Parallel Bayesian Global Optimization of Expensive Functions*

Jialei Wang[†1], Scott C. Clark[‡2], Eric Liu[§3], and Peter I. Frazier[¶1]

[1]School of Operations Research and Information Engineering, Cornell University
[2]SigOpt, 244 Kearny St, San Francisco, CA
[3]Yelp, Inc., 140 New Montgomery, San Francisco, CA

Abstract

We consider parallel global optimization of derivative-free expensive-to-evaluate functions, and propose an efficient method based on stochastic approximation for implementing a conceptual Bayesian optimization algorithm proposed by Ginsbourger et al. (2007). To accomplish this, we use infinitesimal perturbation analysis (IPA) to construct a stochastic gradient estimator and show that this estimator is unbiased. We also show that the stochastic gradient ascent algorithm using the constructed gradient estimator converges to a stationary point of the q-EI surface, and therefore, as the number of multiple starts of the gradient ascent algorithm and the number of steps for each start grow large, the one-step Bayes optimal set of points is recovered. We show in numerical experiments that our method for maximizing the q-EI is faster than methods based on closed-form evaluation using high-dimensional integration, when considering many parallel function evaluations, and is comparable in speed when considering few. We also show that the resulting one-step Bayes optimal algorithm for parallel global optimization finds high-quality solutions with fewer evaluations than a heuristic based on approximately maximizing the q-EI. A high-quality open source implementation of this algorithm is available in the open source Metrics Optimization Engine (MOE).

1 Introduction

We consider derivative-free global optimization of expensive functions, in which (1) our objective function is time-consuming to evaluate, limiting the number of function evaluations we can perform;

*Peter Frazier and Jialei Wang were partially supported by NSF CAREER CMMI-1254298, NSF CMMI-1536895, NSF IIS-1247696, AFOSR FA9550-12-1-0200, AFOSR FA9550-15-1-0038, and AFOSR FA9550-16-1-0046.
[†]jw865@cornell.edu
[‡]scott@sigopt.com
[§]eliu@yelp.com
[¶]pf98@cornell.edu

(2) evaluating the objective function provides only the value of the objective, and not the gradient or Hessian; and (3) we seek a global, rather than a local, optimum. Such problems arise when the objective function is evaluated by running a complex computer code (see, e.g., Sacks et al. 1989), performing a laboratory experiment, or building a prototype system to be evaluated in the real world. In this paper we assume our function evaluations are deterministic, i.e., free from noise.

Bayesian Global Optimization (BGO) methods are one class of methods for solving such problems. They were initially proposed by Kushner (1964), with early work pursued in Mockus et al. (1978) and Mockus (1989), and more recent work including improved algorithms (Boender and Kan 1987, Jones et al. 1998, Huang et al. 2006), convergence analysis (Calvin 1997, Calvin and Žilinskas 2002, Vazquez and Bect 2010), and allowing noisy function evaluations (Calvin et al. 2005, Villemonteix et al. 2009, Frazier et al. 2009, Huang et al. 2006).

The most well-known BGO method is Efficient Global Optimization (EGO) from Jones et al. (1998), which chooses each point at which to evaluate the expensive objective function in the "outer" expensive global optimization problem by solving an "inner" optimization problem: maximize the "expected improvement". Expected improvement is the value of information (Howard 1966) from a single function evaluation, and quantifies the benefit that this evaluation provides in terms of revealing a point with a better objective function value than previously known. If this is the final point that will be evaluated in the outer optimization problem, and if additional conditions are satisfied (the evaluations are free from noise, and the implementation decision, i.e., the solution that will be implemented in practice after the optimization is complete, is restricted to be a previously evaluated point), then the point with largest expected improvement is the Bayes-optimal point to evaluate, in the sense of providing the best possible average-case performance in the outer expensive global optimization problem (Frazier and Wang 2016).

Solving EGO's inner optimization problem is facilitated by an easy-to-compute and differentiate expression for the expected improvement in terms of the scalar normal cumulative distribution function. Fast evaluation of the expected improvement and its gradient make it possible in many applications to solve the inner optimization problem in significantly less time than the time required per evaluation of the expensive outer objective, which is critical to EGO's usefulness as an optimization algorithm.

The inner optimization problem at the heart of EGO and its objective, the expected improvement, was generalized by Ginsbourger et al. (2007) to the parallel setting, in which the expensive objective can be evaluated at several points simultaneously. This generalization, called the "multi-points expected improvement" or the q-EI, is consistent with the decision-theoretic derivation of expected improvement and quantifies the expected utility that will result from the evaluation of a *s*et of points. This work also provided an analytical formula for $q = 2$.

If this generalized inner optimization problem, which is to find the set of points to evaluate next that jointly maximize the q-EI, could be solved efficiently, then this would provide the one-step Bayes-optimal set of points to evaluate in the outer problem, and would create a one-step Bayes-optimal algorithm for global optimization of expensive functions able to fully utilize parallelism.

This generalized inner optimization problem is challenging, however, because unlike the scalar expected improvement used by EGO, the q-EI lacks an easy-to-compute and differentiate expression, and is calculable only through Monte Carlo simulation, high-dimensional numerical integration, or expressions involving high-dimensional multivariate normal cumulative distribution functions (CDFs). This significantly restricts the set of applications in which a naive implementation can solve the inner problem faster than a single evaluation of the outer optimization problem. Stymied by this difficulty, Ginsbourger et al. (2007) and later work (Chevalier and Ginsbourger 2013), propose heuristic methods that are *m*otivated by the one-step optimal algorithm of evaluating the set of points that jointly maximize the q-EI, but that do not actually achieve this gold standard.

Contributions The main contribution of this work is to provide a method that solves the inner optimization problem of maximizing the q-EI efficiently, creating a practical and broadly applicable one-step Bayes-optimal algorithm for parallel global optimization of expensive functions. To accomplish this we use infinitesimal perturbation analysis (IPA) (Ho 1987) to construct a stochastic gradient estimator of the gradient of the q-EI surface, and show that this estimator is unbiased, with a bounded second moment. Our method uses this estimator within a stochastic gradient ascent algorithm, which we show converges to the set of stationary points of the q-EI surface. We use multiple restarts to identify multiple stationary points, and then select the best stationary point found. As the number of restarts and the number of iterations of stochastic gradient ascent within each restart both grow large, the one-step optimal set of points to evaluate is recovered.

Our method can be implemented in both synchronous environments, in which function evaluations are performed in batches and finish at the same time, and asynchronous ones, in which a function evaluation may finish before others are done.

In addition to our methodological contribution, we have developed a high-quality open source software package, the "Metrics Optimization Engine (MOE)" (Clark et al. 2014), implementing our method for solving the inner optimization problem and the resulting algorithm for parallel global optimization of expensive functions. To further enhance computational speed, the implementation takes advantage of parallel computing and achieves 100X speedup over single-threaded computation when deployed on a graphical processing unit (GPU). This software package has been used by Yelp and Netflix to solve global optimization problems arising in their businesses (Clark 2014, Amatriain 2014). For the rest of the paper, we refer to our method as "MOE-qEI" because it is implemented in MOE.

We compare MOE-qEI against several benchmark methods. We show that MOE-qEI provides high-quality solutions to the outer optimization problem using fewer function evaluations than the heuristic CL-mix policy proposed by Chevalier and Ginsbourger (2013), which is motivated by the inner optimization problem. We also show that MOE-qEI provides a substantial parallel speedup over the single-threaded EGO algorithm, which is one-step optimal when parallel resources are unavailable. We also compare our simulation-based method for solving the inner optimization problem against methods based on exact evaluation of the q-EI from Chevalier and Ginsbourger (2013) and Marmin et al. (2015) (discussed in more detail below) and show that our simulation-based approach to solving the inner optimization problem provides solutions to both the inner and outer optimization problem that are comparable in quality and speed when q is small, and superior when q is large.

Related Work Developed independently and in parallel with our work is Chevalier and Ginsbourger (2013), which provides a closed-form formula for computing q-EI, and the book chapter Marmin et al. (2015), which provides a closed-form expression for its gradient. Both require multiple calls to high-dimensional multivariate normal CDFs. These expressions can be used within an existing continuous optimization algorithm to solve the inner optimization problem that we consider.

While attractive in that they provide closed-form expressions, calculating these expressions when q is even moderately large is slow and numerically challenging. This is because calculating the multivariate normal CDF in moderately large dimension is itself challenging, with state of the art methods relying on numerical integration or Monte Carlo sampling as described in Genz (1992). Indeed, the method for evaluating the q-EI from Chevalier and Ginsbourger (2013) requires q^2 evaluations of the $q-1$ dimensional multivariate normal CDF, and the method for evaluating its gradient requires $O(q^4)$ calls to multivariate normal CDFs with dimension ranging from $q-$

3 to q. In our numerical experiments, we demonstrate that our method for solving the inner optimization problem requires less computation time and parallelizes more easily than do these competing methods for $q > 4$, and performs comparably when q is smaller. We also demonstrate that MOE-qEI's improved performance in the inner optimization problem for $q > 4$ translates to improved performance in the outer optimization problem.

Other related work includes the previously proposed heuristic CL-mix from Chevalier and Ginsbourger (2013), which does not solve the inner maximization of q-EI, instead using an approximation. While solving the inner maximization of q-EI as we do makes it more expensive to compute the set of points to evaluate next, we show in our numerical experiments that it results in a substantial savings in the number of evaluations required to find a point with a desired quality. When function evaluations are expensive, this results in a substantial reduction in overall time to reach an approximately optimal solution.

In other related work on parallel Bayesian optimization, Frazier et al. (2011) and Xie et al. (2016) proposed a Bayesian optimization algorithm that evaluates pairs of points in parallel, and is one-step Bayes-optimal in the noisy setting under the assumption that one can only observe noisy function values for single points, or noisy function value differences between pairs of points. This algorithm, however, is limited to evaluating pairs of points, and does not extend to a higher level of parallelism.

There are also other non-Bayesian algorithms for derivative-free global optimization of expensive functions with parallel function evaluations from Dennis and Torczon (1991), Kennedy (2010) and Holland (1992). These are quite different in spirit from the algorithm we develop, not being derived from a decision-theoretic foundation.

Outline We begin in Section 2 by describing the mathematical setting in which Bayesian global optimization is performed, and then defining the q-EI and the one-step optimal algorithm. We construct our stochastic gradient estimator in Section 3.2, and use it within stochastic gradient ascent to define a one-step optimal method for parallel Bayesian global optimization in Section 3.3. Then in Section 4.1 we show that the constructed gradient estimator of the q-EI surface is unbiased under mild regularity conditions, and in Section 4.2 we provide convergence analysis of the stochastic gradient ascent algorithm. Finally, in Section 5 we present numerical experiments: we compare MOE-qEI against previously proposed heuristics from the literature; we demonstrate that MOE-qEI provides a speedup over single-threaded EGO; we show that MOE-qEI is more efficient than optimizing evaluations of the q-EI using closed-form formula provided in Chevalier and Ginsbourger (2013) when q is large; and we show that MOE-qEI computes the gradient of q-EI faster than evaluating the closed-form expression proposed in Marmin et al. (2015).

2 Problem formulation and background

In this section, we describe a decision-theoretic approach to Bayesian global optimization in parallel computing environments, previously proposed by Ginsbourger et al. (2007). This approach was considered to be purely conceptual as it contains a difficult-to-solve optimization sub-problem (our so-called "inner" optimization problem). In this section, we present this inner optimization problem as background, and present a novel method in the subsequent section that solves it efficiently.

2.1 Bayesian Global Optimization

Bayesian global optimization considers optimization of a function f with domain $\mathbb{A} \subseteq \mathbb{R}^d$. The overarching goal is to find an approximate solution to

$$\min_{\boldsymbol{x} \in \mathbb{A}} f(\boldsymbol{x}).$$

We suppose that evaluating f is expensive or time-consuming, and that these evaluations provide only the value of f at the evaluated point and not its gradient or Hessian. We assume that the function defining the domain $\mathbb{A}$ is easy to evaluate and that projections from $\mathbb{R}^d$ into the nearest point in $\mathbb{A}$ can be performed quickly.

Rather than focusing on asymptotic performance as the number of function evaluations grows large, we wish to find an algorithm that performs well, on average, given a limited budget of function evaluations. To formalize this, we model our prior beliefs on the function f with a Bayesian prior distribution, and we suppose that f was drawn at random by nature from this prior distribution, before any evaluations were performed. We then seek to develop an optimization algorithm that will perform well, on average, when applied to a function drawn at random in this way.

2.2 Gaussian process priors

For our Bayesian prior distribution on f, we adopt a Gaussian process prior (see Rasmussen and Williams 2006), which is specified by its mean function $\mu(\boldsymbol{x}) : \mathbb{A} \to \mathbb{R}$ and positive semi-definite covariance function $k(\boldsymbol{x}, \boldsymbol{x}') : \mathbb{A} \times \mathbb{A} \to \mathbb{R}$. We write the Gaussian process as

$$f \sim \mathcal{GP}(\mu, k).$$

Then for a collection of points $\boldsymbol{X} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$, the prior of f at $\boldsymbol{X}$ is

$$f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}), \tag{1}$$

where $\boldsymbol{\mu}_i^{(0)} = \mu(\boldsymbol{x}_i)$ and $\boldsymbol{\Sigma}_{ij}^{(0)} = k(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j \in \{1, \ldots, q\}$.

