Bayesian Optimization of Composite Functions

Raul Astudillo [1] **Peter I. Frazier** [1 2]

Abstract

We consider optimization of *composite* objective functions, i.e., of the form $f(x) = g(h(x))$, where h is a black-box derivative-free expensive-to-evaluate function with vector-valued outputs, and g is a cheap-to-evaluate real-valued function. While these problems can be solved with standard Bayesian optimization, we propose a novel approach that exploits the composite structure of the objective function to substantially improve sampling efficiency. Our approach models h using a multi-output Gaussian process and chooses where to sample using the expected improvement evaluated on the implied non-Gaussian posterior on f, which we call expected improvement for composite functions (EI-CF). Although EI-CF cannot be computed in closed form, we provide a novel stochastic gradient estimator that allows its efficient maximization. We also show that our approach is asymptotically consistent, i.e., that it recovers a globally optimal solution as sampling effort grows to infinity, generalizing previous convergence results for classical expected improvement. Numerical experiments show that our approach dramatically outperforms standard Bayesian optimization benchmarks, reducing simple regret by several orders of magnitude.

1. Introduction

We consider optimization of *composite* objective functions, i.e., of the form $f(x) = g(h(x))$, where h is a black-box expensive-to-evaluate vector-valued function, and g is a simple real-valued function that can be cheaply evaluated. We assume evaluations are noise-free. These problems arise, for example, in calibration of simulators to real-world data (Vrugt et al., 2001; Cullick et al., 2006; Schultz & Sokolov, 2018); in materials and drug design (Kapetanovic, 2008; Frazier & Wang, 2016) when seeking to design a compound with a particular set of physical or chemical properties; when finding maximum *a posteriori* estimators with expensive-to-evaluate likelihoods (Bliznyuk et al., 2008); and in constrained optimization (Gardner et al., 2014; Hernández-Lobato et al., 2016) when seeking to maximize one expensive-to-evaluate quantity subject to constraints on others. (See Section 2 for a more detailed description of these problems.)

One may ignore the composite structure of the objective and solve such problems using Bayesian optimization (BO) (Brochu et al., 2010), which has been shown to perform well compared with other general-purpose optimization methods for black-box derivative-free expensive-to-evaluate objectives (Snoek et al., 2012). In the standard BO approach, one would build a Gaussian process (GP) prior over f based on past observations of $f(x)$, and then choose points at which to evaluate f by maximizing an acquisition function computed from the posterior. This approach would not use observations of $h(x)$ or knowledge of g.

In this paper, we describe a novel BO approach that leverages the structure of composite objectives to optimize them more efficiently. This approach builds a multi-output GP on h, and uses the expected improvement (Jones et al., 1998) under the implied statistical model on f as its acquisition function. This implied statistical model is typically non-Gaussian when g is non-linear. We refer to the resulting acquisition function as expected improvement for composite functions (EI-CF) to distinguish it from the classical expected improvement (EI) acquisition function evaluated on a GP posterior on f.

Intuitively, the above approach can substantially outperform standard BO when $h(x)$ contains information relevant to optimization that is not available from observations of $f(x)$ alone. As one example, suppose x and $h(x)$ are both one-dimensional and $g(y) = y^2$. If h is continuous, $h(0) < 0$, and $h(1) > 0$, then our approach knows that there is a global minimum in the interval $(0, 1)$, while the standard approach does not. This informational benefit is compounded further when h is vector-valued.

While EI-CF is simply the expected improvement under a different statistical model, unlike the classical EI acqui-

[1]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA [2]Uber, San Francisco, CA, USA. Correspondence to: Raul Astudillo <ra598@cornell.edu>, Peter I. Frazier <pf98@cornell.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

sition function, it lacks a closed-form analytic expression and must be evaluated through simulation. We provide a simulation-based method for computing unbiased estimators of the gradient of the EI-CF acquisition function, which we use within multi-start stochastic gradient ascent to allow efficient maximization. We also show that optimizing using EI-CF is asymptotically consistent under suitable regularity conditions, in the sense that the best point found converges to the global maximum of f as the number of samples grows to infinity.

In numerical experiments comparing with standard BO benchmarks, EI-CF provides immediate regret that is several orders of magnitude smaller, and reaches their final solution quality using less than 1/4 the function evaluations.

2. Related Work

2.1. Related Methodological Literature

We work within the Bayesian optimization framework, whose origins date back to the seminal work of Močkus (1975), and which has recently become popular due to its success in hyperparameter optimization of machine learning algorithms (Snoek et al., 2012; Swersky et al., 2013).

Optimizing composite functions has been studied in first- and second-order optimization (Shapiro, 2003; Drusvyatskiy & Paquette, 2016). This literature differs from our paper in that it assumes derivatives are available, and also often assumes convexity and that evaluations are inexpensive. In this setting, leveraging the structure of the objective has been found to improve performance, just as we find here in the setting of derivative-free optimization. However, to the best of our knowledge, ours is the first paper to study composite objective functions within the BO framework and also the first within the more general context of optimization of black-box derivative-free expensive-to-evaluate functions.

Our work is related to Marque-Pucheu et al. (2017), which proposes a framework for efficient sequential experimental design for GP-based modeling of nested computer codes. In contrast with our work, that work's goal is not to optimize a composite function, but instead to learn it as accurately as possible within a limited evaluation budget. A predictive variance minimization sampling policy is proposed and methods for efficient computation are provided. Moreover, it is assumed that both the inner (h) and outer (g) functions are real-valued and expensive-to-evaluate black-box functions, while our method uses the ease-of-evaluation of the outer function for substantial benefit.

Our work is also similar in spirit to Overstall & Woods (2013), which proposes to model an expensive-to-evaluate vector of parameters of a posterior probability density function using a multi-output GP instead of the function directly using a single-output GP. The surrogate model is then used to perform Bayesian inference.

Constrained optimization is a special case of optimization of a composite objective. To see this, take h_1 to be the objective to be maximized and take h_i, for $i > 1$, to be the constraints, all of which are constrained to be non-negative without loss of generality. Then, we recover the original constrained optimization problem by setting

$$g(y) = \begin{cases} y_1, \text{ if } y_i \geq 0 \text{ for all } i > 1, \\ -\infty, \text{ otherwise.} \end{cases}$$

Moreover, when specialized to this particular setting, our EI-CF acquisition function is equivalent to the expected improvement for constrained optimization as proposed by Schonlau et al. (1998) and Gardner et al. (2014).

Within the constrained BO literature, our work also shares several methodological similarities with Picheny et al. (2016), which considers an augmented Lagrangian and models its components as GPs instead of it directly as a GP. As in our work, the expected improvement under this statistical model is used as acquisition function. Moreover, it is shown that this approach outperforms the standard BO approach.

Our method for optimizing the EI-CF acquisition function uses an unbiased estimator of the gradient of EI-CF within a multistart stochastic gradient ascent framework. This technique is structurally similar to methods developed for optimizing acquisition functions in other BO settings without composite objectives, including the parallel expected improvement (Wang et al., 2016) and the parallel knowledge-gradient (Wu & Frazier, 2016).

2.2. Related Application Literature

Optimization of composite black-box derivative-free expensive-to-evaluate functions arises in a number of application settings in the literature, though this literature does not leverage the composite structure of the objective to optimize it more efficiently.

In materials design, it arises when the objective is the combination of multiple material properties via a *performance index* (Ashby & Cebon, 1993), e.g., the specific stiffness, which is the ratio of Young's modulus and the density, or *normalization* (Jahan & Edwards, 2015). Here, $h(x)$ is the set of material properties that results from a particular chemical composition or set of processing conditions x, and g is given by the performance index or normalization method used. Evaluating the material properties $h(x)$ that result from a materials design typically requires running expensive physical or computational experiments that do not provide derivative information, for which BO is appropriate (Kapetanovic, 2008; Ueno et al., 2016; Ju et al., 2017; Griffiths & Hernández-Lobato, 2017).

Optimization of composite functions also arises in calibration of expensive black-box simulators (Vrugt et al., 2001; Cullick et al., 2006; Schultz & Sokolov, 2018), where the goal is to find input parameters, x, to the simulator such that its vector-valued output, $h(x)$, most closely matches a vector data observed in the real world, y_{obs}. Here, the objective to be minimized is $g(h(x)) = ||h(x) - y_{\text{obs}}||$, where $||\cdot||$ is often the L_1 norm, L_2 norm, or some monotonic transformation of the likelihood of observation errors.

If one has a prior probability density p on x, and the log-likelihood of some real-world observation error, ϵ, is proportional to $||\epsilon||$ (as it would be, for example, with independent normally distributed errors taking $||\cdot||$ to be the L_2 norm), then, finding the maximum *a posteriori* estimator of x (Bliznyuk et al., 2008) is an optimization problem with a composite objective: the log-posterior is equal to the sum of a constant and $g(h(x)) = -\beta||h(x) - y_{\text{obs}}||^2 + \log(p(x))$ (In this example, g is actually a function of both $h(x)$ and x. Our framework extends easily to this setting as long as g remains a simple cheap-to-evaluate function.).

