Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

CALIBRATING DIGITAL TWINS VIA BAYESIAN OPTIMIZATION WITH A ROOT FINDING STRATEGY

Yongseok Jeon and Sara Shashaani

Fitts Dept. of Industrial and Systems Eng., North Carolina State University, Raleigh, NC, USA

ABSTRACT

Calibrating digital twins is a challenging tasks that various methodologies have been used to address. Bayesian Optimization is a prominent tool for this purpose albeit with computational limitations. We propose root-finding strategies within the Bayesian Optimization framework that is tailored for digital twin calibration task. Employing root-finding scheme introduces new acquisition functions and offer unique advantages over traditional minimization strategy, particularly when using a continuous surrogate model. We demonstrate our findings through a range of motivating examples and calibration tasks.

1 INTRODUCTION

Digital twins (DT) are emerging technologies that enable users to monitor and optimize the real-world (RW) problems in real time. The idea behind the DT is that, by replicating the RW with a virtual representation, users can explore various alternative actions in the system, which facilitates decision-making processes. Examples of use cases of DT are abundant in manufacturing systems (VanDerHorn and Mahadevan 2021), hospital management (Peng et al. 2020) or predicting the equipments' condition in space programs (Liu et al. 2021). Despite the noticeable achievements, ensuring the fidelity of these virtual models in mirroring the RW is a critical challenge that persists. Since the discrepancies between the DT and RW can severely impact the reliability of the subsequent decision-making, proper calibration techniques are essential. In this paper we propose new ideas to more effectively calibrate DT when the real-time collections of RW system performance does not provide more than single observations at each state.

1.1 Problem Statement

Following the notation by Rhodes-Leader and Nelson (2023), assume a DT is used to make decisions at times $t_j,\ j = 1,\dots,J$. At each time, the state of the RW is $\theta_j^c \in \Theta$ (among a known set of possible states Θ), and the corresponding set of feasible actions that users can take is $x \in \mathcal{X}(\theta_j^c)$. In the remainder of this paper, we will omit x from the notations since decision-making with DT is not in the scope of this paper. The state θ_j^c can be unobservable and might be represented either as a vector or a scalar value, depending on the specific context of the problem. Let $h^c(\theta_j^c)$ be the fixed, observed performance measure of the RW system at state θ_j^c. This is a realized quantity of the random variable $H^c(\theta_j^c)$ that represents the system performance following an unknown probability distribution.

We assume to have a DT in our disposal that is well-constructed to spew out random outputs $H(\theta_j)$ where θ_j denotes the (controllable) state of the DT. We assume that once $\theta_j^c \approx \theta_j$, we can expect somewhat similar behavior of the two systems such that $h^c(\theta_j^c) - H(\theta_j) \sim \mathcal{N}(0,\varepsilon)$ for small variance $\varepsilon > 0$. A typical approach for calibration is to match the expected performance in RW, i.e., $\mathbb{E}[H^c(\theta^c)]$ with that of a simulation model $\mathbb{E}[H(\theta)]$, where each expectation is with respect to the corresponding random variable's probability distribution. But in DT applications, often, the calibration has to happen in real-time and we do not have access to more than one observation in RW. Hence, in lieu of $\min_{\theta_j\in\Theta}\ d(\mathbb{E}[H^c(\theta_j^c)], \mathbb{E}[H(\theta_j)])$, where $d : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is the metric of discrepancy between two real-valued quantities (see Section 1.3), we seek to

$$\min_{\theta_j \in \Theta} \; f(\theta_j) := d(h^c(\theta_j^c), \mathbb{E}[H(\theta_j)]). \tag{1}$$

We, therefore, define our calibration task as finding a state for the DT that results in the smallest discrepancy measure when adapted in DT, i.e., θ_j^* that minimizes the objective function in (1). Within this context, we assume that the state is not changed during the calibration task, as in practice, a practitioner may calibrate DT for every time interval $[t_{j-1}, t_j)$ to align with the evolving state of the RW. The expected performance of the DT can be estimated by $h_n(\theta_j)$ following n independent and identically distributed (iid) DT runs. This reduces problem (1) to a sample average approximation (SAA) variant, i.e.,

$$\min_{\theta_j \in \Theta} \; f_n(\theta_j) := d(h^c(\theta_j^c), h_n(\theta_j)). \tag{2}$$

In the remainder of the paper, we drop the subscript j for the j-th state of the system but take into account the possibility of the time interval for calibration to be small, and therefore, a need for speedy calibration.

1.2 Metamodeling for Calibration

Various approaches have been explored for calibration, each with unique advantages. In this paper, we consider the case where the complexity of the DT imposes significant challenges to conduct iterative analysis, or due to the small amount of allowed time interval, practitioner has to make only a few or no simulation at each $\hat{\theta}$. This means we either let $n = 1$ or $n = 0$ for some of the $\hat{\theta}$ values that we explore. In this setting, employing a simulation metamodeling is a desirable option to draw an approximation to the response surface of $f(\theta)$ (Staum 2009). One popular choice for such metamodeling is a Gaussian Process Regression (GPR), often referred to as kriging in Geostatistics area and surrogate modeling in the optimization domain (Yang et al. 2019). We shall use the term metamodeling and surrogate model interchangeably throughout the paper. GPR's capability of addressing complex structures (e.g., non-convex functions) or noisy observations also makes it a popular choice for optimization.

When using metamodels for estimating the response surface of the objective function, there may be time constraints that may restrain us from running the DT multiple times for each θ that is evaluated. We specifically consider extreme limitations where only a single DT observation is being collected at each θ, denoted by $h(\theta)$. Practically speaking, this situation arises when exploring various θs is more preferred than exploiting limited points due to less concerns about the noise and its heterogeneity, In this setting, stochastic kriging (Staum 2009) will not be easy to apply given that we do not have access to sample variance estimates to augment the covariance function with stochastic noise. As a result, we are forced to yet again alter the objective function from (2) to

$$\min_{\theta \in \Theta} \; \tilde{f}(\theta) := d(h^c(\theta^c), h(\theta)), \tag{3}$$

where, under the assumption that we have control over the random numbers used in DT and can apply common random numbers (CRN) throughout the process, the problem reduces to calibrating the sample path objective function instead. Crucially, in this set-up, we have removed the intrinsic variability by fixing the random number stream and exploring a single realization of the random function $F(\theta) := d(h^c(\theta^c), H(\theta))$. This approach has been adopted by Wang (2021) among others and resembles path-wise optimization or SAA with $n = 1$.

With such surrogate models, a widely appreciated solution method for optimization is Bayesian Optimization (BO). However, BO, like many other optimization algorithms that primarily rely on their surrogate model, is not expected to perform as desired when the surrogate model fails to capture the underlying behavior of the function. Often, this failure is due to the inappropriate assumptions and model misspecification. We shall see some of the misusage cases of employing surrogate models throughout the upcoming sections. Despite that, when appropriately deployed, metamodeling would be a reasonable option for the calibration task.