Our proposed method for choosing the points to evaluate next additionally requires that μ and k satisfy some mild regularity assumptions discussed below, but otherwise adds no additional requirements. In practice, μ and k are typically chosen using an empirical Bayes approach discussed in Brochu et al. (2010), in which first, a parameterized functional form for μ and k is assumed; second, a first stage of data is collected in which f is evaluated at points chosen according to a Latin hypercube or uniform design; and third, maximum likelihood estimates for the parameters specifying m and k are obtained. In some implementations (Jones et al. 1998, Snoek et al. 2012), these estimates are updated iteratively as more evaluations of f are obtained, which provides more accurate inference and tends to reduce the number of function evaluations required to find good solutions but increases the computational overhead per evaluation. We adopt this method below in our numerical experiments in Section 5. However, the specific contribution of this paper, a new method for solving an optimization sub-problem arising in the choice of design points, works with any choice of mean function μ and covariance matrix k, as long as they satisfy the mild regularity conditions discussed below.

In addition to the prior distribution specified in (1), we may also have some previously observed function values $y^{(i)} = f(\boldsymbol{x}^{(i)})$, for $i = 1, \ldots, n$. These might have been obtained through the previously mentioned first stage of sampling, running the second stage sampling method we are about to describe, or from some additional runs of the expensive objective function f performed by

another party outside of the control of our algorithm. If no additional function values are available, we set $n = 0$. We define notation $\boldsymbol{x}^{(1:n)} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)})$ and $y^{(1:n)} = (y^{(1)}, \ldots, y^{(n)})$. We require that all points in $\boldsymbol{x}^{(1:n)}$ be distinct.

We then combine these previously observed function values with our prior to obtain a posterior distribution on $f(\boldsymbol{X})$. This posterior distribution is still a multivariate normal (e.g., see Eq. (A.6) on pp. 200 in Rasmussen and Williams 2006)

$$f(\boldsymbol{X}) \mid \boldsymbol{X}, \boldsymbol{x}^{(1:n)}, y^{(1:n)} \sim \mathcal{N}(\boldsymbol{\mu}^{(n)}, \boldsymbol{\Sigma}^{(n)}), \tag{2}$$

with

$$\begin{aligned} \boldsymbol{\mu}^{(n)} &= \boldsymbol{\mu}^{(0)} + K\left(\boldsymbol{X}, \boldsymbol{x}^{(1:n)}\right) K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}\right)^{-1} \left(y^{(1:n)} - \mu(\boldsymbol{x}^{(1:n)})\right), \\ \boldsymbol{\Sigma}^{(n)} &= K\left(\boldsymbol{X}, \boldsymbol{X}\right) - K\left(\boldsymbol{X}, \boldsymbol{x}^{(1:n)}\right) K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}\right)^{-1} K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{X}\right), \end{aligned} \tag{3}$$

where $\mu(\boldsymbol{x}^{(1:n)})$ is the vector obtained by evaluating the prior mean function at each point in $\boldsymbol{x}^{(1:n)}$, $K\left(\boldsymbol{X}, \boldsymbol{x}^{(1:n)}\right)$ is a $q \times n$ matrix with $K\left(\boldsymbol{X}, \boldsymbol{x}^{(1:n)}\right)_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}^{(j)})$, and similarly for $K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{X}\right)$, $K\left(\boldsymbol{X}, \boldsymbol{X}\right)$ and $K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}\right)$.

2.3 Multi-points expected improvement (q-EI)

In a parallel computing environment, we wish to use this posterior distribution to choose the set of points to evaluate next. Ginsbourger et al. (2007) proposed making this choice using a decision-theoretic approach that considers the utility provided by evaluating a particular candidate set of points in terms of their ability to reveal better solutions than previously known. We review this decision-theoretic approach here, and then present a new algorithm for implementing this choice in the next section.

Let q be the number of function evaluations that we may perform in parallel, and let $\boldsymbol{X}$ be a candidate set of points that we are considering evaluating next. Let $f_n^\star = \min_{m \leq n} f(\boldsymbol{x}^{(m)})$ indicate the value of the best point evaluated, before beginning these q new function evaluations. The value of the best point evaluated after all q function evaluations are complete will be $\min\left(f_n^\star, \min_{i=1,\ldots,q} f(\boldsymbol{x}_i)\right)$. The difference between these two values (the values of the best point evaluated, before and after these q new function evaluations) is called the $i$mprovement, and is equal to $\left(f_n^\star - \min_{i=1,\ldots,q} f(\boldsymbol{x}_i)\right)^+$, where $a^+ = \max(a, 0)$ for $a \in \mathbb{R}$.

We then compute the expectation of this improvement over the joint probability distribution over $f(\boldsymbol{x}_i), i = 1, \ldots, q$, and we refer to this quantity as the $m$ulti-points expected improvement or q-EI from Ginsbourger et al. (2007). This multi-points expected improvement can be written as,

$$q\text{-EI}(\boldsymbol{X}) = \mathbb{E}_n\left[\left(f_n^\star - \min_{i=1,\ldots,q} f(\boldsymbol{x}_i)\right)^+\right], \tag{4}$$

where $\mathbb{E}_n[\cdot] := \mathbb{E}\left[\cdot | \boldsymbol{x}^{(1:n)}, y^{(1:n)}\right]$ is the expectation taken with respect to the posterior distribution.

Ginsbourger et al. (2007) then proposes evaluating next the set of points that maximize the multi-points expected improvement,

$$\operatorname*{argmax}_{\boldsymbol{X} \in H} q\text{-EI}(\boldsymbol{X}), \tag{5}$$

where $H = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q) : \boldsymbol{x}_i \in \mathbb{A}, ||\boldsymbol{x}_i - \boldsymbol{x}_j|| \geq r, ||\boldsymbol{x}_i - \boldsymbol{x}^{(\ell)}|| \geq r, i \neq j, 1 \leq i, j \leq q, 1 \leq \ell \leq n\}$.

This formulation generalizes Ginsbourger et al. (2007) slightly by allowing an optional require-

ment that new evaluation points be a distance of at least $r \geq 0$ from each other and previously evaluated points. Ginsbourger et al. (2007) implicitly set $r = 0$. Our convergence proof requires $r > 0$, which provides a compact feasible domain over which the stochastic gradient estimator has bounded variance. Setting a strictly positive r can also improve numerical stability in inference (see, e.g., Ababou et al. 1994), and evaluating a point extremely close to a previously evaluated point is typically unlikely to provide substantial improvement in the revealed objective value. In our experiments we set $r = 10^{-5}$.

In the special case $q = 1$, which occurs when we are operating without parallelism, the multi-points expected improvement reduces to the expected improvement (Mockus 1989, Jones et al. 1998), which can be evaluated in closed-form in terms of the normal density and CDF as discussed in Section 1. Ginsbourger et al. (2007) provided an analytical expression for q-EI when $q = 2$, but in the same paper the authors commented that computing q-EI for $q > 2$ involves expensive-to-compute q-dimensional Gaussian cumulative distribution functions relying on multivariate integral approximation, which makes solving (5) difficult. Ginsbourger (2009) writes "directly optimizing the q-EI becomes extremely expensive as q and d (the dimension of inputs) grow."

3 Algorithm

In this section we present a new algorithm for solving the inner optimization problem (5) of maximizing q-EI. This algorithm uses a novel estimator of the gradient of the q-EI presented in Section 3.2, used within a multistart stochastic gradient ascent framework as described in Section 3.3. We additionally generalize this technique from synchronous to asynchronous parallel optimization in Section 3.4. We begin by introducing additional notation used to describe our algorithm.

3.1 Notation

In this section we define additional notation to better support construction of the gradient estimator. Justified by (2), we write $f(\boldsymbol{X})$ as

$$f(\boldsymbol{X}) \overset{d}{=} \boldsymbol{\mu}(\boldsymbol{X}) + \boldsymbol{L}(\boldsymbol{X})\boldsymbol{Z}, \tag{6}$$

where $\boldsymbol{L}(\boldsymbol{X})$ is the lower triangular matrix obtained from the Cholesky decomposition of $\boldsymbol{\Sigma}^{(n)}$ in (2), $\boldsymbol{\mu}(\boldsymbol{X})$ is the posterior mean (identical to $\boldsymbol{\mu}^{(n)}$ in (2), but rewritten here to emphasize the dependence on X and de-emphasize the dependence on n), and $\boldsymbol{Z}$ is a multivariate standard normal random vector. We will also use the notation $\boldsymbol{\Sigma}(\boldsymbol{X})$ in place of $\boldsymbol{\Sigma}^{(n)}$ in our analysis.

By substituting (6) into (4), we have

$$q\text{-EI}(\boldsymbol{X}) = \mathbb{E}\left[\left(f_n^* - \min_{i=1,\ldots,q} \boldsymbol{e}_i \left[\boldsymbol{\mu}(\boldsymbol{X}) + \boldsymbol{L}(\boldsymbol{X})\boldsymbol{Z}\right]\right)^+\right], \tag{7}$$

where $\boldsymbol{e}_i$ is a unit vector in direction i and the expectation is over $\boldsymbol{Z}$. To make (7) even more compact, define a new vector $\boldsymbol{m}(\boldsymbol{X})$ and new matrix $\boldsymbol{C}(\boldsymbol{X})$,

$$\begin{aligned} \boldsymbol{m}(\boldsymbol{X})_i &= \begin{cases} f_n^* - \boldsymbol{\mu}(\boldsymbol{X})_i & \text{if } i > 0\ , \\ 0 & \text{if } i = 0\ , \end{cases} \\ \boldsymbol{C}(\boldsymbol{X})_{ij} &= \begin{cases} -\boldsymbol{L}(\boldsymbol{X})_{ij} & \text{if } i > 0\ , \\ 0 & \text{if } i = 0\ , \end{cases} \end{aligned} \tag{8}$$

and (7) becomes

$$q\text{-EI}(\boldsymbol{X}) = \mathbb{E}\left[\max_{i=0,\ldots,q} \boldsymbol{e}_i \left[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}\right]\right]. \tag{9}$$

3.2 Constructing the gradient estimator

We now construct our estimator of the gradient $\nabla q\text{-EI}(\boldsymbol{X})$. Let

$$h(\boldsymbol{X}, \boldsymbol{Z}) = \max_{i=0,\ldots,q} \boldsymbol{e}_i \left[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}\right]. \tag{10}$$

Then

$$\nabla q\text{-EI}(\boldsymbol{X}) = \nabla \mathbb{E} h(\boldsymbol{X}, \boldsymbol{Z}). \tag{11}$$

If gradient and expectation in (11) are interchangeable, the gradient would be

$$\nabla q\text{-EI}(\boldsymbol{X}) = \mathbb{E}\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}), \tag{12}$$

where

$$\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z}) = \begin{cases} \nabla h(\boldsymbol{X}, \boldsymbol{Z}) & \text{if } \nabla h(\boldsymbol{X}, \boldsymbol{Z}) \text{ exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

$\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z})$ can be computed using results on differentiation of the Cholesky decomposition from Smith (1995).

We use $\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{Z})$ as our estimator of the gradient ∇q-EI, and will discuss interchangeability of gradient and expectation, which implies unbiasedness of our gradient estimator, in Section 4.1. As will be discussed in Section 4.2, unbiasedness of the gradient estimator is one of the sufficient conditions for convergence of the stochastic gradient ascent algorithm proposed in Section 3.3.

3.3 Optimization of *q*-EI

Our stochastic gradient ascent algorithm begins with some initial point $\boldsymbol{X}_0 \in H$, and generates a sequence $\{\boldsymbol{X}_t : t = 1, 2, \ldots\}$ using

$$\boldsymbol{X}_{t+1} = \prod_H \left[\boldsymbol{X}_t + \epsilon_t \boldsymbol{G}(\boldsymbol{X}_t)\right], \tag{14}$$

where $\prod_H(\boldsymbol{X})$ denotes the closest point in H to $\boldsymbol{X}$, and if the closest point is not unique, a closest point such that the function $\prod_H(\cdot)$ is measurable. $\boldsymbol{G}(\boldsymbol{X}_t)$ is an estimate of the gradient of q-EI$(\cdot)$ at $\boldsymbol{X}_t$, obtained by averaging M replicates of our stochastic gradient estimator,

$$\boldsymbol{G}(\boldsymbol{X}_t) = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{Z}_{t,m}), \tag{15}$$

where $\{\boldsymbol{Z}_{t,m}$: m=1, ..., M$\}$ are i.i.d. samples generated from the multivariate standard normal distribution, $\boldsymbol{g}(\boldsymbol{X}_t, \boldsymbol{Z}_{t,m})$ is defined in (13). $\{\epsilon_t : t = 0, 1, \ldots\}$ is a stochastic gradient stepsize sequence (Kushner and Yin 2003), typically chosen to be equal to $\epsilon_t = \frac{a}{t^\gamma}$ for some scalar a and $\gamma \in (0, 1]$. Because we use Polyak-Ruppert averaging as described below, we set $\gamma < 1$. Analysis in Section 4 shows that, under certain mild conditions, this stochastic gradient algorithm converges almost surely to the set of stationary points.

