

RESEARCH ARTICLE

Bayesian optimization with informative parametric models via sequential Monte Carlo

Rafael Oliveira^{1,2,*} , Richard Scalzo^{2,3}, Robert Kohn^{2,4}, Sally Cripps^{2,5}, Kyle Hardman^{6,7}, John Close⁷, Nasrin Taghavi⁸ and Charles Lemckert⁹

¹Brain and Mind Centre, The University of Sydney, Sydney, New South Wales, Australia

²Data Analytics for Resources and Environments, Australian Research Council, Sydney, New South Wales, Australia

³School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia

⁴School of Economics, University of New South Wales, Sydney, New South Wales, Australia

⁵Data61, Commonwealth Scientific and Industrial Research Organisation, Sydney, New South Wales, Australia

⁶Nomad Atomics, Canberra, Australian Capital Territory, Australia

⁷Department of Quantum Science & Technology, Australian National University, Canberra, Australian Capital Territory, Australia

⁸School of Engineering and Information Technology, University of New South Wales, Canberra, Australian Capital Territory, Australia

⁹School of Design and the Built Environment, University of Canberra, Canberra, Australian Capital Territory, Australia

*Corresponding author. E-mail: rafael.oliveira@sydney.edu.au

Received: 24 September 2021; **Revised:** 17 December 2021; **Accepted:** 10 February 2022

Keywords: Bayesian inference; Bayesian optimization; inverse problems; sequential Monte Carlo

Abstract

Bayesian optimization (BO) has been a successful approach to optimize expensive functions whose prior knowledge can be specified by means of a probabilistic model. Due to their expressiveness and tractable closed-form predictive distributions, Gaussian process (GP) surrogate models have been the default go-to choice when deriving BO frameworks. However, as nonparametric models, GPs offer very little in terms of interpretability and informative power when applied to model complex physical phenomena in scientific applications. In addition, the Gaussian assumption also limits the applicability of GPs to problems where the variables of interest may highly deviate from Gaussianity. In this article, we investigate an alternative modeling framework for BO which makes use of sequential Monte Carlo (SMC) to perform Bayesian inference with parametric models. We propose a BO algorithm to take advantage of SMC's flexible posterior representations and provide methods to compensate for bias in the approximations and reduce particle degeneracy. Experimental results on simulated engineering applications in detecting water leaks and contaminant source localization are presented showing performance improvements over GP-based BO approaches.

Impact Statement

The methodology we present in this article can be applied to a wide range of problems involving sequential decision making. As presented in a water leak detection experiment, one may apply the algorithm to guide robots in automated monitoring of underground water lines. Other applications include environmental monitoring, chemical synthesis, disease control, and so forth. One of the main advantages of the proposed framework when compared to previous Bayesian optimization approaches is the interpretability of the model, which allows for inferring variables important for analysis and decision support. In addition, practical performance improvements are also observed in experiments.

1. Introduction

Bayesian optimization (BO) offers a principled approach to integrate probabilistic models into processes of decision making under uncertainty (Shahriari et al., 2016). In particular, BO has been successful in optimization problems involving black-box functions and very little prior information, such as smoothness of the objective function. Examples include hyperparameter tuning (Snoek et al., 2012), robotic exploration (Souza et al., 2014), and chemical design (Shields et al., 2021), and disease control (Spooner et al., 2020). Most of its success relies on the flexibility and well-understood properties of nonparametric modeling frameworks, particularly Gaussian process (GP) regression (Rasmussen and Williams, 2006). Although nonparametric models are usually the best approach for problems with scarce prior information, they offer little in terms of interpretability and may be a suboptimal guide when compared to expert parametric models. In these expert models, parameters are attached to variables with physical meaning and reveal aspects of the nature of the problem, since models are derived from domain knowledge.

As a motivating example, consider the problem of localizing leaks in underground water distribution pipes (Mashford et al., 2012). Measurements from pipe monitoring stations are usually sparse and excavations are costly (Sadeghioon et al., 2014). As a possible alternative, microgravimetric sensors have recently allowed detecting gravity anomalies of parts in a billion (Brown et al., 2016; Hardman et al., 2016), making them an interesting data source for subsurface investigations (Hauge and Kolbjørnsen, 2015; Rossi et al., 2015). One could then design a GP-based BO algorithm to localize a leak in a pipe by searching for a maximum in the gravity signal on the surface due to the heavier wet soil. The determined 2D location, however, tells nothing of the depth or the volume of leaked water. In this case, a physics-based probabilistic model of a simulated water leak could better guide BO and the decision-making end users. Bayesian inference on complex parametric models, however, is usually intractable, requiring the use of sampling-based techniques, like Markov chain Monte Carlo (MCMC) (Andrieu et al., 2003), or variational inference methods (Bishop, 2006; Ranganath et al., 2014). Either of these approaches can lead to high computational overheads during the posterior updates in a BO loop.

In this article, we investigate a relatively unexplored modeling approach in the BO community which offers a balanced trade-off between MCMC and approximate inference methods for problems where domain knowledge is available. Namely, we consider sequential Monte Carlo (SMC) (Doucet et al., 2001), also known as particle filtering, algorithms as posterior update mechanisms for BO. SMC has an accuracy-speed trade-off controlled by the number of particles it uses, and it suffers less of the drawbacks of other approximate inference methods (Bishop, 2006), while still enjoying asymptotic convergence guarantees similar to MCMC (Crisan and Doucet, 2002; Beskos et al., 2014). SMC methods have traditionally been applied to state-space models for localization (Thrun et al., 2006) and tracking (Doucet et al., 2000) problems, but SMC has also found use in more general areas, including experimental design (Kuck et al., 2006) and likelihood-free inference (Sisson et al., 2007).

As contributions, we present an approach to efficiently incorporate SMC within BO frameworks. In particular, we derive an acquisition function to take advantage of the flexibility of SMC's particle distributions by means of empirical quantile functions. Our approach compensates for the approximation bias in the empirical quantiles by taking into account SMC's effective sample size (ESS). We also propose methods to reduce the correlation and bias among SMC samples and improve its predictive power. Lastly, experimental results demonstrate practical performance improvements over GP-based BO approaches.

2. Related Work

Other than the GP approach, BO frameworks have applied a few other methods for surrogate modeling, including linear parametric models with nonlinear features (Snoek et al., 2015), Bayesian neural networks (Springenberg et al., 2016), and random forests (Hutter et al., 2011). Linear models and limits of Bayesian neural networks when the number of neurons approaches infinity can be directly related to GPs (Rasmussen and Williams, 2006; Khan et al., 2019). Mostly these approaches, however, consider a black-box setting for the optimization problem, where very little is known about the stochastic process

defining the objective function. In this article, we take BO instead toward a gray-box formulation, where we know a parametric structure which can accurately describe the objective function, but whose true parameters are unknown.

SMC has previously been combined with BO in GP-based approaches. Benassi et al. (2012) applies SMC to the problem of learning the hyperparameters of the GP surrogate during the BO process by keeping and updating a set of candidate hyperparameters according to the incoming observations. Bijl et al. (2016) provide a method for Thompson sampling (Russo and Van Roy, 2016) using SMC to keep track of the distribution of the global optimum. These approaches still use GPs as the main modeling framework for the objective function. Lastly and more related to this article, Dalibard et al. (2017) presents an approach to use SMC for inference with semiparametric models, where one combines GPs with informative parametric models. Their framework is tailored to automatically tuning computer programs following dynamical systems, where the system state transitions. In contrast, our approach is based on a static formulation of SMC, where the system state corresponds to a probabilistic model's parameters vector, which does not change over time. Simply adapting a dynamics-based SMC model to a static system is problematic due to particle degeneracy in the lack of a transition model (Doucet et al., 2000). We instead apply MCMC to move and rejuvenate particles in the static SMC framework, as originally proposed by Chopin (2002).

In the multiarmed bandits literature, whose methods have frequently been applied to BO (Srinivas et al., 2010; Wang and Jegelka, 2017; Berkenkamp et al., 2019), SMC has also been shown as a useful modeling framework (Kawale et al., 2015; Urteaga and Wiggins, 2018). In particular, Urteaga and Wiggins (2018) present a SMC approach to bandits in dynamic problems, where the reward function evolves over time. A generalization of their approach has recently been proposed to include linear and discrete reward functions (Urteaga and Wiggins, 2019) with support by empirical results. Bandit problems seek policies which maximize long-term payoffs. In this article, we instead focus on investigating and addressing the effects of the SMC approximation on a more general class of problems. We also provide experimental results on applications where domain knowledge offers informative models.

Lastly, BO algorithms provide model-based solutions for black-box derivative-free optimization problems (Rios and Sahinidis, 2013). In this context, there are plenty of other model-free approaches, such as evolutionary algorithms, which include the popular covariance matrix adaptation evolution strategy (CMA-ES) algorithm (Arnold and Hansen, 2010). However, these approaches are usually not focused on improving data efficiency as the usually require hundreds or even thousands of steps to converge to a global optimum. In contrast, BO algorithms are usually applied to problems where the number of evaluations of the objective function is very limited, usually to the order of tens of evaluations, due to a high cost of collecting observations (Shahriari et al., 2016), for example, drilling a hole to find natural gas deposits. The algorithm we propose in this article also targets this kind of use case, with the difference that we apply more informative predictive models than the usual GP-based formulations.

