Distributionally Robust Bayesian Optimization with φ-divergences

Hisham Husain
Amazon
hushisha@amazon.com

Vu Nguyen
Amazon
vutngn@amazon.com

Anton van den Hengel
Amazon
hengelah@amazon.com

Abstract

The study of robustness has received much attention due to its inevitability in data-driven settings where many systems face uncertainty. One such example of concern is Bayesian Optimization (BO), where uncertainty is multi-faceted, yet there only exists a limited number of works dedicated to this direction. In particular, there is the work of Kirschner et al. [26], which bridges the existing literature of Distributionally Robust Optimization (DRO) by casting the BO problem from the lens of DRO. While this work is pioneering, it admittedly suffers from various practical shortcomings such as finite contexts assumptions, leaving behind the main question *Can one devise a computationally tractable algorithm for solving this DRO-BO problem*? In this work, we tackle this question to a large degree of generality by considering robustness against data-shift in φ-divergences, which subsumes many popular choices, such as the χ^2-divergence, Total Variation, and the extant Kullback-Leibler (KL) divergence. We show that the DRO-BO problem in this setting is equivalent to a finite-dimensional optimization problem which, even in the continuous context setting, can be easily implemented with provable sublinear regret bounds. We then show experimentally that our method surpasses existing methods, attesting to the theoretical results.

1 Introduction

Bayesian Optimization (BO) [29, 25, 52, 49, 34] allows us to model a black-box function that is expensive to evaluate, in the case where noisy observations are available. Many important applications of BO correspond to situations where the objective function depends on an additional context parameter [27, 57], for example in health-care, recommender systems can be used to model information about a certain type of medical domain. BO has naturally found success in a number of scientific domains [56, 20, 30, 18, 55] and also a staple in machine learning for the crucial problem of hyperparameter tuning [44, 36, 40, 41, 59].

As with all data-driven approaches, BO is prone to cases where the given data *shifts* from the data of interest. While BO models this in the form of Gaussian noise for the inputs to the objective function, the context distribution is assumed to be consistent. This can be problematic, for example in healthcare where patient information shifts over time. This problem exists in the larger domain of operations research under the banner of *distributionally robust optimization* (DRO) [46], where one is interested in being *robust* against shifts in the distribution observed. In particular, for a given *distance* between distributions D, DRO studies robustness against adversaries who are allowed to modify the observed distribution p to another distribution in the set:

$$\{q : \mathsf{D}(p, q) \leq \varepsilon\},$$

for some $\varepsilon > 0$. One can interpret this as a ball of radius ε for the given choice of D and the adversary perturbs the observed distribution p to q where ε is a form of "budget".

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

https://doi.org/10.52202/075280-0883

Distributional shift is a topical problem in machine learning and the results of DRO have been specialized in the context of supervised learning [12, 13, 11, 10, 7, 16, 22], reinforcement learning [21] and Bayesian learning [51], as examples. One of the main challenges however is that the DRO is typically intractable since in the general setting of continuous contexts, involves an infinite dimensional constrained optimization problem. The choice of D is crucial here as various choices such as the Wasserstein distance [6, 7, 9, 48], Maximum Mean Discrepancy (MMD) [53] and φ-divergences [1] [12, 13] allow for computationally tractable regimes. In particular, these specific choices of D have shown intimate links between regularization [22] which is a conceptually central topic of machine learning.

More recently however, DRO has been studied for the BO setting in Kirschner et al. [26], which as one would expect, leads to a complicated minimax problem, which causes a computational burden practically speaking. Kirschner et al. [26] makes the first step and casts the formal problem however develops an algorithm only in the case where D has been selected as the MMD. While, this work makes the first step and conceptualizes the problem of distributional shifts in context for BO, there are two main practical short-comings. Firstly, the algorithm is developed specifically to the MMD, which is easily computed, however cannot be replaced by another choice of D whose closed form is not readily accessible with samples such as the φ-divergence. Secondly, the algorithm is only tractable when the contexts are finite since at every iteration of BO, it requires solving an M-dimensional problem where M is the number of contexts.