After running T iterations of stochastic gradient ascent using (14), we obtain the sequence $\{\boldsymbol{X}_t :$

$t = 1, 2, \ldots, T\}$. From this sequence we extract the average $\overline{X}_T = \frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{X}_t$ and use it as an estimated stationary point. This Polyak-Ruppert averaging approach (Polyak 1990, Ruppert 1988) is more robust to misspecification of the stepsize sequence than using $\boldsymbol{X}_T$ directly.

To find the global maximum of the q-EI, we use multiple restarts of the algorithm from a set of starting points, drawn from a Latin hypercube design (McKay et al. 2000), to find multiple stationary points, and then use simulation to evaluate q-EI at these stationary points and select the point for which it is largest. For simplicity we present our approach using a fixed sample size N to perform this evaluation and selection (Step 8 in Algorithm 1 below) but one one could also use a more sophisticated ranking and selection algorithm with adaptive sample sizes (see, e.g., Kim and Nelson 2007), or evaluate q-EI using the closed-form formula in Chevalier and Ginsbourger (2013). We summarize our procedure for selecting the set of points to sample next, which we call MOE-qEI, in Algorithm 1.

Algorithm 1 MOE-qEI: Optimization of q-EI

Require: number of starting points R; stepsize constants a and γ; number of steps for one run of gradient ascent T; number of Monte Carlo samples for estimating the gradient M; number of Monte Carlo samples for estimating q-EI N.

1: Draw R starting points from a Latin hypercube design in H, $\boldsymbol{X}_{r,0}$ for $r = 1, \ldots, R$.
2: **for** $r = 1$ to R **do**
3: **for** $t = 0$ to $T - 1$ **do**
4: Compute $\boldsymbol{G}_t = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{g}(\boldsymbol{X}_{r,t}, \boldsymbol{Z}_{r,t,m})$ where $\boldsymbol{Z}_{r,t,m}$ is a vector of q i.i.d. samples drawn from the standard normal distribution.
5: Update solution using stochastic gradient ascent $\boldsymbol{X}_{r,t+1} = \prod_H \left[\boldsymbol{X}_{r,t} + \frac{a}{t^\gamma}\boldsymbol{G}_t\right]$.
6: **end for**
7: Compute the simple average of the solutions for $\boldsymbol{X}_{r,t}$, $\overline{\boldsymbol{X}}_{r,T} = \frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{X}_{r,t}$.
8: Estimate q-EI$(\overline{\boldsymbol{X}}_{r,T})$ using Monte Carlo simulation with N i.i.d. samples, and store the estimate as $\widehat{q\text{-EI}}_r$.
9: **end for**
10: **return** $\overline{\boldsymbol{X}}_{r',T}$ where $r' = \text{argmax}_{r=1,\ldots,R}\, \widehat{q\text{-EI}}_r$.

The MOE software package (Clark et al. 2014) implements Algorithm 1, and supplies the following additional optional fallback logic. If $\max_{r=1,\ldots,R} \widehat{q\text{-EI}}_r \leq \epsilon'$, so that multistart stochastic gradient ascent fails to find a point with estimated expected improvement better than ϵ', then it generates L additional solutions from a Latin Hypercube on H, estimates the q-EI at each of these using the same Monte Carlo approach as in Step 8, and selects the one with the largest estimated q-EI. This logic takes two additional parameters: a strictly positive real number ϵ' and an integer L. We turn this logic off in our experiments by setting $\epsilon' = 0$.

3.4 Asynchronous parallel optimization

So far we have assumed synchronous parallel optimization, in which we wait for all q points to finish before choosing a new set of points. However, in some applications, we may wish to generate a new partial batch of points to evaluate next while p points are still being evaluated, before we have their values. This is common in expensive computer evaluations, which do not necessarily finish at the same time.

We can extend Algorithm 1 to solve an extension of (5) proposed by Ginsbourger et al. (2010) for asynchronous parallel optimization: suppose parallelization allows a batch of q points to be evaluated simultaneously; the first p points are still under evaluation, while the remaining $q - p$ points have

finished evaluation and the resources used to evaluate them are free to evaluate new points. We let $\boldsymbol{X'} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be the first p points still under evaluation, and let $\boldsymbol{X} := (\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_q)$ be the $(q-p)$ points ready for new evaluations. Computation of q-EI for these q points remains the same as in (4), but we use an alternative notation, $q\text{-EI}(\boldsymbol{X'}, \boldsymbol{X})$, to explicitly indicate that $\boldsymbol{X'}$ are the points still being evaluated and $\boldsymbol{X}$ are the new points to evaluate. Keeping $\boldsymbol{X'}$ fixed, we optimize q-EI over $\boldsymbol{X}$ by solving this alternative problem

$$\underset{\boldsymbol{X} \in H'}{\operatorname{argmax}}\, q\text{-EI}(\boldsymbol{X'}, \boldsymbol{X}), \tag{16}$$

where $H' = \{(\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_q) : \boldsymbol{x}_i \in \mathbb{A}, ||\boldsymbol{x}_i - \boldsymbol{x}_j|| \geq r, ||\boldsymbol{x}_i - \boldsymbol{x}_k|| \geq r, ||\boldsymbol{x}_i - \boldsymbol{x}^{(m)}|| \geq r, i \neq j, p < i \leq q, p < j \leq q, 1 \leq k \leq p, 1 \leq m \leq n\}$. As we did in the algorithm for synchronous parallel optimization in Section 3.3, we estimate the gradient of the objective function with respect to $\boldsymbol{X}$, i.e., $\nabla_{\boldsymbol{X}}\, q\text{-EI}(\boldsymbol{X'}, \boldsymbol{X})$. The gradient estimator is essentially the same as that in Section 3.2, except that we only differentiate $h(\cdot, \cdot)$ with respect to $\boldsymbol{X}$. Then we proceed according to Algorithm 1.

In practice, one typically sets $p = q - 1$. This is because Bayesian optimization procedures are used most frequently when function evaluation times are large, and asynchronous computing environments typically have a time between evaluation completions that increases with the evaluation time. When this inter-completion time is large relative to the time required to solve (16), it is typically better to solve (16) each time an evaluation completes, i.e., to set $p = q - 1$, rather than waiting and letting a machine sit idle. If the time to perform a function evaluation is small enough, or if the computing environment is especially homogeneous, or if q is large (shortening the inter-completion time), then the time between completions might be smaller than the time to solve (16) and one might wish to set p strictly smaller than $q - 1$.

4 Theoretical analysis

In Section 3, when we constructed our gradient estimator and described the use of stochastic gradient ascent to optimize q-EI, we alluded to conditions under which this gradient estimator is unbiased and this stochastic gradient ascent algorithm converges to the set of stationary points of the q-EI surface. In this section, we describe these conditions and state these results.

4.1 Unbiasedness of the gradient estimator

We now state our main theorem showing unbiasedness of the gradient estimator. Proofs of all results including supporting lemmas are available as supplemental material.

Theorem 1 *If $\boldsymbol{m}(\boldsymbol{X})$ and $\boldsymbol{C}(\boldsymbol{X})$ are continuously differentiable in a neighborhood of $\boldsymbol{X}$, and $\boldsymbol{C}(\boldsymbol{X})$ has no duplicate rows, then $\nabla h(\boldsymbol{X}, \boldsymbol{Z})$ exists almost surely and*

$$\nabla \mathbb{E} h(\boldsymbol{X}, \boldsymbol{Z}) = \mathbb{E} \nabla h(\boldsymbol{X}, \boldsymbol{Z}).$$

Theorem 1 requires continuous differentiability of $\boldsymbol{C}(\boldsymbol{X})$, which may seem difficult to verify. However, using Smith (1995), which shows that mth-order differentiability of a symmetric and non-negative definite matrix implies mth-order differentiability of the lower triangular matrix obtained from its Cholesky factorization, $\boldsymbol{L}(\boldsymbol{X})$ and thus $\boldsymbol{C}(\boldsymbol{X})$ have the same order of differentiability as $\boldsymbol{\Sigma}^{(n)}$, whose order of differentiability can in turn be verified by examination of the prior covariance function $k(\cdot, \cdot)$. In addition, when $\boldsymbol{\Sigma}^{(n)}$ is positive definite, $\boldsymbol{C}(\boldsymbol{X})$ will not have duplicate rows. We will use these facts below in Corollary 1, after first discussing convergence, to provide easy-to-verify conditions under which unbiasedness and convergence to the set of stationary points hold.

4.2 Convergence analysis

In this section, we show almost sure convergence of our proposed stochastic gradient ascent algorithm. We assume that $\mathbb{A}$ is compact and can be written in the form $\mathbb{A} = \{\boldsymbol{x} : a_i'(\boldsymbol{x}) \leq 0, i = 1, \ldots, m'\} \subseteq \mathbb{R}^d$, where $a_i'(\cdot)$ is any real-valued constraint function. Then H can be written in a form more convenient for analysis,

$$H = \{\boldsymbol{X} : a_i(\boldsymbol{X}) \leq 0, i = 1, \ldots, m\} \subseteq \mathbb{R}^{d \times q},$$

where $a_{(i-1)q+j}(\boldsymbol{X}) = a_i'(\boldsymbol{x}_j)$ with $\boldsymbol{x}_j$ being the jth point in $\boldsymbol{X}$, and $a_i(\boldsymbol{X})$ for $i > m'q$ encodes the constraints $||\boldsymbol{x}_i - \boldsymbol{x}_j|| \geq r$ and $||\boldsymbol{x}_i - \boldsymbol{x}^{(\ell)}|| \geq r$ present in (5).

The following theorem shows that Algorithm 1 converges to the set of stationary points under conditions that include those of Theorem 1. The proof is available as supplemental material.

Theorem 2 *Suppose the following assumptions hold,*

1. *$a_i(\cdot), i = 1, \ldots, m$ are continuously differentiable.*
2. *$\epsilon_t \to 0$ for $t \geq 0$; $\sum_{t=1}^{\infty} \epsilon_t = \infty$ and $\sum_{t=0}^{\infty} \epsilon_t^2 < \infty$.*
3. *$\forall \boldsymbol{X} \in H$, $\boldsymbol{\mu}(\boldsymbol{X})$ and $\boldsymbol{\Sigma}(\boldsymbol{X})$ are twice continuously differentiable and $\boldsymbol{\Sigma}(\boldsymbol{X})$ is positive definite.*

Then the sequence $\{\boldsymbol{X}_t : t = 0, 1, \ldots\}$ and its Polyak-Ruppert average $\{\overline{\boldsymbol{X}}_t : t = 0, 1, \ldots\}$ generated by algorithm (14) converges almost surely to a connected set of stationary points of the q-EI surface.

The following corollary of Theorem 2 uses conditions that can be more easily checked prior to running MOE-qEI. It requires that the sampled points are distinct, which can be made true by dropping duplicate samples. Since function evaluations are deterministic, no information is lost in doing so.

Corollary 1 *If the sampled points $\boldsymbol{x}^{(1:n)}$ are distinct and*

1. *the prior covariance function k is positive definite and twice differentiable,*
2. *the prior mean function μ is twice differentiable,*
3. *conditions 1 and 2 in Theorem 2 are met,*

then $\boldsymbol{X}_t$ and its Polyak-Ruppert average $\overline{\boldsymbol{X}}_t$ converge to a connected set of stationary points.

Proof 1 *Proof to Corollary 1 Since $\boldsymbol{x}^{(1:n)}$ are distinct, and $\boldsymbol{X} \in H$, and the prior covariance function is positive definite and twice continuously differentiable, then $K\left(\boldsymbol{X}, \boldsymbol{x}^{(1:n)}\right)$, $K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{X}\right)$, $K\left(\boldsymbol{X}, \boldsymbol{X}\right)$ and $K\left(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}\right)$ in (3) are all positive definite and twice continuously differentiable. Since the prior mean function is also twice continuously differentiable, it follows that $\boldsymbol{\mu}^{(n)} = \boldsymbol{\mu}(\boldsymbol{X})$ and $\boldsymbol{\Sigma}^{(n)} = \boldsymbol{\Sigma}(\boldsymbol{X})$ defined in (3) are twice continuously differentiable, and in addition, $\boldsymbol{\Sigma}^{(n)}$ is positive definite. Thus the conditions of Theorem 2 are verified, and its conclusion holds.*

5 Numerical results

In this section, we present numerical experiments demonstrating the performance of MOE-qEI. The implementation of MOE-qEI follows Algorithm 1, and is available in the open source software package "MOE" (Clark et al. 2014).

We first discuss the choice of constants in Algorithm 1: R, T, M, N, γ, and a.

