Uncertainty quantification using martingales for misspecified Gaussian processes

Willie Neiswanger NEISWANGER@CS.STANFORD.EDU
Stanford University

Aaditya Ramdas ARAMDAS@STAT.CMU.EDU
Carnegie Mellon University

Editors: Vitaly Feldman, Katrina Ligett, and Sivan Sabato

Abstract

We address uncertainty quantification for Gaussian processes (GPs) under misspecified priors, with an eye towards Bayesian Optimization (BO). GPs are widely used in BO because they easily enable exploration based on posterior uncertainty bands. However, this convenience comes at the cost of robustness: a typical function encountered in practice is unlikely to have been drawn from the data scientist's prior, in which case uncertainty estimates can be misleading, and the resulting exploration can be suboptimal. We present a frequentist approach to GP/BO uncertainty quantification. We utilize the GP framework as a working model, but do not assume correctness of the prior. We instead construct a *confidence sequence* (CS) for the unknown function using martingale techniques. There is a necessary cost to achieving robustness: if the prior was correct, posterior GP bands are narrower than our CS. Nevertheless, when the prior is wrong, our CS is statistically valid and empirically outperforms standard GP methods, in terms of both coverage and utility for BO. Additionally, we demonstrate that powered likelihoods provide robustness against model misspecification.

1. Introduction

In Bayesian optimization (BO), a Bayesian model is leveraged to optimize an unknown function f^* (Mockus et al., 1978; Shahriari et al., 2015; Snoek et al., 2012). One is allowed to query the function at various points x in the domain, and get noisy observations of $f^*(x)$ in return. Most BO methods use a Gaussian process (GP) prior, with a chosen kernel function. However, in practice, it may be difficult to specify the prior accurately. A few examples of where misspecification may arise include

- an incorrect kernel choice (e.g. squared exponential versus Matern),
- bad estimates of kernel hyperparameters (e.g. lengthscale or signal variance), and
- heterogenous smoothness of f^* over the domain $\mathcal{X}$.

Each of these can yield misleading uncertainty estimates, which may then negatively affect the performance of BO (Schulz et al., 2016; Sollich, 2002). **This paper instead presents a frequentist approach to uncertainty quantification for GPs** (and hence for BO), which uses martingale techniques to construct a confidence sequence (CS) for f^*, irrespective of misspecification of the prior. A CS is a sequence of (data-dependent) sets that are uniformly valid over time, meaning that $\{C_t\}_{t\geq 1}$ such that $\Pr(\exists t \in \mathbb{N} : f^* \notin C_t) \leq \alpha$. The price of such a robust guarantee is that if the prior was indeed accurate, then our confidence sets are looser than those derived from the posterior.

Outline The next page provides a visual illustration of our contributions. Section 2 provides the necessary background on GPs and BO, as well as on martingales and confidence sequences. Section 3 derives our prior-robust confidence sequence, as well as several technical details needed to implement it in practice. Section 4 describes the simulation setup used in Figure 1 in detail. We end by discussing related work and future directions in Section 5, with additional figures in the appendix.

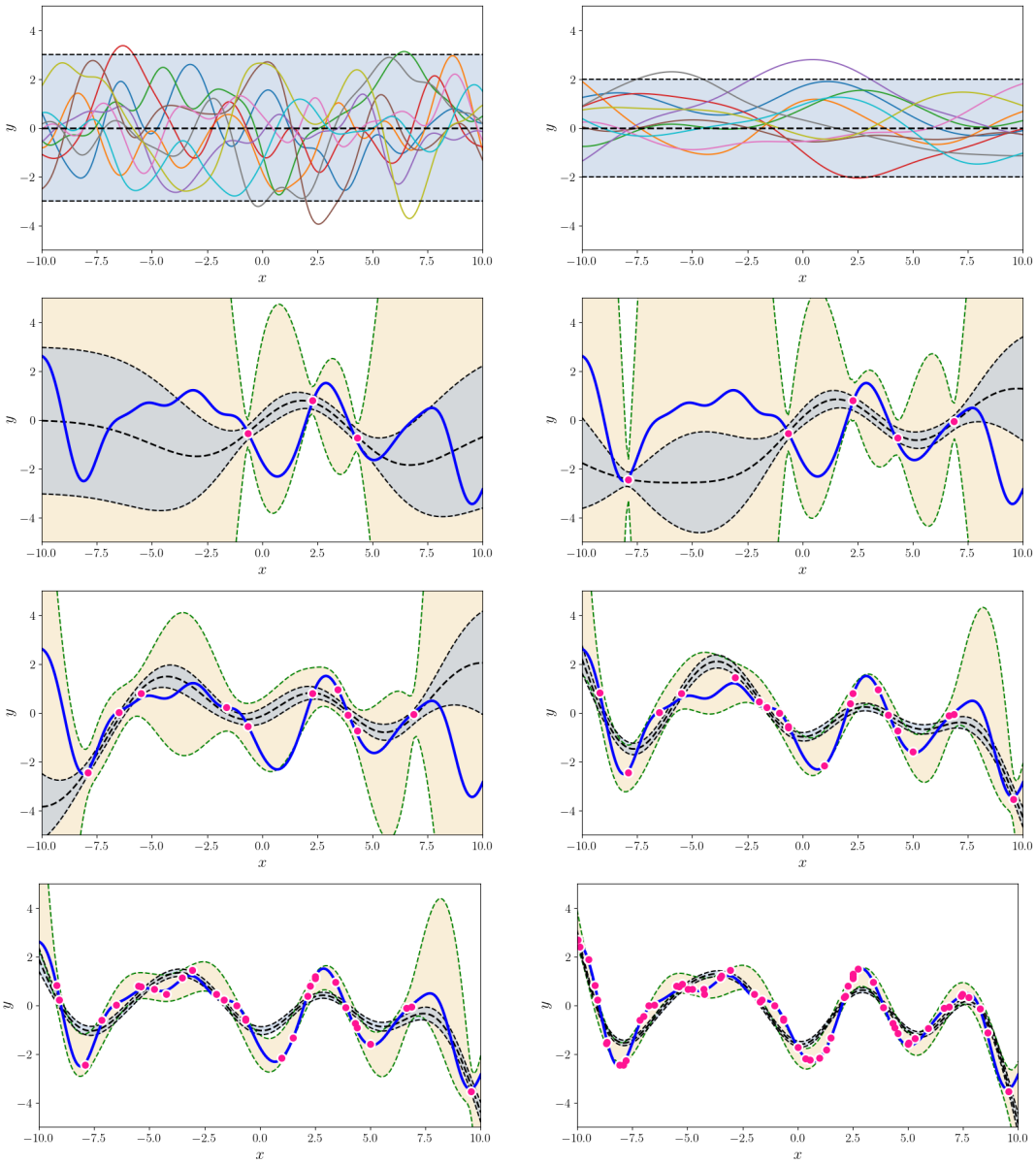

Figure 1: This figure summarizes the paper's contributions. The top two plots show various random functions drawn from a GP prior with hyperparameter settings A (left) and B (right). Then, a single function (blue curve) is drawn using prior A, and is fixed through the experiment. The pink dots are the observations; there are 3, 5, 10, 20, 30, 60 pink dots in the bottom 6 plots. The grey shaded region shows the standard GP posterior when (mistakenly) working with prior B. The brown shaded region shows our new confidence sequence, also constructed with the wrong prior B. The brown region is guaranteed to contain the true function with high probability uniformly over time. The grey confidence band after just 3 observations is already (over)confident, but quite inaccurate, and it never recovers. The brown confidence sequence is very wide early on (perhaps as it should be) but it recovers as more points are drawn. Thus, the statistical price of robustness to prior misspecification is wider bands. Whether this is an acceptable tradeoff to the practitioner is a matter of their judgment and confidence in the prior. The observation that the posterior never converges to the truth (the data does not wash away the wrong prior) appears to be a general phenomenon of failure of the Bernstein-von-Mises theorem in infinite dimensions (Freedman, 1999; Cox, 1993). The rest of this paper explains how this confidence sequence is constructed, using the theory of martingales. We provide more details (such as kernel hyperparameters A and B) for this simulation in Section 4. Simulations for BO are available in the supplement.

