2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

No-Regret Bayesian Optimization with Gradients using Local Optimality-based Constraints: Application to Closed-loop Policy Search

Georgios Makrygiorgos[1,*], Joel A. Paulson[2,*], Ali Mesbah[1]

***Abstract*— Bayesian optimization (BO) has emerged as a data-efficient method for global optimization of expensive black-box functions, which commonly arise in learning-based control applications. Recent work has shown that BO can be augmented with gradient measurements to further improve its convergence behavior. These approaches mostly rely on standard acquisition functions and indirectly incorporate gradient information into a probabilistic surrogate model of the performance function to improve its local predictions. This paper presents a new strategy to simultaneously exploit performance (zeroth-order) and gradient (first-order) data within a single constrained acquisition optimization. This is done by enforcing a set of black-box constraints that mimic the necessary optimality conditions for the original global optimization problem. We establish how the incorporation of these constraints restricts the allowable search space of BO, leading to less exploration than zeroth-order BO. The performance of the proposed method is demonstrated for closed-loop policy search via reinforcement learning on a benchmark LQR problem.**

I. Introduction

The control of complex systems is often associated with the challenge of optimizing black-box functions that are expensive to evaluate and lack an analytical, closed-form structure. These functions may also be subject to noise, further complicating their optimization. Thus, in many real-world applications, we resort to derivative-free global optimization techniques that can effectively handle these challenges. In recent years, there has been a growing interest in the use of black-box optimization methods for various control applications. Specifically, Bayesian optimization (BO) [1] has emerged as an effective strategy for controller auto-tuning [2], [3], [4] and direct policy-search reinforcement learning (RL) [5], [6]. The main idea of BO is to convert a challenging black-box optimization problem into a sequence of easier-to-solve sub-problems that aim at iteratively learning and updating our belief about the objective by querying the system performance. This is achieved by constructing a Gaussian process (GP) model of the objective given the current set of observations and subsequently optimizing over a utility metric, a so-called acquisition function (AF), to determine where to query the system next. AFs use the surrogate model of objective to suggest new evaluation points, balancing the competing aims of exploration and exploitation.

Although BO is by nature a zeroth-order optimization method, recent work has demonstrated that gradient information, when accessible in practice, can be valuable since it provides additional information about the objective function [7], [8], [9], leading to so-called gradient-enhanced BO. Generally, the key idea of these methods is to condition the predictions of the function on gradient observations to reduce the variance in unexplored points in the domain, yielding a more accurate surrogate of the objective and, thus, accelerating the overall convergence of BO. The gradient-enhanced GP can be utilized with typical zeroth-order AFs [7], or with first-order AFs [5], [10].

In contrast to previous work, here we introduce a gradient-enhanced BO method that directly incorporates gradient information into the AF, as opposed to indirectly through the design of a more complicated derivative GP model. The proposed necessary-optimality BO, or NOBO, method uses GP surrogates for the partial derivatives of the objective to approximately enforce the first-order optimality conditions as black-box constraints in the AF. These constraints allow for defining a feasible set that explicitly takes into account the uncertainty present in approximating the partial gradients from data, which is updated by observing new data. Thus, the feasible set enables narrowing down the search of the design space to regions that are jointly informative with respect to both zeroth- and first-order information. Unlike our previous work [11] that relied on an ensemble of AFs with first-order information, the performance of NOBO only depends on scalar exploration hyperparameters that are easier to select. We analyze the theoretical performance of NOBO based on the cumulative regret metric, connecting it to the kernel properties of the GP. The performance of NOBO is demonstrated for policy-based RL on a benchmark LQR problem, and is compared to that of standard BO and REINFORCE.

II. Problem Statement

A. Optimization goal and regularity assumptions

We consider the following black-box optimization problem

$$\max_{x \in X} f(x), \tag{1}$$

where $x \in X$ are decision variables that are restricted to some known compact domain $X \subset \mathbb{R}^d$ and $f : X \to \mathbb{R}$ is an expensive-to-evaluate objective function whose mathematical structure is unknown. We assume that X can be expressed as the level set of a known function $c : \mathbb{R}^d \to \mathbb{R}^c$, i.e.,

$$X = \{x \in \mathbb{R}^d : c(x) \leq 0\}. \tag{2}$$

We consider the bandit feedback setting wherein, at iteration t, a query point x_t is selected for which noisy evaluations

[1]G. Makrygiorgos and A. Mesbah are with the Dept. of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720.
[2]Joel A. Paulson is with the Dept. of Chemical and Biomolecular Engineering, The Ohio State University, Columbus, OH 43210.
∗G. Makrygiorgos and J.A. Paulson have equally contributed to this work.
This work was partially supported by NSF Grants 2130734 and 2237616.

