# Distributionally Robust Bayesian Optimization

**Johannes Kirschner**
ETH Zürich

**Ilija Bogunovic**
ETH Zürich

**Stefanie Jegelka**
MIT

**Andreas Krause**
ETH Zürich

## Abstract

Robustness to distributional shift is one of the key challenges of contemporary machine learning. Attaining such robustness is the goal of distributionally robust optimization, which seeks a solution to an optimization problem that is worst-case robust under a specified distributional shift of an uncontrolled covariate. In this paper, we study such a problem when the distributional shift is measured via the maximum mean discrepancy (MMD). For the setting of zeroth-order, noisy optimization, we present a novel distributionally robust Bayesian optimization algorithm (DRBO). Our algorithm provably obtains sub-linear robust regret in various settings that differ in how the uncertain covariate is observed. We demonstrate the robust performance of our method on both synthetic and real-world benchmarks.

## 1 Introduction

Bayesian optimization (BO) is a framework for model-based sequential optimization of *black-box* functions that are expensive to evaluate and for which noisy point evaluations are available. Bayesian optimization algorithms have been successfully applied in a wide range of applications where the goal is to discover best-performing designs from a small number of trials, e.g., in vaccine and molecular design, gene optimization, automatic machine learning, robotics and control tasks, and many more.

In many practical tasks, the objective also depends on *contextual* covariates of the environment. If this context follows a known distribution, the setting is essentially that of stochastic optimization with the objective to maximize the expected pay-off. Often, however, there exists a distributional mismatch between the covariate distribution that the learner assumes, and the true distribution of the environment. Examples include automated machine learning, where hyperparameters are tuned on training data while the test distribution can differ; recommender systems, where the distribution of the users shifts with time; and robotics, where the simulated environmental variables are only an approximation of the real physical world. In particular, whenever there is a distributional mismatch between the true and the data distribution used at training time, the optimization solutions can result in inferior performance or even lead to unsafe/unreliable execution. The problem of *distributional data shift* has been recently identified as one of the most prevalent concrete challenges of modern AI safety (Amodei et al., 2016). While the connection of robust optimization (RO) and Bayesian optimization has recently been established by Bogunovic et al. (2018), robustness to *distributional data shift* remains unexplored in this field.

| Objective | Formulation |
|---|---|
| **Stochastic (SO)** | $\max_x \mathbb{E}_{c\sim P}[f(x,c)]$ |
| **Worst-case robust (RO)** | $\max_x \min_{c\in\Delta} f(x,c)$ |
| **Distributionally robust (DRO)** | $\max_x \inf_{Q\in\mathcal{U}} \mathbb{E}_{c\sim Q}[f(x,c)]$ |

Table 1: Different optimization objectives considered in Bayesian optimization.

In this paper, we introduce the setting of *distributionally robust Bayesian optimization (DRBO)*: The goal is to track the optimal input that maximizes the expected function value under the worst-case distribution of an external, contextual parameter. In distributionally robust optimization (DRO), such a worst-case distribution belongs to a known *uncertainty set* of distributions that is typically chosen as a ball centered around a given reference distribution. To measure the

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. 

distance between distributions, in this work, we focus on the kernel-based *maximum mean discrepancy* (MMD) distance. This metric fits well with the kernel-based regularity assumptions on the unknown function that are typically made in Bayesian optimization.

### 1.1 Related Work

A large number of Bayesian optimization algorithms have been developed over the years, (e.g. Srinivas et al., 2010; Wang and Jegelka, 2017; Hennig and Schuler, 2012; Chowdhury and Gopalan, 2017; Bogunovic et al., 2016b). Several practical variants of the standard setting were addressed recently, including contextual (Krause and Ong, 2011; Valko et al., 2013; Lamprier et al., 2018; Kirschner and Krause, 2019) and time-varying (Bogunovic et al., 2016a) BO, high-dimensional BO (Djolonga et al., 2013; Kandasamy et al., 2015; Kirschner et al., 2019), BO with constraints (Gardner et al., 2014; Gelbart et al., 2014), heteroscedastic noise (Kirschner and Krause, 2018) and uncertain inputs (Oliveira et al., 2019).

Two classical objectives for optimization under uncertainty are stochastic optimization (SO) (Srinivas et al., 2010; Krause and Ong, 2011; Lamprier et al., 2018; Oliveira et al., 2019; Kirschner and Krause, 2019) and robust optimization (RO) (Bogunovic et al., 2018), see Table 1. SO asks for a solution that performs well in expectation over an uncontrolled, stochastic covariate. Here, the assumption is that the distribution of the contextual parameter is known, or (i.i.d.) samples are provided. Some variants of SO have been considered in the related contextual Bayesian optimization works (Krause and Ong, 2011; Valko et al., 2013; Kirschner and Krause, 2019). RO aims at a solution that is robust with respect to the worst possible realization of the context parameter. The RO objective has recently been studied in Bayesian optimization in (Bogunovic et al., 2018); the authors provide a robust BO algorithm, and obtain strong regret guarantees. In many practical scenarios, however, the solution to the SO problem might be highly *non-robust*, while on the other hand, the worst-case RO solution might be overly *pessimistic*. This motivates us to consider the *distributionally robust optimization* (DRO), which is a "middle ground" between SO and RO.

Distributionally robust optimization (DRO) dates back to the seminal work of Scarf (1957) and since then it has become an important topic in robust optimization (e.g. Bertsimas et al., 2018; Goh and Sim, 2010). It has recently received significant attention in machine learning, in particular due to its relation to regularization, adversarial learning, and generalization (Staib et al., 2018). The full literature on DRO is too vast to be adequately covered here, so we refer the interested reader to the recent review by Rahimian and Mehrotra (2019) and references within. For defining the uncertainty sets of distributions, different DRO works have studied $\phi$-divergences (Ben-Tal et al., 2013; Namkoong and Duchi, 2017), Wasserstein (Gao et al., 2017; Esfahani and Kuhn, 2018; Sinha et al., 2017) and the MMD (Staib and Jegelka, 2019) distances. In this work, we focus on the kernel-based MMD distance, but unlike previous DRO works, we assume that the objective function is *unknown*, and only noisy point evaluations are available.

We conclude this section by mentioning other robust aspects and settings that have been previously considered in Bayesian optimization. BO with outliers has been considered by Martinez-Cantin et al. (2017), while the setting in which sampled points are subject to uncertainty has been studied by Nogueira et al. (2016); Beland and Nair (2017); Oliveira et al. (2019). These settings differ significantly from the one considered in this paper and they do not consider robustness under distributional shift. Finally, we note that another robust BO algorithm has been recently developed for playing unknown repeated games against non-cooperative agents (Sessa et al., 2019).

While this work was under submission, a related approach for distributionally robust Bayesian quadrature appeared online (Nguyen et al., 2020). The authors propose an approach based on Thompson sampling to solve a related robust objective for Bayesian quadrature. Our work captures this scenario in the "simulator setting", detailed below. The main difference in the analysis is that we bound worst-case frequentist regret opposed to the expected Bayesian regret.

**Contributions** We propose a novel, distributionally robust Bayesian optimization (DRBO) algorithm. Our analysis shows that the DRBO achieves sublinear robust regret on several variants of the setting. Finally, we demonstrate robust performance of the DRBO method on synthetic and real-world benchmarks.

## 2 Problem Statement

Let $f : \mathcal{X} \times \mathcal{C} \to \mathbb{R}$ be an *unknown* reward function defined over a parameter space $\mathcal{X} \times \mathcal{C}$ with finite[1] action and context sets, $\mathcal{X}$ and $\mathcal{C}$. The objective is to optimize $f$ from *sequential* and *noisy* point evaluations. In our main setup, at each time step $t$, the learner chooses $x_t \in \mathcal{X}$ whereas the environment provides the context $c_t \in \mathcal{C}$ together with the *noisy* function observation $y_t = f(x_t, c_t) + \xi_t$, where

[1]Our formulation and the theory extend to continuous sets $\mathcal{C}$ and $\mathcal{X}$, but for the algorithm we rely on solving a convex program of size $|\mathcal{C}|$.