2. Mathematical background

Gaussian Processes (GP). A GP is a stochastic process (a collection of random variables indexed by domain $\mathcal{X}$) such that every finite collection of those random variables has a multivariate normal distribution. The distribution of a GP is a distribution over functions $g : \mathcal{X} \mapsto \mathbb{R}$, and thus GPs are often used as Bayesian priors over unknown functions. A GP is itself typically specified by a mean function $\mu : \mathcal{X} \to \mathbb{R}$ and a covariance kernel $\kappa : \mathcal{X}^2 \to \mathbb{R}$. Suppose we draw a function

$$f \sim \text{GP}(\mu, \kappa) \tag{1}$$

and obtain a set of n observations $D_n = \{(X_i, Y_i)\}_{i=1}^n$, where $X_i \in \mathcal{X}$,

$$Y_i = f(X_i) + \epsilon_i \in \mathbb{R}, \text{ and } \epsilon_i \sim \mathcal{N}(0, \eta^2). \tag{2}$$

Then, the posterior process $f|D_n$ is also a GP with mean function μ_n and covariance kernel κ_n, described as follows. Collect the Y_is into a vector $Y \in \mathbb{R}^n$, and define $k, k' \in \mathbb{R}^n$ with $k_i = \kappa(x, X_i), k'_i = \kappa(x', X_i)$, and $K \in \mathbb{R}^{n \times n}$ with $K_{i,j} = \kappa(X_i, X_j)$. We can then write μ_n, κ_n as

$$\mu_n(x) = k^\top (K + \eta^2 I)^{-1} Y, \qquad \kappa_n(x, x') = \kappa(x, x') - k^\top (K + \eta^2 I)^{-1} k'. \tag{3}$$

Further background on GPs can be found in Williams and Rasmussen (2006). In this paper, we describe a simple method for inference when (1) does not hold, but (2) holds; in other words, the prior is arbitrarily misspecified but the model is correct. (If both are correct, GPs work fine, and if both are arbitrarily misspecified, statistical inference is essentially impossible.)

Bayesian Optimization (BO). Suppose we wish to minimize an unknown, fixed, nonrandom, function f^* over a domain $\mathcal{X}$. Bayesian optimization (BO) leverages probabilistic models to perform optimization by assuming that f^* was sampled from a GP.

At time t (we switch from n to t to emphasize temporality), assume we have already evaluated f^* at points $\{X_i\}_{i=1}^{t-1}$ and obtained observations $\{Y_i\}_{i=1}^{t-1}$. To determine the next domain point X_t to evaluate, we first use the posterior GP to define an *acquisition function* $\varphi_t : \mathcal{X} \to \mathbb{R}$, which specifies the utility of evaluating f^* at any $x \in \mathcal{X}$. We then minimize the acquisition function to yield $X_t = \text{argmin}_{x \in \mathcal{X}} \varphi_t(x)$, and evaluate f^* at X_t. One of the most commonly used acquisition functions is the GP lower confidence bound[1] (GP-LCB) by Srinivas et al. (2010), written

$$\varphi_t(x) = \mu_t(x) - \beta_t^{1/2} \sigma_t(x) \tag{4}$$

where μ_t and σ_t are the posterior GP mean and standard deviation, and $\beta_t > 0$ is a tuning parameter that determines the tradeoff between exploration and exploitation.

Due to inheriting their worldview from GPs, theoretical guarantees in the BO literature typically assume correctness of both (1) and (2). These may or may not be reasonable assumptions. In this paper, (1) is used as a working model that is not assumed correct, but (2) is still assumed. We do not provide guarantees on any particular BO algorithm persay, but instead provide correct uncertainty quantification that could be exploited by any BO algorithm, including but not necessarily GP-LCB.

1. Often described as the GP upper confidence bound (GP-UCB), we use the GP lower confidence bound (GP-LCB) since we are performing minimization.

Filtrations and stopping times. To make the sequential aspect of BO explicit, let

$$\mathcal{D}_t = \sigma((X_1, Y_1), \ldots, (X_t, Y_t)) \equiv \sigma(D_t)$$

denote the sigma-field of the first t observations, which captures the information known at time t; $\mathcal{D}_0$ is the trivial sigma-field. Since $\mathcal{D}_t \supset \mathcal{D}_{t-1}$, $\{\mathcal{D}_t\}_{t\geq 0}$ forms a filtration (an increasing sequence of sigma-fields). Using this language, the acquisition function φ_t is then *predictable*, written $\varphi_t \in \mathcal{D}_{t-1}$, meaning that it is measurable with respect to $\mathcal{D}_{t-1}$ and is hence determined with only the data available after $t-1$ steps. As a result, X_t is technically also predictable. However, Y_t is not predictable (it is *adapted*), since $Y_t \in \mathcal{D}_t$ but $Y_t \notin \mathcal{D}_{t-1}$. A stopping time τ is an $\mathbb{N}$-valued random variable such that $\{\tau \geq t\} \in \mathcal{D}_{t-1}$ for all t, or equivalently if

$$\{\tau \leq t\} \in \mathcal{D}_t,$$

meaning that we can tell if we have stopped by time t, using only the information available up to t.

Martingales. An integrable process $\{M_t\}_{t\geq 0}$ is said to be a martingale with respect to filtration $\{\mathcal{D}_t\}_{t\geq 0}$, if $M_t \in \mathcal{D}_t$ and for every $t \geq 1$, we have

$$\mathbb{E}[M_t | \mathcal{D}_{t-1}] = M_{t-1}.$$

If we replaced the $=$ above by an inequality $\leq$, the resulting process is called a supermartingale. Every martingale is a supermartingale but not vice versa. Ville's inequality (Ville, 1939, Pg. 100) states that if $\{M_t\}$ is a nonnegative supermartingale, then for any $x > 0$, we have

$$\Pr(\exists t \in \mathbb{N} : M_t \geq x) \leq \frac{\mathbb{E}[M_0]}{x}. \tag{5}$$

See Howard et al. (2020, Lemma 1) for a measure-theoretic proof and Shafer and Vovk (2019, Proposition 14.8) for a game-theoretic variant. In many statistical applications, M_0 is chosen to deterministically equal one. Ville's may be viewed as a time-uniform version of Markov's inequality for nonnegative random variables. In the next section, we will construct a martingale (hence supermartingale) for GP/BO, and construct a confidence sequence (defined next) for the underlying function f by applying Ville's inequality.

Confidence sequences (CS). Suppose we wish to sequentially estimate an unknown quantity θ^* (a scalar, vector, function, etc.) as we observe an increasing number of datapoints, summarized as a filtration $\mathcal{D}_t$. A CS is defined as a sequence of confidence sets $\{C_t\}_{t\geq 1}$ that contains θ^* at all times with high probability. Formally, for a confidence level $\alpha \in (0, 1)$, we need that $C_t \in \mathcal{D}_t$ and

$$\underbrace{\Pr(\forall t \in \mathbb{N} : \theta^* \in C_t)}_{\text{coverage at all times}} \geq 1 - \alpha \quad \equiv \quad \underbrace{\Pr(\exists t \in \mathbb{N} : \theta^* \notin C_t)}_{\text{error at some time}} \leq \alpha. \tag{6}$$

Here, C_t obviously depends on α, but it is suppressed for simplicity. Importantly, property (6) holds if and only if $\Pr(\theta^* \in C_\tau) \geq 1 - \alpha$ for all possible (potentially infinite) stopping times τ. This allows us to provide correct uncertainty quantification that holds even at data-dependent stopping times. Next, we describe our construction of a confidence sequence for f^*.

3. Deriving our prior-robust confidence sequence

One of the roles of the prior in BO is to restrict the complexity of the function f^*. Since we do not assume the prior is well-specified, we need some other way to control the complexity of f^*—without any restriction on f^*, we cannot infer its value at any point outside of the observed points since it could be arbitrarily different even at nearby points. We do this by assuming that $f^* \in \mathcal{F}$ for some $\mathcal{F}$, that is either explicitly specified—say via a bound on the Reproducing Kernel Hilbert Space (RKHS) norm, or by a bound on the Lipschitz constant—or implicitly specified (via some kind of regularization).

The choice of $\mathcal{F}$ has both statistical and computational implications; the former relates to the size of the class, the sample complexity of identifying the optimum via BO, and the rate at which the confidence bands will shrink, while the latter relates to how much time it takes to calculate and/or update the confidence bands. Ultimately, $\mathcal{F}$ must be specified by the practitioner based on their knowledge of the underlying problem domain. For this section, we treat any arbitrary $\mathcal{F}$, and in the next section we discuss one particular choice of $\mathcal{F}$ for which the computational load is reasonable.

It is worth noting that we have not simply shifted the problem from specifying a prior to specifying $\mathcal{F}$—the latter does not impose a probability structure amongst its elements, while the former does. There are other differences as well; for example comparing a GP prior with a particular kernel, to a bounded RKHS ball for the same kernel, we find that the former is much richer than the latter—as mentioned after Theorem 3 of Srinivas et al. (2010), random functions drawn from a GP have infinite RKHS norm almost surely, making the sample paths much rougher/coarser than functions with bounded RKHS norm.

