGP models on $\mathbb{X}_0$ are obtained by taking $\nu \geq \beta_R(f)$. Proposition 30 shows that reGP selects such models automatically.

The previous remark also applies with $R = \emptyset$, recovering a conclusion of Karvonen (2023, Section 4.1) for Gaussian process interpolation. Consequently, the inequality $\beta(f) = \beta_\emptyset(f) \leq \beta_R(f)$ and (33) suggest an improved fit with reGP compared to a traditional Gaussian process when both models are used with estimated smoothness parameters.

Finally, note that the quasi-uniformity assumption is rather strong and prevents applications to active learning tasks (see Wynne et al., 2021, Section 5.2, for a possible application to a Bayesian optimization policy designed to ensure quasi-uniformity).

Appendix C. Proofs

Lemma 31 *(Aronszajn, 1950, Section 1.5) Let $k\colon \mathbb{X}\times\mathbb{X}\to\mathbb{R}$ be a positive-definite covariance function, $\mathbb{U}\subset\mathbb{X}$, and $\mathscr{H}(\mathbb{U})$ be the RKHS attached to the restriction of k to $\mathbb{U}\times\mathbb{U}$. The RKHS $\mathscr{H}(\mathbb{U})$ is the space of restrictions of functions from $\mathscr{H}(\mathbb{X})$ and the norm of $g\in\mathscr{H}(\mathbb{U})$ is given by*

$$\inf_{h\in\mathscr{H}(\mathbb{X}),\, h_{|\mathbb{U}}=g}\|h\|_{\mathscr{H}(\mathbb{X})}.$$

Proof of Proposition 3. First the existence and the uniqueness of the solution are given by the first statement of Proposition 12 (with $\mathscr{H}_{R,n}=\mathscr{H}_{R,\{x_1,\dots,x_n\}}$).

Furthermore let $\underline{z}\in\mathbb{R}^n$ and write $\underline{\alpha}=K_n^{-1}\underline{z}$, the reproducing property (6) gives

$$\|s_{\underline{z}}\|^2_{\mathscr{H}(\mathbb{X})}=\underline{\alpha}^{\mathsf{T}}K_n\underline{\alpha}=\underline{z}^{\mathsf{T}}K_n^{-1}\underline{z}, \tag{34}$$

and therefore

$$\min_{h\in\mathscr{H}_{R,n}}\|h\|^2_{\mathscr{H}(\mathbb{X})}=\inf_{\underline{z}\in C_{R,n}}\ \min_{h\in\mathscr{H}(\mathbb{X}),h(\underline{x}_n)=\underline{z}}\|h\|^2_{\mathscr{H}(\mathbb{X})}=\inf_{\underline{z}\in C_{R,n}}\underline{z}^{\mathsf{T}}K_n^{-1}\underline{z},$$

where the last infimum is uniquely reached by the evaluation of the solution on $\underline{x}_n$. ■

Proof of Proposition 4. Write $K_{m,n}$ for the covariance matrix of the random vector $\left(\underline{Z}'^{\mathsf{T}}_m,\underline{Z}^{\mathsf{T}}_n\right)^{\mathsf{T}}$. Using the equalities (4) and (34), and a slight abuse of notation by dropping irrelevant constants with respect to $\underline{z}'$ and $\underline{z}\in C_{R,n}$, we have

$$-2\ln\left(p\left(\underline{z}',\underline{z}\,|\,\underline{Z}_n\in C_{R,n}\right)\right)=\left(\underline{z}'^{\mathsf{T}},\underline{z}^{\mathsf{T}}\right)K_{m,n}^{-1}\left(\underline{z}'^{\mathsf{T}},\underline{z}^{\mathsf{T}}\right)^{\mathsf{T}}=\min_{h\in\mathscr{H}(\mathbb{X}),h(\underline{x}_n)=\underline{z},h(\underline{x}'_m)=\underline{z}'}\|h\|^2_{\mathscr{H}(\mathbb{X})}.$$

This gives

$$\inf_{\underline{z}'\in\mathbb{R}^m,\underline{z}\in C_{R,n}}-2\ln\left(p\left(\underline{z}',\underline{z}\,|\,\underline{Z}_n\in C_{R,n}\right)\right)=\min_{h\in\mathscr{H}(\mathbb{X}),h(\underline{x}_n)\in C_{R,n}}\|h\|^2_{\mathscr{H}(\mathbb{X})},$$

which is reached by taking $\underline{z}=\underline{z}^\star_n$ and $\underline{z}'=\left(s_{\underline{z}^\star_n}(x'_1),\dots,s_{\underline{z}^\star_n}(x'_m)\right)^{\mathsf{T}}$. ■

Proof of Proposition 10. First, one has

$$\sup_{x\in\mathbb{X}}\sigma_n(x)=\sup_{x\in\mathbb{X}}\sup_{\|f\|_{\mathscr{H}(\mathbb{X})}=1}|f(x)-s_{\underline{z}_n}(x)|=\sup_{\|f\|_{\mathscr{H}(\mathbb{X})}=1}\|f-s_{\underline{z}_n}\|_{L^\infty(\mathbb{X})}.$$

Now, let $f\in\mathscr{H}(\mathbb{X})$ such that $\|f\|_{\mathscr{H}(\mathbb{X})}=1$, and $\mathbb{X}^o$ be the interior of $\mathbb{X}$. The boundary of $\mathbb{X}^o$ is the one of $\mathbb{X}$ under Assumption 9, and the Sobolev space $W_2^{\alpha+d/2}(\mathbb{X}^o)$ defined by (15) is norm-equivalent to the Sobolev–Slobodeckij space (see, e.g., Proposition 3.4 of Di Nezza et al., 2012, for a statement on $\mathbb{R}^d$ and Theorem 1.4.3.1 of Grisvard, 1985, for the existence of an extension operator).

Then, one can apply Theorem 4.1 from Arcangéli et al. (2007) to $f-s_{\underline{z}_n}$ restricted to $\mathbb{X}^o$ to show that, for h_n lower than some positive h_0 (not depending on f or $(x_n)_{n\geq1}$), we have:

$$\|f-s_{\underline{z}_n}\|_{L^\infty(\mathbb{X})}=\|f-s_{\underline{z}_n}\|_{L^\infty(\mathbb{X}^o)}\lesssim h_n^\alpha\|f-s_{\underline{z}_n}\|_{W_2^{\alpha+d/2}(\mathbb{X}^o)}\lesssim h_n^\alpha\|f-s_{\underline{z}_n}\|_{\mathscr{H}(\mathbb{X})}\lesssim h_n^\alpha \tag{35}$$

by continuity of $f - s_{\underline{z}_n}$, since $\|\cdot\|_{W_2^{\alpha+d/2}(\mathbb{X}^o)} \leq \|\cdot\|_{W_2^{\alpha+d/2}(\mathbb{X})}$ due to the definition (15), $W_2^{\alpha+d/2}(\mathbb{X})$ being norm equivalent to $\mathscr{H}(\mathbb{X})$, and because of the projection interpretation of $s_{\underline{z}_n}$ (see, e.g., Wendland, 2004, Theorem 13.1). Finally, one can get rid of the condition $h_n \leq h_0$ for simplicity by possibly increasing the constant, since $\sup_{x\in\mathbb{X}} \sigma_n(x) \leq \sup_{x\in\mathbb{X}} \sqrt{k(x,x)} < +\infty$ by compacity. ■

Proof of Proposition 12. First observe that $\mathscr{H}_{R,\mathbb{U}}$ is not empty since it contains f. Furthermore, it is easy to verify that $\mathscr{H}_{R,\mathbb{U}}$ is convex and that it is closed since pointwise evaluation functionals are continuous on an RKHS. The problem is then the one of projecting the null function on a convex closed subset; hence the existence and the uniqueness.

Then, the function $s_{R,n}$ is the projection of the null function on the closed convex set $\mathscr{H}_{R,n}$. Moreover, the sequence $(\mathscr{H}_{R,n})_{n\geq 1}$ is non-increasing, so the sequence $(s_{R,n})_{n\geq 1}$ converges in $\mathscr{H}(\mathbb{X})$ to the projection of the null function on $\bigcap_{n\geq 1} \mathscr{H}_{R,n}$ (see, e.g., Brezis, 2011, Exercice 5.5), i.e., the solution of (19), with $\mathbb{U} = \{x_n\}$. But this last solution is also the solution on the closure since it satisfies the constraints by continuity. ■

Proof of Proposition 13. Define $\underline{x}_n^0$ and $\underline{z}_n^0$ to be data points within $\mathbb{X}_0$, and $s_{\underline{x}_n^0,\underline{z}_n^0}$ the associated (interpolation) predictor, i.e., the solution of (7). Take the null function if there are no observations within B. Observing that $s_{\underline{x}_n^0,\underline{z}_n^0}$ interpolates $s_{R,n}$, we have:

$$\begin{aligned}
\left|f(x) - s_{R,n}(x)\right| &\leq \left|f(x) - s_{\underline{x}_n^0,\underline{z}_n^0}(x)\right| + \left|s_{\underline{x}_n^0,\underline{z}_n^0}(x) - s_{R,n}(x)\right| \\
&\leq \sigma_{n,0}(x)\,\|f\|_{\mathscr{H}(\mathbb{X}_0)} + \sigma_{n,0}(x)\,\|s_{R,n}\|_{\mathscr{H}(\mathbb{X}_0)} \\
&\leq 2\sigma_{n,0}(x)\,\|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})},
\end{aligned}$$

since f coincides with $s_{R,\mathbb{X}}$ on $\mathbb{X}_0$, $s_{R,\mathbb{X}} \in \mathscr{H}_{R,n}$, and by Lemma 31. ■

Lemma 32 *Use the notation* (36)*. If k has smoothness $\alpha > 0$, then there exists $h_0 > 0$ depending only on α such that, for all $x, y \in \mathbb{X}$ satisfying $\|x-y\| \leq h_0$, we have*

$$\|\delta_y - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \lesssim \|x-y\|^\alpha, \text{ for } \alpha < 1,$$

$$\|\delta_y - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \lesssim \sqrt{|\ln(\|x-y\|)|}\,\|x-y\|, \text{ for } \alpha = 1,$$

and

$$\|\delta_y - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \lesssim \|x-y\|, \text{ for } \alpha > 1.$$

Proof. Since equivalent norms give equivalent operator norms on the topological dual of a normed space, it suffices to show the result for a unit-variance isotropic Matérn covariance function (3) of regularity α.

In this case, we have

$$\|\delta_y - \delta_x\|^2_{\mathscr{H}^\star(\mathbb{X})} = k(x,x) + k(y,y) - 2k(x,y) = 2\left(1 - r_\alpha(\|x-y\|)\right),$$

with r_α the corresponding isotropic correlation function. Standard results on principal irregular terms (see, e.g., Stein, 1999, Chapter 2.7) give the results. ■

Lemma 33 *Let $B \subset \mathbb{X}$ satisfying Assumption 9 and $h_{n,B}$ be the fill distance of $\mathbb{X}_{n,B} = \{x_1, \ldots, x_n\} \cap B$ within B, with the convention $h_{n,B} = \operatorname{diam}(B)$ if $\mathbb{X}_{n,B}$ is empty. Then, $h_{n,B} \lesssim h_n$.*

Proof. The idea of the proof is given by Wendland (2004, Lemma 11.31), but it is interlinked with a much more sophisticated construction, so we provide a specific version here for completeness. Adams and Fournier (2003, Section 4.11) show that B satisfies a cone condition with radius $\rho > 0$ and angle $\phi \in (0, \pi/2)$. If $\mathbb{X}_{n,B}$ is not empty, then the compacity of B ensures the existence of an $x \in B$ such that $d(x, \mathbb{X}_{n,B}) = h_{n,B}$. (If $\mathbb{X}_{n,B}$ is empty, then the rest of the proof is also valid taking an arbitrary $x \in B$.)

A cone C originating from x with angle ϕ and radius $\delta = \min(h_{n,B}, \rho)$ is contained in B and its interior do not contains observations. Furthermore, Wendland (2004, Lemma 3.7) shows that there exists a $y \in C$ such that the open ball $B\left(y, \delta \sin(\phi)\,(1+\sin(\phi))^{-1}\right)$ is subset of C, and therefore contains no observations as well. Thus, $h_n \geq \delta \sin(\phi)(1+\sin(\phi))^{-1}$. Now, if $h_{n,B} \leq \rho$, then the desired result follows. If not, the result holds as well since $h_{n,B} \leq \operatorname{diam}(B)$. ■

Proof of Proposition 14. For the first assertion, suppose first that B contains observations and let $\sigma_{n,B}$ be the power function using only the observations within B. It can be shown that the restriction of the kernel to B also has regularity α using (15) and Lemma 31. Using Proposition 13, the inequality $\sigma_{n,0} \leq \sigma_{n,B}$ given by the projection residuals interpretation, and applying Proposition 10 to B yields a bound depending on the fill distance $h_{n,B}$ of $\{x_1, \ldots, x_n\} \cap B$ within B. Finally, Lemma 33 allows us to conclude. Dealing with the case where B contains no observations using (6) makes no difficulty.

Regarding the second assertion, f is continuous so the sets $\mathbb{X}_j$ are compact for $j \geq 1$. In addition, they are disjoint so

$$\delta = \min_{1 \leq j < p} \inf_{x \in \mathbb{X}_j, y \in \mathbb{X}_p} \|x - y\| > 0.$$

Suppose now that $h_n < \delta$ and let $j \geq 1$, $x \in \mathbb{X}_j$ and $1 \leq i \leq n$ the index of the closest x_i to x. By definition, $\|x - x_i\| \leq h_n$ and therefore $x_i \in \mathbb{X}_0 \cup \mathbb{X}_j$. Let $\mathscr{H}^\star(\mathbb{X})$ be the (topological) dual of $\mathscr{H}(\mathbb{X})$ and

$$\delta_y \colon h \in \mathscr{H}(\mathbb{X}) \mapsto h(y), \tag{36}$$

which lies in $\mathscr{H}^\star(\mathbb{X})$ for all $y \in \mathbb{X}$. Then using the reproducing property (6), we have

$$|s_{R,n}(x_i) - s_{R,n}(x)| \leq \|\delta_{x_i} - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \|s_{R,n}\|_{\mathscr{H}(\mathbb{X})} \leq \|\delta_{x_i} - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})},$$

and therefore

$$d(s_{R,n}(x), R_j) \leq \|\delta_{x_i} - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})} + d(s_{R,n}(x_i), R_j).$$

Now, if $x_i \in \mathbb{X}_j$, then $d(s_{R,n}(x_i), R_j) = 0$. Otherwise $x_i \in \mathbb{X}_0$ necessarily and, then, using the fact that $s_{R,\mathbb{X}}(x) \in R_j$, we have:

$$d(s_{R,n}(x_i), R_j) \leq |s_{R,n}(x_i) - s_{R,\mathbb{X}}(x)| = |s_{R,\mathbb{X}}(x_i) - s_{R,\mathbb{X}}(x)| \leq \|\delta_{x_i} - \delta_x\|_{\mathscr{H}^\star(\mathbb{X})} \|s_{R,\mathbb{X}}\|_{\mathscr{H}(\mathbb{X})}.$$

So one can use Lemma 32 along with the previous statements to conclude if $h_n < \min(\delta, h_0)$. (It may be necessary to reduce $\min(\delta, h_0)$ to use the monotonicity of $x \mapsto x\sqrt{-\ln(x)}$, for $x > 0$ close to zero.)

