Efficient batch-sequential Bayesian optimization with moments of truncated Gaussian vectors

Sébastien Marmin, Clément Chevalier, David Ginsbourger

Efficient batch-sequential Bayesian optimization with moments of truncated Gaussian vectors

Sébastien Marmin*†‡ , Clément Chevalier§ , David Ginsbourger*¶

September 7, 2016

Abstract

We deal with the efficient parallelization of Bayesian global optimization algorithms, and more specifically of those based on the expected improvement criterion and its variants. A closed form formula relying on multivariate Gaussian cumulative distribution functions is established for a generalized version of the multipoint expected improvement criterion. In turn, the latter relies on intermediate results that could be of independent interest concerning moments of truncated Gaussian vectors. The obtained expansion of the criterion enables studying its differentiability with respect to point batches and calculating the corresponding gradient in closed form. Furthermore, we derive fast numerical approximations of this gradient and propose efficient batch optimization strategies. Numerical experiments illustrate that the proposed approaches enable computational savings of between one and two order of magnitudes, hence enabling derivative-based batch-sequential acquisition function maximization to become a practically implementable and efficient standard. **Keywords:** Kriging, Expected Improvement, Parallel Optimization.

1 Introduction

Since their beginnings about half a century ago [26, 49, 31], Bayesian optimization algorithms have been increasingly used for derivative-free global minimization of expensive to evaluate functions. Typically assuming a continuous objective function $f : \boldsymbol{x} \in D \subset \mathbb{R}^d \longrightarrow f(\boldsymbol{x}) \in \mathbb{R}$, single-objective Bayesian optimization algorithms consist in sequentially evaluating f at promising points under the assumption that f is a sample realization (*path* or *trajectory*) of a random field $(Y(\boldsymbol{x}))_{\boldsymbol{x}\in D}$. Such algorithms are especially popular in the case where evaluating $f(\boldsymbol{x})$ requires heavy high-fidelity numerical simulations (or *computer experiments*, see notably [33, 40, 41, 24]), where $\boldsymbol{x}$ stands for some design parameters to be optimized over. Such expensive simulations are classically encountered in the resolution of partial differential equations from physical sciences, engineering and beyond [13]. In recent years, Bayesian optimization also has attracted a lot of interest from the machine learning community [27, 6, 34, 43], be it to optimize simulation-based objective functions [28, 45, 38] or even to estimate tuning parameters of machine learning algorithms themselves [3, 4, 42]. In both communities, a Gaussian random field (or *Gaussian Process*, GP) model is often used for Y, so that prior information on f is taken into account through a trend function $m : D \longrightarrow \mathbb{R}$ and a covariance kernel $k : (\boldsymbol{x}, \boldsymbol{x}') : D \times D \longrightarrow \mathbb{R}$. Once m and k are specified, possibly up to some parameters to be inferred based on data, the considered GP model can be used as an instrument to locate the next evaluation point(s) via so-called infill sampling criteria, also referred to as *acquisition functions* or simply as *criteria*. While a number of Bayesian

*IMSV, Department of Mathematics and Statistics, University of Bern, Switzerland

†Institut de Mathématiques de Marseille (UMR7373), École Centrale de Marseille, France

‡Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSN-RES, SEMIA, LIMAR, Cadarache, 13115 Saint-Paul-lès-Durance, France

§Institute of Statistics, University of Neuchâtel, Switzerland

¶Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Martigny, Switzerland

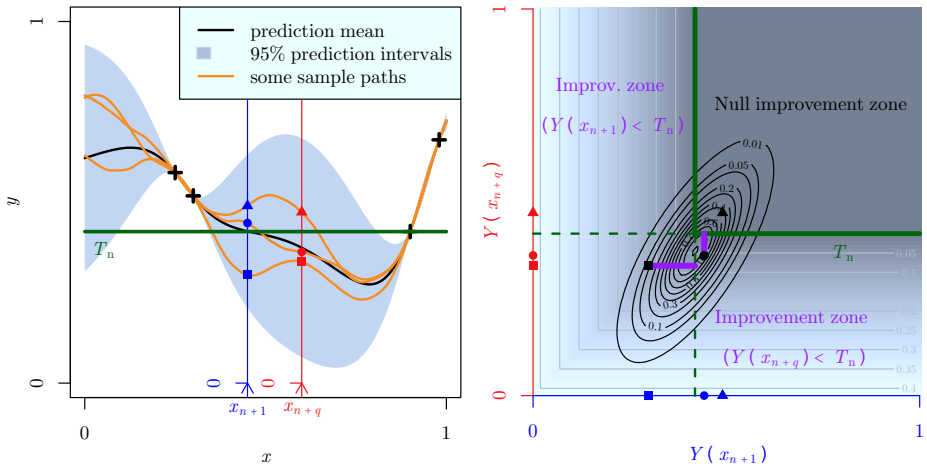

Figure 1: Illustration of the principles underlying q-EI for $d = 1$, $n = 4$, $q = 2$. Left: Gaussian process prediction of a function f from observations $\mathcal{A}_n$ (depicted by black crosses). The green horizontal line stands for T_n, the smallest response value from $\mathcal{A}_n$. Three conditional simulation draws are plotted in orange and various point symbols represent their respective values at two unobserved locations x_{n+1} and x_{n+q}. Right: distribution of the random vector $(Y(x_{n+1}), Y(x_{n+q}))^\top$ knowing $\mathcal{A}_n$ (black contours). For each point symbol, the length of the purple segment represents the improvement realized by the corresponding sample path. The multipoint EI is the expectation of this length, or in other words, it is the integral of the improvement (grey-scale function) with respect to the conditional distribution of $(Y(x_{n+1}), Y(x_{n+q}))^\top$ knowing $\mathcal{A}_n$.

optimization criteria have been proposed in the literature (see, e.g., [23, 15, 46, 43, 9] and references therein), we concentrate here essentially on the *Expected Improvement* (EI) criterion [30, 24] and on variations thereof, with a focus on its use in synchronous batch-sequential optimization. Denoting by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in D$ points were f is assumed to have already been evaluated and by $\boldsymbol{x}_{n+1:n+q} := (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}) \in D^q$ a batch of candidate points where to evaluate f next, the multipoint EI is defined as

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \mathbb{E}_n\left(\left(\min_{i=1,\ldots,n} Y(\boldsymbol{x}_i) - \min_{j=n+1,\ldots,n+q} Y(\boldsymbol{x}_j)\right)_+\right), \tag{1}$$

where $\mathbb{E}_n$ refers to the conditional expectation knowing the event $\mathcal{A}_n := \{Y(\boldsymbol{x}_1) = f(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n) = f(\boldsymbol{x}_n)\}$. One way of calculating such criterion is to rely on Monte Carlo simulations. Figure 1 illustrates both what the criterion means and how to approach it by simulations, relying on three samples from the multivariate Gaussian distribution underlying Equation (1). Our main focus here, in contrast, is on deriving Equation (1) in closed form, studying the criterion's differentiability, and ultimately calculating and efficiently approximating its gradient in order to perform efficient batch optimization using derivative-based deterministic search.

Now, for $q = 1$, it is well known that EI can be expressed in closed form as a function of $m_n(\boldsymbol{x}) = \mathbb{E}_n(Y_{\boldsymbol{x}})$ and $s_n(\boldsymbol{x}) = \sqrt{\mathrm{var}_n(Y(\boldsymbol{x}))}$ as follows

$$EI_n(\boldsymbol{x}) = s_n(\boldsymbol{x})\left(u_n(\boldsymbol{x})\Phi(u_n(\boldsymbol{x})) + \varphi(u_n(\boldsymbol{x}))\right) \text{ if } s_n(\boldsymbol{x}) \neq 0 \text{ and } 0 \text{ else,} \tag{2}$$

where $u_n(\boldsymbol{x}) = (\min_{i=1,\ldots,n} f(\boldsymbol{x}_i) - m_n(\boldsymbol{x}))/s_n(\boldsymbol{x})$ (defined for $s_n(\boldsymbol{x}) \neq 0$) and Φ, φ are the cumulative distribution function and probability density function of the standard Gaussian distribution, respectively.

When deriving Equation (2), Equation (1) happens (hence for $q = 1$) to involve a first order moment of the truncated univariate Gaussian distribution. As shown in [7] and developed further

here, it turns out that Equation (1) can be expanded in a similar way in the multipoint case ($q \geq 2$) relying on moments of truncated Gaussian vectors. This is essential for the open challenges tackled here of efficiently calculating and optimizing the multipoint criterion of Equation (1).

The applied motivation for having batch-sequential EI algorithms is strong, as distributing evaluations of Bayesian optimization algorithms over several computing units allows significantly reducing wall-clock time and with the fast popularization of clouds, clusters and GPUs in recent years it is becoming always more commonplace to launch several calculations in parallel. Even at a slightly inflated price and scripting effort, reducing the total time off is often a primary goal in order to deliver conclusions involving heavy experiments, be they numerical or laboratory experiments, in studies subject to hard time limitations. Obviously, given its practical importance, the question of parallelizing EI algorithms and alike by selecting $q > 1$ points per iteration has been already tackled in a number of works from various disciplinary horizons (including notably [36, 1, 12, 8, 19]). Here we essentially focus on approaches relying on the maximization of Equation (1) and related multipoint criteria. The multipoint EI of Equation (1) has been defined in [30, 41] and first calculated in closed form for the case $q = 2$ in [17]. For the case $q \geq 3$, a Monte Carlo scheme and some sub-optimal batch selection strategies were proposed. Further work on Monte Carlo simulations for multipoint EI estimation can be found in [22, 18]; besides this, stochastic simulation ideas have been explored in [14] for maximizing this multipoint EI criterion via a stochastic gradient algorithm, an approach recently investigated in [47]. Meanwhile, a closed-form formula for the multipoint EI relying on combinations of $(q-1)$- and q-dimensional Gaussian cumulative distribution functions was obtained in [7], a formula which applicability in reasonable time is however restricted to moderate q (say $q \leq 10$) in the current situation. Building upon [7], [29] recently calculated the gradient of the multipoint EI criterion in closed form and obtained some first experimental results on (non-stochastic) gradient-based multipoint EI maximization.

Our aim in the present paper is to present a set of novel analytical and numerical results pertaining to the calculation, the computation, and the maximization of the multipoint EI criterion. As most of these novel results apply to a broader class of criteria, we first present in Section 2 a generalization of the multipoint EI that allows accounting for noise in conditioning observations and also exponentiating the improvement. This generalized criterion is calculated using moments of truncated Gaussian vectors in the flavour of [7]. The obtained formula is then revisited in the standard case (noise-free with an exponent set to 1), leading to a numerical approximation of the multipoint EI with arbitrary precision and very significantly reduced computation time. Next, the (qd)-dimensional maximization of the multipoint EI criterion is discussed in Section 3, where the differentiability of the generalized criterion is studied, its analytical gradient is calculated, and further numerical approaches for fast gradient approximations with controllable accuracy are presented. Finally, Section 4 is dedicated to numerical experiments where, in particular, a multistart derivative-based multipoint EI maximization algorithm highlighting the benefits of the considered methodological principles and the proposed fast approximations is tested and compared to baseline strategies.