3.1. Constructing the prior-posterior-ratio martingale

We first begin with some technicalities. Recall that a GP is interpreted as a prior distribution over functions $g : \mathcal{X} \mapsto \mathbb{R}$. For simplicity, and to avoid measure-theoretic issues, consider the case of $\mathcal{X} = \mathbb{R}^d$ by default, equipped with the Borel sigma algebra. It is clear to us that the following results do hold more generally, albeit at the price of further mathematical machinery, since extra care is needed when dealing with infinite-dimensional measures. Let $\text{GP}_0(f)$ represent the prior "density" at function f, and let $\text{GP}_t(f)$ represent the posterior "density" at f after observing t datapoints. "Density" is in quotes because in infinite dimensional spaces, there is no analog of the Lebesgue measure, and thus it is a priori unclear which measure these are densities with respect to. Proceeding for now, we soon sort this issue out.

Define the *prior-posterior-ratio* for any function f as the following real-valued process:

$$R_t(f) := \frac{\text{GP}_0(f)}{\text{GP}_t(f)}. \tag{7}$$

Note that $R_0(f) = 1$ for all f. Denote the working likelihood of f by

$$\mathcal{L}_t(f) \;:=\; \prod_{i=1}^{t} \frac{1}{\eta\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{Y_i - f(X_i)}{\eta}\right)^2} \;\equiv\; \prod_{i=1}^{t} \frac{1}{\eta}\phi\left(\frac{Y_i - f(X_i)}{\eta}\right), \tag{8}$$

where $\phi(y)$ denotes the standard Gaussian PDF, so that $\phi((y-\mu)/\sigma)/\sigma$ is the PDF of $\mathcal{N}(\mu, \sigma^2)$. Then, for any function f, the working posterior GP is given by

$$\text{GP}_t(f) := \frac{\text{GP}_0(f)\mathcal{L}_t(f)}{\int_g \text{GP}_0(g)\mathcal{L}_t(g)}. \tag{9}$$

Substituting the posterior (9) and likelihood (8) into the definition of the prior-posterior-ratio (7), the latter can be more explicitly written as

$$R_t(f) \;=\; \int_g \mathrm{GP}_0(g)\frac{\mathcal{L}_t(g)}{\mathcal{L}_t(f)} \;\equiv\; \mathbb{E}_{g\sim\mathrm{GP}_0}\left[\frac{\mathcal{L}_t(g)}{\mathcal{L}_t(f)}\right], \tag{10}$$

and it is this last form that we use, since it avoids measure-theoretic issues. Indeed, $\mathcal{L}_t(f), R_t(f)$ are well-defined and finite for every f, as long as f itself is finite, and one is anyway uninterested in considering functions that can be infinite on the domain.

As mentioned at the start of this section, fix a function $f^* \in \mathcal{F}$. Assume that the data are observed according to (2) when the X_is are predictably chosen according to any acquisition function. Despite not assuming (1), we will still use a GP framework to model and work with this data, and we call this our "working prior" to differentiate it from an assumed prior.

Lemma 1. *Fix any arbitrary $f^* \in \mathcal{F}$, and assume data-generating model (2). Choose any acquisition function φ_t, any working prior GP_0 and construct the working posterior GP_t. Then, the prior-posterior-ratio at f^*, denoted $\{R_t(f^*)\}_{t\geq 0}$, is a martingale with respect to filtration $\{\mathcal{D}_t\}_{t\geq 0}$.*

Proof. Evaluating R_t at f^*, taking conditional expectations and applying Fubini's theorem, yields

$$\begin{aligned}
\mathbb{E}_{D_t\sim f^*}[R_t(f^*) \mid \mathcal{D}_{t-1}] &= \mathbb{E}_{D_t\sim f^*}\left[\mathbb{E}_{g\sim\mathrm{GP}_0}\left[\frac{\mathcal{L}_t(g)}{\mathcal{L}_t(f^*)}\right] \mid \mathcal{D}_{t-1}\right] \\
&= \mathbb{E}_{g\sim\mathrm{GP}_0}\left[\mathbb{E}_{D_t\sim f^*}\left[\frac{\mathcal{L}_t(g)}{\mathcal{L}_t(f^*)} \mid \mathcal{D}_{t-1}\right]\right] \\
&\overset{(i)}{=} \mathbb{E}_{g\sim\mathrm{GP}_0}\left[\frac{\mathcal{L}_{t-1}(g)}{\mathcal{L}_{t-1}(f^*)} \cdot \underbrace{\mathbb{E}_{D_t\sim f^*}\left[\frac{\phi(\frac{Y_t-g(X_t)}{\eta})}{\phi(\frac{Y_t-f^*(X_t)}{\eta})} \mid \mathcal{D}_{t-1}\right]}_{=1}\right] = R_{t-1}(f^*),
\end{aligned}$$

where equality (i) follows because $X_t \in \mathcal{D}_{t-1}$ by virtue of the acquisition funtion being predictable. To conclude the proof, we just need to argue that the braced term in the last expression equals one as claimed. This term can be recognized as integrating a likelihood ratio, which equals one because for any two absolutely continuous distributions P, Q, we have $\mathbb{E}_P(dQ/dP) = \int (dQ/dP)dP = \int dQ = 1$. For readers unfamiliar with this fact, we verify it below by direct integration. Once we condition on $\mathcal{D}_{t-1}$, only Y_t is random, and so the relevant term equals

$$\int_y \frac{\phi(\frac{y-g(X_t)}{\eta})}{\phi(\frac{y-f^*(X_t)}{\eta})}\frac{1}{\eta}\phi\left(\frac{y-f^*(X_t)}{\eta}\right)dy = \int_y \frac{1}{\eta}\phi\left(\frac{y-g(X_t)}{\eta}\right)dy = 1,$$

where the last equality holds simply because a Gaussian PDF with any mean integrates to one. □

Also see Waudby-Smith and Ramdas (2020b) for another application of the prior-posterior ratio martingale. The prior-posterior-ratio is related to the marginal likelihood and the Bayes factor, but the latter two terms are typically used in a Bayesian context, so we avoid their use since the guarantee above is fully frequentist: the expectation $\mathbb{E}_{D_t\sim f^*}$ is not averaging over any prior: no prior is even assumed to necessarily exist in generating f^*, or if it exists it may be incorrectly specified. The most

accurate analogy to past work in frequentist statistics is to interpret this statement as saying that the mixture likelihood ratio is a martingale — a well known fact, implicit in Wald (1947), and exploited in sequential testing (Robbins and Siegmund, 1970) and estimation (Howard et al., 2021). Here, the prior GP_0 plays the role of the mixing distribution. However, our language more directly speaks to how one might apply Bayesian methodology towards frequentist goals in other problems.

3.2. Constructing the confidence sequence

Despite the apparent generality of Lemma 1, it is not directly useful. Indeed, $R_t(f^*)$ is a martingale, but not $R_t(f)$ for any other f, and we obviously do not know f^*. This is where Ville's inequality (5) enters the picture: we use Lemma 1 to construct the following confidence sequence and use Ville's inequality to justify its correctness. Define

$$C_t := \left\{ f \in \mathcal{F} : R_t(f) < \frac{1}{\alpha} \right\}. \tag{11}$$

We claim that f^* is an element of the confidence set C_t, through all of time, with high probability.

Proposition 1. *Consider any (fixed, unknown) $f^* \in \mathcal{F}$ that generates data according to (2), any acquisition function φ_t, and any nontrivial working prior GP_0. Then, C_t defined in (11) is a confidence sequence for f^*:*

$$\Pr(\exists t \in \mathbb{N} : f^* \notin C_t) \leq \alpha.$$

Thus, at any arbitrary data-dependent stopping time τ, we have $\Pr(f^* \notin C_\tau) \leq \alpha$.

Proof. First note that $f^* \notin C_t$ if and only if $R_t(f^*) \geq 1/\alpha$. Recall that $R_t(f^*)$ is a nonnegative martingale by Lemma 1, and note that $R_0(f^*) = 1$. Then, Ville's inequality (5) with $x = 1/\alpha$ implies that $\Pr(\exists t \in \mathbb{N} : R_t(f^*) \geq 1/\alpha) \leq \alpha$. □

C_t is our prior-robust confidence sequence for f^*. For the purposes of the following discussion, let $|C_t|$ denote its size, for an appropriate notion of size such as an ϵ-net covering. Intuitively, if the working prior GP_0 was accurate, which in the frequentist sense means that it put a large amount of mass at f^* relative to other functions, then $|C_t|$ will be (relatively) small. If the working prior GP_0 was inaccurate, which could happen because of a poor choice of kernel hyperparameters, or a poor choice of kernel itself, then $|C_t|$ will be (relatively) large. This degradation of quality ($|C_t|$ relative to accuracy of the prior) is smooth, in the sense that as long as small changes in the GP hyperparameters only change the mass at f^* a little bit, then the corresponding confidence sequence (and hence its size) will also change only slightly. Formalizing these claims is possible by associating a metric over hyperparameters, and proving that if the map from hyperparameters to prior mass is Lipschitz, then the map $|C_t|$ is also Lipschitz, but this is beyond the scope of the current work. Such "sensitivity analysis" can be undertaken if the proposed new ideas are found to be of interest.

