
Modulating Surrogates for Bayesian Optimization

Erik Bodin¹ Markus Kaiser^{2,3} Ieva Kazlauskaitė⁴ Zhenwen Dai⁵ Neill D. F. Campbell⁴ Carl Henrik Ek⁶

Abstract

Bayesian optimization (BO) methods often rely on the assumption that the objective function is well-behaved, but in practice, this is seldom true for real-world objectives even if noise-free observations can be collected. Common approaches, which try to model the objective as precisely as possible, often fail to make progress by spending too many evaluations modeling irrelevant details. We address this issue by proposing surrogate models that focus on the well-behaved structure in the objective function, which is informative for search, while ignoring detrimental structure that is challenging to model from few observations. First, we demonstrate that surrogate models with appropriate noise distributions can absorb challenging structures in the objective function by treating them as irreducible uncertainty. Secondly, we show that a latent Gaussian process is an excellent surrogate for this purpose, comparing with Gaussian processes with standard noise distributions. We perform numerous experiments on a range of BO benchmarks and find that our approach improves reliability and performance when faced with challenging objective functions.

1. Introduction

Bayesian optimization (BO) (Snoek et al., 2012) is a method for finding the optimum of functions that are unknown and expensive to evaluate. By fitting a surrogate model to the samples of an unknown objective, the BO procedure iteratively picks the new samples of the objective believed to be the most informative about where the optimum is located.

Model misspecification has significant negative implications

¹University of Bristol, United Kingdom ²Siemens AG, Germany ³Technical University of Munich, Germany ⁴University of Bath, United Kingdom ⁵Spotify Research, United Kingdom ⁶University of Cambridge, United Kingdom. Correspondence to: Erik Bodin <erik.bodin@bristol.ac.uk>.

for any machine learning tasks. This is especially true for sequential decision making tasks such as BO, where the model is used not only to locate the optimum based on the collected data but also to decide where to collect data for future decisions. If the surrogate model is misspecified, it is likely to acquire samples that are less informative about the optimum, which will lead to a less efficient optimization. Therefore the quality of the surrogate model is essential to achieve both efficient and reliable results.

Many works have been done towards avoiding model misspecification in the surrogate model for BO, such as handling non-stationary objective functions with warpings (Snoek et al., 2014), tree-structured dependencies in the search space (Jenatton et al., 2017), and searching the optimum from piecewise comparisons (González et al., 2017). Comparing with the Gaussian process (GP) regression model in the standard BO setting, these methods avoid model misspecification in real-world problems by using more sophisticated surrogate models that are suitable for the corresponding problems. Bayesian inference with more sophisticated surrogate models will often require additional data to reduce uncertainty and confirm beliefs, because it considers more possibilities. Importantly the ultimate goal of BO is to find the optimum, not to model the unknown objective as precisely as possible. In practice, this means that using a surrogate with high complexity might perform worse compared to a simpler class even if the former contains the true objective function.

Instead of building a complex surrogate model with minimal model misspecification, we propose an alternative approach which allows trading off accuracy in modeling the objective with efficiency of capturing informative structures from small amounts of data. For example, we observe that structures such as local oscillations and discontinuities are less important to capture for the purposes of BO. Such details often require a lot of data to be closely captured in a surrogate model but do not help the search for the optimum, unless the search reaches the last stage of pinpointing the exact location of the optimum. To ignore these details, we associate an independent random *input* variable with every evaluation of the unknown function. As the random variables associated with new evaluations are conditionally independent of the posterior random variables associated with observed data given the function, this is referred to as *irreducible*

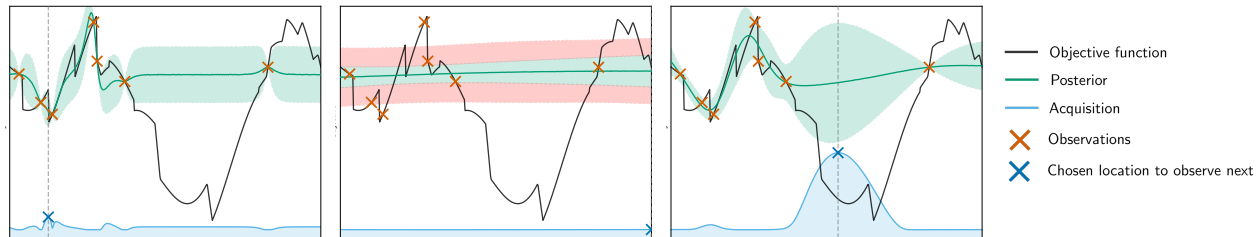


Figure 1. An illustrative example of the posterior surrogate function density obtained given observations of a 1D nonsmooth function using a noise-free GP, a GP with homoscedastic Gaussian noise, and using a LGP model. The posterior belief for the noise-free and homoscedastic GP surrogates results in the EI acquisition function, shown in blue, making poorly informed decisions for the next query. In contrast, the LGP using our proposed setup is able to reduce the influence of the rapid oscillations that do not match the GP prior by explaining part of the variation using the latent input. As a result, the acquisition function can utilize a confidently discovered global trend to increase the efficiency of the search. In this example, σ_h is set to $\frac{1}{50}$ of the domain range to allow ignorance of oscillations at that scale.

uncertainty. Such variables are similar to the noise variables in regression models, which are used to capture measurement noise and the data variance that cannot be attributed to the input variables. In contrast to noise variables for noisy outcomes, where there is irreducible uncertainty about the data, there is now irreducible uncertainty in the model of the function.