1.3 Minimization or Root-finding?

When framing the calibration task as an optimization problem, defining a proper discrepancy measure is indeed a crucial task. Some popular choice for the discrepancy measure include root mean squared percentage error (RMPSE) (Sha et al. 2020), mean squared error (MSE) (Schultz and Sokolov 2018) (Zhan et al. 2022), or some logarithmic form of squared error (Chakrabarty et al. 2023). When multiple observations of both $H^c(\theta^c)$ and $H(\theta)$ are available that correspond to the same number of iid inputs to RW and DT, then the discrepancy can be posed as an empirical risk minimization, whereby the above L_2 type loss functions are reasonable.

Instead, consider the case where $h^c(\theta^c)$ and $h(\theta)$, as defined above, are summary statistics of the performance measure of interests, such as the average sojourn time of multiple products in queuing system, but without the detailed number of outputs corresponding to iid inputs. We assume that there exists at least one $\tilde{\theta}^*$ that makes $\tilde{f}(\tilde{\theta}^*) = h^c(\theta^c) - h(\tilde{\theta}^*) = 0$. When dealing with such problems, instead of minimizing the discrepancy, one may consider finding the roots of the difference function. Often, the root-finding itself can be reformulated as a minimization problem. For instance, if the function $\tilde{f}(\cdot)$ is a strictly monotonic function and has a single root, its root can be found via $\min_{\theta\in\Theta}\ |\tilde{f}(\theta)|$. However, such transformation has to be carefully handled, as depending on the functional structure or optimization algorithm we use, it may introduce additional challenges to the optimization process. In this paper, we examine the applicability of such transformation, particularly when employing a continuous surrogate model. Given that we use metamodeling for contexts where observations are limited, we find that this transformation to the *observed discrepancy* presents additional challenges and therefore not recommend it. We suggest viewing the calibration task as a root finding; this entails finding the roots of the *predicted* discrepancy function, denoted by $\hat{f}(\theta)$ that is often done via $\min_{\theta\in\Theta}\ |\hat{f}(\theta)|$; we argue that this idea improves calibration results within a fixed computational budget. In this paper we focus an implementation of this idea within the BO context. To this end, we derive new variants of famous acquisition functions and present a tailored approach that leverages continuity of sample paths to reduce the search space from each new solutions will be sampled.

To start, we provide some motivating examples that justify our proposition for the use of root-finding in Section 2. We then briefly review BO covering its basic principles and components in Section 3. In Section 4, we introduce BO combined with a root-finding scheme, followed by our proposed approach to effectively accelerate the search. We implement the proposed approaches in Section 5 and compare with the benchmarks across several calibration tasks. We then conclude in Section 6 leaving the reader with several open questions that require further explorations.

2 MOTIVATING EXAMPLES

Before delving into BO and its components, we demonstrate our motivation by Figure 1. In this figure, the true value θ that yields the best calibration is depicted on the x-axis by the star icon between the two values $[a, b]$. Suppose in evaluating function $f(\theta)$ we see that the DT generates an output larger than that of the RW at $\theta = a$ and smaller at $\theta = b$, hence $f(a)f(b) < 0$. According to the Bolzano's theorem, at least one $f(c) = 0,\ c \in [a, b]$ is assured within the observed interval. In the standard BO setup, employing a minimization strategy, requires transforming the function into a one-sided domain (e.g., using absolute value transformation in the right panel) and we shall refer to this as a *positivization*.

Suppose a GPR is fitted to the positivized function values $|f(\theta)|$. Then the sampling process (of solutions) may escape this interval as the positivized function values do not signal that estimates in this interval may be smaller than the others. If GPR was instead fitted to the original $f(\theta)$, then the sampling process would likely concentrate on this interval to find a root of the function (which is where the discrepancy is 0), as the continuity of GPR inherently assures the existence of a root in this interval.

Another case in point is related to studies in which DT is a simulation model, such as Discrete Event Simulation (DES) (Agalianos et al. 2020). One can generate several, say n, iid DT outputs at the fixed θ

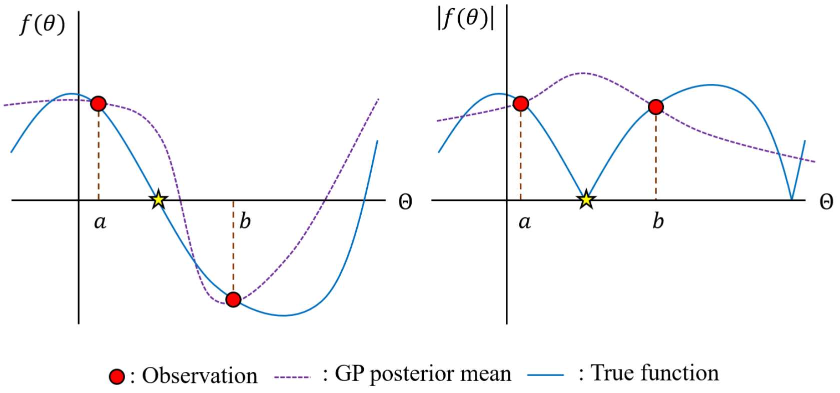

Figure 1: Fitted GPR on original function (left) and positivized case (right)

to estimate $\mathbb{E}[H(\theta)]$ with

$$h_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} H_i(\theta_j) \sim \mathcal{N}(\mu_H(\theta), n^{-1}\sigma_H^2(\theta)) \tag{4}$$

via Central Limit Theorem (CLT), leading to as estimate of the "raw" discrepancy measure

$$f_n(\theta) := h^c(\theta^c) - h_n(\theta) \sim \mathcal{N}(h^c(\theta^c) - \mu_H(\theta), n^{-1}\sigma_H^2(\theta)) \tag{5}$$

that preserves the statistical characteristic of the outputs (normality). On the other hand, positivization via absolute or squared transform would alter the distribution of the discrepancy measure to folded-normal and chi-square distribution, respectively. When employing GPR, there is an underlying belief that the objective function at each θ follows a normal distribution, which will be violated with such positivization.

Based on the insights above, we propose to solve a calibration problem with a root finding perspective in mind rather than searching for the minimum of the positivized discrepancy. We will delve into our proposed modifications to the components in standard BO after briefly reviewing those next.

3 BAYESIAN OPTIMIZATION FOR CALIBRATION

Following the description of objective function in Section 1, suppose we adopt standard BO framework that solves the optimization problem (for calibration) that is of the form $\min\limits_{\theta \in \Theta} |\tilde{f}(\theta)|$, where $\Theta \subset \mathbb{R}^d$ and $\tilde{f} : \Theta \to \mathbb{R}$ is continuous and available through a deterministic black box and therefore with limited structural knowledge. Often one obtains noisy observations of f via a stochastic simulation but in this paper, we limit our scope to the sample-path optimization problem (3) and hence deem the objective function deterministic.