The main question that remains is, *can we devise an algorithm that is computationally tractable for tackling the DRO-BO setting*? We answer this question to a large degree of generality by considering distributional shifts against φ-divergences - a large family of divergences consisting of the extant Kullback-Leibler (KL) divergence, Total Variation (TV) and χ^2-divergence, among others. In particular, we exploit existing advances made in the large literature of DRO to show that the BO objective in this setting for any choice of φ-divergence yields a computationally tractable algorithm, even for the case of continuous contexts. We also present a robust regret analysis that illustrates a sublinear regret. Finally, we show, along with computational tractability, that our method is empirically superior on standard datasets against several baselines including that of Kirschner et al. [26]. In summary, our main contributions are

1. A theoretical result showing that the minimax distributionally robust BO objective with respect to φ divergences is equivalent to a single minimization problem.
2. An efficient algorithm, that works in the continuous context regime, for the specific cases of the χ^2-divergence and TV distance, which admits a conceptually interesting relationship to regularization of BO.
3. A regret analysis that specifically informs how we can choose the DRO ε-budget to attain sublinear regret.

2 Related Work

Due to the multifaceted nature of our contribution, we discuss two streams of related literature, one relating to studies of robustness in Bayesian Optimization (BO) and one relating to advances in Distributionally Robust Optimization (DRO).

In terms of BO, the work most closest to ours is Kirschner et al. [26] which casts the distributionally robust optimization problem over contexts. In particular, the work shows how the DRO objective for any choice of divergence D can be cast, which is exactly what we build off. The main drawback of this method however is the limited practical setting due to the expensive inner optimization, which heavily relies on the MMD, and therefore cannot generalize easily to other divergences that are not available in closed forms. Our work in comparison, holds for a much more general class of divergences, and admits a practical algorithm that involves a finite dimensional optimization problem. In particular, we derive the result when D is chosen to be the χ^2-divergence which we show performs the best empirically. This choice of divergence has been studied in the related problem of Bayesian quadrature [33], and similarly illustrated strong performance, complimenting our results. There also exists work of BO that aim to be robust by modelling adversaries through noise, point estimates or non-cooperative games [37, 32, 3, 39, 47]. The main difference between our work and theirs is

[1] as known as f-divergences in the literature

https://doi.org/10.52202/075280-0883

that the notion of robustness we tackle is at the *distributional* level. Another similar work to ours is that of Tay et al. [54] which considers approximating DRO-BO using Taylor expansions based on the sensitivity of the function. In some cases, the results coincide with ours however their result must account for an approximation error in general. Furthermore, an open problem as stated in their work is to solve the DRO-BO problem for continuous context domains, which is precisely one of the advantages of our work.

From the perspective of DRO, our work essentially is an extension of Duchi et al. [12, 13] which develops results that connect φ-divergence DRO to variance regularization. In particular, they assume φ admits a continuous second derivative, which allows them to connect the φ-divergence to the χ^2-divergence and consequently forms a general connection to constrained variance. While the work is pioneering, this assumption leaves out important φ-divergences such as the Total Variation (TV) - a choice of divergence which we illustrate performs well in comparison to standard baselines in BO. At the technical level, our derivations are similar to Ahmadi-Javid [2] however our result, to the best of our knowledge, is the first such work that develops it in the context of BO. In particular, our results for the Total Variation and χ^2-divergence show that variance is a key penalty in ensuring robustness which is a well-known phenomena existing in the realm of machine learning [12, 11, 10, 22, 1].

3 Preliminaries

Bayesian Optimization We consider optimizing a *black-box* function, $f : \mathcal{X} \to \mathbb{R}$ with respect to the *input* space $\mathcal{X} \subseteq \mathbb{R}^d$. As a black-box function, we do not have access to f directly however receive input in a sequential manner: at time step t, the learner chooses some input $\mathbf{x}_t \in \mathcal{X}$ and observes the *reward* $y_t = f(\mathbf{x}_t) + \eta_t$ where the noise $\eta_t \sim \mathcal{N}(0, \sigma_f^2)$ and σ_f^2 is the output noise variance. Therefore, the goal is to optimize

$$\sup_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}).$$

Additional to the input space $\mathcal{X}$, we introduce the *context* spaces $\mathcal{C}$, which we assume to be compact. These spaces are assumed to be separable completely metrizable topological spaces.[2] We have a reward function, $f : \mathcal{X} \times \mathcal{C} \to \mathbb{R}$ which we are interested in optimizing with respect to $\mathcal{X}$. Similar to sequential optimization, at time step t the learner chooses some input $\mathbf{x}_t \in \mathcal{X}$ and receives a context $c_t \in \mathcal{C}$ and $f(\mathbf{x}_t, c_t) + \eta_t$. Here, the learner can not choose a context c_t, but receive it from the environment. Given the context information, the objective function is written as

$$\sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{c \sim p}[f(\mathbf{x}, c)],$$

where p is a probability distribution over contexts.