3. Problem Description and Standard Approach

As described above, we consider optimization of objectives of the form $f(x) = g(h(x))$, where $h : \mathcal{X} \to \mathbb{R}^m$ is a black-box expensive-to-evaluate continuous function whose evaluations do not provide derivatives, $g : \mathbb{R}^m \to \mathbb{R}$ is a function that can be cheaply evaluated, and $\mathcal{X} \subset \mathbb{R}^d$. As is common in BO, we assume that d is not too large (< 20) and that projections onto $\mathcal{X}$ can be efficiently computed. We also place the technical condition that $\mathbb{E}\left[|g(Z)|\right] < \infty$, where Z is an m-variate standard normal random vector. The problem to be solved is

$$\max_{x \in \mathcal{X}} g(h(x)). \tag{1}$$

As discussed before, one can solve problem (1) by applying standard BO to the objective function, $f := g \circ h$. This approach models f as drawn from a GP prior probability distribution. Then, iteratively, indexed by n, this approach would choose the point $x_n \in \mathcal{X}$ at which to evaluate f next by optimizing an acquisition function, such as the EI acquisition function (Močkus, 1975; Jones et al., 1998). This acquisition function would be calculated from the posterior distribution given $\{(x_i, f(x_i))\}_{i=1}^n$, which is itself a GP, and would quantify the value of an evaluation at a particular point. Although $h(x)$ would be observed as part of this standard approach, these evaluations would be ignored when calculating the posterior distribution and acquisition function.

4. Our Approach

We now describe our approach, which like the standard BO approach is comprised of a statistical model and an acquisition function. Unlike standard BO, however, our approach leverages the additional information in evaluations of h, along with knowledge of g. We argue below and demonstrate in our numerical experiments that this additional information can substantially reduce the number of evaluations required to find good approximate global optima.

Briefly, our statistical model is a multi-output Gaussian process on h (Alvarez et al., 2012) (Section 4.1), and our acquisition function, EI-CF, is the expected improvement under this statistical model (Section 4.2). This acquisition function, unfortunately, cannot be computed in closed form for most functions g. In Section 4.3, under mild regularity conditions, we provide a technique for efficiently maximizing EI-CF. We also provide a theoretical analysis showing that EI-CF is asymptotically consistent (Section 4.4). Finally, we conclude this section by discussing the computational complexity of our approach (Section 4.5).

4.1. Statistical Model

We model h as drawn from a multi-output GP distribution (Alvarez et al., 2012), $\mathcal{GP}(\mu, K)$, where $\mu : \mathcal{X} \to \mathbb{R}^m$ is the mean function, $K : \mathcal{X} \times \mathcal{X} \to \mathcal{S}_{++}^m$ is the covariance function, and $\mathcal{S}_{++}^m$ is the cone of positive definite matrices. Analogously to the single-output case, after observing n evaluations of h, $h(x_1), \ldots, h(x_n)$, the posterior distribution on h is again a multi-output GP, $\mathcal{GP}(\mu_n, K_n)$, where μ_n and K_n can be computed in closed form in terms of μ and K (Liu et al., 2018).

4.2. Expected Improvement for Composite Functions

We define the expected improvement for composite functions analogously to the classical expected improvement, but where our posterior on $f(x)$ is given by the composition of g and the normally distributed posterior distribution on $h(x)$:

$$\text{EI-CF}_n(x) = \mathbb{E}_n\left[\{g(h(x)) - f_n^*\}^+\right], \tag{2}$$

where $f_n^* = \max_{i=1,\ldots,n} f(x_i)$ is the maximum value across the points that have been evaluated so far, $x_1, \ldots, x_n$, $\mathbb{E}_n$ indicates the conditional expectation given the available observations at time n, $\{(x_i, h(x_i))\}_{i=1}^n$, and $a^+ = \max(0, a)$ is the positive part function.

When h is scalar-valued and g is the identity function, EI-CF$_n$ is equivalent to the classical expected improvement computed directly from a GP prior on f, and has an analytic expression that makes it easy to compute and optimize. For general nonlinear functions g, however, EI-CF$_n$ cannot be

computed in closed form. Despite this, as we shall see next, under mild regularity conditions, EI-CF$_n$ can be efficiently maximized.

Figure 1 illustrates the EI-CF and classical EI acquisition functions in a setting where h is scalar-valued, $f(x) = g(h(x)) = h(x)^2$, we have evaluated h and f at four points, and we wish to minimize f. The right-hand column shows the posterior distribution on f and EI acquisition function using the standard approach: posterior credible intervals have 0 width at points where we have evaluated (since evaluations are free from noise), and become wider as we move away from them. The classical expected improvement is largest near the right limit of the domain, where the posterior mean computed using observations of $f(x)$ alone is relatively small and has large variance.

The left-hand column shows the posterior distribution on h, computed using a GP (single-output in this case, since h is scalar-valued), the resulting posterior distribution on f, and the resulting EI-CF acquisition function. The posterior distribution on $f(x)$ (which is not normally distributed) has support only on non-negative values, and places higher probability on small values of $f(x)$ in the regions $x \in [-2, -1] \cup [2.5, 3.5]$, which creates a larger value for EI-CF in these regions.

Examining past observations of $h(x)$, the points with high EI-CF ($x \in [-2, -1] \cup [2.5, 3.5]$) seem substantially more valuable to evaluate than the point with largest EI ($x = 5$). Indeed, for the region $[-2, -1]$, we know that $h(x)$ is below 0 near the left limit, and is above 0 near the right limit. The continuity of h implies that $h(x)$ is 0 at some point in this region, which in turn implies that f has a global optimum in this region. Similarly, f is also quite likely to have a global optimum in $[2.5, 3.5]$. EI-CF takes advantage of this knowledge in its sampling decisions while classical EI does not.

4.3. Computation and Maximization of EI-CF

We now describe computation and efficient maximization of EI-CF. For any fixed $x \in \mathcal{X}$, the time-n posterior distribution on $h(x)$ is multivariate normal. (By the "time-n posterior distribution", we mean the conditional distribution given $\{(x_i, h(x_i))\}_{i=1}^n$.) We let $\mu_n(x)$ denote its mean vector and $K_n(x)$ denote its covariance matrix. We also let $C_n(x)$ denote the lower Cholesky factor of $K_n(x)$. Then, we can express $h(x) = \mu_n(x) + C_n(x)Z$, where Z is a m-variate standard normal random vector under the time-n posterior distribution, and thus

$$\text{EI-CF}_n(x) = \mathbb{E}_n\left[\left\{g(\mu_n(x) + C_n(x)Z) - f_n^*\right\}^+\right]. \quad (3)$$

Thus, we can compute EI-CF$_n(x)$ via Monte Carlo, as summarized in Algorithm 1. We note that (3) and the condition $\mathbb{E}[|g(Z)|] < \infty$ imply that EI-CF$_n(x)$ is finite for all $x \in \mathcal{X}$.

Algorithm 1 Computation of EI-CF

Require: point to be evaluated, x; number of Monte Carlo samples, L
1: **for** $\ell = 1, \ldots, L$ **do**
2: Draw sample $Z^{(\ell)} \sim \mathcal{N}_m(0_m, I_m)$ and compute $\alpha^{(\ell)} := \left\{g\left(\mu_n(x) + C_n(x)Z^{(\ell)}\right) - f_n^*\right\}^+$
3: **end for**
4: Estimate EI-CF$_n(x)$ by $\frac{1}{L}\sum_{\ell=1}^L \alpha^{(\ell)}$

In principle, the above is enough to maximize EI-CF$_n$ using a derivative-free global optimization algorithm (for non-expensive noisy functions). However, such methods typically require a large number of samples, and optimization can be typically performed with much greater efficiency if derivative information is available (Jamieson et al., 2012; Swisher et al., 2000). The following proposition describes a simulation-based procedure for generating such derivative information. A formal statement and proof can be found in the supplementary material.

Proposition 1. *Under mild regularity conditions, EI-CF$_n$ is differentiable almost everywhere and its gradient, when it exists, is given by*

$$\nabla \textit{EI-CF}_n(x) = \mathbb{E}_n\left[\gamma_n(x, Z)\right], \quad (4)$$

where

$$\gamma_n(x, Z) = \begin{cases} 0, \text{ if } g(\mu_n(x) + C_n(x)Z) \le f_n^*, \\ \nabla g(\mu_n(x) + C_n(x)Z), \text{ otherwise.} \end{cases} \quad (5)$$

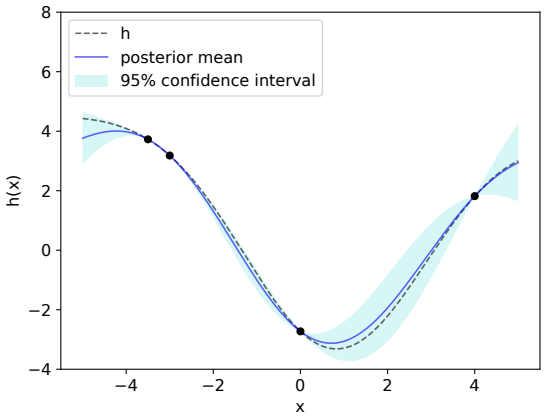
Thus, γ_n provides an unbiased estimator of ∇EI-CF$_n$. To construct such an estimator, we would draw an independent standard normal random vector Z and then compute $\gamma_n(x, Z)$ using (5), optionally averaging across multiple samples as in Algorithm 1. To optimize EI-CF$_n$, we then use this gradient estimation procedure within stochastic gradient ascent, using multiple restarts. The final iterate from each restart is an approximate stationary point of the EI-CF$_n$. We then use Algorithm 1 to select the stationary point with the best value of EI-CF$_n$.