The core of the BO entails two main components. The first component is the surrogate model, typically a GPR, which approximates the structure of the objective function in a probabilistic-manner. Building upon well-defined Bayesian properties, after visiting m solutions $X = [\theta^i \in \mathbb{R}^d,\ i = 1, 2, \cdots, m]$ with corresponding function values $Y = [|\tilde{f}(\theta^1)|, |\tilde{f}(\theta^2)|, \cdots, |\tilde{f}(\theta^m)|]$, the estimates of the function at an unseen θ', $\hat{f}(\theta'|X)$, can be formulated as

$$\begin{aligned} \hat{f}(\theta'|X) &\sim \mathcal{N}(\mu(\theta'), \sigma^2(\theta')), \\ \mu(\theta') &= K(\theta', X; l)[K(X, X; l) + \sigma_m I]^{-1} Y, \\ \sigma^2(\theta') &= K(\theta', \theta'; l) - K(\theta', X; l)\,[K(X, X; l) + \sigma_m I]^{-1}\, K(\theta', X; l). \end{aligned} \tag{6}$$

Note that μ and σ^2, i.e., mean and covariance functions here are different from μ_H and σ_H^2 functions defined for the stochastic simulation in (4) and (5). Here, $K(w,v;l) \in \mathbb{R}^{m\times n}$ is the kernel matrix that defines the covariance function between $w \in \mathbb{R}^{m\times d}$ and $v \in \mathbb{R}^{n\times d}$, with $K_{a,b}(w,v;l) = \exp\left(\frac{\|w_a - v_b\|^2}{2l^2}\right)$ using $w_a \in \mathbb{R}^{1\times d}$ and $v_b \in \mathbb{R}^{1\times d}$ as the a^{th} and b^{th} row of matrices w and v, respectively. I is an identity matrix combined with σ_m which represents the noise term to avoid the overfitting considering uncertainty among observations from the function (RW) (Pedregosa et al. 2011). While the predictor (6) resembles stochastic kriging, we note that here the prediction error $\hat{f}(\theta'|X) - \tilde{f}(\theta')$ only represents extrinsic uncertainty, that is the uncertainty about the response surface at a point not yet visited; in other words, (intrinsic) stochastic uncertainty is not included here.

The second component is the the acquisition function that guides the optimization process to sample the next solution θ in a bid to approach optimality. The idea behind the acquisition function is to convert the probabilistic prediction of the fitted surrogate model into a quantifiable scalar score, which then can be used to identify the most promising areas for further exploration. Some well-known acquisition functions, namely, upper confidence bound (UCB), probability of improvement (PI), and expected improvement (EI), are summarized in Table 1. At each moment of drawing a new sample, $\hat{\theta}^*$ represents the best solution found so far, $\Phi(\cdot)$ and $\varphi(\cdot)$ denote the cdf and pdf of standard normal distribution.

Table 1: Analytical expressions of common acquisition functions within noise-free settings

Acquisition function	Analytical expression
UCB(θ)	$\mu(\theta) + \lambda\sigma(\theta)$
PI(θ)	$\Phi\left(\frac{\lvert\tilde{f}(\hat{\theta}^*)\rvert - \mu(\theta)}{\sigma(\theta)}\right)$
EI(θ)	$(\lvert\tilde{f}(\hat{\theta}^*)\rvert - \mu(\theta))\Phi\left(\frac{\lvert\tilde{f}(\hat{\theta}^*\rvert) - \mu(\theta)}{\sigma(\theta)}\right) + \sigma(\theta)\varphi\left(\frac{\lvert\tilde{f}(\hat{\theta}^*)\rvert - \mu(\theta)}{\sigma(\theta)}\right)$

In the UCB method, exploration and exploitation are balanced with the parameter $\lambda > 0$. When λ is close to 0, UCB prioritizes sampling in regions where the mean estimate is minimized, potentially exploiting the good regions that are discovered. In contrast, large λ makes exploration more preferred, encouraging sampling in regions with more uncertainty, even if the estimated mean function value is not as high. In the PI method, instead of solely utilizing the mean and variance of the surrogate model, the probability of having a better estimate than the current best known value is considered. Since the function estimates follow a normal distribution, the reparametrization trick

$$\hat{f}(\theta|X) = \mu(\theta) + \sigma(\theta)Z \text{ with } Z \sim \mathcal{N}(0,1), \tag{7}$$

yields the expression in Table 1. In the EI approach, instead of relying on the probability, the expected magnitude of improvements over the best known solution is considered. Similarly, using the reparametrization trick, analytical expression is also attainable. Not every acquisition function provides a closed-from expression. When a closed-form expression is unavailable, alternative approaches like Monte Carlo-based methods are often considered (Balandat et al. 2020).

4 BO WITH ROOT-FINDING STRATEGY

In this section, we modify BO in two ways. In the first component that fitting occurs, unlike the standard BO for calibration, we fit the GPR to the original function values rather than to $|\tilde{f}(\theta)|$. In the second component, we solve the problem with a root finding rather than a minimization. It is important to note, while the objective of finding $\min_{\theta\in\Theta} |\tilde{f}(\theta)|$ is identical, standard BO calibration frames the problem where *observations* are positivized, root-finding retains the *observations*. This means rather than a standard calibration problem where *observations* are positivized, here we propose to keep the observations as they are but positivize the *predictions* instead. Hence, instead of searching for a new θ whose $\hat{f}(\theta) \leq \tilde{f}(\hat{\theta}^*)$ with high probability, we search for a θ whose $|\hat{f}(\theta)| \leq |\tilde{f}(\hat{\theta}^*)|$ with high probability. For this component, we derive the new acquisition functions and signify them with subscript "RF" to indicate the use of root-finding. We assume that the positive and negative sign in the function value is equally important. When this assumption does not hold, one can introduce a weighting term if desired. We will begin the new acquisition function variants by discussing the UCB method with the proposed modification:

$$\mathrm{UCB}_{\mathrm{RF}}(\theta) = |\mu(\theta)| + \lambda\sigma(\theta).$$

In this modification, $\mathrm{UCB}_{\mathrm{RF}}$ simply takes the absolute value of the mean prediction $\mu(\theta)$ and is added with user-defined parameter λ and the prediction standard deviation $\sigma(\theta)$ Similar to UCB, λ governs the balance between the exploration and exploitation in $\mathrm{UCB}_{\mathrm{RF}}$. As the expression of $\mathrm{UCB}_{\mathrm{RF}}$ is similar to UCB, we can expect almost identical computational effort for this variants.