C_t is a confidence band for the entire function f^*, meaning that it is uniform over both $\mathcal{X}$ and time, meaning that it provides a confidence interval for $f^*(x)$ that is valid simultaeously for all times and for all x (on the grid, for simplicity). This uniform guarantee is important in practice because the BO algorithm is free to query at any point, and also free to stop at any data-dependent stopping time.

The aforementoned proposition should be compared to Srinivas et al. (2010, Theorem 6, Appendix B), which is effectively a confidence sequence for f (though they did not use that terminology), and yielded the regret bound in their Theorem 3, which is very much in the spirit of our paper.

However, the constants in their Theorems 3, 6 are very loose, and it is our understanding that these are never implemented as such in practice; in contrast, our confidence sequence is essentially tight, with the error probability almost equaling α, because Ville's inequality almost holds with equality for nonnegative martingales (it would be exact equality in continuous time).

Martingales have also been used in other fashions, for example to analyze convergence properties of BO methods; for example, Bect et al. (2019) use (super)martingales to study consistency of sequential uncertainty reduction strategies in the well-specified case.

4. Practical considerations and numerical simulations

Being an infinite dimensional confidence set containing uncountably many functions, even at a fixed time, C_t cannot be explicitly stored on a computer. In order to actually use C_t in practice, two critical questions remains: (a) returning to the very start of Section 3, how should we pick the set of functions $\mathcal{F}$ under consideration? (b) at a fixed time t, and for a fixed new test point x under consideration by the acquisition function for a future query, how can we efficiently construct the confidence interval for $f^*(x)$ that is induced by C_t? These two questions are closely tied together: certain choices of $\mathcal{F}$ in (a) may make step (b) harder. There cannot exist a single theoretically justified way of answering question (a): the type of functions that are "reasonable" will depend on the application.

We describe our approach to tackling these questions in the context of Figure 1. Our answer ties together (a) and (b) using a form of implicit regularization; we suspect there is room for improvement. Our code is available at: https://github.com/willieneis/gp-martingales

4.1. The introductory simulation

In Figure 1, we define two gaussian processes priors, $\text{GP}_0^{(1)}(\mu_1, \kappa_1)$ and $\text{GP}_0^{(2)}(\mu_2, \kappa_2)$. Both covariance matrices κ_1 and κ_2 are defined by a squared exponential kernel, i.e.

$$\kappa(x, x') = \sigma^2 \text{exp}\left(-\frac{(x - x')^2}{2\ell^2}\right), \tag{12}$$

with lengthscale ℓ and signal variance σ^2. In this example, κ_1 has parameters $\{\ell = 1, \sigma^2 = 1.5\}$ and κ_2 has parameters $\{\ell = 3, \sigma^2 = 1\}$. Both GPs have a fixed noise variance $\eta^2 = 0.1$ in model (2). We show the posterior 95% confidence region and posterior samples for $\text{GP}_0^{(1)}(\mu_1, \kappa_1)$ and for $\text{GP}_0^{(2)}(\mu_2, \kappa_2)$ in Figure 1; the top two plots show typical functions drawn from these priors.

Now, we draw a single function from the first prior, $f^* \sim \text{GP}_0^{(1)}(\mu_1, \kappa_1)$ shown as a blue line, which we really treat as a fixed function in this paper. We then draw t observations from this function via

$$X_i \sim \text{Uniform}\left[-10, 10\right], \quad Y_i \sim \mathcal{N}\left(f^*(X_i), \eta^2\right), \quad i = 1, \ldots, t.$$

We compute the posterior GP_t (Eq. 3), under the second prior $\text{GP}_0^{(2)}(\mu_2, \kappa_2)$, and plot the 95% confidence region for $t \in (3, 5, 15, 17, 25, 40)$ in Figure 1, rows 2-4 (shown as blue shaded regions).

We then aim to construct the prior-robust confidence sequence. For each t, we can write the prior-posterior-ratio and confidence sequence for $\alpha = 0.05$ as

$$R_t(f) = \frac{\text{GP}_0^{(2)}(f)}{\text{GP}_t(f)}, \quad \text{and} \quad C_t = \{f \in \mathcal{F} : R_t(f) < 20\}. \tag{13}$$

Next, we describe our procedure for implicitly specifying $\mathcal{F}$ while computing C_t in Section 4.2, and plot it for each $x \in [-10, 10]$ in Figure 1 (shown as yellow/brown shaded regions).

4.2. Implicit specification of $\mathcal{F}$ while computing the confidence interval for $f^*(x)$ at time t

Suppose we are at iteration t of BO, using a Bayesian model with prior $\text{GP}_0(\mu_0, \kappa_0)$. Assume that we have observed data $D_{t-1} = \{(X_i, Y_i)\}_{i=1}^{t-1}$. Assume we have a sequence $X'_1, X'_2, ... \in \mathcal{X}$ over which we'd like to evaluate our acquisition function $\varphi_t(x)$. In BO, this sequence would typically be determined by an acquisition optimization routine, which we can view as some zeroth order optimization algorithm. For each point X' in this sequence we do the following.

(1) Compute the GP posterior. Let $G_t = \{X \in D_{t-1}\} \cup X'$. We will restrict the prior and posterior GP to this set of grid points, making them finite but high-dimensional Gaussians. The infinite-dimensional confidence sequence (or a confidence set at one time instant) for f^* induces a finite-dimensional confidence sequence (set) for its function values at these gridpoints. In other words, for computation tractability, instead of computing the confidence set for the whole function, we can think of each function as $f \in \mathbb{R}^{|G_t|}$, and compute posterior $\text{GP}_t(\mu_t, \kappa_t)$ according to Eq. 3. To avoid unnecessary notation, we will still call the gridded function as f and its induced confidence set as C_t (though in this section they will be G_t-dimensional).

(2) Regularize the posterior-prior ratio. We first define $\widetilde{\text{GP}}_0(\widetilde{\mu}_0, \widetilde{\kappa}_0)$ to be a GP that is very similar to the prior, except slidely wider. More formally, let $\widetilde{\text{GP}}_0 \geq \text{GP}_0$ according to Loewner order, so that $\widetilde{K}_0 - K_0$ is positive semi-definite (where $\widetilde{K}_0$ and K_0 are the covariances matrices associated with $\widetilde{\kappa}_0$ and κ_0). In our experiment, we let $\widetilde{\kappa}_0$ have the same parameters as κ_0, except with a slight larger signal variance (e.g. $(1+\gamma)\sigma^2$, where $\gamma = 10^{-2}$).

One can prove that there exists a Gaussian distribution with density proportional to $\text{GP}_t(f)/\widetilde{\text{GP}}_0(f)$. Define $\widetilde{R}_t^{-1}(f) := \text{GP}_t(f)/\widetilde{\text{GP}}_0(f) = c\mathcal{N}(f|\mu_c, \Sigma_c)$, where $c > 0$. Then

$$\Sigma_c = \left(K_t^{-1} - \widetilde{K}_0^{-1}\right)^{-1}, \quad \mu_c = \Sigma_c \left(K_t^{-1}\mu_t - \widetilde{K}_0^{-1}\widetilde{\mu}_0\right), \quad \text{and} \quad c = \frac{|\widetilde{K}_0|\mathcal{N}(\mu_t|\mu_0, \widetilde{K}_0 - K_t)}{|\widetilde{K}_0 - K_t|}$$

where K_t and $\widetilde{K}_0$ are the covariance matrices associated with κ_t and $\widetilde{\kappa}_0$. Intuitively, $\mathcal{N}(f|\mu_c, \Sigma_c)$ can be viewed as the GP posterior where the prior has been "swapped out" (Neiswanger and Xing, 2017), and replaced with $\text{GP}_0(f)/\widetilde{\text{GP}}_0(f)$. Importantly, note that $\lim_{\gamma \to 0} \widetilde{R}_t(f) = R_t(f)$, the prior-posterior-ratio (Eq. 7), with no restriction on f or $\mathcal{F}$.