Finally, treating the case $h_n \geq \min(\delta, h_0)$ is straightforward using the fact that $\sup_{x\in\mathbb{X}} \sqrt{k(x,x)}$ is finite thanks to the compacity of $\mathbb{X}$ and $d(s_{R,n}(x), R_j) \leq |s_{R,n}(x) - s_{R,\mathbb{X}}(x)|$ for $j \geq 1$ and $x \in \mathbb{X}_j$. ■

Lemma 34 *If $g, h \in W_2^\gamma(\mathbb{X})$ for $\gamma > d/2$, then $gh \in W_2^\gamma(\mathbb{X})$.*

Proof. By the definition (15) of $W_2^\gamma(\mathbb{X})$, the functions g and h can be extended into functions on $\mathbb{R}^d$, having their product in $W_2^\gamma(\mathbb{R}^d)$ (Strichartz, 1967, Theorem 2.1). Taking the restrictions shows the desired result. ■

Proof of Proposition 15. We use a bump function argument. Let $B(x_0, r) \subset \mathbb{X}_j$ (with $r > 0$) be an open ball. There exists a C^∞ function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that

$$\begin{cases} 0 \leq \phi \leq 1, & \\ \phi(x) = 1 & \text{only if} \quad x = 0, \\ \phi(x) = 0 & \text{if} \quad x \in \mathbb{X} \setminus B(0,r). \end{cases}$$

Let $c \in R_j \setminus \{f(x_0)\}$, $\phi_n = \phi(n(\cdot - x_0))$ as a function on $\mathbb{X}$, and $f_n = (1-\phi_n) f + c\phi_n$, for $n \geq 1$. We have $\phi_n \in W_2^{\alpha+d/2}(\mathbb{R}^d)$ as a function on $\mathbb{R}^d$, so it belongs to $W_2^{\alpha+d/2}(\mathbb{X})$ as a function on $\mathbb{X}$, and Lemma 34 ensures that $f_n \in \mathscr{H}(\mathbb{X})$. Moreover, it is easy to check that $f_n \in \mathscr{H}_{R,\mathbb{X}}$. Observe that the sequence $(f_n)_{n\geq 1}$ converges pointwise to a discontinuous function that lies thus outside $\mathscr{H}(\mathbb{X})$.

Suppose now that $\|f_n\|_{\mathscr{H}(\mathbb{X})} \nrightarrow +\infty$. Then, one can extract a bounded subsequence of norms and a classical weak compacity argument would yield a weakly convergent subsequence, which is impossible since the pointwise limit is not in $\mathscr{H}(\mathbb{X})$. ■

Lemma 35 *Use the notations from the proof of Proposition 18 and let $(y_n)_{n\geq 1}$ be a sequence in $\mathbb{X}$. Assume that the sequence $(y_n)_{n\geq 1}$ is convergent, denote by $y^\star$ its limit and assume that $y^\star$ is an adherent point of the set $\{x_n\}$. Let $t_\infty = \liminf t_n$, then*

- *$s_n(y_n) \to f(y^\star)$ if $f(y^\star) < t_\infty$,*
- *$\liminf s_n(y_n) \geq t_\infty$, otherwise.*

In particular, we have

$$\liminf s_n(y_n) \geq \min(f(y^\star),\, t_\infty).$$

Proof. Suppose that $y^\star \notin \{x_n\}$. Then, let $\left(x_{\phi(n)}\right)_{n\geq 1}$ be a subsequence converging to $y^\star$ and let $\psi(n) = \max\{\phi(k), \phi(k) \leq n\}$. We proceed as in Proposition 14 to have:

$$\left|s_n(x_{\psi(n)}) - s_n(y_n)\right| \leq \|\delta_{x_{\psi(n)}} - \delta_{y_n}\|_{\mathscr{H}^\star(\mathbb{X})} \|s_n\|_{\mathscr{H}(\mathbb{X})} \to 0, \tag{37}$$

thanks to the continuity of k and the inequality $\|s_n\|_{\mathscr{H}(\mathbb{X})} \leq \|f\|_{\mathscr{H}(\mathbb{X})}$.

Finally, we have $s_n(x_{\psi(n)}) \geq \min\left(f\left(x_{\psi(n)}\right), t_n\right)$ by construction and therefore $\liminf s_n(x_{\psi(n)}) \geq \min\left(f\left(y^\star\right), t_\infty\right)$, which gives the second assertion thanks to (37). Observe that $f(x_{\psi(n)}) < t_n$ ultimately if $f(y^\star) < t_\infty$ for the first assertion.

If $y^\star \in \{x_n\}$, then the result follows similarly. ■

Lemma 36 *Using the notations from the proof of Proposition 18 and writing* $v_n = \sup_{x\in\mathbb{X}} \rho_{n,t_n}(x)$, *we have* $\liminf v_n = 0$.

Proof. This is an adaptation of Lemma 12 from Vazquez and Bect (2010b).

Let $y^\star$ be a cluster point of $(x_n)_{n\geq 1}$ and let $\left(x_{\phi(n)}\right)_{n\geq 1}$ be a subsequence converging to $y^\star$. We are going to prove that $v_{\phi(n)-1} \to 0$. Define

$$z_{\phi(n)-1} = m_{\phi(n)-1} - s_{\phi(n)-1}(x_{\phi(n)}).$$

Then, Lemma 35 gives[1] $\liminf s_{\phi(n)-1}(x_{\phi(n)}) \geq \min(f(y^\star),\, t_\infty)$. Moreover we have

$$m_{\phi(n)-1} \leq \min\left(f\left(x_{\phi(n-1)}\right),\, t_{\phi(n)-1}\right),$$

since $\phi(n-1) \leq \phi(n)-1$, so $\lim m_{\phi(n)-1} \leq \min(f(y^\star),\, t_\infty)$ holds because $\left(m_{\phi(n)-1}\right)_{n\geq 1}$ is non-increasing. The previous arguments show that $\limsup z_{\phi(n)-1} \leq 0$.

Moreover, one can use Proposition 10 from Vazquez and Bect (2010b, $(i) \Rightarrow (ii)$) similarly to show that $\sigma^2_{\phi(n)-1}(x_{\phi(n)}) \to 0$ and therefore

$$v_{\phi(n)-1} = \gamma\left(z_{\phi(n)-1},\, \sigma^2_{\phi(n)-1}(x_{\phi(n)})\right) \leq \gamma\left(\sup_{k\geq n} z_{\phi(k)-1},\, \sigma^2_{\phi(n)-1}(x_{\phi(n)})\right) \to 0,$$

since γ is non-decreasing with respect to its first argument and continuous. ■

Proof of Proposition 18. This is an adaptation of Theorem 6 from Vazquez and Bect (2010b). Write $s_n = s_{R_n,n}$ for the reGP predictor at step n to avoid cumbersome notations. Then, for $x \in \mathbb{X}$, write $\rho_{n,t_n}(x) = \gamma(m_n - s_n(x),\, \sigma^2_n(x))$ for the expected improvement under the reGP predictive distribution, with γ the function defined in Proposition 1.

Suppose that there exists some $x_0 \in \mathbb{X}$ such that $\sigma^2_n(x_0) \geq C_1 > 0$. The sequence $(m_n)_{n\geq 1}$ converges and the reproducing property (6) yields

$$|s_n(x_0)| \leq \sqrt{k(x_0,\, x_0)}\|s_n\|_{\mathscr{H}(\mathbb{X})} \leq \sqrt{k(x_0,\, x_0)}\|f\|_{\mathscr{H}(\mathbb{X})},$$

so the sequence $(|m_n - s_n(x_0)|)_{n\geq 1}$ is bounded by, say C_2. We then have

$$v_n = \sup_{x\in\mathbb{X}} \rho_{n,t_n}(x) \geq \gamma\left(m_n - s_n(x_0), \sigma^2_n(x_0)\right) \geq \gamma(-C_2, C_1) > 0$$

by Proposition 1. But this yields a contradiction with Lemma 36, so the decreasing sequence $\left(\sigma^2_n\right)_{n\geq 1}$ converges pointwise on $\mathbb{X}$ to zero. Proposition 10 from Vazquez and Bect (2010b) then implies that every $x \in \mathbb{X}$ is adherent to $\{x_n\}$. ■

Lemma 37 *Assume that* b *is finite, and that either* a *is finite too or* $\int_{-\infty}^0 F_P(u)\,\mathrm{d}u = \int_{-\infty}^0 |u|\,P(\mathrm{d}u)$ *is finite. Then*

$$S^{\mathrm{tCRPS}}_{a,b}(P, z) \;=\; (b - a\vee z)_+ + R_2(a,b) - 2R_1(a\vee z, b),$$

where

$$R_q(a,b) \;=\; \int_{-\infty}^{+\infty} \mathbb{1}_{a\leq u\leq b} F_P(u)^q\,\mathrm{d}u.$$

1. Lemma 35 yields $\liminf s_{n-1}(x_{\psi(n)}) \geq \min(f(y^\star),\, t_\infty)$ with $\psi(n) = \max\{\phi(k), \phi(k) \leq n\}$, and the claim follows by extracting a ϕ-subsequence.

Proof.

$$\begin{aligned}
S_{a,b}^{\mathrm{tCRPS}}(P,z) &= \int_{-\infty}^{+\infty} \mathbb{1}_{a\le u\le b}\,(\mathbb{1}_{z\le u}-F_P(u))^2\,\mathrm{d}u \\
&= \int_{-\infty}^{+\infty} \mathbb{1}_{a\le u\le b}\left(\mathbb{1}_{z\le u}+F_P(u)^2-2\,\mathbb{1}_{z\le u}F_P(u)\right)\mathrm{d}u \\
&= \int_{-\infty}^{+\infty} \left(\mathbb{1}_{a\vee z\le u\le b}+\mathbb{1}_{a\le u\le b}F_P(u)^2-2\,\mathbb{1}_{a\vee z\le u\le b}F_P(u)\right)\mathrm{d}u \\
&= (b-a\vee z)_+ + R_2(a,b) - 2R_1(a\vee z,b).
\end{aligned}$$

■

Lemma 38 *Let $a,b\in\mathbb{R}$ with $a\le b$. Let $q\ge 1$. Then*

$$R_q(a,b) = b-a+\mathrm{EI}_q^{\uparrow}(P,b)-\mathrm{EI}_q^{\uparrow}(P,a).$$

Proof.

$$\begin{aligned}
R_q(a,b) &= \int \mathbb{1}_{a\le u\le b}F_P(u)^q\,\mathrm{d}u \\
&= \int \mathbb{1}_{a\le u\le b}\prod_{j=1}^{q}\mathsf{E}\left(\mathbb{1}_{N_j\le u}\right)\mathrm{d}u \qquad \text{with } N_j\overset{\text{iid}}{\sim}P \\
&= \mathsf{E}\left(\int \mathbb{1}_{a\vee N_1\vee\cdots\vee N_q\le u\le b}\,\mathrm{d}u\right) \\
&= \mathsf{E}\left((b-a\vee N_1\vee\cdots\vee N_q)_+\right) \\
&= b-a+\mathrm{EI}_q^{\uparrow}(P,b)-\mathrm{EI}_q^{\uparrow}(P,a).
\end{aligned}$$

■

Proof of Proposition 22. The first result is given by Lemma 37 and Lemma 38.

Then using the dominated convergence theorem, it is easy to see that, when $a\to-\infty$

$$\mathrm{EI}_q^{\uparrow}(P,a) = \mathsf{E}\left(N_1\vee\cdots\vee N_q\right)-a+o(1),$$

and therefore

$$R_q(-\infty,b) = \lim_{a\to-\infty}R_q(a,b) = b+\mathrm{EI}_q^{\uparrow}(P,b)-\mathsf{E}\left(N_1\vee\cdots\vee N_q\right),$$

which gives the second statement.

Finally, a change of variable gives

$$\int_a^{+\infty}\left(F_P(u)-\mathbb{1}_{z\le u}\right)^2\mathrm{d}u=\int_{-\infty}^{-a}\left(F_{\underline{P}}(u)-P(U=-u)-\mathbb{1}_{-z<u}\right)^2\mathrm{d}u,$$

and the last statement follows by observing that a probability measure admits at most a countable number of atoms. ■

Proof of Proposition 25. The proof follows closely those of Theorem 4.2 from Narcowich et al. (2006), Theorem 4.2 from Karvonen et al. (2020), and Theorem 1 from Wynne et al. (2021). It is therefore not entirely repeated for brevity. Note that Theorem 4.1 from Arcangéli et al. (2007) must be used with $q = +\infty$ and $q = 2$ but requires $\mathbb{X}$ to be open. Under Assumption 9, the adaptation for $q = +\infty$ is given by (35) in the proof of Proposition 10. For $q = 2$, use the fact that a locally Lipschitz boundary has zero measure.

The proofs in the references give the result for h_n lower than some h_0 (not dependent of $(x_n)_{n\geq 1}$ and f). Removing this condition by possibly increasing the constant makes no difficulty.