Now we present the modifications of improvement-based acquisition functions. Within the root-finding scheme, the "improvement" needs to quantify the improvement in terms of the amount that the new solution might reduce the distance to zero. Accordingly, we define the improvement event in root-finding as

$$\mathrm{I}_{\mathrm{RF}}(\theta) = \{|\hat{f}(\theta)| \leq |\tilde{f}(\hat{\theta}^*)|\}.$$

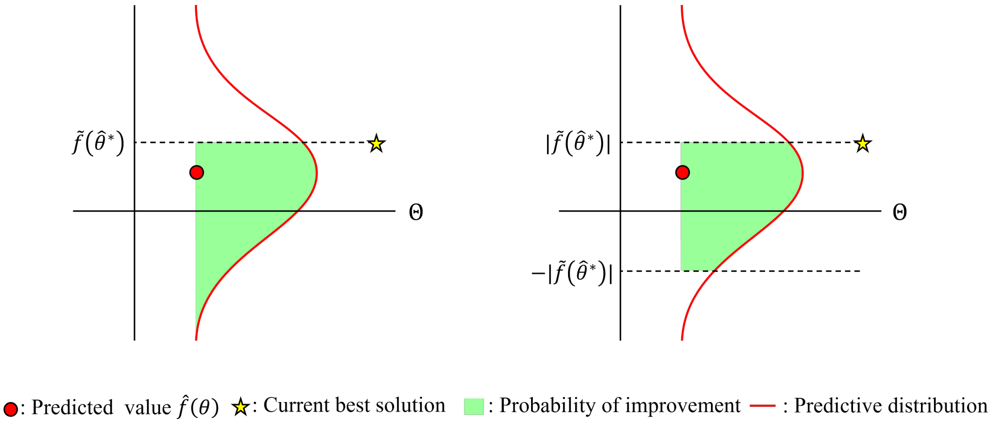
$\tilde{f}(\hat{\theta}^*)$
Θ
$|\tilde{f}(\hat{\theta}^*)|$
Θ
$-|\tilde{f}(\hat{\theta}^*)|$
●: Predicted value $\hat{f}(\theta)$
☆: Current best solution
🟩: Probability of improvement
—: Predictive distribution

Figure 2: Probability of improvement defined in the minimization (left) and root-finding (right) context

Figure 2 illustrates the distinct improvement defined in this context. In the typical minimization strategy, improvement is defined as the portion of having smaller values than $\tilde{f}(\hat{\theta}^*)$ in the predictive distribution, whereas the improvement is bounded within $|\tilde{f}(\hat{\theta}^*)|$ and $-|\tilde{f}(\hat{\theta}^*)|$ for the I_{RF} case. Based on this, we derive $\mathrm{PI}_{\mathrm{RF}}$ using reparameterization trick (7). In the following derivation, recall that $\tilde{f}(\theta^*)$ is the best *observed* function value while $\hat{f}(\theta)$ is the *predicted* function value at a θ whose objective function is not yet evaluated.

$$\mathrm{PI}_{\mathrm{RF}}(\theta) = \mathbb{P}\left(\mathrm{I}_{\mathrm{RF}}\right) = \mathbb{P}\left(-|\tilde{f}(\hat{\theta}^*)| \leq \mu(\theta) + \sigma(\theta)Z \leq |\tilde{f}(\hat{\theta}^*)|\right) = \Phi\left(\mathrm{z}_{\mathrm{ub}}(\theta)\right) - \Phi\left(\mathrm{z}_{\mathrm{lb}}(\theta)\right),$$

where $\mathrm{z}_{\mathrm{ub}}(\theta) = \frac{|\tilde{f}(\hat{\theta}^*)| - \mu(\theta)}{\sigma(\theta)}$ and $\mathrm{z}_{\mathrm{lb}}(\theta) = \frac{-|\tilde{f}(\hat{\theta}^*)| - \mu(\theta)}{\sigma(\theta)}$. Similarly, for the $\mathrm{EI}_{\mathrm{RF}}$, we can write

$$\begin{aligned}
\mathrm{EI}_{\mathrm{RF}}(\theta) &= \int_{-\infty}^{\infty} \mathrm{I}_{\mathrm{RF}}(\theta)\varphi(z)dz = \int_{\mathrm{z}_{\mathrm{lb}}(\theta)}^{\mathrm{z}_{\mathrm{ub}}(\theta)} \left(|\tilde{f}(\hat{\theta}^*)| - |\mu(\theta) + \sigma(\theta)z|\right)\varphi(z)dz \\
&= |\tilde{f}(\hat{\theta}^*)|(\Phi(\mathrm{z}_{\mathrm{ub}}(\theta)) - \Phi(\mathrm{z}_{\mathrm{lb}}(\theta))) - \int_{\mathrm{z}_{\mathrm{lb}}(\theta)}^{\mathrm{z}_{\mathrm{ub}}(\theta)} |\mu(\theta) + \sigma(\theta)z|\varphi(z)dz
\end{aligned}$$

$$
\begin{aligned}
&= |\tilde{f}(\hat{\theta}^*)|(\Phi(\mathrm{z}_{\mathrm{ub}}(\theta)) - \Phi(\mathrm{z}_{\mathrm{lb}}(\theta))) - \int_{\mathrm{z}_{\mathrm{thr}}(\theta)}^{\mathrm{z}_{\mathrm{ub}}(\theta)} (\mu(\theta) + \sigma(\theta)z)\varphi(z)dz \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \int_{\mathrm{z}_{\mathrm{lb}}(\theta)}^{\mathrm{z}_{\mathrm{thr}}(\theta)} (\mu(\theta) + \sigma(\theta)z)\varphi(z)dz \\
&= |\tilde{f}(\hat{\theta}^*)|(\Phi(\mathrm{z}_{\mathrm{ub}}(\theta)) - \Phi(\mathrm{z}_{\mathrm{lb}}(\theta))) + \mu(\theta)(2\Phi(\mathrm{z}_{\mathrm{thr}}(\theta)) - \Phi(\mathrm{z}_{\mathrm{lb}}(\theta)) - \Phi(\mathrm{z}_{\mathrm{ub}}(\theta))) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \sigma(\theta)(2\varphi(\mathrm{z}_{\mathrm{thr}}(\theta)) - \varphi(\mathrm{z}_{\mathrm{lb}}(\theta)) - \varphi(\mathrm{z}_{\mathrm{ub}}(\theta))),
\end{aligned}
$$

where $\mathrm{z}_{\mathrm{thr}}(\theta) = -\mu(\theta)/\sigma(\theta)$.

Moving forward, we now discuss the internal optimization process in BO. Once the acquisition function is chosen, BO needs to seek the maximum value of this function to determine the next observation point at each iteration. A typical approach employed by many practitioners is multi-starting an efficient locally convergent solver such as L-BFGS-B (Frazier 2018) with derivative-free schemes. For our root-finding acquisition functions, first-order derivatives are accessible for all methods and their usage is generally recommended (due to the space limit, we omit their derivatives' closed form expressions).