Remark: **the role of "belief parameter" γ.** The parameter γ plays important computational and statistical roles. Computationally speaking, numerical stability issues related to invertability are reduced by increasing γ. Statistically, γ implicitly defines the function class $\mathcal{F} \equiv \mathcal{F}_\gamma$ under consideration. $\gamma \to 0$ recovers an unrestricted $\mathcal{F}_0$ that allows arbitrarily wiggly functions, and hence necessarily leads to large and pessimistic C_t. At the other extreme, $\gamma \to \infty$ recovers the usual posterior band used in BO, corresponding to the function class $\mathcal{F}_\infty$ created with a full belief in GP_0 (where complexity of a function can be thought of in terms of the mass assigned by the prior GP_0). To summarize, the "belief parameter" γ plays three roles:

(A) computational, providing numerical stability as γ increases);

(B) statistical, adding regularization that restricts the complexity of functions in C_t, and hence size of C_t, by implicitly defining $\mathcal{F}_\gamma$); and

(C) philosophical, trading a (Bayesian) subjective belief in the prior ($\gamma \to \infty$) with (frequentist) robustness against misspecification ($\gamma \to 0$).

Returning to our simulation, the confidence sequence guarantees derived at $\gamma = 0$ provide robustness against *arbitrary* misspecification of the prior, but our choice of $\gamma = 10^{-2}$ seemed more reasonable if we think the prior is not completely ridiculous. An interesting direction for future work is to figure out how to automatically tune γ in light of the aforementioned tradeoffs.

(3) Compute the confidence sequence. We can then use the confidence sequence

$$C_t = \{f \in \mathbb{R}^{|G_t|} : \widetilde{R}_t^{-1}(f) > \alpha\}.$$

Thus we know that C_t is an ellipsoid defined by the superlevel set of $\widetilde{R}_t^{-1}(f)$. To compute C_t, we can traverse outwards from the posterior-prior ratio mean μ_c until we have found the Mahalanobis distance k to the isocontour $\mathcal{I} = \{f \in \mathbb{R}^{|G_t|} : c\mathcal{N}(f|\mu_c, \Sigma_c) = \alpha\}$.

We can therefore view C_t as the k-sigma ellipsoid of the posterior GP (normal distribution) given by $\mathcal{N}(f|\mu_c, \sigma_c))$. Using this confidence ellipsoid over f, we can compute a lower confidence bound for the value of $f(X')$, which we use as a LCB-style acquisition function $\varphi_t(x)$ at input X'.

To summarize the detailed explanations, our simulations use:

$$\widetilde{R}_t(f) = \frac{\widetilde{\text{GP}}_0(f)}{\text{GP}_t(f)} = \frac{\text{GP}_0(f)}{\text{GP}_t(f)} \frac{\widetilde{\text{GP}}_0(f)}{\text{GP}_0(f)} = R_t(f) \frac{\widetilde{\text{GP}}_0(f)}{\text{GP}_0(f)},$$

where $\widetilde{\text{GP}}_0(f)$ is the same as $\text{GP}_0(f)$, except with the signal variance parameter σ^2 set to $\sigma^2(1+\gamma)$.

BO simulations: GP-LCB versus CS-LCB. We demonstrate BO using C_t (following the procedure outlined above, which we call CS-LCB) and compare it against the GP-LCB algorithm. Results for these experiments are given in Appendix A. Briefly, we applied these methods to optimize an unknown function f^* in both the well-specified and misspecified settings. The findings were as expected: under a misspecified prior, GP-LCB is overconfident about its progress and fails to minimize f^*, while CS-LCB mitigates the issue. For a well-specified prior, both algorithms find the minimizer, but GP-LCB finds it sooner than CS-LCB.

Robustness to misspecified likelihood. Throughout this paper, we have assumed correctness of the likelihood model (2), but what if that assumption is suspect? In the supplement, we repeat the experiment in Figure 1, except when the true noise η^* is half the value η used by the working likelihood (Figure 5), as well as when η^* is double of η (Figure 6). As expected, when the noise is smaller than anticipated, our CS remains robust to the prior misspecification, but when the noise is larger, we begin to notice failures in our CS. We propose a simple fix: define $\check{R}_t := R_t^\beta$, for some $\beta \in (0, 1)$, and construct the CS based on $\check{R}_t$. Figure 7 uses $\beta = 0.75$ and reports promising results. This procedure is inspired by a long line of work in Bayesian inference that proposes raising likelihoods to a power less than one in order to increase robustness (Ibrahim and Chen, 2000; Royall and Tsou, 2003; Grünwald, 2012; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019; Wasserman et al., 2020). Since we desire frequentist coverage guarantees for a Bayesian working model (not assuming correctness of a Bayesian prior), we simply point out that $\check{R}_t$ is not a martingale like R_t, and is instead a supermartingale due to Jensen's inequality. Since Ville's inequality applies, the resulting CS is still valid. Thus it appears at first glance, that one can obtain some amount of robustness against both misspecified priors and likelihoods. However, as mentioned below, merging this idea with hyperparameter tuning and a data-dependent choice of β seems critical for practice.

5. Discussion

Confidence sequences were introduced and studied in depth by Robbins along with Darling, Siegmund and Lai (Darling and Robbins, 1967; Robbins and Siegmund, 1970; Lai, 1976a,b). The topic was subsequently somewhat dormant but came back into vogue due to applications to best-arm identification in multi-armed bandits (Jamieson et al., 2014). Techniques related to nonnegative supermartingales, the mixture method, Ville's inequality, and nonparametric confidence sequences have been studied very recently — see Howard et al. (2020, 2021); Kaufmann and Koolen (2018); Howard and Ramdas (2019); Waudby-Smith and Ramdas (2020a,b) and references therein. They are closely tied to optional stopping, continuous monitoring of experiments and scientific reproducibility (Wald, 1947; Balsubramani, 2014; Balsubramani and Ramdas, 2016; Johari et al., 2017; Shafer et al., 2011; Grünwald et al., 2019; Howard et al., 2021). We are unaware of other work that utilizes them to quantify uncertainty in a BO context.

Many important open questions remain. We describe three directions:

- **Hyperparameter tuning.** It is common in BO practice to tune hyperparameters on the fly (Snoek et al., 2012; Shahriari et al., 2015; Kandasamy et al., 2020; Neiswanger et al., 2019). These can alleviate some problems mentioned in the first page of this paper, but probably only if the kernel is a good match and the function has homogeneous smoothness. We would like to explore if hyperparameter tuning can be integrated into confidence sequences.

  The manner in which we estimate hyperparameters is critical, as highlighted by the recent work of Bachoc (2018) who asks: what happens when we estimate hyperparameters of our kernel using (A) maximum likelihood estimation, or (B) cross-validation, when restricting our attention to some prespecified set of hyperparameters which do not actually capture the true covariance function? The answer turns out to be subtle: the Maximum Likelihood estimator asymptotically minimizes a Kullback-Leibler divergence to the misspecified parametric set, while Cross Validation asymptotically minimizes the integrated square prediction error; Bachoc demonstrates that the two approaches could be rather different in practice.

- **The belief parameter γ.** Can γ be tuned automatically, or updated in a data-dependent way? Further, if we move to the aforementioned hyperparameter tuning setup, can we design a belief parameter γ that can smoothly trade off our belief in the tuned prior against robustness to misspecification? Perhaps we would want $\gamma \to \infty$ with sample size so that as we get more data to tune our priors better, we would need less robustness protection. Further, perhaps we may wish to use a convex combination of kernels, with a weight of $1/(1+\gamma)$ for a simpler kernel (like Gaussian) and a weight of $\gamma/(1+\gamma)$ for a more complex kernel, so that as $\gamma \to \infty$, we not only have more faith in our prior, but we may also allow more complex functions.

- **Computationally tractable choices for $\mathcal{F}$.** While the method introduced in Section 3 is general, some care had to be taken when instantiating it in the experiments of Section 4, because the choice of function class $\mathcal{F}$ had to be chosen to make computation of the set C_t easy. Can we expand the set of computational tools so that these ideas are applicable for other choices of $\mathcal{F}$? How do we scale these methods to work in high dimensions?

The long-term utility of our new ideas will rely on finding suitable answers to the above questions. There are other recent works that study the mean-squared error of GPs under prior misspecification (Beckers et al., 2018), or under potentially adversarial noise in the observation model (Bogunovic

et al., 2020). Their goals are orthogonal to ours (uncertainty quantification), but a cross-pollination of ideas may be beneficial to both efforts.

