

Classified Regression for Bayesian Optimization: Robot Learning with Unknown Penalties

Alonso Marco¹
Dominik Baumann¹
Philipp Hennig²
Sebastian Trimpe¹

AMARCO@TUE.MPG.DE
DBAUMANN@TUE.MPG.DE
PH@TUE.MPG.DE
TRIMPE@IS.MPG.DE

1. Intelligent Control Systems

Max Planck Institute for Intelligent Systems
Heisenbergstr. 3
70569 Stuttgart, Germany

2. Computer Science Department, University of Tübingen and

Max Planck Institute for Intelligent Systems
Maria von Linden Str. 6
72076 Tübingen, Germany

Editor:

Abstract

Learning robot controllers by minimizing a black-box objective cost using Bayesian optimization (BO) can be time-consuming and challenging. It is very often the case that some roll-outs result in failure behaviors, causing premature experiment detention. In such cases, the designer is forced to decide on heuristic cost penalties because the acquired data is often scarce, or not comparable with that of the stable policies. To overcome this, we propose a Bayesian model that captures exactly what we know about the cost of unstable controllers prior to data collection: Nothing, except that it should be a somewhat large number. The resulting Bayesian model, approximated with a Gaussian process, predicts high cost values in regions where failures are likely to occur. In this way, the model guides the BO exploration toward regions of stability. We demonstrate the benefits of the proposed model in several illustrative and statistical synthetic benchmarks, and also in experiments on a real robotic platform. In addition, we propose and experimentally validate a new BO method to account for unknown constraints. Such method is an extension of Max-Value Entropy Search, a recent information-theoretic method, to solve unconstrained global optimization problems.

Keywords: Bayesian inference, robot learning, Gaussian process, reinforcement learning, Bayesian optimization

1. Introduction

When learning a policy on a robot, it is very often the case that some roll-outs result in failure behaviors, which forbid the robot from completing the task. In those cases, the system can largely (and often quickly) diverge from the desired behavior, causing the need of a premature experiment detention, for example, by pressing an emergency button. Then, it is unclear what cost should be assigned to that failure case. Intuitively, the designer would heuristically

choose a high penalty so that similar policies are never visited again. For example, (Marco et al., 2016) propose a method to automatically tune controller parameters of a humanoid robot that learns to balance an inverted pole. Therein, unsuccessful balancing experiments were penalized with a large number, chosen heuristically a priori. In (Rai et al., 2017), a walking controller of a bipedal robot is learned from data, using a tailored cost function that penalizes with a high cost roll-outs in which the robot falls down. The authors (Tu and Recht, 2018) learn the parameters of an LQR feedback policy from data in a reinforcement learning (RL) setting, and assign an infinite cost when the policy destabilizes the closed-loop behavior. Finally, in (Kober and Peters, 2009), RL is used for the ball-in-a-cup problem, assigning a constant cost to roll-outs in which the ball does not pass upwards the rim of the cup.

While these methods show successful learning experiments, choosing appropriate penalties for failure cases is a key component of the reward design, and unclear a priori (Chatzilygeroudis et al., 2018). In failure cases (e.g., unstable controllers), deciding on a penalty is non-trivial, because:

- (a) The immediate cost values might be orders of magnitude too large and not comparable against those of the stable policies.
- (b) Due to the premature experiment detention, the collected data can be too scarce, and not comparable with that of the stable policies, whose experiments lasted for longer time (Marco et al., 2016).
- (c) After stopping the experiment, sensor measurements could be difficult, or impossible to access.

To circumvent these problems, the cost designer typically ignores the acquired data at failure cases and commits to *ad hoc* choices, which are decided and fixed before doing any experiment. While this might sound reasonable, *prior to any data collection* and without expert knowledge about the system, the designer does not know how to choose such penalty. Furthermore, arbitrary choices come at the risk of laborious and repeated manual tuning: During learning it is possible that stable roll-outs retrieve a cost worse than the arbitrary penalty chosen a priori, which forces a re-design of the cost function, and thus, possibly a restart of the learning experiments. Generally speaking, choosing an appropriate penalty for failures is problem-dependent and non-trivial. Moreover, failure cases need to be treated separately when designing the cost function, typically requiring manual tuning effort.

In the context of robot learning, Bayesian optimization (BO) (Shahriari et al., 2016) can be seen as a form of RL for policy optimization. Therein, the latent cost function represents an unknown mapping from the parameters of a policy executed on the robot to its performance on completing some task (Berkenkamp et al., 2016; Calandra et al., 2016; Marco et al., 2016; Rai et al., 2017; von Rohr et al., 2018). Such cost is typically modeled in a probabilistic setting and minimized iteratively through experiments. BO incorporates a probabilistic prior about the objective cost typically through a Gaussian process (GP) regression model (Rasmussen and Williams, 2006). This is a non-parametric model that allows for inference on the latent cost at unobserved locations, conditioned on collected data.

The core contribution of this paper is a novel Bayesian model for the cost function, such that the penalty term does not need to be defined a priori, but instead is learned from

data. The likelihood of the Bayesian model that can tackle a *hybrid* dataset, inside which stable policies yield a continuous-valued cost observation, and unstable policies only yield the discrete label “unstable”. The consequent intractable posterior is approximated to obtain a tractable GP model. In addition, the GP model is used within BO to reveal the optimal policy by sequentially selecting experiments that avoid unstable areas with high probability. Because the resulting model serves both, as a probabilistic regressor, and classifies regions as stable or unstable, we call it Gaussian process for *classified regression* (GPCR). We illustrate the problem of learning with failures with a case example.

Example 1 *Consider tuning the policy parameters of a humanoid robot doing a simple task, like squatting. The goal is to find the parameters that minimize the tracking error of the movement. Such error is measured through a cost function that is evaluated after each experiment. During learning, we observe that some controllers are making the robot fall down, and some others cause huge vibrations on the joints. Since those cases are undesired, and considered as unstable, an operator prematurely stops the experiment pressing an emergency button. We consider all these cases as failures, for which no meaningful data is available.*

In Example 1, the unstable cases can be seen as consequence of two external constraints being violated, i.e., the velocity of the joints and the tilting angle of the torso surpassing certain thresholds. In case such thresholds are known, one can know during learning how close is a specific controller from being unstable by measuring the sensor signals. However, prior to *any* data collection, the designer does not know what those thresholds are, i.e., neither the largest joint velocities right before vibrations start to happen, nor the maximum angular position of the torso before the robot falls down. The proposed GPCR model can also be used to model such constraints when the constraint thresholds are unknown. Furthermore, GPCR estimates such unknown thresholds from data. This is the second core contribution of the paper.

In addition, the resulting constrained problem is addressed using Bayesian optimization methods under the presence of unknown constraints (BOC). To this end, recent BOC methods motivated by information theory (Hernández-Lobato et al., 2016) have demonstrated to outperform existing heuristics. However, they are computationally demanding. Instead, we have extended a recent information-theoretic Bayesian optimizer, Max-Value Entropy Search (Wang and Jegelka, 2017), which is computationally cheaper, to account for unknown constraints. The resulting algorithm is called Min-Value Entropy Search with unknown level-set constraints (mESCO). That is the third contribution of this work.

The paper is structured as follows: In Sec. 2, we formulate the constrained controller tuning problem, and define the key mathematical elements of the paper. In Sec. 3, we present the novel Bayesian model (GPCR) in the context of unknown failure penalties. In Sec. 4, a range of applications of the GPCR model are presented, including the extension to model black-box constraints, and specially the case in which constraint thresholds are unknown. In Sec. 5, Bayesian optimization with unknown constraints is explained, along with our extension (mESCO) of Max-Value Entropy Search. In Sec. 6, we discuss our contributions among the existing literature. In Sec. 7, we demonstrate the benefits on GPCR on several numerical simulations and experiments on a real robotic platform, and we conclude with a discussion in Sec. 8.

2. Preliminaries

In this section, we start with a general formulation of the controller tuning problem. We then define mathematically the unknown instability threshold, a key parameter of the GPCR model, and finally introduce notation for Gaussian process (GP).

2.1 Controller tuning problem

Let a system be defined by discrete unknown dynamics $s_{t+1} = h(s_t, u_t, w_t)$, where $s_t \in \mathbb{R}^W$ is the state, $u_t \in \mathbb{R}^V$ is the control signal, and $w_t \in \mathbb{R}^W$ is process noise, at time t . We assume the control signal is obtained from a feedback policy $u_t = \pi(s_t)$. Generally, we are interested in policies π that make the system accomplish a specific user-defined task, for example, picking an object with a robot arm and placing it somewhere else on the table. Such policies drive the system from the initial state s_0 to a terminal state s_T describing a time trajectory as a sequence of states and control inputs $\tau_T = \{s_t, u_t\}_{t=0}^T$. The cost of the system for not accomplishing the desired task is specified via a cost function $l : \mathbb{R}^W \times \mathbb{R}^V \rightarrow \mathbb{R}_{\geq 0}$. The cost of a full trajectory under the policy π is then defined as $f_\tau(\pi) = \mathbb{E} \left[\sum_{t=0}^{T-1} l(s_t, u_t) \right]$. While the true systems dynamics h are unknown, we assume to have access to an approximate model \hat{h} . In cases in which the model \hat{h} and the cost l have certain structure, it is possible to compute an optimal policy

$$\hat{\pi} = \underset{\pi}{\operatorname{argmin}} \quad \mathbb{E} \left[\sum_{t=0}^{T-1} l(x_t, \pi(x_t)) \right] \tag{1}$$

s.t. $s_{t+1} = \hat{h}(s_t, u_t, w_t)$.

For example, when the model \hat{h} is linear and the cost l is quadratic, the optimal policy $\hat{\pi}$ is the linear quadratic regulator (LQR) (Anderson and Moore, 1990). Generally speaking, while such policy $\hat{\pi}$ is optimal for the model \hat{h} , it is suboptimal for the true dynamics h . In order to recover the optimal policy for h , we parametrize $\hat{\pi}(s_t) = \hat{\pi}(s_t; x)$ with parameters $x \in \mathcal{X} \subset \mathbb{R}^D$ and learn the optimal parameters by collecting experimental data from the true system h . To this end, we obtain a different cost of a full trajectory $f_\tau(\tau_T(x))$ for each parametrization x . This induces an objective $f_\tau(\tau_T(x)) : \mathcal{X} \rightarrow \mathbb{R}$, which we write as $f(x)$ to simplify notation. Then, the learning problem is to find the parametrization x_{\min} that minimizes the trajectory cost

$$x_{\min} = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x) \tag{2}$$

on the true dynamics $s_{t+1} = h(s_t, u_t, w_t)$. The function f is unknown, as it depends on unknown dynamics h , but we can query it as a black-box function at location x by doing an experiment on the real system. For instance, in (Marco et al., 2016) the weighting matrices of an LQR controller design are parametrized with parameters x . Therein, the goal is to minimize an unknown cost objective f that quantifies the performance of a real robot balancing an inverted pole. Thus, querying the function f at location x consists of doing an experiment with policy $\hat{\pi}(s_t; x)$ and collecting sensor data to compute the corresponding cost value $f(x)$.

2.2 Instability threshold as a level-set constraint

In the RL literature, the data collected with an unstable controller is typically discarded, and a relatively large value is heuristically chosen to characterize its performance. Whereas we propose to also discard such data, we never choose any heuristic penalty. These ideas can be mathematically reflected in the objective f as follows.

Let us divide the domain \mathcal{X} in two sets: The set of stable controllers \mathcal{X}_s , for which function values $f(x)$ are known, and the set of unstable controllers \mathcal{X}_u , for which function values $f(x)$ are somewhat large, but unknown, with $\mathcal{X}_s \cup \mathcal{X}_u = \mathcal{X}$. To include these concepts in the objective f , we re-define it so that its image is

$$\begin{aligned} f(\mathcal{X}_u) &\in (c, +\infty) \\ f(\mathcal{X}_s) &\in [f_{\min}, c], \end{aligned} \tag{3}$$

where f is lower bounded by the global minimum $f_{\min} = \min_{x \in \mathcal{X}_s} f(x)$, and the image of the objective over the stable set is upper bounded by a threshold c , which is defined next.

Definition 1 *The instability threshold $c = \sup_{x \in \mathcal{X}_s} f(x)$ is the worst possible cost excluding those of the unstable set \mathcal{X}_u .*

In the context of Example 1, the instability threshold c can be seen as the worst squatting possible to observe without destabilizing the robot. From (3), it follows that the threshold c induces a level set $\mathcal{X}_c = \{x : f(x) = c\}$, with $\mathcal{X}_c \subset \mathcal{X}_s$. In addition, we say that c imposes a *level set-constraint* in f that divides the input space \mathcal{X} into the stable set $\mathcal{X}_s = \{x : f(x) \leq c\}$ and the unstable set $\mathcal{X}_u = \{x : f(x) > c\}$. Finally, since the values of $f(x)$ that correspond to stable controllers are “constrained” to live below c , we say that the function f is *self-constrained*.

It is important to remark that without complete knowledge of the function f , and prior to any data collection (i.e., never having realized an experiment on the system), c is unknown to the designer. However, in Sec. 3.4 we show that c can be estimated from data, and that such estimation is useful to target stable regions when searching for the optimum (2).

We have introduced the instability threshold c , a key concept that we use to characterize instability in the context of the image of the objective f . This concept is important, as it will be recurrently used in the remainder of the paper.

2.3 Gaussian process (GP)

The Bayesian model for the objective function f , proposed in Sec. 3, is approximated as a GP (Rasmussen and Williams, 2006). This is a probability distribution over the function f , such that every finite restriction to random function values $[f(x_1), \dots, f(x_N)]$ follows a multivariate Gaussian distribution. That means, the objective function f is modeled as $f \sim \mathcal{GP}(m(x), k(x, x'))$, with prior mean $m : \mathcal{X} \rightarrow \mathbb{R}$, and positive definite covariance function (also called kernel) $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. The kernel encodes the strength of statistical correlation between two latent function values $f(x_i)$ and $f(x_j)$, where $x_i, x_j \in \mathcal{X}$. The model proposed in Sec. 3 assumes a zero-mean prior, $m(x) \equiv 0$.

In the remainder of this paper we will refer to *controller parametrization* x simply as *controller* x . In addition, we will refer to a *stable/unstable controller* x as stable/unstable x .

3. Gaussian process model for classified regression (GPCR)

In the following, we present the Bayesian model for classified regression and how to approximate it as a Gaussian process. Such model for the objective function f is presented in (2). This model inherently encodes the particular distinction between unstable and stable controllers through the instability threshold c defined in Sec. 2.2.

3.1 Likelihood model for unknown penalties

Let us assume that for each parameter x_i we have access to two types of observations: (i) a binary label $l_i \in \{-1, +1\}$ that determines whether the experiment was unstable or stable, and (ii) a noisy cost value $y_i \in \mathbb{R}$ only when the experiment is stable:

$$(y_i, l_i) = \begin{cases} (f_i + \varepsilon, +1), & \text{if } x \in \mathcal{X}_s \\ (\emptyset, -1), & \text{if } x \in \mathcal{X}_u \end{cases}, \quad (4)$$

where we have used the shorthand notation $f_i = f(x_i)$, $y_i = y(x_i)$, and $l_i = l(x_i)$. As stated in Sec. 2.2, we assume that all stable controller parametrizations $x_i \in \mathcal{X}_s$ yield noise-corrupted observations $y_i = f_i + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\varepsilon; 0, \sigma^2)$. In the context of the problem described in Sec. 2.1, the observation noise arises from the process noise w_t in the system, and the need of approximating the expectation from (1) in practice (Marco et al., 2016). For unstable controller parametrizations $x_i \in \mathcal{X}_u$, we assume that no continuous-valued observation was obtained, but only a discrete label.