3. Preliminaries

In this section, we specify our problem setup and review relevant theoretical background. We consider an optimization problem over a function $f: \mathcal{X} \rightarrow \mathbb{R}$ within a compact search space $\mathcal{S} \subset \mathcal{X} \subseteq \mathbb{R}^d$:

$$\mathbf{x}^* \in \underset{\mathbf{x} \in \mathcal{S}}{\operatorname{argmax}} f(\mathbf{x}). \quad (1)$$

We assume a parametric formulation for $f(\mathbf{x}) = h(\mathbf{x}, \theta)$, where $h: \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ has a known form, but $\theta \in \Theta \subset \mathbb{R}^m$ is an unknown parameter vector. The only prior information about θ is a probability distribution $p(\theta)$. We are allowed to collect up to T observations o_t distributed according to an observation (or likelihood) model $p(o_t | \theta, \mathbf{x}_t)$, for $t \in \{1, \dots, T\}$. For instance, in the classical white Gaussian noise setting, we have $o_t = f(\mathbf{x}_t) + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, so that $p(o_t | \theta, \mathbf{x}_t) = \mathcal{N}(o_t; h(\mathbf{x}_t, \theta), \sigma_\varepsilon^2)$. However, our optimization problem is not restricted by Gaussian noise assumptions.

As formulated above, we highlight that the objective function f is a *black-box* function, and the model $h: \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ is simply an assumption over the real function, which is unknown in practice. For example, in one of our experiments, we have f as the gravity measured on the surface above an underground water leak of unknown location. In this case, gradients and analytic formulations for the objective function are not available. Therefore, we need derivative-free optimization algorithms to solve Equation (1). In addition, we assume that the budget of observations T is relatively small (in the orders of tens or a few hundreds) and incrementally built, so that a maximum-likelihood or interpolation approach, as common in response surface methods (Rios and Sahinidis, 2013), would lead to suboptimal results, as it would not properly account for the uncertainty due to the limited amount of data and their inherent noise. We then seek a Bayesian approach to solve Equation (1). In the following, we revise theoretical background on the main components of the method we propose.

3.1. Bayesian optimization

BO approaches the problem in Equation (1) by placing a prior distribution over the objective function f (Shahriari et al., 2016), typically represented by a GP (Rasmussen and Williams, 2006). Under the GP assumption, considering Gaussian observation noise, finite collections of function evaluations and observations are joint normally distributed, which allows deriving closed-form expressions for the posterior $p(f(\mathbf{x})|\{\mathbf{x}_t, o_t\}_{t=1}^T)$. BO sequentially collects a set of observations $\mathcal{D}_T := \{\mathbf{x}_t, o_t\}_{t=1}^T$ by maximizing an acquisition function:

$$\mathbf{x}_t \in \operatorname{argmax}_{\mathbf{x} \in \mathcal{S}} a(\mathbf{x}|\mathcal{D}_{t-1}), \quad t \in \{1, \dots, T\}. \tag{2}$$

The acquisition function informs BO of the utility of collecting an observation at a given location $\mathbf{x} \in \mathcal{S}$ based on the posterior predictions for $f(\mathbf{x})$. For example, with a GP model, one can apply the upper confidence bound (UCB) criterion:

$$a(\mathbf{x}|\mathcal{D}_{t-1}) := \mu_{t-1}(\mathbf{x}) + \beta_t \sigma_{t-1}(\mathbf{x}), \tag{3}$$

where $f(\mathbf{x})|\mathcal{D}_{t-1} \sim \mathcal{N}(\mu_{t-1}(\mathbf{x}), \sigma_{t-1}^2(\mathbf{x}))$ represents the GP posterior at iteration $t \geq 1$, and β_t is a parameter which can be annealed over time to maintain a high-probability confidence interval over $f(\mathbf{x})$ (Chowdhury and Gopalan, 2017). Besides the UCB, the BO literature is filled with many other types of acquisition functions, including expected improvement (Jones et al., 1998; Bull, 2011), Thompson sampling (Russo and Van Roy, 2016), and information-theoretic criteria (Hennig and Schuler, 2012; Hernández-Lobato et al., 2014; Wang and Jegelka, 2017).

3.2. Sequential Monte Carlo

SMC algorithms are Bayesian inference methods which approximate posterior distributions via a finite set of samples (Doucet et al., 2001). The random variables of interest $\theta_t \in \Theta$ are modeled as the state of a dynamical system which evolves over time $t \in \mathbb{N}$ based on a sequence of observations o_1, \dots, o_t . SMC algorithms rely on a conditional independence assumption $p(o_t|\theta_t, o_1, \dots, o_{t-1}) = p(o_t|\theta_t)$, which allows for the decomposition of the posterior over states into sequential updates¹:

$$p(\theta_1, \dots, \theta_t | o_1, \dots, o_t) = p(\theta_1, \dots, \theta_{t-1} | o_1, \dots, o_{t-1}) \frac{p(o_t|\theta_t)p(\theta_t|\theta_{t-1})}{p(o_t|o_1, \dots, o_{t-1})}. \tag{4}$$

¹ In these and the following equations, we omit the dependence on the observation locations $\{\mathbf{x}_j\}_{j=1}^t \in \mathcal{X}$ to avoid notation clutter. However, we will make this dependence explicit whenever needed to emphasize that locations are also part of the observed data.

Algorithm 1: Basic SMC

```

1  $\{\theta_0^i\}_{i=1}^n \sim p(\theta)$ 
2 for  $t \in \{1, \dots, T\}$  do
3    $\hat{\theta}_t^i \sim p(\theta_t^i | \theta_{t-1}^i), \quad i \in \{1, \dots, n\}$  // Move
4    $w_t^i := p(o_t | \hat{\theta}_t^i), \quad i \in \{1, \dots, n\}$  // Re-weight particles
5    $\{\theta_t^i\}_{i=1}^n \sim \sum_{i=1}^n w_t^i \delta_{\hat{\theta}_t^i}$  // Resample (with replacement)
```

Based on this decomposition, SMC methods maintain an approximation of the posterior $p(\theta_t | o_1, \dots, o_t)$ based on a set of *particles* $\{\theta_t^i\}_{i=1}^n \subset \Theta$, where each particle θ_t^i represents a sample from the posterior. Despite the many variants of SMC available in the literature (Doucet et al., 2001; Naesseth et al., 2019), in its basic form, the SMC algorithm is simple and straightforward to implement, given a *transition model* $p(\theta_t | \theta_{t-1})$ and an *observation model* $p(o_t | \theta_t)$. For instance, a time-varying spatial model $h : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ may have a Gaussian observation model $p(o_t | \theta_t) = p(o_t | \theta_t, \mathbf{x}_t) = \mathcal{N}(o_t; h(\mathbf{x}_t, \theta_t), \sigma_\varepsilon^2)$, with $\sigma_\varepsilon > 0$, and the transition model $p(\theta_t | \theta_{t-1})$ may be given by a known stochastic partial differential equation describing the system dynamics.

Basic SMC follows the procedure outlined in Algorithm 1. SMC starts with a set of particles initialized as samples from the prior $\{\theta_0^i\}_{i=1}^n \sim p(\theta)$. At each time step, the algorithm proposes new particles by moving them according to the state transition model $p(\theta_t | \theta_{t-1})$. Given a new observation o_t , SMC updates its particles distribution by first weighing each particle according to its likelihood $w_t^i = p(o_t | \theta_t^i)$ under the new observation. A set of n new particles is then sampled (with replacement) from the resulting weighted empirical distribution. The new particles distribution is then carried over to the next iteration until another observation is received. Additional steps can be performed to reduce the bias in the approximation and improve convergence rates (Chopin, 2002; Naesseth et al., 2019).

4. BO via SMC

In this section, we present a quantile-based approach to solve Equation (1) via BO. The method uses SMC particles to determine a high-probability UCB on f via an empirical quantile function. We start with a description of the proposed version of the SMC modeling framework, followed by the UCB approach.

4.1. SMC for static models

SMC algorithms are typically applied to model time-varying stochastic processes. However, if Algorithm 1 is naively applied to a setting where the state θ_t remains constant over time, as in our case, $p(\theta_t | \theta_{t-1})$ becomes a Dirac delta distribution $D_{\theta_{t-1}}$, and the particles no longer move to explore the state space. To address this shortcoming, Chopin (2002) proposed an SMC variant for static problems that uses a MCMC kernel to promote transitions of the particles distribution. The MCMC kernel is configured with the true posterior at time t , $p(\theta | o_1, \dots, o_t)$, as its invariant distribution, so that the particles transition distribution becomes a Metropolis-Hastings update:

$$p(\theta | \theta') = \min \left\{ 1, \frac{p(o_1, \dots, o_t | \theta) p(\theta) q(\theta' | \theta)}{p(o_1, \dots, o_t | \theta') p(\theta') q(\theta | \theta')} \right\}, \quad \theta, \theta' \in \Theta, \tag{5}$$

where $q(\theta | \theta')$ is any valid MCMC proposal (Andrieu et al., 2003). Efficient proposals, such as those by Hamiltonian Monte Carlo (HMC) (Neal, 2011), can ensure exploration and decorrelate particles.

4.1.1. The static SMC algorithm

The resulting SMC posterior update algorithm is presented in Algorithm 2. In contrast with Algorithm 1, we accumulate old weights and use them alongside the particles to compute an ESS. As suggested by previous approaches (Del Moral et al., 2012), one may keep a good enough posterior

approximation by resampling the particles only when their ESS goes below a certain pre-specified threshold² $n_{\min} \leq n$, which reduces computational costs. Whenever this happens, we also move the particles according to the MCMC kernel. This allows us to maintain a diverse particle set, avoiding particle degeneracy. Otherwise, case the ESS is still acceptable, we simply update the particle weights and continue.