We end with a cautionary quote from Freedman's Wald lecture (Freedman, 1999):

> *With a large sample from a smooth, finite-dimensional statistical model, the Bayes estimate and the maximum likelihood estimate will be close. Furthermore, the posterior distribution of the parameter vector around the posterior mean must be close to the distribution of the maximum likelihood estimate around truth: both are asymptotically normal with mean 0, and both have the same asymptotic covariance matrix. That is the con- tent of the Bernstein–von Mises theorem. Thus, a Bayesian 95%-confidence set must have frequentist coverage of about 95%, and conversely. In particular, Bayesians and frequentists are free to use each other's confidence sets. However, even for the simplest infinite-dimensional models, the Bernstein–von Mises theorem does not hold (see Cox (Cox, 1993))...The sad lesson for inference is this. If frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed. Bayesians, too, need to proceed with caution in the infinite-dimensional case, unless they are convinced of the fine details of their priors. Indeed, the consistency of their estimates and the coverage probability of their confidence sets depend on the details of their priors.*

Our experiments match the expectations set by the above quote: while the practical appeal of Bayesian credible posterior GP intervals is apparent—they are easy to calculate and visualize—they appear to be inconsistent under even minor prior misspecification (Figure 1), and this is certainly seems to be an infinite-dimensional issue. It is perhaps related to the fact that there is no analog of the Lebesgue measure in infinite dimensions, and thus our finite-dimensional intuition that "any Gaussian prior puts nonzero mass everywhere" does not seem to be an accurate intuition in infinite dimensions.

Acknowledgments

AR thanks Akshay Balsubramani for related conversations. AR acknowledges funding from an Adobe Faculty Research Award, and an NSF DMS 1916320 grant. WN was supported by U.S. Department of Energy Office of Science under Contract No. DE-AC02-76SF00515.

References

François Bachoc. Asymptotic analysis of covariance parameter estimation for gaussian processes in the misspecified case. *Bernoulli*, 24(2):1531–1575, 2018.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint, arXiv:1405.2639*, 2014.

Akshay Balsubramani and Aaditya Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016.

T. Beckers, J. Umlauft, and S. Hirche. Mean square prediction error of misspecified gaussian process models. In *IEEE Conference on Decision and Control (CDC)*, 2018.

Julien Bect, François Bachoc, and David Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

Ilija Bogunovic, Andreas Krause, and Scarlett Jonathan. Corruption-tolerant Gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Dennis D Cox. An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics*, pages 903–923, 1993.

Donald A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

David Freedman. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141, 1999.

Peter Grünwald. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pages 169–183. Springer, 2012.

Peter Grünwald and Thijs Van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *arXiv:1906.07801*, June 2019.

Steven R Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B-testing and best-arm identification. *arXiv preprint arXiv:1906.09712*, 2019.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 2021.

Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 423–439, 2014.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Tuning hyperparameters without grad students: Scalable and robust Bayesian optimisation with Dragonfly. *Journal of Machine Learning Research*, 21(81):1–27, 2020.

Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv:1811.11419*, 2018.

Tze Leung Lai. Boundary crossing probabilities for sample sums and confidence sequences. *The Annals of Probability*, 4(2):299–312, 1976a.

Tze Leung Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265–280, 1976b.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129):2, 1978.

Willie Neiswanger and Eric Xing. Post-inference prior swapping. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2594–2602. JMLR. org, 2017.

Willie Neiswanger, Kirthevasan Kandasamy, Barnabas Poczos, Jeff Schneider, and Eric Xing. Probo: a framework for using probabilistic programming in Bayesian optimization. *arXiv preprint arXiv:1901.11515*, 2019.

Herbert Robbins and David Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429, 1970.

Richard Royall and Tsung-Shan Tsou. Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):391–404, 2003.

Eric Schulz, Maarten Speekenbrink, José M Hernández-Lobato, Zoubin Ghahramani, and Samuel J Gershman. Quantifying mismatch in Bayesian optimization. In *NIPS workshop on Bayesian optimization: Black-box optimization and beyond*, 2016.

Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1): 148–175, 2015.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

Peter Sollich. Gaussian process regression with mismatched models. In *Advances in Neural Information Processing Systems*, pages 519–526, 2002.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

J Ville. *Étude Critique de la Notion de Collectif (PhD Thesis)*. Gauthier-Villars, Paris, 1939.

Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 2020.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020a.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33, 2020b.

Christopher KI Williams and Carl E Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Appendix A. Bayesian Optimization Simulations

We demonstrate BO using our confidence sequence C_t (following the procedure outlined in Section 4.2) and compare it against the GP-LCB algorithm. Results for these experiments are shown below, where we apply these methods to optimize a function f in both the misspecified prior (Figure 2) and correctly specified prior (Figure 3) settings. We find that under a misspecified prior, GP-LCB can yield inaccurate confidence bands and fail to find the optimum of f, while BO using C_t (CS-LCB) can help mitigate this issue.

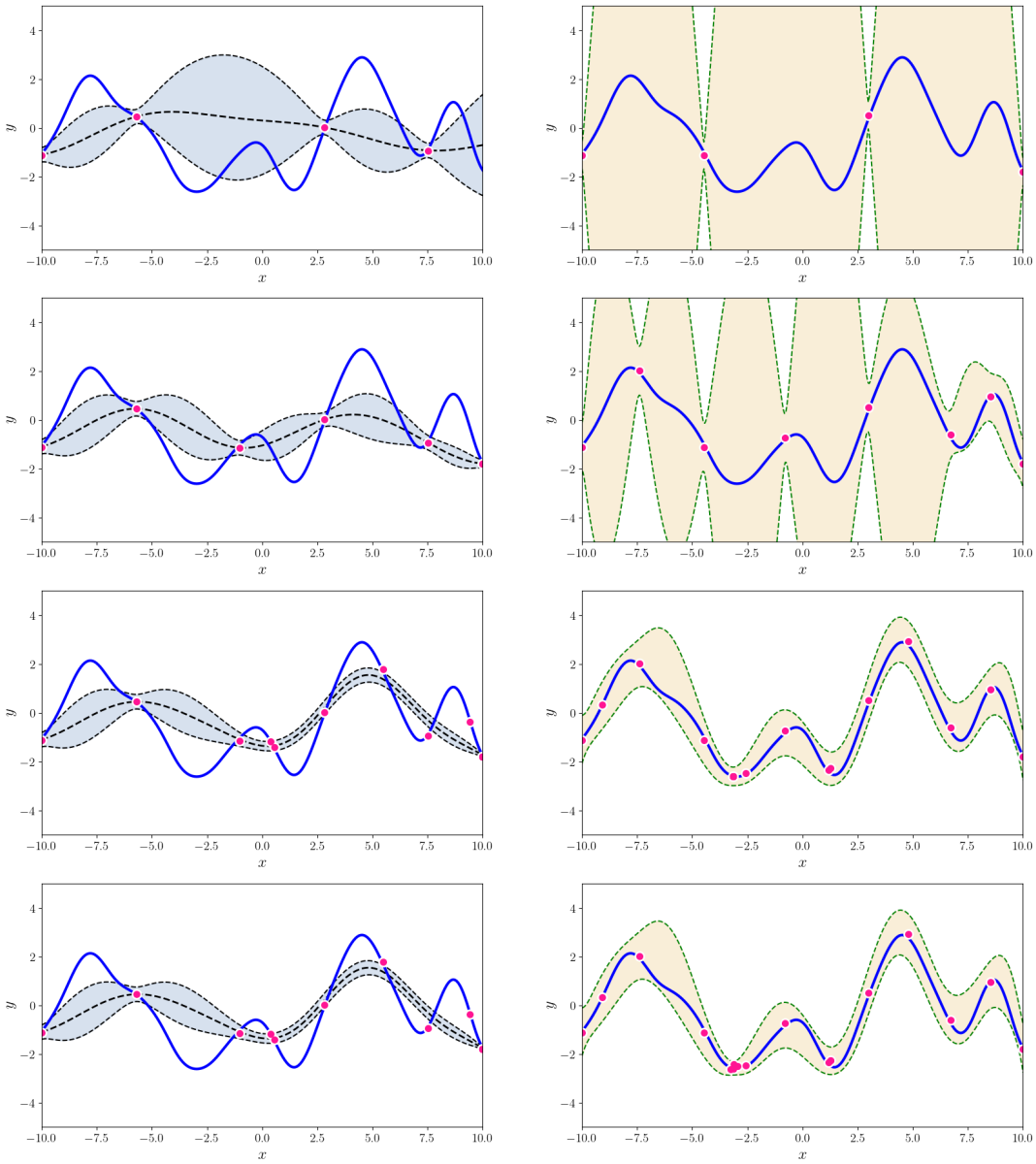

Figure 2: This figure shows GP-LCB (left column) and CS-LCB (right column) for a misspecified prior, showing $t = 3, 7, 18, 25$ (rows 1-4). Here, GP-LCB yields inaccurate confidence bands, repeatedly queries at the wrong point (around $x = 10.0$), and fails to find the minimizer of f, while CS-LCB successfully finds the minimizer (around $x = -3.0$).