1. Number of starting points, R: this should be larger and of the same order as the number of equivalence classes of stationary points of the q-EI surface, where we identify a set of stationary points as in the same class if they can be obtained from each other by permuting $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q$. ($q$-EI is symmetric and such permutations do not change its value.) However, we do not know the number of such equivalence classes, and their number tends to grow with n as the surface grows more modes. Setting R larger increases our chance of finding the global maximum but increases computation. In our numerical experiments, we set $R = n$ to capture this trade-off between runtime and solution quality. As a diagnostic, one can check whether R is large enough by checking the number of unique solutions we obtain; if we obtain the same solution repeatedly from multiple restarts, this suggests R is large enough.

2. Number of steps in stochastic gradient ascent, T, and stepsize sequence parameters a and γ: For simplicity, we set $a = 1$. We set $\gamma = 0.7$, which is significantly below 1, to ensure that the stepsize sequence decreases slowly, as is recommended when using Polyak-Ruppert averaging. We then plotted the q-EI and norm of the gradient from stochastic gradient ascent versus t for a few sample problems. Finding that convergence occurred well before the 100th iterate, we set $T = 100$. As a diagnostic, one may also assess convergence by evaluating the gradient using a large number of Monte Carlo samples at the final iterate T and comparing its norm to 0.

3. Number of Monte Carlo samples M: this determines the accuracy of the gradient estimate and therefore affects stochastic gradient ascent's convergence. We set $M = 1000$ and as discussed in Section 3.3 performed an experiment to justify this setting. While we ran our experiments on a CPU except where otherwise stated to ensure a fair comparison with other competing algorithms, for which a GPU implementation is not available, the "MOE" software package provides a GPU implementation that can be used to increase the amount of parallelism used in MOE-qEI. When using the GPU implementation, we recommend setting $M = 10^6$ because the GPU's parallelism makes averaging a large number of independent replicates fast, and the reduction in noise reduces the number of iterates needed for convergence by stochastic gradient ascent.

4. Number of Monte Carlo samples for estimating q-EI, N: we estimate the q-EI at a limiting solution only once for each restart, i.e., R times, and so setting N large introduces little computational overhead. We set $N = 10^6$ to ensure an essentially noise-free selection of the best of the limiting solutions, and we assess this choice by examining the standard error of our estimates of the q-EI.

For the outer optimization of the objective function, we begin with a small dataset, typically sampled using a Latin hypercube design, to train the Gaussian Process model described in Section 2.2. In our numerical experiments, we use $\mu = 0$ and a squared exponential kernel k whose hyperparameters are estimated using an empirical Bayes approach: we set them to the values that maximize the log marginal likelihood of the observed data. With the trained Gaussian Process model, we perform the inner optimization of MOE-qEI described in Algorithm 1 to find the batch of points to evaluate, and after evaluating them we update the hyperparameters as well as the Gaussian Process model. We repeat this process over a number of iterations and report the best solution found in each iteration.

Noise-free function evaluations may often lead to ill-conditioned covariance matrices $K(\cdot, \cdot)$ in (2). To resolve this problem, we adopt a standard trick from Gaussian process regression (Rasmussen and Williams 2006, Section 3.4.3): we manually impose a small amount of noise $\sim \mathcal{N}(0, \sigma^2)$ where

$\sigma^2 = 10^{-4}$ and use Gaussian Process regression designed for noisy settings, which is almost identical to (2) except that $K(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)})$ is replaced by $K(\boldsymbol{x}^{(1:n)}, \boldsymbol{x}^{(1:n)}) + \sigma^2 I_n$ where I_n is the identity matrix (Rasmussen and Williams 2006, Section 2.2).

5.1 Comparison on the outer optimization problem

Constant Liar is a heuristic algorithm motivated by (9) proposed by Ginsbourger et al. (2007), which uses a greedy approach to iteratively construct a batch of q points. At each iteration of this greedy approach, the heuristic uses the sequential EGO algorithm to find a point that maximizes the expected improvement. However, since the posterior used by EGO depends on the current batch of points, which have not yet been evaluated, Constant Liar imposes a heuristic response (the "liar") at this point, and updates the Gaussian Process model with this "liar" value. The algorithm stops when q points are added, and reports the batch for function evaluation.

There are three variants of Constant Liar (CL), which use three different strategies for choosing the liar value: CL-min sets the liar value to the minimum response observed so far; CL-max sets it to the maximum response observed so far; and CL-mix is a hybrid of the two, computing one set of points using CL-min, another set of points using CL-max, and sampling the set that has the higher q-EI. Among the three methods, CL-mix was shown by Chevalier and Ginsbourger (2013) to have the best overall performance, and therefore we compare MOE-qEI against CL-mix.

We ran MOE-qEI and CL-mix on a range of standard test functions for global optimization (Jamil and Yang 2013): 2-dimensional Branin2; 3-dimensional Hartmann3; 5-dimensional Ackley5; and 6-dimensional Hartmann6. In the experiment, we first draw $(2d + 2)$ points in the domain using a Latin hypercube design, where d is the dimension of the objective function, and fit a Gaussian Process model using the initial points. Thereafter, we let MOE-qEI and CL-mix optimize over the test functions and report for each iteration the regret as regret $= f^* -$ best solution so far for each iteration, where we note that each of the problems considered is a minimization problem. We repeat the experiment 100 times using different initial sets of points, and report the average performance of both algorithms in Figure 1. The result shows that MOE-qEI consistently finds better solutions than the heuristic method on all four test functions.

Next, we compare MOE-qEI and CL-MIX at different levels of parallelism using the same experimental setup as above. The sequential EGO algorithm makes the same decisions as MOE-qEI when $q = 1$ and so this may also be seen as a comparison against EGO. Figure 2 shows that MOE-qEI achieves significant speedup over EGO as q grows, indicating substantial potential time saving using parallelization and MOE-qEI in Bayesian optimization tasks.

5.2 Comparison on the inner optimization problem

Chevalier and Ginsbourger (2013) provided a closed-form formula for q-EI and argued that it computes q-EI "very fast for reasonably low values of q (typically less than 10)". The closed-form formula is provided as follows for reference, modified to use the notation in this paper. Recall (2) and let $f(\boldsymbol{X}) = (Y_1, \ldots, Y_q)$ be a random vector with mean $\boldsymbol{\mu}^{(n)}$ and covariance matrix $\boldsymbol{\Sigma}^{(n)}$. For $k \in \{1, \ldots, q\}$ consider the vectors $\boldsymbol{Z}^k := (Z_1^k, \ldots, Z_q^k)$ defined as follows:

$$
\begin{aligned}
Z_j^k &:= Y_k - Y_j, j \neq k, \\
Z_k^k &:= Y_k.
\end{aligned}
$$

Let $\boldsymbol{m}^k$ and Σ^k denote the mean and covariance matrix of $\boldsymbol{Z}^k$, and define the vector $\boldsymbol{b}^k \in \mathbb{R}^q$ by

$b_k^k = f_n^*$ and $b_j^k = 0$ if $j \neq k$. Then the closed-form formula is

$$q\text{-EI}(\boldsymbol{X}) = \sum_{k=1}^{q} \left((f_n^* - \mu_k^{(n)}) \Phi_q(\boldsymbol{b}^k - \boldsymbol{m}^k, \Sigma^k) + \sum_{i=1}^{q} \Sigma_{ik}^k \phi_{m_i^k, \Sigma_{ii}^k}(b_i^{(k)}) \Phi_{q-1}(\boldsymbol{c}_{\cdot i}^k, \Sigma_{\cdot i}^k) \right), \tag{17}$$

where $\boldsymbol{c}^k$ is as defined in Chevalier and Ginsbourger (2013).

This formula requires q calls of the q-dimensional multivariate normal CDF ($\Phi_q(\cdot, \cdot)$), and q^2 calls of the $q-1$ dimensional multivariate normal CDF ($\Phi_{q-1}(\cdot, \cdot)$). Since computing multivariate normal CDFs, which are often implemented with numerical integration or Monte Carlo sampling (Genz 1992), is expensive to evaluate even for moderate q, calculating this analytically formula quickly becomes slow and numerically challenging as q grows.

While Chevalier and Ginsbourger (2013) did not propose using this closed-form formula to solve the inner optimization problem (5), one can adapt it to this purpose by using it within any derivative-free optimization method. We implemented this approach in the MOE package, where we use the L-BFGS (Liu and Nocedal 1989) solver from SciPy (Jones et al. 2001) as the derivative free optimization solver. We call this approach "Benchmark 1".

We compare MOE-qEI, CL-mix, and Benchmark 1 in solving the inner optimization problem, in terms of both solution quality and runtime, as shown in Figure 3 and 4. Without surprise, MOE-qEI achieves the best solution quality among the three, and its running time is almost comparable to CL-mix, which is expected to be the fastest approach because it sacrifices solution quality for speed. We ran Benchmark 1 with q going only up to 4 because its runtime goes up drastically with q. MOE-qEI's runtime scales well as q grows, making it feasible to run in applications with high parallelism. To our surprise, CL-mix achieves competitive solution quality against Benchmark 1, using only a fraction of Benchmark 1's runtime. Therefore, despite the promise of using the closed-form formula for q-EI to fully solve the inner optimization problem, this formula's long runtime and the slow convergence of L-BFGS due to lack of derivative information make Benchmark 1 a less favorable option than CL-mix in practice.

Figure 3 also includes a method called "MOE high-MC". This method runs MOE-qEI on GPU with the number of Monte Carlo samples M for the gradient estimator set to 10^7, much higher than the default setting of 1000. As shown in the figure, the solution quality for "MOE high-MC" is the same as that of MOE-qEI, which confirms that $M = 1000$ is sufficiently large.

5.3 Comparison on evaluation of ∇q-EI

A recently published book chapter Marmin et al. (2015), developed independently and in parallel to this work, proposed a method for computing ∇q-EI using a closed-form formula derived from (17), and then proposed to use this formula inside a gradient-based optimization routine to solve (5). The formula is complex and therefore we do not reproduce it here.

This formula faces even more severe computational challenges than (17); indeed, it requires $O(q^4)$ calls to multivariate normal CDFs with dimension between $(q-3)$ and q. Because computing high-dimensional multivariate normal CDFs is itself challenging, this closed-form evaluation becomes extremely time-consuming.

MOE-qEI's Monte-Carlo based approach to evaluating ∇q-EI offers three advantages over using the closed-form formula: first, numerical experiments below suggest that computation scales better with q; second, it can be easily parallelized, with significant speedups possible through parallel computing on graphical processing units (GPUs), as is implemented within the MOE library; third, by using a small number of replications to make each iteration run quickly, and by using it within a stochastic gradient ascent algorithm that averages noisy gradient information intelligently across

iterations, we may more intelligently allocate effort across iterations, only spending substantial effort to estimate gradients accurately late in the process of finding a local maximum.

We first show that computation of exact gradients using our gradient estimator with many replications on a GPU scales better with q through numerical experiments. We compare with closed-form gradient evaluation on a CPU as implemented in the "DiceOptim" package (Ginsbourger et al. 2015) and call it "Benchmark 2". We computed ∇q-EI at 200 randomly chosen points from a 2-dimensional design space to obtain a 95% confidence interval for the average computation time. To make the gradient evaluation in MOE-qEI close to exact, we increased the number of Monte Carlo samples used in the gradient estimator to 10^7, which ensures that the variance of each component of the gradient is on the order of 10^{-10} or smaller for all q we consider in our experiments. Given the large number of Monte Carlo samples, we use the GPU option in the MOE package to speed up computation. This GPU implementation is made possible by the trivial parallelism supported by our Monte-Carlo-based gradient estimator, while a massively parallel GPU implementation of closed-form gradient evaluation would be more challenging.

Figure 5 shows that computational time for Benchmark 2 increases quickly as q grows, but increases slowly for MOE-qEI's Monte Carlo estimator, with this Monte Carlo estimator being faster when $q \geq 4$. This difference in performance arises because gradient estimation in MOE-qEI focuses Monte Carlo effort on calculating a single high-dimensional integral, while the closed-form formula decomposes this high-dimensional integral of interest into a collection of other high-dimensional integrals that are almost equally difficult to compute, and the size of this collection grows with q.

We may reduce the number of Monte Carlo samples used by MOE-qEI substantially while still providing high-accuracy estimates: if we reduce M from 10^7 to 10^4, the variance of each component of the gradient remains below 10^{-7}. Since the GPU implementation provides a roughly 100x to 1000x speedup, the CPU-only implementation of our stochastic gradient estimator with this reduced value of M has run-time comparable to or better than the GPU-based results pictured in Figure 5, showing that even without the hardware advantage offered by a GPU our stochastic gradient estimator provides high-accuracy estimates faster than Marmin et al. (2015) for $q \geq 4$.

Moreover, because stochastic gradient ascent is tolerant to noisy gradients, we may obtain additional speed improvements by reducing the number of Monte Carlo samples even further. Using fewer Monte Carlo samples in each iteration has the potential to increase efficiency by only putting effort toward estimating the gradient precisely when we are close to the stationary point, which stochastic gradient ascent performs automatically through its decreasing stepsize sequence. Thus, our stochastic gradient estimator is both faster when using a large number of samples to produce essentially exact estimates, and it offers more flexibility in its ability to produce inexact estimates at low computational cost.