The fact that $h_n \lesssim n^{-1/d}$ for a quasi-uniform sequence (see, e.g., Wendland, 2004, Proposition 14.1) gives the second statement. ■

Proof of Proposition 27. Assume there are observations $(\underline{x}_n^B, \underline{z}_n^B)$ within B and let $s_{\underline{x}_n^B, \underline{z}_n^B}$ be the associated (interpolation) predictor. We have:

$$\|f - s_{R,n}\|_{L^\infty(B)} \leq \|f - s_{\underline{x}_n^B, \underline{z}_n^B}\|_{L^\infty(B)} + \|s_{\underline{x}_n^B, \underline{z}_n^B} - s_{R,n}\|_{L^\infty(B)}.$$

Use the notations of Lemma 33, let $q_{n,B}$ be the separation distance of points within B (with the convention $q_{n,B} = \mathrm{diam}(\mathbb{X})$ if $\mathbb{X}_{n,B}$ has less than two distinct elements), and let $\rho_{n,B} = h_{n,B}/q_{n,B}$. It holds that $f_{|B} \in W_2^{\beta+d/2}(B)$ since $B \subset \mathbb{X}_0$ and f is an R-element of $W_2^{\beta+d/2}(\mathbb{X})$. It can be shown that the restriction of the kernel to B also has regularity α using (15) and Lemma 31. Therefore, we can apply Proposition 25 to $f_{|B}$ to get:

$$\|f - s_{\underline{x}_n^B, \underline{z}_n^B}\|_{L^\infty(B)} \lesssim h_{n,B}^\beta (1 + \rho_{n,B}^{\alpha-\beta}) \|f\|_{W_2^{\beta+d/2}(B)} \lesssim h_n^\beta (1 + \rho_n^{\alpha-\beta}) \|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})},$$

where we used Lemma 33, the inequality $q_n \leq q_{n,B}$, the definition (15), and the fact that $s_{R,\mathbb{X}}^{(\beta)}$ coincides with f on B.

Consider the restrictions on B and observe that $s_{R,n} \in \mathscr{H}(B)$ and that $s_{\underline{x}_n^B, \underline{z}_n^B}$ is the associated interpolation predictor. One can combine (8) and Proposition 10 to get

$$\|s_{\underline{x}_n^B, \underline{z}_n^B} - s_{R,n}\|_{L^\infty(B)} \lesssim h_{n,B}^\alpha \|s_{R,n}\|_{\mathscr{H}(B)} \lesssim h_n^\alpha \|s_{R,n}\|_{\mathscr{H}(\mathbb{X})}$$

using Lemmata 31 and 33.

Observe that the proof of Lemma A.1. from Karvonen et al. (2020) is valid without restrictions on the domain $\mathbb{X}$. Therefore, there exists a constant $C_\beta > 0$ such that, for each n, there exists a function $u_n^{(\beta)} \in W_2^{\alpha+d/2}(\mathbb{X})$ coinciding with $s_{R,\mathbb{X}}^{(\beta)}$ on the design $\underline{x}_n$ and satisfying

$$\|u_n^{(\beta)}\|_{W_2^{\alpha+d/2}(\mathbb{X})} \leq C_\beta q_n^{\beta-\alpha} \|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})}.$$

To conclude, observe that $u_n^{(\beta)} \in \mathscr{H}_{R,n}$ and thus

$$\|s_{R,n}\|_{\mathscr{H}(\mathbb{X})} \leq \|u_n^{(\beta)}\|_{\mathscr{H}(\mathbb{X})} \lesssim \|u_n^{(\beta)}\|_{W_2^{\alpha+d/2}(\mathbb{X})} \lesssim q_n^{\beta-\alpha} \|s_{R,\mathbb{X}}^{(\beta)}\|_{W_2^{\beta+d/2}(\mathbb{X})}. \tag{38}$$

The previous display is also true when B does not contain observations. Coordination with (6) and Lemma 33 shows that (29) also holds in this case.

The fact that $h_n \lesssim n^{-1/d}$ for a quasi-uniform sequence (see, e.g., Wendland, 2004, Proposition 14.1) gives the second statement. ■

Proof of Proposition 28. Let $g \in \mathscr{H}(\mathbb{X})$ make f an R-element of $\mathscr{H}(\mathbb{X})$. Applying Proposition 12 to g proves one implication.

Conversely, the sequence $(\|s_{R,n}\|_{\mathscr{H}(\mathbb{X})})_{n\geq 1}$ is non-decreasing. Suppose that it is bounded. One can then extract a weakly convergent subsequence from $(s_{R,n})_{n\geq 1}$. The sets $\mathscr{H}_{R,n}$ are closed and convex and hence weakly closed. The weakly convergent subsequence is ultimately in each of these sets so its limit lies in their intersection. It can be checked that the existence of a continuous function coinciding with f on $\mathbb{X}_0$ and defining the same $\mathbb{X}_j$s ensures that this intersection is equal to $\mathscr{H}_{R,\mathbb{X}}$. ■

Proof of Proposition 29. The result is given by (38) with $\alpha = \nu$ and $q_n \lesssim n^{-1/d}$ by quasi-uniformity (see, e.g., Wendland, 2004, Proposition 14.1). ■

Proof of Proposition 30. The proof follows the ideas of Karvonen (2023, Theorem 3.11 and Theorem 3.12). Write $K_{n,\nu}$ to underline the dependence of the covariance matrix on the parameter ν. Let $\varepsilon > 0$ be small enough and write

$$\mathscr{M}(\nu) = \min_{\underline{z}\in C_{R,n}} \mathscr{L}(\nu;\underline{z}) \propto \text{constant} + \log\left(\det\left(K_{n,\nu}\right)\right) + \min_{\underline{z}\in C_{R,n}} \underline{z}^{\mathsf{T}} K_{n,\nu}^{-1}\underline{z}.$$

For $\tau = \beta_R(f) + d/2 - \varepsilon/2$, we are going to prove that $\inf_{\nu_{\min}\leq\nu\leq\beta_R(f)+d/2-\varepsilon} \mathscr{M}(\nu) - \mathscr{M}(\tau)$ is ultimately positive.

First, observe that Proposition 3.7 of Karvonen (2023) applies in our setting (by potentially transforming $\mathbb{X}$ to deal with an isotropic covariance function). Our Proposition 10 adapts Proposition 3.6 from the previous reference in the setting of Assumption 9. Proceed as in the reference by using these two propositions and Stirling's formula to get

$$\inf_{\nu_{\min}\leq\nu\leq\beta_R(f)+d/2-\varepsilon} \log(\det(K_{n,\nu})) - \log(\det(K_{n,\tau})) \geq \mathscr{O}(n) + \frac{\varepsilon}{d}\log(n!) \sim \frac{\varepsilon}{d} n\log(n).$$

To conclude, observe that f is an R-element of $W_2^{\beta_R(f)-\varepsilon/2+d/2}(\mathbb{X})$ by definition, so (38) and quasi-uniformity yields:

$$\min_{\underline{z}\in C_{R,n}} \underline{z}^{\mathsf{T}} K_{n,\tau}^{-1}\underline{z} \lesssim n.$$

■

Appendix D. EI Benchmark Results

The results are provided in Figures 11–15, for all the test functions from Table 1, using the same format as in Figure 8.

Observe that all the EGO-R methods yield (sometimes very) substantial improvements on Beale, Goldstein-Price, Six-hump Camel, Dixon-Price (4), all instances of Perm and Rosenbrock, Three-hump Camel, and Zakharov (4) and (6). Focusing on these test functions, the concentration heuristic appears to give the best results, while the constant heuristic brings the least improvement. The spatial heuristic gives intermediate results between the two, sometimes giving results similar to the concentration heuristic and, at other times, results similar to the constant heuristic.

Taking variability (colored areas) into account, EGO-R with the constant and spatial heuristics performs only slightly better than EGO on Cross-in-Tray and Dixon-Price (6) and (10). The concentration heuristic, however, brings substantial improvements on these test functions, though. All EGO-R methods perform as EGO on Branin, Ackley (4) and (6), Log-Goldstein-Price, Michalewicz (4) and (10), all the Shekel instances, Hartman (3), and Zakharov (10).

The performance of the EGO-R methods is rarely worse than EGO, and never dramatically so. Nevertheless, the concentration heuristic yield substantially higher median empirical best minima on Ackley (10), and Michalewicz (6) and (10). This is also the case of the spatial heuristic on Ackley (10).

Remark 39 *The results are spread out for the Hartman* (6) *function. Investigations have shown this is due to a fraction of runs getting trapped in local minima. The proportions of runs reaching, for* $n = 300$*, an empirical best evaluation close to the global minimum are* 72%*,* 68%*,* 67%*, and* 66% *for EGO and EGO-R with the constant, spatial, and concentration heuristics, respectively. (To measure the closeness to the global minimum, a list* $x_1^\star, \ldots, x_N^\star$ *of local minima was estimated by manually analyzing the output of repetitions of a random-start local optimizer. Then, a function value* $f(x)$ *is said to be close to the global minimum if* $|f(x) - f(x_i^\star)|$ *is minimized by* $\arg\min_i f(x_i^\star)$*.)*

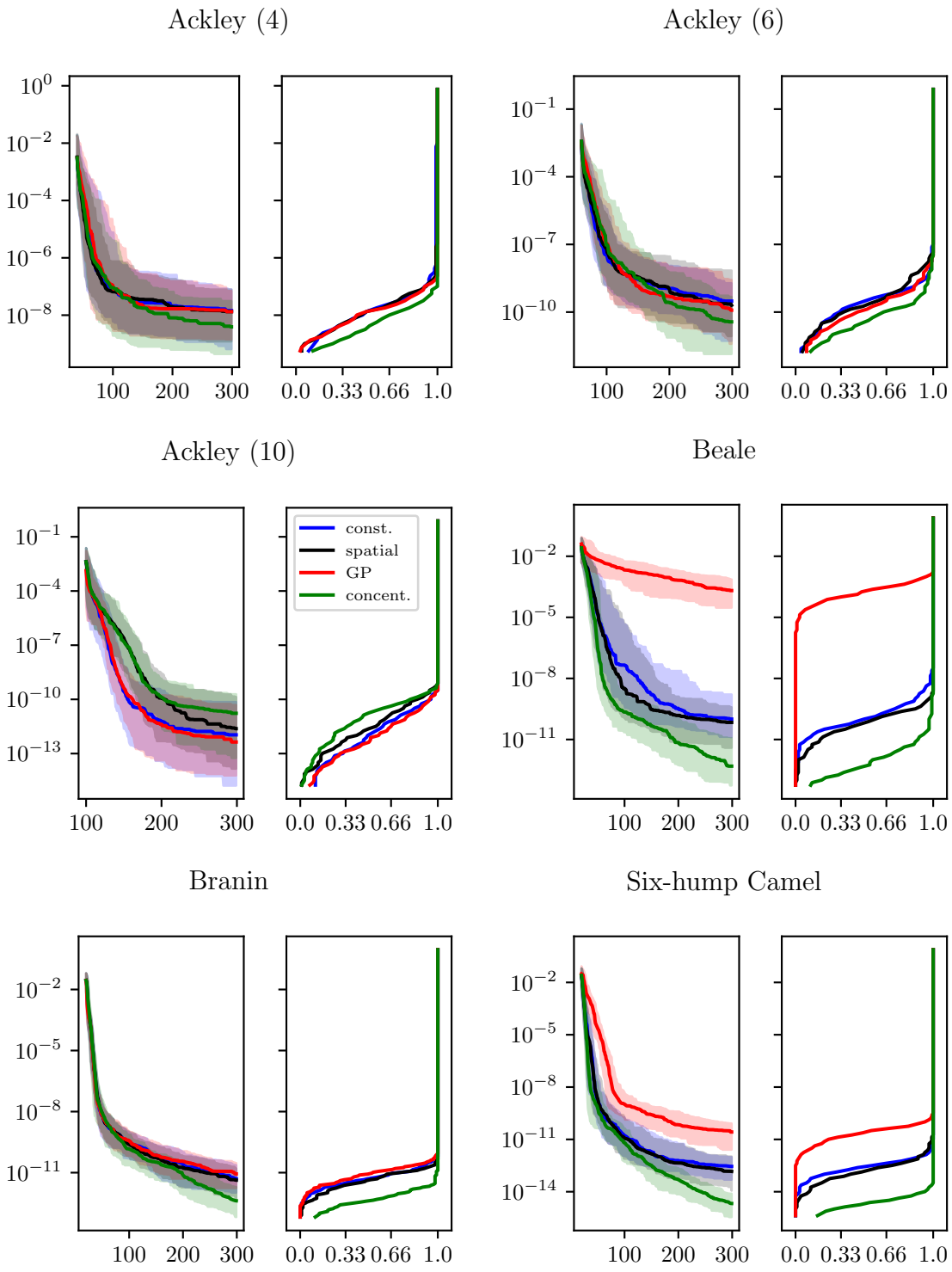

Figure 11: EGO and EGO-R results for a subset of the test functions from Table 1. Same legend as in Figure 8.

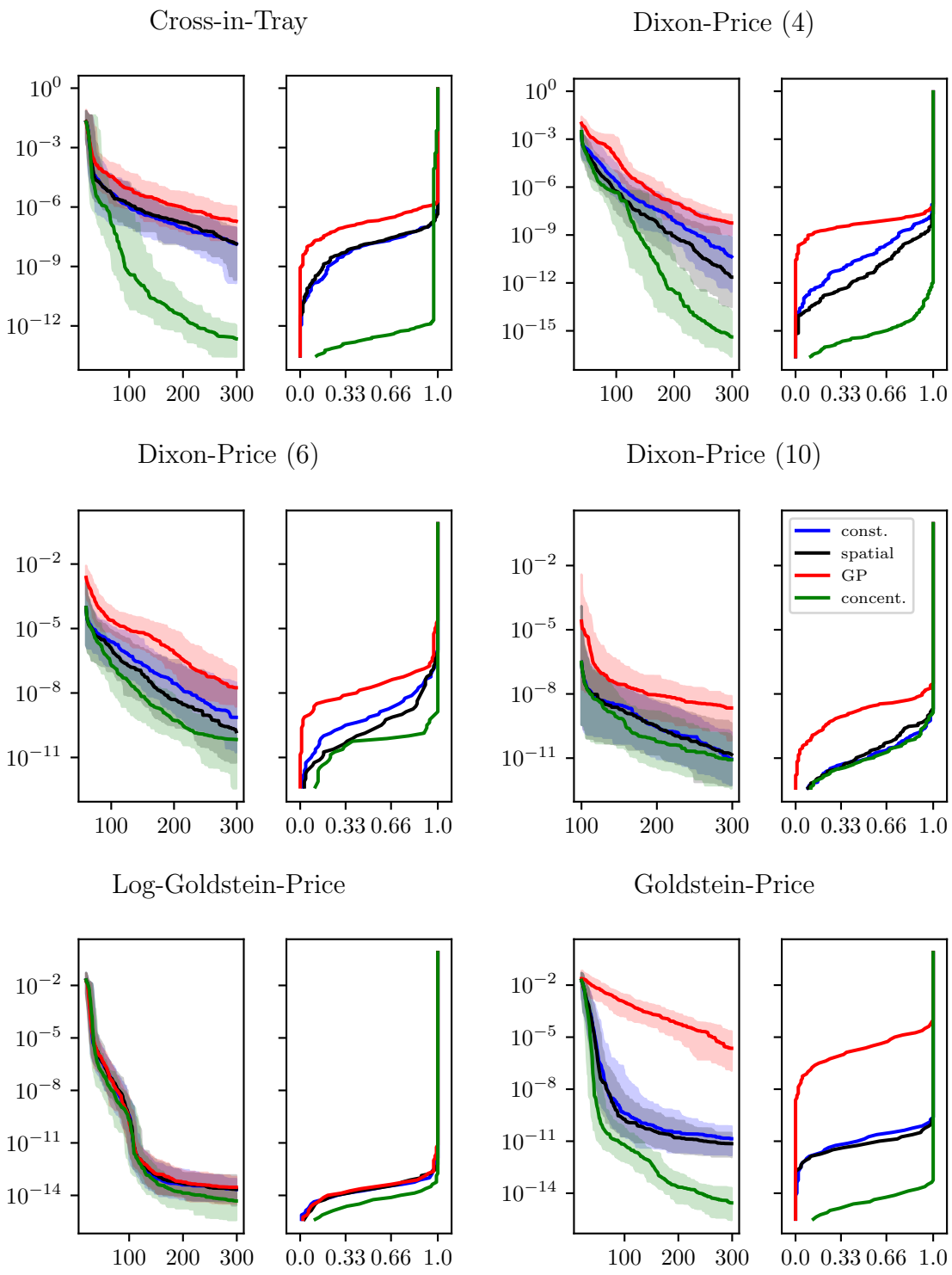

Figure 12: EGO and EGO-R results for a subset of the test functions from Table 1. Same legend as in Figure 8.