Gaussian Processes We follow a popular choice in BO [49] to use GP as a surrogate model for optimizing f. A GP [43] defines a probability distribution over functions f under the assumption that any subset of points $\{\mathbf{x}_i, f(\mathbf{x}_i)\}$ is normally distributed. Formally, this is denoted as:

$$f(\mathbf{x}) \sim \mathrm{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right),$$

where $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are the mean and covariance functions, given by $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ and $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^T\right]$. For predicting $f_* = f(\mathbf{x}_*)$ at a new data point $\mathbf{x}_*$, the conditional probability follows a univariate Gaussian distribution as $p\big(f_* \mid \mathbf{x}_*, [\mathbf{x}_1 ... \mathbf{x}_N], [y_1, ... y_N]\big) \sim \mathcal{N}\left(\mu(\mathbf{x}_*), \sigma^2(\mathbf{x}_*)\right)$. Its mean and variance are given by:

$$\mu(\mathbf{x}_*) = \mathbf{k}_{*,N} \mathbf{K}_{N,N}^{-1} \mathbf{y}, \qquad (1) \qquad \sigma^2(\mathbf{x}_*) = k_{**} - \mathbf{k}_{*,N} \mathbf{K}_{N,N}^{-1} \mathbf{k}_{*,N}^T \qquad (2)$$

where $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$, $\mathbf{k}_{*,N} = [k(\mathbf{x}_*, \mathbf{x}_i)]_{\forall i \leq N}$ and $\boldsymbol{K}_{N,N} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{\forall i,j \leq N}$. As GPs give full uncertainty information with any prediction, they provide a flexible nonparametric prior for Bayesian optimization. We refer to Rasmussen and Williams [43] for further details on GPs.

Distributional Robustness Let $\Delta(\mathcal{C})$ denote the set of probability distributions over $\mathcal{C}$. A *divergence* between distributions $\mathsf{D} : \Delta(\mathcal{C}) \times \Delta(\mathcal{C}) \to \mathbb{R}$ is a dissimilarity measure that satisfies

[2] We remark that this is an extremely mild condition, satisfied by the large majority of considered examples.

https://doi.org/10.52202/075280-0883

$\Delta(p, q) \geq 0$ with equality if and only if $p = q$ for $p, q \in \Delta(\mathcal{C})$. For a function, $h : \mathcal{C} \to \mathbb{R}$, base probability measure $p \in \Delta(\mathcal{C})$, the central concern of Distributionally Robust Optimization (DRO) [4, 42, 5] is to compute

$$\sup_{q \in B_{\varepsilon,\mathsf{D}}(p)} \mathbb{E}_{q(c)}[h(c)], \tag{3}$$

where $B_{\varepsilon,\mathsf{D}}(p) = \{q \in \Delta(\mathcal{C}) : \mathsf{D}(p, q) \leq \varepsilon\}$, is ball of distributions q that are ε away from p with respect to the divergence D. The objective in Eq. (3) is intractable, especially in setting where $\mathcal{C}$ is continuous as it amounts to a constrained infinite dimensional optimization problem. It is also clear that the choice of D is crucial for both computational and conceptual purposes. The vast majority of choices typically include the Wasserstein due to the transportation-theoretic interpretation and with a large portion of existing literature finding connections to Lipschitz regularization [6, 7, 9, 48]. Other choices where they have been studied in the supervised learning setting include the Maximum Mean Discrepancy (MMD) [53] and φ-divergences [12, 13].

Distributionally Robust Bayesian Optimization Recently, the notion of DRO has been applied to BO [26, 54], who consider robustness with respect to shifts in the context space and therefore are interested in solving

$$\sup_{\mathbf{x} \in \mathcal{X}} \inf_{q \in B_{\varepsilon,D}(p)} \mathbb{E}_{c \sim q}[f(\mathbf{x}, c)],$$

where p is the reference distribution. This objective becomes significantly more difficult to deal with since not only does it involve a constrained and possibly infinite dimensional optimization problem however also involves a minimax which can cause instability issues if solved iteratively.