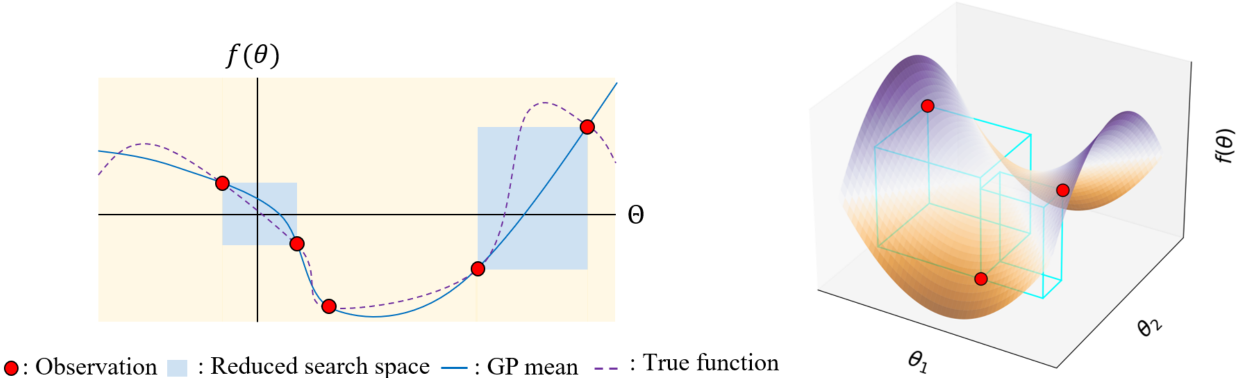

Figure 3: Bounding the search space where sign of the function value changes in 1D (left) and 2D (right).

Moreover, root finding offers an additional advantage to accelerate maximization of the acquisition function. Recall we are using a continuous surrogate model. Due to continuity, once we observe opposite signs of the function values, our fitted GPR inherently passes through the root points, meaning the maximum acquisition function value is guaranteed to exist within an interval between the closest points with opposite function values. Hence, rather than arbitrarily allocating the computational effort across the entire search space Θ, we may only need to focus on these regions, as depicted in Figure 3. That is, without any prior knowledge of the search space, aforementioned multi-start strategy can be inefficient, particularly for the high-dimensional space. As the internal optimization process is the most computationally demanding part in BO (with frequent computation in (6)), effectively narrowing down the search space is indeed a crucial task for the scalability of BO.

To generalize this process, we introduce the Reduced Search Space (RSS) framework, as outlined in Algorithm 1. Here, we define $V(\cdot,\cdot)$ to represent the hypervolume of the region bounded by two points with a differing sign of the function value. This hypervolume is computed based on the spatial distance across the dimensions and the function values. Intuitively, the smaller the hypervolume, the more likely we are to find the roots and maximum acquisition function value within the limited computational resources.

It is important to note that, while RSS is devised to narrow down the search space, RSS does not consider the redundancy between the regions. Numerous questions remain open in this topic, such as determining the optimal number of regions, efficiently allocating the computational burden to the selected regions, or minimizing the redundancy between the regions. In this paper, we have only considered the case of $\gamma = 1$, pursuing the most promising (smallest) region, as we expect the fitted GPR is relatively

Algorithm 1: Reduced Search Space (RSS)

```
Input  : observed solutions: $X \in \mathbb{R}^{m \times d}$, observed function values: $Y \in \mathbb{R}^{m}$, number of targeted regions: $\gamma$.
Output:  $\gamma$ subregions to search within
Initialize $D$ as an empty list to store pairs $(i, j)$ with sign changes;
for each pair $(i, j)$, $i < j$ do
    if $\tilde{f}(\theta^i) \cdot \tilde{f}(\theta^j) < 0$ then
        Append $(i, j)$ to $D$;
        Compute hypervolume $V(\theta_i, \theta_j) = \left(\prod_{l=1}^{d} |\theta_l^i - \theta_l^j|\right) \times |\tilde{f}(\theta^i) - \tilde{f}(\theta^j)|$, where $\theta_b^a$ represents the $b^{th}$
          element of the $a^{th}$ observed solution;
    end
end
Sort pairs in $D$ by $V(\theta_i, \theta_j)$ and return $\gamma$ pairs with the minimum hypervolume;
```

smooth with the use of rbf kernel. However, in the case of using a more complex kernel such as Matérn kernel, increasing the value of γ would be a more plausible choice to capture more intricate structure of the function.

5 EXPERIMENT

In this section, we conduct experiments to compare the calibration accuracy of the proposed BO with root finding compared to standard BO. We compare the proposed BO that fits a surrogate model to the signed discrepancy but uses acquisition functions that tackle root finding versus the standard BO when the surrogate model is fitted to either the absolute or squared discrepancy. In each experiment, we conduct 100 independent macro-replications that each are allowed to use only 100 observations. In the implementation detail, we utilize the GPR configuration used by Pedregosa et al. (2011). Moreover, to ensure the reproducibility and fair comparison, each method uses the same pseudo-random seed and fixed starting point at each macro-replication. For the internal optimization process, we employ 50 multi-start L-BFGS-B that is run up to 10 iterations, following the set-up by Virtanen et al. (2020) and using the first-derivative of each method; see the appendix for the derivations.

5.1 2D Synthetic Data

Suppose our DT model consists of two governing parameters, and the deviation from the RW follows the modified Himmelblau's function (Himmelblau 1972):

$$\tilde{f}(\theta_1, \theta_2) = \log_2\left((\theta_1^2 + \theta_2 - 5)^2 + (\theta_1 + \theta_2^2 - 3)^2\right) - 1, \tag{8}$$

for $\theta_1, \theta_2 \in [-5, 5]$. Here, our optimization objective is to identify the optimal parameter set (state) that minimizes the discrepancy. Figure 4 summarizes the results along with the number of observations. Each point represents the minimum absolute function value achieved after averaging the macro-replications that use the corresponding BO strategy for each category of acquisition functions. Notably, root-finding approach outperforms the standard BO with minimization across cases. The RSS combined root-finding approach more resoundingly improves the convergence thanks to its better guided sampling process.

To provide further insights into this result, we present Figure 5, which displays the sampling trace of each method with 10 observed solutions, all starting from the common starting point. Here, the optimal solutions are distributed in the white borderline areas. Interestingly, standard acquisition functions with minimization strategy spend most of their sampling for exploration. Whereas in the root-finding case, we see the sampling trace being more exploitation-oriented right after the sign difference is captured. We highlighted this tendency in Section 2, where discarding the sign of the observation can potentially complicate the optimization process unless near optimal regions.