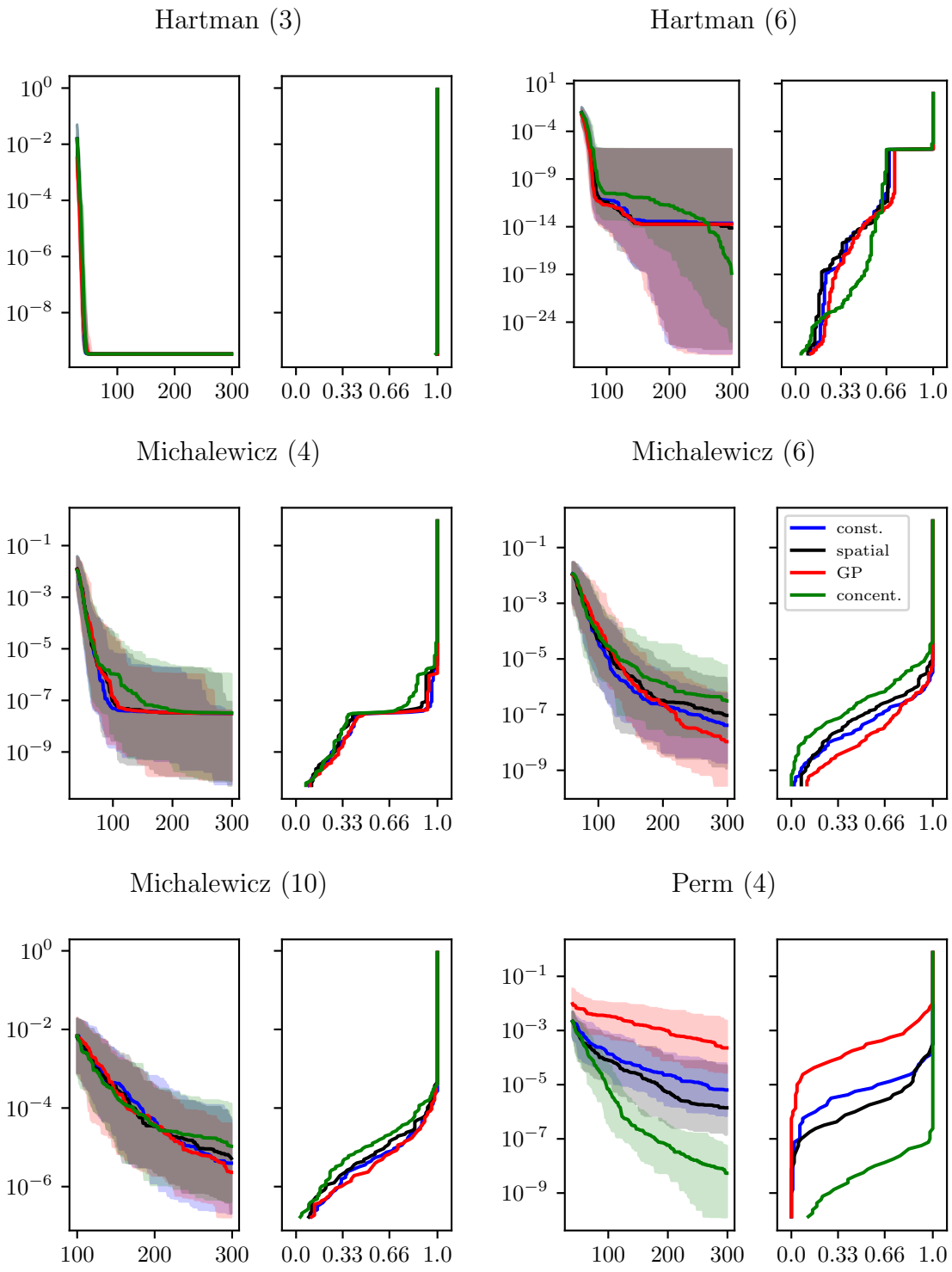

Figure 13: EGO and EGO-R results for a subset of the test functions from Table 1. Same legend as in Figure 8.

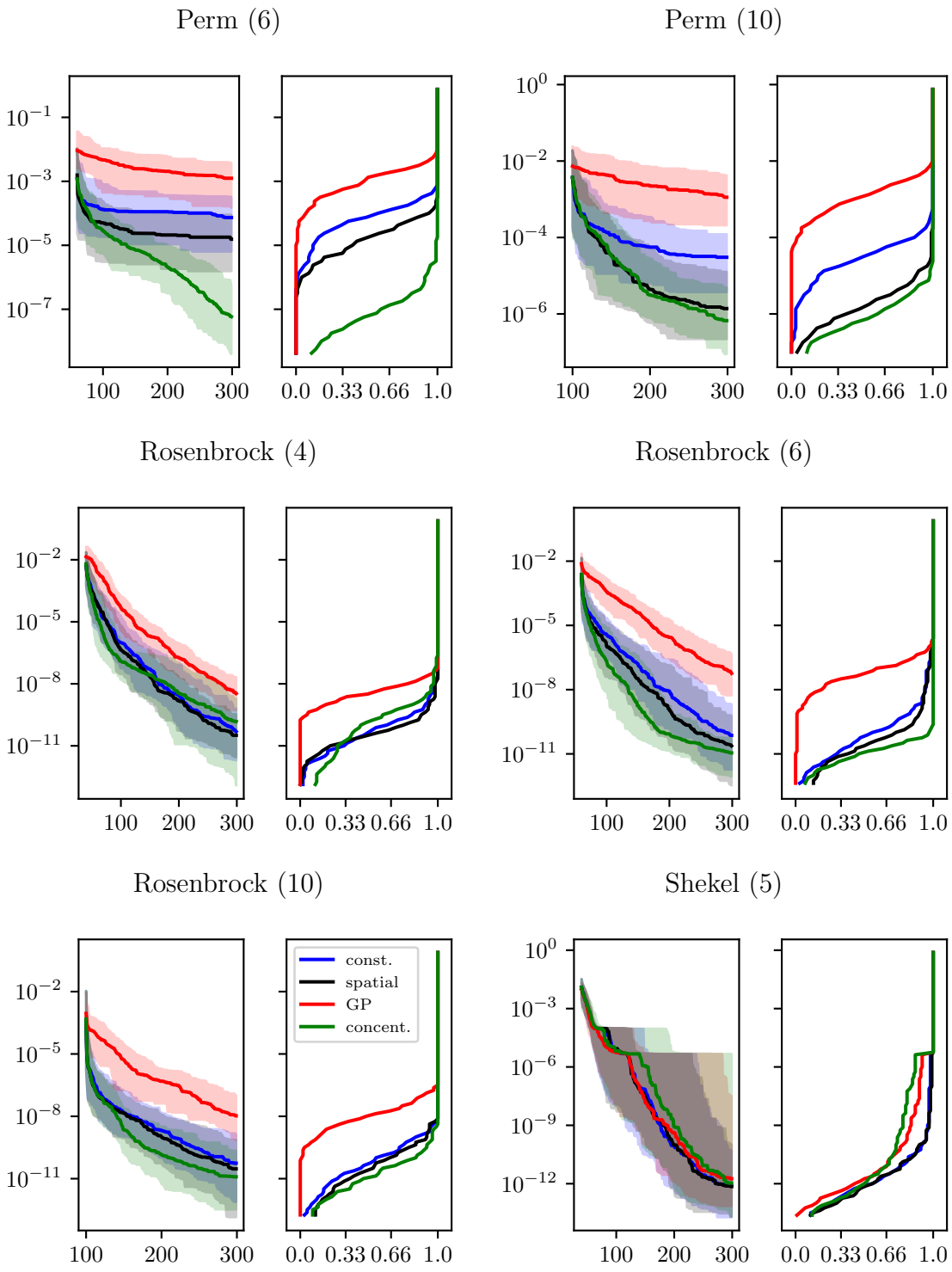

Figure 14: EGO and EGO-R results for a subset of the test functions from Table 1. Same legend as in Figure 8.

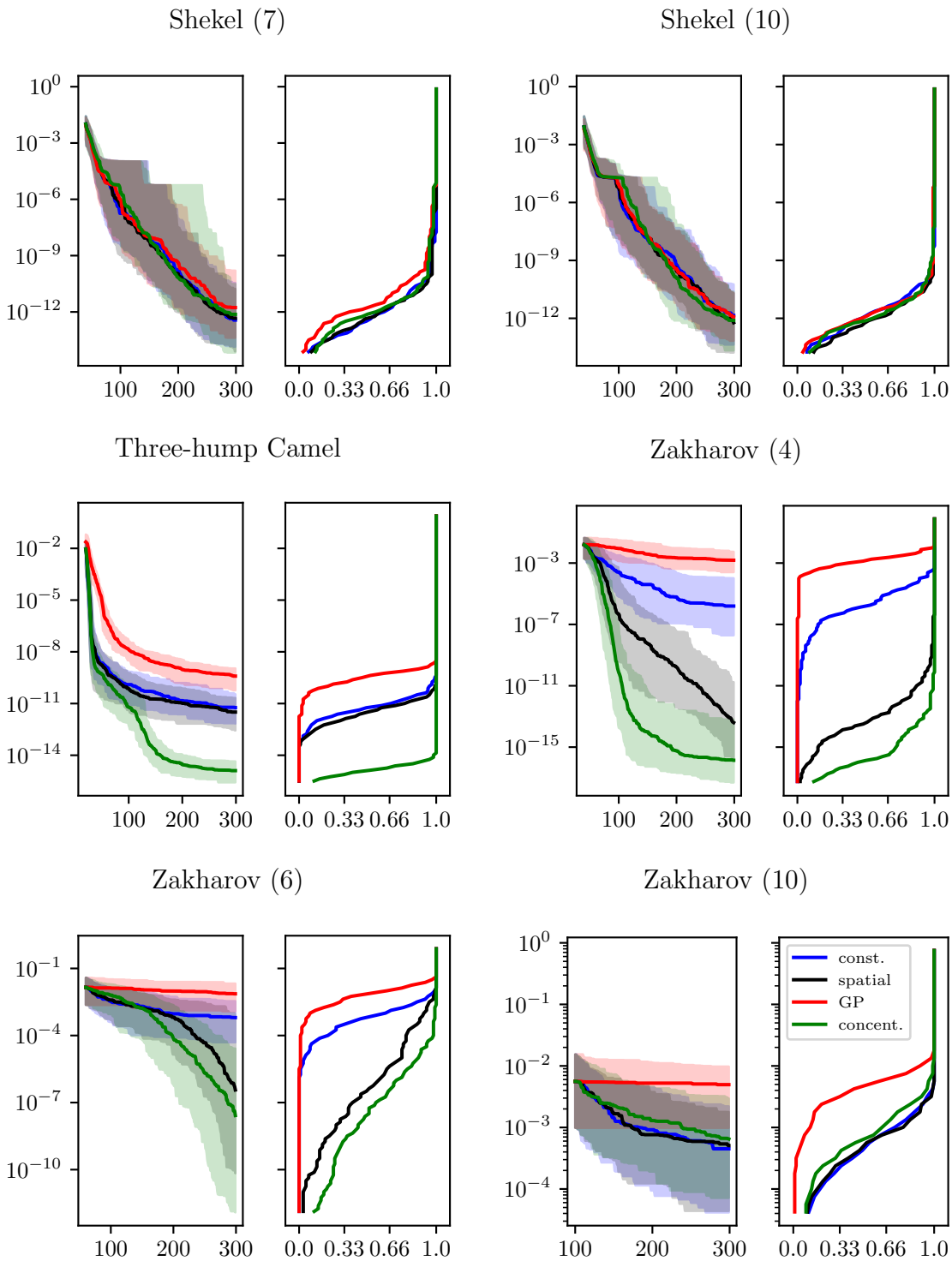

Figure 15: EGO and EGO-R results for a subset of the test functions from Table 1. Same legend as in Figure 8.

Appendix E. UCB Benchmark Results

The Upper Confidence Bound (UCB, Cox and John, 1992; Srinivas et al., 2010) is an other sampling criterion for function minimization. Given a parameter $\beta > 0$, the function f is minimized by choosing sequentially x_{n+1} minimizing

$$\mu_n(x) - \beta\sigma_n(x), \tag{39}$$

with respect to $x \in \mathbb{X}$. In other words, the UCB algorithm minimizes an optimistic quantile of the Gaussian predictive distribution to select the next observation. Large values of β favor exploration.

Varying β slowly with n helps to prove convergence (Srinivas et al., 2010). Nevertheless, a few experiments (not shown here) suggest that, overall, on the test functions from Table 1, the best vanilla GP baselines were obtained with a fixed value. Consequently, this finite-sample study was restricted to a fixed value of β as in Cox and John (1992) and Picheny et al. (2013). The value was chosen so that (39) corresponds to the 10%-quantile of the predictive distribution.

As with the EI criterion, the application of reGP to the UCB algorithm simply involves plugging the reGP predictive distribution into the criterion (39). For choosing the validation threshold $t_n^{(0)}$, we will again compare the three heuristics introduced in Section 5.3.1, namely constant, concentration, and spatial. (Given $t_n^{(0)}$, the relaxation range candidates will also be defined as in Section 5.3.1.) More generally, the assessment was carried out using the methodology described in Section 5.3.1. The criterion (39) was optimized using a subset simulation algorithm. The results are shown in Figures 16–20.