We propose to use the surrogate models that are specified over well-behaved approximations of the objectives, which can be more useful for the search of the optimum (see Figure 1), augmented with flexible “noise” distributions to treat the nuisance parameters. We will demonstrate that, using the same function approximation, a surrogate model with a more flexible nuisance parameter distribution is more robust against challenging structures. In this paper we focus on noise-free objectives with complicated, oscillatory or discontinuous structures. In particular, we propose to use a Latent Gaussian process (LGP) (Pfingsten et al., 2006; Wang & Neal, 2012; Yousefi et al., 2016; Bodin et al., 2017) as the surrogate model due to its flexible nuisance parameter distribution and show that it outperforms the surrogate models with less flexible distributions such as GPs with additive likelihoods. LGP allows us to disentangle the complicated structures a GP surrogate struggles to model while highlighting important structures.

Our main contributions are:

- We propose to address challenging objective functions for BO by using a distribution in the surrogate model to explain structure that is challenging to model with few observations.
- We propose to use latent Gaussian processes (LGP) as surrogate models, which support non-stationary and non-Gaussian residuals.
- With experiments on multiple BO benchmarks, we show that our method significantly outperforms exist-

ing approaches.

2. Modulating Surrogates

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be an unknown, noise-free objective function defined on a bounded subset $\mathcal{X} \subset \mathbb{R}^Q$. The goal of BO is to solve the global optimization problem of finding

$$\mathbf{x}_{\min} = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}). \quad (1)$$

In real world problems, the objective function is often not a well-behaved function and a suitable model is difficult to specify. Instead of applying an automated model selection method (Malkomes & Garnett, 2018), we propose to model only the essential structure of the objective function that is well-behaved and leave the rest of the function details to be absorbed in a noise distribution.

We consider the family of objective functions f that can be represented as a composition of a well-behaved function and another arbitrary, latent function capturing the challenging details, i.e.

$$f(\mathbf{x}) := g(\mathbf{x}, \mathbf{h}), \quad \mathbf{h} := h(\mathbf{x}), \quad (2)$$

where g is a well-behaved function that can be nicely modeled by a surrogate model of choice, which is a Gaussian process (GP) in this paper, and the vector-valued function $h(\mathbf{x})$ encodes the structures which the surrogate model struggles to capture. In general this composition allows for complicated interactions between \mathbf{x} and \mathbf{h} , producing complicated realizations of the function which is observed through data. A simple, special case of a function composition is additive structure¹, i.e. $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$.

Instead of modeling $h(\mathbf{x})$ as part of the surrogate model, we propose to *ignore* the structure of the objective function in $h(\mathbf{x})$ by replacing $h(\mathbf{x})$ with a random variable \mathbf{h} per

¹Note that in the additive case, $h(\mathbf{x})$ must match the output in shape, i.e. be one-dimensional.

data point. The random variables \mathbf{h} for different data points are independent among each other. The objective function becomes a function of two variables $g(\mathbf{x}, \mathbf{h})$, in which \mathbf{h} is a random variable which explain the data variance that cannot be explained by \mathbf{x} . In this paper, we use a normal distribution for the prior of \mathbf{h} , $\mathbf{h} \sim \mathcal{N}(0, \mathbb{I})$. Note that, although the distribution of $h(\mathbf{x})$ induced by the data distribution for \mathbf{x} may not be zero-mean and unit-variance, it is easy to reformulate it as a linear transformation of a normal distribution with zero-mean and unit-variance and the resulting linear transformation can be absorbed into the function g . For further details on the definition, see the supplement.

With the above formulation, a BO method can be developed by constructing a model of the well-behaved function g and a model of h . At each step of the BO optimization, a set of input and output pairs of the objective function has been collected, denoted as $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$ and $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_N)^\top$. The output \mathbf{F} denotes the noise-free observations of the objective function. The Bayesian inference of the model aims at inferring the posterior distribution

$$p(\mathbf{H}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{F}) \propto p(\mathbf{F} | \mathbf{X}, \mathbf{H}, \boldsymbol{\theta}) p(\mathbf{H}) p(\boldsymbol{\theta}) \quad (3)$$

where $\boldsymbol{\theta}$ are the hyperparameters of the surrogate model and $\mathbf{H} = (\mathbf{h}_1, \dots, \mathbf{h}_N)^\top$ is the concatenation of the nuisance parameters associated with the individual data points. The location of the next evaluation is determined according to an acquisition function, which uses the predictive distribution $p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{F})$ of the surrogate model,

$$p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{F}) = \int p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{h}_*, \mathbf{X}, \mathbf{F}, \mathbf{H}, \boldsymbol{\theta}) p(\mathbf{H}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{F}) p(\mathbf{h}_*) d\mathbf{H} d\boldsymbol{\theta} d\mathbf{h}_*, \quad (4)$$

where \mathbf{x}_* is the input of the prediction and \mathbf{f}_* is the noise-free observation at the location \mathbf{x}_* . The predictive distribution of the latent variable $p(\mathbf{h}_*)$ associated with new evaluations is as of the i.i.d. assumption equal to the prior. As such $p(\mathbf{h}_*)$ contains model uncertainty *irreducible* by the active sampling loop, which we suggest to ignore via augmentation, see the supplement for details.

With the predictive distribution Eq. 4, the expectation of the acquisition function is derived as

$$\alpha(\mathbf{x}_*) = \int \mathbb{U}(\mathbf{f}_*, \mathbf{x}_*, \mathbf{X}, \mathbf{F}) p(\mathbf{f}_* | \mathbf{x}_*, \mathbf{X}, \mathbf{F}) d\mathbf{f}_*, \quad (5)$$

where the acquisition function of choice is denoted \mathbb{U} . Note that the predictive distribution due to the marginalization over \mathbf{H} and $\boldsymbol{\theta}$ generally has a complicated form and that the above integral often requires approximate methods.

3. Latent GP surrogates and other choices

In the previous section we presented the BO formulation. We will now proceed to implement the formulation, and ad-

dress the choice of surrogate model for the function (Eq. 2).