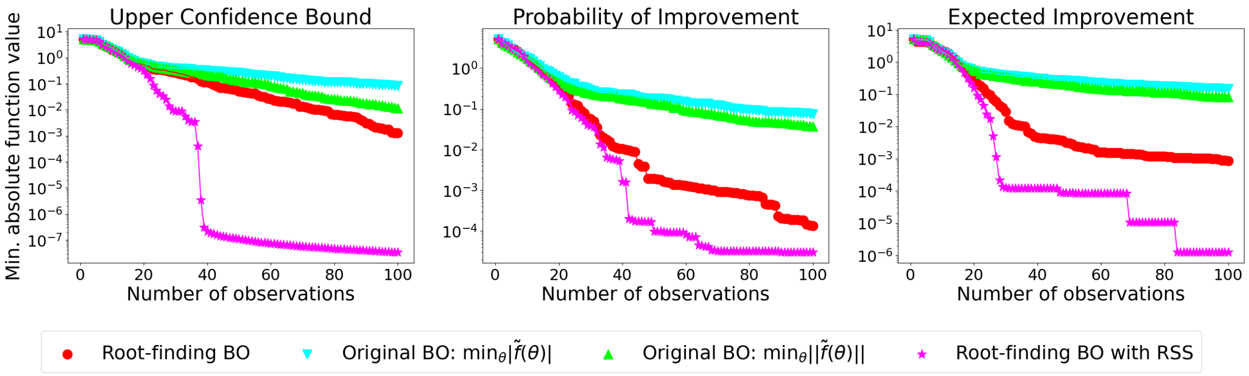

Figure 4: Experimental results on 2D synthetic dataset

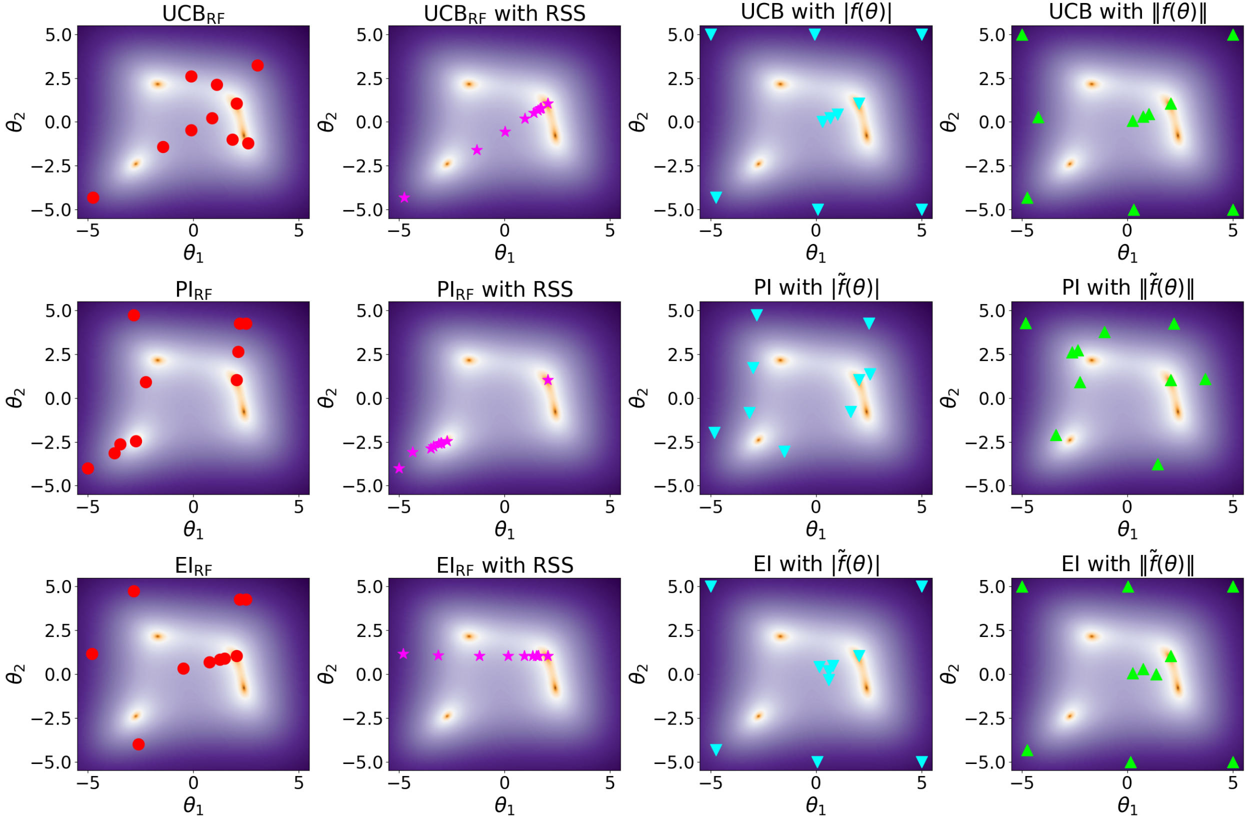

Figure 5: Sampling trace of each method after 10 observed solutions with starting point $[\theta_1, \theta_2] = [2.05, 1.05]$.

5.2 M/M/1 Queue

In this experiment, suppose our DT imitates the RW which is an M/M/1 queue with two parameters, arrival rate $\lambda^c = 4$ per hour and service rate $\mu^c = 5$ per hour. Suppose we know the service rate but we don't know the true arrival rate in RW. Moreover, suppose we are given the average sojourn time of 1000 entities entering the system $h^c(\lambda^c, \mu^c)$ (in one single day) but not each individual entity's arrival or service time. The purpose of calibration in this example is to determine the arrival rate $\lambda \in [1, 10]$ that matches the 1000-average sojourn time of the DT queue with that of the RW. At each λ, we only conduct a single replication (to mimic the time constraints and sample path optimization posed in the problem statement), but we do this multiple times (macro-replications) to study the variability of results. We define the raw discrepancy measure

$$\tilde{f}(\lambda) = h^c(\lambda^c, \mu^c) - h(\lambda, \mu^c). \tag{9}$$

Figure 6 illustrates the distribution of $\tilde{f}(\lambda)$ across different arrival rates. When $\lambda > \lambda^c$, we are likely to observe negative discrepancies as our DT presumes more frequent arrivals, resulting in longer waiting time. Conversely, when $\lambda < \lambda^c$, DT underestimates the average waiting time and yields a positive discrepancy.

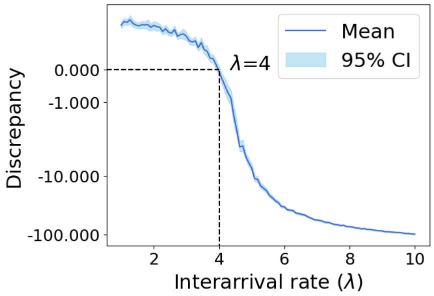

Figure 6: Distribution of discrepancy between RW and DT

Figure 7 reports the overall experimental results in M/M/1 example similar to the 2D synthetic data example. In contrast to the previous experiment, no noticeable differences are observed comparing the minimization and root-finding, except for the UCB case. Nevertheless, root-finding with RSS approach still significantly outperforms all other approaches across acquisition functions both in early and last stages.