$\xi_t \sim \mathcal{N}(0, \sigma^2)$ with known $\sigma^2$ and independence between time steps. More generally, our results hold if the noise is $\sigma$-sub-Gaussian, which allows for non-Gaussian likelihoods (e.g., bounded noise). Further, we assume that $c_t$ is sampled independently from an unknown, time-dependent distribution $P_t^*$.

**Optimization objective.** We consider the *distributionally robust optimization* (DRO) (Scarf, 1957) objective, which asks to perform well simultaneously for a range of problems, each determined by a distribution in some uncertainty set. This is in contrast to SO, where we seek good performance against a single problem instance parametrized by a given distribution.

In DRO, the objective is to find $x \in \mathcal{X}$ that solves

$$\max_{x \in \mathcal{X}} \inf_{Q \in \mathcal{U}_t} \mathbb{E}_{c \sim Q}[f(x, c)]. \tag{1}$$

Here, $\mathcal{U}_t$ is a known *uncertainty set* of distributions over $\mathcal{C}$ that can depend on the step $t$ and contains the true distribution $P_t^* \in \mathcal{U}_t$. Typically, $\mathcal{U}_t$ is chosen as a ball of radius (or margin) $\epsilon_t > 0$, and centered around a given *reference distribution* $P_t$ on $\mathcal{C}$, i.e.,

$$\mathcal{U}_t = \{Q : d(Q, P_t) \leq \epsilon_t\},$$

where $d(\cdot, \cdot)$ measures the discrepancy between two distributions. A possible choice for the reference distribution $P_t$, is the empirical sample distribution $\hat{P}_t = t^{-1} \sum_{s=1}^{t} \delta_{c_s}$, which is an instance of *data-driven DRO* (Bertsimas et al., 2018). Depending on the underlying function and the uncertainty set $\mathcal{U}_t$, the robust solution can significantly differ from the solution to the stochastic objective $\max_{x \in \mathcal{X}} \mathbb{E}_{c \sim P}[f(x, c)]$ for a fixed (and typically known) distribution $P$. We illustrate such a case in Fig. 1.

Hence, at time step $t$, the learner receives a reference distribution $P_t \in \mathcal{P}(C)$ and margin $\epsilon_t > 0$. Our objective is to choose a sequence of actions $x_1, \dots, x_T$ that minimizes *robust cumulative regret*:

$$R_T = \sum_{t=1}^{T} \inf_{Q \in \mathcal{U}_t} \mathbb{E}_Q[f(x_t^*, c)] - \inf_{Q \in \mathcal{U}_t} \mathbb{E}_Q[f(x_t, c)], \tag{2}$$

where $x_t^* = \max_{x \in \mathcal{X}} \inf_{Q \in \mathcal{U}_t} \mathbb{E}_Q[f(x, c)]$. The *robust regret* measures the cumulative loss of the learner on the chosen sequence of actions w.r.t. the worst case distribution over $\mathcal{C}$.

**RKHS Regression.** The main regularity assumption of Bayesian optimization is that $f$ belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with known kernel $k$. We denote the Hilbert norm by $\|\cdot\|_{\mathcal{H}}$ and assume $\|f\|_{\mathcal{H}} \leq B$ for some known $B > 0$. From the observed data $\mathcal{D}_t = \{(x_1, c_1, y_1), \dots, (x_t, c_t, y_t)\}$, we

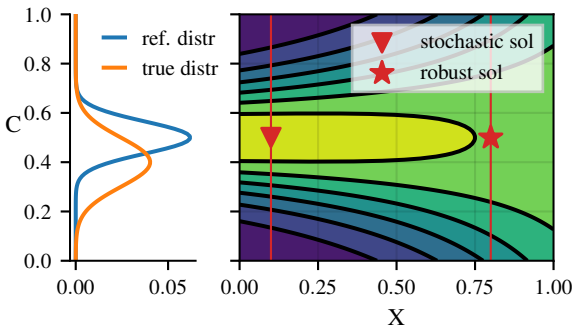



Figure 1: A function where the robust solution significantly differs from the stochastic solution. The learner obtains the blue reference distribution over the context set $\mathcal{C}$ and chooses a design $x \in \mathcal{X}$. If the distribution over the context set is equal to the reference, the solution marked by the triangle maximizes the expected reward. On the other hand, if the true distribution (orange) is shifted away from the reference, the flatter region of the reward function, marked by the star, provides higher expected reward.

can compute a kernel ridge regression estimate with

$$\hat{f}_t = \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{t-1} (g(x_i, c_i) - y_i)^2 + \|g\|_{\mathcal{H}}^2 \,. \tag{3}$$

The representer theorem provides the standard, closed-form solution for the least-squares estimate (Rasmussen and Williams, 2006). The next lemma is a standard result by Srinivas et al. (2010); Abbasi-Yadkori (2013). It provides a frequentist confidence interval of the form $[\hat{f}_t(x, c) \pm \beta_t \sigma_t(x, c)]$ that contains the true function values $f(x, c)$ with high probability. The exact definitions of $\hat{f}_t$ and $\sigma_t$ can be found in Appendix A; we just note here that $\hat{f}_t(x, c)$ and $\sigma_t(x, c)^2$ are the posterior mean and posterior variance functions of the corresponding Bayesian Gaussian process model (Rasmussen and Williams, 2006). We denote the data kernel matrix by $(K_t)_{i,j=1,\dots,t} = k(x_i, c_i, x_j, c_j)$, and assume that $k(x, c, x', c') \leq 1$.

**Lemma 1.** *With probability at least $1 - \delta$, for any $x \in \mathcal{X}$, $c \in \mathcal{C}$ at any time $t \geq 1$,*

$$|\hat{f}_t(x, c) - f(x, c)| \leq \beta_t \sigma_t(x, c)$$

*with* $\beta_t = \sigma \sqrt{\log\det\left(\mathbf{1}_t + K_t\right) + 2\log\frac{1}{\delta}} + B.$

We explicitly define the upper and lower confidence bounds for every $x \in \mathcal{X}$ and $c \in \mathcal{C}$ as follows:

$$\mathrm{ucb}_t(x, c) := \hat{f}_t(x, c) + \beta_t \sigma_t(x, c),$$
$$\mathrm{lcb}_t(x, c) := \hat{f}_t(x, c) - \beta_t \sigma_t(x, c).$$

For a fixed $x$, we use $\mathrm{ucb}_x^t := \mathrm{ucb}_t(x, \cdot)$ and $\mathrm{lcb}_x^t := \mathrm{lcb}_t(x, \cdot)$ to refer to the corresponding vectors in $\mathbb{R}^{|\mathcal{C}|}$.

Finally, we introduce a sample complexity parameter, the *maximum information gain*:

$$\gamma_T := \max_{\{(x_t,c_t)\}_{t=1}^T} \log\det\left(\mathbf{1}_t + K_T\right) . \tag{4}$$

The information gain appears in the regret bounds for Bayesian optimization (Srinivas et al., 2010). Analytical upper bounds are known for a range of kernels, e.g., for the RBF kernel, $\gamma_T \leq \mathcal{O}(\log(T)^{d+1})$ if $\mathcal{X}\times\mathcal{C} \subset \mathbb{R}^d$.