Algorithm 2: Static SMC

Input: $\{\theta_{t-1}^i, w_{t-1}^i\}_{i=1}^n, o_t$
Output: $\{\theta_t^i, w_t^i\}_{i=1}^n$

- 1 $\hat{w}_t^i := w_{t-1}^i p(o_t | \theta^i), i \in \{1, \dots, n\}$ // Re-weight particles
- 2 $\hat{n}_{\text{ESS}} = \text{ESS}(\{\theta^i, \hat{w}_t^i\}_{i=1}^n)$ // Compute effective sample size
- 3 **if** $\hat{n}_{\text{ESS}} < n_{\min}$ **then**
- 4 $\{\hat{\theta}_t^i\}_{i=1}^n \sim \sum_{i=1}^n \hat{w}_t^i \delta_{\theta^i}$ // Resample with replacement
- 5 $\theta_t^i \sim \kappa(\cdot | \hat{\theta}_t^i), w_t^i := 1, i \in \{1, \dots, n\}$ // Move following transition kernel
- 6 **else**
- 7 $\theta_t^i := \theta_{t-1}^i, w_t^i := \hat{w}_t^i, i \in \{1, \dots, n\}$

4.1.2. Computational complexity

The MCMC-based approach incurs an $O(nt)$ computational cost per SMC iteration that grows linearly with the number of particles n and observations t , instead of remaining constant at $O(n)$ across iterations as in basic SMC (Algorithm 1). However, this increased computational cost is still relatively low in terms of run-time complexity when compared to the GP modeling approach. Namely, the time complexity for GP updates is $O(t^3)$, which scales cubically with the number of data points (Rasmussen and Williams, 2006). The SMC approach is also low in terms of memory storage costs, as it does not require storing a covariance matrix (quadratic scaling $O(t^2)$) over the data points, but only the points themselves and a set of particles with fixed size, that is, an $O(n+t)$ memory cost.

4.2. An UCB algorithm for learning with SMC models

We now define an acquisition function based on an UCB strategy. Compared to the GP-based UCB (Srinivas et al., 2010), which is defined in terms of posterior mean and variance, SMC provides us with flexible nonparametric distributions which can adjust well to non-Gaussian cases. A natural approach to take advantage of this fact is to define the UCB in terms of quantile functions (Kaufmann et al., 2012).

The quantile function q of a real random variable \tilde{s} with distribution $P_{\tilde{s}}$ corresponds to the inverse of its cumulative distribution. We define the (upper) quantile function as:

$$q_{\tilde{s}}(\tau) := \inf \{s \in \mathbb{R} \mid P_{\tilde{s}}(\tilde{s} \leq s) \geq \tau\}, \quad \tau \in (0, 1). \quad (6)$$

For probability distributions with compact support, we also define $q_{\tilde{s}}(0)$ as the minimum and $q_{\tilde{s}}(1)$ as the maximum of \tilde{s} , respectively, of the support of $P_{\tilde{s}}$. In essence, a quantile $q_{\tilde{s}}(\tau)$ is an upper bound on the value of \tilde{s} which holds with probability at least τ .

4.2.1. Empirical quantile functions

In our case, we want to bound $f(\mathbf{x})$ at any $\mathbf{x} \in \mathcal{S}$ with high probability. However, we have no direct access to the distribution of $f(\mathbf{x})$. Instead, we use SMC to keep track of its posterior by a set of weighted particles $\{\theta_t^i, w_t^i\}_{i=1}^n$ at each iteration t . As $f(\mathbf{x}) = h(\mathbf{x}, \theta)$, we have a corresponding empirical approximation to the

²Usually n_{\min} is set to a fraction of n , with 50% being a common setting.

posterior probability measure of $f(\mathbf{x})$, denoted by $\hat{P}_{f(\mathbf{x})}^t := \sum_{i=1}^n w_i^t \delta_{h(\mathbf{x}, \theta_i)}$. With the empirical posterior, we approximate the quantile function on $f(\mathbf{x})$ as:

$$q_{f(\mathbf{x})}(\tau) \approx \hat{q}_t(\mathbf{x}, \tau) := \inf \left\{ s \in \mathbb{R} \mid \hat{P}_{f(\mathbf{x})}^t (f(\mathbf{x}) \leq s) \geq \tau \right\}, \quad \tau \in (0, 1). \tag{7}$$

In practice, computing $\hat{q}_t(\mathbf{x}, \tau)$ amounts to sorting realizations of $f(\mathbf{x})$, $s_1 \leq s_2 \leq \dots \leq s_n$, according to the empirical posterior $s_i \sim \hat{P}_t(f(\mathbf{x}))$ and finding the first element s_i whose cumulative weight $\sum_{j \leq i} w_j$ is greater than τ . For $\hat{q}_t(\mathbf{x}, \tau)$ with τ set to 0 or 1, the procedure would reduce to returning the supremum or infimum, respectively, in the support of the distribution of $f(\mathbf{x})$, which correspond to $+\infty$ or $-\infty$ in the unbounded case.

The quality of empirical quantile approximations for confidence intervals can be quantified via the Dvoretzky–Kiefer–Wolfowitz inequality, which bounds the error on the empirical cumulative distribution function (CDF). Massart, 1990 provided a tight bound in that regard.

Lemma 1 (Massart (1990)). *Let $P_n := \frac{1}{n} \sum_{i=1}^n d_{\xi^i}$ denote the empirical measure derived from $n \in \mathbb{N}$ i.i.d. real-valued random variables $\{\xi^i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P$. For any $\zeta \sim P$, assume the cumulative distribution function $s \mapsto P[\zeta \leq s]$ is continuous, $s \in \mathbb{R}$. Then we have that:*

$$\forall c \in \mathbb{R}, \quad \mathbb{P} \left\{ \sup_{s \in \mathbb{R}} |P_n[\zeta \leq s] - P[\zeta \leq s]| > c \right\} \leq 2 \exp(-2nc^2).$$

Although SMC samples are inherently biased with respect to the true posterior, we explore methods to transform SMC particles into approximately i.i.d. samples from the true posterior. For instance, the MCMC moves follow the true posterior, which turn the particles into approximate samples of the true posterior. In addition, other methods such as density estimation and the jackknife (Efron, 1992) allow for decorrelation and bias correction. We explore these approaches in Section 5. As a consequence of Lemma 1, we have the following bound on the empirical quantiles.

Theorem 1 (Szörényi et al. (2015), Proposition 1). Under the assumptions in Lemma 1, given $\mathbf{x} \in \mathcal{X}$ and $\delta \in (0, 1)$, the following holds for every $\rho \in [0, 1]$ with probability at least $1 - \delta$:

$$\forall n \geq 1, \quad \hat{q}_\xi(\rho - c_n(\delta)) \leq q_\xi(\rho) \leq \hat{q}_\xi(\rho + c_n(\delta)), \tag{8}$$

where $c_n(\delta) := \sqrt{\frac{1}{2n} \log \frac{\pi^2 n^2}{3\delta}}$.

Note that Theorem 1 provides a confidence interval around the true quantile based on i.i.d. empirical approximations. In the case of SMC, however, its particles do not follow the i.i.d. assumption. We address this problem in our method, with further approaches for bias reduction presented in Section 5.

4.2.2. Quantile-based UCB

Given the empirical quantile function in Equation (7), we define the following acquisition function:

$$a(\mathbf{x} | \mathcal{D}_t) := \hat{q}_t(\mathbf{x}, \delta_t), \quad t \geq 0, \tag{9}$$

where $\delta_t \in (0, 1)$ is a parameter which can be adjusted to maintain a high-probability upper confidence bound, as in GP-UCB (Srinivas et al., 2010). In contrast to a GP-UCB analogy, we are not using a Gaussian approximation based on the posterior mean and variance of the SMC predictions. We are directly using SMC’s nonparametric approximation of the posterior, which can take arbitrary shape.

Theorem 1 allows us to bound the difference between the quantile function and its empirical approximation in the case of n i.i.d. samples. Given $\mathbf{x} \in \mathcal{X}$ and $\delta \in (0, 1)$, the following holds for every $\tau \in [0, 1]$ with probability at least $1 - \delta$:

$$\forall n \geq 1, \quad \hat{q}(\mathbf{x}, \tau - c_n(\delta)) \leq q(\mathbf{x}, \tau) \leq \hat{q}(\mathbf{x}, \tau + c_n(\delta)), \tag{10}$$

where $c_n(\delta) := \sqrt{\frac{1}{2n} \log \frac{\pi^2 n^2}{3\delta}}$. In the non- i.i.d. case, however, the approximation above is no longer valid. We instead replace n by the ESS n_{ESS} , which is defined as the ratio between the variance of an i.i.d. Monte Carlo estimator and the variance, or mean squared error (Elvira et al., 2018), of the SMC estimator (Martino et al., 2017). With n_{ESS}^t denoting the ESS at iteration t , we set:

$$\delta_t := 1 - \delta + c_{n_{\text{ESS}}^t}(\delta). \tag{11}$$

Several approximations for the ESS are available in the SMC literature (Martino et al., 2017; Huggins and Roy, 2019), which are usually based on the distribution of the weights of the particles. A classical example is $\hat{n}_{\text{ESS}} := \frac{(\sum_{i=1}^n w_i)^2}{\sum_{i=1}^n w_i^2}$ (Doucet et al., 2000; Huggins and Roy, 2019). In practice, the simple substitution of n by \hat{n}_{ESS} defined above can be enough to compensate for the correlation and bias in the SMC samples. In Section 5, we present further approaches to reduce the SMC approximation error with respect to an i.i.d. sample-based estimator.

4.2.3. The SMC-UCB algorithm

In summary, the method we propose is described in Algorithm 3, which we refer to as SMC upper confidence bound (SMC-UCB). The algorithm follows the general BO setup with a few exceptions. We start by drawing i.i.d. samples from the model parameters’ prior and equal weights, following the SMC setup (see Algorithm 2). At each iteration $t \leq T$, we choose a query point by maximizing the acquisition function given by the empirical quantile. We then collect an observation at the selected location. The SMC particles distribution is updated using the new observation, and the algorithm proceeds. As in the usual BO setup, this loop repeats for a given budget of T evaluations.