Additive noise model. As briefly mentioned in the previous section, a simple case of the composition (Eq. 2) is an additive structure, $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$. Following the process of replacing $h(\mathbf{x})$ with the random variable \mathbf{h} , the resulting surrogate model of the objective function is

$$f(\mathbf{x}) = g(\mathbf{x}) + h, \quad h \sim \mathcal{N}(0, \sigma(\mathbf{x})^2), \quad (6)$$

where the variance of h is assumed to be σ^2 in order to adapt to the value range of f . With a GP surrogate model for g , the above model recovers the GP regression model with the Gaussian likelihood.

A typical choice in the above model is to assume σ^2 to be constant, leading to a homoscedastic model. A limitation of noise variances being the same across all the datapoints is that it limits the capability of the model in terms of absorbing irregular variance. A straight-forward extension of the above model is the GP with heteroscedastic noise, in which the noise variance σ^2 is allowed to be different among data points (Goldberg et al., 1998; Lázaro-Gredilla & Titsias, 2011). Another choice could be specifying a GP prior for $h(\mathbf{x})$ and thus recover an additive GP model for f (Bernardo et al., 1998; Duvenaud et al., 2011). Other available choices for an additive noise model include Student's t-distribution (Jylänki et al., 2011), Laplace (Kuss, 2006) or mixture of Gaussian likelihoods as (Kuss, 2006; Stegle et al., 2008; Naish-Guzman & Holden, 2008) where (Naish-Guzman & Holden, 2008) considers the heteroscedastic case.

Latent Gaussian process. A major limitation of the additive noise models in general is the inability to capture the interaction between the input \mathbf{x} and the noise \mathbf{h} . Another choice that produces a more flexible noise distribution is to introduce additive noise in the *input* of a GP (McHutchon & Rasmussen, 2011; Girard et al., 2003; Girard, 2004). This would correspond to the case of $f(\mathbf{x}) = g(\mathbf{x} + \mathbf{h})$. A further more general case of the proposed methodology is to allow non-linear interactions between the random variable \mathbf{h} and \mathbf{x} . This can be formulated as

$$f(\mathbf{x}) = g(\mathbf{x}, \mathbf{h}), \quad g \sim \mathcal{GP}, \quad \mathbf{h} \sim \mathcal{N}(0, \mathbb{I}). \quad (7)$$

This formulation aligns with the general assumptions proposed in the previous section. In particular, the well-behaved function g is assumed to follow a GP prior distribution, and the random variable derived from the challenging details of the objective function \mathbf{h} feeds directly into the GP surrogate model. This allows for an arbitrary interaction between \mathbf{h} and \mathbf{x} , as specified by the covariance function. The introduction of the random variable \mathbf{h} in the input results in a flexible noise distribution, as the GP model can warp the normal distribution of \mathbf{h} into a sophisticated distribution and allow non-linear interactions between \mathbf{h} and \mathbf{x} . This GP model in

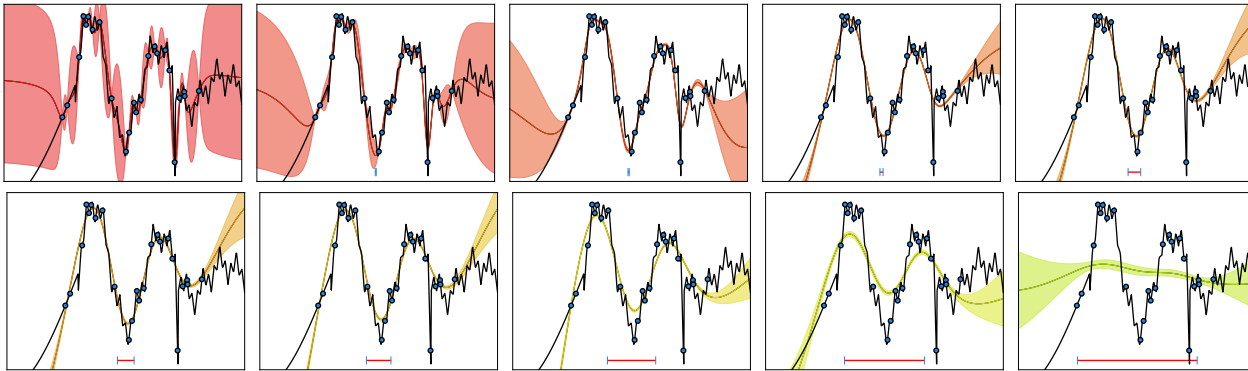


Figure 2. Input-related invariance. Each plot is showing the resulting modulated function posterior using the LGP model and setting σ_h in $p(\mathbf{h})$ to a size corresponding to the red line at the bottom of respective plot. The posterior is shown with mean and two standard deviations. The true function is shown in black. Note how the value of the prior σ_h sets the scale in relation to \mathcal{X} on how much detail is ignored. A connection can be made to low-pass filtering of higher frequencies, but where the filter varies between observations as of the posterior and where each filter is implicitly determined by fit to the function prior.

(7) is also known as a latent Gaussian process (LGP) (Pflingsten et al., 2006; Wang & Neal, 2012; Yousefi et al., 2016; Bodin et al., 2017), which is developed for regression with heteroscedastic noise and non-Gaussian residuals. The non-Gaussian marginals arise as a consequence of the latent covariates and their nonlinear transformation through the covariance function.

Function modulation via \mathcal{H} If we assume a stationary kernel over the product space $\mathcal{X} \times \mathcal{H}$, a constant \mathbf{h}_n for all observations can be interpreted as the \mathcal{H} subspace having no influence. This is due to the stationary property of the kernel, where covariances are determined only by the distances between points.

With everything else held constant, if an observation is moved away from other observations in the \mathcal{H} space, the covariances between that observation and the others are reduced. Similarly, if the length scale in \mathcal{X} -direction is shortened, the covariances between that observation and the others can be equally reduced, but that also reduces the covariances between *all other* observations due to the global influence of the hyperparameter.