**Maximum Mean Discrepancy (MMD).** MMD is a kernel-based discrepancy measure between distributions (e.g., Muandet et al. (2017)). It has been used in various applications, including generative modeling (Sutherland et al., 2016; Bińkowski et al., 2018), DRO (Staib and Jegelka, 2019) and kernel sample tests (Gretton et al., 2012; Chwialkowski et al., 2016). Let $\mathcal{H}_M$ be an RKHS with corresponding kernel $k_M : \mathcal{C}\times\mathcal{C} \to \mathbb{R}$. For two distributions $P$ and $Q$ over $\mathcal{C}$, the *maximum mean discrepancy* (MMD) is

$$d(P,Q) := \sup_{g\in\mathcal{H}_M : \|g\|_{\mathcal{H}_M}\leq 1} \mathbb{E}_{c\sim P}[g(c)] - \mathbb{E}_{c\sim Q}[g(c)] . \tag{5}$$

Note that the kernel $k_M$ over $\mathcal{C}$ that defines the MMD is different from the kernel $k$ over $\mathcal{X}\times\mathcal{C}$ that is used for regression. An equivalent way of writing $d(P,Q)$ is via *kernel mean embeddings* (Muandet et al., 2017, Section 3.5). Specifically, any distribution $P$ over $\mathcal{C}$ can be embedded into $\mathcal{H}_M$ via the mean embedding $m_P = \mathbb{E}_{c\sim P}[k_M(c,\cdot)]$, which satisfies $\langle m_P, k_M(c',\cdot)\rangle = \mathbb{E}_{c\sim P}[k_M(c',c)]$ for all $c' \in \mathcal{C}$. An equivalent expression for the MMD (5) is

$$d(P,Q) = \|m_P - m_Q\|_{\mathcal{H}} . \tag{6}$$

More explicitly, for finite context set $\mathcal{C}$ and probability vectors $w_i = \mathbb{P}_P[c_i]$ and $w'_i = \mathbb{P}_Q[c_i]$, the kernel mean embeddings are $m_P = \sum_{i=1}^n w_i k_M(c_i,\cdot)$ and $m_Q = \sum_{i=1}^n w'_i k_M(c_i,\cdot)$, respectively. With the kernel matrix $(M)_{ij} := k_M(c_i,c_j)$, the MMD becomes

$$d(P,Q) = \sqrt{(w-w')^\top M (w-w')} =: \|w-w'\|_M .$$

# 3 Distributionally Robust Bayesian Optimization

We now introduce a Bayesian optimization algorithm for our main objective (2). We will start with a *general formulation* that allows for time-dependent reference distributions $P_t$ and margins $\epsilon_t$. We then continue with *data-driven* DRO (Bertsimas et al., 2018), where we specialize the general setup and choose the empirical distribution $P_t = \frac{1}{t}\sum_{s=1}^t \delta_{c_s}$ as reference distribution. Hence, our algorithm chooses actions that are robust w.r.t. the estimation error of the true context distribution. Finally, we motivate and discuss the *simulator* setting, where the learner is allowed to choose the context $c_t$ and obtains the corresponding evaluation $y_t = f(x_t,c_t) + \xi_t$.

**Algorithm 1** DRBO - General Setting

Initialize $(K_x)_{i,j} = k(x,c_i,x,c_j), \mathcal{C} = \{c_1,\dots,c_n\}$
**For** step $t = 1,2,\dots,T$:

1. Learner obtains reference distribution $P_t$ with $w_i^t = \mathbb{P}[c = c_i]$, and margin $\epsilon_t$
2. Define $(\mathrm{ucb}_x^t)_j := \hat{f}_t(x,c_j) + \beta_t\sigma_t(x,c_j)$
3. Define $w_x^{\mathrm{ucb}_t} := \arg\min_{w'}\langle \mathrm{ucb}_x^t, w'\rangle$, s.t. $\|w'\|_1 = 1, 0 \leq w'_j \leq 1$ $(\forall j \in [n])$, and $\|w' - w^t\|_M \leq \epsilon_t$
4. Choose action $x_t = \arg\max_{x\in\mathcal{X}}\langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle$
5. Learner observes $c_t \sim P_t^*$ and $y_t = f(x_t,c_t) + \xi_t$.
6. Use $\{x_t,c_t,y_t\}$ to update $\hat{f}_{t+1}(\cdot,\cdot)$ and $\sigma_{t+1}(\cdot,\cdot)$.

## 3.1 General DRBO

In our general DRBO formulation, the interaction protocol at time $t$ is specified by the following steps:

1. The environment chooses a reference distribution $P_t$ and margin $\epsilon_t$. This defines the uncertainty set
$$\mathcal{U}_t = \{Q : d(Q,P_t) \leq \epsilon_t\} . \tag{7}$$
2. The learner observes $P_t$ and $\epsilon_t$, and chooses a robust action $x_t \in \mathcal{X}$.
3. The environment chooses a sampling distribution $P_t^* \in \mathcal{U}_t$ and the context is realized as an independent sample $c_t \sim P_t^*$.
4. The learner observes the reward $y_t = f(x_t,c_t) + \xi_t$ and $c_t \sim P_t^*$.

We make no further assumptions on how the environment chooses the sequences $P_t, P_t^*$ and $\epsilon_t$. The DRBO algorithm for this setting is given in Algorithm 1. Recall that $P_t$ is a distribution over the finite context set $\mathcal{C}$ with $n$ elements, and we use $w^t \in \mathbb{R}^n$ to denote a probability vector with entries $w_i^t = \mathbb{P}_{P_t}[c = c_i]$ for every $i \in [n]$. With this, the inner adversarial problem for a fixed action $x$ can be equivalently written as:

$$\inf_{Q: d(P_t,Q)\leq\epsilon_t} \mathbb{E}_{c\sim Q}[f(x,c)] = \min_{\substack{w':\|w'\|_1=1,\\ 0\leq w'_j\leq 1\ \forall j\in[n],\\ \|w'-w^t\|_M\leq\epsilon_t}} \langle w', f_x\rangle, \tag{8}$$

where $f_x := f(x,\cdot) \in \mathbb{R}^n$, and $M \in \mathbb{R}^{n\times n}$ with $(M)_{ij} := k_M(c_i,c_j)$. In particular the solution to (8)

is the worst-case distribution over $c$ for the objective $f$ if the learner chooses action $x$. Since the constraints are convex, the program (8) can be solved efficiently by standard convex optimization solvers.

Since the true function values $f_x$ are unknown to the learner, we can only obtain an approximate solution to (8). In our algorithm, we hence use an optimistic upper bound instead. Specifically, we substitute $\mathrm{ucb}_x^t$ for $f_x$ to compute the "optimistic" worst-case distribution for every action $x$. Finally, at time $t$, the learner chooses $x_t$ that maximizes the optimistic expected reward under the worst-case distribution.

The DRBO algorithm achieves the following regret bound.

**Theorem 2.** *The robust regret $R_T$ of Algorithm 1, with $\beta_t = \sigma\sqrt{\log\det\big(\mathbf{1}_t + K_t\big) + 2\log\frac{2}{\delta}} + B$, is bounded with probability at least $1-\delta$ by*

$$R_T \leq 4\beta_T\sqrt{T\big(\gamma_T + 4\log\big(\tfrac{12}{\delta}\big)\big)} + 2B'\sum_{t=1}^{T}\epsilon_t \ .$$

*Here, $\gamma_T$ is the maximum information gain defined in Eq. (4), $\|f\|_{\mathcal{H}} \leq B$ and $B' = \max_{x\in\mathcal{X}}\|f_x\|_{M^{-1}}$.*

The complete proof is given in Appendix B.1, and we only sketch the main steps here. Denote by $w_t^*$ the probability vector of the true distribution at time $t$, and by $w_{x_t}^f$ the solution to (8) at $x_t$. The idea is to bound the instantaneous regret at time $t$ by

$$\begin{aligned} r_t &= \inf_{Q:d(P_t,Q)\leq\epsilon_t}\mathbb{E}_Q[f(x^*,c)] - \inf_{Q:d(P_t,Q)\leq\epsilon_t}\mathbb{E}_Q[f(x_t,c)] \\ &\overset{(i)}{\leq} \langle w_t^*, \mathrm{ucb}_{x_t}^t\rangle - \langle w_{x_t}^f, f_{x_t}\rangle \\ &= \langle w_t^*, \mathrm{ucb}_{x_t}^t - f_{x_t}\rangle + \langle w_t^* - w_{x_t}^f, f_{x_t}\rangle \\ &\overset{(ii)}{\leq} 2\beta_t\langle w_t^*, \sigma_t(x_t,\cdot)\rangle + \|w_t^* - w_{x_t}^f\|_M\|f_{x_t}\|_{M^{-1}} \\ &\overset{(iii)}{\leq} 2\beta_T\langle w_t^*, \sigma_t(x_t,\cdot)\rangle + 2\epsilon_t B' \ . \end{aligned}$$

For the first inequality (i), we used that $f_x \leq \mathrm{ucb}_x$, the definition of the UCB action and that $w_t^* \in \mathcal{U}_t$. In step (ii), we use Cauchy-Schwarz and the confidence bounds, and step (iii) follows since $w_{x_t}^f \in \mathcal{U}_t$. From here it remains to sum the instantaneous regret, where we rely on Lemma 3 in (Kirschner and Krause, 2018) to relate the expectation over the true sampling distribution $\langle w_t^*, \sigma_t(x_t,\cdot)\rangle$ to the observed values $\sigma_t(x_t,c_t)$.