Algorithm 3: SMC-UCB

```

1  $\{\theta_0^i\}_{i=1}^n \sim p(\theta)$ 
2  $w_0^i := 1, i \in \{1, \dots, n\}$ 
3 for  $t \in \{1, \dots, T\}$  do
4    $\mathbf{x}_t \in \text{argmax}_{\mathbf{x} \in \mathcal{S}} \hat{q}_{t-1}(\mathbf{x}, \delta_t)$  // Select query point based on quantile
5
6    $o_t \leftarrow \text{Observe } f \text{ at } \mathbf{x}_t$  // Collect observation
7
8    $\{\theta_t^i, w_t^i\}_{i=1}^n \leftarrow \text{SMC}(\{\theta_{t-1}^i, w_{t-1}^i\}_{i=1}^n, o_t)$  // Update SMC estimates

```

5. Drawing Independent Posterior Samples from SMC

Applying to SMC estimates a result similar to Theorem 1 would allow us to provide accurate confidence intervals based on SMC. This result is not directly applicable, however, due to the bias and correlation among SMC samples. Hence we propose an approach to decorrelate samples and remove the bias in SMC estimates.

The approach we propose consists of three main steps. We first decorrelate SMC samples by estimating a continuous density over the SMC particles. The resulting probability distribution is then used as a proposal for importance sampling. The importance sampling estimate provides us with a still biased estimate of expected values under the true posterior. To remove this bias, we finally apply the jackknife method (Efron, 1992).

Throughout this section, our main object of interest will be the expected value of a function of the parameters $u : \Theta \rightarrow \mathbb{R}$, which we will assume to be bounded and continuous. We want to minimize the bias in expected value estimates:

$$\varepsilon_u := \mathbb{E}[u(\theta)|\mathcal{D}_t] - \hat{\mathbb{E}}[u(\theta)|\mathcal{D}_t], \tag{12}$$

where $\hat{\mathbb{E}}[\cdot]$ denotes the SMC-based expected value estimator.

5.1. Density estimation

As a first step, we consider the problem of decorrelating SMC particles by sampling from an estimated continuous probability distribution q over the particles. The simplest approach would be to use a Gaussian approximation to the parameters posterior density. In this case, we approximate $p(\theta|\mathcal{D}_t) \approx \hat{p}_t(\theta) := \mathcal{N}(\theta; \hat{\theta}_t, \Sigma_t)$, where $\hat{\theta}_t$ and Σ_t correspond to the sample mean and covariance of the SMC particles. However, this approach does not properly capture multimodality and asymmetry in the SMC posterior. In our case, instead, we use a kernel density estimator (Wand and Jones, 1994):

$$\hat{p}_t(\theta) := \sum_{i=1}^n w_t^i k_q(\theta, \theta_t^i), \tag{13}$$

with the kernel $k_q : \Theta \times \Theta \rightarrow \mathbb{R}$ chosen so that the constraint $\int_{\Theta} q(\theta) d\theta = 1$ is satisfied. In particular, applying normalized weights, we use a Gaussian kernel for our problems $k_q(\theta, \theta_t^i) := \mathcal{N}(\theta; \theta_t^i, \sigma_q^2 \mathbf{1})$, where $\sigma_q > 0$ corresponds to the kernel’s bandwidth. Machine learning methods for density estimation, such as normalizing flows (Dinh et al., 2017; Kobzyev et al., 2020) and kernel embeddings (Schuster et al., 2020), are also available in the literature. However, we do not apply more complex methods to our framework to preserve the simplicity of the approach, also noticing that simple kernel density estimation provided reasonable results in preliminary experiments.

5.2. Importance reweighting

Having a probability density function \hat{p}_t and a corresponding distribution which we can directly sample from, we draw i.i.d. samples $\{\theta_{\hat{p}_t}^i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} \hat{p}_t$. The importance sampling estimator of $\mathbb{E}[u(\theta)|\mathcal{D}_t]$ using \hat{p}_t as a proposal distribution is given by:

$$\mathbf{x} \in \mathcal{X}, \quad \mathbb{E}[u(\theta)|\mathcal{D}_t] = \int u(\theta) p(\theta|\mathcal{D}_t) d\theta = \int u(\theta) \frac{p(\theta|\mathcal{D}_t)}{\hat{p}_t(\theta)} \hat{p}_t(\theta) d\theta \approx \sum_{i=1}^n u(\theta_{\hat{p}_t}^i) \frac{p(\theta_{\hat{p}_t}^i|\mathcal{D}_t)}{\hat{p}_t(\theta_{\hat{p}_t}^i)}. \tag{14}$$

We, however, do not have access to $p(\mathcal{D}_t)$ to compute $p(\theta|\mathcal{D}_t)$. Therefore, we do another approximation using the same proposal samples:

$$p(\mathcal{D}_t) = \int_{\Theta} p(\mathcal{D}_t|\theta) p(\theta) d\theta = \int_{\Theta} p(\mathcal{D}_t|\theta) \frac{p(\theta)}{\hat{p}_t(\theta)} \hat{p}_t(\theta) d\theta \approx \sum_{i=1}^n p(\mathcal{D}_t|\theta_{\hat{p}_t}^i) \frac{p(\theta_{\hat{p}_t}^i)}{\hat{p}_t(\theta_{\hat{p}_t}^i)}. \tag{15}$$

Setting $\alpha_t^i := p(\mathcal{D}_t|\theta_{\hat{p}_t}^i) \frac{p(\theta_{\hat{p}_t}^i)}{\hat{p}_t(\theta_{\hat{p}_t}^i)}$, we then have:

$$\mathbb{E}[u(\theta)|\mathcal{D}_t] \approx \frac{1}{\eta_t} \sum_{i=1}^n \alpha_t^i u(\theta_{\hat{p}_t}^i), \quad \mathbf{x} \in \mathcal{X}, \tag{16}$$

where $\eta_t := \sum_{i=1}^n \alpha_t^i \approx p(\mathcal{D}_t)$. This approach has been recently applied to estimate intractable marginal likelihoods of probabilistic models (Tran et al., 2021).

5.3. Debiasing via the jackknife

There are different sources of bias in the SMC estimate and in the further steps described above. The jackknife method provides sample-based estimates for statistics of a random variable. The method is based on averaging leave-one-out estimates (Efron, 1992). For the case of bias in estimation, consider a set of samples $\{\theta_i\}_{i=1}^n$ and a function $u : \Theta \rightarrow \mathbb{R}$. Let \hat{u} denote the importance-sampling estimate of the expected value of u according to Equation (16). If we remove a sample $i \in \{1, \dots, n\}$, we can reestimate a proposal density and the importance-weighted expected value of u as \hat{u}_i . The jackknife estimate for the bias in the approximation of $\mathbb{E}[u]$ is then given by:

$$\hat{u}_{\text{bias}} := (n - 1) \left(\frac{1}{n} \sum_{i=1}^n \hat{u}_i - \hat{u} \right). \tag{17}$$

Having an estimate for the bias, we can subtract it from the approximation in Equation (16) to compensate for its bias. We compare the effect of this combined with the previous approaches in preliminary experiments presented next.

5.4. Preliminary experiments on bias correction

We ran preliminary experiments to test the effects of debiasing and decorrelation methods for the estimation of the posterior CDF with SMC. Experiments were run with randomized settings, having parameters for the true model sampled from the prior used by SMC. The test model consists of an exponential-gamma model over $\Theta \subset \mathbb{R}$ with a gamma prior $\theta \sim \Gamma(\alpha, \beta)$ and an exponential likelihood $p(o|\theta) = \theta \exp(-\theta o)$. Given T observations, to compute the approximation error, the true posterior is available in closed-form as:

$$p(\theta|o_1, \dots, o_T) = \Gamma(\theta; \alpha + T, \beta + T\bar{o}_T), \tag{18}$$

where $\bar{o}_T := (1/T) \sum_{i=1}^T o_T$. For this experiment, we set $\alpha = \beta = 1$ as the prior parameters.

Within each trial, we sampled a parameter θ from the prior to serve as the true model, and generated observations from it. We ran Algorithm 2 with an MCMC kernel equipped with a random walk proposal $q(\theta'|\theta) := \mathcal{N}(\theta'; \theta, \sigma^2)$ with $\sigma = 0.1$. We applied Gaussian kernel density estimation to estimate the parameters' posterior density, setting the median distance between particles as the kernel bandwidth.

CDF approximation results are presented in Figures 1 and 2 for the posteriors after $T = 2$ and $T = 5$ observations, respectively. As seen the effects of bias correction methods becomes more evident for SMC

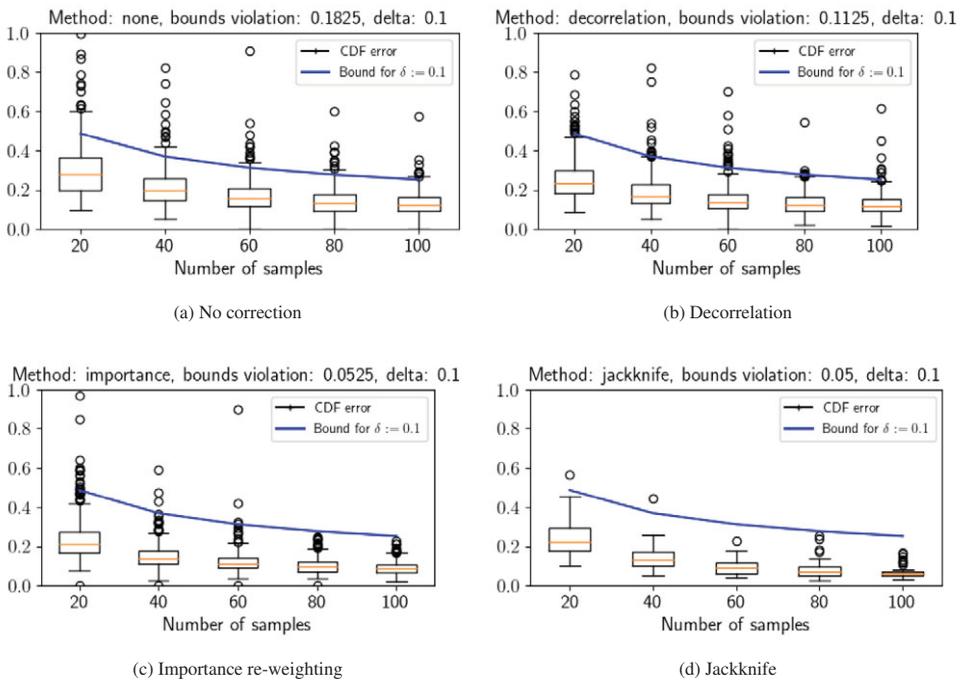


Figure 1. Posterior CDF approximation errors for the exponential-gamma model using $T = 2$ observations. For each sample size, which corresponds to the number of SMC particles, SMC runs were repeated 400 times for each method, except for the jackknife, which was rerun 40 times due to a longer run time. The theoretical upper confidence bound on the CDF approximation error $c_n(\delta)$ (Theorem 1) is shown as the plotted blue line. The frequency of violation of the theoretical bounds for i.i.d. empirical CDF errors is also presented on the top of each plot, alongside the target ($\delta = 0.1$).