When applied to UCB, all reGP variants outperform the standard UCB algorithm on Beale, Six-hump Camel, Dixon-Price (4) and (10), Goldstein-Price, all instances of Perm and Rosenbrock, Three-hump Camel, Zakharov (4) and (6). Furthermore, the concentration heuristic yields the best results on these test functions.

The spatial and constant heuristics perform as standard GPs on Branin, and Dixon-Price (6). Only the concentration heuristic brings a substantial benefit. However, GPs and all reGP variants give similar results on Ackley (4), Michalewicz (6) and (10), Hartman (3), and Zakharov (10). The concentration heuristic underperforms on Ackley (6) and (10).

The results are spread out for all algorithms on Cross-in-Tray, Log-Goldstein-Price, Hartman (6), Michalewicz (4), and Shekel (5), (7), and (10). As with the Hartman (6) function with EI, this is caused by local minima. This is a known feature of the standard UCB algorithm with fixed β (see Jones, 2001, Section 5). (However, as mentioned earlier, this produced the best baselines, overall, in this study.) Table 2 shows the fractions of runs getting close to the global minima on these test functions. Observe that the reGP variants are not routinely more often trapped in local minima. Furthermore, the figures show that the concentration heuristic still benefits runs reaching the global minimum.

Table 2: Proportion of UCB runs reaching, for $n = 300$, an empirical best evaluation close to the global minimum. These proportions are estimated as described in Appendix D.

Problem	GP	constant	spatial	concentration
Hartman (6)	66%	65%	65%	67%
Shekel 5	31%	32%	36%	36%
Shekel 7	41%	44%	46%	47%
Shekel 10	19%	16%	16%	16%
Log-Goldstein-Price	87%	84%	81%	94%
Cross-in-Tray	97%	88%	86%	88%
Michalewicz (4)	26%	22%	16%	24%

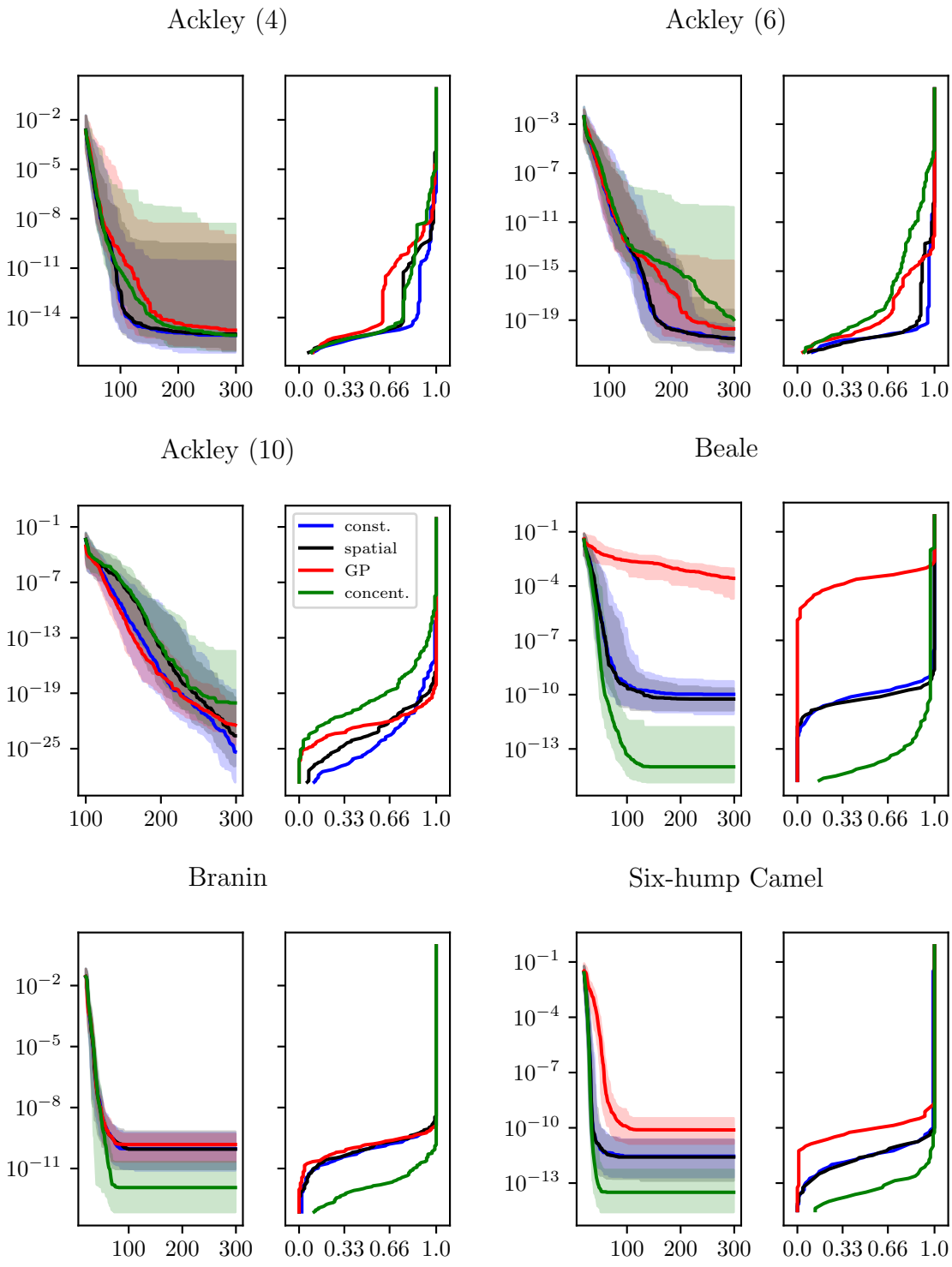

Figure 16: UCB results for a subset of the test functions from Table 1. Same legend as in Figure 8. The red color stands for the UCB algorithm with vanilla GPs and the blue, green, and black colors for UCB with reGP, using the constant, concentration, and spatial heuristics respectively.

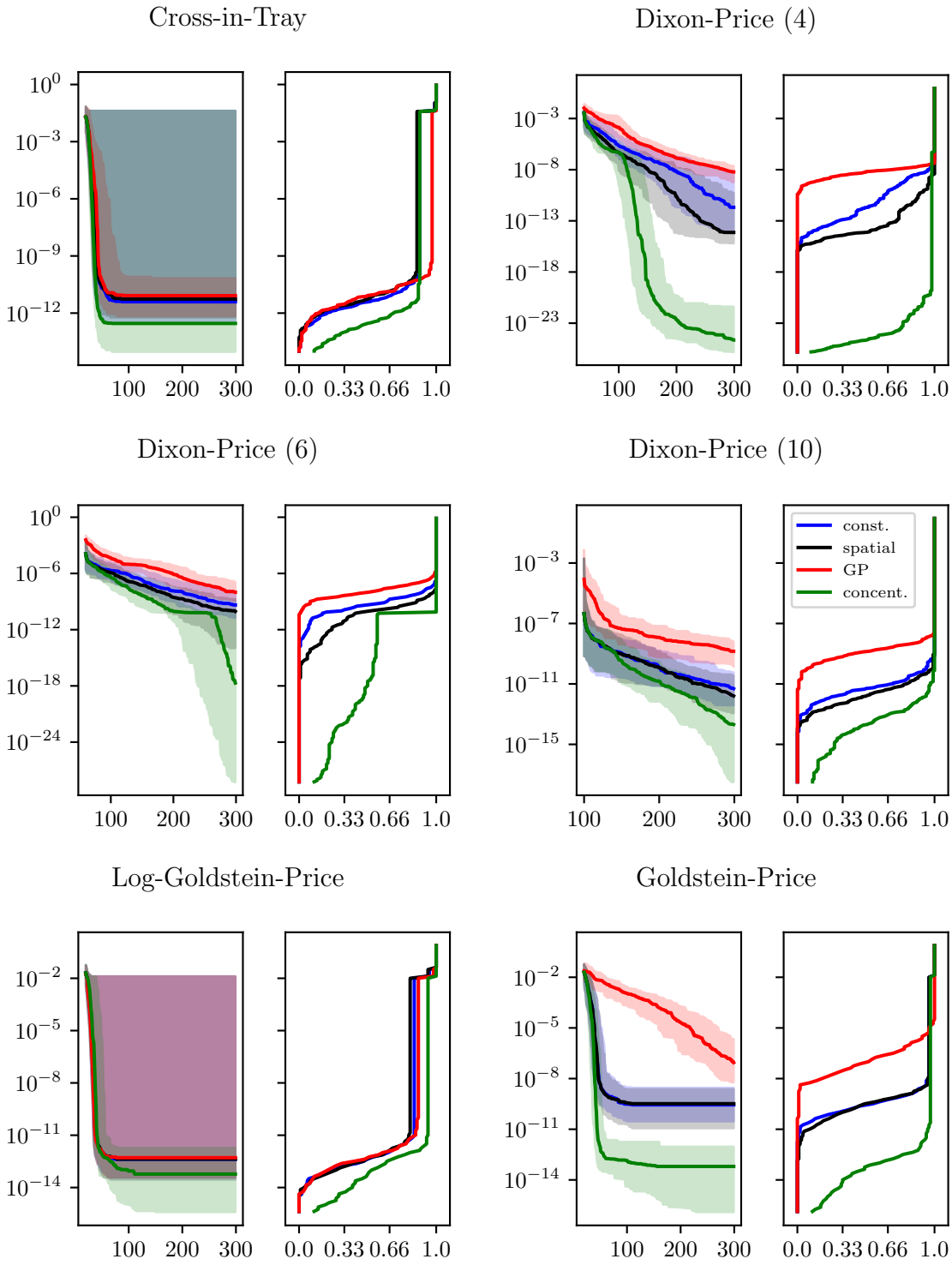

Figure 17: UCB results for a subset of the test functions from Table 1. Same legend as in Figure 16.

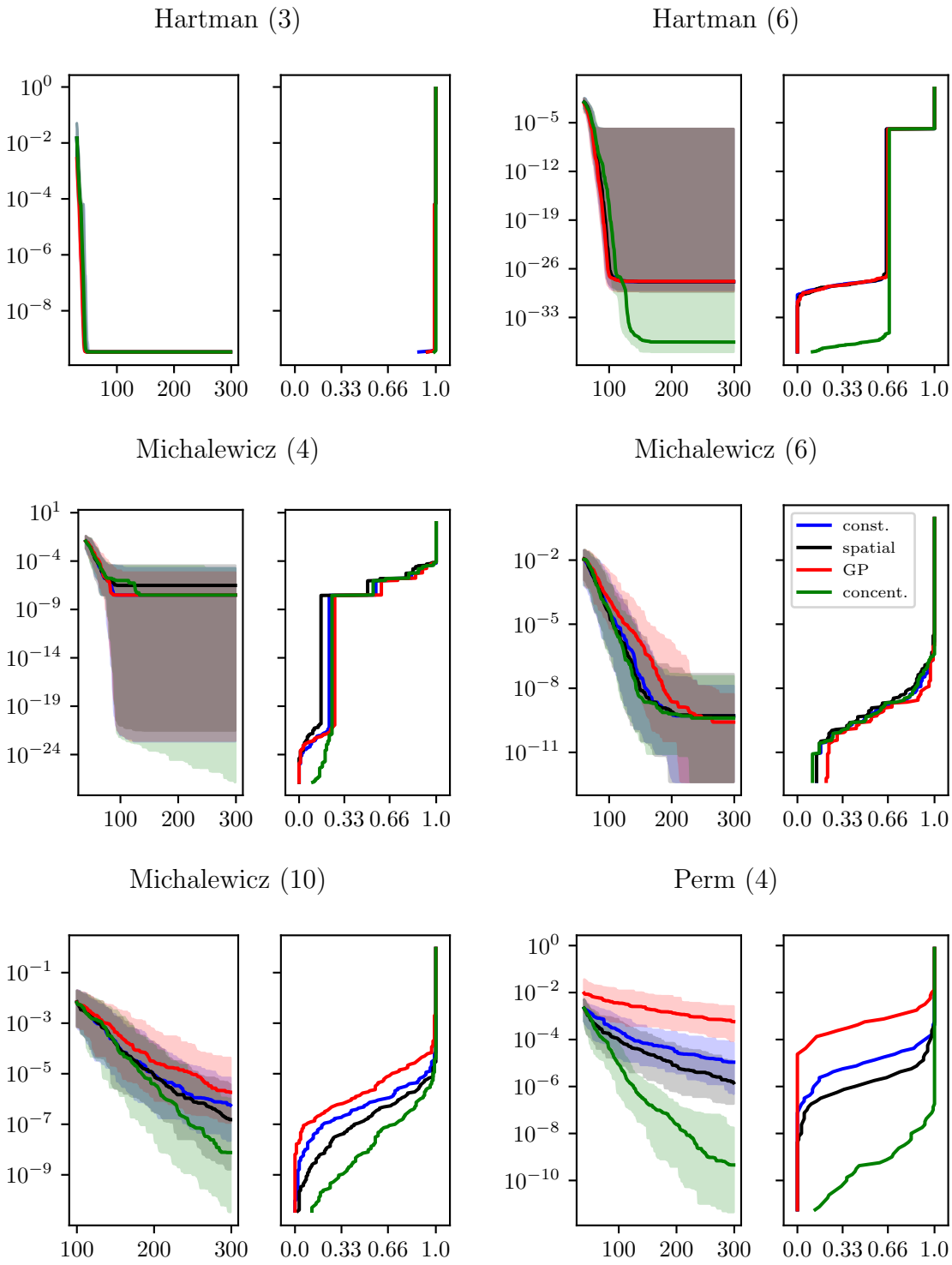

Figure 18: UCB results for a subset of the test functions from Table 1. Same legend as in Figure 16.