of $f(x_t)$ and its gradient $\nabla f(x_t)$ can be observed. That is, we observe

$$y_{0,t} = f(x_t) + \epsilon_{0,t}, \tag{3a}$$
$$y_{i,t} = \partial_{x_i} f(x_t) + \epsilon_{i,t}, \quad \forall i \in \mathbb{N}_1^d, \tag{3b}$$

where $\epsilon_{i,t}$ are i.i.d. R-sub-Gaussian noise terms for a fixed constant $R \geq 0$, meaning they must satisfy

$$\mathbb{E}\left\{e^{\lambda \epsilon_{i,t}} \mid \Sigma_{i,t-1}\right\} \leq e^{\frac{\lambda^2 R^2}{2}}, \ \forall i \in \mathbb{N}_0^d, t \geq 0, \lambda \in \mathbb{R}, \tag{4}$$

where $\Sigma_{i,t-1}$ denotes the σ-algebra generated by the random variables $\{x_k, \epsilon_{i,k}\}_{k=1}^{t-1}$ and x_t. This is a standard assumption in the bandit feedback setting, and is relatively mild since it holds for all distributions bounded in $[-R, R]$ [12]. We note that this differs from traditional bandit feedback problems that only assume the availability of zeroth-order information $y_{0,t}$, which can place a limitation on performance. Here, we look to incorporate gradient information, which can often be observed (or estimated) in control applications such as closed-loop policy optimization (Section V). We further assume that f is "regular" in the following sense.

Assumption 1: Let $H_k(X)$ denote the reproducing kernel Hilbert space (RKHS) of functions $X \to \mathbb{R}$, with a positive semi-definite kernel function $k : X \times X \to \mathbb{R}$. Furthermore, let $\langle \cdot, \cdot \rangle_k$ denote the inner product that obeys the reproducing property $f(x) = \langle f, k(x, \cdot) \rangle_k$ for all $f \in H_k(X)$, which induces the RKHS norm $\|f\|_k = \sqrt{\langle f, f \rangle_k}$. We assume that $\|f\|_{k_0} \leq B_0$ and $\|\partial_{x_i} f\|_{k_i} \leq B_i$ for all $i \in \mathbb{N}_1^d$ have known finite bounds $B_0, \ldots, B_d$ for some known kernels $k_0, \ldots, k_d$.

Assumption 1 allows for the construction of well-behaved confidence bounds on the target functions and is valid as long as $(f, \nabla f)$ satisfy basic properties such as being bounded, continuous, and f being at least once differentiable over X.

B. Gaussian process models

We consider a GP prior $\mathcal{GP}(0, k_i(\cdot, \cdot))$ over the target function f and its partial derivatives $\partial_{x_i} f$ to learn the unknown black-box functions, where k_i is the kernel function associated with the RKHS $H_{k_i}(X)$ (Assumption 1). Additionally, we adopt an i.i.d. Gaussian zero-mean noise model with variance $\eta_i > 0$. The GP model of f enables us to construct analytic expressions for the posterior mean and covariance functions, as well as the maximum information gain in the bandit feedback problem at hand, which will be useful for the ensuing theoretical analysis.

Given t observations $\boldsymbol{y}_{i,t} = (y_{i,1}, \ldots, y_{i,t})$ under the GP prior, the posterior remains a GP with the following mean $\mu_{i,t}$, kernel $k_{i,t}$, and variance $\sigma_{i,t}^2$ functions for all $i \in \mathbb{N}_0^d$

$$\mu_{i,t}(x) = \boldsymbol{k}_{i,t}^\top(x)(\boldsymbol{K}_{i,t} + \eta_i I)^{-1} \boldsymbol{y}_{i,t}, \tag{5a}$$
$$k_{i,t}(x, x') = k_i(x, x') - \boldsymbol{k}_{i,t}^\top(x)(\boldsymbol{K}_{i,t} + \eta_i I)^{-1} \boldsymbol{k}_{i,t}(x'),$$
$$\sigma_{i,t}^2(x) = k_{i,t}(x, x), \tag{5b}$$

where $\boldsymbol{k}_{i,t}(x) = [k_i(x_1, x), \ldots, k_i(x_t, x)]^\top$ and $\boldsymbol{K}_{i,t}$ is the positive definite kernel matrix whose elements are given by $[\boldsymbol{K}_{i,t}]_{n,m} = k_i(x_n, x_m)$ for all $n, m \in \mathbb{N}_1^t$. Note that, in principle, one could replace this set of $d+1$ independent GP models with a joint GP model that captures the correlation between f and ∇f (see, e.g., [10]). Here, we consider the case of independent GPs because (i) it simplifies analysis and model complexity, (ii) established results carry over to the joint GP case due to Slepian's comparison lemma [13], and (iii) it provides more flexibility in the kernel choice.

Next, we define the *maximum information gain* (MIG) for the unknown functions f and ∇f.

Definition 1: Let $A \subset X$ denote any potential subset of points sampled from X. The maximum information gain for the $(i+1)^{\text{th}}$ element of $(f, \nabla f)$ for t noisy measurements is

$$\gamma_{i,t} = \max_{A \subset X : |A| = t} \frac{1}{2} \log\det\left(I + \eta_i^{-1} \boldsymbol{K}_{i,A}\right), \tag{6}$$

where $\boldsymbol{K}_{i,A} = [k_i(x, x')]_{x,x' \in A}$.

Note $\gamma_{i,t}$ depends on both the domain X and the kernel function k_i, and can be interpreted as a measure for the difficulty of the optimization task. Several results exist for bounding the growth of $\gamma_{i,t}$ as a function of the number of iterations t, as used in the theoretical analysis of Section IV.

We now summarize a key result that shows how the posterior GP mean is centered around the unknown functions by a multiplicative factor of the posterior standard deviation.