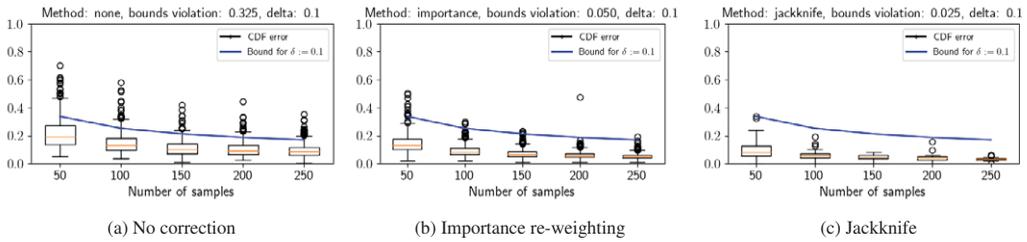


Figure 2. Posterior CDF approximation errors for the exponential-gamma model using $T = 5$ observations. For each sample size, which corresponds to the number of SMC particles, SMC runs were repeated 400 times for each method, except for the jackknife, which was rerun 40 times due to a longer run time. The theoretical upper confidence bound on the CDF approximation error $c_n(\delta)$ (Theorem 1) is shown as the plotted blue line. The frequency of violation of the theoretical bounds for i.i.d. empirical CDF errors is also presented on the top of each plot, alongside the target ($\delta = 0.1$).

posteriors after a larger number of observations is collected. In particular, we highlight the difference between importance reweighting and jackknife results, which is more noticeable for $T = 5$. The main drawback of the jackknife approach, however, is its increased computational cost due to the repeated leave-one-out estimates. In our main experiments, therefore, we present the effects of the importance reweighting correction compared to the standard SMC predictions.

5.5. Limitations

As a few remarks on the limitations of the proposed framework, we highlight that the use of more complex forward models and bias correction methods for the predictions comes at an increased cost in terms of runtime, when compared to simpler models, such as GPs. In practice, one should consider applications where the cost of observations is much larger than the cost of evaluating model predictions. For example, in mineral exploration, collecting observations may involve expensive drilling operations which take a much higher cost than running simulations for a few minutes, or even days, to obtain an SMC estimate. This runtime cost, however, becomes amortized as the number of observations grows, since the computational complexity for GP updates is still $O(t^3)$, which grows cubically with the number of observations, when compared to an $O(nt)$ for SMC. Even when considering importance reweighting or optionally combining it with the jackknife method for bias correction, the computational complexity for SMC updates is only added by $O(nt)$ or $O(n^2t)$, respectively, which are both linear with respect to the number of observations. Lastly, the compromise between accuracy and speed can be controlled by adjusting the number of particles n used by the algorithm. As each particle corresponds to an independent simulation by the forward model, these simulations can also be executed simultaneously in parallel, further reducing the algorithm’s runtime. The following section presents experimental results on practical applications involving water resources monitoring, which in practice present a suitable use case for our framework, given the cost of observations and the availability of informative physics simulation models.

6. Experiments

In this section, we present experimental results comparing SMC-UCB against its GP-based counterpart GP-UCB (Srinivas et al., 2010) and the GP expected improvement (GP-EI) algorithm (Jones et al., 1998), which does not depend on a confidence parameter. We assess the effects of the SMC approximation and its optimization performance in different problem settings. As a performance metric, we use the algorithm’s regret $r_t := \max_{\mathbf{x} \in \mathcal{S}} f(\mathbf{x}) - f(\mathbf{x}_t)$. The averaged regret provides an upper bound on the minimum regret of the algorithm up to iteration T , that is, $\min_{t \leq T} r_t \leq \frac{1}{T} \sum_{t=1}^T r_t$. A vanishing average regret (as $T \rightarrow \infty$) indicates that the choices of the algorithm get arbitrarily close to the optimal solution \mathbf{x}^* in Equation (1).

6.1. Linear Gaussian case

Our first experiment is set with a linear Gaussian model, where the true posterior over the parameters is available in closed form and corresponds to a degenerate (i.e., finite-dimensional) GP. As a synthetic data example, we sample true parameters from their prior distribution $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and set $f(\mathbf{x}) := \theta^T \phi(\mathbf{x})$, where $\phi: \mathcal{X} \rightarrow \mathbb{R}^m$ is a kernel-based feature map $\phi(\mathbf{x}) = [k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_m)]^T$ with $\{\mathbf{x}_i\}_{i=1}^m \sim \mathcal{U}(\mathcal{S})$, and $\mathcal{S} := [0, 1]^d, d := 2$. We set k as a squared-exponential kernel (Rasmussen and Williams, 2006) with length-scale $\ell := 0.2$. The parameters prior $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the likelihood $o|f(\mathbf{x}) \sim \mathcal{N}(f(\mathbf{x}), \sigma_\varepsilon^2)$ are both Gaussian, with $\sigma_\varepsilon = 0.1$. In this case, the posterior given t observations $\mathcal{D}_t = \{\mathbf{x}_t, o_t\}_{i=1}^t$ is available in closed form as:

$$\theta|\mathcal{D}_t \sim \mathcal{N}(\mathbf{A}_t^{-1} \Phi \mathbf{o}_t, \sigma_\varepsilon^2 \mathbf{A}_t^{-1}), \tag{19}$$

where $\mathbf{A}_t := \Phi_t \Phi_t^T + \sigma_\varepsilon^2 \mathbf{I}$, $\Phi_t := [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_t)]$ and $\mathbf{o}_t := [o_1, \dots, o_t]^T$.

As a move proposal for SMC, we use a Gaussian random walk (Andrieu et al., 2003) within a Metropolis-Hastings kernel. We use the same standard normal prior for the parameters and the linear Gaussian likelihood based on the same feature map. For all experiments, we set $\delta := 0.3$ for SMC-UCB.

We first compared SMC-UCB against GP-UCB in the optimization setting with different settings for the number of particles n . GP-UCB (Srinivas et al., 2010) followed Durand et al. (2018, Theorem 2) for the setting of the UCB parameter, and the GP covariance function was set as the same kernel in the feature map. We also set $\delta := 0.3$ for GP-UCB. The resulting regret across iterations is shown in Figure 3a. We observe that SMC-UCB is able to attain faster convergence rates than GP-UCB for $n \geq 300$ particles. A factor driving this improvement is that the quantile approximation error quickly decreases with n , as seen in Figure 3b.

Another set of tests compared SMC-UCB using 400 particles and the importance reweighting method in Section 5.5 against GP-UCB and GP-EI. The optimization performance results are presented in Figure 3c and show that SMC-UCB is able to maintain a clear advantage against both GP-based approaches and improve over its no-importance-correction baseline (Figure 3a). In particular, we notice that GP-EI is outperformed by both UCB methods. Lastly, a main concern regarding SMC methods is the decrease in posterior approximation performance as the dimensionality of the parameter space increases. Figure 3d, however, shows that SMC-UCB is able to maintain reasonable optimization performance up to dimensionality $m = 60$ when compared to baselines. Although this experiment involved a linear model, we conjecture that similar effects in terms of optimization performance should be observed in models which are at least approximately linear, such as when $h(\mathbf{x}, \theta)$ is Lipschitz continuous w.r.t. θ . Yet, a formal theoretical performance analysis is left for future work.

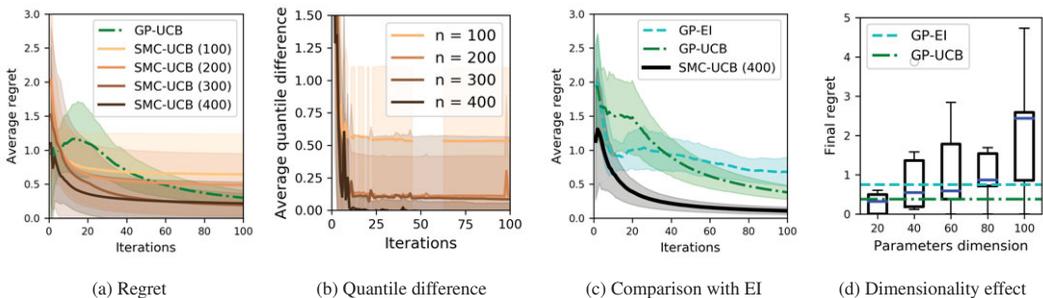


Figure 3. Linear Gaussian case: (a) mean regret of SMC-UCB for different n compared to the GP-UCB baseline with parameter dimension $m := 10$; (b) approximation error between the SMC quantile $\hat{q}_t(\mathbf{x}_t, \delta_t)$ and the true $q_t(\mathbf{x}_t, \delta_t)$ at SMC-UCB’s selected query point \mathbf{x}_t for different n settings (absent values correspond to cases where $\hat{q}_t(\mathbf{x}_t, \delta_t) = \infty$); (c) comparison with the non-UCB, GP-based expected improvement algorithm; and (d) effect of parameter dimension m on optimization performance when compared to the median performance of the GP optimization baselines. All results were averaged over 10 runs. The shaded areas correspond to ± 1 standard deviation.