4.4. Theoretical Analysis

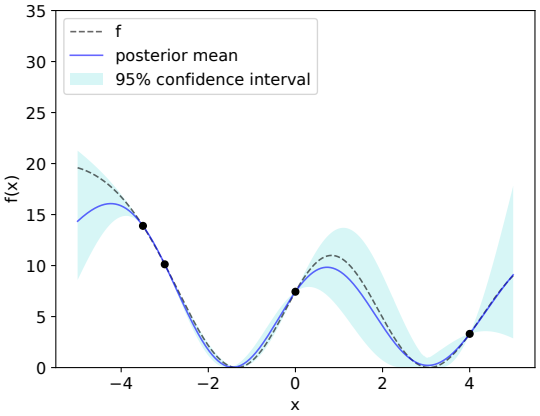
Here we present two results giving insight into the properties of the expected improvement for composite functions. The first result simply states that, when g is linear, EI-CF has a closed form analogous to the one of the classical EI.

Proposition 2. *Suppose that g is given by $g(y) = w^\top y$ for some fixed $w \in \mathbb{R}^m$. Then,*

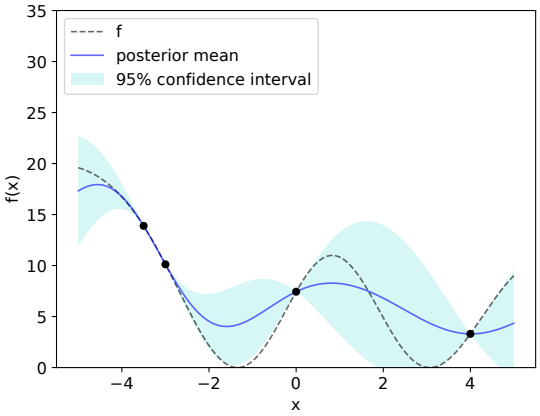
$$\textit{EI-CF}_n(x) = \Delta_n(x)\Phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) + \sigma_n(x)\varphi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right),$$



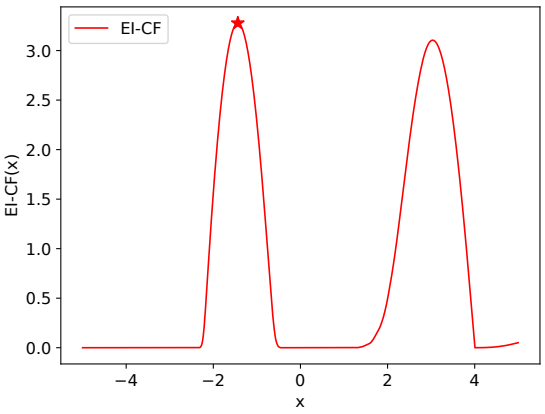
(a) Posterior on h used by our EI-CF acquisition function



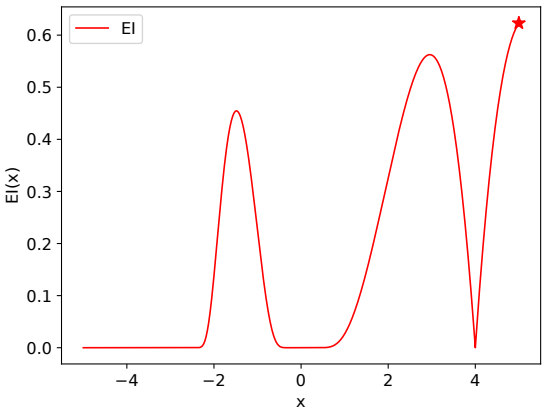
(b) Implied posterior on f used by our EI-CF acquisition function



(c) Posterior on f used by the classical EI acquisition function



(d) EI-CF acquisition function



(e) Classical EI acquisition function

Figure 1. Illustrative example of the EI-CF and classical EI acquisition functions, in a 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, and cause EI-CF to evaluate at points much closer to these global optima.

where $\Delta_n(x) = w^\top \mu_n(x) - f_n^*$, $\sigma_n(x) = \sqrt{w^\top K_n(x) w}$, *and* φ *and* Φ *are the standard normal probability density function and cumulative distribution function, respectively.*

This result can be easily verified by noting that, since the time-n posterior distribution of $h(x)$ is m-variate normal with mean vector $\mu_n(x)$ and covariance matrix $K_n(x)$, the time-n posterior distribution of $w^\top h(x)$ is normal with mean $w^\top \mu_n(x)$ and variance $w^\top K_n(x) w$. Proposition 2 does not, however, mean that our approach is equivalent to the classical one when g is linear. This is because, in general, the posterior distribution given observations of $h(x)$ is different from the one given observations of $w^\top h(x)$. We refer the reader to the supplementary material for a discussion.

Our second result states that, under suitable conditions, our acquisition function is asymptotically consistent, i.e., the solution found by our method converges to the global optimum when the number of evaluations goes to infinity. An analogous result for the classical expected improvement was proved by Vazquez & Bect (2010).

Theorem 1. *Let* $\{x_n\}_{n\in\mathbb{N}}$ *be the sequence of evaluated points and suppose there exists* $n_0 \in \mathbb{N}$ *such that for all* $n \geq n_0$,

$$x_{n+1} \in \arg\max_{x\in\mathcal{X}} \textit{EI-CF}_n(x).$$

Then, under suitable regularity conditions and as $n \to \infty$,

$$f_n^* \to \max_{x\in\mathcal{X}} f(x).$$

A formal statement and proof of Theorem 1 can be found in the supplementary material.

4.5. Computational Complexity of Posterior Inference

The computation required to maximize the classical EI acquisition function is dominated by the computation of the posterior mean and variance and thus in principle scales as $\mathcal{O}(n^2)$ (with a pre-computation of complexity $\mathcal{O}(n^3)$) with respect to the number of evaluations (Shahriari et al., 2016). However, recent advances on approximate fast GP training and prediction may considerably reduce the computational burden (Pleiss et al., 2018).

In our approach, the computational cost is again dominated by the computation of the posterior mean and covariance matrix, $\mu_n(x)$ and $K_n(x)$, respectively. When the outputs of h are modeled independently, the components of $\mu_n(x)$ and $K_n(x)$ can be computed separately ($K_n(x)$ is diagonal in this case) and thus computation of the posterior mean and covariance scales as $\mathcal{O}(mn^2)$. This allows our approach to be used even if h has a relatively large number of outputs. However, in general, if correlation between components of h is modeled, these computations scale as $\mathcal{O}(m^2n^2)$. Therefore, in principle there is a trade-off between modeling correlation between components of h, which presumably allows for a faster learning of h, and retaining tractability in the computation of the acquisition function.

5. Numerical Experiments

We compare the performance of three acquisition functions: expected improvement (EI), probability of improvement (PI) (Kushner, 1964), and the acquisition function that chooses points uniformly at random (Random), both under our proposed statistical model and the standard one, i.e., modeling h using a multi-output GP and modeling f directly using a single-output GP, respectively. We refer the reader to the supplementary material for a formal definition of the probability of improvement under our statistical model, and a discussion of how we maximize this acquisition function in our numerical experiments. To distinguish each acquisition function under our proposed statistical model from its standard version, we append "-CF" to its abbreviation; so if the classical expected improvement acquisition function is denoted EI, then the expected improvement under our statistical model is denoted EI-CF, as previously defined. We also include as a benchmark the predictive entropy search (PES) acquisition function (Hernández-Lobato et al., 2014) under the standard statistical model, i.e., modeling f directly using a single-output GP. For all problems and methods, an initial stage of evaluations is performed using $2(d + 1)$ points chosen uniformly at random over $\mathcal{X}$.

For EI-CF, PI-CF, and Random-CF, the outputs of h are modeled using independent GP prior distributions. All GP distributions involved, including those used by the standard BO methods (EI, PI, Random, and PES), have a constant mean function and ARD squared exponential covariance function; the associated hyperparameters are estimated under a Bayesian approach. As proposed in Snoek et al. (2012), for all methods we use an averaged version of the acquisition function, obtained by first drawing 10 samples of the GP hyperparameters, computing the acquisition function conditioned on each of these hyperparameters, and then averaging the results.

We calculate each method's simple regret at the point it believes to be the best based on evaluations observed thus far. We take this point to be the point with the largest (or smallest, if minimizing) posterior mean. For EI-CF, PI-CF, and Random-CF, we use the posterior mean on f implied by the GP posterior on h, and for EI, PI, Random, and PES we use the GP posterior on f. Error bars in the plots below show the mean of the base-10 logarithm of the simple regret plus and minus 1.96 times the standard deviation divided by the square root of the number of replications. Each experiment was replicated 100 times.

Our code is available at Astudillo (2019).