2 Criteria in parallel Bayesian optimization

2.1 General definition of Expected Improvement

Throughout this section the objective function f may be observed noise-free or in noise, meaning that at some arbitrary iteration i the observed value may be $f(\boldsymbol{x}_i)$ or $f(\boldsymbol{x}_i)+\varepsilon_i$ where ε_i is a realization of a zero mean Gaussian random variable with known (or estimated and plugged-in) variance. f is assumed to be one realization of a random field Y, where Y has a Gaussian random field (GRF) distribution conditionally to events of the form $\mathcal{A}_n := \{Y(\boldsymbol{x}_1) = f(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n) = f(\boldsymbol{x}_1)\}$ (with conditioning on $Y(\boldsymbol{x}_i) + \varepsilon_i$ in the noisy case, see for instance [35]). This setup naturally includes the case where Y is a GRF, but also the so-called Universal Kriging settings where Y is the sum of a trend with an improper prior and a GRF [33, 32]. Note that in noisy cases the ε_i's are generally assumed to be

independent (although the case of ε_i's forming a Gaussian vector is tractable), but more essentially they are assumed independent of Y.

In batch-sequential Bayesian Optimization we are interested in computing sampling criteria J_n depending on $q \geq 1$ new points $\boldsymbol{x}_{n+1:n+q} = (\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q}) \in D^q$. At any step of corresponding synchronous parallel algorithms, the next batch of q points $\boldsymbol{x}^\star_{n+1:n+q}$ is then defined by globally maximizing J_n over all possible batches:

$$\boldsymbol{x}^\star_{n+1:n+q} \in \underset{\boldsymbol{x}_{n+1:n+q} \in D^q}{\arg\max}\, J_n(\boldsymbol{x}_{n+1:n+q}). \tag{3}$$

Values of such criteria typically depend on $\boldsymbol{x}_{n+1:n+q}$ through the conditional distribution $Y(\boldsymbol{x}_{1:n+q})|\mathcal{A}_n$, simplifying to $Y(\boldsymbol{x}_{n+1:n+q})|\mathcal{A}_n$ in the noiseless context. Conditional mean and covariance functions are analytically formulated via the so-called *kriging equations*, see e.g. [39]. Working out these criteria thus generally boils down to Gaussian vector calculus, which may become intricate and quite cumbersome to implement as q (or $n+q$, in noisy settings) increases. Our considered generalized version of the multipoint EI criterion, that allows accounting for a Gaussian noise in the conditioning observations and also for an exponentiation in the definition of the improvement, is defined as:

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \mathbb{E}_n\left(\left(\min_{\ell=1,\ldots,n} Y(\boldsymbol{x}_\ell) - \min_{k=1,\ldots,q} Y(\boldsymbol{x}_{n+k})\right)_+^\alpha\right), \tag{4}$$

where $\alpha \in \mathbb{N}\backslash\{0\}$, $\mathbb{E}_n(\cdot) = \mathbb{E}(\cdot|\mathcal{A}_n)$ and $(\cdot)_+ := \max(0, \cdot)$. This form gathers several sampling criteria notably including q-EI, both in noiseless and noisy settings, and also a multipoint version of the generalized EI of [41]. In addition, the obtained results apply to batch-sequential versions of the Expected Quantile Improvement [35] (EQI) and variations thereof, by a simply change of process from Y to the quantile process. We will show in proposition 2 that such generalized multipoint EI criteria can be formulated as a sum of moments of truncated Gaussian vectors. In the next subsection, in order to get a closed form for the generalized EI we first define these moments and derive some first analytical formulas, that might also be of relevance in further contexts.

2.2 Preliminaries on moments of truncated Gaussian distribution

We fix $\alpha \in \mathbb{N}\backslash\{0\}$ and $p = n + q$ in noisy settings or $p = q$ in noiseless settings.

Definition 1. *Let $\boldsymbol{Z}$ be a Gaussian vector with mean $\boldsymbol{m} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in S^p_{++}$, where S^p_{++} is the cone of positive definite matrices of $\mathbb{R}^{p\times p}$. For all positive integer $k \leq p$, we define the function $\mathcal{M}_{k,\alpha}$ on $\mathbb{R}^p \times S^p_{++}$ by*

$$\mathcal{M}_{k,\alpha} : (\boldsymbol{m}, \Sigma) \mapsto \mathcal{M}_{k,\alpha}(\boldsymbol{m}, \Sigma) = \mathbb{E}_n\left(Z_k^\alpha\ 1_{\{\boldsymbol{Z}\leq \boldsymbol{0}\}}\right), \tag{5}$$

where the inequality $\boldsymbol{Z} \leq \boldsymbol{0}$ is to be interpreted component-wise.

If $\boldsymbol{Z}$ is composed of values of a GRF at a batch of q locations $\boldsymbol{x}_{n+1:n+q}$, we use the notation $\mathcal{M}_{k,\alpha}(\boldsymbol{Z}(\boldsymbol{x}_{n+1:n+q})) := \mathcal{M}_{k,\alpha}(\boldsymbol{m}(\boldsymbol{x}_{n+1:n+q}), \Sigma(\boldsymbol{x}_{n+1:n+q}))$. We obtain the moments $\mathcal{M}_{k,\alpha}(\boldsymbol{m}, \Sigma)$ of a truncated Gaussian distribution by an extension of Tallis' technique [44] to any order, presented in the following proposition:

Proposition 1. *The function $\mathcal{G} : \mathbb{R}^p \times \mathbb{R}^p \times S^p_{++} \to \mathbb{R}$ defined by*

$$\mathcal{G}(\boldsymbol{t}, \boldsymbol{m}, \Sigma) = e^{\frac{1}{2}\left((\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m})^\top \Sigma (\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}) - \boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)}\Phi_{p,\Sigma}\left(-\boldsymbol{m} - \Sigma\boldsymbol{t}\right), \tag{6}$$

where $\Phi_{p,\Sigma}(\cdot)$ is the cumulative distribution function of the centered p-variate normal distribution, is infinitely differentiable, and the moments $\mathcal{M}_{k,\alpha}$ are given by:

$$\mathcal{M}_{k,\alpha}(\boldsymbol{m},\Sigma) = \left.\frac{\partial^\alpha \mathcal{G}(\cdot,\boldsymbol{m},\Sigma)}{\partial t_k^\alpha}\right|_{\boldsymbol{t}=\boldsymbol{0}}. \tag{7}$$

The proof of this Proposition is given in appendix B.1 and relies on calculating the moment generating function $\boldsymbol{t} \to \mathbb{E}\left(\exp\left(\boldsymbol{t}^\top \boldsymbol{Z}\right) 1_{\{\boldsymbol{Z}\preceq \boldsymbol{0}\}}\right)$. Even if an analytical formula can be obtained at any order of differentiation α, the complexity of derivatives in equation (7) increases rapidly. We give below the results for α equals 1 and 2.

Case $\alpha = 1$

Differentiating $\mathcal{G}$ with respect to $\boldsymbol{t}$ yields:

$$\begin{aligned}\frac{\partial \mathcal{G}}{\partial \boldsymbol{t}}(\boldsymbol{t},\boldsymbol{m},\Sigma) = &\exp\left(\frac{1}{2}\left(\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right) - \boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)\right)\times \\ &\left(\Sigma\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)\Phi_{p,\Sigma}\left(-\boldsymbol{m}-\Sigma\boldsymbol{t}\right) - \Sigma\nabla\Phi_{p,\Sigma}\left(-\boldsymbol{m}-\Sigma\boldsymbol{t}\right)\right)\end{aligned}$$

where $\nabla\Phi_{p,\Sigma}$ is the gradient of $\Phi_{p,\Sigma}$ (see appendix A.1 for an analytical derivation). Taking $\boldsymbol{t} = \boldsymbol{0}$ in the previous equation gives

$$\mathcal{M}_{k,1}(\boldsymbol{m},\Sigma) = m_k\Phi_{p,\Sigma}(-\boldsymbol{m}) - \boldsymbol{\Sigma}_k^\top\nabla\Phi_{p,\Sigma}(-\boldsymbol{m}) \tag{8}$$

where $\boldsymbol{\Sigma}_k$ is the k^{th} column of Σ. It is shown in appendix A.1 that computing each of the p components of $\nabla\Phi_{p,\Sigma}$ requires to compute a multivariate CDF of the normal distribution in dimension $p-1$. The number of calls to this function for computing the first moment of the truncated Gaussian distribution is thus of $O(p)$.

Case $\alpha = 2$

Similarly, differentiating $\mathcal{G}$ twice with respect to $\boldsymbol{t}$ yields

$$\begin{aligned}\mathcal{M}_{k,2}(\boldsymbol{m},\Sigma) = &(\Sigma_{kk}+m_k^2)\Phi_{p,\Sigma}(-\boldsymbol{m}) + \boldsymbol{\Sigma}_k^\top\ \nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{\Sigma}_k \\ &+ 2m_k\mathcal{M}_{k,1}(\boldsymbol{m},\Sigma).\end{aligned} \tag{9}$$

For readability, the detailed formula of $\nabla\nabla^\top\Phi_{p,\Sigma}$, the Hessian matrix of $\Phi_{p,\Sigma}$, is sent to Appendix A.2. The number of calls to the multivariate normal CDF is of $O(p^2)$.

2.3 Analytic formulas for generalized q-EI

The previous results obtained for the moments of the truncated normal distribution turn out to be of interest for computing the generalized q-EI introduced in Equation (4), as shown by the following proposition.