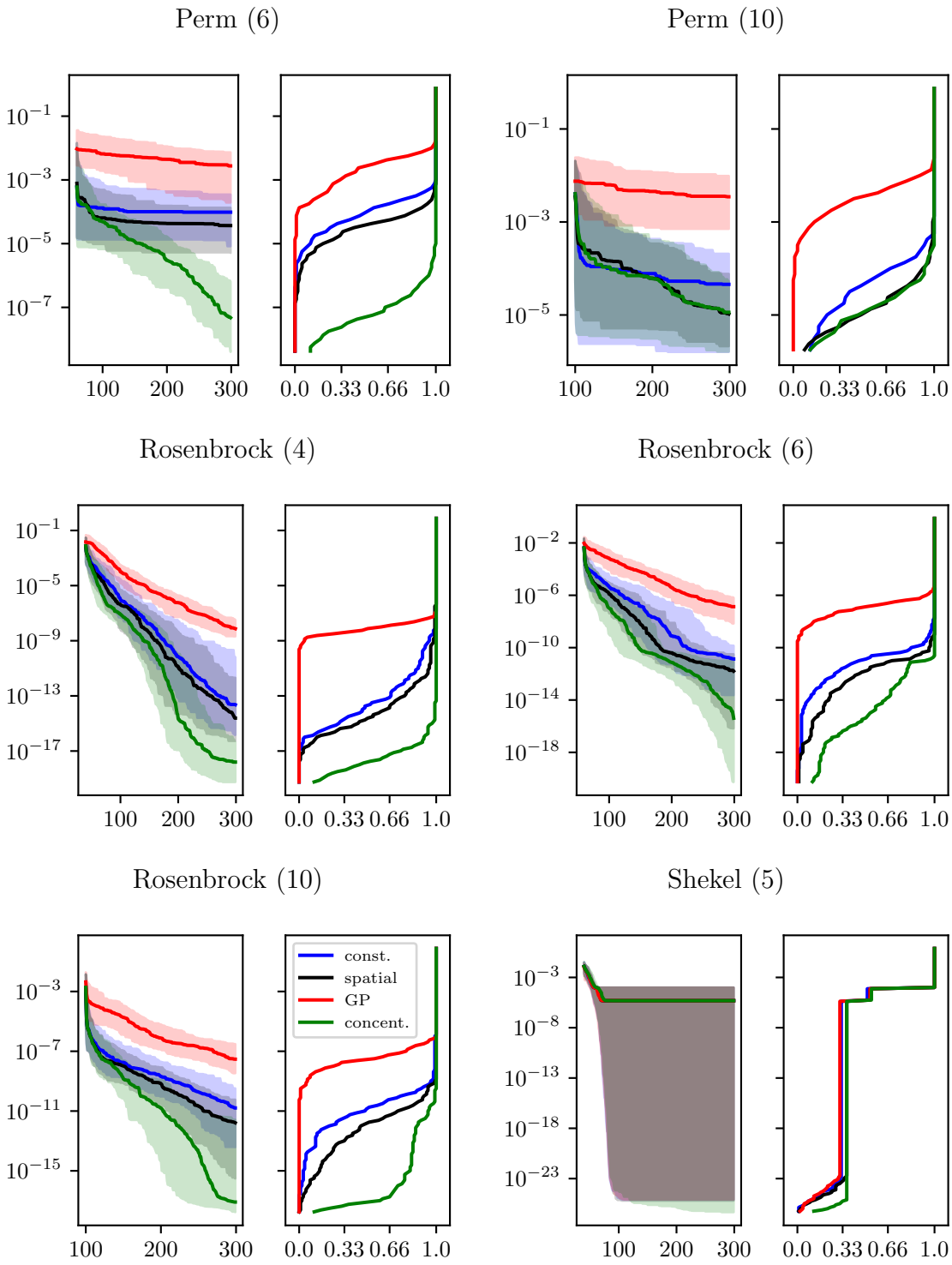

Figure 19: UCB results for a subset of the test functions from Table 1. Same legend as in Figure 16.

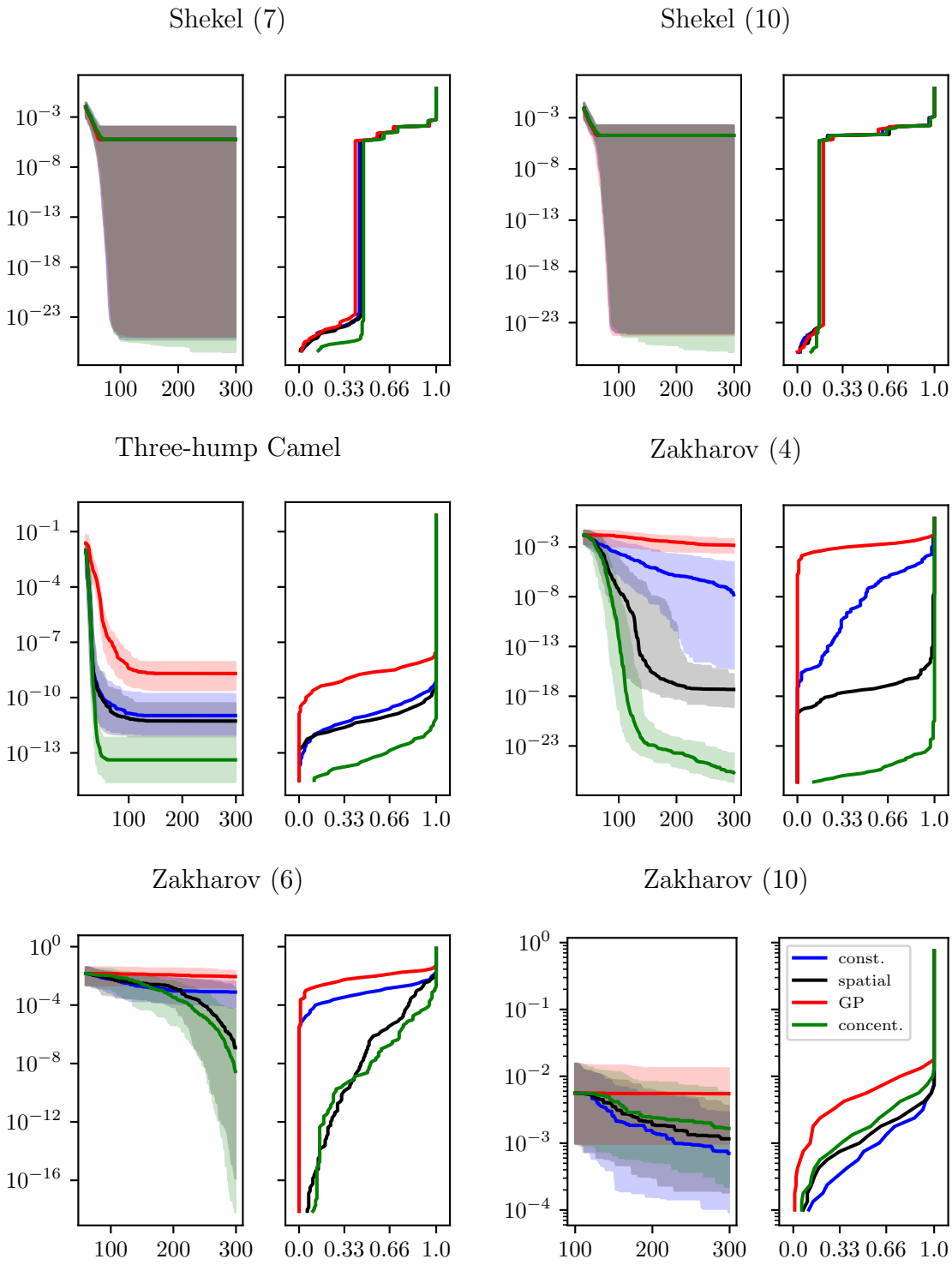

Figure 20: UCB results for a subset of the test functions from Table 1. Same legend as in Figure 16.

Appendix F. Application to the Estimation of an Excursion Set

Consider the problem of estimating the set $\{x \in \mathbb{X}, f(x) \leq u\}$, for some $u \in \mathbb{R}$, as illustrated in Section 4.4. When f is expensive to evaluate, a common approach is to use active learning with Gaussian process surrogates. A large variety of sampling criteria are available in the literature (see, e.g., Bryan et al., 2005; Bect et al., 2012; Chevalier et al., 2014; Bogunovic et al., 2016). To illustrate the benefits of reGP for the estimation of an excursion set, we choose the straddle heuristic of Bryan et al. (2005), which consists in choosing x_{n+1} maximizing

$$1.96\sigma_n(x) - |\mu_n(x) - u|, \quad x \in \mathbb{X}. \tag{40}$$

This simple (but effective) criterion rewards locations where the prediction is both close to u and uncertain.

As with the EI and UCB criteria, the coordination of reGP with the straddle heuristic simply involves plugging the reGP predictive distribution into the criterion (40). Consequently, with observations $(\underline{x}_n, \underline{z}_n)$ and using the framework described by Algorithm 1, we are left to specify a range of interest Q_n and a list $\mathbb{R} \setminus Q_n = R_n^{(0)} \supset \cdots \supset R_n^{(G-1)}$ of relaxation range candidates. As in Section 4.4, symmetric ranges of interest $Q_n = (u - t_n^{(0)}, u + t_n^{(0)})$ are considered. For $t_n^{(0)}$, the three heuristics proposed in Section 5.3.1 are adapted by considering α-quantiles of $|u - z_i|$. Then, given $t_n^{(0)}$, we propose to use $R_n^{(g)} = (-\infty, u - t_n^{(g)}] \cup [u + t_n^{(g)}, +\infty)$ with $t_n^{(g)}$ ranging logarithmically from $t_n^{(0)}$ to $\max_i |f(x_i) - u|$. As was done for optimization, we use the convention that $t_n^{(G-1)} > \max_i |f(x_i) - u|$ so that the G-th model is a vanilla GP.

As in Section 5.3.1, the experiments are conducted using Matern covariance functions with regularity $\nu = 5/2$ and constant mean functions. We also use $G = 10$ and initial designs of size $n_0 = 10d$. The criterion (40) is optimized using a subset simulation algorithm.

Four test cases are considered for the straddle heuristic in this work. First, the two functions c_6 and c_6^7 introduced in Section 4.4, with $u = 0$. Then, as in Section 5.3.1, we also consider the Goldstein-Price function, as well as its log-version. The vanilla Goldstein-Price function ranges from 3 to about 10^6. To depart from optimization, the intermediate value $u = 1000$ is chosen, which is a spatial quantile with a level of about 25.4%. The corresponding value $u = \log(1000)$ is also used for the log-version. For each of these four test cases, we ran $n_{\text{rep}} = 100$ (random) repetitions, each with a budget of $n_{\text{tot}} = 300$ evaluations.

As often done for the estimation of excursion sets, the performance of the sequential strategies is measured using model predictions. More specifically, writing P_n for a predictive distribution, we use a Sobol' sequence $\tilde{x}_1, \ldots, \tilde{x}_N$ to compute the estimate

$$\frac{1}{N} \sum_{i=1}^{N} \mathsf{P}_n\left(\xi(\tilde{x}_i) > u\right) \mathbb{1}_{f(\tilde{x}_i) \leq u} + \mathsf{P}_n\left(\xi(\tilde{x}_i) \leq u\right) \mathbb{1}_{f(\tilde{x}_i) > u} \tag{41}$$

of the (relative) expected volume of the symmetric difference between $\{x \in \mathbb{X}, f(x) \leq u\}$ and the random set $\{x \in \mathbb{X}, \xi(x) \leq u\}$.

Figure 21 shows the evolution of quantiles of (41) across the n_{rep} repetitions for the four test cases. Observe the large improvement brought by reGP on the c_6 test case. Using GPs, the evaluation metric (41) does not improve as n increases, which is consistent with the observations of Feliot et al. (2017). However, the metric drops using reGP. Using the constant heuristic, the metric eventually stagnates at a value substantially lower than that obtained with vanilla GPs, whereas it continues

to decrease using the concentration and spatial heuristics. Figure 7 illustrates the relaxation obtained with $n = 300$ observations after a run of the concentration heuristic.

The performances of all the methods are comparable on c_6^7. The straddle heuristic with vanilla GPs is more effective than on c_6. All reGP instances are a little behind, with slightly higher medians but overlapping colored areas (except at the very beginning). Curiously, the concentration and spatial heuristics are less effective on c_6^7 than on c_6 (while the level sets are the same, since $t \mapsto t^7$ is increasing).

The three reGP instances give very similar results on the Goldstein-Price test case, with a substantial improvement over vanilla GPs. Finally, for the Log-Goldstein-Price test case, the reGP instances begin by giving very similar results to those obtained with GPs. Then, from around $n = 150$, the performances of the concentration and spatial heuristics improve to become substantially better. Observe that this is also the case for a fraction of runs with the constant heuristic.

In conclusion, we observe that reGP is a useful modeling technique for the estimation of excursion sets with the straddle heuristic. It can give clear improvements in the best cases, while giving results comparable to those obtained with GPs in the worst cases.

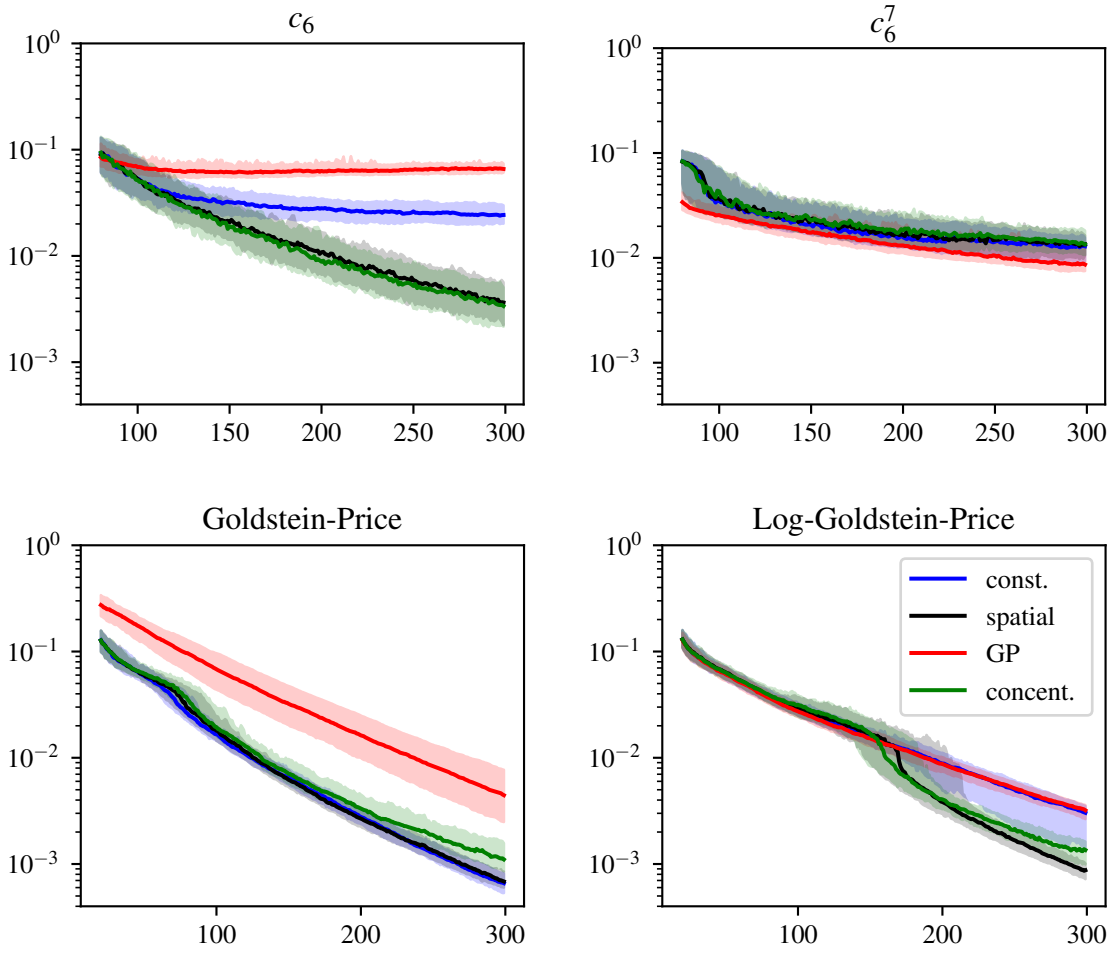

Figure 21: Evolution of quantiles of (41) against n. The solid lines represent the medians and, for each n, the shaded areas are deliminated by the 10% and 90% quantiles. The red color stands for the straddle heuristic with vanilla GPs and the blue, green, and black colors for straddle with reGP, using the constant, concentration, and spatial heuristics respectively.

