In proceedings of the 2021 Winter Simulation Conference

THINKING INSIDE THE BOX: A TUTORIAL ON GREY-BOX BAYESIAN OPTIMIZATION

Raul Astudillo & Peter I. Frazier

School of Operations Research and Information Engineering
Cornell University
Ithaca, NY 14853, USA

ABSTRACT

Bayesian optimization (BO) is a framework for global optimization of expensive-to-evaluate objective functions. Classical BO methods assume that the objective function is a black box. However, internal information about objective function computation is often available. For example, when optimizing a manufacturing line's throughput with simulation, we observe the number of parts waiting at each workstation, in addition to the overall throughput. Recent BO methods leverage such internal information to dramatically improve performance. We call these "grey-box" BO methods because they treat objective computation as partially observable and even modifiable, blending the black-box approach with so-called "white-box" first-principles knowledge of objective function computation. This tutorial describes these methods, focusing on BO of composite objective functions, where one can observe and selectively evaluate individual constituents that feed into the overall objective; and multi-fidelity BO, where one can evaluate cheaper approximations of the objective function by varying parameters of the evaluation oracle.

1 INTRODUCTION

Bayesian optimization (BO) is a framework for global optimization of objective functions that are expensive or time-consuming to evaluate. The standard BO problem is of the form $\max_{x\in\mathbb{X}} f(x)$, where f is an expensive-to-evaluate continuous function and $\mathbb{X}\subset\mathbb{R}^d$ is a simple compact set such as a hyperrectangle or a polytope. Classical BO methods make no other explicit assumptions on the objective function.

The main characteristic of the BO paradigm is that f is modeled as a realization from a Bayesian prior distribution over functions, with Gaussian processes being the most widely used family of distributions. Within an iterative algorithm, this prior distribution along with the evaluations of f performed so far give rise to a posterior distribution which is used via an acquisition function that quantifies the value of information from an objective function evaluation to select the next point at which to evaluate f.

These methods are appealing because they can be easily applied without detailed knowledge of the objective function or derivative evaluations, in contrast with classical nonlinear optimization methods, but nonetheless perform reasonably well across a wide variety of problems (Calandra et al. 2016; Turner et al. 2021). At the same time, they are flexible and permit the introduction of prior information from domain experts in the form of an informative prior distribution, in contrast with non-Bayesian derivative-free methods (Conn et al. 2009).

BO originated with the seminal works of Kushner (1964), Močkus (1975), and Zhilinskas (1975), focused on engineering design, but is best known for its recent success in hyperparameter tuning of machine learning algorithms (Snoek et al. 2012; Swersky et al. 2013; Wu et al. 2020). Beyond engineering design and hyperparameter tuning, BO has also been successful in many other application areas, including operations-focused optimization via simulation applications (Pearce and Branke 2017), drug discovery (Griffiths and Hernández-Lobato 2020), and robotics (Calandra et al. 2016).

While BO has been broadly successful, the expense of evaluations nevertheless remains prohibitive in a number of problem domains. For example, consider optimizing population-level interventions such as social distancing and masking to prevent the spread of a disease based on predictions from an agent-based simulator that models the detailed movements of millions of people. Suppose each evaluation takes several hours on a high-performance computing cluster, the search domain is 10-dimensional, and there are many local maxima. In such problems, it is plausible that a black-box method would need thousands of evaluations or more before it finds a solution close to the global optimum, requiring months of computation. At the same time, suppose this agent-based simulation can be run with a smaller population size to obtain an approximation to the objective in dramatically less time. Then, it becomes appealing to use a method that uses such less accurate but faster approximations to understand how the objective behaves at a high level (see, e.g., Figure 3), and only afterward focuses its attention on high-value regions.

This approach, in which lower fidelity faster evaluations are used in concert with higher fidelity slow evaluations, is called *multi-fidelity optimization* (or *multi-fidelity BO* when specialized to the BO setting) (Huang et al. 2006; Forrester et al. 2007). Such methods generate value by sacrificing the generality of black box optimization, leveraging knowledge and access to the internals of the objective function evaluation to accelerate search. While such methods require more specialization, they can be much faster.

As we articulate here, multi-fidelity BO is just one example within a broader class of methods that leverage knowledge and access to the internals of objective function evaluation to improve efficiency. We refer collectively to such methods as *grey-box Bayesian optimization methods*. Specifically, we refer to any method as a grey-box BO method if it leverages access to the internal computational structure of objective function or constraint evaluation. This can deliver dramatic performance gains, sometimes improving accuracy multiple orders of magnitude at a given level of computational effort (see, e.g., Figure 1).

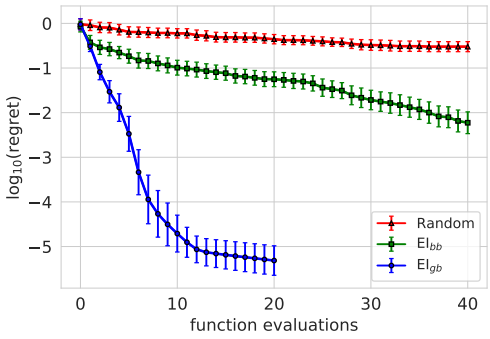


Figure 1: Performance of grey-box BO (EI_{gb}) compared to standard black-box BO (EI_{bb}) and random search (Random) on the calibration test problem described in §5.3 of Astudillo and Frazier (2019). Grey-box BO leverages the composite structure of the objective function and, by doing so, it dramatically improves performance. Grey-box BO achieves the same regret as standard BO after 40 evaluations using only 5 evaluations.

Existing work on grey-box BO can be broadly divided into three classes: BO of composite objective functions; multi-fidelity BO; and BO with objective function constituent evaluations.

- **BO of composite objective functions** uses observations of internal constituents that feed into the overall objective value calculation to improve the predictive model of the objective function. This arises, for example, when calibrating a simulator's parameters to data and in inverse reinforcement learning. The objective function at a given vector of parameters x is the sum of squared errors between the simulator's prediction $h_j(x)$ for an experimental condition and an observation y_j^{obs} of that condition. Rather than simply modeling the overall objective as one monolithic black box, one can model each function h_j as a black box and understand that the objective is the composition of these functions with the function $g(y) = \sum_j (y_j - y_j^{\text{obs}})^2$. Using observations of $h_j(x)$ provides substantially more information that can be used to select good x at which to evaluate (Uhrenholt and Jensen 2019; Astudillo and Frazier 2019). This also arises, for example, in aerospace engineering, where multiple physics-based simulators pass information back and forth, creating an objective function that is a composition of a collection of black-box functions; and when minimizing a cost function that aggregates a simulator's predictions across a variety of environmental conditions.

- **Multi-fidelity BO** modifies the process of evaluating the objective function so that the output is not the objective value itself but is instead a faster-to-compute but less accurate approximation. Modifications to the objective function evaluation process include using smaller mesh sizes when solving partial differential equations (PDEs), reducing run-lengths when the objective function is the output of a steady-state simulation, and reducing the number of training iterations when the objective is the test error for a deep neural network (DNN). This is the most widely studied class of grey-box methods to date.
- **BO with objective constituent evaluations** leverages the ability to evaluate just some of the constituents that make up the objective function to save time while also enabling learning, either delaying the evaluation of other constituents until some future point in time or not evaluating them at all. This is possible in both BO of composite objective functions and multi-fidelity BO. For example, when minimizing average cost over scenarios, one can evaluate cost on a subset of the possible scenarios; or when minimizing the test error for a DNN, an evaluation can be paused at a small number of training iterations (and continued later if desired).