The graphical model in Fig. 1 expresses the relation between variables. For a given x_i , having full information about f_i is sufficient to determine whether the controller is stable or not (l_i), and whether we have access to a cost value observation or not (y_i). Notice that the noisy evaluation y_i is not fully determined by f_i , but also requires the additional information about the associated label l_i (cf. (4)). Such model can be encoded in a likelihood over the observation (y_i, l_i) of the latent function f_i at location x_i as

$$p(y_i, l_i | f_i, x_i) = H(l_i(c - f_i)) (1_{\{l_i=+1\}} \mathcal{N}(y_i; f_i, \sigma^2) + 1_{\{l_i=-1\}}), \quad (5)$$

where the dependency on x_i in the right-hand side is redundant, but it has been kept for clarity. $H(z)$ is the Heaviside function, i.e., $H(z) = 1$, if $z \geq 0$, and $H(z) = 0$ otherwise. The likelihood (5) captures our knowledge about the latent function: When x_i is unstable, all we know about f_i is that it takes any possible value above the threshold c , with all values $f_i \geq c$ being equally likely, but we never specify what that value is. If x_i is stable, all the probability mass falls below c for consistency, and the indicator function $1_{\{\cdot\}}$ toggles the contribution of the Gaussian.

Let the latent objective f be observed at locations $X = \{X_s, X_u\}$, which entails both, stable $X_s = \{x_i\}_{i=1}^{N_s}$ and unstable controllers $X_u = \{x_i\}_{i=N_s+1}^N$, with $N = N_s + N_u$. The corresponding latent objective values at the stable locations X_s are $\mathbf{f}_s = [f_1, f_2, \dots, f_{N_s}]^\top$, and $\mathbf{f}_u = [f_{N_s+1}, f_{N_s+2}, \dots, f_N]^\top$ at X_u . We group the latent objective values in a vector $\mathbf{f} = [\mathbf{f}_s^\top, \mathbf{f}_u^\top]^\top$. The latent function values induce observations grouped in the set $\mathcal{D} =$

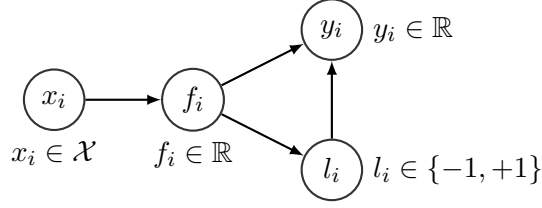


Figure 1: Graphical model for Classified Regression¹

$\{y_i, l_i\}_{i=1}^{N_s} \cup \{l_i\}_{i=N_s+1}^N$. Since \mathcal{D} contains both discrete labels and real scalar values, we refer to it as a *hybrid* set of observations. We assume such observations to be conditionally independent and identically distributed following (5), which allows the likelihood to factorize $p(\mathcal{D}|\mathbf{f}, X) = \prod_{i=1}^N p(y_i, l_i|f_i, x_i)$. Then, the likelihood over the data set \mathcal{D} becomes

$$p(\mathcal{D}|\mathbf{f}, X) = \prod_{i=1}^{N_s} H(c - f_i) \mathcal{N}(y_i|f_i, \sigma^2) \prod_{i=N_s+1}^N H(f_i - c). \quad (6)$$

3.2 Posterior probability

The posterior follows from Bayes theorem

$$p(\mathbf{f}|\mathcal{D}, X) \propto p(\mathcal{D}|\mathbf{f}, X)p(\mathbf{f}|X), \quad (7)$$

with zero-mean Gaussian prior $p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K})$ and likelihood (6). The last term in (6) can be rewritten as a multivariate Gaussian $\prod_{i=1}^{N_s} \mathcal{N}(y_i|f_i, \sigma^2) = \mathcal{N}(\mathbf{y}_s|\mathbf{f}_s, \sigma^2 \mathbf{I})$. Then, we obtain the posterior as

$$p(\mathbf{f}|\mathcal{D}, X) \propto \prod_{i=1}^{N_s} H(c - f_i) \prod_{i=N_s+1}^N H(f_i - c) \mathcal{N}(\mathbf{y}_s|\mathbf{f}_s, \sigma^2 \mathbf{I}) \mathcal{N}\left(\begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} K_{ss} & K_{su} \\ K_{us} & K_{uu} \end{bmatrix}\right), \quad (8)$$

where K_{ss} , $K_{su} = K_{us}$, and K_{uu} are the prior covariance matrices expressing the correlation between function evaluations at stable (s) and/or unstable (u) locations. The last two terms are the product of two multivariate Gaussians of different dimensionality. This is equal to another unnormalized Gaussian $\alpha \mathcal{N}(\mathbf{f}; \tilde{\mathbf{m}}, \tilde{\Sigma})$, whose mean, variance and scaling factor depend on the observations and the noise, i.e., $\tilde{\mathbf{m}} = \tilde{\mathbf{m}}(\mathbf{y}_s, \sigma^2)$, $\tilde{\Sigma} = \tilde{\Sigma}(\sigma^2)$ and $\alpha = \alpha(\mathbf{y}_s, \sigma^2)$, respectively (cf. Appendix A). Hence, the posterior becomes

$$p(\mathbf{f}|\mathcal{D}, X) \propto \prod_{i=1}^{N_s} H(c - f_i) \prod_{i=N_s+1}^N H(f_i - c) \mathcal{N}(\mathbf{f}; \tilde{\mathbf{m}}, \tilde{\Sigma}), \quad (9)$$

where the scaling factor α has been omitted for simplicity. The role of the Heaviside functions is to restrict the support of $p(\mathbf{f}|\mathcal{D}, X)$ to an unbounded hyperrectangle with a single corner at

1. While the objective function (3) is lower bounded by f_{\min} , an instance f_i at location x_i has infinite support. While this might seem contradictory, it is actually possible because (3) represents the true (unknown) objective, and f_i, y_i are random variables, assumed to be Gaussian distributed, and thus $f_i, y_i \in \mathbb{R}$.

location $f_i = c$, $\forall i \in \{1, \dots, N\}$. Thus, (9) can be geometrically understood as the product of a multivariate Gaussian density of dimension N times an activation function that sets to zero any probability outside such hyperrectangle. Computing the right-hand side of (9) is analytically intractable. However, it can be approximated with an unnormalized Gaussian

$$p(\mathbf{f}|\mathcal{D}, X) \propto q(\mathbf{f}|\mathcal{D}, X) = Z_{\text{EP}} \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}_{\text{EP}}, \boldsymbol{\Sigma}_{\text{EP}}), \quad (10)$$

where Z_{EP} , $\boldsymbol{\mu}_{\text{EP}}$, and $\boldsymbol{\Sigma}_{\text{EP}}$ are computed following (Cunningham et al., 2011). Therein, the posterior of a multivariate Gaussian, whose support is restricted to bounded polyhedra (i.e., a more general case than the one from (9)), is approximated using Expectation propagation (EP) (Minka, 2001). Our implementation of such EP approximation performs comparably to computing the posterior using sampling methods, e.g., Elliptical Slice Sampling (Murray et al., 2010). However, we have found EP to need much lower wall-clock time to yield the same results.

3.3 Predictive probability

Our goal is to make inference at an unobserved location x_* , where the predictive latent function value $f_* = f(x_*)$ is jointly distributed with the vector of latent function values at the observations \mathbf{f} , i.e., $p(\mathbf{f}, f_*)$. The joint posterior is

$$p(\mathbf{f}, f_*|\mathcal{D}, X, x_*) \propto p(\mathcal{D}|\mathbf{f}, X)p(\mathbf{f}, f_*|X, x_*), \quad (11)$$

where the likelihood term does not depend on the unobserved location x_* , nor on f_* . The predictive distribution of f_* conditioned on the collected data can then be obtained by marginalizing over the latent function values \mathbf{f} from the joint posterior as

$$p(f_*|\mathcal{D}, X, x_*) \propto \int_{\mathbf{f}} p(\mathcal{D}|\mathbf{f}, X)p(\mathbf{f}, f_*|X, x_*)d\mathbf{f}, \quad (12)$$

where the likelihood term is defined as in (6), and we extend the prior with the unobserved location x_* as

$$p(\mathbf{f}, f_*|X, x_*) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix}; \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} K_{\mathbf{f}\mathbf{f}} & K_{\mathbf{f}f_*} \\ K_{*f} & K_{**} \end{bmatrix}\right). \quad (13)$$

The multivariate integral in (12) is analytically intractable due to the structure of the likelihood (6). However, we can alleviate this by applying some transformations. First, by combining $p(\mathbf{f}, f_*|X, x_*) = p(\mathbf{f}|X)p(f_*|\mathbf{f}, X, x_*)$ and (7), the integral from (12) can be rewritten as $\int_{\mathbf{f}} p(\mathbf{f}|\mathcal{D}, X)p(f_*|\mathbf{f}, X, x_*)d\mathbf{f}$. Then, using the Gaussian approximation (10) of the posterior $p(\mathbf{f}|\mathcal{D}, X)$, the resulting predictive distribution is also Gaussian $q(f_*) = \mathcal{N}(f_*; \mu_*, \sigma_*^2)$, with mean μ_* and variance σ_*^2 as a function of the mean $\boldsymbol{\mu}_{\text{EP}}$ and covariance $\boldsymbol{\Sigma}_{\text{EP}}$ of the posterior (Rasmussen and Williams, 2006, Sec. 3.4.2). Using such approximated posterior (10), the moments of $q(f_*)$ are analytically given by

$$\begin{aligned} \mu_* &= K_{*f} K_{\mathbf{f}\mathbf{f}}^{-1} \boldsymbol{\mu}_{\text{EP}} \\ \sigma_*^2 &= K_{**} - K_{*f} K_{\mathbf{f}\mathbf{f}}^{-1} \left(\mathbf{I} + \boldsymbol{\Sigma}_{\text{EP}} K_{\mathbf{f}\mathbf{f}}^{-1} \right) K_{f_*}. \end{aligned} \quad (14)$$

We illustrate in Fig. 2 (top) the approximate posterior $q(f_*)$ for the setting described in Example 2. A detailed analysis of the quality of the approximation $q(f_*)$ to $p(f_*|\mathcal{D}, X, x_*)$ can be found in Appendix B.

Example 2 Consider the unknown objective $f : \mathcal{X} \rightarrow \mathbb{R}$, $\mathcal{X} = [0, 1]$, for which we have three stable evaluations $X_s = [0.1, 0.3, 0.5]$, $\mathbf{y}_s = [0.5, 2.0, 1.0]$ and two unstable evaluations $X_u = [0.7, 0.9]$. We assume small noise $\sigma_n = 0.02$, and a zero-mean prior with covariance \mathbf{K} , whose entries $K_{ij} = k(x_i, x_j)$ are computed with an ARD Matérn 3/2 kernel k with signal variance 0.5 and lengthscale 0.2.

The univariate distribution $q(f_*)$ is straightforwardly extended to a set of unobserved locations \mathbf{x}_* , where we consider the joint prior $p(\mathbf{f}, \mathbf{f}_*)$. This gives rise to a multivariate approximate Gaussian posterior $q(\mathbf{f}_*) = \mathcal{N}(\mathbf{f}_*; \boldsymbol{\mu}_*, \boldsymbol{\sigma}_*^2)$, i.e., a Gaussian process.

The threshold c influences the predictive distribution through the likelihood model (6), which induces an intractable posterior that, when approximated as a multivariate Gaussian, yields a Gaussian process model. In addition, the threshold c can be seen as a discriminator that distinguishes instability from stability in the axis of the cost value. Because of the closeness to the idea of classification, but reinterpreted in the regressor, we call this model: Gaussian process for *classified regression* (GPCR).

In the next section we give detailed insights about the influence of c in the model, and also how it can be learned from data.

3.4 Instability threshold as a model parameter

The instability threshold c plays an important role in the likelihood model (6). However, prior to data observation, one has in principle no correct notion about an appropriate value for it. Instead, we can compute an estimate \hat{c} from data using GPCR. Herein, we show two possible ways of computing \hat{c} .

3.4.1 ESTIMATION VIA MAXIMUM LIKELIHOOD (ML)

One possible way to estimate c from data is by maximizing the marginal likelihood $p(\mathcal{D}|X)$, which can be computed by integrating out \mathbf{f} from the right-hand side of (9). Since that is analytically intractable, we instead use the integral over the approximate posterior $q(\mathbf{f}|\mathcal{D}, X)$ (10) to compute its marginal

$$\hat{c} = \operatorname{argmax}_c \log \int_{\mathbf{f}} q(\mathbf{f}|\mathcal{D}, X, c) d\mathbf{f}, \quad (15)$$

where the dependency on c has been introduced in $q(\mathbf{f}|\mathcal{D}, X)$ for clarity. From (10) follows that $\int_{\mathbf{f}} q(\mathbf{f}|\mathcal{D}, X, c) d\mathbf{f} = Z_{\text{EP}}(c)$. Thus, solving (15) implies performing a line search, for which every call to the function $Z_{\text{EP}}(c)$ involves solving an EP loop.

3.4.2 ESTIMATION VIA MAXIMUM A POSTERIORI (MAP)

Another possible way of estimating c is to consider it as a stochastic variable distributed with a *hyperprior* $p(c)$, and compute the maximum of the posterior distribution $p(c|\mathcal{D}) = \int_{\mathbf{f}} p(\mathcal{D}|\mathbf{f}, X, c) p(\mathbf{f}|X) p(c) d\mathbf{f}$, where we have assumed that c and \mathbf{f} are uncorrelated. Using the approximate posterior (10) over \mathbf{f} , and Bayes' rule, we obtain the posterior $p(c|\mathcal{D}) \propto \int_{\mathbf{f}} q(\mathbf{f}|\mathcal{D}, X, c) p(c) d\mathbf{f}$. Due to the lack of prior knowledge about c , we propose a wide Gaussian

hyperprior $p(c) = \mathcal{N}(c; \mu_c, \sigma_c^2)$, whose parameters are to be selected by the designer. Similar to (15), we find \hat{c} by solving the optimization problem

$$\hat{c} = \operatorname{argmax}_c \log Z_{\text{EP}}(c) - 0.5(c - \mu_c)^2 / \sigma_c^2. \quad (16)$$

An advantage of ML over MAP is that the former requires no prior reasoning about c at all. However, we have found in our experiments that ML can lead to inconvenient estimations of c when stable evaluations are absent. In such case, we see from (9) that the posterior probability becomes larger since the term $\prod_{i=N_s+1}^N H(f_i - c)$ activates a larger portion of the Gaussian, and has a maximum at $\hat{c} = -\infty$. Such estimation can cause problems when using the model to perform Bayesian optimization under unknown constraints, as explained in Sec. 5. Analogously, when only stable evaluations are present, the maximum is found at $\hat{c} = +\infty$. On the contrary, MAP enjoys the hyperprior over c as a regularizer that avoids such extreme point estimates.

The estimated instability threshold \hat{c} becomes important when determining the probability of a specific unforeseen location x resulting in a stable controller. In the next section, we describe how to compute such probability.

3.5 Probabilistic stability constraint in GPCR

Given our current estimate of the instability threshold \hat{c} , we can compute the probability of a specific point x being a stable observation, i.e., $\Pr[x \in \mathcal{X}_s]$. From Sec. 2.2, it follows that $\Pr[x \in \mathcal{X}_s] = \Pr[f(x) \leq \hat{c}]$. After having observed data \mathcal{D} , we define

$$\Pr[f(x) \leq \hat{c} \mid \mathcal{D}] = \int_{-\infty}^{\hat{c}} p(f \mid \mathcal{D}, x) \, df \simeq \int_{-\infty}^{\hat{c}} q(f) \, df = \Phi(z(x)), \quad (17)$$

where the predictive distribution (12) has been rewritten as $p(f \mid \mathcal{D}, x)$ for simplicity, and its approximation (14) as $q(f) = \mathcal{N}(f; \mu(x), \sigma^2(x))$, where Φ is the CDF of a univariate Gaussian distribution, and $z(x) = (\hat{c} - \mu(x)) / \sigma(x)$. The predictive constraint (17) is depicted in Fig. 2, using the synthetic Example 2.