Appendix G. Extension to the case of noisy observations

G.1 Gaussian process regression and kernel ridge regression

We now consider the case where observations are corrupted by additive noise, a common setting when dealing with stochastic simulators (Baker et al., 2020) or when optimizing validation loss in machine learning tasks with mini-batch training (Snoek et al., 2012).

Let $\underline{x}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$ be the input locations and let $\underline{Z}_n = (Z_1, \ldots, Z_n)^\mathsf{T} \in \mathbb{R}^n$ denote the noisy observations, modeled as

$$Z_i = f(x_i) + \varepsilon_i,$$

where $f\colon \mathbb{X} \to \mathbb{R}$ is the latent function of interest and $\varepsilon_i \sim \mathcal{N}(0, \eta^2)$ are independent Gaussian noise terms with common variance $\eta^2 > 0$.

We place a Gaussian process prior $\xi \sim \mathrm{GP}(\mu, k)$ on f, using the same notations as in Section 2. The posterior distribution of ξ given the observations $\underline{Z}_n$ remains Gaussian:

$$\xi \,|\, \underline{Z}_n \sim \mathrm{GP}\left(\mu_n, k_n\right), \tag{42}$$

where the posterior mean and covariance functions are

$$\begin{aligned} \mu_n(x) &= \mu(x) + k(x, \underline{x}_n)(K_n + \eta^2 I)^{-1}\left(\underline{Z}_n - \mu(\underline{x}_n)\right), \\ k_n(x, y) &= k(x, y) - k(x, \underline{x}_n)(K_n + \eta^2 I)^{-1} k(y, \underline{x}_n)^\mathsf{T}. \end{aligned}$$

Here, K_n is the $n \times n$ covariance matrix with entries $[K_n]_{ij} = k(x_i, x_j)$, and I is the $n \times n$ identity matrix.

The posterior mean in the zero-mean case ($\mu = 0$) coincides with the solution of a kernel ridge regression (KRR) problem. This correspondence is a direct consequence of the representer theorem, as formalized below.

Proposition 40 *(Kimeldorf and Wahba, 1970) The solution to the regularized least-squares problem*

$$\begin{cases} \textit{minimize} & \|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \displaystyle\sum_{i=1}^{n} \left(Z_i - h(x_i)\right)^2 \\ \textit{subject to} & h \in \mathscr{H}(\mathbb{X}) \end{cases}$$

is unique and given by

$$s^{(\eta)}_{\underline{Z}_n} = k(\cdot, \underline{x}_n)(K_n + \eta^2 I)^{-1} \underline{Z}_n.$$

Hence, when $\mu = 0$, the posterior mean μ_n coincides with the KRR solution $s^{(\eta)}_{\underline{Z}_n}$.

G.2 Relaxed Gaussian process regression

As seen in Proposition 40, the posterior mean of a Gaussian process model does not interpolate noisy data. How can the goal-oriented modeling framework introduced in Section 3 be extended to this setting?

A natural extension is to adapt Definition 6 by optimizing the likelihood with respect to noisy observations falling in the relaxation range.

Assume that both the mean and the covariance functions belong to parametric families, indexed by $\theta \in \Theta$. Fix the noise variance $\eta^2 > 0$ (see Appendix G.3 for an estimator). Let $R = \bigcup_{j=1}^{J} R_j$

be a union of disjoint closed intervals and define the corresponding constraint set $C_{R,n} \subset \mathbb{R}^n$ as in Section 3.1.

We define the relaxed GP predictive distribution as the posterior distribution (42) of the GP conditioned on $\underline{Z}_n = \underline{z}_n^\star$ and on the estimated parameter $\widehat{\theta}_n$, where

$$(\widehat{\theta}_n, \underline{z}_n^\star) \;=\; \operatorname*{arg\,min}_{\theta\in\Theta,\, \underline{z}\in C_{R,n}} \mathscr{L}(\theta; \underline{z}),$$

and $\mathscr{L}$ is the (negative) log-likelihood of the observations written as

$$\mathscr{L}(\theta;\underline{z}) \propto \log\det(K_n + \eta^2 I) + (\underline{z} - \mu(\underline{x}_n))^\mathsf{T} (K_n + \eta^2 I)^{-1} (\underline{z} - \mu(\underline{x}_n)) + \text{constant}.$$

Note that we recover Definition 6 when $\eta^2 = 0$.

When the mean function is zero and the covariance function k is fixed, the resulting procedure can be formulated as a regularized regression problem:

Proposition 41 *Let $\mathscr{H}(\mathbb{X})$ be the RKHS associated with k, and let $\eta^2 > 0$ be fixed. Consider the relaxed regularized least-squares problem*

$$\begin{cases} \textit{minimize} & \|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \displaystyle\sum_{i=1}^n e(Z_i, h(x_i); R), \\ \textit{subject to} & h \in \mathscr{H}(\mathbb{X}), \end{cases} \tag{43}$$

where the relaxed squared error $e(y_1, y_2; R)$ is defined for all $y_1, y_2 \in \mathbb{R}$ by

$$e(y_1, y_2; R) = \begin{cases} \inf_{r\in R_j} (y_2 - r)^2 & \textit{if } y_1 \in R_j \textit{ for some } j, \\ (y_2 - y_1)^2 & \textit{otherwise}. \end{cases}$$

Then, the unique solution of (43) *is $s^{(\eta)}_{\underline{z}_n^\star}$, where $\underline{z}_n^\star$ is the unique solution of*

$$\operatorname*{arg\,min}_{\underline{z}\in C_{R,n}} \underline{z}^\mathsf{T} (K_n + \eta^2 I)^{-1} \underline{z}. \tag{44}$$

Proof Write $\Delta = \eta^2 I$. For $\underline{z} = (z_1, \ldots, z_n)^\mathsf{T} \in \mathbb{R}^n$, straightforward calculations (using the identity $I - K_n(K_n + \Delta)^{-1} = \Delta\,(K_n + \Delta)^{-1}$) show that

$$\|s^{(\eta)}_{\underline{z}}\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n \left(z_i - s^{(\eta)}_{\underline{z}}(x_i) \right)^2 = \underline{z}^\mathsf{T} \left(K_n + \Delta \right)^{-1} \underline{z}.$$

The quadratic problem (44) has a unique solution $\underline{z}_n^\star$ since the matrix is positive-definite and the feasible set is closed and convex. Write $L^\star$ for the corresponding minimum value.

For $h \in \mathscr{H}(\mathbb{X})$, the definition of e yields

$$\|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n e\left(Z_i, h(x_i); R\right) = \inf_{\underline{z}=(z_1,\ldots,z_n)^\mathsf{T}\in C_{R,n}} \|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n \left(z_i - h(x_i)\right)^2.$$

The infimum is reached by a $\underline{z}(h) = (z_1(h), \ldots, z_n(h))^\mathsf{T} \in C_{R,n}$. We then have

$$\begin{aligned}
\|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n e\left(Z_i, h(x_i); R\right) &= \|h\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n \left(z_i(h) - h(x_i)\right)^2 \\
&\geq \|s^{(\eta)}_{\underline{z}(h)}\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n \left(z_i(h) - s^{(\eta)}_{\underline{z}(h)}(x_i)\right)^2 \\
&= \underline{z}(h)^\mathsf{T} \left(K_n + \Delta\right)^{-1} \underline{z}(h) \\
&\geq L^\star.
\end{aligned}$$

If the previous inequalities are equalities, then $h = s^{(\eta)}_{\underline{z}^\star_n}$ because of the unicity of the solution of (44) and by Proposition 40.

Finally, writing $\underline{z}^\star_n = (z^\star_1, \ldots, z^\star_n)$, we have

$$\|s^{(\eta)}_{\underline{z}^\star_n}\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n e\left(Z_i, s^{(\eta)}_{\underline{z}^\star_n}(x_i); R\right) \leq \|s^{(\eta)}_{\underline{z}^\star_n}\|^2_{\mathscr{H}(\mathbb{X})} + \eta^{-2} \sum_{i=1}^n \left(z^\star_i - s^{(\eta)}_{\underline{z}^\star_n}(x_i)\right)^2 = L^\star.$$

■

This shows that, in the fixed-kernel case, the relaxed GP regression procedure corresponds to a modified KRR problem: observations Z_i outside the relaxation range R are penalized by squared prediction error, while observations inside R_j are only penalized by the squared distance from the prediction $h(x_i)$ to the interval R_j.

The resulting predictive distribution thus reflects a form of soft constraint on the latent function: fidelity to the data is enforced outside R, while inside R the model is only constrained to remain consistent with the relaxation range.

G.3 Application to Bayesian optimization with the UCB criterion

We now illustrate the use of reGP in Bayesian optimization with noisy observations. We adopt the same UCB sampling strategy as described in Appendix E, using a fixed β parameter. Importantly, we do not inflate the posterior variance with the estimated noise variance η^2; the UCB criterion is applied to the predictive distribution of the latent process ξ.

Two test functions are considered: the Goldstein-Price function and its logarithmic transformation. In both cases, Gaussian noise is added to the evaluations. For the Goldstein-Price function, we use $\eta = 100$ (standard deviation), which is large relative to the local variations of the function near its minimum. Specifically, while the function ranges from 3 to approximately 10^6, the 25% and 5% spatial quantiles are around 1000 and 100, respectively. The noise level is thus significant near the global minimum, whereas extreme values act as outliers. This setting is favorable to the reGP strategy, which focuses on accurate modeling in target regions. For the Log-Goldstein-Price function (value range: 1.10 to 13.83), we set $\eta = 3$.

In both cases, the noise variance η^2 is assumed to be unknown and must be estimated from the data. Jointly estimating η^2, the model parameters, and the relaxed values by maximizing the likelihood is not advisable in the reGP framework, as the relaxed values tend to shrink toward the posterior mean function, potentially leading to a biased underestimation of the noise level. Instead,

we estimate η^2 using standard GP maximum likelihood, restricted to observations within the range of interest, and treat the result as a fixed value in the relaxed procedure.

For reGP, the relaxation threshold is selected from a predefined list using the LOO-tCRPS criterion, as described in Section 5.3.1. The spatial heuristic is adapted by setting the validation threshold $t_n^{(0)}$ to a quantile of the previous model's predictions at the previous design $\underline{x}_{n-1}$. Further adjustments are likely needed for the concentration heuristic.

We run $n_{\text{rep}} = 100$ independent repetitions with random initial designs of size $n_0 = 10d$ and total budgets of $n_{\text{tot}} = 300$ evaluations. Results are reported in Figure 22. We define $x_n^\star = \arg\min \mu_n$ as the predicted minimizer at step n, and use $f(x_n^\star)$ as an oracle benchmark.

On the Goldstein-Price function, reGP yields substantially lower median noisy evaluations up to approximately $n = 150$. This is confirmed by the superior oracle values $f(x_n^\star)$ obtained with reGP during the same phase. After that point, the difference between noisy values and the global minimum becomes comparable to the noise level, and the two approaches yield similar performance. Notably, the estimation error $|f(x_n^\star) - \mu_n(x_n^\star)|$ remains consistently smaller with reGP across all steps.

On the Log-Goldstein-Price function, both methods perform similarly overall. However, standard GP models show higher 90% quantiles for $f(x_n^\star)$ and $f(x_n) + \varepsilon_n$, indicating more variability in the predicted minima.

These results illustrate the robustness of reGP in noisy optimization tasks.