The term "grey-box" originates from physics-based modeling, where black-box models are purely empirical models that include no first-principles theoretical knowledge from physics, white-box models exclusively rely on such detailed physical knowledge, and grey-box models blend the two approaches (Tulleken 1993; Bohlin 2006). In grey-box BO, we adopt this same terminology, except that we seek to model the objective function rather than the real world. We use a blend of empirical data-driven methods (black-box methods) and detailed first-principles knowledge of how the objective is computed (white-box methods). Distinct from grey-box optimization, there is work applying black-box surrogate-based optimization methods (which assume nothing about the structure of the objective function and constraints) to optimize such grey-box physics-based models (Beykal et al. 2018).

To help clarify the use of the term grey-box optimization, we also mention here other methods that go beyond standard BO but that we do not consider to be grey-box BO:

- BO with shape constraints (Jauch and Peña 2016). This uses knowledge of the objective function but does not leverage access to its internal computation.
- High-dimensional BO assuming additive structure (Gardner et al. 2017) or a linear embedding in a low-dimensional space (Letham et al. 2020), when such structure is used or assumed (as it often is) without access to the internals of objective function calculation.
- BO with gradient information (Wu et al. 2017). Gradient information is commonly included in objective function oracles used in classical (non-Bayesian) optimization, and is viewed as an externally-provided output. We take that view here.

Organization of the rest of this tutorial The rest of this tutorial first provides a brief introduction and basic concepts in standard BO in §2. For a more detailed tutorial on standard BO, we refer the reader to Frazier (2018). This tutorial on grey-box BO will be most enjoyable to those who have already read such a tutorial focused on standard BO. Then, §3-5 describe different types of grey-box BO: §3 describes BO of composite objective functions; §4 describes multi-fidelity BO. and §5 describes BO with objective constituent evaluations. §6 concludes while offering directions for future research.

2 STANDARD BAYESIAN OPTIMIZATION

A BO method consists of two main components: a predictive model, given by a Bayesian prior distribution over f that serves as a surrogate equipped with uncertainty estimates; and an acquisition function, which depends on the implied posterior distribution over f given the set of available evaluations so far, and whose value at an arbitrary point $x \in \mathbb{X}$ quantifies the *benefit* of evaluating at this point. In this section, we discuss

these two components in detail, focusing on Gaussian processes (GPs), which is the class of probability distributions most widely used in BO, and reviewing several popular acquisition functions.

2.1 Predictive Model

As mentioned above, the first component of a BO method is a predictive model, given by a Bayesian prior probability distribution over f. Examples of probability distributions used in the BO literature include random forests (Hutter et al. 2011), Bayesian neural networks (Snoek et al. 2015) and GPs. Here, we focus on the latter class of probability distributions, which is arguably the most widely used in practice due to its computational tractability and well-calibrated uncertainty estimates. For a detailed discussion on GPs, we refer the reader to Rasmussen and Williams (2006).

A GP prior distribution over f is fully determined by a prior mean function, $\mu_0 : \mathbb{X} \rightarrow, \mathbb{R}$ and a prior covariance function, $K_0 : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$. It has the property that, for any finite collection of points $x_1, \ldots, x_n \in \mathbb{X}$, the prior distribution of $(f(x_1), \ldots, f(x_n))^\top$ is multivariate normal with mean vector $(\mu_0(x_1), \ldots, \mu_0(x_n))^\top$ and covariance matrix $(K_0(x_i, x_j))_{i,j=1}^n$. Moreover, given a data set of n (potentially noisy) evaluations, $\mathscr{D}_n = \{(x_i, y_i)\}_{i=1}^n$, where $y_i = f(x_i) + \varepsilon_i$ and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. from a normal distribution with mean zero and variance σ^2, the posterior distribution on f is again a GP with mean and covariance functions

$$\mu_n(x) = \mu_0(x) + K_0(x, x_{1:n}) \left[K_0(x_{1:n}, x_{1:n}) + \sigma^2 I_n\right]^{-1} (y_{1:n} - \mu_0(x_{1:n})),$$
$$K_n(x, x') = K_0(x, x') - K_0(x, x_{1:n}) \left[K_0(x_{1:n}, x_{1:n}) + \sigma^2 I_n\right]^{-1} K_0(x_{1:n}, x'),$$

respectively, where I_n is the n-dimensional identity matrix and, making a slight abuse of notation, we define $y_{1:n} = (y_1, \ldots, y_n)^\top$, $\mu_0(x_{1:n}) = (\mu_0(x_1), \ldots, \mu_0(x_n))^\top$, $K_0(x_{1:n}, x_{1:n}) = (K_0(x_i, x_j))_{i,j=1}^n$, $K_0(x, x_{1:n}) = (K_0(x, x_1), \ldots, K_0(x, x_n))$, and $K_0(x_{1:n}, x') = (K_0(x_1, x'), \ldots, K_0(x_n, x'))^\top$.

Following the Bayesian statistics terminology, μ_n and K_n are called the posterior mean and covariance functions, respectively. The function μ_n can be interpreted as a surrogate for f, whereas K_n equips this surrogate with uncertainty estimates. In particular, when evaluations of f are noiseless; i.e., when $\sigma^2 = 0$, the posterior mean function interpolates the evaluations of f collected so far and their corresponding uncertainty estimates are exactly 0; i.e., $\mu_n(x_i) = f(x_i)$ and $K_n(x_i, x_i) = 0$ for $i = 1, \ldots, n$.

2.2 Acquisition Function

The second component of a BO method is an acquisition function, $\alpha_n : \mathbb{X} \rightarrow \mathbb{R}$, where the sub-index n indicates the dependence on the posterior distribution on f at time n. The value of the acquisition function at a particular point $x \in \mathbb{X}$ can be interpreted as a measure of the *benefit* of evaluating at this point, and thus one wishes to evaluate a point with the highest acquisition value possible. Formally, a BO method using an acquisition function α_n chooses the next point to evaluate, x_{n+1}, as a maximizer of α_n; i.e., $x_{n+1} \in \operatorname{argmax}_{x \in \mathbb{X}} \alpha_n(x)$. Importantly, unlike f, α_n is not expensive-to-evaluate and its gradients are typically available. This makes optimizing the acquisition function substantially easier than optimizing f.

Acquisition functions navigate the trade-off between evaluating points whose objective value is still very uncertain and those whose objective value is expected to be high, commonly known as the exploration-exploitation trade-off. Popular acquisition functions include expected improvement (EI) (Zhilinskas 1975; Jones et al. 1998), knowledge gradient (KG) (Frazier et al. 2008; Frazier et al. 2009; Scott et al. 2011), Gaussian process upper confidence bound (GP-UCB) (Srinivas et al. 2012), entropy search (ES) (Hennig and Schuler 2012), predictive entropy search (PES) (Hernández-Lobato et al. 2014), and max-value entropy search (Wang and Jegelka 2017) (MVES). Below, we discuss in detail EI and KG due to their simplicity and also because they have been the most widely generalized to grey-box settings.

2.2.1 Expected Improvement

The most widely used acquisition function in standard BO is the expected improvement (EI), which is defined by

$$\mathrm{EI}_n(x) = \mathbb{E}_n[\{f(x) - f_n^*\}^+],$$

where the sub-index n indicates that the expectation is taken under the posterior distribution at time n, and $f_n^* = \max_{i=1,\dots,n} f(x_i)$ is the best observed objective value so far. Interpreting f_n^* as the reward that would be received if we reported a solution to our optimization problem after n evaluations, $f_{n+1}^* - f_n^* = \{f(x) - f_n^*\}^+$ is the improvement in this reward due to an additional sample at x.

The EI acquisition function was first proposed by Močkus (1975) and popularized by Jones et al. (1998). It is known to perform well in practice, especially when evaluations are noiseless. At the same time, it is also known to be outperformed by other more sophisticated acquisition functions when evaluations are noisy (Frazier 2018) or the objective function is highly multi-modal (Jiang et al. 2020).