Proposition 2. *For $\boldsymbol{x}_{n+1:n+q} \in D^q$, the criterion EI_n defined by* (4) *exists for all α and can be written as a sum of moments of truncated normal distributions*

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{k=1}^{q}\mathcal{M}_{n+k-1,\alpha}\left(\boldsymbol{Z}^{(\ell,k)}(\boldsymbol{x}_{n+1:n+q})\right), \tag{10}$$

with $\boldsymbol{Z}^{(\ell,k)}(\boldsymbol{x}_{n+1:n+q})$ *a vector of size* $n+q-1$ *defined, noting* $Y_i := Y(\boldsymbol{x}_i)$*, by*

$$Z_i^{(\ell,k)} = \begin{cases} Y_\ell - Y_i & \text{if } 1 \leq i \leq \ell - 1, \\ Y_\ell - Y_{i+1} & \text{if } \ell \leq i \leq n-1, \\ Y_k - Y_{i+1} & \text{if } n \leq i \leq n+q-1 \text{ and } i \neq n+k-1, \\ Y_k - Y_\ell & \text{if } i = n+k-1. \end{cases}$$

Moreover, in the noiseless case the random vector $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n))$ *becomes deterministic conditionally to* $\mathcal{A}_n$*. Denoting by* ℓ_0 *the (smallest) index of the minimal observation, i.e.* $Y_{\ell_0} = \min_{\ell=1,\ldots,n} Y_\ell$*, and writing* $\boldsymbol{Z}^{(k)}(\boldsymbol{x}_{n+1:n+q})$ *the vector of the* q *last components of* $\boldsymbol{Z}^{(\ell_0,k)}(\boldsymbol{x}_{n+1:n+q})$*, Equation* (10) *is simplified to:*

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{k=1}^{q} \mathcal{M}_{k,\alpha}\left(\boldsymbol{Z}^{(k)}(\boldsymbol{x}_{n+1:n+q})\right). \tag{11}$$

Remark 1. *In the rest of the article we also use the following compact notation for the* $(n+q-1)-$*dimensional vector* $\boldsymbol{Z}^{(\ell,k)}(\boldsymbol{x}_{n+1:n+q})$*:*

$$\boldsymbol{Z}^{(\ell,k)}(\boldsymbol{x}_{n+1:n+q}) = A^{(\ell,k)}\left(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_{n+q})\right)^\top, \tag{12}$$

where $A^{(\ell,k)}$ *is a matrix implicitly defined by* $Z_i^{(l,k)}$ *of proposition 2.*

The proof of Proposition 2 is relegated to Appendix C for conciseness. Equation (10) highlights that the computation of the generalized q-EI in noisy settings is challenging since it involves computing nq different moments, each requiring $(n+q)^\alpha$ calls to the multivariate normal CDF in a dimension close to $n+q$. Even for $\alpha = 1$ and moderate q, the linear dependence in the number of observations n makes the use of this criterion challenging in application. Regarding the noiseless criterion, the computation of q moments is more affordable, at least for moderate q, but one has to keep in mind that the ultimate goal here is to perform global maximization of the considered criteria. It is thus important to bring further calculation speed-ups in order to perform this optimization in a reasonable time compared to the evaluation time of the objective function f, assumed expensive to evaluate. The next section discusses these matters and proposes faster formulas to compute both q-EI and its gradient.

3 Computing and optimizing the criteria

3.1 Generalities

Maximizing the EI_n expressions given in Equation (10) (noisy settings) or (11) (noiseless settings) is difficult. These maximizations are performed with respect to a batch of q points $\boldsymbol{x}_{n+1:n+q} \in (\mathbb{R}^d)^q$, and are thus optimization problems in dimension dq. In this space, the objective function to be maximized, is not convex in general and has the interesting property that the q points in the batch can be permuted without changing the value of EI_n; i.e. $EI_n((\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+q})) = EI_n((\boldsymbol{x}_{n+\sigma(1)}, \ldots, \boldsymbol{x}_{n+\sigma(q)}))$ for any permutation σ of $\{1, \ldots, q\}$. With this property, one can reduce the measure of the search domain by $q!$, e.g. by imposing that the first coordinate of the q points in the batch are in ascending order. We will restrict our attention here to the use of multi-start gradient based local optimization algorithms acting on the whole input domain $D^q \subset \mathbb{R}^{dq}$, that do not exploit the structure of the problem but do not seem to be affected by this, at least with the chosen settings regarding the starting designs. Our contribution here will be to propose a faster formula for computing the first moments $\mathcal{M}_{k,1}$ previously presented, as well as their derivatives. This will yield an easier computation of both the generalized EI and its dq-dimensional gradient. Besides, a second approximate but faster formula to further reduce the calculation time of the gradient will be introduced.

3.2 Gradient of the generalized q-EI

In this section, we extend the analytical gradient calculation of the q-EI performed in [29] to the case of the generalized noisy and noise-free q-EI, and provide in turn a more concise formula. Again, the presented formulas rely on results on moments of truncated Gaussian distributions.

Proposition 3. *Let $\boldsymbol{x}_{n+1:n+q} \in D^q$ be a batch such that the conditional covariance matrix $(\mathrm{cov}\left(Y\left(\boldsymbol{x}_{n+i}\right), Y\left(\boldsymbol{x}_{n+j}\right) | \mathcal{A}_n\right))_{1\leq i,j\leq q}$ is positive definite and the functions $\mathbb{E}\left(Y(\cdot) | \mathcal{A}_n\right)$ and $(\mathrm{cov}\left(Y(\cdot), Y(\boldsymbol{x}_{n+j}) | \mathcal{A}_n\right))_{j=1,\ldots,q}$ are differentiable at each point $\boldsymbol{x}_{n+i}$ $(1 \leq i \leq q)$. These derivatives are written $\boldsymbol{m}'^{(i)} \in \mathbb{R}^d$ and $\Sigma'^{(i)} \in \mathbb{R}^{q\times d}$ respectively. In this setup, the EI_n function of Equation (10) is differentiable and its derivative with respect to the j^{th} coordinate of the point $\boldsymbol{x}_{n+i}$ is*

$$\frac{\partial EI}{\partial x_{ij}}(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{k=1}^{q} m_j'^{(i)} \boldsymbol{A}_i^{(l,k)\top} \frac{\partial \mathcal{M}_{n+k-1,1}}{\partial \boldsymbol{m}}\left(\boldsymbol{Z}^{(\ell,k)}\right) + \mathrm{tr}\left(A^{(l,k)}\Gamma'^{(i,j)}A^{(l,k)\top}\frac{\partial \mathcal{M}_{n+k-1,1}}{\partial \Sigma}\left(\boldsymbol{Z}^{(\ell,k)}\right)\right), \tag{13}$$

where $\Gamma'^{(i,j)} = \left(\Sigma'^{(i)}_{u,j}\delta_{i,v} + \Sigma'^{(i)}_{v,j}\delta_{i,u}\right)_{u,v} \in \mathbb{R}^{q\times q}$, and δ is the Kronecker symbol. The derivatives $\frac{\partial \mathcal{M}_{n+k-1,1}}{\partial \boldsymbol{m}}$ and $\frac{\partial \mathcal{M}_{n+k-1,1}}{\partial \Sigma}$ are calculated in Appendix B.2.

This new expansion of the gradient of the generalized EI as a sum of derivatives of first order moments will prove to be very useful thanks to formulas presented next.

3.3 Fast numerical estimation of first order moments and their derivatives

Let us now focus on the practical implementation of the closed-form formula of Equation (10). We take $\alpha = 1$ and note $p = n + q$ in noisy settings and $p = q$ in noiseless settings. As mentioned before, the computation of the noisy or noiseless q-EI (see, Eqs. (10),(11)) requires calls to the CDF of the p and $(p-1)-$variate normal distribution, Φ_p and Φ_{p-1}. These CDFs are here computed using the Fortran algorithms of [16] wrapped in the mnormt R package [2]. A quick look at Eqs. (8),(10) suggests that the noisy q-EI requires nq evaluations of Φ_p and nq^2 evaluations of Φ_{p-1}. For the noiseless case (see, Equation (11)), the number of calls are divided by n. In both cases, a slight improvement can be obtained by noticing a symmetry which reduces the number of Φ_{p-1} calls from nq^2 (resp. q^2 in the noiseless case) to $nq(q+1)/2$ (resp. $q(q+1)/2$). This symmetry is justified in Appendix E.

Despite this improvement, and even in the classical noiseless case, the number of Φ_{p-1} calls is still proportional to q^2. We now give new efficient and trustworthy expansion that enables a fast and reliable approximation of first order moments of truncated Gaussian vectors $\mathcal{M}_{k,1}$ by reducing this number of calls to $O(q)$.

Proposition 4. *Let $\varepsilon > 0$, and let $\boldsymbol{Z}$ be a Gaussian random vector with mean vector and covariance matrix $(\boldsymbol{m}, \Sigma) \in \mathbb{R}^p \times S^p_{++}$. Then we have*

$$\mathcal{M}_{k,1}(\boldsymbol{m}, \Sigma) = \frac{1}{\varepsilon}\left(e^{m_k\varepsilon}\Phi_{p,\Sigma}(-\varepsilon\boldsymbol{\Sigma}_k - \boldsymbol{m}) - \Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2). \tag{14}$$

Proof. Let us consider the function $g_k : t \in \mathbb{R} \to e^{m_k t}\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_k t - \boldsymbol{m})$. This function g_k is tangent at $t = 0$ with the function $t \in \mathbb{R} \to \mathcal{G}(t\boldsymbol{e}_k)$, where the function $\mathcal{G}$ is introduced in Proposition 1 and $\boldsymbol{e}_k$ is the k^{th} vector of the canonical basis. It follows from Proposition 1 that

$$\mathcal{M}_{k,1}(\boldsymbol{m}, \Sigma) = \frac{\partial \mathcal{G}}{\partial t_k}(\boldsymbol{t}, \boldsymbol{m}, \Sigma)\bigg|_{t=0},$$

and we obtain the announced result by Taylor expansion of g_k. □

The obtained formula simply uses the approximation of a moment with finite differences of the moment generating function. We showed here that instead of fully computing the moment generating function, we can expand the simpler *tangent* function g_k. For conciseness, we name here the use of this formula "tangent moment method". This formula thus enables approximating the first order moment $\mathcal{M}_{k,1}$ at the cost of only two calls to Φ_p. Hence, from Equation (11), computing a noiseless q-EI can be performed at the cost of $2q$ calls to Φ_q. Besides, a similar approach can be applied to approximate the gradient of q-EI through faster computations of $\frac{\partial \mathcal{M}_{k,1}}{\partial \boldsymbol{m}}$ and $\frac{\partial \mathcal{M}_{k,1}}{\partial \Sigma}$, as shown next:

Proposition 5. *The following equations hold:*

$$\frac{\partial \mathcal{M}_{k,1}}{\partial \boldsymbol{m}} = \Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{e}_k - \frac{1}{\varepsilon}\left(e^{m_k\varepsilon}\nabla\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_k\varepsilon - \boldsymbol{m}) - \nabla\Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2) \tag{15}$$

$$\begin{aligned}\frac{\partial \mathcal{M}_{k,1}}{\partial \Sigma} = &-\left(\frac{\partial \Phi_{p,\Sigma}}{\partial x_v}(-\boldsymbol{m})\,\delta_{u,k} + \frac{\partial \Phi_{p,\Sigma}}{\partial x_u}(-\boldsymbol{m})\,\delta_{v,k}\right)_{u,v\leq p} \\ &+ \frac{1}{\varepsilon}\left(e^{m_k\varepsilon}\nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{\Sigma}_k\varepsilon - \boldsymbol{m}) - \nabla\nabla^\top\Phi_{p,\Sigma}(-\boldsymbol{m})\right) + O(\varepsilon^2)\end{aligned} \tag{16}$$

where $\nabla\nabla^\top\Phi_{p,\Sigma}$ is the Hessian matrix of $\Phi_{p,\Sigma}$ (see Appendix A.2 for details).