4. Applications and extensions of GPCR

The GPCR framework introduced in Sec. 3 can be used to model the objective function from (2), but it is not limited to it. One could think about the class of unstable controllers as something wider, i.e., any controller that results in any undesired behavior. For example, a robot arm achieving a stable trajectory, but hitting an obstacle on the way. Such undesired behavior can be measured, for example, by computing the euclidean distance from the robot arm to the obstacle. This information is much richer than a discrete label to indicate success/failure and could, for instance, be treated as an additional external black-box function, modeled as a standard GP², and same for the objective. In such a case, at first sight it seems like Bayesian optimization under unknown constraints (Gardner et al., 2014; Gelbart et al., 2014) suffices to address the constrained problem and GPCR is not needed. However,

2. By “standard GP” we mean the most common use of a GP, i.e., using a Gaussian likelihood, as opposed to the GPCR model proposed herein.

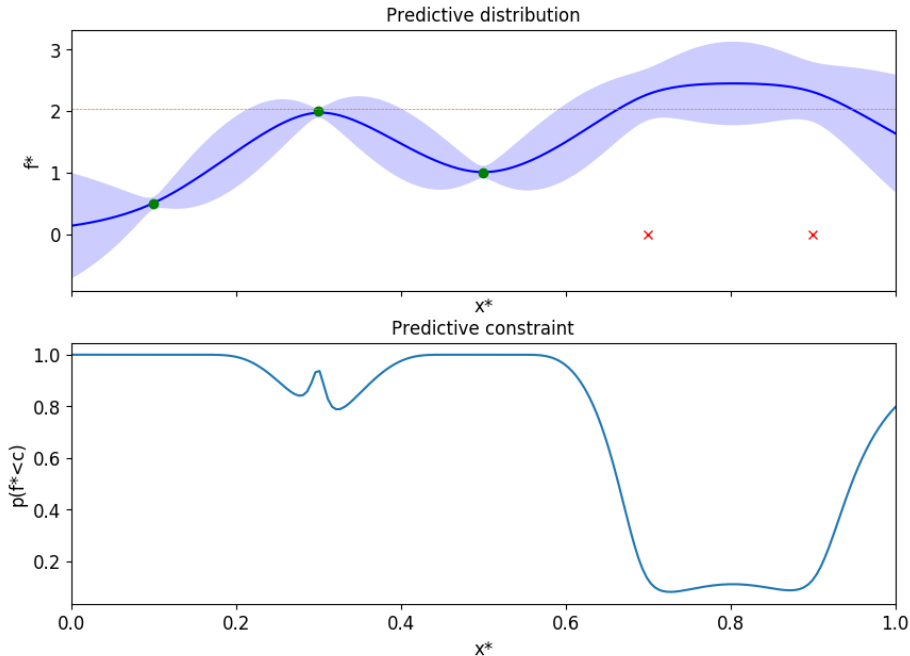


Figure 2: (top) GPCR model posterior conditioned on three stable evaluations (green dots), and two unstable evaluations (red crosses). The estimated instability threshold \hat{c} (orange dashed line) is used by the GPCR model to push up the probability mass in those regions where unstable observations are predicted. (bottom) Probabilistic constraint $\Pr[f(x) < \hat{c} \mid \mathcal{D}]$ of the GPCR model, as detailed in (17). Both plots follow the settings from Example 2.

there exist many scenarios in which GPCR comes at handy; for example, when the constraint threshold is unknown (i.e., in the aforementioned example with the robot arm, we do not know a priori where the obstacle is). In that case, one can model the constraint using GPCR, and see the unknown constraint threshold as the instability threshold introduced in Sec. 3.4. By doing so, we could also learn the constraint threshold from data.

Including the aforementioned case, we have identified in total four different cases where the GPCR model plays an important role in either modeling the constraint, the objective, or both. However, before explaining each case, we need to introduce some new notation and concepts that will make the differences clear. Let us assume the optimization problem (2) is constrained under $M \geq 1$ black-box external constraints, which results in the constrained minimization problem

$$\begin{aligned} \tilde{f}_{\min} &= \min_{x \in \mathcal{X}_s} f(x) \\ &\text{with } \mathcal{X}_s := \{x : g_1(x) \leq c_1, \dots, g_M(x) \leq c_M\}, \end{aligned} \quad (18)$$

where $c_j \in \mathbb{R}$, $\forall j = \{1, \dots, M\}$ are the constraint thresholds. From now on, we assume that evaluations are *coupled*, i.e., when evaluating at x_i , we obtain measurements from both

the objective, and all the constraints (Hernández-Lobato et al., 2016). Additionally, we distinguish between “stable” and “safe” controllers: A controller x_i is considered unsafe, but stable, when it violates at least one of the constraints, but it keeps the system stable (e.g., the robot arm performs a stable trajectory, but it finds an obstacle in the way). Similarly, x_i is considered safe, but unstable, when it does not violate any constraint, but destabilizes the system (e.g., the robot arm executes a poor controller, which destabilizes the system, but it does not hit any obstacle). Consequently, we redefine \mathcal{X}_s more broadly as to include both, stable and safe controllers, as opposed to the definition from Sec. 2.2. Although the evaluations are coupled “in location”, we assume they are uncoupled “in observation”. For example, an unsafe but stable controller yields a discrete label when evaluating the constraint, but a continuous-valued observation when evaluating the objective.

The nature of the constraints (18) is important when deciding whether the GPCR model is useful or not. More concretely, when it comes to obtain a query from a constraint g_j , we distinguish two characteristic cases: (i) the query value $g_j(x)$ exists when it is satisfied, and it does not exist when it is violated (instead a binary label is obtained), and (ii) the query value $g_j(x)$ does not exist in either case, and we only obtain a binary outcome that determines constraint satisfaction/violation. In the first case, we characterize g_j as a *level-set constraint*, while in the second case, we characterize it as a *binary constraint*. For modeling a level-set constraint, GPCR has benefits over standard GPs, because it can estimate the constraint threshold during learning. For modeling a *binary constraint*, GPCR has no special benefit, as the dataset is not hybrid (cf. Sec. 3.1), but binary. Given these two clearly distinctive groups, let us assume M_L level-set constraints and M_B binary constraints, with $M = M_L + M_B$.

Using the aforementioned details, we discuss separately the benefits of GPCR in four different scenarios that arise from particular characterizations of the constraint and the objective. The explanations are supported with illustrations (Fig. 3) on a simple one dimensional setting with one objective function and one constraint. In all cases, we also show the decision-making of the Bayesian optimizer that addresses the constrained minimization problem (18), and will be detailed in Sec. 5. As a quick reference for the reader, we briefly summarize here the four cases:

- Case 1) Single constraint equal to the objective; $M = M_L = 1$; self-constrained objective (Sec. 4.1).
- Case 2) Multiple binary constraints; $M = M_B \geq 1$; constraints absorbed by the objective (Sec. 4.2).
- Case 3) Multiple level-set constraints; $M = M_L$; objective modeled as a standard GP (Sec. 4.3).
- Case 4) Self-constrained objective, and multiple level-set constraints; $M_L \geq 1$ and $M_B \geq 1$ (Sec. 4.4).

4.1 Case 1) Self-constrained objective

The general problem formulation (18) finds its simplest case when the objective is also the constraint, i.e., $M = M_L = 1$, with $f(x) = g_1(x)$ and $c = c_1$. This case is equivalent to the one explained in Sec. 2.2: The threshold c imposes a level-set constraint on f (we say f is *self-constrained*), with stability region $\mathcal{X}_s = \{x : g_1(x) \leq c_1\} = \{x : f(x) \leq c\}$. Fig. 3a shows a possible situation corresponding to such a scenario. The objective is modeled with GPCR, and the success/failure binary information is captured by it. The true constraint is also shown for clarity to emphasize that it is identical to the true objective, but it does not play any role. We see that GPCR pushes the probability mass above the current estimate of the instability threshold \hat{c} , in areas with unstable evaluations. The first clear benefit of GPCR in this case is that no heuristic cost needs to be assigned to unstable controllers by the designer.

The second, is that the Bayesian optimizer (bottom plot of Fig. 3a) sees those regions as less interesting because the cost is predicted to be high.

4.2 Case 2) Multiple binary constraints

In this case, we assume all constraints to be binary ($M = M_B \geq 1$). Such constraint provides only discrete labels indicating success/failure. Thus, instead of modeling it separately (e.g., using a Gaussian process classifier), its discrete information is directly absorbed in the GPCR model used to model the objective. In Fig. 3b, the region of stability is given by $\mathcal{X}_s = \{x : g_1(x) \leq c_1\}$. When evaluating $f(x)$ in the middle region, where the constraint is violated, three unstable controllers are obtained. Therefore, the GPCR model pushes the probability mass above \hat{c}_1 . That way, those regions have a high probable cost. The first clear benefit of GPCR in this case is that one model handles both, the information of the objective and the constraint. The second clear benefit of GPCR is that it automatically makes the unstable region less interesting for the Bayesian optimizer, which selects the next evaluation in a stable area. The estimated instability threshold \hat{c}_1 is found above all collected stable points, but in this case it has no meaning beyond its role as a parameter within the GPCR model. It merely serves to push the cost probability above already collected stable evaluations.

4.3 Case 3) Multiple level-set constraints

Herein, we assume all constraints to be level-set constraints ($M = M_L$). In Fig. 3c, the constraint g_1 is a level-set constraint, modeled with GPCR, with the stable set determined by $\mathcal{X}_s = \{x : g_1(x) < c_1\}$, and c_1 being unknown a priori. On the contrary, the objective function itself is not level-set constrained, and is modeled with a standard GP. The Bayesian optimizer searches the constrained minimum \tilde{f}_{\min} in areas where the constraint is satisfied with high probability. The estimated instability threshold \hat{c}_1 (orange horizontal line) herein estimates the *true* constraint threshold c_1 (grey horizontal line). The main benefit of GPCR in this case is that c_1 does not need to be known by the designer a priori. Instead, it can be estimated while collecting data.

4.4 Case 4) Multiple binary and level-set constraints

Last, we discuss the case in which $M_L \geq 1$ and $M_B \geq 1$. Generally, we model each of the M_L level-set constraints independently with a GPCR model, and we consider the M_B binary constraints absorbed into the GPCR model for f . In total, this comprises $M_L + 1$ models. This is a more general case than the ones before, and mixes Sec. 4.2 and Sec. 4.3 in one. Fig. 3d shows three main advantages of GPCR in this scenario: (i) The designer does not need to know a priori the threshold for the the level-set constraints, (ii) the binary constraints do not need to be modeled in addition, but they are rather absorbed by the model for f , and (iii) the designer does not need to define a penalty value for f . It is worth remarking that this case is equivalent to having $M_L \geq 1$ level-set constraints and a self-constrained objective. The reasoning is that, given a self-constrained objective, one can always construct a set of virtual binary constraints, which combined together, divide the space into the same

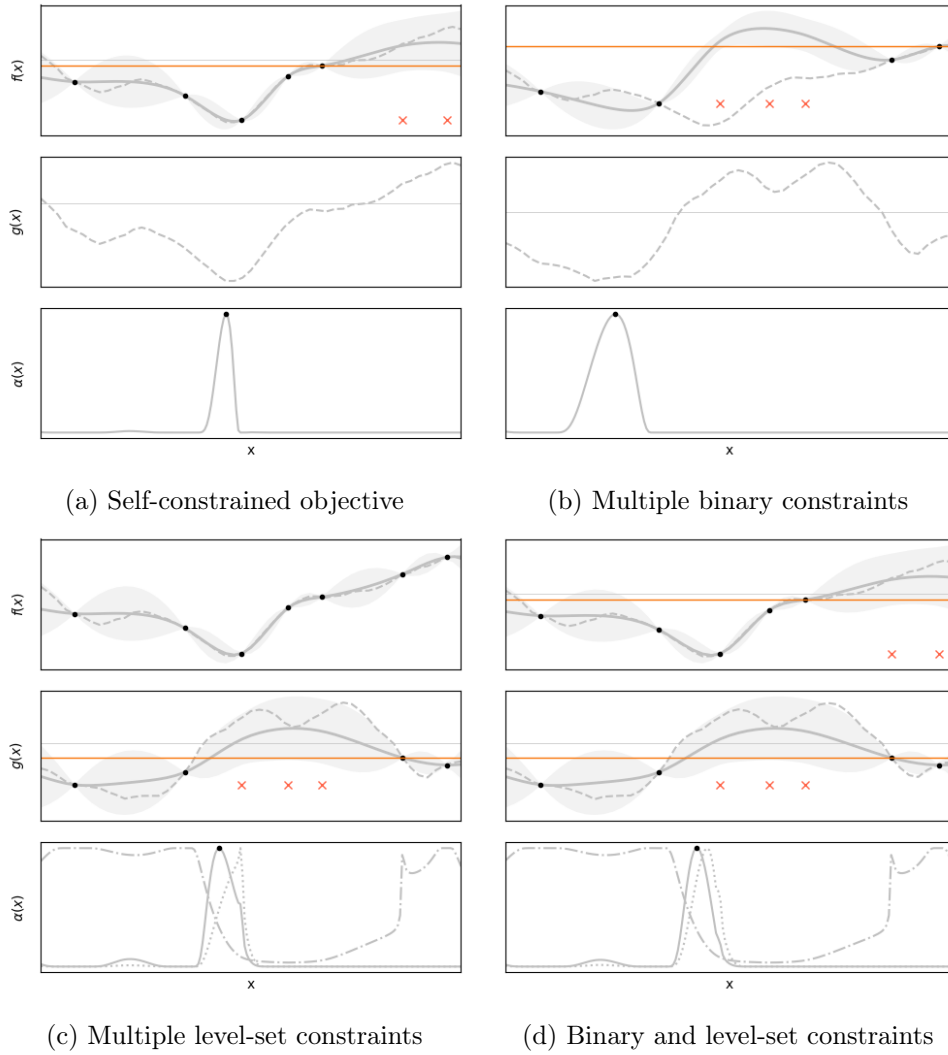


Figure 3: One dimensional synthetic example that shows the use of the GPCR model in four different scenarios. For each scenario, the top plot shows the objective function f and the middle plot shows the constraint g . When possible, the top and middle plots include the true function (dashed grey line) and the collected evaluations: stable (black dots) and unstable (red crosses). Also, when possible, the estimated instability threshold \hat{c} (orange horizontal line), and the true threshold c (grey horizontal line) are also shown. The bottom plot shows the acquisition function (grey solid line) of the Bayesian optimizer. In plots (c) and (d), the probability of constraint satisfaction (dash-dotted line), and the unconstrained acquisition function (dotted line) are also shown as a reference. All three quantities are normalized to fit to the full range of the vertical axis.

stable/unstable areas implied by the instability threshold of the self-constrained objective.

We have illustrated the benefits of GPCR and its capability to handle multiple constraints in four different simple scenarios. However, it still remains a question how to solve (18) for each of the aforementioned four cases. In the next section, we show how constrained Bayesian optimization can be leveraged to address (18), and explain the nuisances for each of the four cases.

5. Constrained Bayesian optimization and GPCR

Standard Bayesian optimization (BO) attempts to solve problem (2) by sequentially deciding which controller x_{next} shall be evaluated in the next experiment. Such decision is made based on a specific criterion, e.g., maximize information gain about the optimum (Hennig and Schuler, 2012), trade-off exploration vs. exploitation (Srinivas et al., 2010), or improve upon the best point observed so far (Jones et al., 1998). For an extensive review on the recent BO methods, see (Shahriari et al., 2016). The aforementioned criteria are represented by an acquisition function $\alpha : \mathcal{X} \rightarrow \mathbb{R}$, typically computed through the Gaussian process model on f , updated with the function values revealed up to the current iteration. The maximizer of α is used to decide on the next experiment $x_{\text{next}} = \arg \max_{x \in \mathcal{X}} \alpha(x)$.

In presence of M black-box constraints, *Bayesian optimization with unknown constraints* (BOC) (Gardner et al., 2014; Gelbart, 2015; Gelbart et al., 2014; Griffiths and Hernández-Lobato, 2017; Schonlau et al., 1998; Snoek, 2013) has been proposed to address problem (18). Therein, the main idea is to target experiments where the constraints are satisfied with high probability.