In addition to its good empirical performance, another property contributing to EI's popularity is that it admits an analytic expression when f is modeled using a GP and evaluations are noiseless. This analytic expression, obtained by noting that $\{f(x) - f_n^*\}^+$ is a truncated normal random variable, is given by

$$\mathrm{EI}_n(x) = \Delta_n(x)\Phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) + \sigma_n(x)\varphi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right),$$

where $\Delta_n(x) = \mu_n(x) - f_n^*$, $\sigma_n(x) = \sqrt{K_n(x,x)}$, and φ and Φ are the standard normal probability density function (PDF) and cumulative distribution function (CDF), respectively. Importantly, if μ_n and K_n are differentiable, so is EI_n, and thus (deterministic) gradient-based optimization methods can be used to maximize EI_n. A common choice in practice is to use L-BFGS-B (Byrd et al. 1995) with multiple restarts.

2.2.2 Knowledge Gradient

The knowledge gradient (KG) acquisition function was proposed by Frazier et al. (2008) for Bayesian ranking and selection among a finite number of alternatives and later adapted to the BO setting by Scott et al. (2011). Since then, it has been generalized to handle parallel evaluations (Wu and Frazier 2016), derivative information (Wu et al. 2017), multiple information sources (Poloczek et al. 2017), and also adapted to multiple grey-box settings that we describe in detail in §3, §4, and §5.

KG modifies the reward in EI's definition. EI assumes that the reward for reporting a solution to our optimization problem at time n is the maximum objective value across evaluated points, f_n^*. KG instead assumes this reward is the maximum (expected) objective value across the whole feasible space, $\mu_n^* = \max_{x\in\mathbb{X}} \mathbb{E}_n[f(x)] = \max_{x\in\mathbb{X}} \mu_n(x)$. The reward f_n^* is only improved by full evaluations of the objective function, but μ_n^* can be improved by any information about the objective function. This allows KG to generalize easily to grey-box settings. Analogously to EI, KG is defined as the expected change in reward from one additional evaluation; i.e.,

$$\mathrm{KG}_n(x) = \mathbb{E}_n[\mu_{n+1}^* - \mu_n^* \mid x_{n+1} = x].$$

While μ_n^* is deterministic given the information available up to time n (and thus can be pulled out of the expectation above), μ_{n+1}^* is random due to its dependence on the yet unobserved value of y_{n+1}. Moreover, when $\mathbb{X}$ is continuous, KG_n does not admit a simple analytic expression, and thus maximizing it requires solving a nested stochastic optimization problem. This makes KG significantly harder to maximize than acquisition functions with simple analytic expressions like EI. Very often, however, the extra computation required to maximize KG is justified by its superior performance. This is particularly true in grey-box settings where KG often admits natural generalizations, whereas other acquisition functions are adapted to these settings via less principled heuristics or approximations which often lead to a worse performance.

Several approaches have been proposed in the literature to maximize KG. Scott et al. (2011) proposes to approximate KG_n by replacing μ_m^* by $\widetilde{\mu}_m^* = \max_{i=1,\dots,n+1} \mu_m(x_i)$, for $m = n, n+1$. The key advantage of this approximation is that it admits an analytic expression which, like EI, can be maximized using deterministic gradient-based optimization methods. However, it becomes computationally burdensome for problems of moderate dimension. Wu et al. (2017) proposes an *exact* approach to maximize KG by computing stochastic gradients of KG_n which are then used in a multi-start stochastic gradient ascent (SGA) routine. This approach scales better to problems of moderate dimension. However, using SGA requires choosing its learning rate, which can be non-trivial. Balandat et al. (2020) proposes a *one-shot optimization* approach that effectively replaces the original problem of maximizing KG_n with a sample average approximation (SAA). This approximate problem is not only deterministic but it can also be cast as a non-nested optimization problem over a higher dimensional space, thus allowing again the use of deterministic gradient-based optimization methods. However, the dimension of this approximate problem grows linearly with the number of samples used, restricting the number that can be used in practice.

The above three approaches to maximize KG (implicitly or explicitly) rely on the so-called reparametrization trick for acquisition functions (Wilson et al. 2018), which consists on rewriting an acquisition function as an expectation of a deterministic transformation (depending on the posterior mean and covariance functions) of a standard normal random variable. Such an approach has also been key to extending other acquisition functions such as EI or GP-UCB to settings where they no longer have an analytic expression such as batch evaluations (Wilson et al. 2018; Wang et al. 2020) and composite objective functions (Astudillo and Frazier 2019). The reparameterized expression of KG is given by

$$\text{KG}_n(x) = \mathbb{E}_n \left[\max_{x' \in \mathbb{X}} \mu_n(x') + \widetilde{\sigma}_n(x'; x_{n+1}) Z \mid x_{n+1} = x \right] - \mu_n^*,$$

where $\widetilde{\sigma}(x'; x_{n+1}) = K_n(x', x_{n+1}) / \sqrt{K_n(x_{n+1}, x_{n+1}) + \sigma^2}$, and the (conditional) distribution of Z is standard normal. We refer the reader to Frazier et al. (2009) and Wu and Frazier (2016) for a derivation.

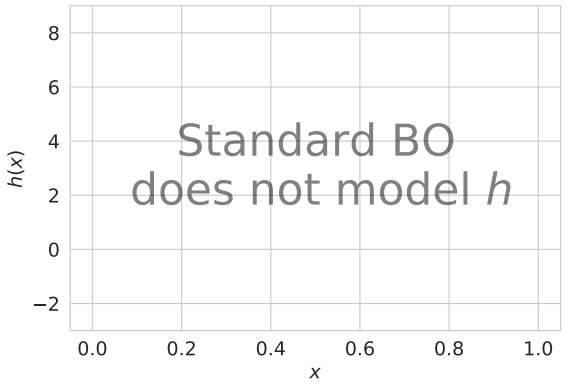
3 BAYESIAN OPTIMIZATION OF COMPOSITE OBJECTIVE FUNCTIONS

The first grey-box BO setting we consider is the *composite objective functions* setting, where the objective function is a known transformation of a vector-valued black-box function. Formally, we assume that the objective function, $f : \mathbb{X} \to \mathbb{R}$, is known to be of the form $f(x) = g(h(x))$, where $h : \mathbb{X} \to \mathbb{R}^k$ is a black-box expensive-to-evaluate vector-valued function, and $g : \mathbb{R}^k \to \mathbb{R}$ is a cheap-to-evaluate scalar function, typically known in closed form. This occurs, for example, in simulation calibration and inverse reinforcement learning, where $h(x)$ is a vector containing predictions for reality and the goal is to find the design variables x so that these predictions most closely matches a vector data observed in the real world, $y^{\text{obs}} \in \mathbb{R}^k$. In this case, a common choice is to minimize $f(x) = g(h(x))$, where $g(y) = \|y - y^{\text{obs}}\|_2^2$.

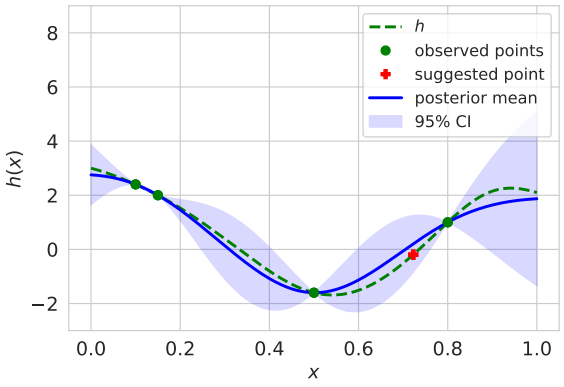
3.1 Predictive Model

Notably, although evaluations of h are available when computing the objective function, the standard BO approach does not use this information (directly). Intuitively, using this information can be beneficial, especially when h carries information relevant for optimization that is not available from f alone. For example, suppose we wish to minimize $f(x) = h(x)^2$, where x and $h(x)$ are both scalars. If $h(x_1) < 0 < h(x_2)$ for some $x_1 < x_2$ and h is continuous, then we know there exists $x^* \in (x_1, x_2)$ such that $h(x^*) = 0$, making x^* a global minimizer of f. This valuable information, however, is ignored by standard BO methods. Figure 2 shows that, in this example, a grey-box BO method that explicitly models h can indeed make a much better sampling decision than a standard black-box BO method that ignores it.