As before, these formulas enable reducing the number of calls to the multivariate CDF by an order q. For the computation of q-EI this number goes from $O(q^2)$ to $O(q)$. For computing its dq-dimensional gradient, it goes from $O(q^4)$ to $O(q^3)$. The latter complexity suggests restricting to moderate values of q in applications. In the next section we present further results that enable further reducing the complexity for numerically estimating the gradient.

3.4 A slightly biased but fast proxy of the gradient

The key idea to obtain further computational savings is summarized in this section. We first strategically decompose the gradient of moments as a sum of two terms.

Proposition 6. *Let us consider a Gaussian multivariate random field $\boldsymbol{Z}$ from $\mathbb{R}^d$ to $\mathbb{R}^p$. For $\boldsymbol{x} \in \mathbb{R}^d$, let us denote by $\boldsymbol{m}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$ the mean and the covariance matrix of $\boldsymbol{Z}(\boldsymbol{x})$. Let $\boldsymbol{x}_a \in \mathbb{R}^d$ and assume that $\Sigma(\boldsymbol{x}_a)$ is positive definite. Also, assume that the functions $\boldsymbol{x} \to \boldsymbol{m}(\boldsymbol{x})$, $\boldsymbol{x} \to \Sigma(\boldsymbol{x})$ and $\boldsymbol{x} \to (\mathrm{cov}(Z_i(\boldsymbol{x}), Z_j(\boldsymbol{x}_a)))_{i,j\leq p}$ are differentiable at $\boldsymbol{x} = \boldsymbol{x}_a$. Then the following decomposition holds for $k = 1, \ldots, p$.*

$$\begin{aligned}&\nabla_{\boldsymbol{x}}\left[\mathcal{M}_{k,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} := \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z_k^\alpha(\boldsymbol{x})1_{\{\boldsymbol{Z}(\boldsymbol{x})\leq \boldsymbol{0}\}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} \\ &= \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z_k^\alpha(\boldsymbol{x})1_{\{\boldsymbol{Z}(\boldsymbol{x}_a)\leq \boldsymbol{0}\}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} + \nabla_{\boldsymbol{x}}\left[\mathbb{E}\left(Z_k^\alpha(\boldsymbol{x}_a)1_{\{\boldsymbol{Z}(\boldsymbol{x})\leq \boldsymbol{0}\}}\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a}.\end{aligned} \tag{17}$$

Proof. $\Sigma(\cdot)$ is continuous at $\boldsymbol{x}_a$, so there exists a neightborhood $V_{\boldsymbol{x}_a}$ of $\boldsymbol{x}_a$ such that for all $\boldsymbol{x} \in V_{\boldsymbol{x}_a}$, $\Sigma(\boldsymbol{x})$ is positive definite. Let us define on $V_{\boldsymbol{x}_a} \times V_{\boldsymbol{x}_a}$:

$$g(\boldsymbol{u}, \boldsymbol{v}) = \mathbb{E}\left(Z_k^\alpha(\boldsymbol{u})1_{\{\boldsymbol{Z}(\boldsymbol{v})\leq \boldsymbol{0}\}}\right).$$

Applying equation (22) of appendix B.1, for all $\boldsymbol{u}$ and $\boldsymbol{v}$, $g(\boldsymbol{u}, \boldsymbol{v})$ is a moment generated by differentiation of the following function:

$$M_{\boldsymbol{u},\boldsymbol{v}} : t \to e^{\frac{1}{2}\left(\Sigma_{kk}(\boldsymbol{u})t^2 + 2tm_k(\boldsymbol{u})\right)}\Phi_{p,\Sigma(\boldsymbol{v})}\left(-\boldsymbol{m}(\boldsymbol{v}) - t\left(\mathrm{cov}(Z_k(\boldsymbol{u}), Z_j(\boldsymbol{v}))\right)_{j\leq p}^\top\right). \tag{18}$$

The analytical form of equation (18) and the assumed differentiability at $\boldsymbol{x}_a$ ensure existence of partial derivatives of $g = (\boldsymbol{u}, \boldsymbol{v}) \to \frac{\mathrm{d}^\alpha M_{\boldsymbol{u},\boldsymbol{v}}}{\mathrm{d}t^\alpha}(0)$ at $(\boldsymbol{x}_a, \boldsymbol{x}_a)$. So to conclude,

$$\begin{aligned}\nabla_{\boldsymbol{x}} \left[\mathcal{M}_{k,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\right)\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} &= \nabla_{\boldsymbol{x}} \left[g(\boldsymbol{x}, \boldsymbol{x})\right]\big|_{\boldsymbol{x}=\boldsymbol{x}_a} \\ &= \frac{\partial}{\partial \boldsymbol{u}} \left[g(\boldsymbol{u}, \boldsymbol{x}_a)\right]\Big|_{\boldsymbol{u}=\boldsymbol{x}_a} + \frac{\partial}{\partial \boldsymbol{v}} \left[g(\boldsymbol{x}_a, \boldsymbol{v})\right]\Big|_{\boldsymbol{v}=\boldsymbol{x}_a}\end{aligned}$$

□

The latter decomposition can be interpreted as follows: infinitesimal variations of $(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x}))$ around $(\boldsymbol{m}(\boldsymbol{x}_a), \Sigma(\boldsymbol{x}_a))$ modify the moments $\mathcal{M}_{k,\alpha}\left(\boldsymbol{m}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\right)$ in two ways. First, it modifies the distribution of $Z_k^\alpha(\boldsymbol{x})$, second it changes the distribution of the truncation $1_{\{\boldsymbol{Z}(\boldsymbol{x}) \leq \boldsymbol{0}\}}$. For the particular case of q-EI, we propose to neglect this second variation. Applying this approximation to (11) gives for $X_0 \in D^q$,

$$\begin{aligned}&\nabla_{\boldsymbol{x}_{n+j}} EI(\boldsymbol{x}_{n+1:n+q})\big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= \sum_{k=1}^{q} \nabla_{\boldsymbol{x}_{n+j}} \mathbb{E}\left(\left(T - Y(\boldsymbol{x}_{n+k})\right)^\alpha 1_{\left\{A^{(k)} Y(\boldsymbol{x}_{n+1:n+q}) \leq \boldsymbol{0}\right\}}\right)\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &\approx \sum_{k=1}^{q} \nabla_{\boldsymbol{x}_{n+j}} \mathbb{E}\left(\left(T - Y(\boldsymbol{x}_{n+k})\right)^\alpha 1_{\left\{A^{(k)} Y(X_0) \leq \boldsymbol{0}\right\}}\right)\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= -\nabla_{\boldsymbol{x}_{n+j}} \mathbb{E}\left(Y(\boldsymbol{x}_{n+j})^\alpha 1_{\left\{A^{(j)} Y(X_0) \leq \boldsymbol{0}\right\}}\right)\Big|_{\boldsymbol{x}_{n+1:n+q}=X_0} \\ &= -\mathbb{E}\left(\nabla_{\boldsymbol{x}_{n+j}} Y(\boldsymbol{x}_{n+j})^\alpha|_{\boldsymbol{x}_{n+1:n+q}=X_0} 1_{\left\{A^{(j)} Y(X_0) \leq \boldsymbol{0}\right\}}\right), \end{aligned} \tag{19}$$

where the last step is obtained by mean square differentiability of the process $\boldsymbol{x} \to Y(\boldsymbol{x})^\alpha 1_{\{B\}}$, with B an event constant with respect to $\boldsymbol{x}$, see Appendix D. We can observe that this approximation makes a summation term disappear. The computation of this formula requires $(d+1)$ evaluations of q-variate Gaussian CDF. Indeed, Equation (19) indicates that each component of the gradient vector can be considered as a moment of a truncated Gaussian vector, so we can apply the results of section 2. In particular, when $\alpha = 1$, applying Proposition 4, two Gaussian CDF calls are needed for each of the d components, leading to $2d$ evaluations. Besides, from Equation (14), the second CDF call does not depend on k, which implies that this term is common for every dimension. Thus the gradient of Equation (19) finally comes with $d+1$ CDF evaluations instead of $2d$. For a full gradient with respect to all q points of the batch, we then need $q(d+1)$ CDF evaluations – a substantial improvement compared to the $O(q^4)$ obtained in [29] and the $O(q^3)$ obtained in the previous section. The complexities for computing moments, q-EI and its gradients, expressed in terms of number of calls to the Φ function, are summarized in Table 1. These new computational savings come at the price of a non-exact gradient calculation. A first numerical validation is represented in Figure 2. On this example, we observe small (1×10^{-2}) relative errors between the exact and approximate gradient of dimension $q \times d = 4$ (the biggest difference vector has a norm of 0.13, compared to an exact gradient norm of 13.1). We also observe that the relative error appears to be typically smaller with higher q-EI, which is promising for q-EI maximisations. However, this apparently trustful but non-exact calculation naturally raises the question of the impact of such an approximation on the performances of gradient-based q-EI maximization algorithms. As we will see in the next section, this proxy gradient turned out to enable quite competitive q-EI maximization performances based on numerical experiments.

4 Application

The goal of this section is to illustrate the usability of the proposed gradient-based q-EI maximization schemes and in particular the improvements brought by the fast formulas detailed in the previous

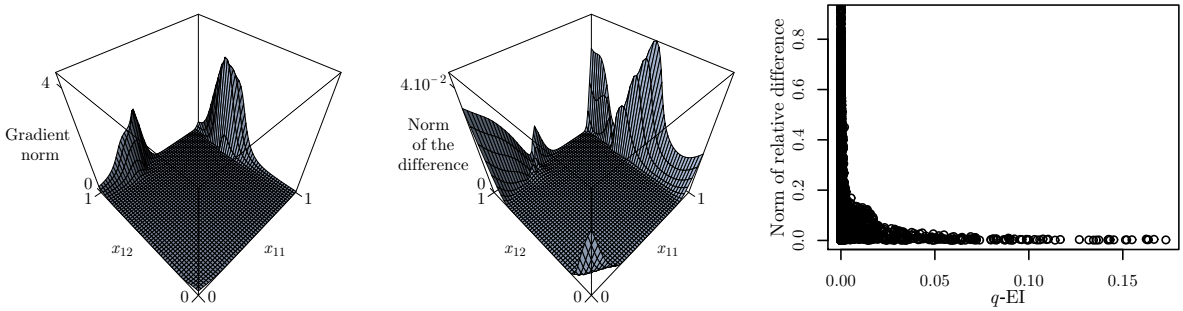

Figure 2: Numerical validation of the approximation from equation (19), with $\alpha = 1$, $q = 2$, $d = 2$. From left to right: 1) Norm of the q-EI gradient, with respect to the first batch point (the other point is fixed in the center of $[0,1]^d$) ; 2) Norm of the difference vector between the analytical gradient and its approximation ; 3) Relative error (norm of the difference divided by the real norm) computed on 3000 random batches sampled uniformly in $[0,1]^{d\times q}$, with respect to their q-EI.