Lemma 1 (Theorem 2, [12]): Let $X \subset \mathbb{R}^d$, $\{\epsilon_{i,t}\}_{t=1}^{\infty}$ be R-sub-Gaussian noise, and Assumption 1 holds. Then, for any $\delta \in (0, 1)$, the following holds for all $x \in X$ and $t \geq 1$

$$|\mu_{i,t-1}(x) - g_i(x)| \tag{7}$$
$$\leq \left(B_i + R\sqrt{2(\gamma_{i,t-1} + 1 + \ln((d+1)/\delta))}\right) \sigma_{i,t-1}(x),$$

with probability at least $1 - \delta/(d+1)$, where g_i denotes the $(i+1)^{\text{th}}$ element of $(f, \nabla f)$ and $\mu_{i,t-1}(x)$, $\sigma_{i,t-1}(x)$, and $\gamma_{i,t-1}$ are given in (5) and (6).

Note that the value of δ in [12, Theorem 2] is replaced by $\delta/(d+1)$ above since we will require joint confidence bounds on $(f, \nabla f)$, as in [14].

C. Performance metrics

We now define the key performance metrics that will be used to analyze the effectiveness of the proposed approach. As in the standard bandit feedback setting, we look to minimize the gap of $f(x_t)$ to the optimal value $f^\star = \max_{x \in X} f(x)$, i.e., the *instantaneous regret*

$$r_t = f^\star - f(x_t), \tag{8}$$

where x_t is the selected query point at iteration $t \geq 1$. Given that gradient information is available, we can also quantify

$$v_t = \|\nabla f(x_t) - \nabla c(x_t) \lambda_t\|_1, \tag{9}$$

where $\lambda_t \in \mathbb{R}_+^c$ will be Lagrange multipliers selected by our approach at iteration t. As shown in Section III, v_t is the distance from a first-order stationarity condition being satisfied. Ideally, our approach would be able to achieve zero regret and violation in a single step; however, this is only possible when $(f, \nabla f)$ are perfectly known. In the black-box setting, we aim to minimize the cumulative regret

$$R_T = \textstyle\sum_{t=1}^T r_t = \sum_{t=1}^T \left(f^\star - f(x_t)\right), \tag{10}$$

over T iterations. Formally, minimizing R_T requires one to solve an intractable dynamic programming problem (see, e.g., [15], [16]). Thus, this paper presents an efficient, simple-to-implement *no-regret* approach that ensures $R_T/T \to 0$ as $T \to \infty$. The no-regret property not only guarantees vanishing per-round instantaneous regret, but also ensures convergence to the global solution. Similarly, we can also define the cumulative violation of stationarity as

$$V_T = \sum_{t=1}^{T} v_t = \sum_{t=1}^{T} \left(\|\nabla f(x_t) - \nabla c(x_t)\lambda_t\|_1 \right). \quad (11)$$

III. Necessary Optimality-constrained Bayesian Optimization (NOBO)

In this section, we present the proposed necessary optimality-constrained Bayesian optimization (NOBO) approach for solving (1). The key observation that motivates NOBO is that we can reformulate (1) as

$$\max_{x \in X} f(x) \quad \text{s.t.} \quad \nabla f(x) \in N_X(x), \quad (12)$$

where $N_X(x) = \{z \in \mathbb{R}^d : z^\top (y - x) \leq 0, \forall y \in X\}$ denotes the normal cone to the set X at the point x. The newly added constraint $\nabla f(x) \in N_X(x)$ implies x is a "stationary point," which constitutes the first-order necessary optimality conditions for x to be a (local) maximum as long as the set X ensures constraint qualifications are satisfied. At the first glance, (12) may not appear useful since the necessary optimality conditions are typically solved numerically to identify possible solutions to (1). This would make $\nabla f(x) \in N_X(x)$ redundant; however, this is only true when the function f is exactly known. In the black-box setting of this work, these constraints provide additional independent information that can be exploited to restrict the set of possible query points.

Assuming the linear independence constraint qualification (LICQ) holds, we can equivalently represent the feasible set of (12) using the Karush-Kuhn-Tucker (KKT) conditions

$$\mathcal{F} = \{x \mid \exists \lambda : \nabla f(x) = \nabla c(x)\lambda,\ 0 \leq \lambda \perp c(x) \leq 0\},$$

where the notation "$0 \leq \lambda \perp c(x) \leq 0$" is shorthand for the complementary constraints, i.e., $c(x) \leq 0$, $\lambda \geq 0$, $\lambda^\top c(x) = 0$. Since neither the target function f nor the feasible set $\mathcal{F}$ are known in the black-box setting, we rely on constructing high probability relaxations using GP models. To this end, we introduce the lower and upper confidence bound functions.

Definition 2: The lower confidence bound (LCB) and upper confidence bound (UCB) for the $(i+1)^{\text{th}}$ element of $(f, \nabla f)$ at iteration t are given by

$$l_{i,t}(x) = \mu_{i,t-1}(x) - \beta_{i,t}^{1/2} \sigma_{i,t-1}(x), \quad (13a)$$

$$u_{i,t}(x) = \mu_{i,t-1}(x) + \beta_{i,t}^{1/2} \sigma_{i,t-1}(x), \quad (13b)$$

where $\beta_{i,t}^{1/2} = B_i + R\sqrt{2(\gamma_{i,t-1} + 1 + \ln((d+1)/\delta))}$.