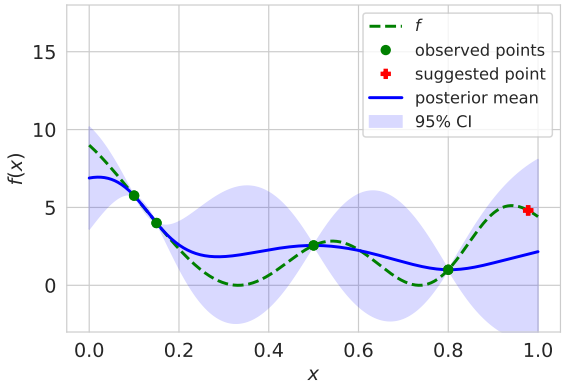
Astudillo and Frazier (2019) addresses this shortcoming by modeling h using a multi-output GP, instead of f using a (single-output) GP as in the standard BO approach. Formally, this approach places a multi-output GP prior distribution on h (Alvarez et al. 2012), which is again characterized by a prior mean function,



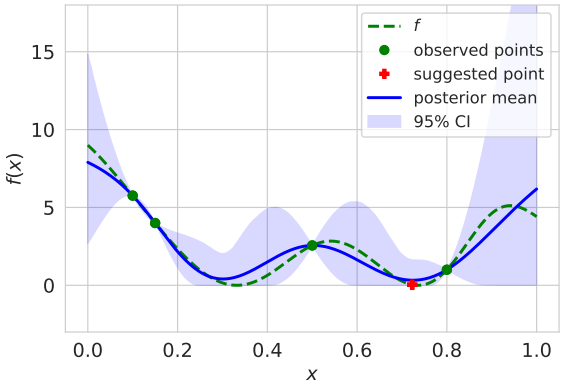
(a) Standard black-box BO does not model h.



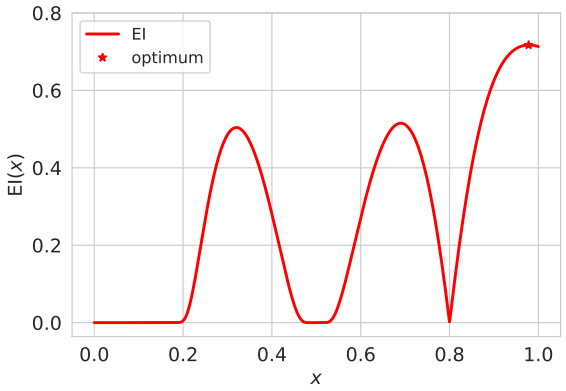
(b) GP posterior on h used by grey-box BO. The optimum occurs when $h(x) = 0$.



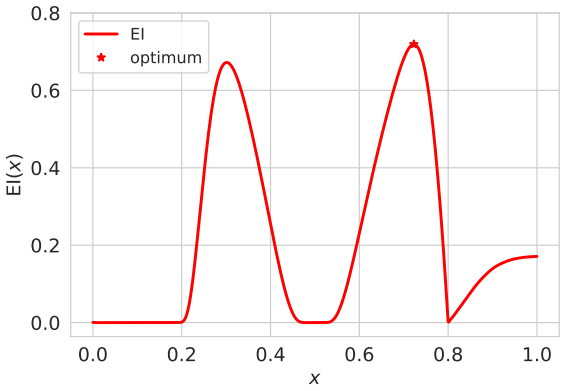
(c) GP posterior on f used by standard black-box BO.



(d) Implied non-Gaussian posterior on f used by grey-box BO. This posterior puts all mass on positive values.



(e) EI computed with respect to the GP posterior on f used by standard black-box BO.



(f) EI computed with respect to the non-Gaussian posterior on f used by grey-box BO.

Figure 2: Illustrative example of BO of a composite objective function in a (minimization) problem where h is scalar-valued and $g(h(x)) = h(x)^2$. Observations of $h(x)$ provide a substantially more accurate view of where global optima of f reside as compared with observations of $f(x)$ alone. This allows grey-box BO (right) to evaluate at points much closer to these global optima compared to standard black-box BO (left).

$\mu_0 : \mathbb{X} \to \mathbb{R}^k$, and a prior covariance function, $K_0 : \mathbb{X} \times \mathbb{X} \to \mathbb{R}^{k \times k}$. Analogous to the single-output case, the posterior distribution on h given n of its evaluations is again a multi-output GP with posterior mean and covariance functions, μ_n and K_n, respectively, which can be computed in closed form. This posterior distribution on h in turn implies a posterior distribution on f, which is in general non-Gaussian.

3.2 Acquisition Functions

Having specified a Bayesian prior probability distribution over f, it remains to specify an acquisition function. When g is non-linear, however, the posterior distribution over f is no longer a GP. Therefore, classical acquisition functions such as PI, EI or GP-UCB no longer have a closed form, thus making them more challenging to compute and maximize.

Uhrenholt and Jensen (2019) considers minimization of $g(h(x)) = \|h(x) - y^{\text{obs}}\|_2^2$, under which the implied posterior distribution on $f(x)$ is a generalized chi-squared distribution. The EI acquisition function under this distribution does not have a closed form expression. However, it is argued that, when the outputs of h are modeled using independent GPs, this distribution can be well approximated by a scaled non-central chi-squared distribution with the same degrees of freedom (k) and non-centrality parameter, where the multiplying factor is chosen so that its expected value matches the one from the true distribution. Under this approximated distribution, EI has a closed-form analytical expression in terms of non-central chi-squared CDFs and can be efficiently optimized using deterministic gradient-based optimization methods.

While the approach proposed by Uhrenholt and Jensen (2019) is appealing due to the closed form analytical expression it provides, the performed approximation has unclear effects. Furthermore, it does not naturally extend to other functions g. Astudillo and Frazier (2019) addresses the more general case by noting that, for arbitrary g, the reparametrization trick can be used to rewrite EI_n as

$$\text{EI}_n(x) = \mathbb{E}_n[\{g(\mu_n(x) + C_n(x)Z_k) - f_n^*\}^+], \tag{1}$$

where $C_n(x)$ is the lower Cholesky factor of $K_n(x,x)$, and Z_k is a k-dimensional standard normal random vector. (Astudillo and Frazier (2019) refers to (1) as the expected improvement for composite functions (EI-CF) to distinguish it from the classical expected improvement.) Equation (1) is then used to show that, under mild regularity conditions, EI_n is differentiable almost everywhere and its gradient, when it exists, is given by $\nabla_x \text{EI}_n(x) = \mathbb{E}_n[\gamma_n(x;Z)]$, where

$$\gamma_n(x;Z_k) = \begin{cases} \nabla_x g(\mu_n(x) + C_n(x)Z_k), \text{ if } g(\mu_n(x) + C_n(x)Z_k) > f_n^*, \\ 0, \text{ otherwise.} \end{cases}$$

In particular, γ_n provides an unbiased estimator of the gradient of EI_n which can be used within SGA with multiple restarts to maximize EI_n.

As a second approach, Balandat et al. (2020) notes that (1) also unlocks the adoption of a SAA scheme to approximately maximize EI_n. This approach consists of fixing M samples from a k-dimensional standard normal distribution, $Z_k^{(1)}, \ldots, Z_k^{(M)}$, and considering the Monte Carlo (MC) estimate of EI_n given by

$$\widehat{\text{EI}}_n\left(x; Z_k^{(1:M)}\right) = \frac{1}{M} \sum_{m=1}^{M} \left\{ g\left(\mu_n(x) + C_n(x) Z_k^{(m)}\right) - f_n^* \right\}^+.$$

Then, the problem

$$\max_{x \in \mathbb{X}} \widehat{\text{EI}}_n\left(x; Z_k^{(1:M)}\right) \tag{2}$$

is solved, and its solution is used as a proxy for the maximizer of EI_n.