Structures in the data could be explained solely by reducing the \mathcal{X} -direction length scale adequately. In that case, evaluating the posterior at $\mathbf{h}_* = \mathbf{0}$ would yield exactly the posterior of a standard GP. Conversely, structures could be explained solely as observations being adequately far from each other in the \mathcal{H} space while maintaining a longer \mathcal{X} length scale. Evaluating the posterior at $\mathbf{h}_* = \mathbf{0}$ then yields a posterior that is both influenced by the longer length scale and which has lower covariances with the data, effectively producing a posterior over smoother functions. If the posterior inputs \mathbf{H} are sufficiently far away with respect to the \mathcal{H} -direction length scale, all data variation will be captured

in \mathcal{H} and the posterior of the function at $\mathbf{h}_* = \mathbf{0}$ will in effect ignore the data.

The posterior weighting over this range of solutions is determined by the trade-off between the GP function prior and the prior of the latent inputs. As such, by controlling this trade-off, we can control properties of structures to be ignored and the ones to be used for search (see Figure 2). Important to note is that the mentioned data ignorance effect affects *individual* data points via the posterior of the corresponding latent input \mathbf{h}_n , which is influenced by the local and global fit of the function prior.

Reparameterization of LGP for ease of specifying the modulation prior

In a BO setting, some prior knowledge about what constitutes a significant change in the input space is often available. We would like to specify a joint prior of the GP and the latent inputs to ignore structures at the appropriate scale. In order to do this, we (re)-parameterise it in the following way. We set the lengthscale in the \mathcal{H} -direction to be the same as in the \mathcal{X} -direction and parameterize the latent input prior as $\mathcal{N}(\mathbf{h}_n | \mathbf{0}, \sigma_h^2 \mathbb{I})$ instead of a unit Gaussian. There is an equivalence between parameterizing σ_h or setting this trade-off via a separate lengthscale for \mathcal{H} as in (Wang & Neal, 2012). However, by using the above parameterization there is a direct correspondence in covariance reduction from moving an observation in \mathcal{H} as in \mathcal{X} , and the prior for the latent inputs can be interpreted as a prior over the coarseness of the function we wish to exploit for search. As such, intuitions about the scale in \mathcal{X} directly translate into the parameterization of the prior. See Figure 2 for a visualization of how changing σ_h in the prior affects the modulated function posterior. To make the σ_h parameterization relevant across input sizes and dimensionalities, we rescale the input domain \mathcal{X} to be a unit hyper-cube and

set σ_h proportionally to the length of the diagonal of the domain \sqrt{Q} (where Q is the number of dimensions).

Posterior inference and acquisition calculation. In BO we assume that N pairs of inputs and outputs $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$ and $\mathbf{F} = (\mathbf{f}_1, \dots, \mathbf{f}_N)^\top$ have been collected. To suggest the location for the next evaluation, we first need to infer the posterior distribution of the latent variables, which are \mathbf{H} and $\boldsymbol{\theta}$ in LGP, and then search for the maximum of the acquisition function $\alpha(\mathbf{x})$.

Given the observed data, the probabilistic model of LGP is formulated as

$$\begin{aligned} p(\mathbf{F} | \mathbf{X}, \mathbf{H}, \boldsymbol{\theta}) &= \mathcal{N}(\mathbf{F} | \mathbf{0}, \mathbf{K}), \\ p(\mathbf{H}) &= \prod_{n=1}^N \mathcal{N}(\mathbf{h}_n | \mathbf{0}, \sigma_h^2 \mathbb{I}), \end{aligned} \quad (8)$$

where \mathbf{K} is the covariance matrix computed using a chosen kernel function $k(\cdot, \cdot)$ over the set of data points $\{\bar{\mathbf{x}}_n\}_{n=1}^N$, and $\bar{\mathbf{x}}_n$ is the concatenation of two vectors $(\mathbf{x}_n^\top, \mathbf{h}_n^\top)^\top$. Because \mathbf{h}_n^\top enters the kernel function non-linearly, it is clear that the posterior distribution $p(\mathbf{H}, \boldsymbol{\theta} | \mathbf{F}, \mathbf{X})$ is intractable. To ensure the quality of the acquisition function, usually, BO methods draw posterior samples of latent variables via Markov Chain Monte Carlo (MCMC) methods such as slice sampling (Snoek et al., 2012) or Hamiltonian Monte Carlo (Duane et al., 1987). We follow this practice and provide details in the supplement.

With the approximate posterior samples $\{\mathbf{H}_i, \boldsymbol{\theta}_i\}_{i=1}^M$, we approximate the acquisition function with LGP in (5) with Monte Carlo samples,

$$\alpha(\mathbf{x}_*) \approx \frac{1}{M} \sum_{i=1}^M \hat{\alpha}(\mathbf{x}_*, \mathbf{H}_i, \boldsymbol{\theta}_i), \quad (9)$$

where $\hat{\alpha}(\mathbf{x}_*, \mathbf{H}_i, \boldsymbol{\theta}_i)$ is the acquisition function given the latent variables of LGP, which is closed-form for common acquisition functions such as expected improvement (EI) and upper confidence bound (UCB).

4. Related Work

Performing BO on an objective function that is not well-behaved is very challenging. Our method takes a Bayesian approach by incorporating a flexible noise distribution and utilising Bayesian inference to assign the challenging details of the objective function to the noise distribution. An alternative approach to this problem is to perform a model selection for the surrogate model, such that the choice of the surrogate model becomes a trade-off between the complexity of the model and the ability to locate the optimum under limited data, which has been explored in (Malkomes & Garnett, 2018). The approach uses a compositional kernel grammar from (Duvenaud et al., 2013) to induce a space of

GP models to choose from. Although this and other model selection procedures (Malkomes et al., 2016; Duvenaud et al., 2013; Grosse et al., 2012; Gardner et al., 2017) themselves have shown promise, in addition to the computational overhead, the procedures are still reliant on the existence of suitable models in this space. It remains challenging to handle cases where the objective function contains structure that is both hard to specify a-priori, and that is unhelpful in guiding the search to the optimum.