In the following, we discuss the benefits of GPCR in the context of BOC and BO for the four cases described in Sec. 4. In Sec. 5.2 we discuss that, when all constraints are binary (i.e., Sec. 4.2), these can be absorbed into the GPCR model, and thus standard BO suffices to solve the constrained optimization problem. Furthermore, in Sec. 5.1 we show that, when at least one of the constraints is a level-set constraint, we can incorporate it into the optimization problem, and address it using constrained Bayesian optimization.

Although existing BOC methods have proven to be successful in practice (Gelbart et al., 2014), the underlying decision-makers are heuristics developed based on intuition. Alternatively, entropy-based methods (Hennig and Schuler, 2012), constructed over the principle of information gain about the optimum, have proven to outperform the aforementioned heuristics. However, entropy-based methods able to handle multiple constraints are scarce and computationally expensive, i.e., only (Hernández-Lobato et al., 2016), to the best of our knowledge. Therefore, we propose to extend a recent entropy-based criterion (Wang and Jegelka, 2017), which is computationally cheaper than (Hernández-Lobato et al., 2016), to account for unknown constraints.

5.1 Bayesian optimization with unknown level-set constraints

Sec. 4.3 and Sec. 4.4 illustrate the case in which at least one of the M constraints of (18) is a level-set constraint, modeled with GPCR. Since such model describes the constraint probabilistically, we cannot guarantee that the constraint will be satisfied/violated in locations x where observations have not been collected. Furthermore, even at locations where the

constraint is satisfied and a continuous-valued observation is acquired, we are unable to guarantee constraint satisfaction/violation, because the observation is contaminated with noise. As a result, it is not possible to ensure that the constraints are satisfied for any x . In light of these challenges, the minimizer of problem (18) can be approximated by

$$\operatorname{argmin}_{x \in \mathcal{X}} \mathbb{E}[f(x)|\mathcal{D}], \text{ s.t. } \prod_{j=1}^M \Pr[g_j(x) \leq c_j | \mathcal{D}_j] \geq 1 - \delta, \quad (19)$$

where $1 - \delta$ represents some user-defined level of confidence, with $\delta \in [0, 1]$, and we have assumed that the constraints are modeled independently. This formulation has been proposed by (Gelbart et al., 2014; Gramacy and Lee, 2011). The idea is to minimize the objective function in expectation while satisfying the constraints with high probability. The same authors also propose a powerful procedure to include the information provided by the probabilistic constraints into the exploration strategy given by the acquisition function α . Although in their framework they consider expected improvement (EI) (Mockus et al., 1978) as the acquisition function, they also state that their formulation is valid for any generic acquisition function α . We use herein their formulation to propose a failures-aware acquisition function $\tilde{\alpha}$, which tends to explore in regions where the constraints are satisfied with high probability

$$\tilde{\alpha}(x) = \begin{cases} \alpha(x) \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j], & \text{if } \exists x, \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j] \geq 1 - \delta \\ \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j], & \text{otherwise.} \end{cases} \quad (20)$$

The acquisition (20) has two modes of operation: (a) If there exists at least one x for which all constraints are satisfied with high probability, we downsize $\alpha(x)$ with the probability of constraint satisfaction (i.e., points that are more likely to satisfy the constraints will get a higher weight); (b) if the constraints are violated for all x with high probability, whatever point is suggested by $\alpha(x)$ will violate the constraint as well. In this case, it is better to exclude $\alpha(x)$ and select the point that has the highest probability of satisfying the constraint (Griffiths and Hernández-Lobato, 2017). The latter is merely an exploratory strategy that will seek for some region where the constraint is satisfied. Once such a region is found, the strategy switches to (a).

An important difference of (20) with respect to (Gelbart et al., 2014) is that we assume the true constraint thresholds c_j to be unknown. Instead, we consider the estimated thresholds \hat{c}_j , available through the GPCR model (see Sec. 3.4). In fact, when modeling g_j with GPCR, the probability $\Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j]$ is a natural consequence of the model and can be computed following (17).

Fig. 3c and 3d show the operation mode (a) of the acquisition function $\tilde{\alpha}$ in light of the constraint g . In both scenarios, the probability of constraint satisfaction $\Pr[g(x) < \hat{c}_g | \mathcal{D}_g]$, which multiplies the unconstrained acquisition function α , are also shown for clarity. In both cases it is clear that downsizing α by the probability of constraint satisfaction changes the decision making to areas where it is less likely to violate the constraint.

5.2 Bayesian optimization with unknown binary constraints

Herein, we consider the case in which all the constraints g_j from (18) are binary. As detailed in Sec. 4.2, we propose to absorb them by the GPCR model of f . This makes the

model push up the probability of the cost at regions predicted to be unstable. Then, those regions will automatically look less appealing to the acquisition function of standard BO algorithms. Therefore, it becomes sufficient to deploy standard BO instead of BOC, although the problem (18) is a *constrained* optimization problem. This is the main benefit of using the GPCR model in presence of only binary constraints.

Interestingly enough, one could anyways use the BOC strategy (20) in this scenario by considering $\Pr[f(x) < \hat{c} \mid \mathcal{D}_g]$ as the only probabilistic constraint. However, in our experiments such constraint seemed to be redundant, since regions where $\Pr[f(x) < \hat{c} \mid \mathcal{D}_g]$ is small are anyways rarely explored.

The acquisition function α in (20) is a placeholder that can be replaced by any improvement criterion (Gelbart et al., 2014), in any problem with level-set constraints. In the following, we propose a Bayesian optimization criterion that accounts for unknown level-set constraints.

5.3 Min-Value Entropy Search with Unknown Level-Set Constraints (mESCO)

Max-Value Entropy Search (Wang and Jegelka, 2017) is a Bayesian optimizer that addresses the problem (2) from an information-theoretic perspective. Whereas (Wang and Jegelka, 2017) consider a maximization problem in their formulation, we characterize it from the point of view of a minimization problem (mES), and extend it to account for unknown level-set constraints (mESCO). The resulting algorithm is called *min-Value Entropy Search with Unknown Level-Set Constraints* (mESCO). In the following, we first introduce mES and then explain the pertinent modifications that lead to mESCO.

5.3.1 mES

The acquisition function proposed by mES is defined as the gain in mutual information between the next point to query $\{x, y\}$ and the minimum f_{\min}

$$\begin{aligned} \alpha_{\text{mES}}(x) &= \text{MI}(\{x, y\}; f_{\min} \mid \mathcal{D}) \\ &= H(p(f \mid \mathcal{D}, x)) - \mathbb{E}_{f_{\min} \sim p(f_{\min} \mid \mathcal{D})} [H(p(f \mid \mathcal{D}, x, f_{\min}))] \\ &\simeq \frac{1}{S} \sum_{i=1}^S \left[-\frac{z_i(x)\phi(z_i(x))}{2\Phi(-z_i(x))} - \log \Phi(-z_i(x)) \right], \end{aligned} \tag{21}$$

where Φ is the cumulative density function, ϕ is the density function of a normal distribution, and $z_i = (f_{\min}^i - \mu(x))/\sigma(x)$. The distribution $p(f \mid \mathcal{D}, x, f_{\min})$ is a truncated Gaussian with support $(f_{\min}, +\infty)$. The acquisition function reveals areas where there is significant variance falling below f_{\min} . Acquiring S samples $f_{\min}^i \sim p(f_{\min} \mid \mathcal{D})$ is challenging because the distribution $p(f_{\min} \mid \mathcal{D})$ is unknown a priori. To this end, (Wang and Jegelka, 2017) propose two different approaches to obtain samples from $p(f_{\min} \mid \mathcal{D})$, which involve numerical approximations that will be briefly explained in the following, while explaining mESCO, for convenience.

Extending mES to account for M unknown constraints requires implementing some modifications to the original algorithm, which we detail next.

5.3.2 mESCO

When addressing the constrained optimization problem (19), mES cannot be used directly because: (i) it gathers samples from the unconstrained minimum f_{\min}^i , while we are interested in the samples of the constrained minimum $\tilde{f}_{\min}^i \sim p(\tilde{f}_{\min}^i|\mathcal{D})$, and (ii) using directly samples \tilde{f}_{\min}^i on (21) makes the acquisition function have extremely large peaks when the sample value \tilde{f}_{\min}^i is near the values of the observations. For these reasons, we modify the original α_{mES} strategy by applying the following steps:

1. Since the true constraint thresholds c_j are unknown, we update the current approximates \hat{c}_j , given by the GPCR model of each constraint g_j , as indicated in Sec. 3.4.2.
2. In (Wang and Jegelka, 2017), it is proposed to either approximate $p(f_{\min}|\mathcal{D})$ using a Gumbel distribution, which involves discretizing the input domain \mathcal{X} , or approximate random function realizations of the GP posterior and minimize them using local optimization. The approximation in the latter proposition comes from applying Bochner’s theorem, which provides a callable object, but as a counterpart, it creates undesired harmonics in the approximated realization that can mislead the search. Additionally, it requires computing the spectral density of the kernel, which might not always be available (see (Hernández-Lobato et al., 2014; Wang and Jegelka, 2017) for details).

While (Wang and Jegelka, 2017) sample from $p(f_{\min}|\mathcal{D})$, we need to sample from the distribution over the constrained minimum $p(\tilde{f}_{\min}|\mathcal{D})$. For this, we follow a method that does not require numerical approximations. Herein, we describe how to obtain one sample \tilde{f}_{\min}^i by running local optimization with random restarts on the problem (18) using the approximated thresholds \hat{c}_j , and virtual evaluations. Every time the local optimizer requests an observation of f or g_j at a location x , we

- a) sample such observation from the predictive distribution of the corresponding GP, i.e., $f_r \sim p(f|\mathcal{D}, x)$, which moments are given in (14),
- b) contaminate it with Gaussian noise, $y_r = f_r + \varepsilon$, $\varepsilon \sim \mathcal{N}(\varepsilon; 0, \sigma^2)$, and
- c) include the sampled point in a virtual dataset $\mathcal{I}_{r+1} \leftarrow \mathcal{I}_r \cup \{x, y_r\}$, which is initialized with the current set of N evaluations, i.e., $\mathcal{I}_0 \leftarrow \mathcal{D}$ and expanded as virtual evaluations are collected.

This operation is repeated for every new point the local optimizer request in each one of the functions f , g_j . Collecting virtual evaluations in this way is equivalent to collecting them from a pre-sampled realization of the GP, with the advantage that it does not need numerical approximations. The caveat of this approach is that the prior kernel matrix $\tilde{\Sigma}$ from (9) needs to be augmented and inverted every time a new virtual evaluation is added, which has a maximum cost of $O(\sum_{r=1}^R (N+r)^3)$, where R is the maximum number of evaluations we allow per random restart. To alleviate this problem, we expand the inverse of the prior covariance using the Woodbury identity (Rasmussen and Williams, 2006, A.3), which has maximum cost of $O(RN^2)$.

Once the local optimization has finished, we return the constrained optimum \tilde{f}_{\min}^i as a sample. For each random restart, we reset $\tilde{\Sigma} \leftarrow \tilde{\Sigma}_{\text{ext}}$ and repeat this operation until S samples are collected.

3. We modify (21) to include the samples from $p(\tilde{f}_{\min}|\mathcal{D})$ by replacing z_i with $\tilde{z}_i = (\tilde{f}_{\min}^i - \mu(x))/\sigma(x)$
4. We drive the acquisition function toward regions that satisfy the constraints with high probability following the strategy (20), where α is to be replaced by α_{mESCO} . As a side effect, the peaks originated by samples \tilde{f}_{\min}^i being near existing evaluations are effectively downsized by the probability of success, which is low in unsafe areas, where the peaks typically occur. This, in practice they only do not affect the decision-making.

The resulting acquisition function is smooth and can be maximized using analytical gradients:

$$\alpha_{\text{mESCO}}(x) = \begin{cases} \alpha_{\text{mES}}(x) \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j], & \text{if } \exists x, \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j] \geq 1 - \delta \\ \prod_{j=1}^M \Pr[g_j(x) \leq \hat{c}_j | \mathcal{D}_j], & \text{otherwise.} \end{cases} \quad (22)$$

A pseudocode for **mESCO**³ is presented in Algorithm 1, and additional implementation details are given in Appendix C. An intuitive idea of the decision-making of **mESCO** is visually illustrated in the one-dimensional synthetic examples (Fig. 3c and 3d). A more thorough analysis in different benchmarks and dimensionality is provided in Sec. 7.

It is worth remarking that the proposed algorithm is not tied to the use of GPCR. For example, if the constraints are not level-set constraints (i.e., unknown thresholds), but standard constraints (i.e., known thresholds), then all the GPCR models can be replaced by standard GPs, and Algorithm 1 can equally be used by simply skipping the update steps from lines 3 and 4, and using the true constraint thresholds instead.

6. Related work and contributions

There has been recent interest in Bayesian optimization (BO) (Shahriari et al., 2016) for learning robot controllers (Berkenkamp et al., 2016; Calandra et al., 2016; Marco et al., 2016; Rai et al., 2017; von Rohr et al., 2018). While in these works, unstable evaluations have been included in the data set during learning in form of a high penalty, in this paper this issue is fundamentally addressed using GPCR. In the following, we show close connections between the core ideas of this work with other methodologies similar or derivate from BO.

A similar methodology to BO is active learning, where the next controller is selected using an acquisition criterion. For example, in (Schreiter et al., 2015) the optimization problem (18) is approached using safe exploration. More specifically, a negative label is assigned to failure cases, and a positive label is assigned to success cases. A Gaussian process classifier (GPC) is used to learn a probabilistic binary constraint that allows for making probabilistic statements about whether a region is likely to be stable or unstable. While this method appears to be simple, it involves carrying two models: One Gaussian process regression model (GPR) for the objective, and one GPC to model the constraint. The authors show that the input space is safely explored, while the unknown constraint g is learned alongside. Observations of such constraint are modeled with a heterogeneous likelihood that perceives (i) continuous valued observations, near the stability boundary, and (ii) discrete labels (unstable/stable), far from it. While this approach shares similarities with the one

3. Our python implementation of both, GPCR and **mESCO** will be made available upon publication.

presented herein, we propose a Bayesian model that explicitly bypasses the need of an extra GPC model by including the discrete labels directly in the GPR, as detailed in in Sec. 2.3. Such model is mathematically different from that of (Schreiter et al., 2015) and carries less hyperparameters to estimate or choose, while it retains the same flexibility. Finally, the authors (Schreiter et al., 2015) assume that continuous valued observations of g are only possible near the transition boundary, while we do not need that assumption.

Another closely related work (Gotovos et al., 2013) in the context of active learning proposes a method that solves a classification problem for function values to lie above or below a specific threshold. Such problem is posed as a level set estimation problem, where the threshold is either explicitly given by an expert, or implicitly defined within a confidence interval. In the latter case, the authors give convergence guarantees for such interval to decrease below a desired accuracy after a number of evaluations. While our problem formulation shares similarities with this one, there are fundamental differences: (i) Their target is to classify the input space in stable vs. unstable areas, while ours is to find the optimum avoiding unstable regions, revealed during exploration; (ii) in their formulation, the threshold does not impose a critical constraint and thus, observations of the objective can be obtained above it. In our formulation, constraint violation is critical in the sense that no observations of the objective are obtained above the threshold, but just a label indicating that the evaluation was unstable.

In the context of RL for robotics, BO has also been combined with policy search methods to learn feedback policies (Drieß et al., 2017; Englert and Toussaint, 2018). Therein, the unknown constraint is modeled as a GPC, which is then used to compute a safe boundary region in the input domain. They include the information of the boundary directly into the Bayesian optimization criterion, which then trades off exploration, near the current boundary estimate, with exploitation, in a region inside the boundary estimate. While such method is relevant in the context of safe learning, we observe two main aspects: (i) it needs a stable policy as a starting point, and (ii) it expands the safe region around this initial controller, leaving unexplored potential better optima that could exist outside of it. Instead, our method does not assume a single safe region that gets expanded, but can identify multiple safe regions for which the probability of constraint satisfaction is high. Additionally, although they propose safe exploration, their method is not exempt of failures when identifying the safety boundary on a real robot. When failures occur, we see them as a benefit, since they are informative about regions that shall not be visited again. The interesting case in which the optimal solution lies close to the boundary can be found by our method, as well.