Using Lemma 1, we can then establish the following result on the joint relaxation of (12).

Theorem 1: Let the assumptions of Lemma 1 hold. Then, with probability at least $1 - \delta$, the following bounds hold simultaneously for all $x \in X$ and $t \geq 1$

$$f(x) \in [l_{0,t}(x), u_{0,t}(x)] \text{ and } \mathcal{F} \subseteq \mathcal{F}_t^u, \quad (14)$$

Algorithm 1 The relaxation-based **N**ecessary **O**ptimality-constrained **B**ayesian **O**ptimization (NOBO) algorithm.

Input: The compact domain X; GP priors $(\mu_i, k_i)_{i=0}^d$, parameters $\{\beta_{i,t}\}_{i \in \mathbb{N}_0^d, t \geq 1}$; and total number of iterations T.

1: **for** $t = 1$ to T **do**
2: Solve $(x_t, \lambda_t) \in \operatorname{argmax}_{x,\lambda} u_{0,t}(x)$ s.t. $(x, \lambda) \in \mathcal{R}_t^u$.
3: Get noisy observations of f and ∇f at x_t.
4: Update GP posteriors (5) with new observations.
5: **end for**

where $\mathcal{F}_t^u = \{x \mid \exists \lambda : (x, \lambda) \in \mathcal{R}_t^u\}$ is the relaxed feasible region defined in terms of the set

$$\mathcal{R}_t^u = \left\{ \begin{bmatrix} x \\ \lambda \end{bmatrix} \;\middle|\; \begin{array}{c} |\boldsymbol{\mu}_{d,t-1}(x) - \nabla c(x)\lambda| \leq \boldsymbol{\beta}_{d,t}^{1/2} \boldsymbol{\sigma}_{d,t-1}(x) \\ 0 \leq \lambda \perp c(x) \leq 0 \end{array} \right\},$$

with the following definitions

$$\begin{aligned} \boldsymbol{\mu}_{d,t-1}(x) &= (\mu_{1,t-1}(x), \ldots, \mu_{d,t-1}(x)) && \in \mathbb{R}^{d \times 1}, \\ \boldsymbol{\sigma}_{d,t-1}(x) &= (\sigma_{1,t-1}(x), \ldots, \sigma_{d,t-1}(x)) && \in \mathbb{R}^{d \times 1}, \\ \boldsymbol{\beta}_{d,t}^{1/2} &= \operatorname{diag}(\beta_{1,t}^{1/2}, \ldots, \beta_{d,t}^{1/2}) && \in \mathbb{R}^{d \times d}. \end{aligned}$$

Proof: The confidence bounds $l_{i,t}(x)$ and $u_{i,t}(x)$ are random variables since they depend on observations $y_{i,t}$ that are corrupted by random noise. Therefore, we can define the following events that the unknown functions respect the confidence bounds for all $x \in X$ and $t \geq 1$

$$\begin{aligned} \mathcal{E}_0 &= \cap_{x \in X} \cap_{t \geq 1} \{l_{0,t}(x) \leq f(x) \leq u_{0,t}(x)\}, \\ \mathcal{E}_i &= \cap_{x \in X} \cap_{t \geq 1} \{l_{i,t}(x) \leq \partial_{x_i} f(x) \leq u_{i,t}(x)\},\ \forall i \in \mathbb{N}_1^d. \end{aligned}$$

We can then establish the following sequence of inequalities

$$\begin{aligned} \mathbb{P}\left\{\cap_{i=0}^d \mathcal{E}_i\right\} &= 1 - \mathbb{P}\left\{\cup_{i=0}^d \overline{\mathcal{E}_i}\right\} \geq 1 - \textstyle\sum_{i=0}^d \mathbb{P}\left\{\overline{\mathcal{E}_i}\right\} \\ &\geq 1 - \textstyle\sum_{i=0}^d \frac{\delta}{d+1} = 1 - \delta, \end{aligned}$$

where the second inequality follows from Boole's inequality and the third inequality follows from Lemma 1. The first part of (14) directly follows. To see that $\mathcal{F} \subset \mathcal{F}_t^u$ must also hold, the stationarity condition $\nabla f(x) = \nabla c(x)\lambda$ can be represented by two inequalities $\nabla f(x) - \nabla c(x)\lambda \leq 0$ and $\nabla f(x) - \nabla c(x)\lambda \geq 0$, which can be relaxed by replacing the elements of ∇f by their lower and upper confidence bounds, respectively. After a few algebraic manipulations, one can derive $\mathcal{F}_t^u$ as an equivalent representation. ■

The proposed NOBO method is summarized in Algorithm 1, which is conceptually straightforward in that only a single auxiliary problem is solved at each iteration. This auxiliary problem in line 2 is an instance of a mathematical program with complementarity constraints (MPCC) [17]. Given the focus on expensive-to-evaluate functions f, we assume that the cost of solving the MPCC is small relative to the cost of a function query. Also, a direct consequence of Theorem 1 is that the set of global solutions $x^\star$ must be contained within $\mathcal{F}_t^u$ with probability at least $1 - \delta$, so that the "size" of $\mathcal{F}_t^u$ provides a measure for progress of NOBO as t increases.