Problem	$\mathcal{X}$	g	m
1	$[0,1]^4$	$g(y) = -\|\|y - y_{\text{obs}}\|\|_2^2$	5
2	$[0,1]^3$	$g(y) = -\sum_j \exp(y_j)$	4

Table 1. Description of GP-generated test problems

5.1. GP-Generated Test Problems

The first two problems used functions h generated at random from GPs. Each component of h was generated by sampling on a uniform grid from independent GP distributions with different fixed hyperparameters and then taking the resulting posterior mean as a proxy; the hyperparameters were not known to any of the algorithms. The details of each problem, including the function g used, are summarized in Table 1.

Results are shown on a logarithmic scale in Figures 2 and 3, where the horizontal axis indicates the number of samples following the initial stage. EI-CF outperforms the other methods significantly. Regret is smaller than the best of the standard BO benchmarks throughout and by several orders of magnitude after 50 evaluations (5 orders of magnitude smaller in test 1, and 2 in test 2). It also requires substantially fewer evaluations beyond the first stage to reach the regret achieved by the best of the standard BO benchmarks in 100 evaluations: approximately 30 in test 1, and 10 in test 2. Random-CF performs surprisingly well in type-2 GP-generated problems, suggesting that a substantial part of the benefit provided by our approach is the value of the additional information available from observing $h(x)$. In type-1 problems it does not perform as well, suggesting that high-quality decisions about where to sample are also important.

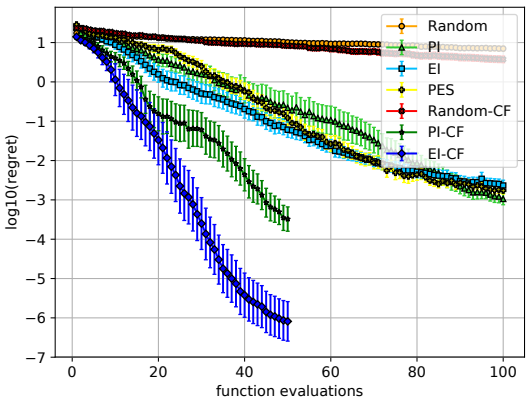


Figure 2. Expected log10(regret) in type-1 GP-generated test problems, estimated from 100 independent replications. These problems use $\mathcal{X} = [0,1]^4$, $g(y) = -||y - y_{\text{obs}}|_2^2$, and $m = 5$. EI-CF outperforms other methods by a large margin.

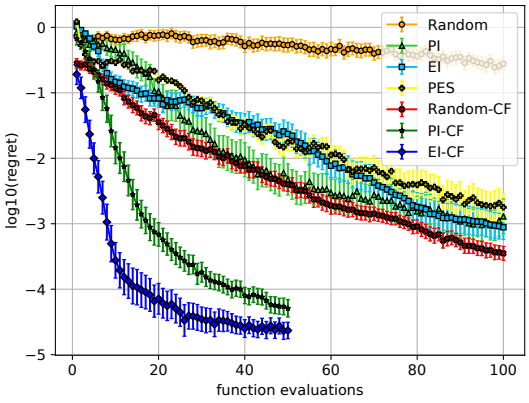


Figure 3. Expected log10(regret) in type-2 GP-generated test problems, estimated from 100 independent replications. These problems use $\mathcal{X} = [0,1]^3$, $g(y) = -\sum_j \exp(y_j)$, and $m = 4$.

5.2. Standard Global Optimization Test Problems

We assess our approach's performance on two standard benchmark functions from the global optimization literature: the Langermann (Surjanovic & Bingham, a) and Rosenbrock (Surjanovic & Bingham, b) functions. We refer the reader to the supplementary material for a description of how these functions are adapted to our setting.

Results of applying our method and benchmarks to these problems are shown on a logarithmic scale in Figures 4 and 5. As before, EI-CF outperforms competing methods with respect to the final achieved regret. PI-CF and Random-CF also perform well compared to methods other than EI-CF. Moreover, in the Langermann test problem, PI-CF outperforms EI-CF during the first 20 evaluations.

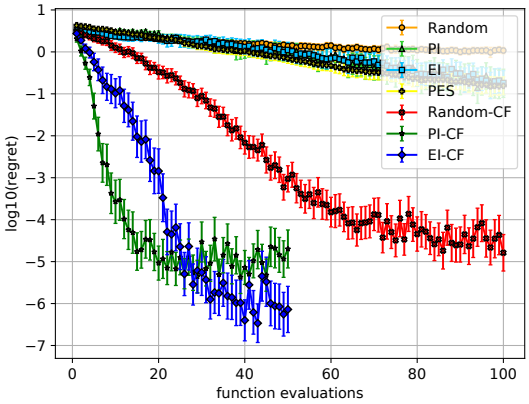


Figure 4. Expected log10(regret) in the Langermann test function, estimated from 100 independent replications.

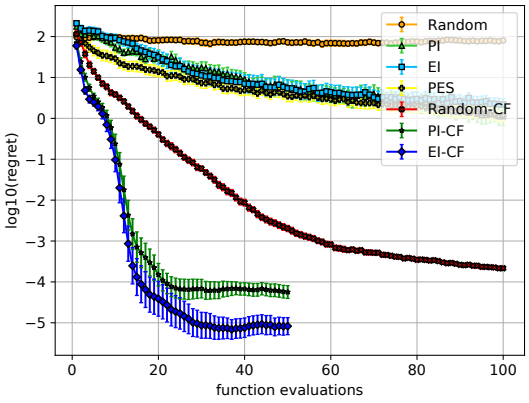


Figure 5. Expected log10(regret) in the Rosenbrock test problem, estimated from 100 independent replications.

5.3. Environmental Model Function

The environmental model function was originally proposed by Bliznyuk et al. (2008) and is now a well-known test problem in the literature of Bayesian calibration of expensive computer models. It models a chemical accident that has caused a pollutant to spill at two locations into a long and narrow holding channel, and is based on a first-order approach to modeling the concentration of substances in such channels under the assumption that the channel can be approximated by an infinitely long one-dimensional system with diffusion as the only method of transport. This leads to the concentration representation

$$c(s,t;M,D,L,\tau) = \frac{M}{\sqrt{4\pi Dt}}\exp\left(\frac{-s^2}{4Dt}\right) + \frac{\mathbb{I}\{t>\tau\}M}{\sqrt{4\pi D(t-\tau)}}\exp\left(\frac{-(s-L)^2}{4D(t-\tau)}\right),$$

where M is the mass of pollutant spilled at each location, D is the diffusion rate in the channel, L is the location of the second spill, and τ is the time of the second spill.

We observe $c(s,t;M_0,D_0,L_0,\tau_0)$ in a 3×4 grid of values; specifically, we observe $\{c(s,t;M_0,D_0,L_0,\tau_0) : (s,t) \in S\times T\}$, where $S=\{0,1,2.5\}$, $T=\{15,30,45,60\}$, and (M_0,D_0,L_0,τ_0) are the underlying true values of these parameters. Since we assume noiseless observations, the calibration problem reduces to finding (M,D,L,τ) so that the observations at the grid minimize the sum of squared errors, i.e., our goal is to minimize

$$\sum_{(s,t)\in S\times T}(c(s,t;M_0,D_0,L_0,\tau_0)-c(s,t;M,D,L,\tau))^2.$$

In our experiment, we take $M_0=10$, $D_0=0.07$, $L_0=1.505$ and $\tau_0=30.1525$. The search domain is $M\in[7,13]$, $D\in[0.02,0.12]$, $L\in[0.01,3]$ and $\tau\in[30.01,30.295]$.

Results from this experiment are shown in Figure 6. As above, EI-CF performs best, with PI-CF and Random-CF also significantly outperforming benchmarks that do not leverage the composite structure.

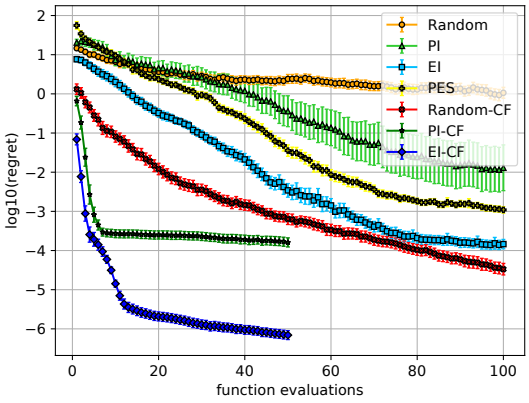


Figure 6. Expected log10(regret) in the environmental model function test problem, estimated from 100 independent replications.

6. Conclusion and Future Work

We have proposed a novel Bayesian optimization approach for objective functions of the form $f(x)=g(h(x))$, where h is a black-box expensive-to-evaluate vector-valued function, and g is a simple real-valued function that can be cheaply evaluated. Our numerical experiments show that this approach may substantially outperform standard Bayesian optimization, while retaining computational tractability.

There are several relevant directions for future work. Perhaps the most evident is to understand whether other well-known acquisition functions can be generalized to our setting in a computationally tractable way. We believe this to be true for predictive entropy search (Hernández-Lobato et al., 2014) and knowledge gradient (Scott et al., 2011). Importantly, these acquisition functions would allow noisy and decoupled evaluations of the components of h, thus increasing the applicability of our approach. However, in the standard Bayesian optimization setting, they are already computationally intensive and thus a careful analysis is required to make them computationally tractable in our setting.