		Number of CDF evaluations				
		Φ_{q-3}	Φ_{q-2}	Φ_{q-1}	Φ_q	Total
$\mathcal{M}_{k,1}$	analytic			q	1	$O(q)$
	tangent moment				2	2
EI	analytic			$\binom{q+1}{2}$	q	$O(q^2)$
	tangent moment				$2q$	$O(q)$
$\nabla\mathcal{M}_{k,1}$	analytic	$3\binom{q}{3}$	$3\binom{q}{2}$	$2q$	1	$O(q^3)$
	tangent moment		$2\binom{q}{2}$	$2q$	2	$O(q^2)$
	proxy				$d+1$	$O(d)$
∇EI	analytic	$6\binom{q+1}{4}$	$3\binom{q+1}{3}$	$(3q^2+q)/2$	q	$O(q^4)$
	tangent moment		$q^2(q-1)$	$2q^2$	$2q$	$O(q^3)$
	proxy				$q(d+1)$	$O(qd)$

Table 1: In noiseless settings, total number of calls to the CDF of the multivariate Gaussian distribution for computing $\mathcal{M}_{k,1}$, q-EI, their gradients and their approximations, depending on q and d. For q-EI in noisy setting, replace q by $p = n + q$ and multiply each number of calls by n.

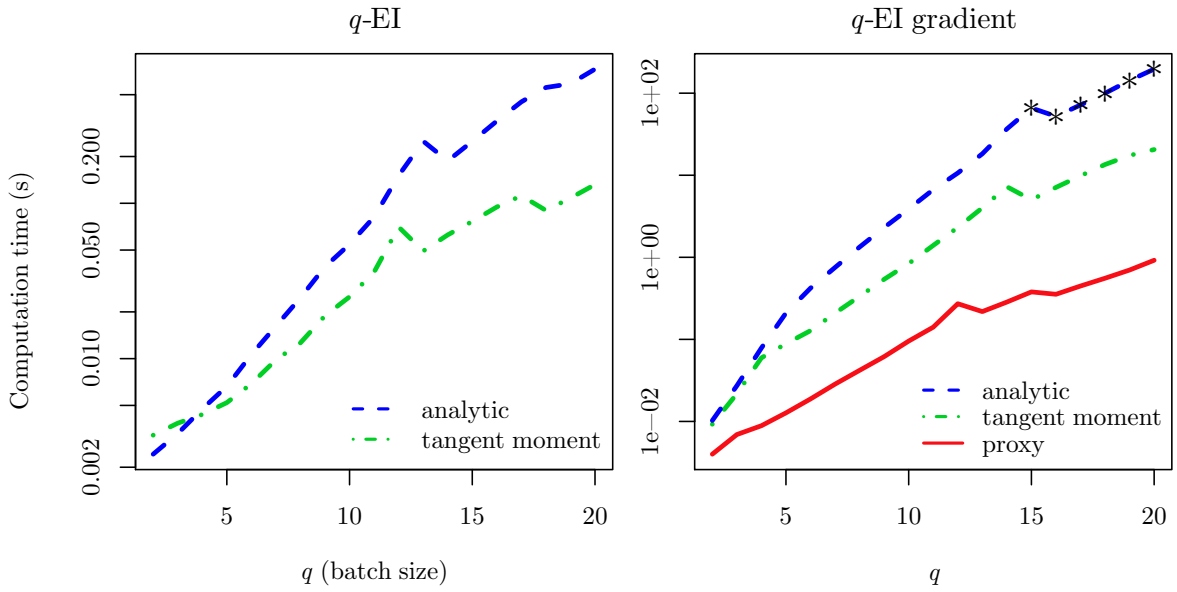

Figure 3: Computation times for q-EI or its gradient as a function of the batch size q (logarithmic scale). We take an averaged computation time over 1000 batches (except for points marked with a $*$, averaged over 150 batches).

sections. The relevance of using sequential sampling strategies based on the q-EI maximization has already been investigated before (see, [7, 47, 29]) and all these articles pointed out the importance of calculation speed which often limits the use of q-EI based strategies to moderate q. We do not aim again at proving the performance of q-EI based sequential strategies. Instead we aim at illustrating the gain, in computation time, brought by the fast formulas and show that using the approximate gradient obtained in Equation (19) does not impair the ability to find batches with (close to) maximal q-EI.

4.1 Objective function and pure calculation speed

The objective function is the so-called Borehole function [21]. It has been previously used for testing methods using a surrogate model [48, 20]. The function computes a rate of water flow, ϕ, through a borehole. The problem is described by $d = 8$ input variables, $r_w \in [0.05, 0.15]$, $r \in [100, 50000]$, $T_u \in [63070, 115600]$, $H_u \in [990, 1110]$, $T_l \in [63.1, 116]$, $H_l \in [700, 820]$, $L \in [1120, 1680]$, $K_w \in [1500, 15000]$ and is given below

$$\phi = \frac{2\pi T_u (H_u - H_l)}{\ln\left(\frac{r}{r_w}\right)\left(1 + \frac{2LT_u}{\left(\ln\left(\frac{r}{r_w}\right) r_w^2 K_w\right)} + \frac{T_u}{T_l}\right)}. \tag{20}$$

Here, the objective function f is obtained by rescaling ϕ on the input domain $D = [0, 1]^8$. An analytical study of variations shows that there is a unique global minimum at $\boldsymbol{x}^* = (0, 1, 0, 0, 0, 1, 1, 0)^\top$, with $f(\boldsymbol{x}^*) \approx 1.1918$.

Before using sequential strategies to minimize f, we look at empirical computation times for evaluating q-EI and its gradient as a function of the batch size q. For the computations, the so-called "analytic" method relies on the state of the art formulas of [7, 29] with a number of calls to the multivariate normal cdf of respectively $O(q^2)$ and $O(q^4)$. The "tangent moment" method uses our formula for moment calculation to yield q-EI and its gradient (see, Equations (14) and (15),(16)). Finally, for computing the gradient only, the "proxy" method relies on Equation (19).

Figure 3 exhibits computation times averaged over 1000 batches drawn uniformly. The Gaussian process model is based on an initial design of $n_0 = 10d = 80$ points drawn from a optimum-LHS procedure [25]. We use the Matérn ($\nu = 3/2$) tensor product covariance function and estimate the hyperparameters by maximum likelihood using the DiceKriging R package [11]. Figure 3 shows significant computational savings. For instance with $q = 8$, one gradient computation takes respectively $0.04s$, $0.33s$ and $1.33s$ using respectively the proxy, the tangent moment and analytic methods. Since the complexity for computing a gradient with the proxy is of $O(qd)$ against $O(q^3)$ and $O(q^4)$ for the two other methods, the computational savings of the proxy tend to increase with q. It should also be noted that these savings will be larger with decreasing domain dimension d. If we look at q-EI computations, the tangent moment method is 3.3 times faster than the analytic one when $q = 8$ and 6.5 times faster when $q = 20$; thanks to an $O(q)$ complexity against $O(q^2)$.

4.2 Experimental setup: sequential minimization strategies

We now perform a total of 50 minimizations of f, each using an initial design of experiments of $n_0 = 80$ points drawn from an optimum-LHS procedure with a different seed. Three different batch-sequential strategies are investigated.

The first one – serving as a benchmark – is a variation of the "Constant Liar Mix" heuristic [7, 47] where, at each iteration, the batch of size q is chosen among several batches obtained from the Constant Liar heuristic [17] with different lie levels. We use 7 lie levels fixed to the current maximum observation the current minimum observation, and the $2.5\%, 10\%, 50\%, 90\%, 97.5\%$ quantiles of the conditional distribution of the point selected in the batch. A total of 7 batches are proposed at each iteration and the CL-mix heuristic picks the one with maximum q-EI.

The two other strategies considered here rely on pure q-EI maximization using a multistart BFGS algorithm with a stopping criterion of precision 2.2×10^{-7} (parameter `control$factr` of the R function `optim` [37]). The gradients involved in the optimization are computed either with the tangent moment formula or the proxy. For the gradient-based q-EI maximization, we use a total of 10 starting batches obtained, again, using a Constant Liar heuristic with random lies sampled from the conditional distribution at the selected point. Finally we use two different batch sizes. When $q = 8$ we run a total of 10 iterations and when $q = 4$ we run 20 iterations. The hyperparameters of the GP model are re-estimated at each iteration after having incorporated the new observations.

4.3 First q-EI maximization

We first compare the performances, in terms of q-EI, of the multistart BFGS algorithm when the proxy gradient and the tangent moment methods are used. Table 2 compares the results at iteration 1 for these two methods and the CL-mix strategy. The results are averaged over the 50 initial designs.

	$q = 4$	$q = 8$
tangent moment	12.45 (22.6 s)	15.35 (700.2 s)
proxy	12.46 (14.3 s)	15.35 (127.0 s)
CL-mix	11.80 (7.7 s)	14.34 (15.6 s)

Table 2: Average q-EI value of the optimal batches found for each of the 50 initial designs. The numbers between brackets are the average computation times.

As expected, the CL-mix heuristic yields batches with lower q-EI than the strategies directly maximizing q-EI. Also, for both $q = 4$ and $q = 8$, the two q-EI based methods have the same performance, which stresses out the relevance of the proxy method since the latter is about 1.6 times faster when $q = 4$ and 5.5 faster when $q = 8$.

4.4 Several q-EI maximization steps

We now compare the performances of the different q-EI maximization approaches after multiple batch evaluations. Figure 4 displays the average regret as a function of the iteration number (first row) and the total computation time (i.e. the time to evaluate f and find the next batch to evaluate) assuming respectively that the computation time of f is 0 seconds (i.e. instantaneous), two minutes and one hour (rows $2, 3, 4$ respectively). Looking at the performances as a function of the iteration number (first row on Figure 4), the CL-mix heuristic, which samples a batch with lower q-EI at each step, leads in average to a slower convergence than the two other methods, for both $q = 4$ and $q = 8$. In contrast, the two strategies based on q-EI maximization have similar performances.