Remark 1: When $x^\star$ is known to lie strictly in the interior of X, we can set $\lambda = 0$, which simplifies the auxiliary

problem in Algorithm 1. This is equivalent to simplifying the necessary optimality conditions to $\nabla f(x) = 0$.

IV. Theoretical Analysis of NOBO

In this section, we analyze the theoretical performance of NOBO. Our goal is to establish bounds on the cumulative regret R_T and the stationarity violation V_T that depend on the MIG of the unknown functions and the number of iterations T, similar to [14]. We can then use established bounds on the MIG in [18, Theorem 5] to bound the MIG growth over T, which will allow us to establish convergence of NOBO. First, a lemma is introduced to bound r_t and v_t.

Lemma 2: If the inequalities (14) hold, then the auxiliary problem in line 4 of Algorithm 1 will always be feasible and the instantaneous regret and stationary violation will satisfy

$$r_t \leq 2\beta_{0,t}^{1/2}\sigma_{0,t-1}(x_t), \tag{15a}$$

$$v_t \leq \textstyle\sum_{i=1}^{d} 2\beta_{i,t}^{1/2}\sigma_{i,t-1}(x_t), \tag{15b}$$

for all $x \in X$ and $t \geq 1$.

Proof: From (8) and x_t in line 4 of Algorithm 1, we have

$$\begin{aligned} r_t &\leq u_{0,t}(x^\star) - l_{0,t}(x_t) \\ &\leq u_{0,t}(x_t) - l_{0,t}(x_t) = 2\beta_{0,t}^{1/2}\sigma_{0,t-1}(x_t), \end{aligned}$$

where the first inequality follows from the assumed upper and lower bounds on the target function and the second inequality follows from the fact that x_t maximizes $u_{0,t}(x)$ over a set $\mathcal{F}_t^u \subseteq X$ that contains $x^\star$ under the assumed bounds on the gradient ∇f. Feasibility of $x^\star \in \mathcal{F}_t^u$ directly implies feasibility of the auxiliary problem. We now consider the violation of the stationarity condition. Let $q_t = \nabla c(x_t)\lambda_t \in \mathbb{R}^d$. We can rewrite (9) as

$$v_t = \textstyle\sum_{i=1}^{d} |\partial_{x_i} f(x_t) - [q_t]_i|.$$

We look to bound each element of this sum

$$\begin{aligned} &|\partial_{x_i} f(x_t) - [q_t]_i| \\ &\leq |\partial_{x_i} f(x_t) - \mu_{i,t-1}(x_t)| + |\mu_{i,t-1}(x_t) - [q_t]_i| \\ &\leq \beta_{i,t}^{1/2}\sigma_{i,t-1}(x_t) + \beta_{i,t}^{1/2}\sigma_{i,t-1}(x_t) = 2\beta_{i,t}^{1/2}\sigma_{i,t-1}(x_t), \end{aligned}$$

where the first inequality follows from $|a+b| \leq |a| + |b|$ and the second inequality follows from the assumed bounds in (14) and the fact that $(x_t, \lambda_t) \in \mathcal{R}_u$. ■

We can now combine these results, along with results from [14], to establish the main theorem on the cumulative regret and stationarity violation for NOBO.

Theorem 2: Under the assumptions of Lemma 1, we have, with probability at least $1-\delta$, that the sample points $\{x_t\}_{t\geq 1}$ generated by NOBO (Algorithm 1) satisfy

$$R_T \leq 4\beta_{0,T}^{1/2}\sqrt{(T+2)\gamma_{0,T}}, \tag{16a}$$

$$V_T \leq \textstyle\sum_{i=1}^{d} 4\beta_{i,T}^{1/2}\sqrt{(T+2)\gamma_{i,T}}. \tag{16b}$$

Proof: Combining Lemma 1 and Theorem 1, the following event must hold with probability $\geq 1-\delta$

$$\{r_t \leq 2\beta_{0,t}^{1/2}\sigma_{0,t-1}(x_t)\} \cup \{v_t \leq \textstyle\sum_{i=1}^{d} 2\beta_{i,t}^{1/2}\sigma_{i,t-1}(x_t)\}.$$

From the definition of cumulative regret, we can establish the following inequalities that must hold with probability $\geq 1-\delta$

$$\begin{aligned} R_T = \textstyle\sum_{t=1}^{T} r_t &\leq 2\beta_{0,T}^{1/2} \textstyle\sum_{t=1}^{T} \sigma_{0,t-1}(x_t) \\ &\leq 4\beta_{0,T}^{1/2}\sqrt{(T+2)\gamma_{0,T}}, \end{aligned}$$

where the first inequality follows from the monotonicity of $\{\beta_{0,t}\}_{t\geq 1}$ and the second inequality follows from [14, Lemma 4], which shows that $\sum_{t=1}^{T}\sigma_{0,t-1}(x_t) \leq \sqrt{4(T+2)\gamma_{0,T}}$. The stated result follows by applying the same analysis to the cumulative stationarity violation V_T. ■

The following corollary to Theorem 2 is immediately established for the convergence rate of NOBO to $f^\star$.