Kirschner et al. [26] tackle these problems by letting D be the kernel Maximum Mean Discrepancy (MMD), which is a popular choice of discrepancy motivated by kernel mean embeddings [19]. In particular, the MMD can be efficiently estimated in $O(n^2)$ where n is the number of samples. Naturally, this has two main drawbacks: The first is that it is still computationally expensive since one is required to solve two optimization problems, which can lead to instability and secondly, the resulting algorithm is limited to the scheme where the number of contexts is finite. In our work, we consider D to be a φ-divergence, which includes the Total Variance, χ^2 and Kullback-Leibler (KL) divergence and furthermore show that minmax objective can be reduced to a single maximum optimization problem which resolves both the instability and finiteness assumption. In particular, we also present a similar analysis, showing that the robust regret decays sublinearly for the right choices of radii.

4 φ-Robust Bayesian Optimization

In this section, we present the main result on distributionally robustness when applied to BO using φ-divergence. Therefore, we begin by defining this key quantity.

Definition 1 (φ-divergence) *Let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be a convex, lower semi-continuous function such that $\varphi(1) = 0$. The φ-divergence between $p, q \in \Delta(\mathcal{C})$ is defined as*

$$\mathsf{D}_\varphi(p, q) = \mathbb{E}_{q(c)}\left[\varphi\left(\frac{dp}{dq}(c)\right)\right],$$

where dp/dq is the Radon-Nikodym derivative if $p \ll q$ and $\mathsf{D}_\varphi(p, q) = +\infty$ otherwise.

Popular choices of the convex function φ include $\varphi(u) = (u-1)^2$ which yields the χ^2 and, $\varphi(u) = |u-1|$, $\varphi(u) = u \log u$ which correspond to the χ^2 and KL divergences respectively. At any time step $t \geq 1$, we consider distributional shifts with respect to an φ-divergence for any choice of φ and therefore relevantly define the DRO ball as

$$B_\varphi^t(p_t) := \{q \in \Delta(\mathcal{C}) : \mathsf{D}_\varphi(q, p_t) \leq \varepsilon_t\},$$

where $p_t = \frac{1}{t}\sum_{s=1}^t \delta_{c_s}$ is the reference distribution and ε_t is the distributionally robust radius chosen at time t. We remark that for our results, the choice of p_t is flexible and can be chosen based on the specific domain application. The φ divergence, as noted from the definition above, is only defined

https://doi.org/10.52202/075280-0883

finitely when the measures p, q are absolutely continuous to each other and there is regarded as a *strong* divergence in comparison to the Maximum Mean Discrepancy (MMD), which is utilized in Kirschner et al. [26]. The main consequence of this property is that the geometry of the ball B_φ^t would differ based on the choice of φ-divergence. The φ-divergence is a very popular choice for defining this ball in previous studies of DRO in the context of supervised learning due to the connections and links it has found to variance regularization [12, 13, 11].

We will exploit various properties of the φ-divergence to derive a result that reaps the benefits of this choice such as a reduced optimization problem - a development that does not currently exist for the MMD [26]. We first define the convex conjugate of φ as $\varphi^\star(u) = \sup_{u' \in \text{dom}_\varphi} (u \cdot u' - \varphi(u'))$, which we note is a standard function that is readily available in closed form for many choices of φ.

Theorem 1 *Let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be a convex lower semicontinuous mapping such that $\varphi(1) = 0$. Let f be measurable and bounded. For any $\varepsilon > 0$, it holds that*

$$\sup_{\mathbf{x} \in \mathcal{X}} \inf_{q \in B_\varphi^t(p)} \mathbb{E}_{c \sim q}[f(\mathbf{x}, c)] = \sup_{\mathbf{x} \in \mathcal{X}, \lambda \geq 0, b \in \mathbb{R}} \left(b - \lambda \varepsilon_t - \lambda \mathbb{E}_{p_t(c)} \left[\varphi^\star \left(\frac{b - f(\mathbf{x}, c)}{\lambda} \right) \right] \right).$$

Proof (Sketch) The proof begins by rewriting the constraint over the φ-divergence constrained ball with the use of Lagrangian multipliers. Using existing identities for f-divergences, a minimax swap yields a two-dimensional optimization problem, over $\lambda \geq 0$ and $b \in \mathbb{R}$.