6 Conclusions

We proposed an efficient method based on stochastic approximation for implementing a conceptual parallel Bayesian global optimization algorithm proposed by Ginsbourger et al. (2007). To accomplish this, we used infinitessimal perturbation analysis (IPA) to construct a stochastic gradient estimator and showed that this estimator is unbiased. We also provided convergence analysis of the stochastic gradient ascent algorithm with the constructed gradient estimator. Through numerical experiments, we demonstrate that our method outperforms the existing state-of-the-art approximation methods.

References

Ababou, R., Bagtzoglou, A. C., and Wood, E. F. (1994). On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26(1):99–133.

Amatriain, X. (2014). 10 lessons learned from building ml systems. https://www.youtube.com/watch?v=WdzWPuazLA8. Recording of presentation from MLconf 2014, Accessed: 2015-11-26.

Boender, C. G. E. and Kan, A. R. (1987). Bayesian stopping rules for multistart global optimization methods. *Mathematical Programming*, 37(1):59–80.

Brochu, E., Brochu, T., and de Freitas, N. (2010). A bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 103–112. Eurographics Association.

Calvin, J. et al. (2005). One-dimensional global optimization for observations with noise. *Computers & Mathematics with Applications*, 50(1):157–169.

Calvin, J. M. (1997). Average performance of a class of adaptive algorithms for global optimization. *The Annals of Applied Probability*, 7(3):711–730.

Calvin, J. M. and Žilinskas, A. (2002). One-dimensional global optimization based on statistical models. In *Stochastic and Global Optimization*, pages 49–63. Springer.

Chevalier, C. and Ginsbourger, D. (2013). Fast computation of the multi-points expected improvement with applications in batch selection. In *Learning and Intelligent Optimization*, pages 59–69. Springer.

Clark, S. (2014). Introducing "MOE": metric optimization engine; a new open source, machine learning service for optimal experiment design. http://engineeringblog.yelp.com/2014/07/introducing-moe-metric-optimization-engine-a-new-open-source-machine-learning-service-for-optimal-ex.html. Accessed: 2015-11-26.

Clark, S. C., Liu, E., Frazier, P. I., Wang, J., Oktay, D., and Vesdapunt, N. (2014). Metrics optimization engine. http://yelp.github.io/MOE/. Accessed: 2017-09-17.

Dennis, Jr, J. E. and Torczon, V. (1991). Direct search methods on parallel machines. *SIAM Journal on Optimization*, 1(4):448–474.

Frazier, P., Powell, W., and Dayanik, S. (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613.

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

Frazier, P. I., Xie, J., and Chick, S. E. (2011). Value of information methods for pairwise sampling with correlations. In *Proceedings of the Winter Simulation Conference*, pages 3979–3991. Winter Simulation Conference.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Ginsbourger, D. (2009). Two advances in gaussian process-based prediction and optimization for computer experiments. In *MASCOT09 Meeting*.

Ginsbourger, D., Le Riche, R., and Carraro, L. (2007). A multi-points criterion for deterministic parallel global optimization based on kriging. In *NCP07*.

Ginsbourger, D., Le Riche, R., and Carraro, L. (2010). Kriging is well-suited to parallelize optimization. *Computational Intelligence in Expensive Optimization Problems*, 2:131–162.

Ginsbourger, D., Picheny, V., Roustant, O., et al. (2015). Diceoptim: Kriging-based optimization for computer experiments. https://cran.r-project.org/web/packages/DiceOptim/index.html. Accessed: 2016-02-13.

Glasserman, P. (1991). *Gradient estimation via perturbation analysis*. Springer Science & Business Media.

Ho, Y.-C. (1987). Performance evaluation and perturbation analysis of discrete event dynamic systems. *IEEE Transactions on Automatic Control*, 32(7):563–572.

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

Howard, R. A. (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26.

Huang, D., Allen, T. T., Notz, W. I., and Zeng, N. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. *Journal of global optimization*, 34(3):441–466.

Jamil, M. and Yang, X.-S. (2013). A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation*, 4(2):150–194.

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

Jones, E., Oliphant, T., Peterson, P., et al. (2001). SciPy: Open source scientific tools for Python. [Online; accessed 2014-12-01].

Kennedy, J. (2010). Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer.

Kim, S.-H. and Nelson, B. L. (2007). Recent advances in ranking and selection. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 162–172. IEEE Press.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Kushner, H. J. (1964). A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Fluids Engineering*, 86(1):97–106.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Marmin, S., Chevalier, C., and Ginsbourger, D. (2015). Differentiating the multipoint expected improvement for optimal batch design. In *Machine Learning, Optimization, and Big Data*, pages 37–48. Springer.

McKay, M. D., Beckman, R. J., and Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61.

Mockus, J. (1989). *The bayesian approach to local optimization*. Springer.

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

Polyak, B. T. (1990). New stochastic approximation type procedures. *Automat. i Telemekh*, 7(98-107):2.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Ruppert, D. (1988). Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Smith, S. P. (1995). Differentiation of the cholesky algorithm. *Journal of Computational and Graphical Statistics*, 4(2):134–147.

Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

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

Villemonteix, J., Vazquez, E., and Walter, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534.

Xie, J., Frazier, P. I., and Chick, S. E. (2016). Bayesian optimization via simulation with pairwise sampling and correlated prior beliefs. *Operations Research*, 64(2):542–559.

Appendix: proofs of results in the main paper

Lemma 1 *Suppose functions $\lambda_i(x) : \mathbb{R} \mapsto \mathbb{R}$, $i = 1, \ldots, m$ are continuously differentiable on a compact interval $\mathcal{X}$. Let $\Lambda(x) = \max_{i=1}^m \lambda_i(x)$, and $\widetilde{\mathcal{X}}$ be the set of points where $\Lambda(x)$ fails to be differentiable. Then $\widetilde{\mathcal{X}}$ is countable.*

Proof 2 *Proof of Lemma 1. We first consider $m = 2$, and later extend to $m > 2$. Let x_0 be a point of non-differentiability of $\Lambda(x)$. Then $\lambda_1(x_0) = \lambda_2(x_0)$ and $\lambda_1'(x_0) \neq \lambda_2'(x_0)$. (If the first condition were not true, and suppose $\lambda_1(x_0) > \lambda_2(x_0)$, then continuity of λ_1 and λ_2 would imply $\Lambda(x) = \lambda_1(x)$ in an open neighborhood of x_0. If the first condition were true but not the second, then $\Lambda'(x_0) = \lambda_1'(x_0) = \lambda_2'(x_0)$.)*

By continuity of λ_1' and λ_2', $\exists \delta > 0$ such that $\lambda_1'(x) > \lambda_2'(x)$ for all $x \in (x_0 - \delta, x_0 + \delta)$. Therefore $\Lambda(x) = \lambda_1(x)$ at $x \in (x_0, x_0 + \delta)$ and $\Lambda(x) = \lambda_2(x)$ at $x \in (x_0 - \delta, x_0)$. Thus Λ is differentiable on $(x_0 - \delta, x_0 + \delta) \backslash \{x_0\}$.

Let $n(x_0)$ be the smallest integer $n \geq 1$ such that Λ is differentiable on $(x_0 - 1/n, x_0 + 1/n) \backslash \{x_0\}$ and let $D(n)$ be the set of non-differentiable points x such that $n(x) = n$. In an interval of length L, there can be at most $Ln + 1$ points in $D(n)$. Hence the set of all non-differentiable points $\widetilde{\mathcal{X}} = \cup_{n=1}^{\infty} D(n)$ is countable.

Now let $m > 2$. We show that all points of discontinuity of $\Lambda(x)$ are also points of discontinuity of $\max(\lambda_i(x), \lambda_j(x))$ for at least one pair of i, j. Let $S(x) = \operatorname{argmax}_i \lambda_i(x)$. Using Taylor's theorem, for $\Delta \in \mathbb{R}$,

$$\lambda_i(x + \Delta) = \lambda_i(x) + \lambda_i'(x)\Delta + h_i(x + \Delta)\Delta,$$

where $h_i(\cdot)$ is a function such that $\lim_{\Delta \to 0} h_i(x + \Delta) = 0$. We write the left and right derivative of Λ at x as

$$\begin{aligned}
\lim_{\Delta \to 0^+} \frac{\Lambda(x + \Delta) - \Lambda(x)}{\Delta} &= \lim_{\Delta \to 0^+} \frac{\max\{\lambda_i(x) + \lambda_i'(x)\Delta + h_i(x + \Delta)\Delta\} - \lambda_{i^*}(x)}{\Delta}, \\
&= \lim_{\Delta \to 0^+} \max\{\frac{\lambda_i(x) - \lambda_{i^*}(x)}{\Delta} + \lambda_i'(x) + h_i(x)\}, \\
&= \max\{\lambda_i'(x) : i \in S\},
\end{aligned}$$

and

$$\begin{aligned}
\lim_{\Delta \to 0^+} \frac{\Lambda(x) - \Lambda(x - \Delta)}{\Delta} &= \lim_{\Delta \to 0^+} \frac{\lambda_{i^*}(x) - \max\{\lambda_i(x) - \lambda_i'(x)\Delta - h_i(x - \Delta)\Delta\}}{\Delta}, \\
&= \lim_{\Delta \to 0^+} \min\{\frac{\lambda_{i^*}(x) - \lambda_i(x)}{\Delta} + \lambda_i'(x) + h_i(x)\}, \\
&= \min\{\lambda_i'(x) : i \in S\}.
\end{aligned}$$

If the left and right derivative are equal, Λ is differentiable at x. If not, let $i^+ \in \operatorname{argmax}\{\lambda_i'(x) : i \in S\}$ and $i^- \in \operatorname{argmin}\{\lambda_i'(x) : i \in S\}$. Then $\max(\lambda_{i^+}(x), \lambda_{i^-}(x))$ fails to be differentiable at x.

Thus the non-differentiable points of $\Lambda'(x)$ are a subset of the union of the non-differentiable points of $\max(\lambda_i(x), \lambda_j(x))'$ over all i, j, and so it is a subset of a finite union of countable sets, which is countable.

Lemma 2 *If $\boldsymbol{m}(\boldsymbol{X})$ and $\boldsymbol{C}(\boldsymbol{X})$ are differentiable in a neighborhood of $\boldsymbol{X}$, and there are no duplicated rows in $\boldsymbol{C}(\boldsymbol{X})$, then $P\left(\boldsymbol{e}_i[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}] = \boldsymbol{e}_j[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}]\right) = 0$ for any $i \neq j$, and $\nabla h(\boldsymbol{X}, \boldsymbol{Z})$ exists almost surely for any $\boldsymbol{X}$.*

Proof 3 *Proof of Lemma 2. Observe that* $h(\boldsymbol{X},\boldsymbol{Z}) = \boldsymbol{e}_{I^*}\left[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}\right]$*, where* $I^* \in \operatorname{argmax}_{i=0,\ldots,q} \boldsymbol{e}_i\left[\boldsymbol{m}(\boldsymbol{X}) + \right.$ $\mathcal{S}$*.* $\nabla h(\boldsymbol{X},\boldsymbol{Z})$ *can fail to exist only if* $\exists I_1, I_2 \in \mathcal{S}$ *with* $\boldsymbol{e}_{I_1}(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial x_{ik}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial x_{ik}}\boldsymbol{Z}) \neq \boldsymbol{e}_{I_2}(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial x_{ik}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial x_{ik}}\boldsymbol{Z})$*. Thus,*

$$\begin{aligned} P(\nabla h(\boldsymbol{X},\boldsymbol{Z}) \text{ does not exist}) &\leq P(|\mathcal{S}| \geq 2), \\ &\leq \frac{1}{2}\sum_{i\neq j} P\left(\boldsymbol{e}_i[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}] = \boldsymbol{e}_j[\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}]\right), \\ &= \frac{1}{2}\sum_{i\neq j} P\left(\left(\boldsymbol{C}(\boldsymbol{X})_{i\cdot} - \boldsymbol{C}(\boldsymbol{X})_{j\cdot}\right)\boldsymbol{Z} = m(\boldsymbol{X})_j - m(\boldsymbol{X})_i\right), \end{aligned}$$

where $\boldsymbol{C}_{i\cdot}(\boldsymbol{X})$ *is the* i*th row of* $\boldsymbol{C}(\boldsymbol{X})$*. Since* $\boldsymbol{C}_{i\cdot}(\boldsymbol{X}) \neq \boldsymbol{C}_{j\cdot}(\boldsymbol{X})$*,* $\{\boldsymbol{Z} : \left(\boldsymbol{C}_{i\cdot}(\boldsymbol{X}) - \boldsymbol{C}_{j\cdot}(\boldsymbol{X})\right)\boldsymbol{Z} = m_j(\boldsymbol{X}) - m_i(\boldsymbol{X})\}$ *is subspace of* $\mathbb{R}^q$ *with dimension smaller than* q*, and*

$$P\left(\left(\boldsymbol{C}_{i\cdot}(\boldsymbol{X}) - \boldsymbol{C}_{j\cdot}(\boldsymbol{X})\right)\boldsymbol{Z} = m_j(\boldsymbol{X}) - m_i(\boldsymbol{X})\right) = 0 \quad \forall i \neq j.$$

