Bayesian Optimization of Composite Functions: Supplementary Material

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

1. Unbiased Estimator of the Gradient of EI-CF

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

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

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

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

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

where the expectation is with respect to Z and

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

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

We end this section by making a few remarks.

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

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

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

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

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

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

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

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

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

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

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

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

3. Probability of Improvement for Composite Functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4. Description of Langermann and Rosenbrock Test Problems

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

4.1. Langermann Function

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

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

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

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

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

4.2. Rosenbrock Function

The Rosenbrock function (Surjanovic & Bingham, b) is

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

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

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

5. Asymptotic Consistency of the Expected Improvement for Composite Functions

5.1. Basic Definitions and Assumptions

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

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

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

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

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

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

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

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

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

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

5.2. Preliminary Results

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

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

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

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

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

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

almost surely as $k \to \infty$.

Now observe that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Thus,

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

since h is continuous. □

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

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

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

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

but

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

and thus the desired conclusion follows. □

5.3. Proof of the Main Result

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

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

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

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

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

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

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

Finally, using Fatou's lemma we obtain

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

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

6. Auxiliary Results

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

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

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

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

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

but we claim that, indeed,

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

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

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

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

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

□

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

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

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

Proof. By the law of total covariance, we have

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

which implies that

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

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

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

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

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

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

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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