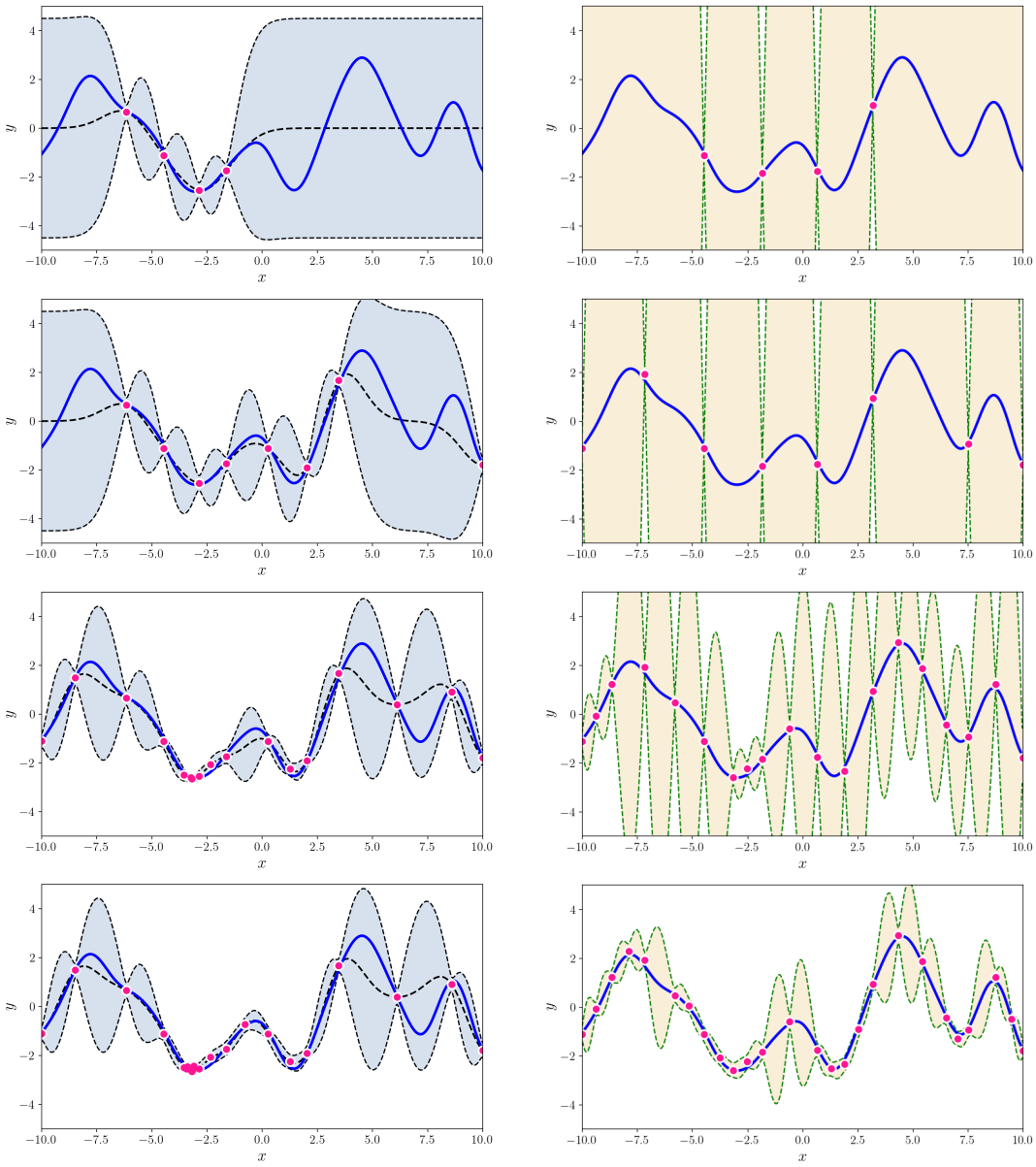

Figure 3: This figure shows GP-LCB (left column) and CS-LCB (right column) for a correctly specified prior, showing $t = 3, 7, 18, 25$ (rows 1-4). Here, both methods find the minimizer of f, though GP-LCB has tighter confidence bands and finds the minimizer sooner than CS-LCB.

Two dimensional benchmark function We also perform a Bayesian optimization experiment on the two dimensional benchmark *Branin* function.[2] In this experiment, we first run Bayesian optimization using the GP-LCB algorithm on a model with a misspecified prior, setting $\{\ell = 7, \sigma^2 = 0.1\}$, and compare it with our CS-LCB algorithm. In both cases, we run each algorithms for 50 steps, and repeat each algorithm over 10 different seeds. We plot results of both algorithms in Fig. 4, along with the optimal objective value. We find that in this misspecified prior setting, CS-LCB converges to the minimal objective value more quickly than GP-LCB.

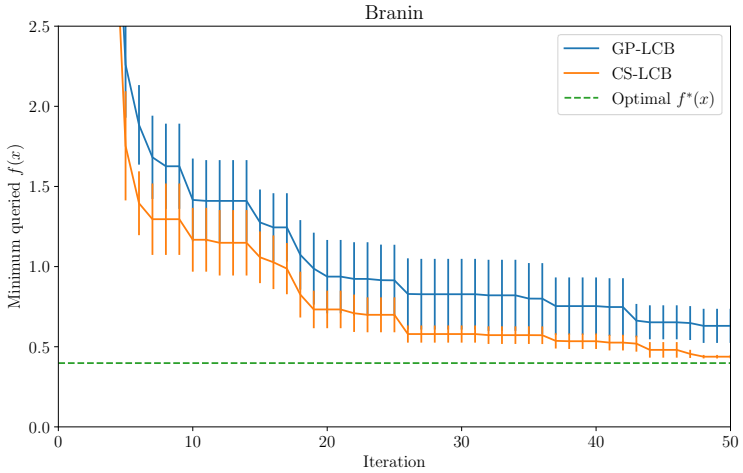

Figure 4: Bayesian optimization using CS-LCB and GP-LCB on the Branin function.

2. Details about this function can be found here: https://www.sfu.ca/~ssurjano/branin.html

Appendix B. Misspecified Likelihood: low/high noise, and powered likelihoods

We next demonstrate BO in the setting where the likelihood is misspecified. In particular, we are interested in the setting where the model assumes noise η, which is not equal to the true noise η^* from which the data is generated. In this case, we demonstrate the fix proposed in Section 4, using powered likelihoods. We show results of this adjustment by repeating the experiment of Figure 1 for $\eta > \eta^*$ (Figure 5) and $\eta < \eta^*$ (Figures 6 and 7).