However, these conclusions do not hold when the regret is plotted as a function of the total computation time (rows $2, 3, 4$ on Figure 4). First, when the computation time t_{eval} of f is null (row 2) it is clear that q-EI-based sequential strategies are not adapted since they are too expensive. In this case, the CL-mix heuristic performs better and some other optimization strategies which are not metamodel-based would probably be more relevant. Second, when f is moderately expensive (i.e. $t_{\text{eval}} = 2$ minutes), the proxy method and CL-mix have comparable performances when $q = 8$, but the proxy outperforms when $q = 4$. Besides, the proxy shows a much faster convergence than the tangent moment method when $q = 8$. The use of q-EI based strategies thus becomes relevant when t_{eval} is larger than a few minutes, if the proxy is used. Finally, when t_{eval} is equal to one hour, the use of q-EI based strategies is particularly recommended. In that case the relative improvement of the proxy compared to the tangent moment method tends to naturally vanish because of the long computation time of f. When f is extremely expensive to compute, using the proxy is thus not essential. However, since it does not impair the ability to find a batch with large q-EI we still recommend to use it, especially when q is large.

Conclusion

In this article we provide a closed-form expression of generalized q-points Expected Improvement criterion for batch-sequential Bayesian global optimization. An interpretation based on moments of truncated Gaussian vectors yields fast q-EI formulas with arbitrary precision. Furthermore an new approximation for the gradient is shown to be even faster while preserving ability to find batches close to maximal q-EI. As the use of these strategies was previously considered cumbersome from a dozen of batch points, these formulas happen to be of particular interest to run q-EI based batch-sequential strategies for larger batch sizes. We show that these methods are implementable and efficient on a classic 8-dimensional test case. Additionally, some of the intermediate results established here might be of interest for other research questions involving moments of truncated Gaussian vectors and their gradients. Perspectives include deriving second order derivatives of q-EI and fast numerical estimates thereof. Also, we aim at improving the sampling of initial batches in multistart derivative-based q-EI maximization.

Acknowledgements: Part of this work has been conducted within the frame of the ReDice consortium, gathering industrial (CEA, EDF, IFPEN, IRSN, Renault) and academic (École des Mines de Saint-Étienne, INRIA, Universität Bern) partners around advanced methods for computer experiments.

A Differentiating multivariate Gaussian CDF

We consider the CDF dimension $p \geq 2$. We use the convention $\Phi_0 = 1$.

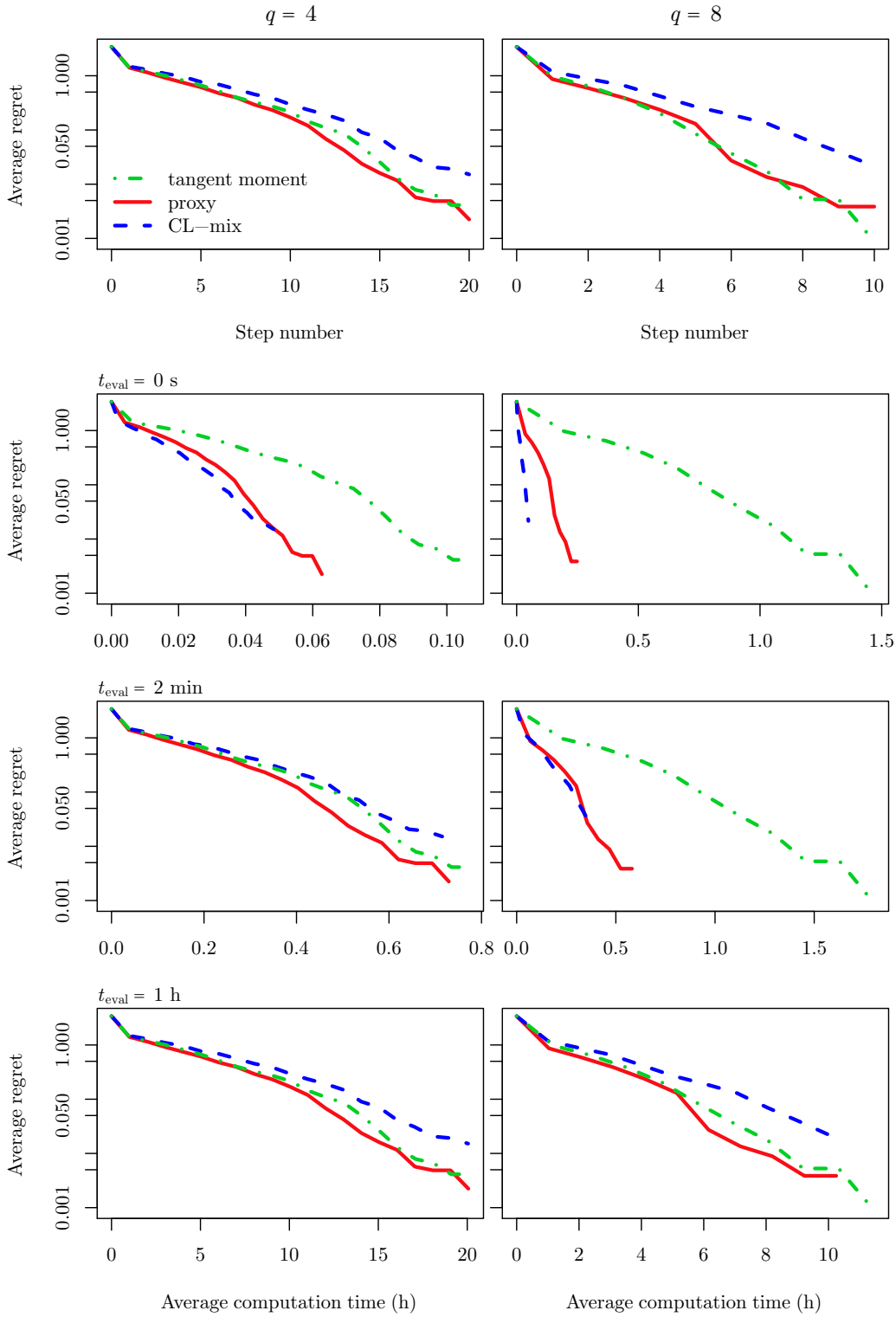

Figure 4: Log-scaled average regret of the three considered optimization strategies as a function of the iteration number (row 1) and the total computation time (rows $2, 3, 4$) assuming that the computation times of f, t_{eval}, are respectively 0 seconds, 2 minutes and 1 hour. Experiments are performed with $q = 4$ (left column) and $q = 8$ (right column).

A.1 Gradient

Using the following identity, derived from conditional distributions of a Gaussian vector,

$$\forall i = 1, \dots, p,\ \varphi_{p,\Sigma}(\boldsymbol{x}) = \varphi_{1,\Sigma_{ii}}(x_i)\, \varphi_{p-1,\Sigma_{|i}}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,x_i}\right),$$

with $\boldsymbol{m}_{|i,u} = \frac{u}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}$ and $\Sigma_{|i} = \Sigma_{-i,-i} - \frac{1}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{-i,i}^{\top}$, we reformulate the integral of the Gaussian CDF:

$$\forall i = 1, \dots, p, \Phi_{p,\Sigma}(\boldsymbol{x}) = \int_{-\infty}^{x_i} \varphi_{1,\Sigma_{ii}}(u_i)\, \Phi_{p-1,\Sigma_{|i}}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,u_i}\right) \mathrm{d}u_i.$$

Here indexed minus symbols, e.g. in $\boldsymbol{\Sigma}_{-i,i}$, refer to exclusions of a line or a column.

Finally we have

$$\nabla \Phi_{p,\Sigma}(\boldsymbol{x}) = \left(\varphi_{1,\Sigma_{ii}}(x_i)\, \Phi_{p-1,\Sigma|i}\left(\boldsymbol{x}_{-i} - \boldsymbol{m}_{|i,x_i}\right)\right)_{i=1,\dots,p}. \tag{21}$$

A.2 Hessian

As for the computation of the gradient, we write
$\forall i, j = 1, \dots, p, i \neq j$,

$$\Phi_{p,\Sigma}(\boldsymbol{x}) = \int_{-\infty}^{x_i}\int_{-\infty}^{x_j} \varphi_{2,\Sigma_{ij,ij}}\left(\begin{bmatrix} u_i \\ u_j \end{bmatrix}\right) \Phi_{p-2,\Sigma_{|ij}}\left(\boldsymbol{x}_{-\{i,j\}} - \boldsymbol{m}_{|(i,j),(u_i,u_j)}\right) \mathrm{d}u_j \mathrm{d}u_i,$$

with $\boldsymbol{m}_{|(i,j),(u,u')} = \Sigma_{-\{ij\},ij}\Sigma_{ij,ij}^{-1}\begin{bmatrix} u \\ u' \end{bmatrix}$ and $\Sigma_{|ij} = \Sigma_{-\{ij\},-\{ij\}} - \Sigma_{-\{ij\},ij}\Sigma_{ij,ij}^{-1}\Sigma_{-\{ij\},ij}^{\top}$.

So $\forall i, j = 1, \dots, p, i \neq j$,

$$\frac{\partial^2 \Phi_q}{\partial x_i \partial x_j}(\boldsymbol{x}) = \varphi_{2,\Sigma_{ij,ij}}\left(\begin{bmatrix} x_i \\ x_j \end{bmatrix}\right)\ \Phi_{p-2,\Sigma_{|ij}}(\boldsymbol{x}_{-\{i,j\}} - \boldsymbol{m}_{|(i,j),(x_i,x_j)}).$$

When $i = j$, the differentiation of equation (21) gives,

$$\frac{\partial^2 \Phi_q}{\partial x_i^2}(\boldsymbol{x}) = -\frac{1}{\Sigma_{ii}}\left(x_i \frac{\partial \Phi_{p,\Sigma}}{\partial x_i}(\boldsymbol{x}) + \sum_{\substack{j=1 \\ j\neq i}}^{p} \Sigma_{ij} \frac{\partial^2 \Phi_{p,\Sigma}}{\partial x_i \partial x_j}(\boldsymbol{x})\right).$$

B Moments of truncated multivariate Gaussian distribution

B.1 Analytical formula (propositions 1 and 6)

We see here why we can derive an analytical formula of $\mathcal{M}_{k,\alpha}(\boldsymbol{m}, \Sigma)$, with $k \leq s \in \mathbb{N}\backslash\{0\}$, $\boldsymbol{m} \in \mathbb{R}^s$ and $\Sigma \in S_{++}^s$, by differentiating $\mathcal{G}$, defined in equation (6). It is known, see e.g. [10], that moments can be obtained differentiating the moment generating function $G_{\boldsymbol{m},\Sigma,s}$:

$$\mathcal{M}_{k,\alpha}(\boldsymbol{m}, \Sigma) = \frac{\partial^{\alpha} G_{\boldsymbol{m},\Sigma,s}}{\partial t_k^{\alpha}}(\boldsymbol{0}),$$

with, for $r \in \{1, \dots, s\}$, $G_{\boldsymbol{m},\Sigma,r} : \boldsymbol{t} \to \mathbb{E}\left(\exp\left(\boldsymbol{t}^{\top}\boldsymbol{Z}\right) 1_{\{(Z_1,\dots,Z_r)^{\top} \leq \boldsymbol{0}\}}\right)$, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{m}, \Sigma)$. We derive now an analytical formula for $G_{\boldsymbol{m},\Sigma,r}$. As needed in Proposition 6, we derive an analytical formula for any

r, and not only for $r = s$.