Hence $P(\nabla h(\boldsymbol{X},\boldsymbol{Z}) \text{ does not exist}) = 0$*.*

Proof 4 *Proof of Theorem 1. Without loss of generality, we consider the partial derivative with respect to the* k*th component of the* m*th point in* $\boldsymbol{X}$*, that is,* $\frac{\partial h(\boldsymbol{X},\boldsymbol{Z})}{\partial X_{mk}}$*. We use the following result in Theorem 1.2. from Glasserman (1991), restated here for convenience:*

Suppose the following conditions (1), (2), (3) and (4) hold on a compact interval Θ*, then* $\mathbb{E}\left[\xi'(\boldsymbol{W}(\theta)\right] = \ell'(\theta)$ *on* Θ*, where* $\ell(\theta) = \mathbb{E}\left[\xi(\boldsymbol{W}(\theta))\right]$*.*

(1) For all $\theta \in \Theta$ *and* $i = 1, \ldots, n$*,* W_i *is a.s. differentiable at* θ*.*

(2) Define D_ξ *to be the subset of* $\mathbb{R}^n$ *on which* ξ *is continuously differentiable. For all* $\theta \in \Theta$*,* $P\left(\boldsymbol{W}(\theta) \in D_\xi\right) = 1$*.*

(3) $\xi\left(\boldsymbol{W}(\cdot)\right)$ *is a.s. continuous and piecewise differentiable throughout* Θ*.*

(4) $\widetilde{D}$ *is countable and* $\mathbb{E}\left[\sup_{\theta \notin \widetilde{D}} |\xi'(\boldsymbol{W}(\theta)|\right] < \infty$*, where* $\widetilde{D}$ *is the random collection of points in* Θ *at which* $\xi(\boldsymbol{W}(\cdot))$ *fails to be differentiable.*

We apply this result with $\theta = X_{mk}$*,* $\boldsymbol{W}(\cdot)$ *equal to the random function mapping* X_{mk} *to the random vector* $\boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{C}(\boldsymbol{X})\boldsymbol{Z}$*,* $\xi(w) = \max_{i=0,1,\ldots,q} w_i$*, and* $\widetilde{D}$ *equal to the set of* X_{mk} *at which* $h'(\boldsymbol{X},\boldsymbol{Z})$ *does not exist.*

Condition (1) is satisfied because $\boldsymbol{m}(\cdot)$ *and* $\boldsymbol{C}(\cdot)$ *are assumed differentiable.*

For condition (2), the set of points D_ξ *at which* ξ *is continuously differentiable is* $D_\xi = \{w \in \mathbb{R}^{q+1} : |\operatorname{argmax}_{i=0,1,\ldots,q} w_i| = 1\}$*. Lemma 2 implies that the probability of equality between two components of* $W(\theta)$ *is 0, and so* $P(W(\theta) \in D_\xi) = 0$*.*

For condition (3), it is obvious that $\xi\left(\boldsymbol{W}(\cdot)\right)$ *is a.s. continuous. Lemma 1 implies that the set of non-differentiable points is countable, and therefore* $\xi\left(\boldsymbol{W}(\cdot)\right)$ *is a.s. piecewise differentiable.*

For condition (4), first $\widetilde{D}$ *is countable by Lemma 1. We now show the second part of condition (4). Fix* $\boldsymbol{X}$ *except for* X_{mk}*. Since the interval is compact and* $\boldsymbol{m}(\boldsymbol{X})$ *and* $\boldsymbol{C}(\boldsymbol{X})$ *are continuously*

differentiable,

$$\sup_{X_{mk}} \left| \frac{\partial \boldsymbol{m}_i(\boldsymbol{X})}{\partial X_{mk}} \right| = m_i^* < \infty,$$
$$\sup_{X_{mk}} \left| \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}} \right| = C_{ik}^* < \infty.$$

Then

$$\mathbb{E}\left[\sup_{X_{mk} \notin \widetilde{D}} |h'(\boldsymbol{X}, \boldsymbol{Z})| \right] \leq m^{**} + qC^{**}\mathbb{E}[|Z|] = m^{**} + \sqrt{\frac{2}{\pi}} qC^{**} < \infty,$$

where $m^{**} = \max_i m_i^*$ *and* $C^{**} = \max_{i,j} C_{ij}^*$. *Therefore, condition (4) is satisfied.*

Thus the conditions of Theorem 1.2 from Glasserman (1991) are satisfied and $\nabla \mathbb{E} h(\boldsymbol{X}, \boldsymbol{Z}) = \mathbb{E} \nabla h(\boldsymbol{X}, \boldsymbol{Z})$.

Proof 5 *Proof of Theorem 2. We use a convergence analysis result from Section 5, Theorem 2.3 of Kushner and Yin (2003) to prove our theorem, which we first state using our notation and setting: the sequence* $\{\boldsymbol{X}_n\}$ *produced by algorithm* (14) *converges to a stationary point almost surely if the following assumptions hold,*

1. $\epsilon_n \to 0$ *for* $n \geq 0$ *and* $\epsilon_n = 0$ *for* $n < 0$; $\sum_{n=1}^{\infty} \epsilon_n = \infty$
2. $\sup_n \mathbb{E} \left| \boldsymbol{G}(\boldsymbol{X}_n) \right|^2 < \infty$
3. *There are functions* $\lambda_n(\cdot)$ *of* $\boldsymbol{X}$, *which are continuous uniformly in* n, *a continuous function* $\overline{\lambda}(\cdot)$ *and random variables* β_n *such that*

$$\mathbb{E}_n \boldsymbol{G}(\boldsymbol{X}_n) = \lambda_n(\boldsymbol{X}_n) + \beta_n,$$

and for each $\boldsymbol{X} \in H$,

$$\lim_n \left| \sum_{i=n}^{m(t_n+t)} \epsilon_i [\lambda_i(\boldsymbol{X}) - \overline{\lambda}(\boldsymbol{X})] \right| = 0$$

for each $t > 0$, *and* $\beta_n \to 0$ *with probability one. The function* $m(t_n + \cdot)$ *is defined in (Kushner and Yin 2003, Section 5.1).*

4. $\sum_i \epsilon_i^2 < \infty$.
5. *There is a continuously differentiable real-valued function* $\phi(\cdot)$ *such that* $\bar{\lambda}(\cdot) = -\nabla \phi(\cdot)$ *and* $\phi(\cdot)$ *is constant on each connected subset* S_i *of the set of stationary points.*
6. $a_i(\cdot), i = 1, \ldots, p$ *are continuously differentiable.*

Kushner and Yin (2003) shows that if these conditions are satisfied, then $\{X_n\}$ *converges to a unique* S_i. *Now we prove that the 5 conditions stated above are indeed satisfied if the assumptions in Theorem 2 hold, where* $\phi(\boldsymbol{X}) = -q\text{-}EI(\boldsymbol{X})$.

1. *Condition 1 is satisfied by assumption 2 in Theorem 2. Construction of this sequence has been discussed in Section 3.3.*

2. *First we assume $M = 1$ and treat $M > 1$ below. Then $\boldsymbol{G}(\boldsymbol{X}_n) = \boldsymbol{g}(\boldsymbol{X}_n, \boldsymbol{Z})$, and*

$$\begin{aligned}
\mathbb{E}|\boldsymbol{G}(\boldsymbol{X}_n)|^2 &= \mathbb{E}\sum_{m=1}^{q}\sum_{k=1}^{d} \boldsymbol{e}_{m,k}\boldsymbol{G}(\boldsymbol{X}_n)^2, \\
&= \sum_{m=1}^{q}\sum_{k=1}^{d} \mathbb{E}\left(\frac{\partial h(\boldsymbol{X},\boldsymbol{Z})}{\partial X_{mk}}\bigg|_{\boldsymbol{X}=\boldsymbol{X}_n}\right)^2, \\
&= \sum_{m=1}^{q}\sum_{k=1}^{d} \mathbb{E}\left[\boldsymbol{e}_{I_{\boldsymbol{Z}}^*}\left(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial X_{mk}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}}\boldsymbol{Z}\right)\bigg|_{\boldsymbol{X}=\boldsymbol{X}_n}\right]^2, \\
&\leq \sum_{m=1}^{q}\sum_{k=1}^{d} \mathbb{E}\sum_{i=0}^{q}\left[\boldsymbol{e}_{i}\left(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial X_{mk}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}}\boldsymbol{Z}\right)\bigg|_{\boldsymbol{X}=\boldsymbol{X}_n}\right]^2, \\
&= \sum_{m=1}^{q}\sum_{k=1}^{d}\sum_{i=0}^{q} \mathbb{E}\left[\boldsymbol{e}_{i}\left(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial X_{mk}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}}\boldsymbol{Z}\right)\bigg|_{\boldsymbol{X}=\boldsymbol{X}_n}\right]^2,
\end{aligned}$$

where $I_{\boldsymbol{Z}}^* = \underset{i=0,\ldots,q}{\operatorname{argmax}} \boldsymbol{e}_i\left(\boldsymbol{m}(\boldsymbol{X}_n) + \boldsymbol{C}(\boldsymbol{X}_n)\boldsymbol{Z}\right)$. Since $\boldsymbol{m}(\boldsymbol{X})$ and $\boldsymbol{C}(\boldsymbol{X})$ are continuously differentiable for $\forall \boldsymbol{X} \in H$ and H is compact, $\sup_{\boldsymbol{X}_n}\left|\left|\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial X_{mk}}\right|\right|_\infty < \infty$ and $\sup_{\boldsymbol{X}_n}\left|\left|\frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}}\right|\right|_\infty < \infty$. Thus $\sup_{\boldsymbol{X}_n}\mathbb{E}\left[\left(\boldsymbol{e}_i\left(\frac{\partial \boldsymbol{m}(\boldsymbol{X})}{\partial X_{mk}} + \frac{\partial \boldsymbol{C}(\boldsymbol{X})}{\partial X_{mk}}\boldsymbol{Z}\right)\right)^2\right] < \infty$, and we can conclude that $\sup_n \mathbb{E}|\boldsymbol{G}(\boldsymbol{X}_n)|^2 < \infty$.

If $M > 1$, $\boldsymbol{G}(\boldsymbol{X}_n)$ is an average of i.i.d. samples of $g(X_n, Z)$. Then $\mathbb{E}|g(X_n,Z)|^2 = \frac{1}{M}\mathbb{E}|g(X_n,Z)|^2$. We have just showed that $\sup_n \mathbb{E}|\boldsymbol{G}^1(\boldsymbol{X}_n)|^2$ is finite, and thus $\sup_n \mathbb{E}|\boldsymbol{G}(\boldsymbol{X}_n)|^2$ is finite. Therefore, condition 2 is satisfied.

3. *Define a function $\bar{\boldsymbol{g}}(\cdot)$ on H by $\bar{\boldsymbol{g}}(X) = \mathbb{E}\boldsymbol{g}(\boldsymbol{X},\boldsymbol{Z})$. Then $\mathbb{E}_n\boldsymbol{G}(\boldsymbol{X}_n) = \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_n,\boldsymbol{Z})$. Then, since our assumptions meet the requirements for Theorem 1 ($\boldsymbol{\Sigma}(\boldsymbol{X})$ being positive definite implies that $\boldsymbol{C}(\boldsymbol{X})$ has no duplicate rows), we know $\bar{\boldsymbol{g}}(\boldsymbol{X}) = \nabla\mathbb{E}h(\boldsymbol{X},\boldsymbol{Z})$. We will show $\nabla\mathbb{E}h(\boldsymbol{X},\boldsymbol{Z})$ is continuous on H. Letting $\lambda_n(\cdot) \equiv \overline{\lambda}(\cdot) \equiv \overline{\boldsymbol{g}}(\cdot)$, and $\beta_n = 0$, the first half of condition 3 will then be satisfied. Since $\lambda_n(\cdot) \equiv \overline{\lambda}(\cdot)$, the second half of condition 3 is satisfied from the fact that the summand is 0.*

We now show $\nabla\mathbb{E}h(\boldsymbol{X},(Z)$ is continuous. First, we let $\boldsymbol{m}'(\boldsymbol{X}) = f_n^* - \boldsymbol{\mu}(\boldsymbol{X})$, and $\boldsymbol{C}'(\boldsymbol{X}) = -\boldsymbol{L}(\boldsymbol{X})$, which are the first through the qth entries and rows of $\boldsymbol{m}(\boldsymbol{X})$ and $\boldsymbol{C}(\boldsymbol{X})$ respectively. Note that $\boldsymbol{\Sigma}(\boldsymbol{X}) = \boldsymbol{C}'(\boldsymbol{X})\boldsymbol{C}'^T(\boldsymbol{X})$. Then

$$\mathbb{E}\left[h(\boldsymbol{X},\boldsymbol{Z})\right] = \mathbb{E}\left[\mathbb{E}\left[h(\boldsymbol{X},\boldsymbol{Z}) \mid t_{-i}\right]\right], \tag{18}$$

where $t_{-i} = \left\{\boldsymbol{e}_\ell\left(\boldsymbol{m}'(\boldsymbol{X}) + \boldsymbol{C}'(\boldsymbol{X})\boldsymbol{Z}\right), \forall \ell \neq i, \ell = 1,\ldots,q\right\}$ for some $i = 1,\ldots,q$.