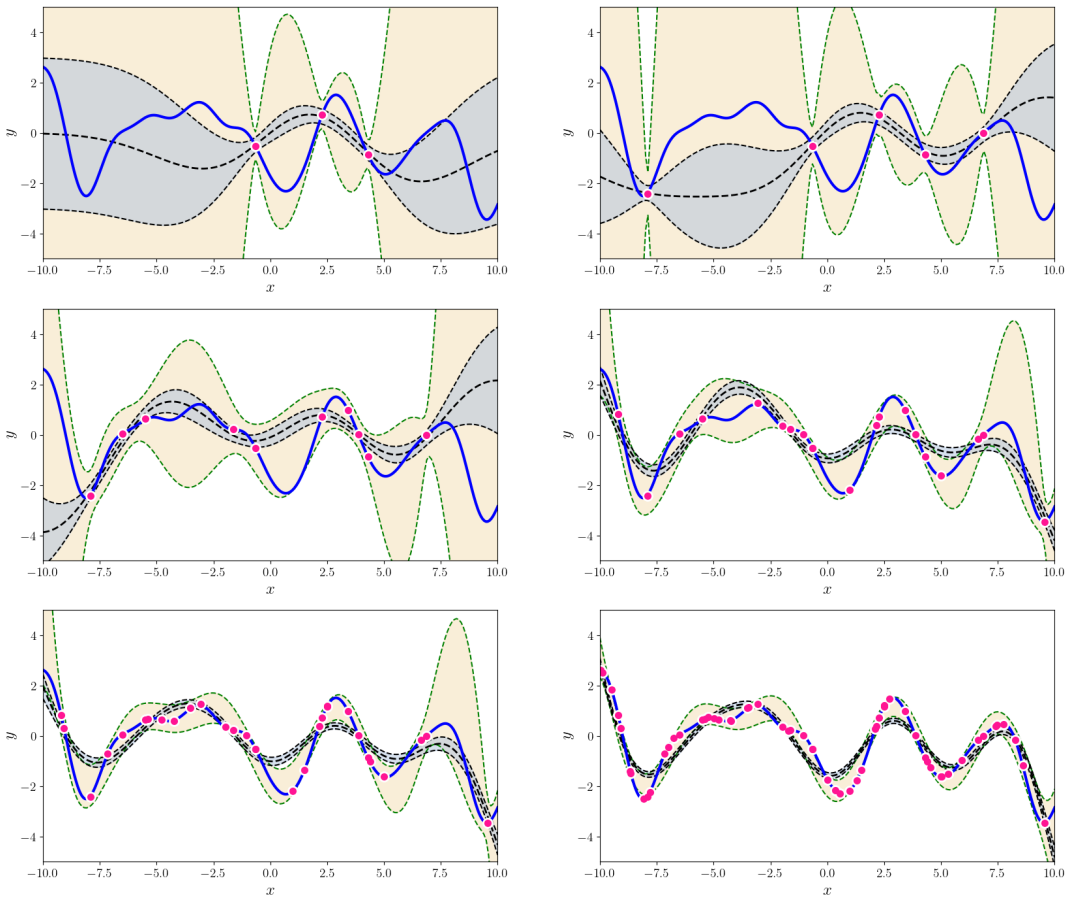

Figure 5: [Low noise setting] We repeat the experiment of Figure 1, but with the true noise η^* of the data being one quarter of the assumed noise η in the working model likelihood (2). Perhaps as expected, the observed behavior is almost indistinguishable from Figure 1 for both the standard GP posterior, which remains incorrectly overconfident, and our method, which covers the true function at all times.

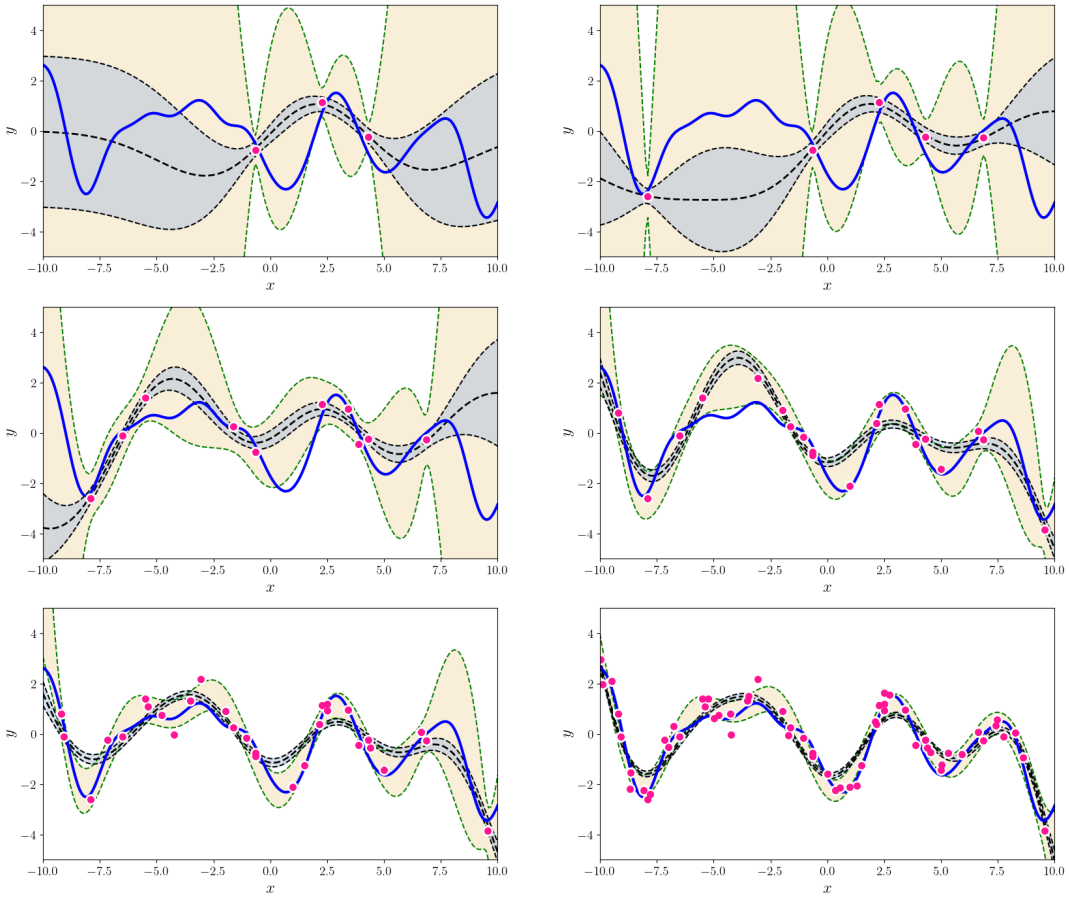

Figure 6: [High noise setting] We repeat the experiment of Figure 1, but with the true noise η^* of the data being four times the assumed noise η in the working model likelihood (2). In these plots, we can see incorrect confidence estimates for our prior-robust CS—for example, when the number of observations $t = 10$ (second row, first column), and when $t = 20$ (second row, second column). As expected, our prior-robust CS is not robust to misspecification of the likelihood.

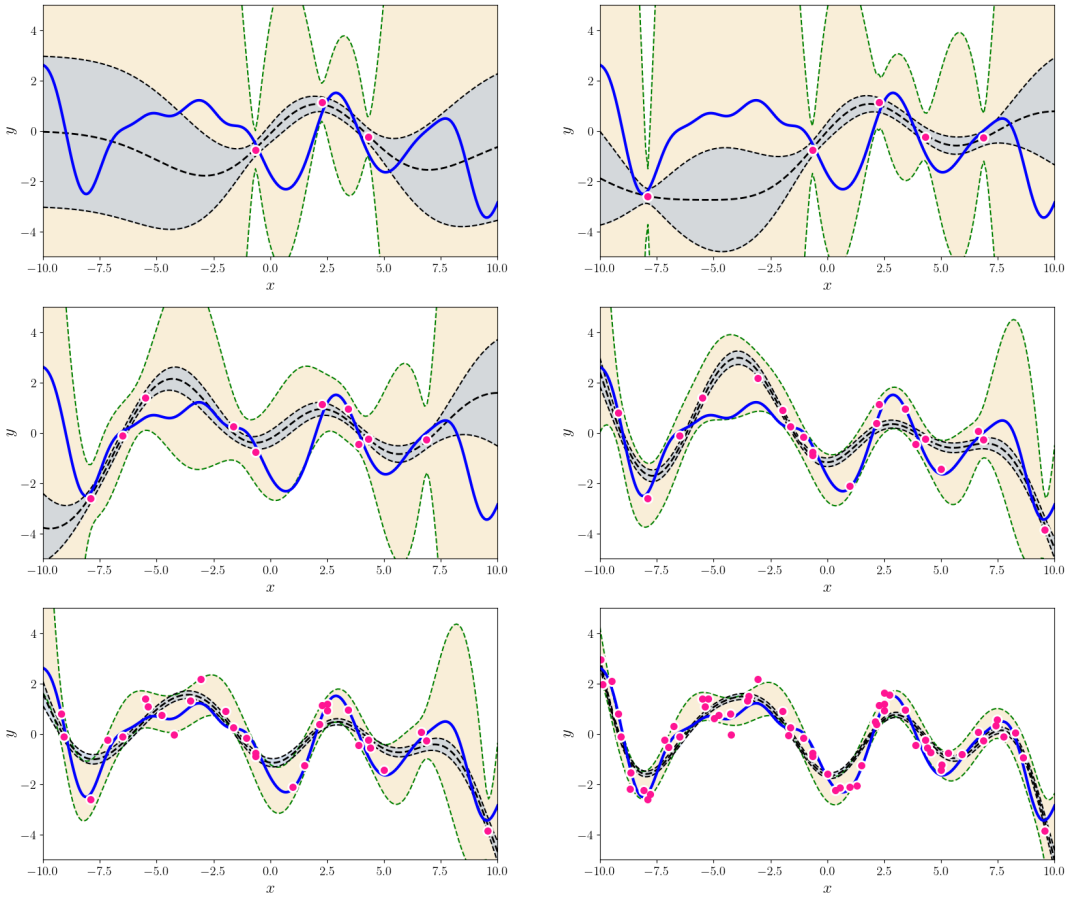

Figure 7: [High-noise setting with our 'powered likelihood' CS] We consider the same setting of Figure 6 when the noise of the data is multiplied by four while the assumed noise in the working model likelihood remains the same. Here, we use a powered likelihood of $\beta = 0.75$ for a more robust confidence sequence, as described at the end of Section 4. Note that the earlier issues at $t = 10$ (second row, first column) and $t = 20$ (second row, second column) are now resolved.