Bayesian optimization with unknown constraints (BOC) (Gardner et al., 2014; Gelbart et al., 2014; Gramacy and Lee, 2011; Griffiths and Hernández-Lobato, 2017; Hernández-Lobato et al., 2016; Picheny, 2014; Schonlau et al., 1998; Schreiter et al., 2015; Snoek, 2013) has shown to be effective in well-known optimization benchmarks (Gelbart, 2015; Gelbart et al., 2014) and in applications such as learning the hyperparameters of a neural network or Markov chain Monte Carlo methods (Hernández-Lobato et al., 2016). The target of the constrained optimization problem (18) is to optimize an unknown objective while satisfying multiple unknown constraints with high probability. When it comes to determine constraint satisfaction, there are two main differences between the aforementioned BOC techniques and our approach: First, while in BOC it is typically assumed that the constraint threshold c_j

is known a priori ⁴, we do not assume so. In our work, we explicitly emphasize a generic unknown threshold c_j , which does not need to be defined a priori, but can be learned from data, either estimated as a model parameter, or by treating it from a Bayesian perspective. This aspect is a core contribution of this work, novel in the context of BOC, and not explored by the aforementioned papers. Second, a prevailing assumption in BOC is that noisy observations of the constraints are always available, even after constraint violation, while in this work, we assume that no continuous-valued observation is obtained upon constraint violation, as stated in Sec. 4. Alternatively, we assign to those measurements a negative label, which is captured by the proposed Bayesian model, and allows to make inference and improve the knowledge about the unknown threshold.

To summarize, in this work we propose a general Bayesian model that can be used to model objective and constraints in BO when unstable evaluations yield no continuous-valued observations. In addition, we extend an existing BO method to account for unknown constraints. In the context of the related work, a detailed list of the contributions is summarized below.

1. **One model for both, classification and regression:** The proposed GPCR model bypasses the need of training an additional GP classifier, as done in (Drief et al., 2017; Englert and Toussaint, 2018; Schreiter et al., 2015; Snoek, 2013), resulting in an equally flexible model, but with a lower number of parameters.
2. **No failure penalties, and no constraint thresholds:** The proposed Bayesian model circumvents the need of defining an arbitrary penalty a priori for failures, as commonly done in RL and BO. Instead, whenever the roll-out becomes unstable, the model pushes the predictive probability mass above the unknown stability threshold c . In that way, failure points and their vicinity have lower chances to be re-visited again during the optimization search. The unknown stability threshold c is a key parameter of the model, and is re-learned from data at each iteration, instead of being heuristically chosen by a designer.

When additional indicators of failure are accessible, they can be included in the problem as *additional* constraints g_j . Such constraints can then be modeled using the same aforementioned model, for which the constraint threshold c_j is unknown, but can also be learned from data.

To the best of our knowledge, GPCR is the first Bayesian model able to estimate the threshold from a hybrid data set that combines discrete labels and observations of the objective, in the context of GP regression and Bayesian optimization.

3. **Min-Value Entropy Search with unknown level-set constraints (mESCO):** Besides the GPCR model, we propose a novel extension of a recent entropy-based Bayesian optimizer: Max-Value Entropy Search (Wang and Jegelka, 2017) to account for unknown constraints: mESCO. Such method is used to validate GPCR, first in several

4. While in the aforementioned papers, the constrained optimization problem is generally formulated as $\min_{x \in \mathcal{X}} f(x)$, s.t. $\tilde{g}_1(x) \geq 0, \dots, \tilde{g}_M(x) \geq 0$, we can show that our formulation (18) is equivalent by defining $\tilde{g}_j(x) = c_j - g_j(x)$. It is important to notice that the constraint $\tilde{g}_j(x) \geq 0$, as presented in BOC, implicitly assumes that the threshold (zero) is known, while we do not assume so.

synthetic benchmarks, and second, using a humanoid robot balancing an inverted pole as experimental platform.

7. Experimental results

In this section, we validate GPCR in the context of black-box constrained optimization. We assess the capability of the model to learn the objective and the constraints when their corresponding thresholds are unknown. At the same time, we test the performance of mESCO in finding the constrained global minimum. To this end, we present three different settings. First, in Sec. 7.1, we illustrate the benefits and caveats of GPCR in finding the global minimum in two simulated optimization problems. Second, in Sec. 7.2, we report on the consistency of GPCR and our implementations of mES and mESCO by benchmarking the same simulated scenarios. Last, in Sec. 7.3, we show the benefit of GPCR and mESCO in finding the controller parameters of a real robot balancing an inverted pole.

7.1 Illustrative examples

In this section, GPCR is demonstrated in two different simulated scenarios in 2D. The first simulation illustrates the usability of GPCR when the objective is self-constrained (cf. Sec. 4.1), and the second simulation shows the case of one unconstrained objective and one level-set constraint (cf. Sec. 4.3).

7.1.1 EXPERIMENT DESIGN CHOICES

For both simulations, the domain of the objective functions and the constraints is scaled to the unit square $\mathcal{X} = [0, 1]^2$. We consider an isometric Matérn 3/2 kernel, and all lengthscales and variances fixed a priori. In both cases, we assume a Gaussian hyperprior $p(c) = \mathcal{N}(c; 0, \sigma_c^2)$ over the instability threshold in Simulation I and over the constraint threshold in Simulation II. Such thresholds are re-estimated after each iteration using MAP, as explained in Sec. 3.4.2. The user confidence level is fixed with $\delta = 0.05$ for all the experiments. The initial point x_0 was randomly sampled from a uniform distribution for all cases. For all evaluations, we consider small evaluation noise, i.e., $\sigma = 0.01$.

7.1.2 SIMULATION I

We consider the 2D objective function

$$f(x_1, x_2) = \cos(10x_1) \cos(5x_2) + \sin(10x_1) + 2, \quad (23)$$

proposed in (Gardner et al., 2014) as a benchmark to test constrained Bayesian optimization. While they consider f , together with an external constraint g , we decide instead to consider that f is self-constrained (cf. Sec. 4.1), with instability threshold $c = 1.5$. This induces a region of unstable controllers $\mathcal{X}_u = \{x : f(x_1, x_2) > c\}$, depicted in Fig. 4a as a shadowed area. We set a wide hyperprior on the threshold, i.e., $\sigma_c = 5$, and run mES for 30 iterations. Fig. 4b and 4c show that mES finds the optimal value after 17 stable and 13 unstable evaluations. The last 12 evaluations are stable and targeted in the area around the global minimum. At the final iteration, the estimated best guess is $x_{\text{bg}} = (0.9468, 0)$, $f(x_{\text{bg}}) = 0.001177$, while the true global minimum is $x_{\text{min}} = (0.942, 0)$, $f(x_{\text{min}}) = 0$. The instability threshold is estimated

to be $\hat{c} = 1.35$. While in (Gardner et al., 2014), an algorithm for constrained Bayesian optimization is proposed, it is not possible to fairly compare our framework with theirs because we define g differently (i.e., we consider constraint and objective identical). However, we can highlight two clear benefits of using GPCR. First, the GPCR model on the objective absorbs the failures, while in (Gardner et al., 2014), any existing black-box constraint needs to be modeled with an additional standard GP. Furthermore, we can search for the global minimum using an unconstrained Bayesian optimization method, such as mES, without the need of using a constrained Bayesian optimizer (like the one proposed in (Gardner et al., 2014) or mESCO).

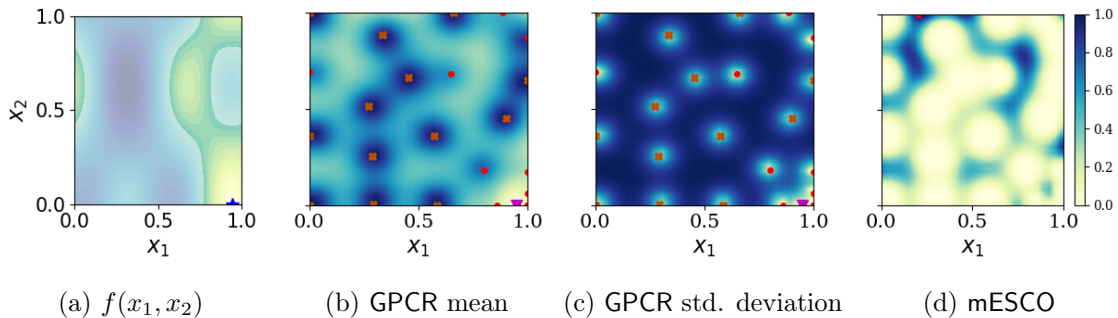


Figure 4: Global search on the objective function (23) using mES. (a) shows the unstable region \mathcal{X}_u (shaded area), and the true feasible global minimum (blue star). (b) and (c) show the predictive mean and standard deviation of GPCR, respectively, stable evaluations (red dots) unstable evaluations (orange crosses), and the best guess estimation of the global minimum x_{bg} (magenta triangle). (d) shows the mES acquisition function, and the next suggestion x_{next} (red dot). Subsequent figures in the paper follow the same legend.

7.1.3 SIMULATION II

We consider the well-known 2D Branin-Hoo benchmark, constrained to a centered circle

$$\begin{aligned}
 f(x_1, x_2) &= (15x_2 - \frac{5.1}{4\pi^2}(15x_1 - 5)^2 + \frac{5}{\pi}(15x_1 - 5) - 6)^2 + 9.6 \cos(15x_1 - 5) + 10)^2 \\
 g(x_1, x_2) &= -\sqrt{2/9 - (x_1 - 0.5)^2 - (x_2 - 0.5)^2} \leq c_1,
 \end{aligned} \tag{24}$$

with $c_1 = 0$, which has also been used in (Gelbart et al., 2014). The constraint violation region is depicted as a shadowed area in Fig. 5e, and the objective is shown in Fig. 5a. The objective has three minima, marked in Fig. 5a and 5e, but the constraint hides two of them, leaving only one as the solution to the constrained optimization problem. The squared root in the constraint function g is deliberately introduced to make it undefined in the outer part of the circle, i.e., if violated, no continuous-valued observation can be obtained, but only a discrete label. This fact characterizes g as a level-set constraint (cf. Sec. 4.3). We model g using GPCR, while f is modeled with a standard GP. We have chosen a wide hyperprior on c_1 , with $\sigma_{c_1} = 2.0$. Fig. 5 shows our proposed constrained Bayesian optimization algorithm, mESCO, after 50 iterations, from which 9 were unstable and 41 were stable. At the final

iteration, the estimated best guess is $x_{\text{bg}} = (0.5448, 0.1527)$, $f(x_{\text{bg}}) = 0.4039$, while the true constrained minimum is $x_{\text{min}} = (0.5428, 0.1517)$, $f(x_{\text{min}}) = 0.40$. The instability threshold is estimated to be $\hat{c}_1 = -0.09842$. These results are comparably better than (Gelbart et al., 2014), where their estimated global minimum is reported at $(0.5340, 0.1573)$, with value 0.48, and known threshold. Furthermore, the advantage of using GPCR on the constraint as compared to a standard GP classifier (Gelbart et al., 2014) is that GPCR can afford the constraint threshold c_1 being unknown a priori, and it can be estimated over iterations.

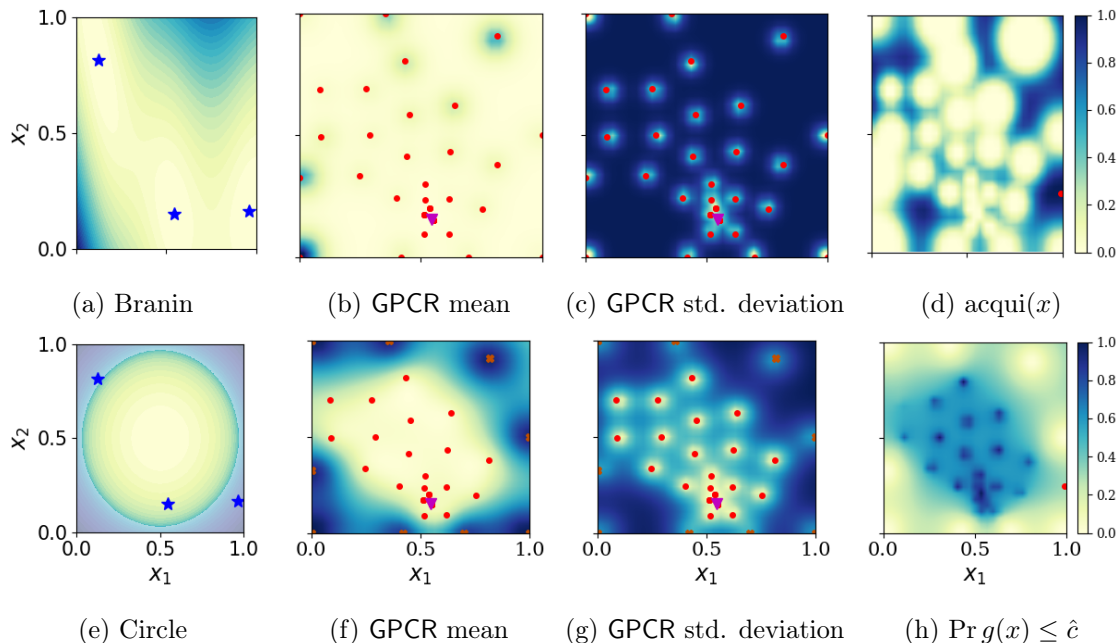


Figure 5: Global search on the objective function (24) using mESCO. (a) and (e) show the objective function f and the constraint g from (24), respectively. In (e), the shadowed area represents constraint violation. (b),(c) show the predictive mean and std. deviation for f , and (f),(g) for g . (d) shows the acquisition function and (h) shows the probability of constraint satisfaction using the current estimate \hat{c}_1 of the true constraint threshold c_1 .

7.2 Statistical comparison

After having shown the general applicability of the proposed framework, we now report on the consistency of GPCR and our implementations of both, mES and mESCO. To this end, we run the experiments from the previous section multiple times and measure the performance of the global optimization methods using the inference regret R_i at each iteration i , defined as $R_i = y(x_{\text{bg},i}) - \min_{x \in \mathcal{X}_s} f(x)$ (Wang and Jegelka, 2017), where $y(x_{\text{bg},i})$ is a function evaluation at the best guess at iteration i .

We benchmark the experiments described in Sec. 7.1 over 20 runs for each case and report the averaged R_i in Fig. 6a and Fig. 6b. We can see the results from benchmarking

Simulation I in the green curve in Fig. 6a, and the results from benchmarking Simulation II in the green curve in Fig. 6b, which correspond to cases 1) and 3) described in Sec. 4.1 and Sec. 4.3, respectively. The remaining cases 2) and 4) are also addressed herein, by reusing the benchmark on Simulation II, but with a few variants, depending on the case.