Fix i, and letting $t_i = \boldsymbol{e}_i\left(\boldsymbol{m}'(\boldsymbol{X}) + \boldsymbol{C}'(\boldsymbol{X})\boldsymbol{Z}\right)$, we know that t_i given t_{-i} has a normal distribution:

$$t_i \mid t_{-i} \sim \mathcal{N}\left(\mu(\boldsymbol{X}, t_{-i}), \sigma^2(\boldsymbol{X})\right),$$

where

$$\begin{aligned}\mu(\boldsymbol{X}, t_{-i}) &= m_i'(\boldsymbol{X}) + \Sigma_{i,-i}(\boldsymbol{X})\Sigma_{-i,-i}^{-1}(\boldsymbol{X})\left(t_{-i} - m_{-i}'(\boldsymbol{X})\right) \\ &= a(\boldsymbol{X}) + \sum_{\ell \neq i} b_\ell(\boldsymbol{X}) t_\ell, \\ \sigma^2(\boldsymbol{X}) &= \Sigma_{i,i}(\boldsymbol{X}) - \Sigma_{i,-i}(\boldsymbol{X})\Sigma_{-i,-i}^{-1}(\boldsymbol{X})\Sigma_{-i,i}(\boldsymbol{X}).\end{aligned}$$

Note that $\sigma^2(\boldsymbol{X})$ *is the Schur complement of* $\boldsymbol{\Sigma}(\boldsymbol{X})$*, and since* $\boldsymbol{\Sigma}(\boldsymbol{X})$ *is positive definite, we know that both* $\boldsymbol{\Sigma}_{-i,-i}(\boldsymbol{X})$ *and* $\sigma^2(\boldsymbol{X})$ *are positive definite. Knowing the distribution of* t_i *given* t_{-i}*, we can write the inner expectation of* (18) *as*

$$\begin{aligned}f(\boldsymbol{X}, t_{-i}) &= \mathbb{E}\left[h(\boldsymbol{X}, \boldsymbol{Z}) \mid t_{-i}\right], \\ &= \left(\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*\right)\Phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right) + \sigma(\boldsymbol{X})\phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right) + t_{-i}^*,\end{aligned} \tag{19}$$

where $t_{-i}^* = \max(t_{-i}, 0)$*. Without loss of generality, we only look at* j*th component of the gradient, and we have*

$$\begin{aligned}\frac{\partial f(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j} &= \frac{\partial \mu(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j}\Phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right) + \frac{\partial \sigma(\boldsymbol{X})}{\partial \boldsymbol{X}_j}\phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right), \\ &= \left(\frac{\partial a(\boldsymbol{X})}{\partial \boldsymbol{X}_j} + \sum_{\ell \neq i}\frac{\partial b_\ell(\boldsymbol{X})}{\partial \boldsymbol{X}_j} t_\ell\right)\Phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right) + \frac{\partial \sigma(\boldsymbol{X})}{\partial \boldsymbol{X}_j}\phi\left(\frac{\mu(\boldsymbol{X}, t_{-i}) - t_{-i}^*}{\sigma(\boldsymbol{X})}\right).\end{aligned} \tag{20}$$

Since $\boldsymbol{\Sigma}_{-i,-i}(\boldsymbol{X})$ *is positive definite, and the matrix inverse is a continuous function when restricted to the set of positive definite matrices, and the composition of two continuous functions is continuous, we have that* $\boldsymbol{\Sigma}_{-i,-i}^{-1}(\boldsymbol{X})$ *is continuously differentiable. Moreover,* $\boldsymbol{\mu}(\boldsymbol{X})$ *and* $\boldsymbol{\Sigma}(\boldsymbol{X})$ *are assumed continuously differentiable in the statement of the theorem. This together implies continuous differentiability of* $a(\boldsymbol{X}), b_\ell(\boldsymbol{X})$ *and* $\sigma(\boldsymbol{X})$*. Then*

$$\left|\frac{\partial f(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j}\right| \leq a^* + \sum_{\ell \neq i} b_\ell^* |t_\ell| + \sigma^*$$

for all $\boldsymbol{X} \in H$ *and* t_{-i}*, where*

$$\begin{aligned}a^* &= \sup_{\boldsymbol{X} \in H}\left|\frac{\partial a(\boldsymbol{X})}{\partial \boldsymbol{X}_j}\right|, \\ b_\ell^* &= \sup_{\boldsymbol{X} \in H}\left|\frac{\partial b_\ell(\boldsymbol{X})}{\partial \boldsymbol{X}_j}\right|, \\ \sigma^* &= \sup_{\boldsymbol{X} \in H}\left|\frac{\partial \sigma(\boldsymbol{X})}{\partial \boldsymbol{X}_j}\right|,\end{aligned}$$

and a^*, b_ℓ^*, σ^* *are finite because* H *is compact.*

Since t_ℓ has a normal distribution, $\mathbb{E}[|t_\ell|] < \infty$ and therefore

$$\mathbb{E}\left[|\frac{\partial f(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j}|\right] < \infty.$$

With the conditions above, we can apply Theorem 1.2 in Glasserman (1991), and have

$$\frac{\partial \mathbb{E}\left[f(\boldsymbol{X}, t_{-i})\right]}{\partial \boldsymbol{X}_j} = \mathbb{E}\left[\frac{\partial f(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j}\right]. \tag{21}$$

Moreover, we can write (21) as

$$\mathbb{E}\left[\frac{\partial f(\boldsymbol{X}, t_{-i})}{\partial \boldsymbol{X}_j}\right] = \mathbb{E}\left[\frac{\partial f(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_j}\phi(t'_{-i})\right],$$

where each component of t'_{-i} is an independent uniform random variable on $(-\infty, \infty)$, and $\phi(\cdot)$ is the multivariate normal probability density function for t_{-i}. Define the function

$$G(\boldsymbol{X}, t'_{-i}) = \frac{\partial f(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_j}\phi(t'_{-i}),$$

From Lemma 3, $|\frac{\partial G(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_k}|$ is bounded by a finite constant for all t'_{-i}. Thus $G(\boldsymbol{X}, t'_{-i})$ is Lipschitz continuous in $\boldsymbol{X}$ with some constant K.

Given any $\epsilon > 0$, we let $\delta = \epsilon/K$, and for any $\boldsymbol{X}'$ such that $|\boldsymbol{X}' - \boldsymbol{X}| < \delta$, $|G(\boldsymbol{X}', t'_{-i}) - G(\boldsymbol{X}, t'_{-i})| < K \cdot \delta = \epsilon$. Hence, $|\mathbb{E}\left[G(\boldsymbol{X}', t'_{-i})\right] - \mathbb{E}\left[G(\boldsymbol{X}, t'_{-i})\right]| \leq \mathbb{E}\left[|G(\boldsymbol{X}', t'_{-i}) - G(\boldsymbol{X}, t'_{-i})|\right] < \epsilon$ by Jensen's inequality. Therefore, $\mathbb{E}\left[G(\boldsymbol{X}, t'_{-i})\right] = \partial \mathbb{E}[h(\boldsymbol{X}, \boldsymbol{Z})]/\partial \boldsymbol{X}_j$ is continuous at any $\boldsymbol{X} \in H$.

4. *Condition 4 is satisfied by assumption 2 in Theorem 2.*

5. *From the proof of condition 3, we know $\bar{\lambda}(\cdot) = \overline{\boldsymbol{g}}(\cdot) = \nabla q\text{-}EI(\cdot)$, and thus $\phi(\cdot) = -q\text{-}EI(\cdot)$. We have shown that $\bar{\boldsymbol{g}}(\cdot)$ is continuous, and it is also trivial to see $\phi(\cdot)$ is constant on each S_i. Therefore, condition 5 is satisfied.*

6. *This is satisfied by assumption 1 in Theorem 2.*

In conclusion, all conditions are satisfied and therefore $\{\boldsymbol{X}_n\}$ converges to a connected set of stationary points almost surely. From Lemma 4, the Polyak-Ruppert average $\overline{\boldsymbol{X}_n(\omega)}$ of the sequence $\{\boldsymbol{X}_n(\omega)\}$ converges to the same set as the sequence $\{\boldsymbol{X}_n(\omega)\}$ for every ω.

Lemma 3 *If $\boldsymbol{\mu}(\boldsymbol{X})$, $\boldsymbol{\Sigma}(\boldsymbol{X})$ are twice differentiable, and $\boldsymbol{\Sigma}(\boldsymbol{X})$ is positive definite, then $|\frac{\partial G(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_k}|$ is bounded by a finite constant for all t'_{-i}.*

Proof 6 *Proof of Lemma 3 We can write $G(\boldsymbol{X}, t'_{-i})$ as*

$$G(\boldsymbol{X}, t'_{-i}) = \frac{1}{\sqrt{(2\pi)^{q-1}|\Sigma_{-i,-i}(\boldsymbol{X})|}} \frac{\partial f(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_j} e^{-\frac{1}{2}\left(t'_{-i} - \boldsymbol{m}'_{-i}(\boldsymbol{X})\right)^T \Sigma^{-1}_{-i,-i}(\boldsymbol{X})\left(t'_{-i} - \boldsymbol{m}'_{-i}(\boldsymbol{X})\right)}.$$

Since $\boldsymbol{\mu}(\boldsymbol{X})$, $\boldsymbol{\Sigma}(\boldsymbol{X})$ are twice differentiable, and $\boldsymbol{\Sigma}(\boldsymbol{X})$ is positive definite, we can take the partial

derivative with respect to $\boldsymbol{X}_k$. With some algebra, we can show

$$\left|\frac{\partial G(\boldsymbol{X}, t'_{-i})}{\partial \boldsymbol{X}_k}\right| < \sum_r c_r |P_r(t'_{-i})| e^{-\frac{1}{2}\left(t'_{-i}-\boldsymbol{m}'_{-i}(\boldsymbol{X})\right)^T \Sigma^{-1}_{-i,-i}(\boldsymbol{X})\left(t'_{-i}-\boldsymbol{m}'_{-i}(\boldsymbol{X})\right)}, \tag{22}$$

where each $P_r(t'_{-i})$ is a monomial in components of t'_{-i} with coefficient 1 and order ranging between 0 and 2, and $0 < c_r < \infty$. Let $\boldsymbol{L}_{-i,-i}(\boldsymbol{X})$ be the Cholesky decomposition of $\boldsymbol{\Sigma}^{-1}_{-i,-i}(\boldsymbol{X})$, and $z_{-i} = \boldsymbol{L}_{-i,-i}(\boldsymbol{X})\left(t'_{-i} - m'_{-i}(\boldsymbol{X})\right)$. Invertibility of $\Sigma^{-1}_{-i,-i}(\boldsymbol{X})$ implies that t'_{-i} can be written in terms of z_{-i} Substitute z_{-i} into (22) and, and we get

$$\left|\frac{\partial G(\boldsymbol{X}, z_{-i})}{\partial \boldsymbol{X}_k}\right| < \sum_r c'_r |P'_r(z_{-i})| e^{-\frac{1}{2} z^T_{-i} z_{-i}},$$

where each c'_r is a finite constant and each $P'_r(\cdot)$ is a monomial with coefficient 1 and order between 0 and 2. Without loss of generality, we assume the first component, z_0, has the largest absolute value among z_{-i}. Then

$$\left|\frac{\partial G(\boldsymbol{X}, z_{-i})}{\partial \boldsymbol{X}_k}\right| < \sum_r c'_r |P'_r(z_0)| e^{-\frac{1}{2} z_0^2}. \tag{23}$$

We can show that $|x^p| e^{-\frac{1}{2}x^2} \leq p^{\frac{p}{2}} e^{-\frac{1}{2}p}, \forall x \in (-\infty, \infty)$. Therefore, each summand in (23) is bounded by a constant.