6.2. Water leak detection

As an example of a realistic application scenario, we present an approach to localize leaks in water pipes via SMC and BO. A long rectilinear underground pipe presents a 5-day water leak at an unknown location within a porous soil medium. We have access to a mobile gravimetric sensor on the surface with a very low noise level $\sigma_\varepsilon := 5 \times 10^{-8} \text{ m s}^{-2}$ for gravitational acceleration. The wet soil mass presents higher density than the dry soil in its surroundings, leading to a localized maximum on the gravitational readings. We assume that prior information on gravitational measurements in the area allows us to cancel out the gravity from the surrounding buildings and other underground structures. We performed computational fluid dynamics (CFD) simulations³ to provide “ground-truth” gravity readings $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), g_3(\mathbf{x})]^T$ at the surface, and took the vertical gravity component $g_3(\mathbf{x})$ as our objective function for BO. Figure 5a shows the relative gravity measurements. The true pipe is 2 m in diameter and buried 3 m underground, quantities which are not revealed to BO.

6.2.1. Pipe leak simulation

To provide an objective function consisting of realistic gravity measurements, we simulated a 3D model in the COMSOL Multiphysics software environment⁴ using some basic assumptions related to soil properties. This tool is a finite-element analysis and simulation software for solving and post-processing computational models of physical systems. For our simulations, in particular, we used the subsurface flow module,⁵ which allows simulating fluid flows in porous media. In detail, the soil porosity is set to 0.339, residual saturation to 1×10^{-3} , soil storage coefficient to $3.4466 \times 10^{-5} \text{ Pa}^{-1}$, saturated hydraulic conductivity to $5.2546 \times 10^{-6} \text{ ms}^{-1}$ and the leak from the pipe to $10 \text{ kg m}^{-2} \text{ s}^{-1}$. For the experiments we use the gravity measured on the surface after 5 days of leakage. Figure 4 shows the spatial arrangement of the simulation. To provide observations to the optimization algorithms, gravity data is corrupted with simulated noise following a quantum gravimeter noise characteristics (Hardman et al., 2016). The resulting data is shown in Figure 5a.

6.2.2. Gravity forward model

To derive an informative model from domain knowledge for SMC, we model the pipe as an infinite cylinder and the leak as a spherical mass, which have closed-form expressions for their gravitational field (Young and Freedman, 2015). The gravity of an infinite cylinder of density ρ_P and radius r_P passing by a location $\mathbf{x}_P \in \mathbb{R}^3$ can be derived as:

$$\mathbf{g}_P(\mathbf{x}) = -\frac{2G\pi\rho_P r_P^2 \mathbf{u}_P \times (\mathbf{x} - \mathbf{x}_P) \times \mathbf{u}_P}{\|(\mathbf{x} - \mathbf{x}_P) \times \mathbf{u}_P\|_2^2}, \quad (20)$$

where $G := 6.674 \times 10^{-11} \text{ m}^3 \text{ kg}^{-1} \text{ s}^{-2}$ represents Newton’s gravitation constant. For the leak, the gravity of a uniform sphere of density ρ_S and radius r_S centered at location $\mathbf{x}_S \in \mathbb{R}^3$ is given by:

$$\mathbf{g}_S(\mathbf{x}) = -\frac{4G\pi\rho_S r_S^3}{3\|\mathbf{x} - \mathbf{x}_S\|_2^2}. \quad (21)$$

6.2.3. Likelihood

As an observation model, we use Equations (20) and (21) to compose a Gaussian likelihood model based on the gravity of the pipe \mathbf{g}_P and of the wet-soil sphere \mathbf{g}_S . As gravity is an integral over distributions of mass, Gaussian noise is a well-suited assumption due to the central limit theorem (Bauer, 1981). So we set $p(o_t | \boldsymbol{\theta}, \mathbf{x}) = \mathcal{N}(o_t; g_{P,3}(\mathbf{x}) + g_{S,3}(\mathbf{x}), \sigma_\varepsilon^2)$ for the gravity model parameters $\boldsymbol{\theta} := [\mathbf{x}_P^T, r_P, \rho_P, \mathbf{u}_P^T, \mathbf{x}_S^T, r_S, \rho_S]^T$

³ Simulations based on COMSOL Multiphysics software environment: <https://www.comsol.com>

⁴ Available at: <https://www.comsol.com>

⁵ Details at: <https://www.comsol.com/subsurface-flow-module>

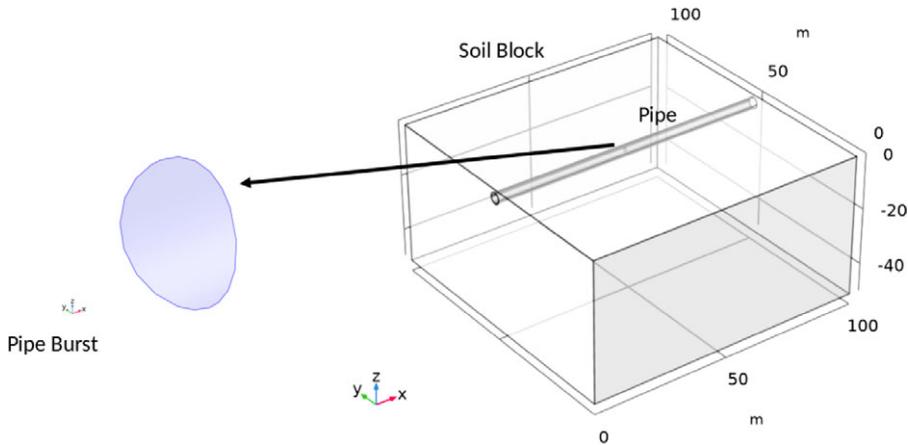


Figure 4. Pipe simulation diagram: A water pipe of 2 in diameter is buried 3 underground in a large block of soil (100 × 100 × 50m).

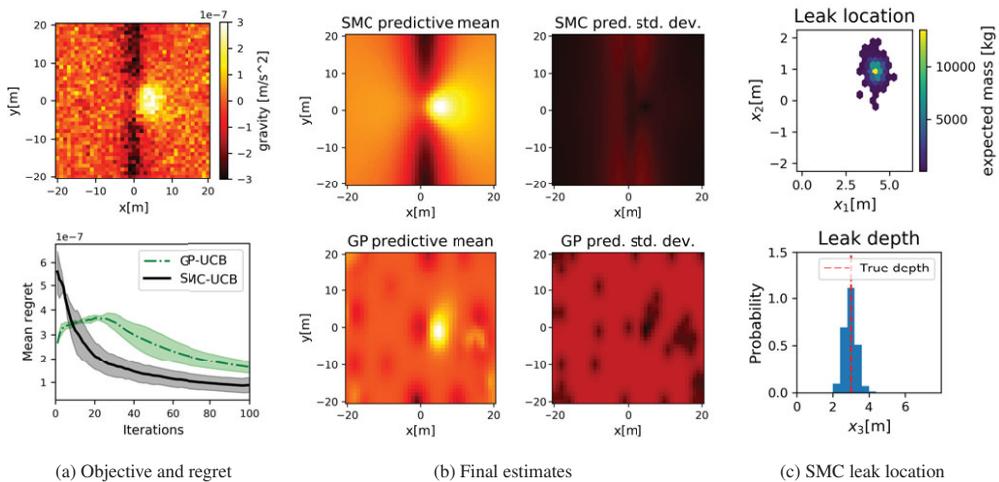


Figure 5. Performance results for water leak detection experiment: (a) The gravity objective function generated by CFD simulations and the mean regret curves for each algorithm. The shaded areas in the plot correspond ± 1 standard deviation results from results which were averaged over 10 trials. (b) The gravity estimates according to the final SMC and GP posteriors after 100 iterations. (c) SMC estimates for the parameters concerning the location of the leak. The upper plot in (c) is colored according to an estimate for the mass of leaked water.

based on the vertical-axis gravity measurements with $\sigma_\epsilon := 5 \times 10^{-8} \text{ m s}^{-2}$, the noise level of the gravity sensor we consider (Hardman et al., 2016).

6.2.4. Priors on the pipe

We assume uncertainty on the pipe’s location $\mathbf{x}_P := [x_{P,1}, x_{P,2}, x_{P,3}]^T$, though not on its orientation \mathbf{u}_P .⁶ We place a Gaussian prior on the pipe’s horizontal coordinate with $x_{P,1} \sim \mathcal{N}(0, 1)$ and a gamma prior $\Gamma(\alpha_z, \beta_z)$

⁶ The pipe can, for example, be assumed to be aligned with a street of known location.

with $\alpha = 2$ and $\beta = 1$ on the pipe's depth, that is, $-x_{P,3}$. For the radius, we set $r_P \sim \Gamma(\alpha_r, \beta_r)$ with $\alpha_r = 2.25$ and $\beta_r = 0.75$, yielding a 95% confidence interval of $r_P \in [0.4, 8]$ m. The pipe's density is set as tightly around the water density $\rho_P \sim \mathcal{N}(\hat{\rho}_P, \sigma_{\rho_P}^2)$ with $\hat{\rho}_P := 1000 \text{ kg m}^{-3}$ and $\sigma_{\rho_P} := 10 \text{ kg m}^{-3}$, which allows for small variations in the pipe's mass distribution due to the outflow of water with the leak.