In the regret bound in Theorem 2, the first term is the same as the standard regret bound for GP-UCB (Srinivas et al., 2010; Abbasi-Yadkori, 2013) and reflects the statistical convergence rate for estimating the RKHS function. The additional term $B'T\epsilon$ (for $\epsilon_t = \epsilon$) is specific to our setting. First, the complexity parameter $B' = \max_{x\in\mathcal{X}}\|f_x\|_{M^{-1}}$ quantifies how much the distributional shift can increase the regret on the given objective $f$. A crude upper bound is $B' \leq B\sqrt{\lambda_{\max}(M^{-1})|\mathcal{C}|}$, but in general $B'$ can be much smaller. The linear scaling $\mathcal{O}(\epsilon T)$ of the regret bound is arguably unsatisfying, but seems unavoidable without further assumptions. A problematic case is when the true distribution $P_t^*$ is supported on a single context, e.g., $P_t^* = \delta_{c_1}$, and the learner is not able to learn the function values at different contexts $c_i$ for $i > 1$. In this case, the learner can never infer the robust solution exactly from the data and consequently incurs constant regret of order $\epsilon_t$ per round. In practice, we do not expect that this severely affects the performance of our algorithm if the true distribution sufficiently *covers* the context space. We leave a precise formulation of this intuition for future work.

Instead, in the following sections we explore two different ways of controlling the additional regret that the learner incurs in the general DRBO setting. First, for the *data-driven* setting, we will set the reference distribution to the empirical distribution of the observed context samples. In this case, the margin $\epsilon_t$ is the distance to the true sampling distribution, which for the MMD is of order $1/\sqrt{t}$ and results in $\sum_{t=1}^T \epsilon_t = \mathcal{O}(\sqrt{T})$. In the second variant, the learner is allowed to also choose $c_t$, which circumvents the estimation problem outlined above and avoids the linear regret term.

### 3.2 Data-Driven DRBO

In *data-driven* DRBO, we assume there is a fixed but unknown distribution $P^*$ on $\mathcal{C}$. In each round, the learner first chooses an action $x_t \in \mathcal{X}$, and then observes a context sample $c_t \sim P^*$ together with the corresponding observation $y_t = f(x_t, c_t) + \xi_t$. At the beginning of round $t$, the learner compute the empirical distribution $\hat{P}_t = \frac{1}{t-1}\sum_{s=1}^{t-1}\delta_{c_s}$ using the observed contexts $\{c_1,\dots,c_{t-1}\}$. The objective is to choose a sequence of actions $x_t$, which is robust to the estimation error in $\hat{P}_t$. This corresponds to minimizing the robust regret (2), where we set $P_t = \hat{P}_t$ for every $t$.

As the learner observes more context samples, she becomes more confident about the true unknown $P^*$. It is therefore reasonable to shrink the uncertainty set of distributions $\mathcal{U}_t = \{Q : d(Q,\hat{P}_t) \leq \epsilon_t\}$ over time. We make use of the following lemma.

**Lemma 3** (Muandet et al. (2017), Theorem 3.4)**.** *Assume $k(c_i,c_j) \leq 1$ for all $c_i, c_j \in \mathcal{C}$. Let $P^*$ be the true context distribution over $\mathcal{C}$, and let $\hat{P}_t = t^{-1}\sum_{s=1}^t \delta_{c_s}$ be the empirical sample distribution. Then, with probability at least $1-\delta$,*

$$d(P^*,\hat{P}_t) \leq \frac{1}{\sqrt{t}}\left(2 + \sqrt{2\log(1/\delta)}\right).$$

Lemma 3 shows how to set the margin $\epsilon_t$ such that, at time $t$, the true distribution is contained with high probability in the uncertainty set around the empirical distribution. The interaction protocol at time $t$ is then:

1. The learner computes the empirical distribution $\hat{P}_t$ and corresponding margin $\epsilon_t$ according to Lemma 3, and defines the uncertainty set
$$\mathcal{U}_t = \{Q : d(Q, \hat{P}_t) \leq \epsilon_t\} \ .$$
2. The learner chooses a robust action $x_t$.
3. The learner observes reward $y_t = f(x_t, c_t) + \xi_t$ and context sample $c_t \sim P^*$.

We follow Algorithm 1, and set the reference distribution and margin as outlined above. As a consequence of Theorem 2 we obtain the following regret bound.

**Corollary 4.** *The robust regret $R_T$ of Algorithm 1, with $\beta_t = \sigma\sqrt{\log\det\left(\mathbf{1}_t + K_t\right) + 2\log\frac{3}{\delta}} + B$ and $\epsilon_t = \frac{1}{\sqrt{t}}\left(2 + \sqrt{2\log(\frac{6t^2}{\delta})}\right)$ is bounded in the data-driven scenario with probability at least $1-\delta$ by*

$$R_T \leq 2\beta_T\sqrt{T}\sqrt{\gamma_T(1+\log(3/\delta)} + 4B'\sqrt{T}\big(2+\sqrt{2\log\big(\tfrac{6T^2}{\delta}\big)}\big) \ , \qquad (9)$$

*where $\gamma_T$ is the maximum information gain as defined in (4), $\|f\|_{\mathcal{H}} \leq B$ and $B' = \max_{x\in\mathcal{X}} \|f_x\|_{M^{-1}}$.*

The proof can be found in Appendix B.2. We just note that we increased the value of $\epsilon_t$ such that Lemma 3 holds simultaneously over all time steps. In the data-driven contextual setting *without the robustness requirement*, several related approaches have been proposed (Lamprier et al., 2018; Kirschner and Krause, 2019). These are based on computing a UCB score directly at the kernel mean embedding of the empirical distribution $\hat{P}_t$ . To account for the estimation error, an additional exploration bonus is added. We note that as $t \to \infty$ and $\hat{P}_t$ becomes an accurate estimation of $P^*$, both robust and non-robust approaches converge to the stochastic solution. The advantage of the *robust* formulation is that we explicitly minimize the loss under the worst-case estimation error in the context distribution. As we demonstrante in our experiments (in Section 4), DRBO obtains significantly smaller regret when the robust and stochastic solutions are different.

### 3.3 Simulator DRBO

In our second variant of the general setup, the learner is allowed to choose $c_t$ in addition to $x_t$ and then obtains the observation $y_t = f(x_t, c_t) + \xi_t$.

**Algorithm 2** DRBO - Simulator Setting

Initialize $(K_x)_{i,j} = k(x, c_i, x, c_j), \mathcal{C} = \{c_1, \ldots, c_n\}$
**For** step $t = 1, 2, \ldots, T$:

1. Obtain reference distribution $P_t$ with $w_i^t = \mathbb{P}[c = c_i]$, margin $\epsilon_t$
2. Define $(\mathrm{ucb}_x^t)_j := \hat{f}_t(x, c_j) + \beta_t\sigma_t(x, c_j)$
3. $w_x^{\mathrm{ucb}_t} := \arg\min_{w'} \langle \mathrm{ucb}_x^t, w'\rangle$, s.t. $\|w'\|_1 = 1, 0 \leq w_j \leq 1\ (\forall j \in [n]), \|w' - w\|_M \leq \epsilon_t$
4. $x_t = \arg\max_{x\in\mathcal{X}} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle$
5. $c_t = \arg\max_{c\in\mathcal{C}} \sigma_t(x_t, c)$.
6. Observe $y_t = f(x_t, c_t) + \xi_t$ from simulator

One example of this setting, previously considered in the context of RO (Bogunovic et al., 2018), is when the learner tunes control parameters with a simulator of the environment (e.g. for a building heating system). The simulator gives the learner the ability to evaluate the objective at any specific context $c_t$. The objective is to simultaneously (or only at the final time $T$) deploy a robust solution $x_T$ on the real system, where the covariate $c_t$ is uncontrolled. Again, the learner's objective is to be robust with respect to an uncertainty set of distributions on $c_t$ *on the real environment* (e.g. for heating control, we want robustness on predicted weather conditions that effect the building's state). With this motivation in mind, we refer to this setup as *simulator* DRBO. Formally, the interaction protocol is:

1. The environment provides a reference distribution $P_t$, margin $\epsilon_t$ and uncertainty set $\mathcal{U}_t$ as before.
2. The learner chooses an action $x_t$ and a context $c_t \in \mathcal{C}$.
3. The learner observes reward $y_t = f(x_t, c_t) + \xi_t$ from the simulator.
4. The learner deploys a robust action $x_t$ on the real system (or possibly only at the final step $T$).