Acknowledgements

This work was partially supported by NSF CAREER CMMI-1254298, NSF CMMI-1536895 and AFOSR FA9550-15-1-0038. The authors also thank Eytan Bakshy for helpful comments.

References

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Ashby, M. F. and Cebon, D. Materials selection in mechanical design. *Le Journal de Physique IV*, 3(C7):C7–1, 1993.

Astudillo, R. BOCF, 2019. URL https://github.com/RaulAstudillo06/BOCF.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, 17 (2):270–294, 2008.

Brochu, E., Cora, V. M., and De Freitas, N. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

Cullick, A. S., Johnson, W. D., Shi, G., et al. Improved and more rapid history matching with a nonlinear proxy and global optimization. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 2006.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *arXiv preprint arXiv:1605.00125*, 2016.

Frazier, P. I. and Wang, J. Bayesian optimization for materials design. In *Information Science for Materials Discovery and Design*, pp. 45–75. Springer, 2016.

Gardner, J., Kusner, M., Zhixiang, Weinberger, K., and Cunningham, J. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 937–945, 2014.

Griffiths, R.-R. and Hernández-Lobato, J. M. Constrained bayesian optimization for automatic chemical design. *arXiv preprint arXiv:1709.05501*, 2017.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in Neural Information Processing Systems*, pp. 918–926, 2014.

Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z. A general framework for constrained bayesian optimization using information-based search. *Journal of Machine Learning Research*, 17(160):1–53, 2016.

Jahan, A. and Edwards, K. L. A state-of-the-art survey on the influence of normalization techniques in ranking: Improving the materials selection process in engineering design. *Materials & Design (1980-2015)*, 65:335–342, 2015.

Jamieson, K. G., Nowak, R., and Recht, B. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2012.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

Ju, S., Shiga, T., Feng, L., Hou, Z., Tsuda, K., and Shiomi, J. Designing nanostructures for phonon transport via bayesian optimization. *Physical Review X*, 7(2):021024, 2017.

Kapetanovic, I. Computer-aided drug discovery and development (caddd): in silico-chemico-biological approach. *Chemico-Biological Interactions*, 171(2):165–176, 2008.

Kushner, H. J. 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, 1964.

Liu, H., Cai, J., and Ong, Y.-S. Remarks on multi-output gaussian process regression. *Knowledge-Based Systems*, 144:102–121, 2018.

Marque-Pucheu, S., Perrin, G., and Garnier, J. Efficient sequential experimental design for surrogate modeling of nested codes. *arXiv preprint arXiv:1712.01620*, 2017.

Močkus, J. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404. Springer, 1975.

Overstall, A. M. and Woods, D. C. A strategy for bayesian inference for computationally expensive models with application to the estimation of stem cell properties. *Biometrics*, 69(2):458–468, 2013.

Picheny, V., Gramacy, R. B., Wild, S., and Le Digabel, S. Bayesian optimization under mixed constraints with a slack-variable augmented lagrangian. In *Advances in Neural Information Processing Systems*, pp. 1435–1443, 2016.

Pleiss, G., Gardner, J. R., Weinberger, K. Q., and Wilson, A. G. Constant-time predictive distributions for gaussian processes. *arXiv preprint arXiv:1803.06058*, 2018.

Schonlau, M., Welch, W. J., and Jones, D. R. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pp. 11–25, 1998.

Schultz, L. and Sokolov, V. Bayesian optimization for transportation simulators. *Procedia Computer Science*, 130: 973–978, 2018.

Scott, W., Frazier, P., and Powell, W. The correlated knowledge gradient for simulation optimization of continuous parameters using gaussian process regression. *SIAM Journal on Optimization*, 21(3):996–1026, 2011.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2016.

Shapiro, A. On a class of nonsmooth composite functions. *Mathematics of Operations Research*, 28(4):677–692, 2003.

Snoek, J., Larochelle, H., and Adams, R. P. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

Surjanovic, S. and Bingham, D. Langermann function, a. URL https://www.sfu.ca/~ssurjano/langer.html.

Surjanovic, S. and Bingham, D. Rosenbrock function, b. URL https://www.sfu.ca/~ssurjano/rosen.html.

Swersky, K., Snoek, J., and Adams, R. P. Multi-task bayesian optimization. In *Advances in Neural Information Processing Systems*, pp. 2004–2012, 2013.

Swisher, J. R., Hyden, P. D., Jacobson, S. H., and Schruben, L. W. A survey of simulation optimization techniques and procedures. In *Proceedings of the 2000 Winter Simulation Conference*, volume 1, pp. 119–128. IEEE, 2000.

Ueno, T., Rhone, T. D., Hou, Z., Mizoguchi, T., and Tsuda, K. Combo: An efficient bayesian optimization library for materials science. *Materials Discovery*, 4:18–21, 2016.

Vazquez, E. and Bect, J. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095, 2010.

Vrugt, J., Hopmans, J., and Šimunek, J. Calibration of a two-dimensional root water uptake model. *Soil Science Society of America Journal*, 65(4):1027–1037, 2001.

Wang, J., Clark, S. C., Liu, E., and Frazier, P. I. Parallel Bayesian global optimization of expensive functions. *arXiv preprint arXiv:1602.05149*, 2016.

Wu, J. and Frazier, P. The parallel knowledge gradient method for batch Bayesian optimization. In *Advances in Neural Information Processing Systems*, pp. 3126–3134, 2016.

Bayesian Optimization of Composite Functions: Supplementary Material

Raul Astudillo [1] **Peter I. Frazier** [1 2]

1. Unbiased Estimator of the Gradient of EI-CF

In this section we prove that, under mild regularity conditions, EI-CF$_n$ is differentiable and an unbiased estimator of its gradient can be efficiently computed. More concretely, we prove the following.

Proposition 1.1. *Suppose that g is differentiable and let $\mathcal{X}_0$ be an open subset of $\mathcal{X}$ so that μ_n and K_n are differentiable on $\mathcal{X}_0$ and there exists a measurable function $\eta : \mathbb{R}^m \to \mathbb{R}$ satisfying*

1. $\|\nabla g\left(\mu_n(x) + C_n(x)z\right)\| < \eta(z)$ *for all* $x \in \mathcal{X}_0,\ z \in \mathbb{R}^m$,
2. $\mathbb{E}[\eta(Z)] < \infty$, *where Z is a m-variate standard normal random vector.*

Further, suppose that for almost every $z \in \mathbb{R}^m$ the set $\{x \in \mathcal{X}_0 : g\left(\mu_n(x) + C_n(x)z\right) = f_n^*\}$ is countable. Then, EI-CF$_n$ is differentiable on $\mathcal{X}_0$ and its gradient is given by

$$\nabla \textit{EI-CF}_n(x) = \mathbb{E}_n\left[\gamma_n(x, Z)\right],$$

where the expectation is with respect to Z and

$$\gamma_n(x,z) = \begin{cases} \nabla g\left(\mu_n(x) + C_n(x)z\right), \text{ if } g\left(\mu_n(x) + C_n(x)z\right) > f_n^*, \\ 0, \text{ otherwise.} \end{cases}$$

Proof. Since g is differentiable and μ_n and K_n are differentiable on $\mathcal{X}_0$, for any fixed $z \in \mathbb{R}^m$ the function $x \mapsto g\left(\mu_n(x) + C_n(x)z\right)$ is differentiable on $\mathcal{X}_0$ as well. This in turn implies that the function $x \mapsto \{g\left(\mu_n(x) + C_n(x)z\right) - f_n^*\}^+$ is continuous on $\mathcal{X}_0$ and differentiable at every $x \in \mathcal{X}_0$ such that $g\left(\mu_n(x) + C_n(x)z\right) \neq f_n^*$, with gradient equal to $\gamma(x, z)$. From our assumption that for almost every $z \in \mathbb{R}^m$ the set $\{x \in \mathcal{X} : g\left(\mu_n(x) + C_n(x)z\right) = f_n^*\}$ is countable, it follows that for almost every z the function $x \mapsto \{g\left(\mu_n(x) + C_n(x)z\right) - f_n^*\}^+$ is continuous on $\mathcal{X}_0$ and differentiable on all $\mathcal{X}_0$, except maybe on a countable subset. Using this, along with conditions 1 and 2, and Theorem 1 in L'Ecuyer (1990), the desired result follows. □

We end this section by making a few remarks.

- If μ and K are differentiable on int$(\mathcal{X})$, then one can show that μ_n and K_n are differentiable on int$(\mathcal{X}) \setminus \{x_1, \ldots, x_n\}$.
- If one imposes the stronger condition $\mathbb{E}[\eta(Z)^2] < \infty$, then γ_n has finite second moment, and thus this unbiased estimator of ∇EI-CF$_n(x)$ can be used within stochastic gradient ascent to find a stationary point of EI-CF$_n$ (Bottou, 1998).
- In Proposition 1.1, the condition that for almost every $z \in \mathbb{R}^m$ the set $\{x \in \mathcal{X}_0 : g\left(\mu_n(x) + C_n(x)z\right) = f_n^*\}$ is countable, can be weakened to the following more technical condition: for almost every $z \in \mathbb{R}^m$, every $x \in \mathcal{X}_0$ and every $i \in \{1, \ldots, d\}$, there exists $\epsilon > 0$ such that the set $\{x + he_i : |h| < \epsilon$ and $g\left(\mu_n(x + he_i) + C_n(x + he_i)z\right) = f_n^*\}$ is countable, where e_i denotes the i-th canonical vector in $\mathbb{R}^d$.