Importantly, since the samples $Z_k^{(1)}, \ldots, Z_k^{(M)}$ are fixed, (2) is a deterministic optimization problem. Moreover, under mild regularity conditions, $\widehat{\text{EI}}_n\left(\cdot; Z_k^{(1:M)}\right)$ is differentiable, allowing the use of deterministic

gradient-based optimization. Balandat et al. (2020) shows empirically that this approach produces better results than SGA with multiple restarts at a lower computational cost. In addition, it shows that, under suitable regularity conditions, any solution of (2) converges in probability exponentially fast to a maximizer of EI_n as $M \to \infty$, thus suggesting that in practice it is safe to use low values of M.

While the discussion has focused on extending EI, it is possible to extend other acquisition functions following similar approaches. For example, Balandat et al. (2020) derives an extension of the KG acquisition function for composite objective functions following an analogous SAA approach.

3.3 Other Related Work

The approach of Astudillo and Frazier (2019) was recently extended by Astudillo and Frazier (2021) to a more general class of composite objectives, evaluated via a series of functions, arranged in a directed acyclic network so that each function in the network takes as input the output of its parent nodes. Composite objective functions have also been considered outside the BO framework. For example Wild (2017) developed a trust-region method for derivative-free optimization of a composite objective function where $g(y) = \|y - y^{\text{obs}}\|_2^2$. In contrast with BO, this method is designed for local rather than global optimization. There is also a broad literature on gradient-based methods for optimizing composite objective functions (Burke and Ferris 1995; Shapiro 2003; Drusvyatskiy and Paquette 2019). In addition to derivatives, these methods often rely on convexity. Finally, composite (a.k.a. nested) functions have also been considered in GP-based sequential design of experiments with the goal of prediction rather than optimization (Marque-Pucheu et al. 2019).

4 MULTI-FIDELITY BAYESIAN OPTIMIZATION

The second grey-box setting we consider is multi-fidelity BO, where it is possible to evaluate cheaper approximations of the objective function by varying evaluation oracle parameters. This arises, for example, when optimizing steady-state peraformance as estimated by simulating over a long time horizon: we can simulate over a shorter time to quickly approximate the objective (see Figure 3). It also arises in hyperparameter tuning of DNNs trained via stochastic gradient descent (SGD): we can run SGD for a small number of iterations and obtain a proxy of the accuracy of the DNN that would result if it were trained until convergence using a potentially larger number of iterations.

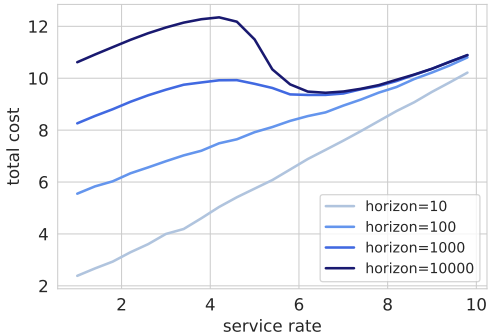


Figure 3: Multi-fidelity output for a steady-state queuing control problem, plotting objective value (total cost) versus the decision variable (service rate) at a collection of time horizons. Our goal is to choose the service rate to minimize the total cost at the longest time horizon pictured. Shorter time horizons offer approximations to the objective with less computational effort. Computational effort is approximately proportional to the time horizon.

The multi-fidelity BO problem is formalized by assuming that the objective function is given by $f(x) = h(x, w^{\text{tf}})$, where $h : \mathbb{X} \times \mathbb{W} \to \mathbb{R}$ is a black-box function,$\mathbb{W}$ is the space of fidelity-control parameters, and w^{tf} is the *target fidelity*. We also assume that evaluating h at (x, w) has a cost $c(x, w)$. When the function $c : \mathbb{X} \times \mathbb{W} \to (0, \infty)$ is unknown (i.e., a black box), a common choice is to model $\log c$ using a GP.

4.1 Predictive Model

To effectively leverage the information provided by cheaper approximations of the objective function, it is key to use a predictive model able to capture the correlation between these approximations and the

objective function itself. As an illustrative example, below we discuss the model proposed by Poloczek et al. (2017), which is as a special case of the *semi-parametric latent factor model* (Alvarez et al. 2012).

Suppose that $\mathbb{W}$ is finite, say $\mathbb{W} = \{w_1, \ldots, w_k\}$. Making a slight abuse of notation, let $h(x) = (h(x, w_j) : j = 1, \ldots, k)$, and assume that h_k is the objective to optimize (i.e., $w^{\text{tf}} = w_k$). A sensible modeling choice is then to assume that h_k and $h_j - h_k$ for $j = 1, \ldots, k-1$ are drawn from independent GPs, which implicitly assumes that the bias of lower fidelity approximations and the objective function are all independent. This translates in a multi-output GP model over h with prior mean and covariance functions of the form

$$\mu_0(x) = (\nu_0^{(1)}(x) + \mu_0^{(k)}(x), \ldots, \nu_0^{(k-1)}(x) + \mu_0^{(k)}(x),\ \mu_0^{(k)}(x))$$

$$K_0(x, x') = \left(\mathbb{I}\{j = j' \text{ and } j \neq k\} \Xi_0^{(j)}(x, x') + K_0^{(k)}(x, x') \right)_{j, j' = 1, \ldots, k},$$

where $\nu_0^{(j)}$ and $\Xi_0^{(j)}$ are the prior mean and covariance functions of $h_j - h_k$, for $j = 1, \ldots, k-1$, and $\mu_0^{(k)}$ and $K_0^{(k)}$ are the prior mean and covariance functions of h_k.

4.2 Acquisition Functions

Since the EI acquisition function is defined via a quantity that can only be measured when the evaluation of the objective function is performed fully (namely, the improvement that would be obtained from such evaluation), it is difficult to extend it in a principled way to multi-fidelity evaluations. Despite this difficulty, previous work (e.g., Huang et al. 2006) has supported multi-fidelity extensions of EI via heuristic arguments. However, other acquisition functions can be naturally extended to this setting. Wu et al. (2020), for example, extends the KG acquisition function to measure incremental reward per unit cost,

$$\text{KG}_n(x, w) = \mathbb{E}_n \left[\frac{\mu_{n+1}^* - \mu_n^*}{c(x_{n+1}, w_{n+1})} \mid (x_{n+1}, w_{n+1}) = (x, w) \right],$$

where $\mu_m^* = \max_{x \in \mathbb{X}} \mathbb{E}_m[f(x)] = \max_{x \in \mathbb{X}} \mu_m(x, w^{\text{tf}})$ for $m = n, n+1$.

Other acquisition functions have been extended to the multi-fidelity evaluations setting. Swersky et al. (2013), for example, extends the ES acquisition function by dividing the expected information gain that would be obtained from evaluating a pair (x_{n+1}, w_{n+1}) by its cost $c(x_{n+1}, w_{n+1})$. An analogous extension of the max-value entropy search acquisition function was proposed by Takeno et al. (2020).

4.3 Other Related Work

Multi-fidelity optimization, including both Bayesian approaches and those that use non-Bayesian surrogate models, is the longest thread of literature of those grey-box problem classes considered in this tutorial. Early work in this area focused on problems in engineering design, especially those where PDE mesh size could be controlled, and includes Huang et al. (2006) and Forrester et al. (2007). This work also owes a great deal to non-optimization-focused work that developed surrogate models using multi-fidelity computer codes (Kennedy and O'Hagan 2000). More recently, while interest in the use of multi-fidelity methods for engineering design (especially in aerospace) has been sustained (see, e.g., Peherstorfer et al. 2018), interest has also grown in its use in machine learning, especially for hyperparameter optimization (Wu et al. 2020; Takeno et al. 2020). This remains an exciting area with great potential for value delivered through intelligent application of existing methods and the development of new methods and theory, especially aligned with the challenges of novel application domains.