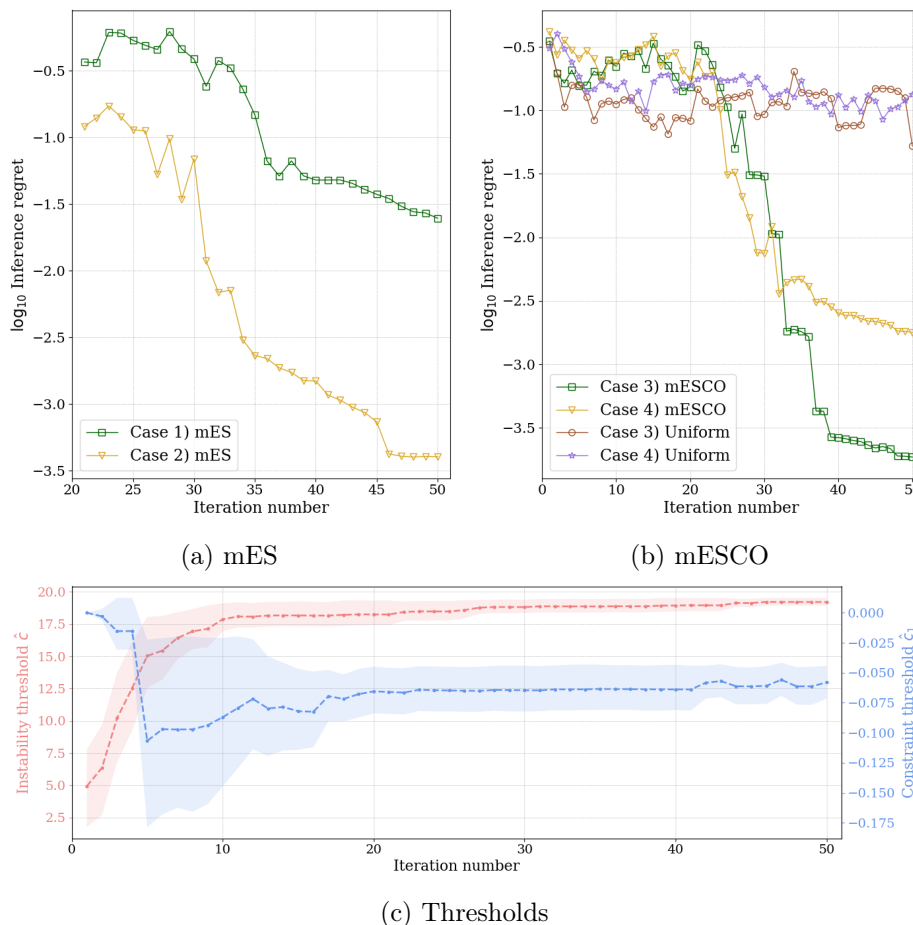


Figure 6: Statistical comparison

To address case 2) (Sec. 4.2), we assume that g is a binary constraint, and that its binary information is absorbed by GPCR, used to model the objective f . To address case 4) (Sec. 4.4), we assume that f is self-constrained, with instability threshold $c = 20$, and g is a level set constraint. Fig. 6a shows the performance of mES in case 2), and Fig. 6b shows the performance of mESCO in case 4). To assess the performance of mESCO, we compare against random search using uniform sampling in Fig. 6b. As shown, mESCO achieves the lowest regret in case 3). We also see that mESCO achieves better minima than the one reported in (Gelbart et al., 2014), on average, i.e., $f(x_{\text{bg}}) = 0.4717 \pm 0.1392$. For all cases, the solution is approximately found with a low error.

In Fig. 6c, we depict the evolution of \hat{c} and \hat{c}_1 over iterations for case 4). While initially the estimation shows a broad transient, it eventually exhibits a convergent behavior, obtaining on average $\hat{c} = 19.2459 \pm 0.6714$ and $\hat{c}_1 = -0.0618 \pm 0.0339$ at the last iteration. This shows

that only a rough approximation to the true threshold is needed for GPCR to work in practice, when using it for BO.

7.3 Robot experiments

The simulation study revealed the possible benefits of GPCR and the effectiveness of mESCO in synthetic scenarios. In this section, we will evaluate its performance on a real robotic platform: The humanoid robot Apollo balancing an inverted pole. This is a much more challenging scenario than the aforementioned simulations because (i) the latent objective and constraint functions are unknown a priori, (ii) collecting evaluations is time expensive, (iii) stopping the experiments due to the controller being unstable or unsafe implies restarting the platform, which is also time consuming, and (iv) the noise is not synthetically generated, but propagated from the sensors to the cost observation we obtain.

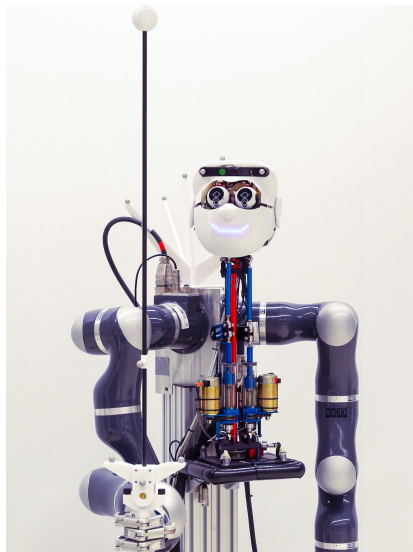


Figure 7: Robot Apollo balancing an inverted pole

Such platform has been used as a testbench for controller learning (Marco et al., 2016). The experiments presented in this section build on top of those from (Marco et al., 2016), in which the parameters of a Linear Quadratic Regulator (LQR) controller were automatically tuned in order to improve the balancing performance. Therein, an important caveat of the learning setup was committing to a fixed heuristic cost f_{heur} , prior to starting the learning experiments. Such fixed f_{heur} is a penalty value chosen *ad hoc* for unstable controllers. As detailed in (Marco et al., 2016), during the learning run, a stable controller (with poor performance) that yielded a higher cost than f_{heur} was found. Because a stable controller should not be penalized more than an unstable one, we stopped the learning run, selected a higher f_{heur} , and restarted the learning from the beginning. Since the choice of the threshold f_{heur} affects the decision-making of the Bayesian optimizer used therein, the data acquired with the older f_{heur} had to be discarded for fairness. The clear practical advantage of GPCR is that it can overcome such situation by adapting \hat{c} to the worst observed stable cost, maintaining the probability mass of unstable areas always above it, and avoiding the need of restarting the learning run and discarding the data.

Additionally, in (Marco et al., 2016) it is reported that there were two main situations that lead to instability, and, in consequence, to premature experiment detention: (i) the endeffector position leaving a safety region and (ii) the endeffector acceleration reaching a maximum value. In the context of this work, those restrictions can be seen as two external safety constraints that determine the shape of \mathcal{X}_s .

In this section, we show the advantages of using GPCR to take into account such constraints, and bypass the need of defining a heuristic cost. For this, we show three different tuning scenarios: (i) a 2D tuning problem with two binary constraints, (ii) idem, with one binary constraint, and one level-set constraint, and (iii) a 5D tuning problem with two binary constraints.

7.3.1 CONTROLLER TUNING WITH UNKNOWN CONSTRAINTS

In (Marco et al., 2016), the design parameters of the well known Linear Quadratic Regulator (LQR) (Anderson and Moore, 1990, Sec. 2.4) are learned from data. Therein, the performance of a balancing experiment over a finite time horizon T is quantified with a quadratic cost $f(x) = \frac{1}{T} \left[\sum_{t=0}^{T-1} s_t^\top Q s_t + u_t^\top R u_t \right]$, where the state and control input trajectories are collected during an balancing experiment with control parameters x . Such parameters enter the diagonal of the control design weights $W_s(x)$ and $W_u(x)$, used to compute the feedback gain matrix $F(x)$ that characterizes the control strategy $u_t = F(x)s_t$. This LQR tuning problem is a particular case of Sec. 2.1, which holds in general for a wider range of control strategies. The GPCR model is agnostic to the control strategy used, and thus it is also applicable to the LQR tuning. For a detailed explanation of automatic LQR tuning, we refer the reader to (Marco et al., 2016).

Contrary to (Marco et al., 2016), we consider herein a constrained tuning problem. The first constraint $g_1(x) = \max \{s_t^{\text{endeff}}\}_{t=0}^T$ is the maximum distance reached by the endeffector in one experiment, relative to the starting position, with constraint satisfaction determined as $g_1(x) < c_1$, where $c_1 = 13$ cm. The second constraint $g_2(x) = \max \{\ddot{s}_t^{\text{endeff}}\}_{t=0}^T$ is the maximum acceleration reached by the endeffector position s_t^{endeff} , with constraint satisfaction determined as $g_2(x) < c_2$, where $c_2 = 3 \text{ m s}^{-2}$. Although the thresholds c_1 and c_2 are known to us, they are unknown to the model, which refines the estimates \hat{c}_1, \hat{c}_2 over iterations. Generally speaking, such safety thresholds are commonly deployed in robotic setups, either by the manufacturer, or by the end user, to prevent the robot from damage. While in this case we know c_1 and c_2 , it is often the case that the manufacturer does not reveal such information to the end user, which motivates the need of using GPCR to learn such unknown thresholds.

7.3.2 IMPLEMENTATION CHOICES

We characterize the objective cost with $Q = \text{diag}(1, 6, 1, 1)$ and $R = 0.25$. In the 2D learning experiments, the design weights are normalized within the domain $\mathcal{X} = [0, 1]^D$ and parametrized as $W_s(x) = \text{diag}(1, 10^{4x_2}, 1, 1)$, $W_u(x) = 10^{4x_2-2}$, and in the 5D experiments as $W_s(x) = \text{diag}(10^{4x_1}, 10^{4x_2}, 10^{4x_3}, 10^{4x_4})$, $W_u(x) = 10^{4x_5-2}$, with $x_i \in [0, 1]$. In all experiments, the first point is chosen randomly in the input space. The number of BO iterations is fixed a priori. The user confidence is determined by $\delta = 0.05$. In all cases we use an isometric 3/2 Matérn kernel (Rasmussen and Williams, 2006), with fixed prior variance $\nu = 25$. The lengthscales are fixed to $\lambda = 0.1$ in the 2D experiments and to $\lambda = 0.2$ in the 5D experiments. The noise for the objective is fixed to $\sigma^2 = 0.03^2$ for f and to $\sigma^2 = 0.01^2$ for g_1 . After completing a learning run, we test the performance of the best guess for the global minimum x_{bg} on the real robot by averaging over 5 experiments, and obtaining μ_{bg} and σ_{bg}^2 .

7.3.3 2D CONTROLLER TUNING

In the first learning run, we assume g_1 and g_2 are binary constraints, and, thus, absorbed by the GPCR model of f . The hyperprior over the instability threshold c is $p(c) = \mathcal{N}(c; 0, 10^2)$. We run our implementation of mES for 30 iterations, and report the results in Fig. 8. As can be seen, the best safe area appears to be found around the center of the domain. The best guess for the global minimum is found at $x_{\text{bg}} = (0.45, 0.61)$. The performance of the global minimum is $\mu_{\text{bg}} = 0.75$, $\sigma_{\text{bg}}^2 = 0.058^2$. We obtained 17 unstable evaluations and 13 stable evaluations.

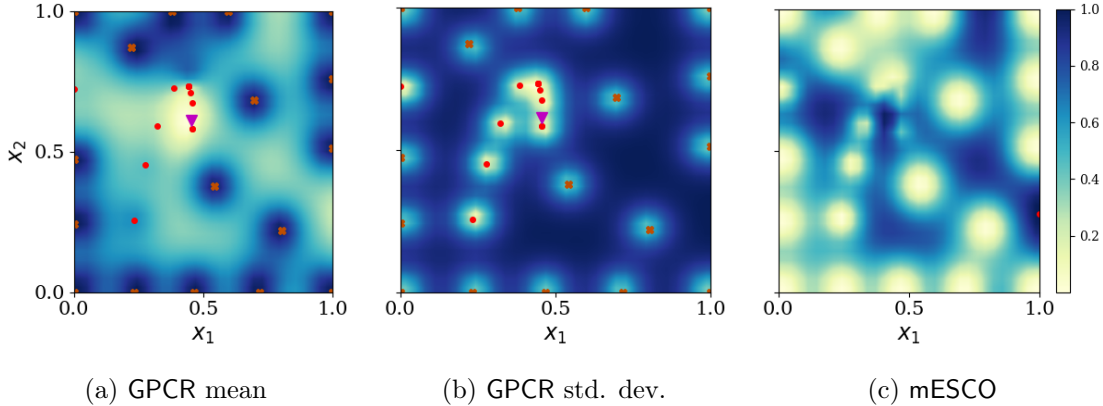


Figure 8: Results of the 2D controller tuning after 30 iterations using mES and assuming binary constraints. (a) and (b) show the predictive mean and standard deviation of GPCR, stable evaluations (red dots), unstable evaluations (orange crosses), and the best guess estimation of the global minimum x_{bg} (magenta triangle).

In the second learning run, we assume a mixed case, in which g_1 is a level set-constraint and g_2 is a binary constraint. While g_2 is absorbed by the GPCR model for f , g_1 is modeled with another GPCR, as shown in Fig. 9e and 9f. The corresponding hyperpriors are assumed identical for both c and c_1 , with $p(c) = \mathcal{N}(c; 0, 10^2)$. Of special interest it is that the constraint (Fig. 9e) shows unsafe evaluations in the top part of the domain, while that part appears to be stable for the objective (Fig. 9a). Interestingly, in the bottom part of the domain, it happens the other way around. As a result, only the area in the center of the domain yields safe and stable controllers, which is coherent with the safe area found in the previous experiment (Fig. 8). mESCO outputs a global minimum at location $x_{\text{bg}} = (0.64, 0.65)$, with $\mu_{\text{bg}} = 0.41$, $\sigma_{\text{bg}}^2 = 0.063^2$. Evaluations to the objective yielded 23 stable and 7 unstable controllers. Evaluations on the constraint yielded (by coincidence) the same numbers. The threshold of the constraint is estimated at $\hat{c}_1 = 12.75$ cm, while the true threshold is $c_1 = 13$ cm.

7.3.4 5D CONTROLLER TUNING

Herein, we assume again that the two binary constraints are absorbed in the GPCR model for f . The best guess for the global minimum is found at $x_{\text{bg}} = (0.51, 1.00, 0.52, 0.47, 1.00)$,

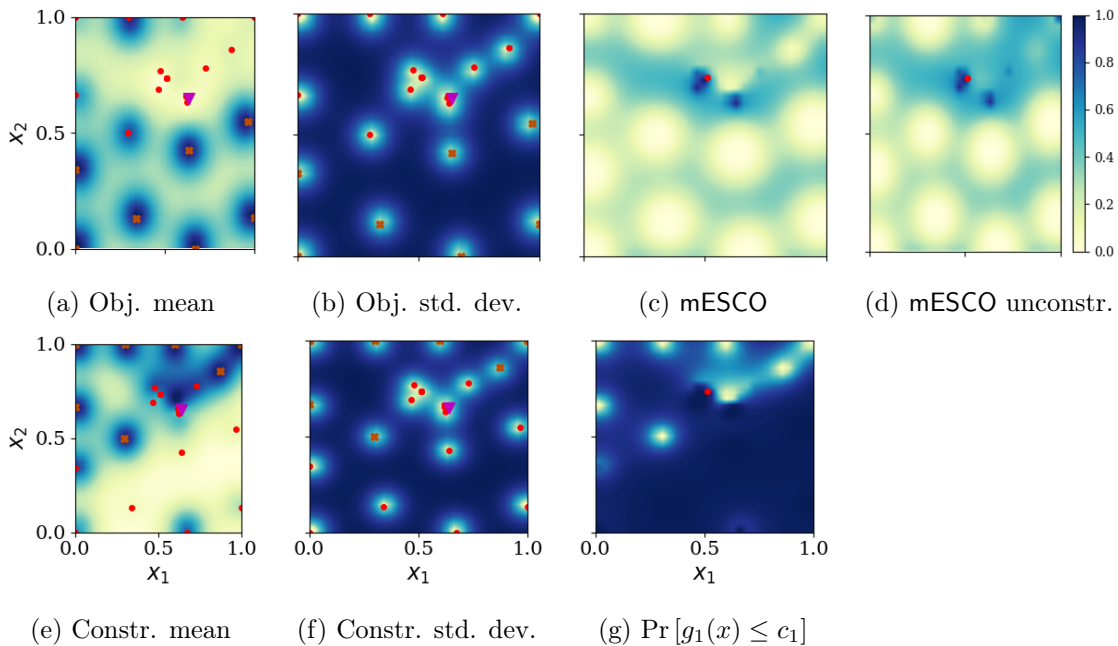


Figure 9: Results of the 2D controller tuning after 30 iterations with a level set constraint (g_1) and a binary constraint (g_2), using **mESCO**. (a) and (b) show mean and standard deviation of the objective function and (e) and (f) show the same for the constraint g_1 . In all 4 plots, stable evaluations (red dots), unstable evaluations (orange crosses), and the best guess estimation of the global minimum x_{bg} (magenta triangle) are highlighted. (c) and (d) show the acquisition function of constrained and unconstrained **mESCO**, respectively, and x_{next} (red dot). (g) shows the probability of constraint satisfaction using the current estimate of the constraint threshold.

with $\mu_{\text{bg}} = 0.34$, $\sigma_{\text{bg}}^2 = 0.045^2$. As shown in Fig. 10, we obtained 65 unstable evaluations and 15 stable evaluations. The performance of the best guess is better than the one of the best guess found in the other 2D cases. At iteration 52, the first stable evaluation took place, with a cost $y_{52} = 0.569$, which changed the value of the estimated threshold to $\hat{c} = 0.602$. At iteration 69, a marginally stable controller yielded $y_{69} = 51.64$. However, the GPCR model handled this situation gracefully with a new threshold estimate $\hat{c} = 51.52$. With this, the probability of unstable regions is pushed above the threshold, which resolves the need of resetting the learning experiments explained in Sec. 7.3, and also the major caveat observed in the approach from (Marco et al., 2016).