[1]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY, USA [2]Uber, San Francisco, CA, USA. Correspondence to: Raul Astudillo <ra598@cornell.edu>.

2. EI-CF and EI Do Not Coincide When g Is Linear

Recall the following result that was stated in the main paper.

Proposition 2.1. *Suppose that g is given by $g(y) = w^\top y$ for some fixed $w \in \mathbb{R}^m$. Then,*

$$\textit{EI-CF}_n(x) = \Delta_n(x)\Phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) + \sigma_n(x)\varphi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right)$$

The resemblance of the above expression to the classical EI acquisition function may make one think that, in the above case, EI-CF coincides, in some sense, with the classical EI under an appropriate choice of the prior distribution.

Indeed, suppose that we set a single-output GP prior with mean $w^\top \mu(x)$ and covariance function $w^\top K_n(x) w$ of f (and fix its hyperparameters), then

$$\mathbb{E}\left[\left\{w^\top h(x) - f_n^*\right\}^+ \mid x_i, w^\top h(x_i) = y_i : i = 1, \dots n\right] = \mathbb{E}\left[\left\{f(x) - f_n^*\right\}^+ \mid x_i, f(x_i) = y_i : i = 1, \dots n\right].$$

However, if we condition on $h(x_i)$ rather than $w^\top h(x_i)$ in the left-hand side, then the equality is no longer true, even if the values on which we condition satisfy $w^\top h(x_i) = y_i$:

$$\mathbb{E}\left[\left\{w^\top h(x) - f_n^*\right\}^+ \mid x_i, h(x_i) : i = 1, \dots n\right] \neq \mathbb{E}\left[\left\{f(x) - f_n^*\right\}^+ \mid x_i, f(x_i) = y_i : i = 1, \dots n\right].$$

Thus, even if we initiate optimization using EI-CF and a parallel optimization using EI with a single-output Gaussian process as described above, their acquisition functions will cease to agree once we condition on the results of an evaluation.

3. Probability of Improvement for Composite Functions

In this section, we formally define the probability of improvement for composite functions (PI-CF) acquisition function and specify its implementation details used within our experimental setup.

Analogously to EI-CF, PI-CF is simply defined as the probability of improvement evaluated with respect to the implied posterior distribution on f when we model h as a multi-output GP:

$$\text{PI-CF}(x) = \mathbb{P}_n\left(g(h(x)) \geq f_n^* + \delta\right),$$

where $\mathbb{P}_n$ denotes the conditional probability given the available observations at time n, $\{(x_i, h(x_i))\}_{i=1}^n$, and $\delta > 0$ is a parameter to be specified. As we did with EI-CF, we can express PI-CF(x) as

$$\text{PI-CF}(x) = \mathbb{P}_n\left(g(\mu_n(x) + C_n(x)Z) \geq f_n^* + \delta\right),$$

where Z is a m-variate standard normal random vector under the time-n posterior distribution.

We can further rewrite PI-CF(x) using an indicator function $\mathbb{I}$ as

$$\text{PI-CF}(x) = \mathbb{E}_n\left[\mathbb{I}\{g(\mu_n(x) + C_n(x)Z) \geq f_n^* + \delta\}\right],$$

which implies that PI-CF can be computed with arbitrary precision following a Monte Carlo approach as well:

$$\text{PI-CF}(x) \approx \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{I}\left\{g\left(\mu_n(x) + C_n(x)Z^{(\ell)}\right) \geq f_n^* + \delta\right\},$$

where $Z^{(1)}, \dots, Z^{(L)}$ are draws of an m-variate standard normal random vector. However, an unbiased estimator of the gradient of PI-CF cannot be computed following an analogous approach to the one used with EI-CF. In fact, $\nabla\mathbb{I}\{g(\mu_n(x) + C_n(x)Z) \geq f_n^* + \delta\} = 0$ at those points for which the function $x \mapsto \mathbb{I}\{g(\mu_n(x) + C_n(x)Z) >\geq f_n^* + \delta\}$ is differentiable. Thus, even if $\nabla\mathbb{I}\{g(\mu_n(x) + C_n(x)Z) \geq f_n^* + \delta\}$ exists, in general

$$\nabla\text{PI-CF}(x) \neq \mathbb{E}_n\left[\nabla\mathbb{I}\{g(\mu_n(x) + C_n(x)Z) \geq f_n^* + \delta\}\right],$$

unless $\nabla\text{PI-CF}(x) = 0$.

In our experiments, we adopt a sample average approximation (Kim et al., 2015) scheme for approximately maximizing PI-CF. At each iteration we fix $Z^{(1)}, \ldots, Z^{(L)}$ and choose the next point to evaluate as

$$x_{n+1} \in \arg\max_{x \in \mathcal{X}} \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{I}\left\{ g\left(\mu_n(x) + C_n(x) Z^{(\ell)} \right) \geq f_n^* + \delta \right\},$$

where $L = 50$ and $\delta = 0.01$. We solve the above optimization problem using the derivative-free optimization algorithm, CMA-ES (Hansen, 2016).

4. Description of Langermann and Rosenbrock Test Problems

The following pair of test problems are standard benchmarks in the global optimization literature. In this section, we describe in detail how they are adapted our setting, i.e., how we express them as composite functions.

4.1. Langermann Function

The Langermann function (Surjanovic & Bingham, a) is defined by $f(x) = g(h(x))$ where

$$h_j(x) = \sum_{i=1}^{d} (x_i - A_{ij})^2, \; j = 1, \ldots m,$$
$$g(y) = -\sum_{j=1}^{m} c_j \exp(-y_j/\pi) \cos(\pi y_j).$$

In our experiment we set $d = 2$, $m = 5$, $c = (1, 2, 5, 2, 3)$,

$$A = \begin{pmatrix} 3 & 5 & 2 & 1 & 7 \\ 5 & 2 & 1 & 4 & 9 \end{pmatrix},$$

and $\mathcal{X} = [0, 10]^2$.

4.2. Rosenbrock Function

The Rosenbrock function (Surjanovic & Bingham, b) is

$$f(x) = -\sum_{j=1}^{d-1} 100(x_{j+1} - x_j^2)^2 + (x_j - 1)^2$$

We adapt this problem to our framework by taking $d = 5$ and defining h and g by

$$h_j(x) = x_{j+1} - x_j^2, \; j = 1, \ldots, 4,$$
$$h_{j+4}(x) = x_j, \; j = 1, \ldots, 4,$$
$$g(y) = -\sum_{j=1}^{4} 100 y_j^2 + (y_{j+4} - 1)^2.$$

5. Asymptotic Consistency of the Expected Improvement for Composite Functions

5.1. Basic Definitions and Assumptions

In this section we prove that, under suitable conditions, the expected improvement sampling policy is asymptotically consistent in our setting. In the standard Bayesian optimization setting, this was first proved under quite general conditions by Vazquez & Bect (2010). Later, Bull (2011) provided convergence rates for several expected improvement-type policies both with fixed hyperparameters and hyperparameters estimated from the data in suitable way. Here, we restrict to prove

asymptotic consistency, under fixed hyperparameters, following a similar approach to Vazquez & Bect (2010). In particular, we provide a generalization of the No-Empty-Ball (NEB) condition, under which the expected improvement sampling policy is guaranteed to be asymptotically consistent in our setting. In the reminder of this work $\{x_n\}_{n\in\mathbb{N}}$ denotes the sequence of points at which h is evaluated, which is not necessarily given by the expected improvement acquisition function, unless explicitly stated.

Definition 5.1 (Generalized-No-Empty-Ball property)**.** *We shall say that a kernel, K, satisfies the Generalized-No-Empty-Ball (GNEB) property if, for all sequences $\{x_n\}_{\mathbb{N}}$ in $\mathcal{X}$ and all $\tilde{x} \in \mathcal{X}$, the following assertions are equivalent:*

1. *$\tilde{x}$ is a limit point of $\{x_n\}_{n\in\mathbb{N}}$.*
2. *There exists a subsequence of $\{K_n(\tilde{x})\}_{n\in\mathbb{N}}$ converging to a singular matrix.*

We highlight that, if K is diagonal, i.e. if the output components are independent of each other, the GNEB property holds provided that at least one of its components satisfies the standard NEB property. In particular, the following result is a corollary of Proposition 10 in Vazquez & Bect (2010).

Corollary 5.2. *Suppose K is diagonal and at least one of its components has a spectral density whose inverse has at most polynomial growth. Then, K satisfies the GNEB property.*

Thus, the GNEB property holds, in particular, if K is diagonal and at least one of its components is a Matérn kernel (Stein, 2012).