6.2.5. Priors on the sphere

For the wet soil sphere, we place conditional priors $x_{S,1}|x_{P,1}, r_P \sim \mathcal{N}(x_{P,1}, r_P^2/4)$ and $(x_{P,3} - x_{S,3})/x_{P,3} \sim \text{Beta}(2, 5)$, indicating that the leak is most likely concentrated somewhere around the pipe's border. However, the leak's $x_{S,2}$ coordinate, which goes along the pipe, is only known to be within a bounded segment of the pipe, $x_{S,2} \sim \mathcal{U}(-20, 20)$. As a leak is a rare phenomenon, we place priors with high probability around zero for the leak's radius and density. In particular, we set a 1-degree-of-freedom chi-square distribution for $r_S \sim \chi^2(1)$, and the density as a gamma random variable $\rho_S \sim \Gamma(\alpha_\rho, \beta_\rho)$ with $\alpha_\rho = 2$ and $\beta_\rho = 1/250$, so that $\rho_S \in [100, 1000]$ kg with approximately 90% probability.

6.2.6. SMC settings

A HMC kernel within SMC was configured with 100 leapfrog steps and per-parameter tuned step sizes, according to their prior distribution.⁷ The number of particles is set to $n := 1000$.

6.2.7. Optimization settings

GP-UCB and SMC-UCB were run with similar settings. GP hyperparameters were learnt offline via maximum a posteriori estimation. The GP covariance function was a Matérn function with smoothness parameter $\nu := 2.5$ (see Rasmussen and Williams, 2006, Ch. 4) SMC-UCB was configured with $\delta := 0.1$. As we do not have enough information about the objective function to derive theory-recommended UCB parameters, GP-UCB was set with an equivalent UCB parameter based on the inverse CDF of a normal distribution. For both methods, the acquisition function optimization is performed over a discretized fine grid covering the search space.

6.2.8. Results

Figure 5 presents results for the leak detection experiment. As the plot in Figure 5a shows, the proposed SMC-UCB algorithm is able to overcome GP-UCB in this setting with a clearly lower regret from the beginning. The SMC method also presents final predictions better approximating the provided CFD data with low uncertainty on the sphere position parameters, as seen in Figure 5b. Due to exploration–exploitation trade-off of BO algorithms (Brochu et al., 2010; Shahriari et al., 2016), it is natural that the uncertainty regarding the pipe's parameters is slightly higher, as its gravity does not directly contribute to the maximum of the gravity objective function. Yet we can see that SMC's estimate on the pipe's gravity signature is still evident when compared to the GP's estimated gravity predictions (see Figure 5c). In addition, SMC is able to directly estimate the leak's location online based on its empirical posterior parameters distribution. With GPs the same would require running a second inference scheme using all the observations.

6.3. Contaminant source localization

As a final experiment to demonstrate the potential of SMC for BO, we present the problem of localizing a source of contamination within an aquifer with strong property contrasts (Piro et al., 2019). We transform the source localization problem into that of finding the maximum contaminant concentration over a set of 25 possible well locations. We have access to hydraulic simulations⁸ (Cornaton, 2007) made available by Piro et al. (2019) predicting the contaminant concentration at $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^L$, where $L := 25$ possible well

⁷ Parameters with wider priors are set with larger step sizes.

⁸ Data available at: <https://github.com/gpirot/BGICLP>

locations over a window of 300 days with candidate contaminant source locations $\theta := \mathbf{x}_C$ spread over a 51-by-51, 3-meter cell grid, totalling 2,601 possible locations.

6.3.1. Data

The contaminant concentration data is composed of two datasets, one with the simulated concentration values for all 25 wells and all 301 measurement times over the 51-by-51 source locations grid, while the other dataset consists of reference “real” concentration values over the 25 wells and the 301 time steps. The “real” concentration values are in fact simulated concentration values at locations not revealed to the BO algorithms. To simulate different scenarios, each experiment trial was run on a random day, sampled uniformly among the 300 days. To provide observations for each algorithm, we also add Gaussian noise with $\sigma_\epsilon := 1 \times 10^{-3}$, corresponding to roughly a 10% measurement error level.

6.3.2. SMC model

We use the hydraulic simulations to form a model for SMC-UCB using $n := 400$ particles. A separate dataset of simulated concentrations is used as the objective function. SMC is configured with a Metropolis-Hastings move kernel using a proposal over the discrete source locations space. At each move, we compute the current posterior probabilities at each point in the location space and form an empirical distribution with it, which is our proposal distribution.

6.3.3. Optimization settings

For this experiment, we set SMC-UCB with $\delta := 0.3$. GP-UCB is set with a fixed value for the UCB parameter $\beta_t := 3$, which was the best performing for this experiment after tuning with preliminary runs. To provide a tuned GP model, we compute an empirical covariance matrix for the well measurements data from a random subset of the 2,601 possible source locations. With this empirical covariance matrix, we form a zero-mean GP over the discrete search space. This procedure is similar to common practice in the GP-UCB literature when real datasets are provided (Srinivas et al., 2010; Chowdhury and Gopalan, 2017).

6.3.4. Results

Figure 6 presents our experimental results for the contaminant source localization problem. As Figure 6a shows, SMC-UCB is able to achieve a better performance than GP-UCB by using the informative prior coming from hydraulic simulations. In addition, SMC final parameter estimates (Figure 6b) present high probability mass around the true source location.

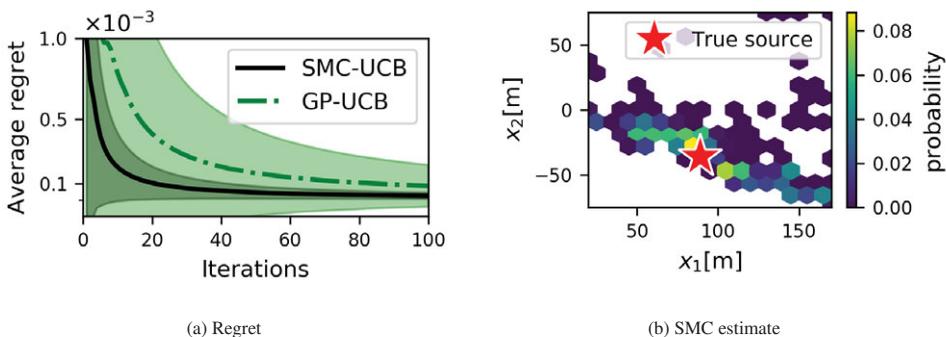


Figure 6. Contaminant source localization problem: (a) The optimization performance of each algorithm in terms of regret. Results were averaged over 10 runs. (b) A final SMC estimate for the source location, while the true location is marked as a red star.

7. Conclusion

This article presented SMC as an alternative to GP based approaches for Bayesian optimization when domain knowledge is available in the form of informative computational models. We presented a quantile-based acquisition function for BO which is adjusted for the effects of the approximation bias in the SMC estimates and allows for inference based on non-Gaussian posterior distributions. The resulting SMC algorithm is shown to outperform the corresponding GP-based baseline in different problem scenarios.

The work in this article can be seen as a starting point toward a wider adoption of SMC algorithms in BO applications. Further studies with theoretical assessments of the SMC approximation should strengthen the approach and allow for the derivation of new algorithms which can make an impact on the practical use of BO in data-driven science and engineering applications. Another interesting direction of future research is the incorporation of likelihood-free SMC (Sisson et al., 2007) methods into the BO loop in cases where forward models do not allow for a closed-form expression of the observation model.

Acknowledgment. Rafael Oliveira was supported by the Medical Research Future Fund Applied Artificial Intelligence in Health Care grant (MRFAI000097).

Data Availability Statement. The data for the contaminant source localization experiment is available at <https://github.com/gpirot/BGICLP>.

Author Contributions. Conceptualization: R.O., R.S., R.K., S.C., J.C.; Data curation: K.H., N.T.; Formal analysis: R.O.; Funding acquisition: R.S., R.K., S.C., J.C., C.L.; Investigation: R.O., K.H., N.T., C.L.; Methodology: R.O., R.K.; Software: R.O.; Supervision: R.S., R.K., S.C., C.L.; Writing—original draft: R.O. All authors approved the final submitted draft.

Funding Statement. Parts of this research were conducted by the Australian Research Council Industrial Transformation Training Centre in Data Analytics for Resources and the Environment (DARE), through project number IC190100031.

Competing Interests. The authors declare no competing interests exist.

References

- Andrieu C, De Freitas N, Doucet A and Jordan MI (2003) An introduction to MCMC for machine learning. *Machine Learning* 50 (1–2), 5–43.
- Arnold DV and Hansen N (2010) Active covariance matrix adaptation for the (1+1)-CMA-ES. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation - GECCO '10, Portland, OR*. New York: ACM.
- Bauer H (1981) *Probability Theory and Elements of Measure Theory. Probability and Mathematical Statistics*. New York: Academic Press.
- Benassi R, Bect J and Vazquez E (2012) Bayesian optimization using sequential Monte Carlo. In Hamadi Y and Schoenauer M (eds), *Learning and Intelligent Optimization*. Berlin: Springer, pp. 339–342.
- Berkenkamp F, Schoellig AP and Krause A (2019) No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research (JMLR)* 20, 1–24.
- Beskos A, Crisan DO, Jasra A and Whiteley N (2014) Error bounds and normalising constants for sequential Monte Carlo samplers in high dimensions. *Advances in Applied Probability* 46(1), 279–306.
- Bijl H, Schön TB, van Wingerden J-W and Verhaegen M (2016) *A sequential Monte Carlo approach to Thompson sampling for Bayesian optimization*. arXiv.
- Bishop CM (2006) Chapter 10: Approximate inference. In *Pattern Recognition and Machine Learning*. Berlin: Springer, pp. 461–522.
- Brochu E, Cora VM and de Freitas N (2010) *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. Technical Report. Vancouver, BC: University of British Columbia.
- Brown G, Ridley K, Rodgers A and de Villiers G (2016) Bayesian signal processing techniques for the detection of highly localised gravity anomalies using quantum interferometry technology. In *Emerging Imaging and Sensing Technologies, 9992*. Bellingham, WA: SPIE.
- Bull AD (2011) Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research (JMLR)* 12, 2879–2904.
- Chopin N (2002) A sequential particle filter method for static models. *Biometrika* 89(3), 539–551.