The idea of making use of noise models for dealing with model mismatch to noise-free data is not in itself new. In (Gramacy & Lee, 2012) it was shown that introducing noise in the modelling of noise-free computer experiments can lead to models with better statistical properties such as predictive accuracy and coverage. In that work, homoscedastic noise was addressed and used in a regression context.

In this paper we consider noise-free functions and address model misspecification of the function surrogate, but many works have been done to make BO resilient to noisy experiments. For example, robust noise distributions such as Student’s t-distribution have been used to make BO more resilient to noise outliers (Martinez-Cantin et al., 2017a;b). Approaches to address noisy experiments, via the addition of likelihood functions, can be combined with our approach.

Hierarchical surrogate models with input warpings have been proposed to tackle BO for non-stationary objective functions (Snoek et al., 2014; Oh et al., 2018; Calandra et al., 2016). A particularly successful application is hyperparameter optimization for machine learning methods, in which the parameters are often presented in logarithmic scales. In this case, the Beta cumulative distribution function, which only has two parameters, serves as a good warping function (Snoek et al., 2014). Such augmentation in surrogate models requires strong domain knowledge of the objective function, and one often still has to control the increased complexity of the surrogate model, which is orthogonal to our approach.

5. Experiments

In this section we will demonstrate the benefit of our approach empirically. As the approach is motivated by robustness to the presence of challenging structures in the objective function, we will test its ability to improve search efficiency on a range of functions exhibiting such structure. Visual examples of functions with typical properties are shown in Figure 3. As we will show, our approach increases reliability in the search when faced with detrimental structure (see Figure 4) that has a large negative impact on traditional surrogates.

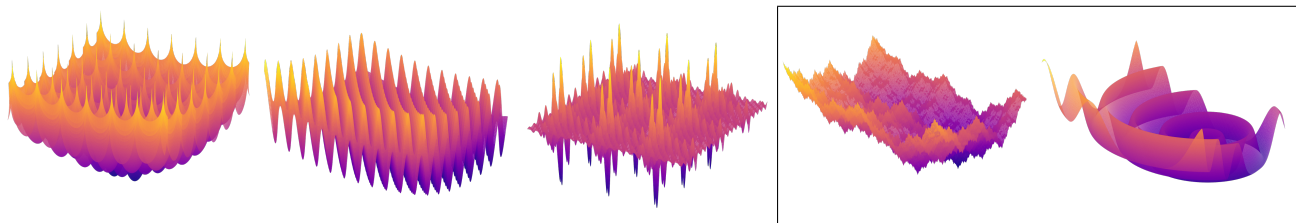


Figure 3. Surface plots of the benchmark functions Cross In Tray, Griewank, Shubert, Weierstrass and Deflected Corrugated Spring (McCourt, 2016), from the left. The two right-most functions are available in multiple dimensionalities, where 8D and 10D is used in the experiments, respectively.

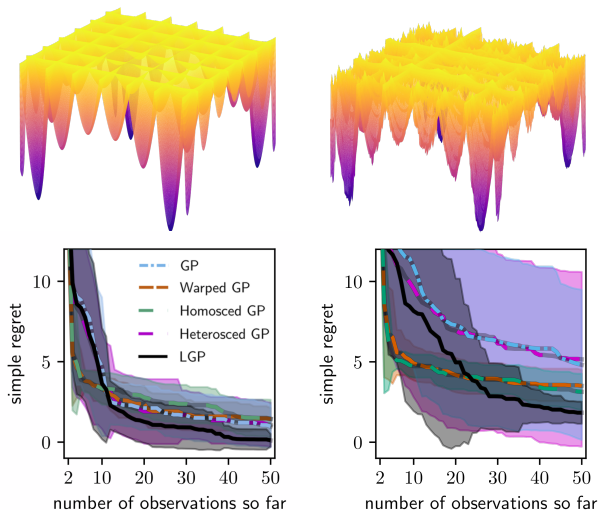


Figure 4. A comparison of experiments on the Holder Table benchmark (McCourt, 2016) (left) and a corrupted version with added nonsmooth structure (right). We show plots of the respective functions and performance in terms of regret so far (over 20 repetitions). We show both mean regret (line) and standard deviation (shading). The nonsmooth structure is challenging for a noiseless GP to model and leads to a high variance in-between runs. Warped and homoscedastic GPs explain away the corruption, but their performance plateaus as no informative trends can be identified. LGP reliably identifies these trends and reliably finds good solutions.

Baselines and metric We compare with and without function modulation (Section 2) - implemented as in Section 3 - on a popular GP model setup for BO. In addition, we compare the LGP against other methods of handling challenging structure in the objective function, namely (i) a noiseless GP, (ii) a GP with homoscedastic noise, (iii) a GP with heteroscedastic noise and (iv) a non-stationary, Warped GP (Snoek et al., 2014).

We follow the standard practice to compare across benchmarks and provide the *mean gap* estimated over 20 runs as in (Malkomes & Garnett, 2018). The gap measure is defined as $\frac{f(x_{\text{first}}) - f(x_{\text{best}})}{f(x_{\text{first}}) - f(x_{\text{optimum}})}$, where $f(x_{\text{first}})$ is the minimum function value among the first initial random points, $f(x_{\text{best}})$

is the best function value found within the evaluation budget and $f(x_{\text{optimum}})$ is the function’s true optimum. Methods are judged to have very similar or equivalent performance to the best performing if not significantly different, determined by a two-sided paired Wilcoxon signed-rank test at 5% significance (Malkomes & Garnett, 2018). We also report regret (with mean and standard deviation) in the supplement.