7.4 Discussion

We have reported results on the usage of GPCR and **mESCO** in different 2D synthetic benchmarks, and also in 2D and 5D experiments on a real robotic scenario. In summary, the benefits of GPCR are (i) its flexibility to learn online the instability threshold, which

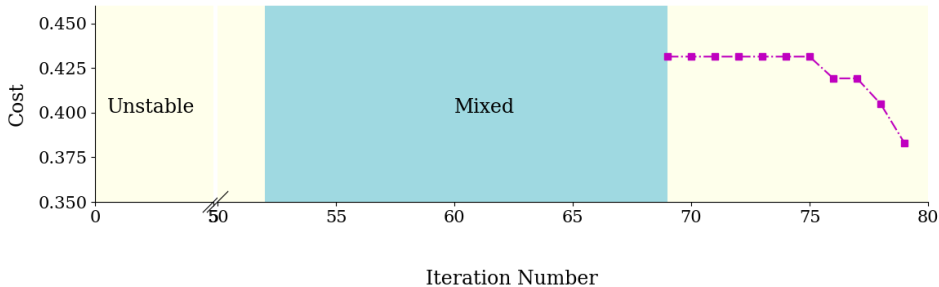


Figure 10: Results of the 5D controller tuning

bypasses the need of defining a priori a heuristic penalty, and also avoids the risk of having to restart the learning experiments, (ii) its adaptability to learn online the constraint threshold whenever this is not available, and (iii) its capability of absorbing binary constraints in the model for the objective, which reduces the number of needed models, and thus, the number of hyperparameters. In addition to this, our implementation of **mESCO** has shown successful results in both, synthetic benchmarks and a real robot application.

In the following, we discuss about some minor issues found during the experiments, as well as implementation choices.

Instead of learning the hyperparameters, we fix them to conservative values in order to ensure a wide coverage. The reason is that for such tuning scenarios, where gathering data is prohibitive, there is a high risk of overfitting to the data observed during an early stage of BO. This would yield long lengthscales and small variance, which would make the acquisition function completely myopic to other potentially interesting regions, targeting the current area as the only interesting one. On the other hand, fixing the hyperparameters yields a slow search, but the optimizer has a wider coverage over the domain, which ensures that interesting regions will not be missed.

In the 2D examples the global minimum is estimated at two nearby, but different locations. On one hand, this indicates that a higher number of evaluations would have been needed to find an agreement between the two results. Nonetheless, in both cases the Bayesian optimizer targets the area where stable controllers are present. Additionally, both **mES** and **mESCO** are shown to consistently find the global minimum in the numerical benchmarks from Sec. 7.2.

In order to speed up the learning experiments on the robot, we have implemented a “back-to-safety” routine: When a new controller is detected as unstable, we automatically switch back to a *safe controller*, which is a priori known to yield a stable controller with very poor performance. Such controller does not have any impact in the learning experiments, as it is never included as a data point in the exploration carried out by the Bayesian optimizer. It is just used to mitigate the interaction time with the robot. Even then, some controllers were too aggressive for the stable balancing to be recovered, and the robot had to be prematurely stopped by triggering safety mechanisms in place.

In practice, we found that a wide hyperprior $p(c) = \mathcal{N}(c; 0, \sigma_c^2)$ works well. However, it frequently occurs that during the initial iteration of **mES** and **mESCO**, only unstable

evaluations are collected. This pushes the \hat{c} towards negative values, which can cause overall numerical instability issues when computing the approximate posterior (10). To prevent this, we assign $\hat{c} = 0$ in the absence of stable controllers. With this choice, and small δ , the probabilistic constraint will very unlikely be satisfied, which will force an exploratory strategy until the constraint is satisfied somewhere (cf. Sec. 5.1).

In some situations, a stable evaluation is found at a very close to a stable one. In order to avoid numerical instability, we relax the evaluation noise of the stable point, so that the model can better adapt to such situation. For all the experiments presented in this section, we never encountered such issue, except for the random exploration shown in Sec. 7.2.

8. Conclusions

We have proposed a novel non-parametric Bayesian model that solves a fundamental problem in robot learning: How to choose penalty values for unstable controllers. While normally such values are chosen *ad hoc* after a pre-training phase of data collection, the proposed model infers them directly during learning. This alleviates the usual need of the aforementioned pre-training phase, normally used to tailor appropriate cost functions.

To this end, the Bayesian model handles two types of data: Discrete labels that indicate whether controllers are stable or unstable, and continuous-valued observations, only present when controllers are stable. The Bayesian model is approximated as a GP, which classifies unsafe regions by pushing the predictive probability mass toward high costs. Because of the tight connection existing between classification and regression, we call this model *Gaussian process for classified regression* (GPCR).

We have also shown how this model helps to solve black-box constrained optimization problems using Bayesian optimization. Therein, unsafe regions are defined by those areas where the controller is unstable, together with those others where external constraints are violated. Our experiments show that GPCR can also be used to model black-box constraints and learns its threshold online, in case it is unknown.

In addition, we have shown the benefits of GPCR on a robotic platform, where modeling the threshold and the penalty terms for the objective would otherwise be cumbersome, and would involve collecting extra experiments during a pre-training phase.

For conducting the experiments, we have extended Max-Value Entropy Search (Wang and Jegelka, 2017), a recent information-theoretic Bayesian optimization method, to account for unknown black-box constraints, and have tested its effectiveness in simulations and real hardware.

In summary, GPCR is beneficial whenever designing penalty terms for the reward/cost function is problematic or requires a pre-training phase, for example learning to grasp objects with a robot arm. In addition, we have shown that GPCR is useful when the constraint threshold is unknown, for instance learning to pick-and-place with a robot arm under the presence of obstacles. Finally, GPCR is a powerful tool that can be used to learn robot controllers mitigating human intervention in the learning loop.

Acknowledgement

We thank Manuel Wüthrich for his feedback and for the interesting discussions. We thank Felix Grimminger for the design of the pole and for his support on the hardware. We thank Vincent Berenz for his support with the software infrastructure and Julian Viereck for a key contribution to the robot code. This work was supported in part by the Max Planck Society, the Cyber Valley Initiative, the IMPRS-IS, and the German Research Foundation (DFG, grant TR 1433/1-1). Philipp Hennig gratefully acknowledges support through ERC StG Action 757275 / PANAMA.

Appendix A. Product of two multivariate Gaussians of different dimensionality

The product p of two multivariate Gaussians of dimensions N_1 and N_2 , with $N_2 > N_1$ yields another Gaussian with dimensions N_2 . We prove this for the case in which the second Gaussian is zero-mean. To prove it, we resort to the completion of squares in the vectorial case.

$$p = \mathcal{N}(\mathbf{y}_s | \mathbf{f}_s, \sigma^2 \mathbf{I}) \mathcal{N} \left(\begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} K_{ss} & K_{su} \\ K_{us} & K_{uu} \end{bmatrix} \right) = a_0 \exp \left\{ -\frac{1}{2} a_1 \right\}, \quad (25)$$

where $a_0 = [(2\pi)^{N_s} (2\pi)^{N_s+N_u} \det(\sigma^2 \mathbf{I}) \det(K)]^{-1/2}$ and

$$a_1 = [\mathbf{y}_s - \mathbf{f}_s]^\top \sigma^{-2} \mathbf{I} [\mathbf{y}_s - \mathbf{f}_s] + \begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix}^\top \underbrace{\begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}}_{K^{-1}} \begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix}.$$

We can reorder the terms to obtain a quadratic expresion

$$\begin{aligned} a_1 &= \mathbf{f}_s^\top (\mathbf{V}_{11} + \sigma^{-2}) \mathbf{f}_s - 2\sigma^2 \mathbf{y}_s^\top \mathbf{f}_s + \mathbf{f}_s^\top \mathbf{V}_{12} \mathbf{f}_u + \mathbf{f}_u^\top \mathbf{V}_{21} \mathbf{f}_s + \mathbf{f}_u^\top \mathbf{V}_{22} \mathbf{f}_u + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s \\ &= \begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix}^\top \underbrace{\begin{bmatrix} \mathbf{V}_{11} + \sigma^{-2} \mathbf{I} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}}_C \underbrace{\begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix}}_f + \underbrace{[-2\sigma^{-2} \mathbf{y}_s^\top \quad \mathbf{0}^\top]}_{\mathbf{b}^\top} \begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \end{bmatrix} + \underbrace{\sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s}_d \\ &= \mathbf{f}^\top C \mathbf{f} + \mathbf{b}^\top \mathbf{f} + d, \end{aligned} \quad (26)$$

which can be more conveniently written using the square completion in matrix form⁵

$$a_1 = [\mathbf{f} - C^{-1} \mathbf{b}]^\top C [\mathbf{f} - C^{-1} \mathbf{b}] + \frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s. \quad (27)$$

5. Square completion in matrix form (Peterson et al., 2008): $\mathbf{f}^\top C \mathbf{f} + \mathbf{b}^\top \mathbf{f} = (\mathbf{f} - C^{-1} \mathbf{b})^\top C (\mathbf{f} - C^{-1} \mathbf{b}) + \frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b}$.

Using this result in (25), and defining mean $\tilde{\mathbf{m}} = C^{-1}\mathbf{b}$ and variance $\tilde{\Sigma} = C^{-1}$, we can rewrite the product of two Gaussians as unnormalized Gaussian

$$\begin{aligned} p &= a \cdot \exp \left\{ -\frac{1}{2} [\mathbf{f} - C^{-1}\mathbf{b}]^\top C [\mathbf{f} - C^{-1}\mathbf{b}] \right\} \cdot \exp \left\{ -\frac{1}{2} \left(\frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s \right) \right\} \\ &= a \left[(2\pi)^{N_s + N_u} \det(\tilde{\Sigma}) \right]^{1/2} \mathcal{N}(\mathbf{f}; \tilde{\mathbf{m}}, \tilde{\Sigma}) \cdot \exp \left\{ -\frac{1}{2} \left(\frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s \right) \right\} \\ &= \alpha \mathcal{N}(\mathbf{f}; \tilde{\mathbf{m}}, \tilde{\Sigma}), \end{aligned}$$

where the constant α can be simplified as

$$\begin{aligned} \alpha &= a \left[(2\pi)^{N_s + N_u} \det(\tilde{\Sigma}) \right]^{1/2} \exp \left\{ -\frac{1}{2} \left(\frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s \right) \right\} \\ &= \left[(2\pi)^{N_s} \det(\tilde{\Sigma}^{-1}) \det(K) \det(\sigma^2 \mathbf{I}) \right]^{-1/2} \exp \left\{ -\frac{1}{2} \left(\frac{1}{4} \mathbf{b}^\top C^{-1} \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s \right) \right\} \\ &= \left[e^v (2\pi\sigma^2)^{N_s} \det(\tilde{\Sigma}^{-1}) \det(K) \right]^{-1/2}, \end{aligned} \tag{28}$$

with $v = \frac{1}{4} \mathbf{b}^\top K \mathbf{b} + \sigma^{-2} \mathbf{y}_s^\top \mathbf{y}_s$, and $\tilde{\Sigma}^{-1} = K^{-1} + \begin{bmatrix} \sigma^{-2} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$.

Appendix B. Analysis of the predictive distribution $p(f_* | \mathcal{D}, X, x_*)$

The integral in the right-hand side of (12) can yield different results, depending on the assumptions we make. In Sec. 3.3, we have followed (Rasmussen and Williams, 2006, Sec. 3.4.2) and approximated the posterior as Gaussian, which makes it possible to approximate the integral from (12) also as a Gaussian distribution $q(f_*)$, with moments (14). Whereas such approximation yields a tractable GPCR model, and works well in practice, it is also interesting to analyze how close this approximation is to the exact distribution $p(f_* | \mathcal{D}, X, x_*)$. For this, one would need to have the exact solution to $p(f_* | \mathcal{D}, X, x_*)$, which is not possible. Instead, we compare $q(f_*)$ with a better approximation $p(f_*)$ to $p(f_* | \mathcal{D}, X, x_*)$, obtained by relaxing some of the assumptions made in Sec. 3.3, and solving the integral from (12) in a different way. The resulting distribution $p(f_*)$ is closer to the true distribution $p(f_* | \mathcal{D}, X, x_*)$ than $q(f_*)$, but also not exact. In addition to this, it is computationally more demanding, which makes it impractical for Bayesian optimization. To simplify notation, we refer to the predictive probability $p(f_* | \mathcal{D}, X, x_*)$ (12) as $p(f_*)$.

We depart from $p(f_*)$, and use the extended prior (13) and the likelihood terms presented in (8)

$$\begin{aligned} p(f_*) &\propto \\ &\underbrace{\int_c^{+\infty} \cdots \int_c^{+\infty}}_{\text{Unstable}} \underbrace{\int_{-\infty}^c \cdots \int_{-\infty}^c}_{\text{Stable}} \mathcal{N}(\mathbf{y}_s | \mathbf{f}_s, \sigma^2 \mathbf{I}) \mathcal{N} \left(\begin{bmatrix} \mathbf{f}_s \\ \mathbf{f}_u \\ \mathbf{f}_* \end{bmatrix}; \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} K_{ss} & K_{su} & K_{s*} \\ K_{us} & K_{uu} & K_{u*} \\ K_{*s} & K_{*u} & K_{**} \end{bmatrix} \right) d\mathbf{f}_s d\mathbf{f}_u, \end{aligned} \tag{29}$$

where the Heaviside functions from (8) have been absorbed in the integration limits. The product of two Gaussian densities gives rise to another Gaussian density, as shown in Appendix A. Applying such transformation, (29) can be rewritten as

$$p(f_*) \propto \underbrace{\int_c^{+\infty} \cdots \int_c^{+\infty}}_{\text{Unstable}} \underbrace{\int_{-\infty}^c \cdots \int_{-\infty}^c}_{\text{Stable}} \mathcal{N} \left(\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix}; \hat{\mathbf{m}}, \hat{\Sigma} \right) d\mathbf{f}, \quad (30)$$

where $\hat{\mathbf{m}} = \hat{\mathbf{m}}(\mathbf{y}_s, \sigma^2)$ and $\hat{\Sigma} = \hat{\Sigma}(\sigma^2)$ are functions of the continuous-valued observations \mathbf{y}_s and the evaluation noise σ^2 . The proportionality constant obtained from the applied transformation has been omitted for simplicity. In order to numerically solve the integral from (30), we first apply the rule of conditional probability to the Gaussian, i.e., we want the conditional on \mathbf{f} after revealing f_*

$$\mathcal{N} \left(\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix}; \underbrace{\begin{bmatrix} \hat{\mathbf{m}}_{\mathbf{f}} \\ \hat{m}_* \end{bmatrix}}_{\hat{\mathbf{m}}}, \underbrace{\begin{bmatrix} \hat{\Sigma}_{\mathbf{f}\mathbf{f}} & \hat{\Sigma}_{\mathbf{f}f_*} \\ \hat{\Sigma}_{*f} & \hat{\Sigma}_{**} \end{bmatrix}}_{\hat{\Sigma}} \right) = \mathcal{N}(\mathbf{f}; \mathbf{a}(f_*), \mathbf{B}) \mathcal{N}(f_*; \hat{m}_*, \hat{\Sigma}_{**}), \quad (31)$$

where $\mathbf{a}(f_*) = \hat{\mathbf{m}}_{\mathbf{f}} + \hat{\Sigma}_{\mathbf{f}f_*} \hat{\Sigma}_{**}^{-1} (f_* - \hat{m}_*)$ is a linear mapping on f_* , and $\mathbf{B} = \hat{\Sigma}_{\mathbf{f}\mathbf{f}} - \hat{\Sigma}_{\mathbf{f}f_*} \hat{\Sigma}_{**}^{-1} \hat{\Sigma}_{*f}$. Then, we rewrite (30) using (31) as

$$p(f_*) \propto \mathcal{N}(f_*; \hat{m}_*, \hat{\Sigma}_{**}) \underbrace{\int_c^{+\infty} \cdots \int_c^{+\infty}}_{\text{Unstable}} \underbrace{\int_{-\infty}^c \cdots \int_{-\infty}^c}_{\text{Stable}} \mathcal{N}(\mathbf{f}; \mathbf{a}(f_*), \mathbf{B}) d\mathbf{f} \quad (32)$$

$$= \mathcal{N}(f_*; \hat{m}_*, \hat{\Sigma}_{**}) Z_{\text{EP}}(c, f_*), \quad (33)$$

where we have encapsulated the multivariate integral (32) in a functional $Z_{\text{EP}}(c, f_*)$, similarly as done for (16). We resort also here to the method from the authors (Cunningham et al., 2011), only that EP needs to be called for each pair of values f_* and c . Of course, this involves discretizing f_* for each test point x_* , which is computationally inefficient. We illustrate $p(f_*)$ for the simple one-dimensional Example 2 in Fig. 11. Therein, f_* has been uniformly discretized on a fixed interval $[-1.0, +3.5]$ with 200 divisions, and x_* has also been uniformly discretized in the unit interval with 200 divisions. We use an estimated value for the threshold $\hat{c} = 2.03$, computed by maximizing the marginal likelihood (15).