Corollary 1: Under the assumptions of Theorem 1, we have, with probability at least $1-\delta$, that there exists some $\tilde{x}_T \in \{x_1, x_2, \ldots, x_T\}$ such that

$$f^\star - f(\tilde{x}_T) \leq \frac{4\beta_{0,T}^{1/2}\sqrt{(T+2)\gamma_{0,T}}}{T} = \mathcal{O}\left(\frac{\gamma_{0,T}}{\sqrt{T}}\right). \tag{17}$$

Proof: Let $S_T = \min_{t\in\{1,\ldots,T\}} r_t$ be the minimum of the regret sequence. Since r_t is non-negative and the minimum of a sequence must be less than or equal to the average, we have $0 \leq S_T \leq R_T/T$. The claim follows from Theorem 2 and letting $\tilde{x}_T$ be the point that minimizes S_T. ■

We note that the point $\tilde{x}_T$ cannot be identified by minimizing the regret sequence unless the noise variance is zero, i.e., $\eta_0 = 0$. In the noisy case, we can resort to the following recommendation procedure

$$\tilde{x}_T = \underset{x_t\in\{x_1,\ldots,x_T\}}{\operatorname{argmax}} \; l_{0,t}(x_t), \tag{18}$$

which can be interpreted as a pessimistic estimate of the maximum value of f due to the noise in the observations. This will not affect the result shown in Corollary 1 since Theorem 2 holds for the pessimistic estimate of the regret $r_t \leq \bar{r}_t = f^\star - l_{0,t}(x_t)$. It is also interesting to note that this result implies that NOBO has at least the same worst-case convergence rate as the traditional zeroth-order GP-UCB algorithm. However, since we are optimizing over a restricted set $\mathcal{F}_t^u \subset X$, NOBO is expected to provide a faster convergence rate in practice.

Theorem 2 and Corollary 1 are given in terms of the MIG for general kernels. They imply convergence of $\tilde{x}_T \to x^\star$ as $T \to \infty$ as long as $\gamma_{0,T} = o(\sqrt{T})$, which can be guaranteed for the common types of kernels. This is summarized below.

Lemma 3 (Theorem 5, [18]): Let X be compact and convex, $d \in \mathbb{N}$, and assume $k(x, x') \leq 1$. Then,

- Linear: $\gamma_T = \mathcal{O}(d \log T)$;
- Squared exponential: $\gamma_T = \mathcal{O}((\log T)^{d+1})$;
- Matern ($\nu > 1$): $\gamma_T = \mathcal{O}(T^{d(d+1)/(2\nu+d(d+1)}(\log T))$.

Substituting the results of Lemma 3 into (17), it is evident that NOBO converges for the linear, squared exponential, and Matern kernel with smoothness parameter $\nu > 1$. These results can also be used to derive kernel specific bounds as a function of T, e.g., $f^\star - f(\tilde{x}_T) = \mathcal{O}((\log T)^{d+1}/\sqrt{T})$ for the squared exponential kernel.

V. Numerical Illustration of NOBO for Closed-loop Policy Search

The performance of NOBO is demonstrated in the context of policy-based RL, where an "agent" must learn how to take actions in an "environment" by maximizing some reward. This problem can be cast as a stochastic optimal control problem [19]

$$\begin{aligned} \max_{\pi_{0:N-1}} \quad & \mathbb{E}_{w_{0:N-1}} \left\{ \sum_{k=0}^{N-1} r_k(z_k, u_k, w_k) + r_N(z_N) \right\}, \quad (19) \\ \text{s.t.} \quad & z_{k+1} = g_k(z_k, u_k, w_k), \quad u_k = \pi_k(\tau_k), \end{aligned}$$

where z_k, u_k, and w_k are the system state, control input, and disturbance at time step k, respectively; g_k is the (unknown) state transition function; r_k is the reward gained at k; π_k is the feedback control policy at k that can be any feasible function of the observed data $\tau_k = (u_0, \ldots, u_{k-1}, z_0, \ldots, z_k)$; N is the time horizon; and $\mathbb{E}\{\cdot\}$ is the expectation operator with respect to disturbance realizations over the horizon 0 to $N-1$. Since optimizing over a general control policy is intractable, we look to learn the optimal parametrized policy function $p(\tau_k; x)$, where x refers to adjustable policy parameters. This way, the overall reward function in (19) becomes a function of x only [11]. Accordingly, the cost function of (19) can be rewritten as (1)

$$f(x) = \mathbb{E}_{p(\tau;x)}\{R(\tau)\} = \int R(\tau) p(\tau; x) d\tau, \quad (20)$$

where $R(\tau)$ is the overall reward function computed over a single dynamic trajectory τ, which is a random variable with parametrized probability distribution $p(\tau; x)$. Hence, the policy gradient theorem [19] can be used to obtain noisy estimates of gradient of the reward as follows

$$\nabla f(x) = \mathbb{E}_{p(\tau;x)}\{R(\tau) \nabla_x \log p(\tau; x)\}. \quad (21)$$

This gradient information can be readily used in NOBO to search for optimal closed-loop control policies.