Now we introduce some additional notation. We denote by $\mathcal{H}$ to the the reproducing kernel Hilbert space associated with K (Alvarez et al., 2012). As is standard in Bayesian optimization, we make a slight abuse of notation and denote by h both a fixed element of $\mathcal{H}$ and a random function distributed according to a Gaussian process with mean μ and kernel K (below we assume $\mu = 0$); we shall explicitly state whenever h is held fixed. As before, we denote $K_n(x, x)$ by $K_n(x)$. Finally, we make the following standing assumptions.

1. $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, for some $d \geq 1$.
2. The prior mean function is identically 0.
3. K is continuous, positive definite, and satisfies the GNEB property.
4. $g : \mathbb{R}^m \to \mathbb{R}$ is continuous.
5. For any bounded sequences $\{a_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^m$ and $\{A_n\}_{n\in\mathbb{N}} \subset \mathbb{R}^{m\times m}$, $\mathbb{E}[\sup_n |g(a_n + A_n Z)|] < \infty$, where the expectation is over Z and Z is a m-dimensional standard normal random vector.

The assumption that g is continuous guarantees that $f = g \circ h$ is continuous, provided that h is continuous as well. Moreover, in this case, since $\mathcal{X}$ is compact, f attains its maximum value in $\mathcal{X}$; we shall denote this maximum value by M, i.e., $M = \max_{x\in\mathcal{X}} f(x)$.

5.2. Preliminary Results

Before proving asymptotic consistency, we prove several auxiliary results. We begin by proving that EI-CF$_n$ is continuous.

Proposition 5.3. *For any $n \in \mathbb{N}$, the function EI-CF$_n$: $\mathcal{X} \to \mathbb{R}$ defined by*

$$\textit{EI-CF}_n(x) = \mathbb{E}[\{g\left(\mu_n(x) + C_n(x)Z\right) - f_n^*\}^+],$$

where the expectation is over Z and Z is a m-dimensional standard normal random vector, is continuous.

Proof. Let $\{x'_k\}_{k\in\mathbb{N}} \subset \mathcal{X}$ be a convergent sequence with limit x'_∞. Since K is continuous, μ_n and C_n are both continuous functions of x, and thus $\mu_n(x'_k) \to \mu_n(x'_\infty)$ and $C_n(x'_k) \to C_n(x'_\infty)$ as $k \to \infty$. Moreover, since g is continuous too, it follows by the continuous mapping theorem (Billingsley, 2013) that

$$\{g(\mu_n(x'_k) + C_n(x'_k)Z) - f_n^*\}^+ \to \{g(\mu_n(x'_\infty) + C_n(x'_\infty)Z) - f_n^*\}^+$$

almost surely as $k \to \infty$.

Now observe that

$$\{g(\mu_n(x'_k) + C_n(x'_k)Z) - f_n^*\}^+ \leq \sup_k |g(\mu_n(x'_k) + C_n(x'_k)Z)| + |f_n^\star|.$$

Moreover, the sequences $\{\mu_n(x'_k)\}_{k\in\mathbb{N}}$ and $\{C_n(x'_k)\}$ are both convergent (with finite limits) and thus are bounded. Hence, the above inequality, along with assumption 5 and the dominated convergence theorem (Williams, 1991), imply that

$$\mathbb{E}[\{g(\mu_n(x'_k) + C_n(x'_k)Z) - f_n^*\}^+] \to \mathbb{E}[\{g(\mu_n(x'_\infty) + C_n(x'_\infty)Z) - f_n^*\}^+],$$

as $k \to \infty$, i.e., EI-CF$_n(x'_k) \to$ EI-CF$_n(x'_\infty)$. Hence, EI-CF$_n$ is continuous. □

Lemma 5.4. *Let $\{x_n\}_{n\in\mathbb{N}}$ and $\{x'_n\}_{n\in\mathbb{N}}$ be two sequences in $\mathcal{X}$. Assume that $\{x'_n\}_{n\in\mathbb{N}}$ is convergent, and denote by x'_∞ its limit. Then, each of the following conditions implies the next one:*

1. *x'_∞ is a limit point of $\{x_n\}_{n\in\mathbb{N}}$.*
2. *$K_n(x'_n) \to 0$ as $n \to \infty$.*
3. *For any fixed $h \in \mathcal{H}$, $\mu_n(x'_n) \to h(x'_\infty)$ as $n \to \infty$.*

Proof. First we prove that 1 implies 2. If x'_∞ is an element of $\{x_n\}_{n\in\mathbb{N}}$, say $x'_\infty = x_{n_0}$, then, for $n \geq n_0$, we have

$$K_n(x'_n) \lesssim K_{n_0}(x'_n) \to K_{n_0}(x'_\infty) = K_{n_0}(x_{n_0}) = 0,$$

where we use Lemma 6.3 and the fact that K_{n_0} is continuous. Now assume x'_∞ is not an element of $\{x_n\}_{n\in\mathbb{N}}$. Let $\{x_{k_n}\}_{n\in\mathbb{N}}$ be a subsequence of $\{x_n\}_{n\in\mathbb{N}}$ converging to x'_∞ and let $m_n = \max\{k_\ell : k_\ell \leq n\}$. Then, by Lemmas 6.1 and 6.2 we obtain

$$K_n(x'_n) = \text{Cov}(h(x'_n) - \mu_n(x'_n)) \lesssim \text{Cov}(h(x'_n) - h(x_{m_n})).$$

Finally, since x'_∞ is not an element of $\{x_n\}_{n\in\mathbb{N}}$, $m_n \to \infty$, and it follows from the continuity of K that

$$\text{Cov}(h(x'_n) - h(x_{m_n})) = K(x'_n, x'_n) + K(x_{m_n}, x_{m_n}) - 2K(x'_n, x_{m_n}) \to 0,$$

and thus $K_n(x'_n) \to 0$.

Now we prove that 2 implies 3. Using the Cauchy-Schwarz inequality in $\mathcal{H}$, we obtain

$$\|h(x'_n) - \mu_n(x'_n)\|_2 \leq \|K_n(x'_n)\|_2^{\frac{1}{2}} \|h\|_{\mathcal{H}},$$

Thus,

$$\begin{aligned}\|h(x'_\infty) - \mu_n(x'_n)\|_2 &\leq \|h(x'_\infty) - h(x'_n)\|_2 + |h(x'_n) - \mu_n(x'_n)\|_2 \\ &\leq \|h(x'_\infty) - h(x'_n)\|_2 + \|K_n(x'_n)\|_2^{\frac{1}{2}} \|h\|_{\mathcal{H}} \to 0\end{aligned}$$

since h is continuous. □

Lemma 5.5. *Let $\nu_n = \max_{x\in\mathcal{X}}$ EI-CF$_n(x)$. Then, for all $h \in \mathcal{H}$, $\liminf_{n\to\infty} \nu_n \to 0$.*

Proof. Fix $h \in \mathcal{H}$ and let $\{x_n\}_{n\in\mathbb{N}}$ be the sequence of points generated by the expected improvement policy, i.e., $x_{n+1} \in \arg\max_{x\in\mathcal{X}}$ EI-CF$_n(x)$. Let $\tilde{x}$ be a limit point of $\{x_n\}_{n\in\mathbb{N}}$ and let $\{x_{k_n}\}_{n\in\mathbb{N}}$ be any subsequence converging to $\tilde{x}$. Consider the sequence $\{x'_n\}_{n\in\mathbb{N}}$ given by $x'_n = x_{k_\ell}$ for all $k_{\ell-1} \leq n < k_\ell$, $n \in \mathbb{N}$. Clearly, $x'_n \to \tilde{x}$, and thus Lemma 5.4 implies that $\mu_n(x'_n) \to h(\tilde{x})$ and $K_n(x'_n) \to 0$. In particular, $\mu_{k_n-1}(x'_{k_n-1}) \to h(\tilde{x})$ and $C_{k_n-1}(x'_{k_n-1}) \to 0$, i.e., $\mu_{k_n-1}(x_{k_n}) \to h(\tilde{x})$ and $C_{k_n-1}(x_{k_n}) \to 0$. Moreover, $\{f_n^*\}_{n\in\mathbb{N}}$ is a bounded increasing sequence, and thus has a finite limit, f_∞^*, which satisfies $f_\infty^* \geq f(\tilde{x})$ as $\tilde{x}$ is a limit point of $\{x_n\}_{n\in\mathbb{N}}$ and f is continuous.

The sequences $\{\mu_{k_n-1}(x_{k_n})\}_{n\in\mathbb{N}}$ and $\{C_{k_n-1}(x_{k_n})\}_{n\in\mathbb{N}}$ are convergent and thus bounded. Hence, from assumption 5 and the dominated convergence theorem we obtain that

$$\begin{aligned}\mathbb{E}\left[\{g(\mu_{k_n-1}(x_{k_n})+C_{k_n-1}(x_{k_n})Z)-f^*_{k_{n-1}}\}^+\right] &\to \mathbb{E}\left[\{g(h(\tilde{x}))-f^*_\infty\}^+\right]\\ &= \mathbb{E}\left[\{f(\tilde{x})-f^*_\infty\}^+\right]=0,\end{aligned}$$

but

$$\nu_{k_{n-1}} = \mathbb{E}\left[\{g(\mu_{k_n-1}(x_{k_n})+C_{k_n-1}(x_{k_n})Z)-f^*_{k_{n-1}}\}^+\right],$$

and thus the desired conclusion follows. □

5.3. Proof of the Main Result

We are now in position to prove that the expected improvement acquisition function is asymptotically consistent in the composite functions setting.