$$\begin{aligned}
&\forall \boldsymbol{t} \in \mathbb{R}^s, \\
G_{\boldsymbol{m},\Sigma,r}(\boldsymbol{t}) &= \overbrace{\int_{-\infty}^{0}\!\!\dots\int_{-\infty}^{0}}^{r \text{ times}} \int_{-\infty}^{\infty}\!\!\dots\int_{-\infty}^{\infty} \exp\left(\boldsymbol{t}^\top \boldsymbol{z}\right)\varphi_\Sigma\left(\boldsymbol{z}-\boldsymbol{m}\right)\ \mathrm{d}z_1 \cdots \mathrm{d}z_s \\
&= \varphi_\Sigma(\boldsymbol{0}) \int_{-\infty}^{0}\!\!\dots\int_{-\infty}^{0} \int_{-\infty}^{\infty}\!\!\dots\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\left(\boldsymbol{z}-\boldsymbol{m}\right)^\top \Sigma^{-1}\left(\boldsymbol{z}-\boldsymbol{m}\right) - 2\boldsymbol{t}^\top \boldsymbol{z}\right)\right)\mathrm{d}\boldsymbol{z} \\
&= e^{-\frac{1}{2}\left(-\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma \left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)+\boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)} \\
&\qquad \varphi_\Sigma(\boldsymbol{0}) \int_{-\infty}^{0}\!\!\dots\int_{-\infty}^{0} \int_{-\infty}^{\infty}\!\!\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left((\boldsymbol{z}-\boldsymbol{m}-\Sigma\boldsymbol{t})^\top\Sigma^{-1}(\boldsymbol{z}-\boldsymbol{m}-\Sigma\boldsymbol{t})^\top\right)}\mathrm{d}\boldsymbol{z} \\
&= e^{\frac{1}{2}\left(\left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)^\top \Sigma \left(\boldsymbol{t}+\Sigma^{-1}\boldsymbol{m}\right)-\boldsymbol{m}^\top\Sigma^{-1}\boldsymbol{m}\right)} \Phi_{r,(\Sigma_{ij})_{i,j\le r}}\left(-\boldsymbol{m} - (\Sigma_{ij})_{i\le r, j\le s}\,\boldsymbol{t}\right). \qquad (22)
\end{aligned}$$

In the frame of the proof of Proposition 6,

- if $\Sigma_k(\boldsymbol{u},\boldsymbol{v})$, the covariance matrix of $(\boldsymbol{Z}(\boldsymbol{v})^\top, Z_k(\boldsymbol{u}))^\top$, is positive definite, we take

$$M_{\boldsymbol{u},\boldsymbol{v}} = t \to G_{(\boldsymbol{m}(\boldsymbol{v}),m_k(\boldsymbol{u})),\Sigma_k(\boldsymbol{u},\boldsymbol{v}),p}((0,\dots,0,t)^\top),$$

- else, as $\Sigma(\boldsymbol{v})$ is definite positive, there exists only one index k_0 such as $Z_k(\boldsymbol{u}) = Z_{k_0}(\boldsymbol{v})$ almost surely (for example $k = k_0$ when $\boldsymbol{u} = \boldsymbol{v}$), and we have

$$M_{\boldsymbol{u},\boldsymbol{v}} = t \to G_{(\boldsymbol{m}(\boldsymbol{v})),\Sigma(\boldsymbol{v}),p}((0,\dots,\underset{\substack{\uparrow\\ k_0^{\text{th}}\text{ position}}}{t},\dots,0)^\top).$$

In both cases, equation (22) leads to equation (18).

B.2 Differentiation with respect to mean and covariance

We differentiate here the equation (8) with respect to $\boldsymbol{m}$ and Σ.

With respect to the mean $\boldsymbol{m}$

$$\frac{\partial \mathcal{M}_{k,1}}{\partial \boldsymbol{m}}(\boldsymbol{m},\Sigma) = \Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{e}_k - m_k \nabla\Phi_{p,\Sigma}(-\boldsymbol{m}) + \nabla\nabla^\top \Phi_{p,\Sigma}(-\boldsymbol{m})\boldsymbol{\Sigma}_k. \qquad (23)$$

With respect to the covariance Σ

$$\begin{aligned}
\frac{\partial \mathcal{M}_{k,1}}{\partial \Sigma}(\boldsymbol{m},\Sigma) = m_k \frac{\partial}{\partial\Sigma}\Phi_{p,\Sigma}(-\boldsymbol{m}) - \sum_{i=1}^{p} &\varphi_{\Sigma_{ii}}\left(-m_i\right)\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i})E^{(k,i)} \\
&+\Sigma_{ki}\frac{\partial}{\partial\Sigma_{ii}}\varphi_{\Sigma_{ii}}\left(-m_i\right)\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i})E^{(i,i)} \\
&+\Sigma_{ki}\varphi_{\Sigma_{ii}}\left(-m_i\right)\frac{\partial}{\partial\Sigma}\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i}). \qquad (24)
\end{aligned}$$

with $\boldsymbol{m}|_i = \boldsymbol{m}_{-i} - \frac{m_i}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}$ and $\Sigma|_i = \Sigma_{-i,-i} - \frac{1}{\Sigma_{ii}}\boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{-i,i}^\top$. Writting $\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right]$ $\mathrm{d}_\Sigma\left[\Sigma|_i\right]$ the differential of the functions $\Sigma \to \boldsymbol{m}|_i$ and $\Sigma \to \Sigma|_i$, we have:

$$\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right](H) = \frac{m_i}{\Sigma_{ii}}\boldsymbol{H}_{-i,i} \tag{25}$$

$$\mathrm{d}_\Sigma\left[\Sigma|_i\right](H) = H_{-i,-i} + \frac{H_{ii}}{\Sigma_{ii}^2}\boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{-i,i}^\top - \frac{2}{\Sigma_{ii}}\boldsymbol{H}_{-i,i}\boldsymbol{\Sigma}_{-i,i}^\top \tag{26}$$

$$\frac{\partial}{\partial\Sigma}\Phi_{p-1,\Sigma|i}(-\boldsymbol{m}_{|i}) = \sum_{r=1}^{p}\sum_{s=1}^{p}\Big(-\mathrm{d}_\Sigma\left[\boldsymbol{m}|_i\right](E^{(r,s)}).\nabla\Phi_{p-1,\Sigma|_i}(-\boldsymbol{m}|_i) + \mathrm{tr}\left(\frac{\partial}{\partial\Gamma}\Phi_{p-1,\Sigma|i}\left(-\boldsymbol{m}_{|i}\right).\mathrm{d}_\Sigma\left[\Sigma_{|i}\right]\left(E^{(r,s)}\right)\right)\Big)E^{(r,s)}$$

with:

- $E^{(r,s)} = (\delta_{ij})_{i,j=1,\ldots,p}$,
- $\frac{\partial}{\partial\Gamma}\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i})$ the derivative of $\Gamma \to \Phi_{p-1,\Gamma}(-\boldsymbol{m}_{|i})$ evaluated at $\Sigma_{|i}$. We use the Plackett's differential equation, extended by [5], to find

$$\frac{\partial}{\partial\Gamma}\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i}) = \nabla\nabla^\top\Phi_{p-1,\Sigma_{|i}}(-\boldsymbol{m}_{|i}),$$

  $\nabla\nabla^\top\Phi$ is given in appendix A.2.

C Generalized *q*-EI as a sum of moments

Proof. For given (ℓ,k) in $\{1,\ldots,n\}\times\{1,\ldots,q\}$, we consider $E_{\ell,k}$ the event that the random variable inside the expectation term of equation (4) equals $(Y(\boldsymbol{x}_\ell) - Y(\boldsymbol{x}_{n+k}))^\alpha$. We have

$$\begin{aligned} E_{\ell,k} = \{Y(\boldsymbol{x}_{n+k}) \le Y(\boldsymbol{x}_\ell)\} \;&\cap\; \{\forall i \le n, i \ne \ell; Y(\boldsymbol{x}_\ell) \le Y(\boldsymbol{x}_i)\} \\ &\cap\; \{\forall j \le q, j \ne k; Y(\boldsymbol{x}_{n+k}) \le Y(\boldsymbol{x}_{n+j})\} \end{aligned}$$

Considering all pairs $(\ell,\, k)$, we have:

$$EI_n(\boldsymbol{x}_{n+1:n+q}) = \sum_{\ell=1}^{n}\sum_{k=1}^{q}\mathbb{E}_n\left(\left(Y(\boldsymbol{x}_\ell) - Y(\boldsymbol{x}_{n+k})\right)^\alpha 1_{\{E_{\ell,k}\}}\right).$$

For each term (ℓ,k) of the sum, the conditioning event can be rewritten $E_{\ell,k} = \{\boldsymbol{Z}^{(\ell,k)}(\boldsymbol{x}_{n+1:n+q}) \le \boldsymbol{0}\}$, with $\boldsymbol{Z}^{(\ell,k)}$ a random vector of size $n+q-1$, defined by the following linear transformation of $\boldsymbol{Y} = (Y(\boldsymbol{x}_1),\ldots,Y(\boldsymbol{x}_{n+q}))^\top$:

$$\forall i = 1,\ldots,n+q-1, Z_i^{(\ell,k)} = \begin{cases} Y_\ell - Y_i & \text{if } 1 \le i \le \ell-1 \\ Y_\ell - Y_{i+1} & \text{if } \ell \le i \le n-1 \\ Y_{n+k} - Y_{i+1} & \text{if } n \le i \le n+q-1, i \ne n+k-1 \\ Y_{n+k} - Y_\ell & \text{if } i = n+k-1 \end{cases}$$

Indeed, the first $n-1$ components of $\boldsymbol{Z}^{(l,k)} \le \boldsymbol{0}$ reflect $\{\forall i \le n, i \ne \ell; Y(\boldsymbol{x}_\ell) \le Y(\boldsymbol{x}_i)\}$, and the last components reflect $\{\forall j \le q, j \ne k; Y(\boldsymbol{x}_{n+k}) \le Y(\boldsymbol{x}_{n+j})\}$ and $\{Y(\boldsymbol{x}_{n+k}) \le Y(\boldsymbol{x}_\ell)\}$. □

D Mean square differentiability of $Y(x)^\alpha 1_{\{B\}}$

Let B be an event, Y be a mean-squared differentiable Gaussian process and $\alpha \in \mathbb{N}$. Then we have:

$$\begin{aligned}&\mathbb{E}\left(\left(\frac{Y(x+h)^\alpha - Y(x)^\alpha}{h}1_{\{B\}} - \frac{\mathrm{d}Y^\alpha}{\mathrm{d}x}(x)1_{\{B\}}\right)^2\right)\\ &\leq \mathbb{E}\left(\left(\frac{Y(x+h)^\alpha - Y(x)^\alpha}{h} - \frac{\mathrm{d}Y^\alpha}{\mathrm{d}x}(x)\right)^2\right) \underset{h\to 0}{\longrightarrow} 0\end{aligned}$$

by mean-squared differentiability of Y^α.