5 BAYESIAN OPTIMIZATION WITH OBJECTIVE CONSTITUENT EVALUATIONS

Evaluation of just some of the objective function's constituents is possible in both BO of composite objective functions and multi-fidelity BO, and allows learning from these partial evaluations (and optionally later

continuing paused evaluations). This strategy is already implicit in our discussion of multi-fidelity BO. For example, consider multi-fidelity BO approaches for hyperparameter tuning of DNNs, where lower fidelities are obtained by using fewer training iterations to optimize the weights of the DNN. An evaluation can be paused at a small number of iterations, the information obtained thus far can be incorporated into the predictive model, and then it can be continued later if desired. We thus focus here on describing objective constituent evaluations in the context of composite objective functions, where only some constituents of the multivariate output combined to produce the objective are evaluated at each iteration.

To motivate constituent evaluations for composite objective functions, suppose we seek to find a configuration of bicycle docks within a city to maximize the average number of trips taken in a bike-sharing system (Freund et al. 2019). A simulator takes historical demand (in the form of request times and desired origin and destination) and simulates bike availability and the number of trips taken. The objective is $\sum_j h(x,w_j)$, where x is a candidate configuration of bike docks, j indexes days from which historical demands are taken, w_j contains historical context about the day (e.g., the amount of rainfall), and $h(x,w_j)$ is the number of trips taken on that day.

This problem can be tackled using BO of composite objective functions, where $h(x) = (h(x,w_j) : j)$ is the inner function and the outer function $g(h(x)) = \sum_j h_j(x)$ is the sum of this vector. One could then apply approaches from the previous section. There is an opportunity, however, to improve efficiency further. One can easily evaluate just one term in this sum, $h(x,w_j)$. This would be much faster than evaluating the entire sum and might give information nearly as useful for optimization. An algorithm could selectively evaluate just one term at a time in this sum (just one constituent, $h_j(x) = h(x,w_j)$) to identify promising values of x before spending the effort to evaluate the whole sum. This saves substantial time if, e.g., the sum is over 100 terms and each term takes one hour to compute.

By placing a GP prior over h that models its dependence on both x and w, we can perform inference over $g(h(x))$. We can then use an acquisition function to value an additional evaluation of h at a single pair (x,w), toward the goal of solving $\max_x g(h(x))$. This is the approach taken for outer functions g that are sums or integrals in Williams et al. (2000), Xie et al. (2012), and Toscano-Palmerin and Frazier (2018); risk measures in Cakmak et al. (2020), and Nguyen et al. (2021); and arbitrary functionals of a control-dependent PDF modeled using a spatial logistic GP in Gautier et al. (2021).

5.1 Predictive Model

As argued earlier, constituent evaluations (recall, one objective constituent is $h(\cdot,w)$ for a single w) can often provide information nearly as good as complete evaluations of the objective function at a much lower cost. This is particularly true when the constituents are highly correlated, and, therefore, using a statistical model capable of capturing such correlation is paramount.

When $h(x,w)$ is continuous in both x and w, it is convenient to simply model h with a GP prior that uses a standard distance-based covariance function $K : (\mathbb{X}\times\mathbb{W})\times(\mathbb{X}\times\mathbb{W}) \to \mathbb{R}$. This is true even if the objective is a sum that only depends on $h(x,w)$ at finitely many values of w.

When $h(x,w)$ lacks smoothness in w, an effective choice is to use a multi-output GP model with *intrinsic corregionalization*, whose covariance function, $K : \mathbb{X}\times\mathbb{X} \to \mathbb{R}^{k\times k}$, is of the form $K(x,x') = \Sigma K'(x,x')$, where $\Sigma \in \mathbb{R}^{k\times k}$ is a positive definite matrix, and $K' : \mathbb{X}\times\mathbb{X} \to \mathbb{R}$ is a covariance function. The matrix Σ can be estimated with the other hyperparameters of the GP model. Other popular multi-output GP models include the semi-parametric latent factor model, discussed before, and the *linear corregionalization model*. See Alvarez et al. (2012) for details.

Regardless of the choice of the covariance function, a key property that has been leveraged to develop efficient acquisition functions when the objective function is a sum or integral of the individual constituents is that the implied posterior distribution on the objective distribution is again a GP. More concretely, if h is modeled using a multi-output GP with posterior mean function $\mu_n : \mathbb{X} \to \mathbb{R}^k$ and covariance function $K_n : \mathbb{X}\times\mathbb{X} \to \mathbb{R}^{k\times k}$, then, using elementary properties of the multivariate normal distribution, it can be

shown that, for any fixed $p \in \mathbb{R}^k$, the implied posterior distribution on $f = p^\top h$ is a GP with mean function $p^\top \mu_n$ and covariance function $p^\top K_n p$. A similar statement holds true for integrals (O'Hagan 1991).

5.2 Acquisition Functions

As in the case of multi-fidelity evaluations, it is unclear how to extend EI to handle constituent evaluations in a composite objective framework. However, several attempts have been made in the literature. For example, when f is a sum or integral of the components of h, Williams et al. (2000) proposes to select the next point to evaluate, x_{n+1}, by maximizing a variant of the classical EI acquisition function: it is computed with respect to the implied GP posterior distribution on f, taking f_n^* to be the best $f(x)$ across all previously evaluated x. All previously evaluated x are included, even those for which $f(x)$ is not well-estimated because $h(x,w)$ has been evaluated for only one w. Having chosen x_{n+1}, the w_{n+1} determining the constituent $h(x_{n+1}, w_{n+1})$ to evaluate is chosen to minimize the posterior variance of $f(x_{n+1})$ after this evaluation is complete. As argued by Toscano-Palmerin and Frazier (2018), however, this policy is unsatisfactory. First, x_{n+1} is chosen without considering w_{n+1}, while the best choice should be made jointly across x and w: when observing $h(x,w)$ at a single w produces a significant variance reduction, this should increase our willingness to evaluate at this x. Second, in discrete problems, once $h(x,w)$ has been evaluated once at each x for at least one w, this variant of EI becomes identically zero, producing no guidance and potentially leading to a lack of consistency.

While EI does not have a natural extension allowing for constituent evaluations, other acquisition functions do. For example, again for the case where f is a sum or integral of the components of h, Toscano-Palmerin and Frazier (2018) proposes an extension of the KG acquisition function, derived following the same decision-theoretic approach. More concretely, this acquisition function is defined as

$$\mathrm{KG}_n(x,w) = \mathbb{E}_n\left[\mu_{n+1}^* - \mu_n^* \mid (x_{n+1}, w_{n+1}) = (x,w)\right],$$

where $\mu_m^* = \max_{x\in\mathbb{X}} \mathbb{E}_m[f(x)] = \max_{x\in\mathbb{X}} \sum_{j=1}^k \mu_m(x, w_j)$, for $m = n, n+1$. Importantly, in contrast with Williams et al. (2000), this acquisition function chooses the pair (x,w), representing the input and constituent to be evaluated, jointly in a one-step optimal way. In numerical experiments, this acquisition function delivers significantly superior performance to the approach proposed by Williams et al. (2000) and other approaches that select x and w separately.

The above acquisition function can be maximized by virtually unmodified versions of the approaches used to maximize KG for standard BO discussed earlier. This is due to the linear nature of the transformation mapping h to f, causing the GP distribution on h to imply a GP distribution on f. However, in many settings, the transformation that maps h onto f is non-linear (see §5.3). For such settings, approaches similar to those described in §3 can be employed to efficiently maximize MC-based acquisition functions.