We provide Algorithm 2 for this setting. As before, $x_t$ is an optimistic action under the worst-case distribution. In addition, the learner chooses $c_t = \arg\max_{c\in\mathcal{C}} \sigma_t(x_t, c_t)$ as the context with the largest estimation uncertainty at $x_t$. We bound the robust regret in the next theorem.

**Theorem 5.** *In the simulator setting, Algorithm 2, with $\beta_t = \sigma\sqrt{\log\det\left(\mathbf{1}_t + K_t\right) + 2\log\frac{1}{\delta}} + B$, obtains bounded robust regret w.p. at least $1-\delta$,*

$$R_T \leq 2\beta_T\sqrt{\gamma_T T} \ .$$

We provide the proof of Theorem 5 in Appendix B.3.

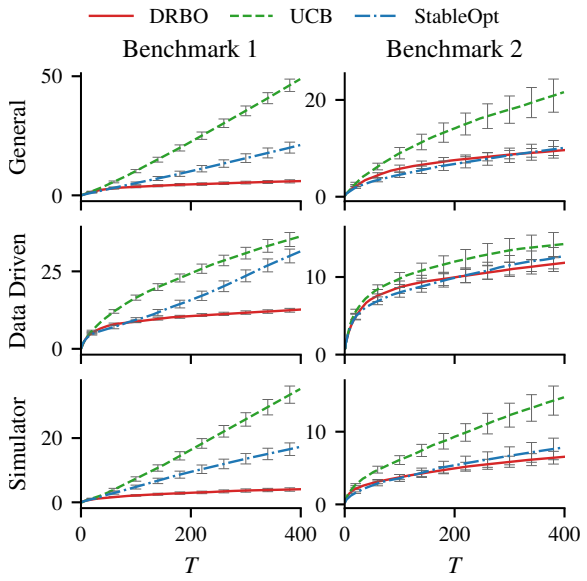



Figure 2: Results for two synthetic benchmarks, where the stochastic, worst-case robust and distributionally robust solution are all different *(left)* or coincide *(right)*. All plots show robust regret, averaged over 50 independent runs and the error bars indicate the standard error.

Perhaps surprisingly, this rate is the same as for GP-UCB in the standard setting (a similar result was obtained for RO (Bogunovic et al., 2018)). This is because now the learner can estimate $\hat{f}_t$ globally at any input $(x_t, c_t) \in \mathcal{X} \times \mathcal{C}$, and the sample complexity to infer the robust solution only depends on the sample complexity of estimating $f$.

In the simulator setting, the performance of the *final solution* can be of significant interest if we aim to deploy the obtained parameter on the real system. To this end, we allow the final solution $\hat{x}_T$ to be different from the last evaluation $x_T$. The metric of interest is then the robust simple regret,

$$r_T = \max_{x \in \mathcal{X}} \inf_{Q} \mathbb{E}_{c \sim Q} f(x, c) - \inf_{Q} \mathbb{E}_{c \sim Q} f(\hat{x}_T, c).$$

To obtain a bound on the simple regret, we assume that the margin $\epsilon = \epsilon_t$ and the reference distribution $P = P_t$ are fixed. This is a natural requirement, which allows the learner to optimize the simple regret for the final solution $\hat{x}_T$ w.r.t. $P$ and $\epsilon$. We choose the final solution $\hat{x}_T := x_{\hat{t}}$ among the iterates $x_1, \dots, x_T$ from Algorithm 2 with

$$\hat{t} := \arg\max_{t=1,\dots,T} \min_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w^t\|_M \le \epsilon}} \langle w', \mathrm{lcb}^t_{x_t} \rangle \ . \tag{10}$$

The program computes the best robust solution among the iterates $\{x_1, \dots, x_T\}$ using the *conservative* function values $\mathrm{lcb}^t_x$ of the corresponding time steps $t$. It is easy to maintain $\hat{x}_T$ iteratively by computing the conservative, worst-case payoff of the action $x_t$ and comparing to the previous solution $\hat{x}_{t-1}$.

**Corollary 6** (Simple Regret)**.** *With probability at least* $1 - \delta$*, the solution* $\hat{x}_T$ *obtains simple regret*

$$r_T \le 2\beta_T \sqrt{\gamma_T / T}. \tag{11}$$

This result is a consequence of the fact that the simple regret of $\hat{x}_t$ is upper bounded by the simple regret of each iterate $x_t$. The guarantee then follows from the proof of Theorem 5. We provide the complete argument in Appendix B.4.

## 4 Experiments

We evaluate the proposed DRBO in the general, data-driven and simulator setting on two synthetic test functions, and on a recommender task based on a real-world crop yield data set. In our experiments, we compare to StableOpt (Bogunovic et al., 2018) and a stochastic UCB variant (Srinivas et al., 2010; Kirschner and Krause, 2019).

**Baselines** The first baseline is a stochastic variant of the UCB approach (Srinivas et al., 2010; Kirschner and Krause, 2019), which chooses actions according to optimistic expected payoff w.r.t. the reference distribution,

$$x_t^{\mathrm{UCB}} = \arg\max_{x \in \mathcal{X}} \mathbb{E}_{P_t}[\mathrm{ucb}_t(x, c)] \ .$$

Our second baseline is StableOpt (Bogunovic et al., 2018), an approach for worst-case robust optimization. It chooses actions according to

$$x_t^{\mathrm{STABLE}} = \arg\max_{x \in \mathcal{X}} \min_{c \in \Delta_t} \mathrm{ucb}_t(x, c) \ ,$$

for a robustness set of possible context values $\Delta_t \subset \mathcal{C}$. There is no canonical way of choosing $\Delta_t$ in our setting, and we use $\Delta_t = \{c \in \mathcal{C} : \|c - \mathbb{E}_{c' \sim P_t}[c']\|_2 \le \epsilon_t\}$. With the decreasing margin and the discretization of the context domain, it can happen that $\Delta_t$ is an empty set. In this case we explicitly set $\Delta_t = \{\arg\min_{c \in \mathcal{C}} \|c - \mathbb{E}_{c' \sim P_t}[c']\|_2\}$.

UCB and StableOpt optimize for the *stochastic and worst-case robust solutions* respectively, and therefore can exhibit *linear* regret for the robust regret (unless $\epsilon_t \to 0$ as in the data-driven setting). For all approaches we use the same RKHS hyper-parameters. In particularly we set $\beta_t = 2$, which is a common practice to improve performance over the (conservative) theoretical values.

**Benchmarks** Our *first synthetic benchmark* is the function illustrated in the introduction. The reference distribution is $P_t = \mathcal{N}(0.5, 0.05)$ and the true sampling distribution is $P^* = \mathcal{N}(0.45, 0.1)$. For simplicity, we set the margin to the exact MMD distance $\epsilon_t := d(P_t, P^*)$. On this function, the stochastic, worst-case robust and distributionally robust solution all differ, which leads to linear robust regret for UCB and StableOpt. The *second synthetic benchmark* is chosen such that stochastic, worst-case and distributionally robust solutions coincide, with the same choice of $P_t, P^*$ and $\epsilon_t$ as before. See Appendix C, Fig. 4a for a contour plot. Fig. 2 illustrates the results.