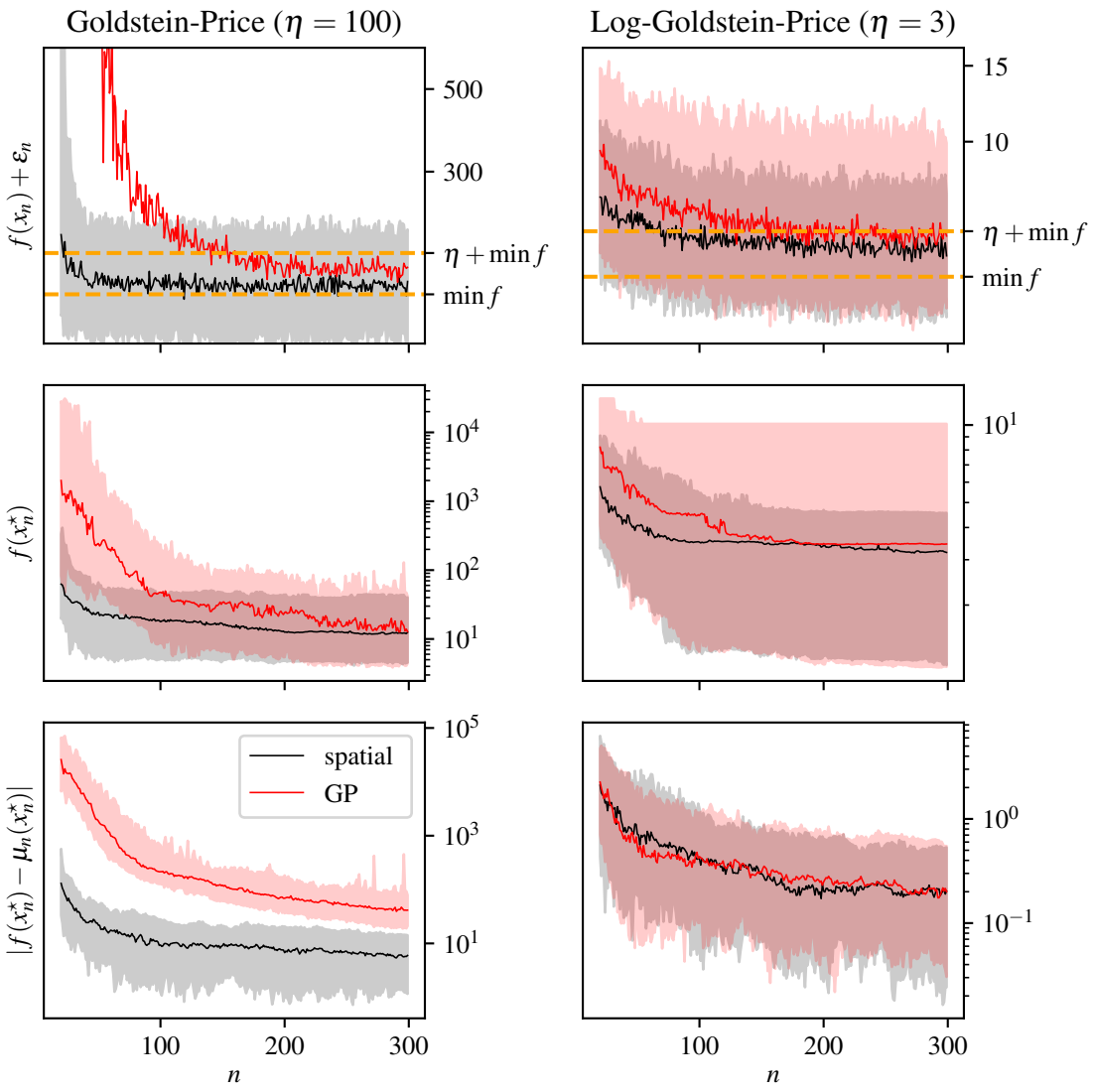

Figure 22: Bayesian optimization with noisy evaluations: performance of the UCB strategy. Left: Goldstein-Price function. Right: Log-Goldstein-Price function. Each plot shows medians (solid lines) and 10%/90% quantiles (shaded areas) as functions of the number of evaluations n. Top: noisy evaluations $f(x_n) + \varepsilon_n$. Middle: oracle values $f(x_n^\star)$ at predicted minimizers $x_n^\star = \arg\min \mu_n$. Bottom: absolute estimation errors $|f(x_n^\star) - \mu_n(x_n^\star)|$.

References

R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Elsevier, 2003.

R. Arcangéli, M. C. López de Silanes, and J. J. Torrens. An extension of a bound for functions in Sobolev spaces, with applications to (m, s)-spline interpolation and smoothing. *Numerische Mathematik*, 107(2):181–211, 2007.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

S.-K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic engineering mechanics*, 16(4):263–277, 2001.

S. Ba and V. R. Joseph. Composite Gaussian process models for emulating expensive functions. *The Annals of Applied Statistics*, 6(4):1838 – 1860, 2012.

F. Bachoc and A. Lagnoux. Posterior contraction rates for constrained deep Gaussian processes in density estimation and classification. *Communications in Statistics-Theory and Methods*, 54(3): 774–811, 2025.

E. Baker, P. Barbillon, A. Fadikar, R. B. Gramacy, R. Herbei, D. Higdon, J. Huang, L. R. Johnson, A. Mondal, B. Pires, J. Sacks, and V. Sokolov. Stochastic simulators: An overview with opportunities. *arXiv preprint arXiv:2002.01321*, 2020.

J. Y. Bao, F. Ye, and Y. Yang. Screening effect in isotropic Gaussian processes. *Acta Mathematica Sinica*, 36(5), 2020.

J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22:773–793, 2012.

R. Benassi, J. Bect, and E. Vazquez. Bayesian optimization using sequential Monte Carlo. In *Learning and Intelligent Optimization – 6th International Conference, LION 6*, pages 339–342, Paris, France, January 16-20 2012. Springer.

A. Benavoli, D. Azzimonti, and D. Piga. A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with Skew Gaussian Processes. *Machine Learning*, 110(11):3095–3133, 2021.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2004.

M. J. Best and B. Ding. On the continuity of the minimum in parametric quadratic programs. *Journal of Optimization Theory and Applications*, 86(1):245–250, 1995.

E. Bodin, M. Kaiser, I. Kazlauskaite, Z. Dai, N. Campbell, and C. H. Ek. Modulating surrogates for Bayesian optimization. In *International Conference on Machine Learning*, pages 970–979. PMLR, 2020.

I. Bogunovic, J. Scarlett, A. Krause, and V. Cevher. Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(1):125–148, 2017.

H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, volume 2. Springer, 2011.

B. Bryan, R. C. Nichol, C. R. Genovese, J. Schneider, C. J. Miller, and L. Wasserman. Active learning for identifying function threshold boundaries. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

C. Chevalier and D. Ginsbourger. Fast computation of the multi-points expected improvement with applications in batch selection. In *International Conference on Learning and Intelligent Optimization*, pages 59–69. Springer, 2013.

C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(7), 2005.

D. D. Cox and S. John. A statistical method for global optimization. In *IEEE International Conference on Systems, Man, and Cybernetics*, pages 1241–1246. IEEE, 1992.

S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 21, pages 529–555, 2012.

A. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR, 2013.

E. Di Nezza, G. Palatucci, and E. Valdinoci. Hitchhiker's guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathématiques*, 136(5):521–573, 2012.

O. Dubrule. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15:687–699, 1983.

M. M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup. How deep are deep Gaussian processes? *Journal of Machine Learning Research*, 19(54):1–46, 2018.

D. Eriksson, M. Pearce, J. R. Gardner, R. C. Turner, and M. Poloczek. Scalable global optimization via local Bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

P. Feliot, J. Bect, and E. Vazquez. A Bayesian approach to constrained single-and multi-objective optimization. *Journal of Global Optimization*, 67(1-2):97–133, 2017.

P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

A. Gessner, O. Kanjilal, and P. Hennig. Integrals over Gaussians under linear domain constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2764–2774. PMLR, 2020.

M. N. Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, Citeseer, 1998.

T. G. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

T. G. Gneiting and R. Ranjan. Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3):411–422, 2011.

T. G. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.

R. B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, 1985.

A. Hebbal, L. Brevault, and N. Melab. Bayesian optimization using deep Gaussian processes with applications to aerospace system design. *Optimization and Engineering*, 22(1):321–361, 2021.

D. Higdon. A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5:173–190, 1998.

D. Higdon. Space and space-time modeling using process convolutions. *Quantitative Methods for Current Environmental Issues*, 04 2002.

A. Iske. *Approximation Theory and Algorithms for Data Analysis*. Springer, 2018.

K. Jakkala. Deep Gaussian processes: A survey. *arXiv preprint arXiv:2106.12135*, 2021.

D. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 12 1998.

T. Karvonen. Error bounds and the asymptotic setting in kernel-based approximation. *Dolomites Research Notes on Approximation*, 15(3):65–77, 2022.

T. Karvonen. Asymptotic bounds for smoothness parameter estimates in Gaussian process interpolation. *SIAM/ASA Journal on Uncertainty Quantification*, 11(4):1225–1257, 2023.

T. Karvonen, G. Wynne, F. Tronarp, C. Oates, and S. Sarkka. Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958, 2020.

G. S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495 – 502, 1970.

S. Lacoste–Julien, F. Huszár, and Z. Ghahramani. Approximate inference for the loss-calibrated Bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pages 416–424. PMLR, 11–13 Apr 2011.

M. Lázaro-Gredilla. Bayesian warped Gaussian processes. *Advances in Neural Information Processing Systems*, 25:1619–1627, 2012.

S. Lerch and T. L. Thorarinsdottir. Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography*, 65(1):21206, 2013.

J. R. Lewis, S. N. MacEachern, and Y. Lee. Bayesian restricted likelihood methods: Conditioning on insufficient statistics in Bayesian regression. *Bayesian Analysis*, 1(1), 2021.

A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6 (3):1224–1255, 2018.

M. Lukić and J. Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

S. Marmin, D. Ginsbourger, J. Baccou, and J. Liandrat. Warped Gaussian processes and derivative-based sequential designs for functions with heterogeneous variations. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):991–1018, 2018.

G. Matheron. The theory of regionalized variables and its applications. Technical Report Les cahiers du CMM de Fontainebleau, Fasc. 5, Ecole des Mines de Paris, 1971.

J. E. Matheson and R. L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

M. McIntire, D. Ratner, and S. Ermon. Sparse Gaussian processes for Bayesian optimization. In *UAI*, 2016.

M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1): 55–61, 2000.

E. Meeds and S. Osindero. An alternative infinite mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 883–890. MIT Press, 2006.

E. Merrill, A. Fern, X. Z. Fern, and N. Dolatnia. An empirical study of Bayesian optimization: Acquisition versus partition. *The Journal of Machine Learning Research*, 22:4–1, 2021.

J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. *Towards Global Optimization*, 2:117–129, 1978.

M. J. Morais and J. W. Pillow. Loss-calibrated expectation propagation for approximate Bayesian decision-making. *arXiv preprint arXiv:2201.03128*, 2022.

H. B. Moss, S. W. Ober, and V. Picheny. Inducing point allocation for sparse Gaussian processes in high-throughput Bayesian optimisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5213–5230. PMLR, 2023.

S. Nadarajah and S. Kotz. Exact distribution of the max/min of two Gaussian random variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(2):210–212, 2008.

F. J. Narcowich, J. D. Ward, and H. Wendland. Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions. *Constructive Approximation*, 24(2): 175–186, 2006.

J. Nocedal and S. J. Wright. Quadratic programming. *Numerical Optimization*, pages 448–492, 2006.

D. J. Nott and William T. M. Dunsmuir. Estimation of nonstationary spatial covariance structure. *Biometrika*, 89(4):819–829, 2002.

C. Park and D. Apley. Patchwork kriging for large-scale Gaussian process regression. *The Journal of Machine Learning Research*, 19(1):269–311, 2018.

S. J. Petit. An asymptotic study of the joint maximum likelihood estimation of the regularity and the amplitude parameters of a periodized Matérn model. *Electronic Journal of Statistics*, 19(1): 2052–2094, 2025.

S. J. Petit, J. Bect, P. Feliot, and E. Vazquez. Parameter selection in Gaussian process interpolation: an empirical study of selection criteria. *SIAM/ASA Journal on Uncertainty Quantification*, 11(4): 1308–1328, 2023.

V. Picheny, T. Wagner, and D. Ginsbourger. A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization*, 48:607–626, 2013.

V. Picheny, S. Vakili, and A. Artemev. Ordinal Bayesian optimisation. *arXiv preprint arXiv:1912.02493*, 2019.

L. Pronzato and M. J. Rendas. Bayesian local kriging. *Technometrics*, 59(3):293–304, 2017.

C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. *Advances in Neural Information Processing Systems*, 2:881–888, 2002.

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

R. G. Regis. Constrained optimization by radial basis function interpolation for high-dimensional expensive black-box problems with infeasible initial points. *Engineering Optimization*, 46(2): 218–243, 2014.

J. Rivoirard and T. Romary. Continuity for kriging with moving neighborhood. *Mathematical Geosciences*, 43(4):469–481, 2011.

I. Rychlik, P. Johannesson, and M. R. Leadbetter. Modelling and statistical analysis of ocean-wave data using transformed Gaussian processes. *Marine Structures*, 10(1):13–47, 1997.

M. Scheurer, R. Schaback, and M. Schlather. Interpolation of spatial data – a stochastic or a deterministic problem? *European Journal of Applied Mathematics*, 24(4):601–629, 2013.

K. Schittkowski. The nonlinear programming method of Wilson, Han, and Powell with an augmented Lagrangian type line search function: Part 1: Convergence analysis. *Numerische Mathematik*, 38:83–114, 1982.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped Gaussian processes. *Advances in Neural Information Processing Systems*, 16:337–344, 2004.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pages 1015–1022, 2010.

M. L. Stein. Nonstationary spatial covariance functions. *Unpublished technical report*, 2005.

M. L. Stein. 2010 Rietz lecture: When does the screening effect hold? *The Annals of Statistics*, 39 (6):2795–2819, 2011.

M.L. Stein. *Interpolation of Spatial Data: Some Theory for kriging*. Springer Series in Statistics. Springer New York, 1999.

I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

R. S. Strichartz. Multipliers on fractional Sobolev spaces. *Journal of Mathematics and Mechanics*, 16(9):1031–1060, 1967.

S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets. Retrieved October 13, 2020, from `http://www.sfu.ca/~ssurjano/branin.html`, 2013.

S. Takeno, M. Nomura, and M. Karasuyama. Towards practical preferential Bayesian optimization with skew Gaussian processes. In *International Conference on Machine Learning*, pages 33516–33533. PMLR, 2023.

V. Tresp. Mixtures of Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 654–660, 2001.

E. Vazquez and J. Bect. Pointwise consistency of the kriging predictor with known mean and covariance functions. In *mODa 9–Advances in Model-Oriented Design and Analysis*, pages 221–228. Springer, 2010a.

E. Vazquez and J. Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and Inference*, 140(11):3088–3095, 2010b.

E. Vazquez and J. Bect. A new integral loss function for Bayesian optimization. *arXiv preprint arXiv:1408.4622*, 2014.

J. Ver Hoef, N. Cressie, and R. Barry. Flexible spatial models for kriging and cokriging using moving averages and the fast Fourier transform (FFT). *Journal of Computational and Graphical Statistics*, 13:265–282, 2004.

J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia, 1990.

W. Wang and B.-Y. Jing. Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23(193):1–67, 2022.

H. Wendland. *Scattered Data Approximation*, volume 17. Cambridge university press, 2004.

G. Wynne, F. X. Briol, and M. Girolami. Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22(123):1–40, 2021.

Y. Xiong, W. Chen, D. Apley, and X. Ding. A non-stationary covariance-based kriging method for metamodelling in engineering design. *International Journal for Numerical Methods in Engineering*, 71(6):733–756, 2007.

A. Yang, C. Li, S. Rana, S. Gupta, and S. Venkatesh. Sparse spectrum Gaussian process for Bayesian optimization. In *Advances in Knowledge Discovery and Data Mining*, pages 257–268. Springer, 2021.

Y. Yang and J. Ma. An efficient EM approach to parameter learning of the mixture of Gaussian processes. In *International Symposium on Neural Networks*, pages 165–174. Springer, 2011.

C. Yuan and C. Neubauer. Variational mixture of Gaussian process experts. In *Advances in Neural Information Processing Systems*, pages 1897–1904, 2009.

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012.

H. Zhang and Y. Wang. Kriging and cross-validation for massive spatial data. *Environmetrics*, 21 (3–4):290–304, 2010.