Here, we consider the problem of policy search in the case of a linear-quadratic regulator (LQR) with a reward function $J_k(z_k, u_k, w_k) = -z_k^\top Q_z z_k - u_t^\top R_u u_k$, linear dynamics $z_{k+1} = Az_k + Bu_k + w_k$ with $w_k \sim \mathcal{N}(0, 10^{-4} I)$, no terminal reward, and horizon $N = 10$; system matrices are given in [11]. We consider a Gaussian policy $p(z_k; x) = \mathcal{N}(-x^\top z_k, \sigma^2)$, which is typical in continuous-action spaces, where $x \in X = [0, 2]^4 \subset \mathbb{R}^4$ are the four linear policy parameters and variance $\sigma^2 = 10^{-4}$. `UQLab` [20] is used to construct the the GP surrogates with a zero prior mean and a squared-exponential kernel. To get an accurate estimate of the cost, as well as to reduce the variance of the gradient estimates, a "mini-batch" size of $N_s = 2^8$ samples is used during each episode. In addition, we use a baseline value in the estimate of the gradient, as discussed in [21], [22].

First, we examine the performance of NOBO in comparison with standard BO and REINFORCE [21], which is a commonly used RL strategy. Standard BO leverages zeroth-order (function) information at each step t, whereas REINFORCE relies only on first-order (gradient) information. Here, BO, similarly to Algorithm 1, utilizes the UCB AF.

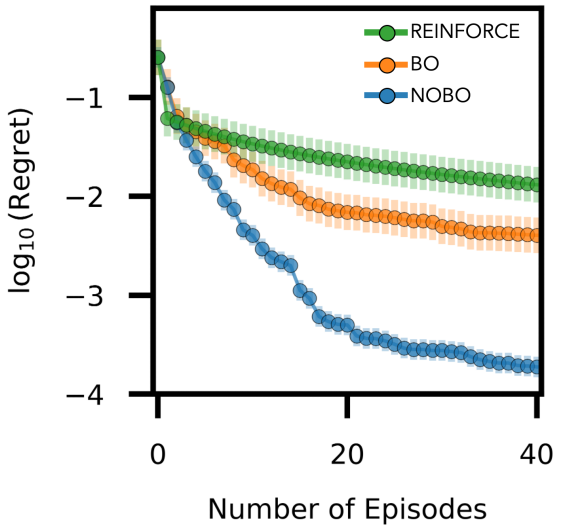

Fig. 1. Simple regret (current best) for REINFORCE (green), standard BO (orange), and NOBO (blue) over 40 closed-loop episodes. Solid lines represent the average regret over 100 trials with different initial datasets. The shaded regions show the one standard deviation about the average regret.

Moreover, since $\lambda_t = 0$ given that the solution lies in the interior of the domain, we resort to an exhaustive grid search (10^4 points) for locating the optimum of line 2 in Algorithm 1, as well as for BO. The initial training dataset is composed of $d+1$ points (here, $d = 4$) chosen uniformly at random from the design space X. We quantify the average closed-loop performance by repeating each algorithm 100 times for different initial datasets. For both standard BO and NOBO, we use $\beta_{0,t} = 0.1 + 0.01 d \ln(1 + 0.01t)$ for the objective, while we set $\beta_{i,t} = 2 + 0.05 d \ln(1 + 0.05t)$ for the constraints in NOBO. These choices were found to work reasonably well in our previous work [2]; however, we observed that our results here were not particularly sensitive to these choices.

Fig. 1 shows the estimated average simple regret $S_T = \min_{t \in \{1,\ldots,T\}} r_t$ versus the number of episodes T for BO, REINFORCE, and NOBO up to $T = 40$. REINFORCE initially yields a regret that is lower than the best solution, but it quickly stalls at a relatively large value after only a few episodes. This can be attributed to the fact that REINFORCE is only utilizing noisy gradient information to update the parameters x. Both standard BO and NOBO outperform REINFORCE, and demonstrate continual improvement as T increases. NOBO, however, consistently outperforms both REINFORCE and standard BO over all episodes by up to two orders of magnitude in simple regret. Furthermore, Fig. 1 shows that the rate of convergence with NOBO is faster than standard BO, suggesting that the incorporation of the necessary optimality conditions into the search process leads to an improved query point selection. It also serves as a validation of Theorem 1 since the global solution was not eliminated from the estimated feasible region.

To further investigate the performance of NOBO, we

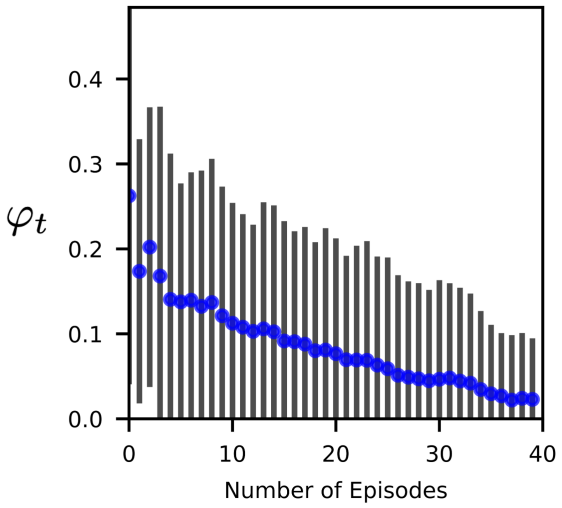

Fig. 2. The percentage of feasible candidate points in the set $\mathcal{F}_t^u$ based on 30,000 uniform random samples drawn from X over the number of closed-loop episodes T. The circles denote the sample mean based on 100 randomly generated initial datasets, with the error bars representing the sample standard deviation (cutoff at zero).