Further, we evaluate the methods on real-world wind power data (Data Package Time Series, 2019). Wind power forecasting is an important task (Wang et al., 2011) as power sources that can be effectively scheduled are valuable on the global energy market. In our problem setup, we take hourly recorded wind power data from 2013/14 and use a 48h sliding window to compute an empirical reference distribution for each time step. The decision variable $x$ is the amount of energy that is guaranteed to be delivered in the next hour after the end of the window. The contextual variable $c$ is the actual power generation which we take from the data set. We choose the reward (revenue) function:

$$f(x, c) = 0.1 \max(c - x, 0) + \min(x, c) - 5 \max(x - c, 0)\,.$$

There is a 0.1 reward/energy that was not committed ahead of time, 1 reward/energy for committed energy and $-5$ penalty for committed energy that is not delivered (if the wind generation was too low). For each time step, we use the simulator scenario to compute the robust/stochastic/stable solution; and evaluate the performance on the data set. In Figure 3, we report the cumulative revenue of the different solutions deployed at each time step; this corresponds to the total revenue obtained during the year. The additional baseline is a "zero commitment strategy" ($x_t = 0$). The figure also shows cumulative robust regret. Clearly, the stochastic solution is different from the robust one, hence UCB obtains linear robust regret. In fact, in this case if the DRO objective is solved exactly for each step, the DRBO method would obtain zero robust regret (we compute the solution according to (10) after $T = 100$ steps, therefore an optimization error may remain).

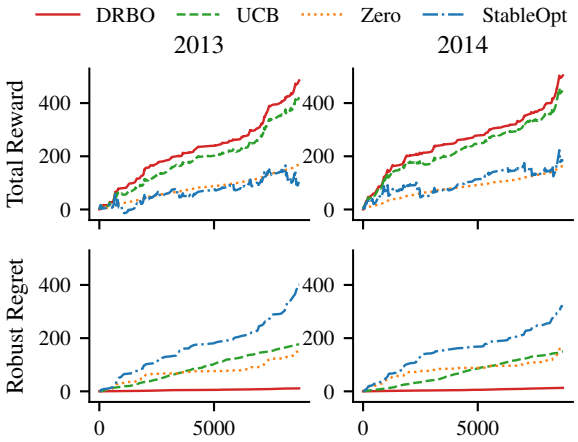



Figure 3: *Wind power prediction.* We show cumulative revenue (top) and robust regret (bottom). StableOpt is too conservative to perform well on either objective. UCB does not account for the distributional shift on the sliding window. By definition, DRBO chooses the robust solution most of the time, therefore achieves (almost) zero robust regret.

## 5 Conclusion

In this work, we introduced and studied distributionally robust Bayesian optimization, where the goal is to be robust against the worst-case contextual distribution among a specified uncertainty set of distributions. Specifically, we focused on uncertainty sets determined by the MMD distance. For a few settings of interest that differ in how the contextual parameter is realized, we provided the first DRBO algorithms with theoretical guarantees. In the experimental study, we demonstrated improvements in terms of robust expected regret over stochastic and worst-case BO baselines.

Our algorithms rely on solving the inner adversary problem, which, in our case, is a linear program with convex constraints. This program can be solved efficiently but is of size $|\mathcal{C}|$, which currently limits the method to relatively small context sets. The formulation and the theory continue to hold for large or continuous context sets, but finding a tractable algorithmic approximation is an interesting direction for future work. Finally, while the considered kernel-based MMD distance fits well with the kernel-based regularity assumptions used in BO, an interesting direction is to extend the ideas to other uncertainty sets used in machine learning, such as the ones defined by $\phi$-divergences and Wasserstein distance. In fact, our approach is still applicable in the case of other divergences, as long as the uncertainty set of distributions is convex and the inner problem can be solved efficiently.

## Acknowledgement

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research, innovation programme grant agreement No 815943, and NSF CAREER award 1553284. IB is supported by ETH Zürich Postdoctoral Fellowship 19-2 FEL-47.

## References


Abbasi-Yadkori, Y. (2013). Online learning for linearly parametrized control problems.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Beland, J. J. and Nair, P. B. (2017). Bayesian optimization under uncertainty. NIPS BayesOpt 2017 workshop.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Bertsimas, D., Gupta, V., and Kallus, N. (2018). Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Bogunovic, I., Scarlett, J., and Cevher, V. (2016a). Time-varying Gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 314–323.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. (2018). Adversarially robust optimization with Gaussian processes. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5760–5770.

Bogunovic, I., Scarlett, J., Krause, A., and Cevher, V. (2016b). Truncated variance reduction: A unified approach to Bayesian optimization and level-set estimation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1507–1515.

Chowdhury, S. R. and Gopalan, A. (2017). On kernelized multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 844–853.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings.

Data Package Time Series, O. (2019). Open power system data. `https://doi.org/10.25832/time_series/2019-06-05`. Version 2019-06-05 (Primary data from various sources, for a complete list see URL).

Djolonga, J., Krause, A., and Cevher, V. (2013). High-dimensional Gaussian process bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1025–1033.

Esfahani, P. M. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166.

Gao, R., Chen, X., and Kleywegt, A. J. (2017). Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

Gardner, J. R., Kusner, M. J., Xu, Z. E., Weinberger, K. Q., and Cunningham, J. P. (2014). Bayesian optimization with inequality constraints. In *ICML*, pages 937–945.

Gelbart, M. A., Snoek, J., and Adams, R. P. (2014). Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*.

Goh, J. and Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations research*, 58(4-part-1):902–917.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Hennig, P. and Schuler, C. J. (2012). Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(Jun):1809–1837.

Kandasamy, K., Schneider, J., and Póczos, B. (2015). High dimensional Bayesian optimisation and bandits via additive models. In *International Conference on Machine Learning (ICML)*, pages 295–304.

Kirschner, J. and Krause, A. (2018). Information directed sampling and bandits with heteroscedastic noise. In *Proc. International Conference on Learning Theory (COLT)*.

Kirschner, J. and Krause, A. (2019). Stochastic bandits with context distributions.

Kirschner, J., Mutnỳ, M., Hiller, N., Ischebeck, R., and Krause, A. (2019). Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. *arXiv preprint arXiv:1902.03229*.

Krause, A. and Ong, C. S. (2011). Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2447–2455.

Lamprier, S., Gisselbrecht, T., and Gallinari, P. (2018). Profile-based bandit with unknown profiles. *The Journal of Machine Learning Research*, 19(1):2060–2099.

Martinez-Cantin, R., Tee, K., and McCourt, M. (2017). Practical Bayesian optimization in the presence of outliers. *arXiv preprint arXiv:1712.04567*.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding

of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Namkoong, H. and Duchi, J. C. (2017). Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980.

Nguyen, T. T., Gupta, S., Ha, H., Rana, S., and Venkatesh, S. (2020). Distributionally robust bayesian quadrature optimization. *arXiv preprint arXiv:2001.06814*.

Nogueira, J., Martinez-Cantin, R., Bernardino, A., and Jamone, L. (2016). Unscented Bayesian optimization for safe robot grasping. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1967–1972. IEEE.

Oliveira, R., Ott, L., and Ramos, F. (2019). Bayesian optimisation under uncertain inputs. *arXiv preprint arXiv:1902.07908*.

Rahimian, H. and Mehrotra, S. (2019). Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Scarf, H. E. (1957). A min-max solution of an inventory problem. Technical report, RAND CORP SANTA MONICA CALIF.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. (2019). No-regret learning in unknown games with correlated payoffs. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Sinha, A., Namkoong, H., and Duchi, J. (2017). Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning (ICML)*, pages 1015–1022.

Staib, M. and Jegelka, S. (2019). Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Staib, M., Wilder, B., and Jegelka, S. (2018). Distributionally robust submodular maximization. *arXiv preprint arXiv:1802.05249*.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2016). Generative models and model criticism via optimized maximum mean discrepancy. *arXiv preprint arXiv:1611.04488*.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*.

Wang, X., Guo, P., and Huang, X. (2011). A review of wind power forecasting models. *Energy procedia*, 12:770–778.

Wang, Z. and Jegelka, S. (2017). Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning (ICML)*, pages 3627–3635.