We use the Matérn 5/2 kernel for all surrogates, the expected improvement acquisition function (where not otherwise stated) and Bayesian hyperparameter marginalisation as in (Snoek et al., 2012). For the maximization of the expected utility with respect to input location, we use δ -cover sampling, as in (De Freitas et al., 2012). The Warped GP implementation and inference is from the Spearmint package (Snoek et al., 2014). For further details, we refer to the supplement.

Benchmark datasets We perform the comparisons on benchmarks from (McCourt, 2016; Head et al., 2018) using the default domains provided by respective benchmark, detailed in the supplement. In addition, problems are marked with the descriptive properties given in (McCourt, 2016) and in the supplement that can reflect the relative difficulty of the task.

Priors on the latent input variables The prior $p(\mathbf{h}_n) = \mathcal{N}(\mathbf{0}, \sigma_h^2 \mathbb{I})$ can be parameterized in relation to the relative scale of the characteristics to be ignored. We specify the function prior over the product space $\mathcal{X} \times \mathcal{H}$ using a kernel with common parameters for \mathbf{x}_n and \mathbf{h}_n . Thus, the standard deviation of the prior σ_h relates directly to distances in the \mathcal{X} -direction. When domain-specific knowledge is available, $p(\mathbf{H})$ may be specified at an appropriate scale. However, we often do not have access to such knowledge. In all our experiments, we adopt a hierarchical prior approach whereby σ_h is sampled uniformly from a small candidate set at each evaluation. Specifically, $\sigma_h \sim \mathbb{U}(\{0.1d, 0.01d, 0\})$ where $d = \sqrt{Q}$, the length of the diagonal of the unit Q -dimensional hypercube. We found that this approach performed well empirically and is applied consistently across all our experiments where not otherwise specified. A choice

Table 1. Mean gap performance for various test functions; higher is better. The upper table shows the results after 50 objective function evaluations and the lower table after 100 evaluations. Due to computational cost, Warped GP results are only reported for 50 evaluations. Methods not significantly different from the best performing method with respect by a two-sided paired Wilcoxon signed-rank test at a 5% significance level over 20 repetitions are shown in bold (Malkomes & Garnett, 2018). For results in terms of regret, see the supplement.

Benchmark	Evals	Dim	Properties	GP	Warped GP	Homosced GP	Heterosced GP	LGP
Hartmann	50	6	boring	0.959	0.537	0.881	0.973	0.937
Griewank	50	2	oscillatory	0.914	0.493	0.752	0.913	0.897
Shubert	50	2	oscillatory	0.378	0.158	0.378	0.480	0.593
Ackley $[-10, 30]^d$	50	2	complicated, oscillatory	0.924	0.274	0.892	0.912	0.927
Cross In Tray	50	2	complicated, oscillatory	0.954	0.385	0.929	0.977	0.945
Holder table	50	2	complicated, oscillatory	0.939	0.896	0.900	0.931	0.993
Corrupted Holder Table	50	2	complicated, oscillatory	0.741	0.798	0.826	0.729	0.896
Branin01	100	2	none	1.000		1.000	1.000	1.000
Branin02	100	2	none	0.991		0.964	0.990	0.981
Beale	100	2	boring	0.987		0.982	0.987	0.988
Hartmann	100	6	boring	0.987		0.947	0.984	0.979
Griewank	100	2	oscillatory	0.967		0.875	0.969	0.946
Levy	100	2	oscillatory	0.997		0.999	0.998	0.998
Deflected Corrugated Spring	100	10	oscillatory	0.347		0.840	0.406	0.697
Shubert $[-10, 10]^d$	100	2	oscillatory	0.510		0.511	0.672	0.877
Weierstrass	100	8	complicated	0.600		0.704	0.577	0.625
Cross In Tray	100	2	complicated, oscillatory	1.000		0.995	1.000	1.000
Holder Table	100	2	complicated, oscillatory	0.971		0.963	0.964	1.000
Ackley $[-10, 30]^d$	100	2	complicated, oscillatory	0.971		0.914	0.980	0.974
Ackley $[-10, 30]^d$	100	6	complicated, oscillatory	0.459		0.789	0.442	0.712
Corrupted Holder Table	100	2	complicated, oscillatory	0.844		0.889	0.822	0.918
Corrupted Exponential	100	8	complicated, oscillatory	0.580		0.847	0.581	0.806
HPO: NN Boston	100	9	unknown	0.720		0.761	0.810	0.770
HPO: NN Climate Model Crashes	100	9	unknown	0.629		0.717	0.683	0.678
Active learning: Robot Pushing	100	4	unknown	0.877		0.745	0.907	0.932

of $\sigma_h \rightarrow 0$ corresponds to a noiseless GP without latent covariates.

Evaluation on benchmark suite Table 1 presents results across a wide range of benchmark functions consisting of the SigOpt benchmark suite (McCourt, 2016). Three additional real-world benchmarks (Head et al., 2018; Malkomes & Garnett, 2018; Kaelbling & Lozano-Pérez, 2017) are included in the bottom section of the table. The benchmarks from (McCourt, 2016) are popular functions used in both black-box optimization as well as classic optimization literature. As of the focus of the paper, benchmarks from the literature exhibiting challenging properties such as oscillatory local structures were included, in addition to simpler functions for reference.