Theorem 5.6 (Asymptotic consistency of EI-CF)**.** *Assume that the covariance function, K, satisfies the GNEB property. Then, for any fixed $h \in \mathcal{H}$ and $x_{\text{init}} \in \mathcal{X}$, any (measurable) sequence $\{x_n\}_{n\in\mathbb{N}}$ with $x_1 = x_{\text{init}}$ and $x_{n+1} \in \arg\max_{x\in\mathcal{X}}$ EI-CF$_n(x)$, $n \in \mathbb{N}$, satisfies $f^*_n \to M$.*

Proof. First note that if $\{x_n\}_{n\in\mathbb{N}}$ is dense in $\mathcal{X}$, then, by continuity of f, $f^*_n \to M$. Thus, we may assume that $\{x_n\}_{n\in\mathbb{N}}$ is not dense in $\mathcal{X}$. For the sake of contradiction, we also assume that $f^*_\infty := \lim_{n\to\infty} f^*_n < M$, which implies that we can find $\epsilon > 0$ such that $f^*_\infty \le M - 2\epsilon$.

Since $\{x_n\}_{n\in\mathbb{N}}$ is not dense in $\mathcal{X}$, there exists $x_\star \in \mathcal{X}$ that is not a limit point of $\{x_n\}_{n\in\mathbb{N}}$. Applying the Cauchy-Schwarz inequality in $\mathcal{H}$, we obtain

$$\|\mu_n(x_\star) - h(x_\star)\|_2 \le \|K_n(x_\star)\|^{\frac{1}{2}}\|h\|_{\mathcal{H}} \le \|K(x_\star)\|_2^{\frac{1}{2}}\|h\|_{\mathcal{H}},$$

where in the last inequality we use that the sequence $\{K_n(x_\star)\}_{n\in\mathbb{N}}$ satisfies $K_{n+1}(x_\star) \lesssim K_n(x_\star) \lesssim K(x_\star)$ for all $n \in \mathbb{N}$. It follows that both sequences $\{\mu_n(x_\star)\}_{n\in\mathbb{N}}$ and $\{K_n(x_\star)\}_{n\in\mathbb{N}}$ are bounded and thus we can find convergent subsequences $\{\mu_{k_n}(x_\star)\}_{n\in\mathbb{N}}$ and $\{K_{k_n}(x_\star)\}_{n\in\mathbb{N}}$, say with limits $\mu_\star$ and $K_\star$, respectively. The GNEB property implies that $K_\star$ is nonsingular. Let $C_\star$ be the upper cholesky factor of $K_\star$ and let $S_\epsilon = \{y \in \mathbb{R}^m : M - \epsilon \le g(y) \le M\}$. By continuity of g, S_ϵ has positive Lebesgue measure, and since $K_\star$ is nonsingular, $\mu_\star + C_\star Z$ is a multivariate normal random vector with full support. Hence, $\mathbb{P}(\mu_\star + C_\star Z \in S_\epsilon) > 0$. Moreover,

$$\begin{aligned}\mathbb{E}\left[\{g(\mu_\star + C_\star Z) - f^*_\infty\}^+\right] &\ge \mathbb{E}[\epsilon\mathbb{I}\{\mu_\star + C_\star Z \in S_\epsilon\}]\\ &= \epsilon\mathbb{P}(\mu_\star + C_\star Z \in S_\epsilon) > 0.\end{aligned}$$

Finally, using Fatou's lemma we obtain

$$\liminf_{n\to\infty} \mathbb{E}\left[\{g(\mu_{k_n}(x_\star) + C_{k_n}(x_\star)Z) - f^*_{k_n}\}^+\right] \ge \mathbb{E}\left[\{g(\mu_\star + C_\star Z) - f^*_\infty\}^+\right] > 0,$$

i.e., $\liminf_{n\to\infty}$ EI-CF$_{k_n}(x_\star) > 0$, which contradicts lemma 5.5. □

6. Auxiliary Results

Here we prove some basic results on multi-output Gaussian processes. Most of them are simple generalizations of well-known facts for single-output Gaussian processes but are included here for completeness.

Lemma 6.1. *Suppose that the sequence $\{x_n\}_{n\in\mathbb{N}}$ is deterministic. Then, for any fixed $x \in \mathcal{X}$*

$$K_n(x) = \text{Cov}(h(x) - \mu_n(x)).$$

Proof. This result may seem obvious at first sight, but it requires careful interpretation. By definition we have

$$K_n(x) = \text{Cov}_n(h(x) - \mu_n(x)),$$

but we claim that, indeed,

$$K_n(x) = \text{Cov}(h(x) - \mu_n(x)),$$

i.e., the same equality holds even if we do not condition on the information available at time n. To see this, it is enough to recall that $K_n(x)$ only depends on $x_1, \ldots, x_n$, but not on the values of h at these points. Thus, the tower property of the expectation yields

$$\begin{aligned} K_n(x) &= \mathbb{E}[K_n(x)] \\ &= \mathbb{E}\left[\mathbb{E}_n\left[(h(x) - \mu_n(x))(h(x) - \mu_n(x))^\top\right]\right] \\ &= \mathbb{E}\left[(h(x) - \mu_n(x))(h(x) - \mu_n(x))^\top\right] \\ &= \text{Cov}(h(x) - \mu_n(x)), \end{aligned}$$

where in the last equality we use that $\mathbb{E}[h(x) - \mu_n(x)] = 0$, which can be verified similarly:

$$\mathbb{E}[h(x) - \mu_n(x)] = \mathbb{E}\left[\mathbb{E}_n[h(x) - \mu_n(x)]\right] = \mathbb{E}\left[0\right] = 0.$$

□

We emphasize that the sequence of points generated by the expected improvement acquisition function is deterministic once h (the function to be evaluated, not the Gaussian process) is fixed and thus satisfies the conditions of lemma 6.1.

Lemma 6.2. *For any $x \in \mathcal{X}$, $n \in \mathbb{N}$ and $i \in \{1, \ldots, n\}$,*

$$\text{Cov}(h(x) - \mu_n(x)) \lesssim \text{Cov}(h(x) - h(x_i)).$$

Proof. By the law of total covariance, we have

$$\mathbb{E}\left[\text{Cov}_n(h(x) - h(x_i))\right] + \text{Cov}\left(\mathbb{E}_n[h(x) - h(x_i)]\right) = \text{Cov}(h(x) - h(x_i)),$$

which implies that

$$\mathbb{E}\left[\text{Cov}_n(h(x) - h(x_i))\right] \lesssim \text{Cov}(h(x) - h(x_i)),$$

Moreover, conditioned on the information at time n, both $\mu_n(x)$ and $h(x_i)$ are deterministic. Hence,

$$\text{Cov}_n(h(x) - \mu_n(x)) = \text{Cov}_n(h(x) - h(x_i)),$$

but by lemma 6.1 we know that $\text{Cov}_n(h(x) - \mu_n(x)) = \text{Cov}(h(x) - \mu_n(x))$, and thus $\mathbb{E}[\text{Cov}_n(h(x) - h(x_i))] = \text{Cov}(h(x) - \mu_n(x))$, which completes the proof. □

Lemma 6.3. *For any fixed $x \in \mathcal{X}$ and $n \in \mathbb{N}$,*

$$K_{n+1}(x) \lesssim K_n(x) \lesssim K(x)$$

Proof. Let $K_0 = K$. The standard formula for the posterior covariance matrix applied to the case where only one additional point is observed yields

$$K_{n+1}(x) = K_n(x) - K_n(x, x_{n+1}) K_n(x_{n+1}, x_{n+1})^{-1} K_n(x_{n+1}, x)$$

for all $n \geq 0$, from which the desired conclusion follows. □

Lemma 6.4. *For any fixed $h \in \mathcal{H}$, $n \in \mathbb{N}$ and $x \in \mathcal{X}$,*

$$\|h(x) - \mu_n(x)\|_2 \leq \|K_n(x)\|_2^{\frac{1}{2}} \|h\|_{\mathcal{H}},$$

where $\|K_n(x)\|_2$ denotes the spectral norm of the matrix $K_n(x)$.

References

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Billingsley, P. *Convergence of probability measures*. John Wiley & Sons, 2013.

Bottou, L. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

Bull, A. D. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct): 2879–2904, 2011.

Hansen, N. The CMA evolution strategy: A tutorial. *arXiv preprint arXiv:1604.00772*, 2016.

Kim, S., Pasupathy, R., and Henderson, S. G. A guide to sample average approximation. In *Handbook of simulation optimization*, pp. 207–243. Springer, 2015.

L'Ecuyer, P. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science*, 36(11):1364–1383, 1990.

Stein, M. L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Surjanovic, S. and Bingham, D. Langermann function, a. URL https://www.sfu.ca/~ssurjano/langer.html.

Surjanovic, S. and Bingham, D. Rosenbrock function, b. URL https://www.sfu.ca/~ssurjano/rosen.html.

Vazquez, E. and Bect, J. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095, 2010.

Williams, D. *Probability with Martingales*. Cambridge University Press, 1991.