5.3 Other Related Work

As mentioned earlier, a related line of work studies problems analogous to those discussed in this section, where the transformation that maps h onto f is non-linear. Such transformations often arise when seeking solutions x that poses some form of *risk aversion* to variations in w. Concrete examples include optimization of worst-case performance (Marzat et al. 2013; Bogunovic et al. 2018); distributionally-robust optimization (Kirschner et al. 2020); and optimization of risk measures (Cakmak et al. 2020; Nguyen et al. 2021).

6 CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

Grey-box BO trades generality for performance gains that are, in some cases, quite dramatic. Since grey-box BO is a young research area and grey-box BO methods are customized to a problem class, or even to an individual problem, there are many questions that remain open. We believe that the following research questions are ones that are particularly interesting and likely to bear fruit over the coming years.

- **Applications**: Applying grey-box BO in important and novel application domains can provide significant benefits while at the same time inspiring new methodological questions. Calibration of simulators and inverse reinforcement learning are exciting areas where composite objective functions can clearly provide benefits. In addition, atomistic simulation of chemical systems (Gillespie 2007) is computationally intensive, important, and likely amenable to grey-box BO.
- **Methods for new problem classes**: Working on novel applications is likely to identify other broad classes of grey-box structures and novel methods that productively leverage such structures.
- **Many constituents**: Existing BO methods for objective constituent evaluations become slow when there are many constituents. There is an opportunity to add value by developing more computationally efficient methods, e.g., by leveraging a known correlation structure between constituents that allows for faster predictive computations (Maddox et al. 2021), or by intelligently selecting which constituents to model individually and which to aggregate.
- **Non-myopic BO**: Non-myopic BO, which has shown promise in standard BO (Jiang et al. 2020), is likely to unlock even more value in grey-box BO. In grey-box BO, constituent evaluations do not provide a direct myopic benefit, and it is thus important to use knowledge-gradient or other methods that look further ahead than expected improvement to derive value. Looking further ahead will allow a method to understand when several pieces of information together can provide much more value than any one piece of information individually.
- **Theoretical understanding of grey-box BO methods**: There is much to be done deepening our theoretical understanding of grey-box BO methods by deriving regret bounds, convergence rates, and understanding how problem structure determines the value derived from a grey-box approach. For example, when does leveraging composite objective structure perform better than standard BO?

REFERENCES

Alvarez, M. A., L. Rosasco, and N. D. Lawrence. 2012. "Kernels for Vector-Valued Functions: A Review". *Foundations and Trends® in Machine Learning* 4(3):195–266.

Astudillo, R., and P. Frazier. 2019. "Bayesian Optimization of Composite Functions". In *Proceedings of the 36th International Conference on Machine Learning*. June 9th-15th, Long Beach, California, USA, 354–363.

Astudillo, R., and P. Frazier. 2021. "Bayesian Optimization of Function Networks". In *Advances in Neural Information Processing Systems*. Red Hook, New York: Curran Associates, Inc.

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

Beykal, B., F. Boukouvala, C. A. Floudas, N. Sorek, H. Zalavadia, and E. Gildin. 2018. "Global Optimization of Grey-Box Computational Systems Using Surrogate Functions and Application to Highly Constrained Oil-Field Operations". *Computers and Chemical Engineering* 114:99–110.

Bogunovic, I., J. Scarlett, S. Jegelka, and V. Cevher. 2018. "Adversarially Robust Optimization with Gaussian Processes". In *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 5760–5770. Red Hook, New York: Curran Associates, Inc.

Bohlin, T. P. 2006. *Practical Grey-Box Process Identification: Theory and Applications*. London: Springer-Verlag.

Burke, J. V., and M. C. Ferris. 1995. "A Gauss-Newton Method for Convex Composite Optimization". *Mathematical Programming* 71(2):179–194.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu. 1995. "A Limited Memory Algorithm for Bound Constrained Optimization". *SIAM Journal on Scientific Computing* 16(5):1190–1208.

Cakmak, S., R. Astudillo, P. Frazier, and E. Zhou. 2020. "Bayesian Optimization of Risk Measures". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, and M. F. Balcan, 20130–20141. Red Hook, New York: Curran Associates, Inc.

Calandra, R., A. Seyfarth, J. Peters, and M. P. Deisenroth. 2016. "Bayesian Optimization for Learning Gaits Under Uncertainty: An Experimental Comparison on a Dynamic Bipedal Walker". *Annals of Mathematics and Artificial Intelligence* 76(1):5–23.

Conn, A. R., K. Scheinberg, and L. N. Vicente. 2009. *Introduction to Derivative-Free Optimization*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Drusvyatskiy, D., and C. Paquette. 2019. "Efficiency of Minimizing Compositions of Convex Functions and Smooth Maps". *Mathematical Programming* 178(1):503–558.

Forrester, A. I., A. Sóbester, and A. J. Keane. 2007. "Multi-fidelity Optimization via Surrogate Modelling". *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 463(2088):3251–3269.

Frazier, P., W. Powell, and S. Dayanik. 2009. "The Knowledge-Gradient Policy for Correlated Normal Beliefs". *INFORMS Journal on Computing* 21(4):599–613.

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

Frazier, P. I., W. B. Powell, and S. Dayanik. 2008. "A Knowledge-Gradient Policy for Sequential Information Collection". *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Freund, D., S. G. Henderson, E. O'Mahony, and D. B. Shmoys. 2019. "Analytics and Bikes: Riding Tandem with Motivate to Improve Mobility". *INFORMS Journal on Applied Analytics* 49(5):310–323.

Gardner, J. R., C. Guo, K. Q. Weinberger, R. Garnett, and R. Grosse. 2017. "Discovering and exploiting additive structure for Bayesian optimization". In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 1311–1319. April 20$^{\text{th}}$-22$^{\text{nd}}$, Fort Lauderdale, Florida, USA, 1311–1319.

Gautier, A., D. Ginsbourger, and G. Pirot. 2021. "Goal-Oriented Adaptive Sampling under Random Field Modelling of Response Probability Distributions". *arXiv preprint arXiv:2102.07612*.

Gillespie, D. T. 2007. "Stochastic Simulation of Chemical Kinetics". *Annual Review of Physical Chemistry* 58:35–55.

Griffiths, R. R., and J. M. Hernández-Lobato. 2020. "Constrained Bayesian Optimization for Automatic Chemical Design Using Variational Autoencoders". *Chemical Science* 11(2):577–586.

Hennig, P., and C. J. Schuler. 2012. "Entropy Search for Information-Efficient Global Optimization". *Journal of Machine Learning Research* 13(6):1809–1837.

Hernández-Lobato, J. M., M. W. Hoffman, and Z. Ghahramani. 2014. "Predictive Entropy Search for Efficient Global Optimization of Black-box Functions". In *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, 918–926. Red Hook, New York: Curran Associates, Inc.

Huang, D., T. T. Allen, W. I. Notz, and R. A. Miller. 2006. "Sequential Kriging Optimization Using Multiple-Fidelity Evaluations". *Structural and Multidisciplinary Optimization* 32(5):369–382.

Hutter, F., H. H. Hoos, and K. Leyton-Brown. 2011. "Sequential Model-Based Optimization for General Algorithm Configuration". In *Learning and Intelligent Optimization*, edited by C. A. C. Coello, 507–523. Heidelberg: Springer.

Jauch, M., and V. Peña. 2016. "Bayesian Optimization with Shape Constraints". *arXiv preprint arXiv:1612.08915*.

Jiang, S., H. Chai, J. Gonzalez, and R. Garnett. 2020. "BINOCULARS for Efficient, Nonmyopic Sequential Experimental Design". In *Proceedings of the 37th International Conference on Machine Learning*. July 13$^{\text{th}}$-18$^{\text{th}}$, Virtual, 4794–4803.