measure the reduction in the volume of the feasible region

$$\varphi_t = \text{Vol}(\mathcal{F}_t^u)/\text{Vol}(X), \tag{22}$$

where $\text{Vol}(A) = \int_A dx$ is the volume of a set. The value φ_t, which quantifies the relative size in the feasible region of standard BO and NOBO, can be straightforwardly estimated at any iteration t by Monte Carlo integration. The evolution of φ_t over 50 closed-loop episodes is shown in Fig. 2 (averaged over 100 random initial datasets). Even in the early episodes, φ_t is about 0.2, implying only 20% of the points in X have the potential to satisfy the necessary optimality conditions (i.e., 80% of the points in X have been confidently eliminated from the search process). Furthermore, as NOBO progresses, φ_t continually decreases, which highlights its ability to learn from the collected gradient information. As such, NOBO systematically excludes points that are inconsistent with the necessary optimality conditions, leading to a substantial reduction in the search space, without compromising performance.

VI. Conclusion

This paper presented a gradient-enhanced Bayesian optimization method, NOBO, that simultaneously leverages zeroth- and first-order information to sequentially maximize an expensive black-box objective function. The primary advantage of NOBO is its ability to conduct a more focused search within the design space, as compared to standard BO, by excluding points that cannot satisfy necessary optimality conditions. We established convergence and upper cumulative and simple regret bounds for NOBO. We also demonstrated NOBO's superior performance over standard BO and the REINFORCE algorithm on a benchmark closed-loop policy optimization problem. Our future work will focus on the incorporation of black-box safety constraints and demonstrations on high-dimensional problems.

References

[1] P. I. Frazier, "A tutorial on Bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[2] J. A. Paulson, G. Makrygiorgos, and A. Mesbah, "Adversarially robust Bayesian optimization for efficient auto-tuning of generic control structures under uncertainty," *AIChE Journal*, vol. 68, no. 6, p. e17591, 2022.

[3] G. Makrygiorgos, A. D. Bonzanini, V. Miller, and A. Mesbah, "Performance-oriented model learning for control via multi-objective Bayesian optimization," *Computers & Chemical Engineering*, vol. 162, p. 107770, 2022.

[4] M. Khosravi, V. N. Behrunani, P. Myszkorowski, R. S. Smith, A. Rupenyan, and J. Lygeros, "Performance-driven cascade controller tuning with Bayesian optimization," *IEEE Transactions on Industrial Electronics*, vol. 69, pp. 1032–1042, 2021.

[5] S. Müller, A. von Rohr, and S. Trimpe, "Local policy search with Bayesian optimization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 708–20 720, 2021.

[6] B. Letham and E. Bakshy, "Bayesian optimization for policy search via online-offline experimentation." *Journal of Machine Learning Research*, vol. 20, pp. 145–1, 2019.

[7] J. Wu, M. Poloczek, A. G. Wilson, and P. Frazier, "Bayesian optimization with gradients," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[8] A. Wu, M. C. Aoi, and J. W. Pillow, "Exploiting gradients and Hessians in Bayesian optimization and Bayesian quadrature," *arXiv preprint arXiv:1704.00060*, 2017.

[9] S. Shekhar and T. Javidi, "Significance of gradient information in Bayesian optimization," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2021, pp. 2836–2844.

[10] S. Penubothula, C. Kamanchi, S. Bhatnagar *et al.*, "Novel first order Bayesian optimization with an application to reinforcement learning," *Applied Intelligence*, vol. 51, no. 3, pp. 1565–1579, 2021.

[11] G. Makrygiorgos, J. A. Paulson, and A. Mesbah, "Gradient-enhanced Bayesian optimization via acquisition ensembles with application to reinforcement learning," *IFAC-PapersOnLine*, pp. 698–703, 2023.

[12] S. R. Chowdhury and A. Gopalan, "On kernelized multi-armed bandits," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 844–853.

[13] D. Slepian, "The one-sided barrier problem for Gaussian noise," *Bell System Technical Journal*, vol. 41, no. 2, pp. 463–501, 1962.

[14] W. Xu, Y. Jiang, B. Svetozarevic, and C. Jones, "Constrained efficient global optimization of expensive black-box functions," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 38 485–38 498.

[15] J. A. Paulson, F. Sorouifar, and A. Chakrabarty, "Efficient multi-step lookahead Bayesian optimization with local search constraints," in *Proceedings of the 61st IEEE Conference on Decision and Control*, 2022, pp. 123–129.

[16] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, 2012, vol. 1.

[17] H. Scheel and S. Scholtes, "Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity," *Mathematics of Operations Research*, vol. 25, no. 1, pp. 1–22, 2000.

[18] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[19] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, pp. 1057–1063, 1999.

[20] S. Marelli and B. Sudret, "UQLab: A framework for uncertainty quantification in Matlab," in *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, 2014, pp. 2554–2563.

[21] B. Recht, "A tour of reinforcement learning: The view from continuous control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 253–279, 2019.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2018.