# A RKHS Regression

Recall that at step $t$, we have data $\mathcal{D}_t = \{(x_1, c_1, y_1), \dots, (x_t, c_t, y_t)\}$. The kernel ridge regression estimate is defined by,

$$\hat{f}_t = \arg\min_{g \in \mathcal{H}} \sum_{i=1}^{t} (g(x_i, c_i) - y_i)^2 + \|g\|_{\mathcal{H}}^2 \ . \tag{12}$$

Denote by $\mathbf{y}_t = [y_1, \dots, y_t]^\top$ the vector of observations, $(K_t)_{i,j=1,\dots,t} = k(x_i, c_i, x_j, c_j)$ the data kernel matrix, and $k_t(x, c) = [k(x, c, c_1, x_1), \dots, k(x, c, x_t, c_t)]^\top$ the data kernel features. We then have

$$\hat{f}_t(x, c) = k_t(x, c)^\top (K_t + \mathbf{1}_t)^{-1} \mathbf{y}_t \ . \tag{13}$$

We further have the posterior variance $\sigma_t(x, c)^2$ that determines the width of the confidence intervals,

$$\sigma_t(x, c)^2 = k(x, c, x, c) - k_t(x, c)^T (K_t + \mathbf{1}_t)^{-1} k_t(x, c) \ . \tag{14}$$

# B Proofs

## B.1 Proof of Theorem 2

The robust cumulative regret is

$$R_T = \sum_{t=1}^{T} \max_{x \in \mathcal{X}} \inf_{Q : d(Q, P_t) \le \epsilon_t} \mathbb{E}_Q[f(x, c)] - \inf_{Q : d(Q, P_t) \le \epsilon_t} \mathbb{E}_Q[f(x_t, c)]. \tag{15}$$

For the proof, we first bound the *instantaneous robust regret*,

$$r_t = \inf_{Q : d(P_t, Q) \le \epsilon_t} \mathbb{E}_Q[f(x_t^*, c)] - \inf_{Q : d(P_t, Q) \le \epsilon_t} \mathbb{E}_Q[f(x_t, c)] \ , \tag{16}$$

where we denote $x_t^* = \arg\max_{x \in \mathcal{X}} \inf_{Q : d(P_t, Q) \le \epsilon_t} \mathbb{E}_Q[f(x, c)]$ the true robust solution at time $t$. We recall the following notation, $f_x = f(x, \cdot)$, $\mathrm{lcb}_x^t = \mathrm{lcb}_t(x, \cdot)$ and $\mathrm{ucb}_x^t = \mathrm{ucb}_t(x, \cdot)$ are vectors in $\mathbb{R}^n$, and $(M)_{i,j} = k(c_i, c_j)$. Further, $w_i = \mathbb{P}_P[c = c_i]$ is a probability vector in $\mathbb{R}^n$, where $n$ is used to denote the size of the contextual set, i.e., $n = |\mathcal{C}|$. With this, note that

$$\inf_{Q : d(P_t, Q) \le \epsilon_t} \mathbb{E}[f(x, c)] = \inf_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w_t\|_M \le \epsilon_t}} \langle w', f_x \rangle \tag{17}$$

The solution to this linear program is the worst case distribution over $c$ if we choose action $x$. Define worst-case distributions $w_x^f$, $w_x^{\mathrm{lcb}_t}$ and $w_x^{\mathrm{ucb}_t}$ for exact, optimistic and pessimistic function values (the dependence on $t$ is implicit),

$$w_x^f = \arg\min_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w_t\|_M \le \epsilon_t}} \langle w', f_x \rangle, \qquad w_x^{\mathrm{lcb}_t} = \arg\min_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w_t\|_M \le \epsilon_t}} \langle w', \mathrm{lcb}_x^t \rangle, \qquad w_x^{\mathrm{ucb}_t} = \arg\min_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w_t\|_M \le \epsilon_t}} \langle w', \mathrm{ucb}_x^t \rangle \ . \tag{18}$$

By combining (8) and (18), we can upper and lower bound the objective as follows:

$$\langle w_x^{\mathrm{lcb}_t}, \mathrm{lcb}_x^t \rangle \le \inf_{\substack{w' : \|w'\|_1 = 1, \\ 0 \le w'_j \le 1 \ \forall j \in [n], \\ \|w' - w_t\|_M \le \epsilon_t}} \langle w', f_x \rangle \le \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t \rangle. \tag{19}$$

Recall that Algorithm 1 takes actions $x_t = \arg\max_{x \in \mathcal{X}} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t \rangle$, and note that $\|w_{x_t}^{\mathrm{lcb}_t} - w_t^*\|_M \le \epsilon_t$ where $(w_t^*)_i = \mathbb{P}_{P_t^*}[c = c_i]$ is the probability vector from the true sampling distribution at time $t$. For any $x \in \mathcal{X}$, we proceed to bound the instantaneous regret,

$$r_t = \inf_{Q:d(P,Q)\leq\epsilon} \mathbb{E}_Q[f(x_t^*, c)] - \inf_{Q:d(P,Q)\leq\epsilon} \mathbb{E}_Q[f(x_t, c)] \tag{20}$$

$$\overset{(i)}{\leq} \langle w_{x_t^*}^{\mathrm{ucb}_t}, \mathrm{ucb}_{x_t^*}^t \rangle - \langle w_{x_t}^f, f_{x_t} \rangle \tag{21}$$

$$\overset{(ii)}{\leq} \langle w_{x_t}^{\mathrm{ucb}_t}, \mathrm{ucb}_{x_t}^t \rangle - \langle w_{x_t}^f, f_{x_t} \rangle \tag{22}$$

$$\overset{(iii)}{\leq} \langle w_t^*, \mathrm{ucb}_{x_t}^t \rangle - \langle w_{x_t}^f, f_{x_t} \rangle \tag{23}$$

$$= \langle w_t^*, f_{x_t} \rangle + \langle w_t^*, \mathrm{ucb}_{x_t}^t - f_{x_t} \rangle - \langle w_{x_t}^f, f_{x_t} \rangle \tag{24}$$

$$= \langle w_t^*, \mathrm{ucb}_{x_t}^t - f_{x_t} \rangle + \langle w_t^* - w_{x_t}^f, f_{x_t} \rangle \tag{25}$$

$$\overset{(iv)}{\leq} 2\beta_t \langle w_t^*, \sigma_t(x_t, \cdot) \rangle + \|w_t^* - w_{x_t}^f\|_M \|f_{x_t}\|_{M^{-1}} \tag{26}$$

$$\overset{(v)}{\leq} 2\beta_t \langle w_t^*, \sigma_t(x_t, \cdot) \rangle + 2\epsilon_t \|f_{x_t}\|_{M^{-1}} \tag{27}$$

$$\leq 2\beta_T \langle w_t^*, \sigma_t(x_t, \cdot) \rangle + 2\epsilon_t B' \ . \tag{28}$$

Here, (i) follows from the definition of the upper bound (19), (ii) is by the choice of $x_t$, (iii) follows from the fact that $w_{x_t}^{\mathrm{ucb}_t}$ is a minimizer, (iv) uses again the confidence bounds and (v) follows from $d_{x_t}(P^*, Q) \leq \epsilon_t$. Finally, the following holds for the instantaneous regret: $r_t = \max_{x\in\mathcal{X}} r_t(x) \leq 2\beta_T \langle w_t^*, \sigma_t(x_t, \cdot) \rangle + 2\epsilon_t B'$.