Fig. 11 shows that $p(f_*)$ pushes 95% of the probability mass above the threshold c , near the unstable evaluations, while it stays below the threshold, near the stable ones. In areas close to the stable evaluations, $p(f_*)$ overlaps almost perfectly with $q(f_*)$. The main differences between them arise in areas close to unstable evaluations. First, we note that $p(f_*)$ becomes an asymmetric distribution in regions close to unstable evaluation points, while it remains symmetric in regions near the stable evaluations. Second, it can be seen that $q(f_*)$ incorrectly puts some non-trivial probability mass below the threshold. When modeling a constraint g_j (cf. Sec. 4.3 and Sec. 4.4), and using mESCO (22) to perform a Bayesian optimization search, this issue could be problematic: The probability of constraint satisfaction (see (17)), computed using $q(f_*)$ is overestimated as compared to using $p(f_*)$, specially in areas close to failure points. This means that regions that are completely unsafe will not

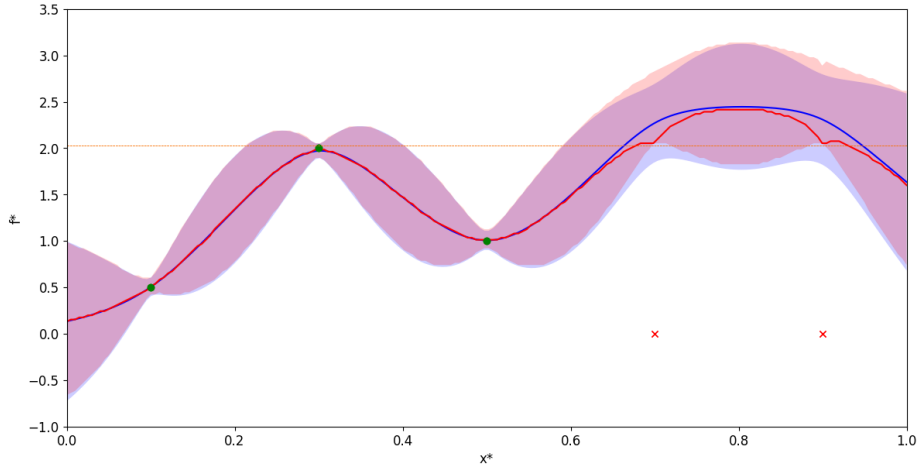


Figure 11: (red) True predictive distribution $p(f_*)$ of the Classified Regression model, conditioned on data from Example 2. We show the mode (red line) and the 95% confidence interval (red surface) at each x_* inside a discretization of the unit interval. Stable evaluations (green dots) and unstable evaluations (red crosses) are included, together with the optimal threshold $\hat{c} = 2.03$ (dashed orange line). The unstable evaluations (red crosses) have been represented at $f_* = 0$ for convenience. (blue) Gaussian approximation $q(f_*)$ to the predictive distribution (12) using the GPCR model. We show the mean (blue line) and the 95% confidence interval (blue surface).

be considered. However, in practice, this issue can be solved by using more conservative user-defined thresholds δ in (22).

This analysis shows that $q(f_*)$ is a reasonable approximation to $p(f_*)$, and thus to the true distribution (30).

Appendix C. Implementation details of mESCO

Herein, we provide additional implementation details of mESCO to those already given in Sec. 5.3.2.

In Algorithm 1, the function `CONDUCTEXPERIMENT()` returns a real evaluation from the objective and all constraints at the same location x_{next} . Lines 3 and 4 are optional (i.e., the threshold is only to be updated when the corresponding function is modeled with GPCR). Inside function `SAMPLECONSTRAINEDMINIMUM()`, the local optimization routine (lines 24-27) terminates when the maximum number of function evaluations R is reached, or when the relative precision κ for the optimum location is small enough. In our experiments, we set $\kappa = 10^{-3}$.

The function `LOCALOPTIMIZATIONSTEP()` (line 25) is not specifically described because it is a placeholder for any local optimization routine that can handle non-linear inequality constraints. Such routine takes as input arguments the thresholds c , and $\{c_j\}_{j=1}^M$ and the

callable functions $\text{Ve}(x_r, \mathcal{I}_r)$ and $\{\text{Ve}(x_r, \mathcal{I}_r^j), \hat{c}_j\}_{j=1}^M$, and returns the estimated location of the constrained minimum \tilde{x}_{\min} . In our experiments, we use the C++ implementation of SLSQP from the NLOPT toolbox ([Johnson](#)). Importantly, `LOCALOPTIMIZATIONSTEP()` also receives as input argument the initial location x_0 (line 22), which is randomly sampled within the unit domain $\mathcal{X} = [0, 1]^D$ every time a new sample \tilde{f}_{\min}^i is requested. Line 36 updates the virtual data set, and keeps it as a persistent variable (or internal state), used to update the covariance matrix as described in Sec. [5.3.2](#). The randomness involved in the resulting sample \tilde{f}_{\min}^i comes from the fact that both, x_0 (line 22) and y (line 35) are random variates.

Algorithm 1 min-Value Entropy Search with Unknown Level-Set Constraints (mESCO)

Initialize: $D, S, T, R, \delta, \mu_c, \sigma_c, \{\mu_{c_j}, \sigma_{c_j}\}_{j=1}^M$

Initialize: $\mathcal{D} \leftarrow \emptyset, \{\mathcal{D}_j \leftarrow \emptyset\}_{j=1}^M$

```
1: function MESCO
2:   for  $t \leftarrow 1$  to  $T$  do
3:      $\hat{c} \leftarrow \text{OptimizeThreshold}(\mu_c, \sigma_c, \mathcal{D})$  ▷ Instability threshold
4:      $\hat{c}_j \leftarrow \text{OptimizeThreshold}(\mu_{c_j}, \sigma_{c_j}, \mathcal{D}_j) \forall j \in \{1, \dots, M\}$  ▷ Constraint threshold
5:      $\alpha_{\text{mESCO}}(\cdot) \leftarrow \text{SampleConstrainedMinimum}(\hat{c}, \mathcal{D}, \{\hat{c}_j, \mathcal{D}_j\}_{j=1}^M, S)$ 
6:      $x_{\text{next}} \leftarrow \arg \max_{x \in \mathcal{X}} \alpha_{\text{mESCO}}(x)$  ▷ Obtain next point as in (22)
7:      $y, \{y_j\}_{j=1}^M \leftarrow \text{ConductExperiment}(x_{\text{next}})$ 
8:      $\mathcal{D} \leftarrow \{y, x_{\text{next}}\}$ 
9:      $\mathcal{D}_j \leftarrow \{y_j, x_{\text{next}}\} \forall j \in \{1, \dots, M\}$ 
10:     $x_{\text{bg}} \leftarrow \arg \min_{x \in \mathcal{X}} \mu(x|\mathcal{D}), \text{ s.t. } \prod_{j=1}^M \Pr[g_j(x) \leq c_j | \mathcal{D}_j] \geq 1 - \delta,$  ▷ (19)
11:  end for
12:   $y_{\text{bg}} \leftarrow \text{ConductExperiment}(x_{\text{bg}})$  ▷ Optional
13:  return  $x_{\text{bg}}, (y_{\text{bg}})$ 
14: end function
```

```
15: function OPTIMIZE_THRESHOLD( $\mu_c, \sigma_c, \mathcal{D}$ )
16:   return  $\arg \max_c \log Z_{\text{EP}}(c|\mathcal{D}) - 0.5(c - \mu_c)^2 / \sigma_c^2$  ▷ MAP optimization (16)
17: end function
```

```
18: function SAMPLECONSTRAINEDMINIMUM( $\hat{c}, \mathcal{D}, \{\hat{c}_j, \mathcal{D}_j\}_{j=1}^M, S$ )
19:   for  $i \leftarrow 1$  to  $S$  do
20:      $\mathcal{I}_1 \leftarrow \mathcal{D}$  ▷ Initialize virtual data set  $\mathcal{I}$  for the objective
21:      $\mathcal{I}_1^j \leftarrow \mathcal{D}_j \forall j = \{1, \dots, M\}$  ▷ Initialize virtual data sets  $\mathcal{I}^j$  for the constraints
22:      $x_0 \leftarrow \text{Uniform}([0, 1]^D)$  ▷ Uniformly sample the starting point
23:      $r \leftarrow 1$ 
24:     while  $r \leq R$  or  $\|x_{r+1} - x_r\| < \kappa$  do
25:        $x_{r+1} \leftarrow \text{LOCALOPTIMIZATIONSTEP}(\text{Ve}(x_r, \mathcal{I}_r), \hat{c}, \{\text{Ve}(x_r, \mathcal{I}_r^j), \hat{c}_j\}_{j=1}^M, x_0)$ 
26:        $r \leftarrow r+1$ 
27:     end while
28:      $\tilde{x}_{\min} \leftarrow x_r$ 
29:      $\tilde{f}_{\min}^i \leftarrow \text{Ve}(\tilde{x}_{\min}, \mathcal{I}_r)$ 
30:   end for
31:   return  $\alpha_{\text{mESCO}}(\cdot) \leftarrow [\tilde{f}_{\min}^i]_{i=1}^S$ 
32: end function
```

```
33: function VE( $x, \mathcal{I}_r$ ) ▷ Function to perform a “virtual evaluation”
34:   Sample  $f \sim p(f|\mathcal{I}_r, x)$  ▷ Sample from  $q(f_*)$  (14)
35:   Add noise  $y = f + \varepsilon, \varepsilon \sim \mathcal{N}(\varepsilon; 0, \sigma^2)$ 
36:    $\mathcal{I}_{r+1} \leftarrow \mathcal{I}_r \cup \{y, x\}$  ▷ Consider  $\mathcal{I}_r$  a persistent variable inside  $\text{Ve}()$ 
37:   return  $y$ 
38: end function
```

References

- Brian DO Anderson and John B Moore. *Optimal Control: Linear Quadratic Methods*. Dover Publications, 1990.
- Felix Berkenkamp, Angela P. Schoellig, and Andreas Krause. Safe controller optimization for quadrotors with Gaussian processes. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 493–496, 2016.
- Roberto Calandra, André Seyfarth, Jan Peters, and Marc P. Deisenroth. Bayesian optimization for learning gaits under uncertainty. *Annals of Mathematics and Artificial Intelligence*, 76(1-2):5–23, 2016.
- Konstantinos Chatzilygeroudis, Vassilis Vassiliades, Freek Stulp, Sylvain Calinon, and Jean-Baptiste Mouret. A survey on policy search algorithms for learning robot controllers in a handful of trials. *arXiv preprint arXiv:1807.02303*, 2018.
- John P Cunningham, Philipp Hennig, and Simon Lacoste-Julien. Gaussian probabilities and expectation propagation. *arXiv preprint arXiv:1111.6832*, 2011.
- Danny Drieß, Peter Englert, and Marc Toussaint. Constrained bayesian optimization of combined interaction force/task space controllers for manipulations. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 902–907. IEEE, 2017.
- Peter Englert and Marc Toussaint. Learning manipulation skills from a single demonstration. *The International Journal of Robotics Research*, 37(1):137–154, 2018.
- Jacob R. Gardner, Matt J. Kusner, Zhixiang E. Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *International Conference on Machine Learning (ICML)*, pages 937–945, 2014.
- Michael A. Gelbart. *Constrained Bayesian Optimization and Applications*. PhD thesis, Harvard University, Graduate School of Arts and Sciences, 2015.
- Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. Bayesian optimization with unknown constraints. In *Conference on Uncertainty in Artificial Intelligence*, pages 250–259, 2014.
- Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1344–1350, 2013.
- Robert B. Gramacy and Herbert Lee. Optimization under unknown constraints. *Bayesian Statistics 9*, 2011.
- Ryan-Rhys Griffiths and José M. Hernández-Lobato. Constrained Bayesian optimization for automatic chemical design. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.
- P. Hennig and CJ. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research (JMLR)*, 13:1809–1837, 2012.

- José M. Hernández-Lobato, Michael A. Gelbart, Ryan P. Adams, Matthew W. Hoffman, and Zoubin Ghahramani. A general framework for constrained Bayesian optimization using information-based search. *The Journal of Machine Learning Research (JMLR)*, 17(1): 5549–5601, 2016.
- José Miguel Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, pages 918–926, 2014.
- Steven G. Johnson. The NLOpt nonlinear-optimization package. URL <http://github.com/stevengj/nlopt>.
- Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.
- Jens Kober and Jan Peters. Learning motor primitives for robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2009.
- Alonso Marco, Philipp Hennig, Jeannette Bohg, Stefan Schaal, and Sebastian Trimpe. Automatic LQR tuning based on Gaussian process global optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.
- Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.
- Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. Toward global optimization, volume 2, chapter Bayesian Methods for Seeking the Extremum, 1978.
- Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 541–548, 2010.
- Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.
- Victor Picheny. A stepwise uncertainty reduction approach to constrained global optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 787–795, 2014.
- Akshara Rai, Rika Antonova, Seungmoon Song, William Martin, Hartmut Geyer, and Christopher G. Atkeson. Bayesian optimization using domain knowledge on the ATRIAS biped. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.
- Carl E. Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Matthias Schonlau, William J. Welch, and Donald R. Jones. Global versus local search in constrained optimization of computer models. *Lecture Notes-Monograph Series*, pages 11–25, 1998.

- Jens Schreiter, Duy Nguyen-Tuong, Mona Eberts, Bastian Bischoff, Heiner Markert, and Marc Toussaint. Safe exploration for active learning with Gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 133–149. Springer, 2015.
- Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.
- Jasper Snoek. *Bayesian optimization and semiparametric models with applications to assistive technology*. PhD thesis, University of Toronto, 2013.
- Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, pages 1015–1022, 2010.
- Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning (ICML)*, pages 5005–5014, 2018.
- Alexander von Rohr, Sebastian Trimpe, Alonso Marco, Peer Fischer, and Stefano Palagi. Gait learning for soft microrobots controlled by light fields. In *International Conference on Intelligent Robots and Systems*, 2018.
- Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning (ICML)*, pages 3627–3635, 2017.