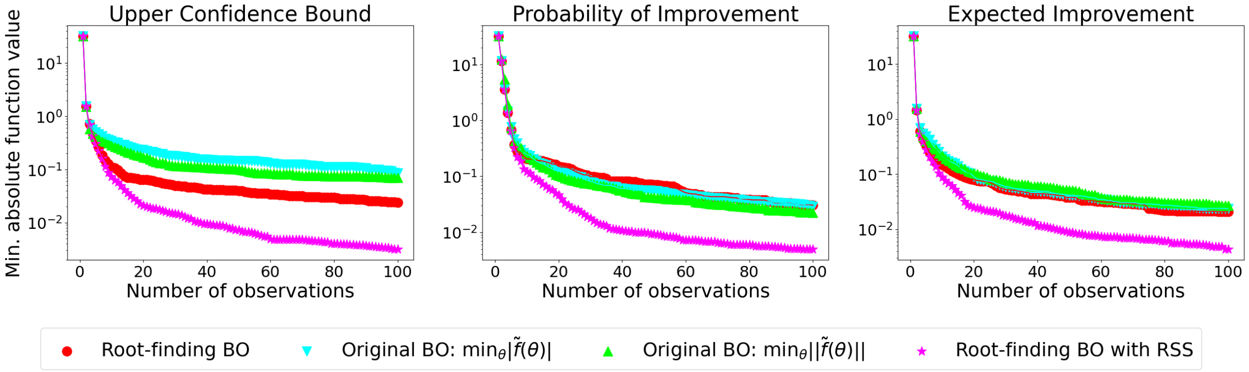

Figure 7: Experimental results on M/M/1 queue example

6 CONCLUSIONS

This paper introduces a novel approach to calibrate DT by leveraging a root-finding strategy in BO. Our proposition is particularly valid in cases where there is a single observation from the RW and to save time we only afford a single observation from the DT at each visited solution. In our experiments, we empirically demonstrate the underperforming behavior of the minimization strategy particularly when using GPR surrogate models. Our suggestion is therefore to fit GPR on the signed discrepancy values but modify acquisition functions to find roots of the prediction function. We believe that our methodology can easily extend to cases where multiple observations from the DT is available. On the other hand, our BO with root-finding strategy would also be applicable in scenarios where the sign of the discrepancy imposes distinct meaning that affects the decision-making procedure for the practitioner, and thereby the calibration task has to be carried out with the sign information. Lastly, we note that this topic is closely related to the modern trends of combining local optimization strategies into BO (Eriksson et al. 2019; Nguyen et al. 2022). With the extension of this work, this root-finding based BO can be applied into finding the root of derivatives, pursuing the local optimum of the function.

APPENDIX

In this section, we present the first-order derivatives of the proposed methods discussed in Section 4. To do this, we first need to derive the gradient of the estimates (6). The gradient of the mean estimates can be derived as

$$\frac{\partial \mu(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left(K(\theta, X; l)[K(X,X;l)+\sigma_m I]^{-1}Y\right) = \frac{\partial K(\theta, X; l)}{\partial \theta}[K(X,X;l)+\sigma_m I]^{-1}Y$$

and the gradient of standard deviation is

$$\begin{aligned}\frac{\partial \sigma(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\sqrt{\sigma^2(\theta)} = \frac{1}{2\sqrt{\sigma^2(\theta)}}\frac{\partial \sigma^2(\theta)}{\partial \theta} \\ &= -\frac{1}{2\sqrt{\sigma^2(\theta)}}\left(\frac{\partial K(\theta, X; l)}{\partial \theta}[K(X,X;l)+\sigma_m I]^{-1}K(X,\theta;l)\right. \\ &\qquad\qquad\qquad\qquad \left. + K(\theta, X; l)[K(X,X;l)+\sigma_m I]^{-1}\frac{\partial K(X,\theta;l)}{\partial \theta}\right),\end{aligned}$$

where $\frac{\partial K(w,v;l)}{\partial w} = \frac{\|w-v\|}{l^2}\exp\left(\frac{\|w-v\|^2}{2l^2}\right)$, following the notation used in Section 3. With this, the first-order derivative of UCB_{RF} becomes

$$\frac{\partial}{\partial \theta}\text{UCB}_{\text{RF}} = \begin{cases} \frac{\partial \mu(\theta)}{\partial \theta} + \lambda\frac{\partial \sigma(\theta)}{\partial \theta}, \text{ if } \mu(\theta) > 0 \\ -\frac{\partial \mu(\theta)}{\partial \theta} + \lambda\frac{\partial \sigma(\theta)}{\partial \theta}, \text{ else} \end{cases}$$

Accordingly, the first-order derivative of PI_{RF} is

$$\frac{\partial}{\partial \theta}\text{PI}_{\text{RF}}(\theta) = \frac{\partial}{\partial \theta}\left(\Phi(\mathrm{z}_{\text{ub}}(\theta)) - \Phi(\mathrm{z}_{\text{lb}}(\theta))\right) = \varphi(\mathrm{z}_{\text{ub}}(\theta))\frac{\partial \mathrm{z}_{\text{ub}}(\theta)}{\partial \theta} - \varphi(\mathrm{z}_{\text{lb}}(\theta))\frac{\partial \mathrm{z}_{\text{lb}}(\theta)}{\partial \theta},$$

where $\frac{\partial \mathrm{z}_{\text{ub}}(\theta)}{\partial \theta} = \frac{-\frac{\partial \mu(\theta)}{\partial \theta}\sigma(\theta) - \left(|\tilde{f}(\hat{\theta}^*)| - \mu(\theta)\right)\frac{\partial \sigma(\theta)}{\partial \theta}}{\sigma^2(\theta)}$ and $\frac{\partial \mathrm{z}_{\text{lb}}(\theta)}{\partial \theta} = \frac{-\frac{\partial \mu(\theta)}{\partial \theta}\sigma(\theta) + \left(|\tilde{f}(\hat{\theta}^*)| + \mu(\theta)\right)\frac{\partial \sigma(\theta)}{\partial \theta}}{\sigma^2(\theta)}$. Likewise, we derive the first order deritative of EI_{RF} as