We now continue to bound the cumulative regret $R_T = \sum_{t=1}^T r_t$. To this end, we first apply the Cauchy-Schwarz inequality as in the standard proof, and then Jensen's inequality to find,

$$R_T \leq 2\beta_T \sum_{t=1}^T \langle w_t^*, \sigma_t(x_t, \cdot) \rangle + 2B' \sum_{t=1}^T \epsilon_t \tag{29}$$

$$\leq 2\beta_T \sqrt{T \sum_{t=1}^T \langle w_t^*, \sigma_t(x_t, \cdot) \rangle^2} + 2B' \sum_{t=1}^T \epsilon_t \tag{30}$$

$$\leq 2\beta_T \sqrt{T \sum_{t=1}^T \langle w_t^*, \sigma_t(x_t, \cdot)^2 \rangle} + 2B' \sum_{t=1}^T \epsilon_t \ . \tag{31}$$

To complete the proof, we need to relate $\sum_{t=1}^T \langle w_t^*, \sigma_t(x_t, \cdot) \rangle^2$ to the observed posterior variance $\sum_{t=1}^T \sigma_t(x_t, c_t)^2$. For this, we apply Lemma 7 below. Note that $\sigma_t(x_t, c_t)^2 \leq 1$ by our assumption that $k(x, c, x', c') \leq 1$. Hence, w.p. at least $1 - \delta$,

$$\sum_{t=1}^T \langle w_t^*, \sigma_t(x_t, \cdot)^2 \rangle \leq 2 \sum_{t=1}^T \sigma_t(x_t, c_t)^2 + 8 \log\left(\frac{6}{\delta}\right) \tag{32}$$

Finally, using that $x \leq 2\alpha \log(1 + x)$ for all $x \in [0, \alpha]$,

$$\sum_{t=1}^T \sigma_t(x_t, c_t)^2 \leq \sum_{t=1}^T 2 \log\left(1 + \sigma_t(x_t, c_t)^2\right) \leq 2\gamma_T \ . \tag{33}$$

The last inequality follows from Lemma 3 in Chowdhury and Gopalan (2017). The final regret bound is therefore,

$$R_T \leq 4\beta_T \sqrt{T\left(\gamma_T + 4 \log\left(\tfrac{6}{\delta}\right)\right)} + 2B' \sum_{t=1}^T \epsilon_t, \tag{34}$$

A union bound over the events such that both Lemma 1 and Lemma 7 hold, yields probability $\geq 1 - 2\delta$ for the complete statement and completes the proof.

**Lemma 7** (Concentration of conditional mean, Lemma 3 in Kirschner and Krause (2018)). *Let $S_t \geq 0$ be non-negative stochastic process adapted to a filtration $\{\mathcal{F}_t\}$, and define $m_t = \mathbb{E}[S_t|\mathcal{F}_{t-1}]$. Further assume that $S_t \leq B$ for $B \geq 1$. Then, for any $T \geq 1$, with probability at least $1-\delta$ it holds that,*

$$\sum_{t=1}^{T} m_t \leq 2\sum_{t=1}^{T} S_t + 4B\log\frac{1}{\delta} + 8B\log(4B) + 1$$
$$\leq 2\sum_{t=1}^{T} S_t + 8B\log\frac{6B}{\delta}$$

### B.2 Proof of Corollary 4

Note that with $\delta_t = \frac{2\delta}{\pi^2 t^2}$, the result from Lemma 3 holds with probability at least $1-\delta/3$ by the union bound over all steps $t = 1, 2, 3\ldots$, i.e.

$$d(\hat{P}_t, P^*) \leq \frac{1}{\sqrt{t}}\left(2 + \sqrt{2\log\frac{\pi^2 t^2}{2\delta}}\right) = \epsilon_t \tag{35}$$

Therefore, with probability at least $1-\delta/3$, $P^* \in \mathcal{U}_t$. The results follows from Theorem 2 and another application of the union bound. Finally we use that $\sum_{t=1}^{T} t^{-1/2} \leq 2\sqrt{T}$ to complete the proof of the corollary.

### B.3 Proof of Theorem 5

Recall that Algorithm 2 takes the actions $x_t = \arg\max_{x\in\mathcal{X}} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle$ and $c_t = \arg\max_{c\in\mathcal{C}} \sigma_t(x_t, c)$. We begin to bound $r_t$ similar as in the proof of the general regret bound.

$$r_t(x) = \inf_{Q:d(P_t,Q)\leq\epsilon_t} \mathbb{E}_Q[f(x,c)] - \inf_{Q:d(P_t,Q)\leq\epsilon_t} \mathbb{E}_Q[f(x_t,c)] \tag{36}$$
$$\overset{(i)}{\leq} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle - \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{lcb}_{x_t}^t\rangle \tag{37}$$
$$\overset{(ii)}{\leq} \langle w_{x_t}^{\mathrm{ucb}_t}, \mathrm{ucb}_{x_t}^t\rangle - \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{lcb}_{x_t}^t\rangle \tag{38}$$
$$\overset{(iii)}{\leq} \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{ucb}_{x_t}^t\rangle - \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{lcb}_{x_t}^t\rangle \tag{39}$$
$$\overset{(iv)}{\leq} 2\beta_t\sigma_t(x_t, c_t) \tag{40}$$

Here, as before, (i) replaces the function values by over/under-estimated values of the upper/lower confidence bounds, (ii) uses the choice of the UCB action and (iii) uses that $w_{x_t}^{\mathrm{ucb}_t}$ is a minimizer of $\langle w_{x_t}^{\mathrm{ucb}_t}, \mathrm{ucb}_{x_t}^t\rangle$. The last inequality (iv) uses that $c_t$ maximizes $\sigma_t(x_t, c_t)$ as well as that $w_{x_t}^{\mathrm{lcb}_t}$ is a probability vector. With that, the cumulative regret bound follows via the standard argument.

### B.4 Proof of Corollary 6

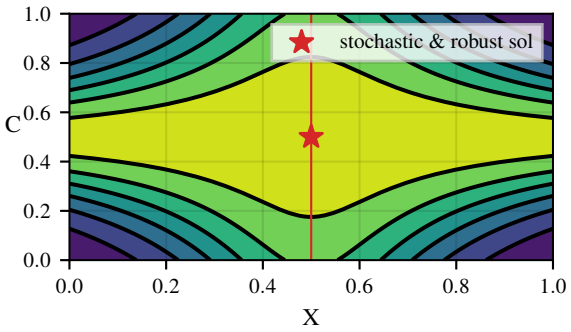
To bound the simple regret, recall that $\hat{t} = \arg\max_{t=1,\ldots,T} \inf_Q \mathbb{E}_Q[\mathrm{lcb}_t(x_t, c)]$ and $\hat{x}_T = x_{\hat{t}}$. For any $t = 1, \ldots, T$ we have that,

$$r_T = \max_{x\in\mathcal{X}} \inf_{Q:d(Q,P)\leq\epsilon} \mathbb{E}_Q f(x,c) - \inf_{Q:d(Q,P)\leq\epsilon} \mathbb{E}_Q f(\hat{x}_T, c) \tag{41}$$
$$\overset{(i)}{\leq} \max_{x\in\mathcal{X}} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle - \langle w_{\hat{x}_T}^{\mathrm{lcb}_{\hat{t}}}, \mathrm{lcb}_{\hat{x}_T}^{\hat{t}}\rangle \tag{42}$$
$$\overset{(ii)}{=} \max_{x\in\mathcal{X}} \langle w_x^{\mathrm{ucb}_t}, \mathrm{ucb}_x^t\rangle - \max_{s=1,\ldots,T} \langle w_{x_s}^{\mathrm{lcb}_s}, \mathrm{lcb}_{x_s}^s\rangle \tag{43}$$
$$\overset{(iii)}{\leq} \langle w_{x_t}^{\mathrm{ucb}_t}, \mathrm{ucb}_{x_t}^t\rangle - \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{lcb}_{x_t}^t\rangle \tag{44}$$
$$\overset{(iv)}{\leq} \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{ucb}_{x_t}^t\rangle - \langle w_{x_t}^{\mathrm{lcb}_t}, \mathrm{lcb}_{x_t}^t\rangle \tag{45}$$
$$\overset{(v)}{\leq} 2\beta_t\sigma_t(x_t, c_t) \tag{46}$$

Here, (i) bounds the function values by the upper and lower confidence bound respectively, (ii) uses the definition of $\hat{x}_T$, (iii) uses the definition of the UCB action and drops the maximum and finally, (iv,v) as before uses the choices of $x_t$, $c_t$, $w_{x_t}^{\mathrm{ucb}_t}$ and that $w_{x_t}^{\mathrm{lcb}_t}$ is a probability vector. With this we are able to leverage the cumulative regret bound, and find

$$r_T \leq \frac{1}{T}\sum_{t=1}^{T} 2\beta_t \sigma_t(x_t, c_t) \leq \frac{R_T}{T} \ . \tag{47}$$

## C Details on the Experiments




(a) Second synthetic benchmark. Here stochastic and robust solutions coincide at the marked design.