In general, the noise-free, homoscedastic and heteroscedastic noise GPs tend to either share best place with the LGP or be outperformed by it. The Warped GP, which warps the input space to obtain a tight fit to the data, consistently struggle with the complicated and oscillatory benchmarks. On some benchmarks there are large differences in favour of using noisy surrogates on the noise-free benchmarks. Such an example is Ackley 6D, which in the dataset is described as “technically oscillatory, but with such a short wavelength

that its behavior probably seems more like noise” (McCourt, 2016). Another example is the Shubert function, which has multiple sharp local optima surrounded by large oscillations. On 2 of 18 benchmarks, the LGP was not best (or within the two-sided Wilcoxon test), but instead the homoscedastic GP. These functions were Weierstrass, which has a homoscedastic characteristic (see Figure 3), and Deflected Corrugated Spring, on both of which the LGP obtained the second highest mean gap. In contrast to the LGP and the heteroscedastic GP, the noise model of the homoscedastic GP sometimes hurt performance in relation to the noise-free GP. Given the black-box nature of functions in BO, it is important that the surrogate noise model ‘turns off’ adequately when not needed. The heteroscedastic GP provided significant benefit on two benchmarks over the GP, whereas the LGP provided such benefit on *eight* benchmarks.

Real-world Apart from widely used synthetic functions, we also compare our method on three real-world problems. The results are shown in the last three rows of Table 1. One of the benchmarks is an active learning task of a robot pushing a box within a simulation. As we show in Figure 5, the benchmark’s response surface is both nonsmooth and oscillatory. The LGP reliably found good solutions on the

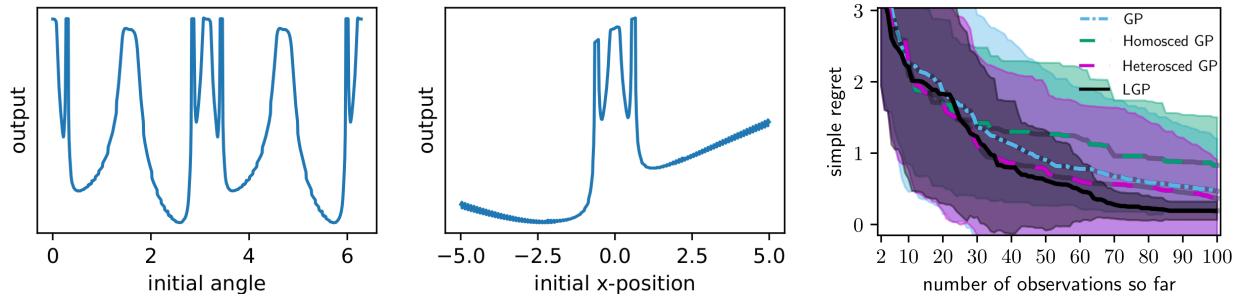


Figure 5. Active learning: Robot pushing. The objective function to be optimized takes as input the pushing action of a robot within a simulation, and outputs the distance of the pushed object to the goal location. The two plots on the left show that the task surface, resulting from the dynamic system, is nonsmooth and non-trivial. The right plot shows the mean regret and standard deviations for the different surrogates. As can be seen, the LGP found good solutions with low variance, improving reliability of the search. The 1D slices of the 4D function (the two from the left) was generated by fixing the initial y-position (param.) to the one of the goal position, the simulation steps (param.) to the center of its domain, and varying the initial angle (param.) or the x-position (param.), respectively, while keeping the other fixed at zero. Slicing the 4D function differently produced similar nonsmooth response curves.

benchmark, while the other surrogates sometimes failed, resulting in high variances. The homoscedastic GP performed the worst, which we suspect is due to the nonsmooth and heavily oscillatory structures forces a high global noise level, which may lead to failure in utilising informative structure in other regions.

Other acquisition functions We suggest that the problem with structures challenging to model is relevant to address irrespective of the acquisition function. To confirm that the method is applicable also using other acquisition functions we ran the Corrupted Exponential benchmark using both Expected Improvement (EI) and Lower Confidence Bound (LCB) with the default exploration weight ($= 2.0$) from GPyOpt (GPyOpt, 2016). As can be seen in Table 1, in the case of EI, the GP and the heteroscedastic GP performed worse than the homoscedastic GP and the LGP. The homoscedastic GP achieved the highest mean gap, but the difference was not significant under the Wilcoxon test to the LGP which obtained a similar mean gap. Using the LCB acquisition function the performance for the homoscedastic GP decreased to 0.818 and the LGP increased to 0.858, and their difference in rank in favour of the LGP was significant under the test. The heteroscedastic GP increased using LCB to 0.797 and the GP to 0.751, remaining as worst performing.

As the experimental evaluation demonstrates, our suggested approach for handling challenging structures in the objective function consistently improved reliability and performance over the traditional surrogate on a wide range of benchmarks. Importantly, on benchmarks where the extended methodology were not needed the performance aligned with that of the traditional surrogate. When it was needed, it was shown

to often have a large positive impact on overall efficiency of the search.

6. Conclusion

We have presented an approach to Bayesian Optimization where the surrogate model is alleviated from needing to explain the observed objective function values perfectly, which is challenging for complicated or nonsmooth functions. Instead, we model the essential structure of the objective function that is well-behaved and leave the rest of the function details to be absorbed in a noise distribution. We show experimentally how our approach is able to solve synthetic and real-world benchmarks with challenging local structures reliably. Importantly our methodology can be applied to any surrogate model used for BO, and the specific case addressed in the paper can be included in any Gaussian process-based surrogate.

7. Acknowledgements

This project was supported by Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership, the Hans Werthn Fund at The Royal Swedish Academy of Engineering Sciences, EPSRC CDE (EP/L016540/1), EPSRC CAMERA project (EP/M023281/1), the German Federal Ministry of Education and Research (project 01 IS 18049 A), and the Royal Society.

References

Bernardo, J., Berger, J., Dawid, A., Smith, A., et al. Regression and classification using gaussian process priors. *Bayesian statistics*, 6:475, 1998.