Lemma 4 *Let $\{\boldsymbol{X}_n : n \geq 1\}$ be a sequence in H, where H is compact, converging to a set A. Let $\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{m=1}^n \boldsymbol{X}_m$. Then $\{\overline{\boldsymbol{X}}_n : n \geq 1\}$ also converges to A.*

Proof 7 *Proof of Lemma 4 Let $\rho_A(\boldsymbol{X}) := \inf\{||\boldsymbol{X} - \boldsymbol{X'}|| : \boldsymbol{X'} \in A\}$ denote the distance to A, where $||\cdot||$ denotes the L_2 norm. $\rho_A(\cdot)$ is convex.*

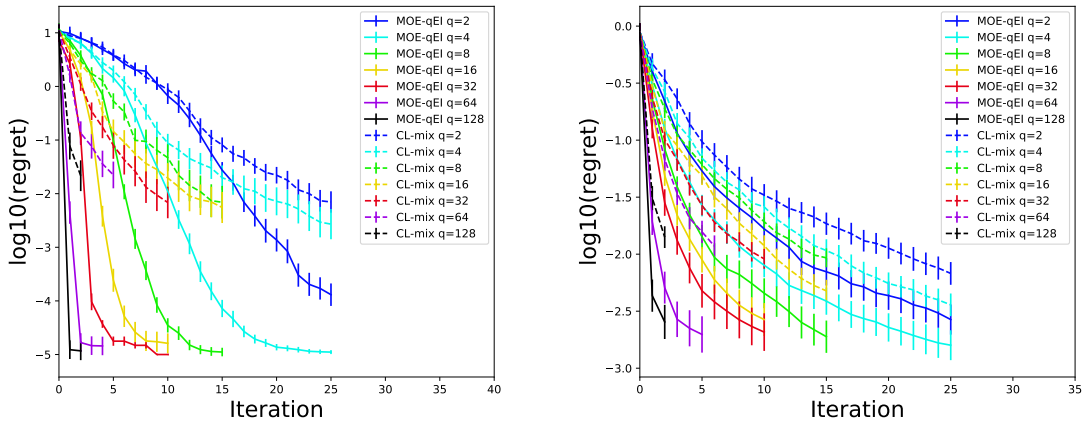
Let $\epsilon > 0$. Since $\boldsymbol{X}_n$ converges to A, $\exists N_\epsilon$ such that $\rho_A(\boldsymbol{X}_n) < \epsilon$ for all $n > N_\epsilon$. For $n > N_\epsilon$,

$$\overline{\boldsymbol{X}}_n = \frac{N_\epsilon}{n}\overline{\boldsymbol{X}}_{N_\epsilon} + \left(1 - \frac{N_\epsilon}{n}\right) \cdot \frac{1}{n - N_\epsilon} \sum_{m=N_\epsilon+1}^n \boldsymbol{X}_m.$$

Then

$$\begin{aligned}
\rho_A(\overline{\boldsymbol{X}}_n) &= \rho_A\left(\frac{N_\epsilon}{n}\overline{\boldsymbol{X}}_{N_\epsilon} + \left(1 - \frac{N_\epsilon}{n}\right)\frac{1}{n - N_\epsilon}\sum_{m=N_\epsilon+1}^n \boldsymbol{X}_m\right) \\
&\leq \frac{N_\epsilon}{n}\rho_A(\overline{\boldsymbol{X}}_{N_\epsilon}) + \left(1 - \frac{N_\epsilon}{n}\right)\frac{1}{n - N_\epsilon}\sum_{m=N_\epsilon+1}^n \rho_A\left(\boldsymbol{X}_m\right) \\
&\leq \frac{N_\epsilon}{n} \cdot C + \left(1 - \frac{N_\epsilon}{n}\right) \cdot \epsilon,
\end{aligned} \tag{24}$$

where $C := \sup_{\boldsymbol{X} \in H} \rho_A(\boldsymbol{X})$ is finite. Let $\epsilon' > 0$, and choose $\epsilon = \frac{\epsilon'}{2}$. Let $n > N_\epsilon$ be such that $\frac{N_\epsilon}{n}C + (1 - \frac{N_\epsilon}{n})\epsilon \leq \epsilon' = 2\epsilon$. Then $\forall n' > n$, $\rho_A(\overline{\boldsymbol{X}_{n'}}) \leq \epsilon'$.



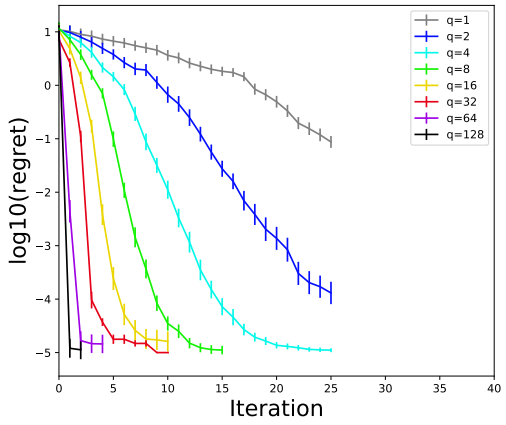
(a) Branin2

(b) Hartmann3

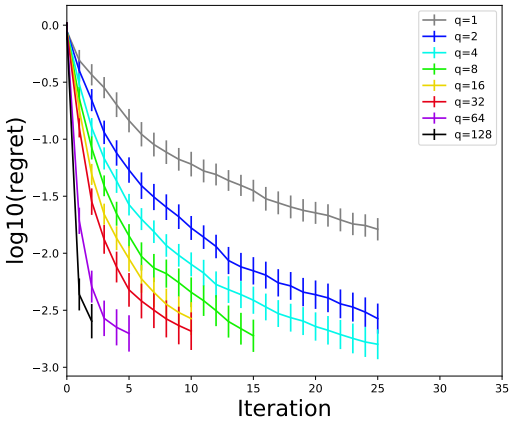
(c) Ackley5

(d) Hartmann6

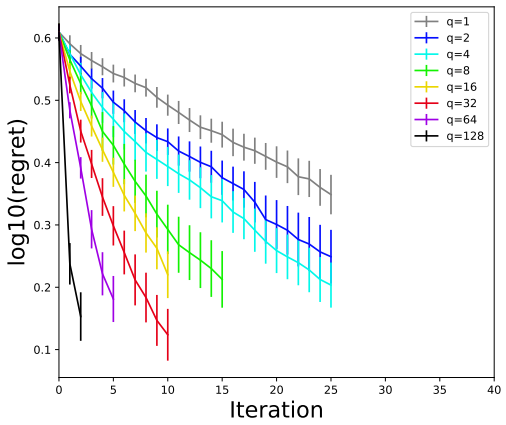
Figure 1: Comparison against Constant Liar: $\log_{10}$(regret) vs. iteration for MOE-qEI (solid line) and CL-mix (dashed line), where the error bars show 95% confidence intervals obtained from 100 repeated experiments with different sets of initial points. MOE-qEI converges faster with better solution quality than the heuristic method CL-mix for all q.



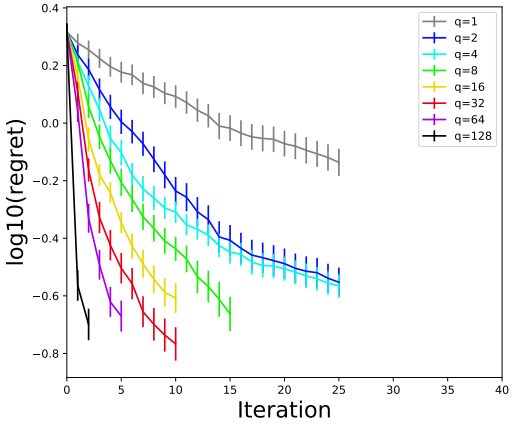
(a) Branin2



(b) Hartmann3

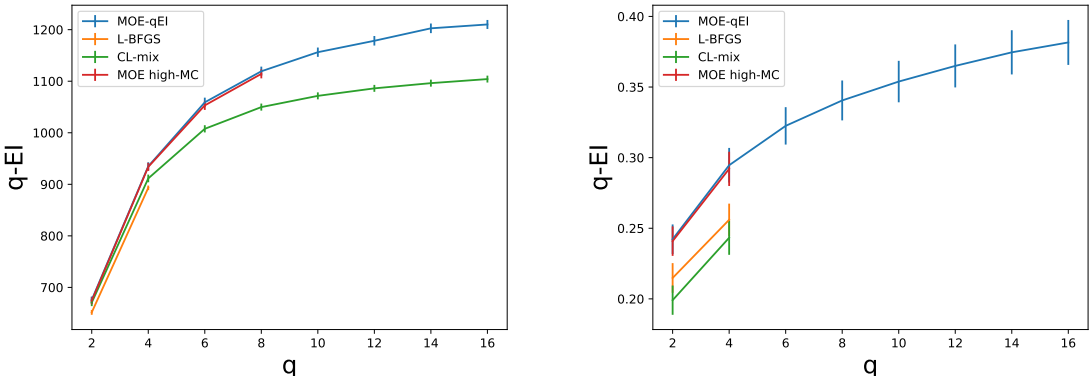


(c) Ackley5



(d) Hartmann6

Figure 2: Comparison against EGO: $\log_{10}$(regret) vs. iteration for different q, where the error bars show 95% confidence intervals obtained from 100 repeated experiments with different sets of initial points.



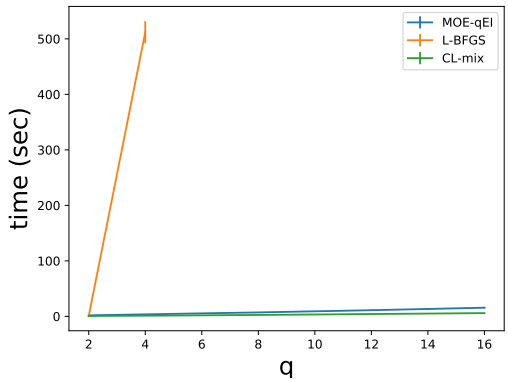
(a) Branin2

(b) Hartmann3

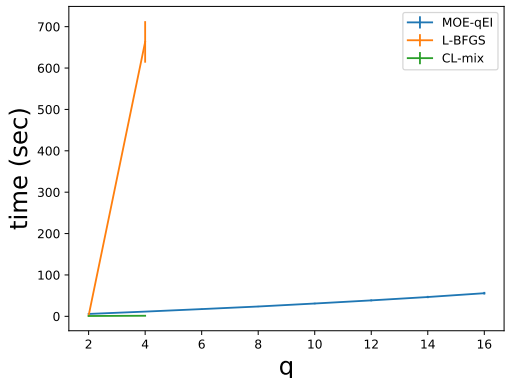
(c) Ackley5

(d) Hartmann6

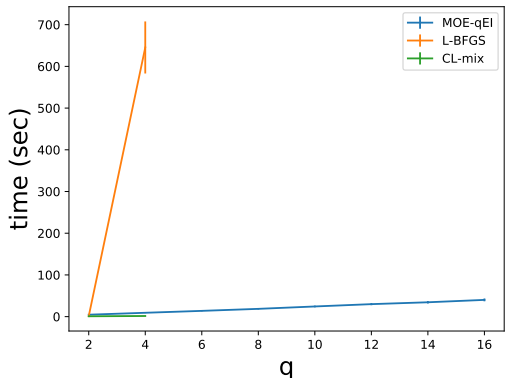
Figure 3: Comparison of algorithms for solving the inner optimization problem: solution quality (maximum q-EI) vs. q for different algorithms solving the inner optimization problem. For each test function, we generated 500 instances of the inner optimization problem by randomly sampling $(2d + 2)$ points, and the plot shows the average solution quality and 95% confidence interval for the expected solution quality over 500 problem instances.



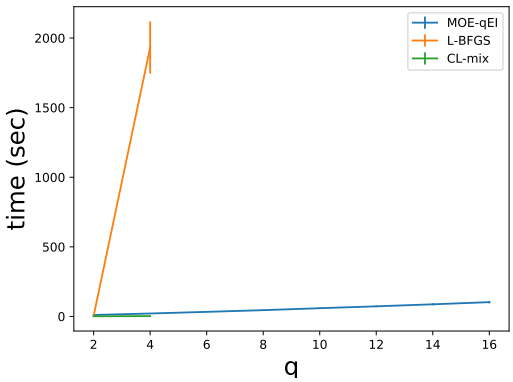
(a) Branin2



(b) Hartmann3



(c) Ackley5



(d) Hartmann6

Figure 4: Comparison of algorithms for solving the inner optimization problem: Runtime vs. q for different algorithms solving the inner optimization problem.

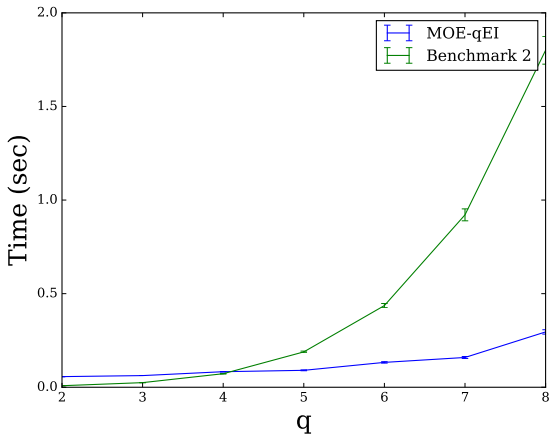


Figure 5: Comparison against closed-form evaluation of ∇q-EI: average time to compute ∇q-EI with high precision v.s. q, comparing the gradient-based estimator from MOE-qEI using a large number of samples (10^7) in a parallel GPU implementation with the closed-form formula from Marmin et al. (2015). The stochastic gradient estimator in MOE-qEI scales better in q and is faster when $q \geq 4$.