- Chowdhury SR and Gopalan A** (2017) On Kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia: Proceedings of Machine Learning Research.
- Cornaton F** (2007) *Ground Water: A 3-d Ground Water and Surface Water Flow, Mass Transport and Heat Transfer Finite Element Simulator*. Technical Report. Neuchâtel: University of Neuchâtel.
- Crisan D and Doucet A** (2002) A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on Signal Processing* 50(3), 736–746.
- Dalibard V, Schaarschmidt M and Yoneki E** (2017) Boat: Building auto-tuners with structured bayesian optimization. In *Proceedings of the 26th International Conference on World Wide Web, WWW '17*. Geneva: Republic and Canton of Geneva, CHE, International World Wide Web Conferences Steering Committee, pp. 479–488.
- Del Moral P, Doucet A and Jasra A** (2012) On adaptive resampling strategies for sequential Monte Carlo methods. *Bernoulli* 18(1), 252–278.
- Dinh L, Sohl-Dickstein J and Bengio S** (2017) Density estimation using real NVP. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France.
- Doucet A, de Freitas N and Gordon N** (eds) (2001) *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics, 1st Edn. New York: Springer.
- Doucet A, Godsill S and Andrieu C** (2000) On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing* 10, 197–208.
- Durand A, Maillard O-A and Pineau J** (2018) Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research* 19, 1–48.
- Efron B** (1992) Bootstrap methods: Another look at the Jackknife. In Kotz S and Johnson NL (eds), *Breakthroughs in Statistics: Methodology and Distribution*. New York: Springer, pp. 569–593.
- Elvira V, Martino L and Robert CP** (2018) Rethinking the Effective Sample Size. *arXiv e-prints*, arXiv:1809.04129.
- Hardman KS, Everitt PJ, McDonald GD, Manju P, Wigley PB, Sooriyabandara MA, Kuhn CCN, Debs JE, Close JD and Robins NP** (2016) Simultaneous precision gravimetry and magnetic gradiometry with a Bose-Einstein condensate: A high precision, quantum sensor. *Physical Review Letters* 117, 138501.
- Hauge VL and Kolbjørnsen O** (2015) Bayesian inversion of gravimetric data and assessment of CO₂ dissolution in the Utsira formation. *Interpretation* 3(2), SP1–SP10.
- Hennig P and Schuler CJ** (2012) Entropy search for information-efficient global optimization. *Journal of Machine Learning Research (JMLR)* 13, 1809–1837.
- Hernández-Lobato JM, Hoffman MW and Ghahramani Z** (2014) Predictive entropy search for efficient global optimization of black-box functions. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS 2014)*, Montréal, Canada: Curran Associates.
- Huggins JH and Roy DM** (2019) Sequential Monte Carlo as approximate sampling: Bounds, adaptive resampling via ∞ -ESS, and an application to particle Gibbs. *Bernoulli* 25(1), 584–622.
- Hutter F, Hoos HH and Leyton-Brown K** (2011) Sequential model-based optimization for general algorithm configuration. In *International Conference on Learning and Intelligent Optimization*. Cham: Springer, pp. 507–523.
- Jones DR, Schonlau M and Welch WJ** (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13(4), 455–492.
- Kaufmann E, Cappé O and Garivier A** (2012) On Bayesian upper confidence bounds for bandit problems. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Vol. 22. La Palma. JMLR: W&CP, pp. 592–600.
- Kawale J, Bui H, Kveton B, Thanh LT and Chawla S** (2015) Efficient Thompson sampling for online matrix-factorization recommendation. In *Advances in Neural Information Processing Systems, Montréal, Canada*. Curran Associates, Curran Associates, Inc. pp. 1297–1305.
- Khan ME, Immer A, Abedi E and Korzepa M** (2019) Approximate inference turns deep networks into Gaussian processes. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada: Curran Associates.
- Kobyzev I, Prince S and Brubaker M** (2020) Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3964–3979.
- Kuck H, de Freitas N and Doucet A** (2006) SMC samplers for Bayesian optimal nonlinear design. In *2006 IEEE Nonlinear Statistical Signal Processing Workshop, Cambridge, UK*: IEEE.
- Martino L, Elvira V and Louzada F** (2017) Effective sample size for importance sampling based on discrepancy measures. *Signal Processing* 131, 386–401.
- Mashford J, De Silva D, Burn S and Marney D** (2012) Leak detection in simulated water pipe networks using SVM. *Applied Artificial Intelligence* 26(5), 429–444.
- Massart P** (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability* 18(3), 1269–1283.
- Naeseth CA, Lindsten F and Schön TB** (2019) Elements of sequential Monte Carlo. *Foundations and Trends in Machine Learning* 12(3), 187–306.
- Neal RM** (2011) MCMC using Hamiltonian dynamics. In Brooks S, Gelman A, Jones G and Meng X-L (eds), *Handbook of Markov Chain Monte Carlo*. New York: Chapman & Hall, pp. 113–162.
- Pirot G, Krityakierne T, Ginsbourger D and Renard P** (2019) Contaminant source localization via Bayesian global optimization. *Hydrology and Earth System Sciences* 23, 351–369.

- Ranganath R, Gerrish S and Blei DM** (2014) Black box variational inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Reykjavik, Iceland: Proceedings of Machine Learning Research.
- Rasmussen CE and Williams CKI** (2006) *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.
- Rios LM and Sahinidis NV** (2013) Derivative-free optimization: A review of algorithms and comparison of software implementations. *Journal of Global Optimization* 56(3), 1247–1293.
- Rossi L, Reguzzoni M, Sampietro D and Sansò F** (2015) Integrating geological prior information into the inverse gravimetric problem: The Bayesian approach. In *VIII Hotine-Marussi Symposium on Mathematical Geodesy*. Cham: International Association of Geodesy Symposia, pp. 317–324.
- Russo D and Van Roy B** (2016) An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research (JMLR)* 17, 1–30.
- Sadeghioon AM, Metje N, Chapman DN and Anthony CJ** (2014) SmartPipes: Smart wireless sensor networks for leak detection in water pipelines. *Journal of Sensor and Actuator Networks* 3(1), 64–78.
- Schuster I, Mollenhauer M, Klus S and Muandet K** (2020) Kernel conditional density operators. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Italy Vol. 108 Proceedings of Machine Learning Research.
- Shahriari B, Swersky K, Wang Z, Adams RP and De Freitas N** (2016) Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* 104(1), 148–175.
- Shields BJ, Stevens J, Li J, Parasram M, Damani F, Alvarado JI, Janey JM, Adams RP and Doyle AG** (2021) Bayesian reaction optimization as a tool for chemical synthesis. *Nature* 590(7844), 89–96.
- Sisson SA, Fan Y and Tanaka MM** (2007) Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* 104(6), 1760–1765.
- Snoek J, Larochelle H and Adams RP** (2012) Practical bayesian optimization of machine learning algorithms. In Pereira F, Burges JC, Bottou L and Weinberger KQ (eds), *Advances in Neural Information Processing Systems 25*. Red Hook, NY: Curran Associates, pp. 2951–2959.
- Snoek J, Rippel O, Swersky K, Kiros R, Satish N, Sundaram N, Patwary M, Prabhat, and Adams R** (2015) Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning (ICML)*, Lille, France: Proceedings of Machine Learning Research (PMLR).
- Souza JR, Marchant R, Ott L, Wolf DF and Ramos F** (2014) Bayesian Optimisation for active perception and smooth navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China: IEEE.
- Spooner T, Jones AE, Fearnley J, Savani R, Turner J and Baylis M** (2020) Bayesian optimisation of restriction zones for bluetongue control. *Scientific Reports* 10(1), 1–18.
- Springenberg JT, Aaron K, Falkner S and Hutter F** (2016) Bayesian optimization with robust Bayesian neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain: Curran Associates.
- Srinivas N, Krause A, Kakade SM and Seeger M** (2010) Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, Haifa, Israel. Omnipress, pp. 1015–1022.
- Szörényi B, Busa-Fekete R, Weng P and Hüllermeier E** (2015) Qualitative multi-armed bandits: A quantile-based approach. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France: Proceedings of Machine Learning Research.
- Thrun S, Burgard W and Fox D** (2006) *Probabilistic Robotics*. Cambridge, MA: The MIT Press.
- Tran M-N, Scharth M, Gunawan D, Kohn R, Brown SD and Hawkins GE** (2021) Robustly estimating the marginal likelihood for cognitive models via importance sampling. *Behavior Research Methods* 53(3), 1148–1165.
- Urteaga I and Wiggins CH** (2018) Sequential Monte Carlo for dynamic softmax bandits. In *1st Symposium on Advances in Approximate Bayesian Inference (AABI)*, Montréal, Canada.
- Urteaga I and Wiggins CH** (2019) (Sequential) Importance Sampling Bandits. *arXiv e-prints*, arXiv:1808.02933.
- Wand MP and Jones MC** (1994) *Kernel Smoothing*. Boca Raton, FL: CRC Press.
- Wang Z and Jegelka S** (2017) Max-value entropy search for efficient Bayesian optimization. In *34th International Conference on Machine Learning, ICML 2017*, Sydney, Australia, Vol. 7. Proceedings of Machine Learning Research.
- Young H and Freedman R** (2015) *University Physics with Modern Physics*. Hoboken, NJ: Pearson Education.

Cite this article: Oliveira R, Scalzo R, Kohn R, Cripps S, Hardman K, Close J, Taghavi N and Lemckert C (2022). Bayesian optimization with informative parametric models via sequential Monte Carlo. *Data-Centric Engineering*, 3: e5. doi:10.1017/dce.2022.5