$$\begin{aligned}\frac{\partial}{\partial \theta}\text{EI}_{\text{RF}} = \frac{\partial}{\partial \theta}(&|\tilde{f}(\hat{\theta}^*)|(\Phi(\mathrm{z}_{\text{ub}}(\theta)) - \Phi(\mathrm{z}_{\text{lb}}(\theta)) + \mu(\theta)(2\Phi(\mathrm{z}_{\text{thr}}(\theta)) - \Phi(\mathrm{z}_{\text{lb}}(\theta)) - \Phi(\mathrm{z}_{\text{ub}}(\theta))) \\ &\qquad - \sigma(\theta)(2\varphi(\mathrm{z}_{\text{thr}}(\theta)) - \varphi(\mathrm{z}_{\text{lb}}(\theta)) - \varphi(\mathrm{z}_{\text{ub}}(\theta)))) \\ = |\tilde{f}(\hat{\theta}^*)|&\left(\frac{\partial \mathrm{z}_{\text{ub}}(\theta)}{\partial \theta}\varphi(\mathrm{z}_{\text{ub}}(\theta)) - \frac{\partial \mathrm{z}_{\text{lb}}(\theta)}{\partial \theta}\varphi(\mathrm{z}_{\text{lb}}(\theta))\right) \\ &+ \frac{\partial \mu(\theta)}{\partial \theta}\Big(2\Phi(\mathrm{z}_{\text{thr}}(\theta)) - \Phi(\mathrm{z}_{\text{lb}}(\theta)) - \Phi(\mathrm{z}_{\text{ub}}(\theta))\Big) \\ &+ \mu(\theta)\left(2\frac{\partial \mathrm{z}_{\text{thr}}(\theta)}{\partial \theta}\varphi(\mathrm{z}_{\text{thr}}(\theta)) - \frac{\partial \mathrm{z}_{\text{lb}}(\theta)}{\partial \theta}\varphi(\mathrm{z}_{\text{lb}}(\theta)) - \frac{\partial \mathrm{z}_{\text{ub}}(\theta)}{\partial \theta}\varphi(\mathrm{z}_{\text{ub}}(\theta))\right) \\ &- \frac{\partial \sigma(\theta)}{\partial \theta}\left(2\varphi(\mathrm{z}_{\text{thr}}(\theta)) - \varphi(\mathrm{z}_{\text{lb}}(\theta)) - \varphi(\mathrm{z}_{\text{ub}}(\theta))\right)\end{aligned}$$

$$- \sigma(\theta) \left(-2 \frac{\partial \mathrm{z}_{\mathrm{thr}}(\theta)}{\partial \theta} \frac{\mathrm{z}_{\mathrm{thr}}(\theta)}{\sqrt{2\pi}} e^{-\frac{\mathrm{z}^2_{\mathrm{thr}}(\theta)}{2}} + \frac{\partial \mathrm{z}_{\mathrm{lb}}(\theta)}{\partial \theta} \frac{\mathrm{z}_{\mathrm{lb}}(\theta)}{\sqrt{2\pi}} e^{-\frac{\mathrm{z}^2_{\mathrm{lb}}(\theta)}{2}} + \frac{\partial \mathrm{z}_{\mathrm{ub}}(\theta)}{\partial \theta} \frac{\mathrm{z}_{\mathrm{ub}}(\theta)}{\sqrt{2\pi}} e^{-\frac{\mathrm{z}^2_{\mathrm{ub}}(\theta)}{2}} \right),$$

where $\frac{\partial \mathrm{z}_{\mathrm{thr}}(\theta)}{\partial \theta} = \frac{-\frac{\partial \mu(\theta)}{\partial \theta}\sigma(\theta) + \mu(\theta)\frac{\partial \sigma(\theta)}{\partial \theta}}{\sigma^2(\theta)}$.

REFERENCES

Agalianos, K., S. Ponis, E. Aretoulaki, G. Plakas and O. Efthymiou. 2020. “Discrete Event Simulation and Digital Twins: Review and Challenges for Logistics”. *Procedia Manufacturing* 51:1636–1641.

Balandat, M., B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson *et al*. 2020. “BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization”. In *Advances in Neural Information Processing Systems 33*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, 21524–21538. New York: Curran Associates Inc.

Chakrabarty, A., S. A. Bortoff, and C. R. Laughman. 2023. “Simulation Failure-Robust Bayesian Optimization for Data-Driven Parameter Estimation”. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53(5):2629–2640.

Eriksson, D., M. Pearce, J. Gardner, R. D. Turner and M. Poloczek. 2019. “Scalable Global Optimization via Local Bayesian Optimization”. In *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. Fox, and R. Garnett, 5496–5507. New York: Curran Associates Inc.

Frazier, P. I. 2018. “A Tutorial on Bayesian Optimization”. *arXiv preprint arXiv: 1807.02811*.

Himmelblau, D. M. 1972. *Applied Nonlinear Programming*. 1st ed. New York: McGraw-Hill Inc.

Liu, M., S. Fang, H. Dong, and C. Xu. 2021. “Review of Digital Twin about Concepts, Technologies, and Industrial Applications”. *Journal of Manufacturing Systems* 58:346–361.

Nguyen, Q., K. Wu, J. Gardner, and R. Garnett. 2022. “Local Bayesian Optimization via Maximizing Probability of Descent”. In *Advances in Neural Information Processing Systems 35*, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 13190–13202. New York: Curran Associates Inc.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel *et al*. 2011. “Scikit-learn: Machine Learning in Python”. *The Journal of Machine Learning Research* 12:2825–2830.

Peng, Y., M. Zhang, F. Yu, J. Xu and S. Gao. 2020. “Digital Twin Hospital Buildings: An Exemplary Case Study through Continuous Lifecycle Integration”. *Advances in Civil Engineering* 2020:1–13.

Rhodes-Leader, L. A. and B. L. Nelson. 2023. “Tracking and Detecting Systematic Errors in Digital Twins”. In *2023 Winter Simulation Conference (WSC)*, 492–503 https://doi.org/10.1109/WSC60868.2023.10408052.

Schultz, L. and V. Sokolov. 2018. “Bayesian Optimization for Transportation Simulators”. *Procedia computer science* 130:973–978.

Sha, D., K. Ozbay, and Y. Ding. 2020. “Applying Bayesian Optimization for Calibration of Transportation Simulation Models”. *Transportation Research Record* 2674(10):215–228.

Staum, J. 2009. “Better Simulation Metamodeling: The Why, What, and How of Stochastic Kriging”. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, 119–133 https://doi.org/10.1109/WSC.2009.5429320.

VanDerHorn, E. and S. Mahadevan. 2021. “Digital Twin: Generalization, Characterization and Implementation”. *Decision Support Systems* 145:113524.

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau *et al*. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. *Nature Methods* 17(3):261–272.

Wang, W. 2021. “On the Inference of Applying Gaussian Process Modeling to a Deterministic Function”. *Electronic Journal of Statistics* 15(2):5014–5066.

Yang, X., D. Barajas-Solano, G. Tartakovsky, and A. M. Tartakovsky. 2019. “Physics Informed CoKriging: A Gaussian Process Regression based Multifidelity Method for Data Model Convergence”. *Journal of Computational Physics* 395:410–431.

Zhan, S., G. Wichern, C. Laughman, A. Chong and A. Chakrabarty. 2022. “Calibrating Building Simulation Models Using Multi-Source Datasets and Meta-Learned Bayesian Optimization”. *Energy and Buildings* 270:112278.

AUTHOR BIOGRAPHIES

YONGSEOK JEON is a Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests include optimization, simulation and data science. His e-mail address is yjeon@ncsu.edu.

SARA SHASHAANI is an Assistant Professor and Bowman Faculty Scholar in the Edward P. Fitts Department of Industrial and System Engineering at North Carolina State University. Her research interests are probabilistic data-driven models and simulation optimization. She is a co-creator of SimOpt. Her email address is sshasha2@ncsu.edu and her homepage is https://shashaani.wordpress.ncsu.edu/.