- Bodin, E., Campbell, N. D. F., and Ek, C. H. Latent Gaussian Process Regression. 2017. URL <http://arxiv.org/abs/1707.05534>.
- Calandra, R., Peters, J., Rasmussen, C. E., and Deisenroth, M. P. Manifold gaussian processes for regression. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 3338–3345. IEEE, 2016.
- De Freitas, N., Smola, A., and Zoghi, M. Exponential regret bounds for gaussian process bandits with deterministic observations. *arXiv preprint arXiv:1206.6457*, 2012.
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics Letters B*, 195(2), 1987.
- Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. Structure discovery in nonparametric regression through compositional kernel search. *arXiv preprint arXiv:1302.4922*, 2013.
- Duvenaud, D. K., Nickisch, H., and Rasmussen, C. E. Additive gaussian processes. In *Advances in neural information processing systems*, pp. 226–234, 2011.
- Gardner, J., Guo, C., Weinberger, K., Garnett, R., and Grosse, R. Discovering and exploiting additive structure for bayesian optimization. In *Artificial Intelligence and Statistics*, pp. 1311–1319, 2017.
- Girard, A. *Approximate methods for propagation of uncertainty with Gaussian process models*. PhD thesis, Citeseer, 2004.
- Girard, A., Rasmussen, C. E., Candela, J. Q., and Murray-Smith, R. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. In *Advances in neural information processing systems*, pp. 545–552, 2003.
- Goldberg, P. W., Williams, C. K., and Bishop, C. M. Regression with input-dependent noise: A gaussian process treatment. In *Advances in neural information processing systems*, pp. 493–499, 1998.
- González, J., Dai, Z., Damianou, A., and Lawrence, N. D. Preferential Bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1282–1291, 2017.
- GPyOpt. Gpyopt: A bayesian optimization framework in python. <http://github.com/SheffieldML/GPyOpt>, 2016.
- Gramacy, R. B. and Lee, H. K. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.
- Grosse, R., Salakhutdinov, R. R., Freeman, W. T., and Tenenbaum, J. B. Exploiting compositionality to explore a large space of model structures. *arXiv preprint arXiv:1210.4856*, 2012.
- Head, T., MechCoder, Louppe, G., Shcherbatyi, I., fcharras, Vincius, Z., cmmalone, Schrder, C., nel215, Campos, N., Young, T., Cereda, S., Fan, T., rene rex, Shi, K. K., Schwabedal, J., carlosdanielcsantos, Hvass-Labs, Pak, M., SoManyUsernamesTaken, Callaway, F., Estve, L., Besson, L., Cherti, M., Pfannschmidt, K., Linzberger, F., Cauet, C., Gut, A., Mueller, A., and Fabisch, A. scikit-optimize/scikit-optimize: v0.5.2, March 2018. URL <https://doi.org/10.5281/zenodo.1207017>.
- Jenatton, R., Archambeau, C., González, J., and Seeger, M. Bayesian optimization with tree-structured dependencies. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1655–1664, 2017.
- Jylänki, P., Vanhatalo, J., and Vehtari, A. Robust gaussian process regression with a student-t likelihood. *Journal of Machine Learning Research*, 12(Nov):3227–3257, 2011.
- Kaelbling, L. P. and Lozano-Pérez, T. Learning composable models of parameterized skills. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 886–893. IEEE, 2017.
- Kuss, M. *Gaussian process models for robust regression, classification, and reinforcement learning*. PhD thesis, Technische Universität, 2006.
- Lázaro-Gredilla, M. and Titsias, M. K. Variational heteroscedastic gaussian process regression. In *ICML*, pp. 841–848, 2011.
- Malkomes, G. and Garnett, R. Automating bayesian optimization with bayesian optimization. In *Advances in Neural Information Processing Systems*, pp. 5984–5994, 2018.
- Malkomes, G., Schaff, C., and Garnett, R. Bayesian optimization for automated model selection. In *Advances in Neural Information Processing Systems*, pp. 2900–2908, 2016.
- Martinez-Cantin, R., McCourt, M., and Tee, K. Robust bayesian optimization with student-t likelihood. *arXiv preprint arXiv:1707.05729*, 2017a.
- Martinez-Cantin, R., Tee, K., and McCourt, M. Practical bayesian optimization in the presence of outliers. *arXiv preprint arXiv:1712.04567*, 2017b.
- McCourt, M. Optimization test functions. <https://github.com/sigopt/evalset>, 2016.

- McHutchon, A. and Rasmussen, C. E. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, pp. 1341–1349, 2011.
- Naish-Guzman, A. and Holden, S. Robust regression with twinned gaussian processes. In *Advances in neural information processing systems*, pp. 1065–1072, 2008.
- Oh, C., Gavves, E., and Welling, M. Bock : Bayesian optimization with cylindrical kernels. In *ICML*, 2018.
- Pfingsten, T., Kuss, M., and Rasmussen, C. E. Nonstationary gaussian process regression using a latent extension of the input space. In *ISBA Eighth World Meeting on Bayesian Statistics*, 2006.
- Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pp. 2951–2959, 2012.
- Snoek, J., Swersky, K., Zemel, R., and Adams, R. Input warping for bayesian optimization of non-stationary functions. In *International Conference on Machine Learning*, pp. 1674–1682, 2014.
- Stegle, O., Fallert, S. V., MacKay, D. J., and Brage, S. Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering*, 55 (9):2143–2151, 2008.
- Wang, C. and Neal, R. M. Gaussian process regression with heteroscedastic or non-gaussian residuals. *CoRR*, 2012. URL <http://arxiv.org/abs/1212.6246v1>.
- Yousefi, F., Dai, Z., Ek, C. H., and Lawrence, N. Un-supervised Learning with Imbalanced Data via Structure Consolidation Latent Variable Model. 2016. URL <http://arxiv.org/abs/1607.00067>.