E Symmetry argument

The term $\frac{q(q+1)}{2}$ comes from a symmetry occurring when summing terms with different index but actually equal. At fixed summation index ℓ in (10), we denote ω_{ki} the i^{th} term in the scalar product in (8) for each $\mathcal{M}_{m+k-1,1}$ required for q-EI:

$$\forall i,k = 1,\dots,q,\ \omega_{ki} = \Sigma_{ki}^{(\ell,k)}\left[\nabla\Phi_{p,\Sigma^{(\ell,k)}}(-\boldsymbol{m}^{(\ell,k)})\right]_i.$$

Then the following symmetry between indices i and k occurs:

$$\forall i,k = 1,\dots,q,\ \frac{\omega_{ki}}{\Sigma_{ki}^{(\ell,k)}\varphi_{1,\Sigma_{ii}^{(\ell,k)}}(-m_i^{(\ell,k)})} = \frac{\omega_{ik}}{\Sigma_{ik}^{(\ell,i)}\varphi_{1,\Sigma_{kk}^{(\ell,i)}}(-m_k^{(\ell,i)}))}$$

Indeed, using the formula of the derivative of CDF, (appendix A.1), leads to:

$$\begin{aligned}\frac{\omega_{ki}}{\Sigma_{ki}^{(\ell,k)}\varphi_{\Sigma_{ii}^{(\ell,k)}}(-m_i^{(\ell,k)}))} &= \Phi_{p-1,\Sigma_{|i}^{(\ell,k)}}(-\boldsymbol{m}_{|i}^{(\ell,k)})\\ &= \mathbb{P}\left(\begin{matrix}Y(\boldsymbol{x}_{n+k}) \leq Y(\boldsymbol{x}_\ell),\\ Y(\boldsymbol{x}_{n+j}) \leq Y(\boldsymbol{x}_{n+k}), \forall j = 1\dots q, j\neq k, j\neq i\end{matrix}\middle|\begin{matrix}Y(\boldsymbol{x}_{n+i}) =\\ Y(\boldsymbol{x}_{n+k})\end{matrix}\right),\end{aligned}$$

which is clearly symmetrical between i and k.

References

[1] J. Azimi, A. Fern, and X. Fern, *Batch bayesian optimization via simulation matching*, in Advances in Neural Information Processing Systems, 2010.

[2] A. Azzalini, *mnormt: The multivariate normal and t distributions*, 2012. R package version 1.4-5.

[3] T. Bartz-Beielstein, C. Lasarczyk, and M. Preuss, *Proc. of CEC-05*, IEEE Press, ch. Sequential parameter optimization, p. 773780.

[4] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, *Algorithms for hyper-parameter optimization*, in Advances in Neural Information Processing Systems, 2011.

[5] S.M. Berman, *An extension of plackett's differential equation for the multivariate normal density*, SIAM Journal on Algebraic Discrete Methods, 8 (1987), pp. 196–197.

[6] E. Brochu, V.M. Cora, and N. de Freitas, *A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning.*, tech. report, Dept. of Computer Science, University of British Columbia, 2009.

[7] C. Chevalier and D. Ginsbourger, Learning and Intelligent Optimization - 7th International Conference, Lion 7, Catania, Italy, January 7-11, 2013, Revised Selected Papers*, chapter fast computation of the multipoint expected improvement with applications in batch selection, pages 59-69*, Springer, (2014.).

[8] E Contal, D. Buffoni, A. Robicquet, and N. Vayatis, *Parallel gaussian process optimization with upper confidence bound and pure exploration*, in ECML, 2013.

[9] E. Contal, V. Perchet, and N. Vayatis, *Gaussian process optimization with mutual information*, in Proceedings of the 31st International Conference on Machine Learning, 2014, pp. 253–261.

[10] N. Cressie, A.S. Davis, and J. Leroy Folks, *The moment-generating function and negative integer moments*, The American Statistician, 35 (3) (1981), pp. 148–150.

[11] D. Ginsbourger and V. Picheny and O. Roustant and with contributions by C. Chevalier and S. Marmin and T. Wagner, *DiceOptim: Kriging-Based Optimization for Computer Experiments*, 2015. R package version 1.5.

[12] T. Desautels, A. Krause, and J. Burdick, *Parallelizing exploration-exploitation trade-offs with Gaussian process bandit optimization*, in Proceedings of ICML, 2012.

[13] A. I. J. Forrester, A. Sóbester, and A. J. Keane, *Engineering design via surrogate modelling: a practical guide*, Wiley, 2008.

[14] P. I. Frazier, *Parallel global optimization using an improved multi-points expected improvement criterion*, in INFORMS Optimization Society Conference, Miami FL, 2012.

[15] P. I. Frazier, W. B. Powell, and S. Dayanik, *A knowledge-gradient policy for sequential information collection*, SIAM Journal on Control and Optimization, 47 (2008), pp. 2410–2439.

[16] A. Genz, *Numerical computation of multivariate normal probabilities*, Journal of Computational and Graphical Statistics, 1 (1992), pp. 141–149.

[17] D. Ginsbourger, R. Le Riche, and Carraro L., *Kriging is well-suited to parallelize optimization*, in Computational Intelligence in Expensive Optimization Problems, vol. 2 of Adaptation Learning and Optimization, Springer, 2010, pp. 131–162.

[18] R. Girdziusas, R. Le Riche, F. Viale, and D. Ginsbourger, *Parallel budgeted optimization applied to the design of an air duct*, tech. report, 2012.

[19] J. González, Z. Dai, P. Hennig, and N.D. Lawrence, *Batch bayesian optimization via local penalization.* arXiv:1505.08052.

[20] R. Gramacy and H. Lian, *Gaussian process single-index models as emulators for computer experiments*, Technometrics, 54 (2012), pp. 30–41.

[21] W.V. Harper and S.K. Gupta, *Sensitivity/uncertainty analysis of a borehole scenario comparing latin hypercube sampling and deterministic sensitivity approaches*, Battelle Memorial Institute, Columbus, USA, (1983).

[22] J. Janusevskis, R. Le Riche, D. Ginsbourger, and R. Girdziusas, *Expected improvements for the asynchronous parallel global optimization of expensive functions : Potentials and challenges*, in LION 6 Conference (Learning and Intelligent OptimizatioN), Paris : France, 2012.

[23] D.R. Jones, *A taxonomy of global optimization methods based on response surfaces*, Journal of Global Optimization, 21 (2001), pp. 345–383.

[24] D. R. Jones, M. Schonlau, and J. William, *Efficient global optimization of expensive black-box functions*, Journal of Global Optimization, 13 (1998), pp. 455–492.

[25] Q. Y. Kenny, W. Li, and A. Sudjianto, *Algorithmic construction of optimal symmetric latin hypercube designs*, Journal of statistical planning and inference, 90 (2000), pp. 145–159.

[26] H. J. Kushner, *A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise*, J. Basic Engineering, 86 (1964), pp. 97–106.

[27] D. Lizotte, *Practical Bayesian Optimization*, PhD thesis, University of Alberta, Canada, 2008.

[28] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans, *Automatic gait optimization with gaussian process regression*, in Proceedings of IJCAI, 2007.

[29] S. Marmin, C. Chevalier, and D. Ginsbourger, *Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers*, Springer International Publishing, Cham, 2015, ch. Differentiating the Multipoint Expected Improvement for Optimal Batch Design, pp. 37–48.

[30] J. Mockus, *Bayesian Approach to Global Optimization. Theory and Applications*, Kluwer Academic Publisher, Dordrecht, 1989.

[31] J. Mockus, V. Tiesis, and A. Zilinskas, *The application of Bayesian methods for seeking the extremum.*, in Towards Global Optimization, L. Dixon and Eds G. Szego, eds., vol. 2, Elsevier, 1978, pp. 117–129.

[32] J. Oakley and A. O'Hagan, *Bayesian inference for the uncertainty distribution of computer model outputs*, Biometrika, 89 (2002).

[33] A. O'Hagan, *Curve fitting and optimal design for prediction*, Journal of the Royal Statistical Society. Series B (Methodological), 40 (1978), pp. 1–42.

[34] M. Osborne, *Bayesian Gaussian Processes for Sequential Prediction, Optimization and Quadrature*, PhD thesis, University of Oxford, 2010.

[35] V. Picheny, D. Ginsbourger, Y. Richet, and G. Caplin, *Quantile-based optimization of noisy computer experiments with tunable precision*, Technometrics, 55 (2013), pp. 2–13.

[36] N.V. Queipo, A. Verde, S. Pintos, and R.T. Haftka, *Assessing the value of another cycle in surrogate-based optimization*, in 11th Multidisciplinary Analysis and Optimization Conference, AIAA, 2006.

[37] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[38] P.A. Romero, A. Krause, and F.H. Arnold, *Navigating the protein fitness landscape with gaussian processes*, Proceedings of the National Academy of Sciences, 110 (3) (2013).

[39] O. Roustant, D. Ginsbourger, and Y. Deville, *DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-Based Metamodelling and Optimization*, Journal of Statistical Software, 51 (1) (2012), pp. 1–55.

[40] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science, 4 (1989), pp. 409–435.

[41] M. Schonlau, *Computer Experiments and global optimization*, PhD thesis, University of Waterloo, 1997.

[42] J. Snoek, H. Larochelle, and R.P. Adams, *Practical Bayesian optimization of machine learning algorithms*, in Advances in Neural Information Processing Systems, 2012.

[43] N. Srinivas, A. Krause, S.M. Kakade, and M. Seeger, *Gaussian process optimization in the bandit setting: No regret and experimental design*, in International Conference on Machine Learning, 2010, pp. 1015–1022.

[44] G.M. Tallis, *The moment generating function of the truncated multi-normal distribution*, J. Roy. Statist. Soc. Ser. B, 23 (1961), pp. 223–229.

[45] M. Tesch, J. Schneider, and H. Choset, *Using response surfaces and expected improvement to optimize snake robot gait parameters*, in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011.

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

[47] J. Wang, S.C. Clark, E. Liu, and P.I. Frazier, *Parallel bayesian global optimization of expensive functions*. Working paper (http://people.orie.cornell.edu/pfrazier/publications.html).

[48] BA Worley, *Deterministic uncertainty analysis*, Trans. Am. Nucl. Soc.;(United States), 55 (1987).

[49] A. Zhilinskas, *Single-step bayesian search method for an extremum of functions of a single variable*, Cybernetics and Systems Analysis, 11 (1975), pp. 160–166.