Jiang, S., D. Jiang, M. Balandat, B. Karrer, J. Gardner, and R. Garnett. 2020. "Efficient Nonmyopic Bayesian Optimization via One-Shot Multi-Step Trees". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, and M. F. Balcan, 18039–18049. Red Hook, New York: Curran Associates, Inc.

Jones, D. R., M. Schonlau, and W. J. Welch. 1998. "Efficient Global Optimization of Expensive Black-Box Functions". *Journal of Global Optimization* 13(4):455–492.

Kennedy, M. C., and A. O'Hagan. 2000. "Predicting the Output from a Complex Computer Code when Fast Approximations are Available". *Biometrika* 87(1):1–13.

Kirschner, J., I. Bogunovic, S. Jegelka, and A. Krause. 2020. "Distributionally Robust Bayesian Optimization". In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*. June 3$^{\text{rd}}$-5$^{\text{th}}$, Palermo, Sicily, Italy, 2174–2184.

Kushner, H. J. 1964. "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise". *Journal of Basic Engineering* 86(1):97–106.

Letham, B., R. Calandra, A. Rai, and E. Bakshy. 2020. "Re-Examining Linear Embeddings for High-Dimensional Bayesian Optimization". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 1546–1558. Red Hook, New York: Curran Associates, Inc.

Maddox, W. J., M. Balandat, A. G. Wilson, and E. Bakshy. 2021. "Bayesian Optimization with High-Dimensional Outputs". *arXiv preprint arXiv:2106.12997*.

Marque-Pucheu, S., G. Perrin, and J. Garnier. 2019. "Efficient Sequential Experimental Design for Surrogate Modeling of Nested Codes". *ESAIM: PS* 23:245–270.

Marzat, J., E. Walter, and H. Piet-Lahanier. 2013. "Worst-Case Global Optimization of Black-Box Functions Through Kriging and Relaxation". *Journal of Global Optimization* 55(4):707–727.

Močkus, J. 1975. "On Bayesian Methods for Seeking the Extremum". In *Optimization Techniques IFIP Technical Conference*, edited by G. I. Marchuk, Volume 27 of *Lecture Notes in Computer Science*, 400–404: Springer-Verlag.

Nguyen, Q. P., Z. Dai, B. K. H. Low, and P. Jaillet. 2021. "Value-at-Risk Optimization with Gaussian Processes". In *Proceedings of the 38th International Conference on Machine Learning*. July 18$^{\text{th}}$-24$^{\text{th}}$, Virtual, 8063–8072.

O'Hagan, A. 1991. "Bayes–Hermite Quadrature". *Journal of Statistical Planning and Inference* 29(3):245–260.

Pearce, M., and J. Branke. 2017. "Bayesian Simulation Optimization with Input Uncertainty". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, G. W. N. Mustafee, and E. Page, 2268–2278. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Peherstorfer, B., K. Willcox, and M. Gunzburger. 2018. "Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization". *SIAM Review* 60(3):550–591.

Poloczek, M., J. Wang, and P. I. Frazier. 2017. "Multi-Information Source Optimization". In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4289–4299. Red Hook, New York: Curran Associates, Inc.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: MIT Press.

Scott, W., P. Frazier, and W. Powell. 2011. "The Correlated Knowledge Gradient for Simulation Optimization of Continuous Parameters Using Gaussian Process Regression". *SIAM Journal on Optimization* 21(3):996–1026.

Shapiro, A. 2003. "On a Class of Nonsmooth Composite Functions". *Mathematics of Operations Research* 28(4):677–692.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms". In *Advances in Neural Information Processing Systems*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 2951–2959. Red Hook, New York: Curran Associates, Inc.

Snoek, J., O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams. 2015. "Scalable Bayesian Optimization Using Deep Neural Networks". In *Proceedings of the 32nd International Conference on Machine Learning*. July 7$^{\text{th}}$-9$^{\text{th}}$, Lille, France, 2171–2180.

Srinivas, N., A. Krause, S. M. Kakade, and M. W. Seeger. 2012. "Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting". *IEEE Transactions on Information Theory* 58(5):3250–3265.

Swersky, K., J. Snoek, and R. P. Adams. 2013. "Multi-Task Bayesian Optimization". In *Advances in Neural Information Processing Systems*, edited by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, 2004–2012. Red Hook, New York: Curran Associates, Inc.

Takeno, S., H. Fukuoka, Y. Tsukada, T. Koyama, M. Shiga, I. Takeuchi, and M. Karasuyama. 2020. "Multi-Fidelity Bayesian Optimization with Max-value Entropy Search and its Parallelization". In *Proceedings of the 37th International Conference on Machine Learning*. July 13$^{\text{th}}$-18$^{\text{th}}$, Virtual, 9334–9345.

Toscano-Palmerin, S., and P. I. Frazier. 2018. "Bayesian Optimization with Expensive Integrands". *arXiv preprint arXiv:1803.08661*.

Tulleken, H. J. A. F. 1993. "Grey-box Modelling and Identification Using Physical Knowledge and Bayesian Techniques". *Automatica* 29(2):285–308.

Turner, R., D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon. 2021. "Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020". *arXiv preprint arXiv:2104.10201*.

Uhrenholt, A. K., and B. S. Jensen. 2019. "Efficient Bayesian Optimization for Target Vector Estimation". In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. April 16$^{\text{th}}$-18$^{\text{th}}$, Naha, Okinawa, Japan, 2661–2670.

Wang, J., S. C. Clark, E. Liu, and P. I. Frazier. 2020. "Parallel Bayesian Global Optimization of Expensive Functions". *Operations Research* 68(6):1850–1865.

Wang, Z., and S. Jegelka. 2017. "Max-value Entropy Search for Efficient Bayesian Optimization". In *Proceedings of the 34th International Conference on Machine Learning*. August 6$^{\text{th}}$-11$^{\text{th}}$, Sydney, Australia, 3627–3635.

Wild, S. M. 2017. "POUNDERS in TAO: Solving Derivative-Free Nonlinear Least-Squares Problems with POUNDERS". In *Advances and Trends in Optimization with Engineering Applications*, edited by T. Terlaky, M. F. Anjos, and S. Ahmed, 529–539. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Williams, B. J., T. J. Santner, and W. I. Notz. 2000. "Sequential Design of Computer Experiments to Minimize Integrated Response Functions". *Statistica Sinica* 10(4):1133–1152.

Wilson, J., F. Hutter, and M. Deisenroth. 2018. "Maximizing Acquisition Functions for Bayesian Optimization". In *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 9884–9895. Red Hook, New York: Curran Associates, Inc.

Wu, J., and P. Frazier. 2016. "The Parallel Knowledge Gradient Method for Batch Bayesian Optimization". In *Advances in Neural Information Processing Systems*, edited by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, 3134–3142. Red Hook, New York: Curran Associates, Inc.

Wu, J., M. Poloczek, A. G. Wilson, and P. Frazier. 2017. "Bayesian Optimization with Gradients". In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 5268–5279. Red Hook, New York: Curran Associates, Inc.

Wu, J., S. Toscano-Palmerin, P. I. Frazier, and A. G. Wilson. 2020. "Practical Multi-Fidelity Bayesian Optimization for Hyperparameter Tuning". In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. July 22$^{\text{nd}}$-25$^{\text{th}}$, Tel Aviv, Israel, 788–798.

Xie, J., P. I. Frazier, S. Sankaran, A. Marsden, and S. Elmohamed. 2012. "Optimization of Computationally Expensive Simulations with Gaussian Processes and Parameter Uncertainty: Application to Cardiovascular Surgery". In *50th Annual*

Allerton Conference on Communication, Control, and Computing. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhilinskas, A. G. 1975. "Single-Step Bayesian Search Method for an Extremum of Functions of a Single Variable". *Cybernetics* 11(1):160–166.