We remark that similar results exist for other areas such as supervised learning [50], robust optimization [4] and certifying robust radii [14]. However this is, to the best of our knowledge, the first development when applied to optimizing expensive black-box functions, the case of BO. The above Theorem is practically compelling for three main reasons. First, one can note that compared to the left-hand side, the result converts this into a single optimization (max) over three variables, where two of the variables are 1-dimensional, reducing the computational burden significantly. Secondly, the notoriously difficult max-min problem becomes only a max, leaving behind instabilities one would encounter with the former objective. Finally, the result makes very mild assumptions on the context parameter space $\mathcal{C}$, allowing infinite spaces to be chosen, which is one of the challenges for existing BO advancements. We show that for specific choices of φ, the optimization over b and even λ can be expressed in closed form and thus simplified. All proofs for the following examples can be found in the Appendix Section 8.

Example 2 (χ^2-divergence) *Let $\varphi(u) = (u-1)^2$, then for any measurable and bounded f we have for any choice of ε_t*

$$\sup_{\mathbf{x} \in \mathcal{X}} \inf_{q \in B_\varphi^t(p_t)} \mathbb{E}_{c \sim q}[f(\mathbf{x}, c)] = \sup_{\mathbf{x} \in \mathcal{X}} \left(\mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \sqrt{\varepsilon_t \cdot \text{Var}_{p_t(c)}[f(\mathbf{x}, c)]} \right).$$

The above example can be easily implemented as it involves the same optimization problem however now appended with a variance term. Furthermore, this objective admits a compelling conceptual insight which is that, by enforcing a penalty in the form of variance, one attains robustness. The idea that regularization provides guidance to robustness or generalization is well-founded in machine learning more generally for example in supervised learning [12, 13]. We remark that this penalty and its relationship to χ^2-divergence has been developed in the similar yet related problem of Bayesian quadrature [33]. Moreover, it can be shown that if φ is twice differentiable then D_φ can be approximated by the χ^2-divergence via Taylor series, which makes χ^2-divergence a centrally appealing choice for studying robustness. We now derive the result for a popular choice of φ that is not differentiable.

Example 3 (Total Variation) *Let $\varphi(u) = |u-1|$, then for any measurable and bounded f we have for any choice of ε_t*

$$\sup_{\mathbf{x} \in \mathcal{X}} \inf_{q \in B_\varphi^t(p_t)} \mathbb{E}_{c \sim q}[f(\mathbf{x}, c)] = \sup_{\mathbf{x} \in \mathcal{X}} \left(\mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \frac{\varepsilon_t}{2} \left(\sup_{c \in \mathcal{C}} f(\mathbf{x}, c) - \inf_{c \in \mathcal{C}} f(\mathbf{x}, c) \right) \right).$$

Similar to the χ^2-case, the result here admits a variance-like term in the form of the difference between the maximal and minimal elements. We remark that such a result is conceptually interesting

https://doi.org/10.52202/075280-0883

since both losses admit an objective that resembles a mean-variance which is a natural concept in ML, but advocates for it from the perspective of distributional robustness. This result exists for the supervised learning in Duchi and Namkoong [11] however is completely novel for BO and also holds for a choice of non-differentiable φ, hinting at the deeper connection between φ-divergence DRO and variance regularization.

4.1 Optimization with the GP Surrogate

To handle the distributional robustness, we have rewritten the objective function using φ divergences in Theorem 1. In DRBO setting, we sequentially select a next point $\mathbf{x}_t$ for querying a black-box function. Given the observed context $c_t \sim q$ coming from the environment, we evaluate the black-box function and observe the output as $y_t = f(\mathbf{x}_t, c_t) + \eta_t$ where the noise $\eta_t \sim \mathcal{N}(0, \sigma_f^2)$ and σ_f^2 is the noise variance.

As a common practice in BO, at the iteration t, we model the GP surrogate model using the observed data $\{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$ and make a decision by maximizing the acquisition function which is build on top of the GP surrogate:

$$\mathbf{x}_t = \arg\max_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x}).$$

Algorithm 1 DRBO with φ-divergence

1: **Input:** Max iteration T, initial data D_0, η
2: **for** $t = 1, \ldots, T$ **do**
3: Fit and estimate GP hyperparameter given D_{t-1}
4: Select a next input $\mathbf{x}_t = \arg\max \alpha(\mathbf{x})$
5: χ^2-divergence: $\alpha(\mathbf{x}) := \alpha^{\chi^2}(\mathbf{x})$ from Eq. (4)
6: Total Variation: $\alpha(\mathbf{x}) := \alpha^{TV}(\mathbf{x})$ from Eq. (5)
7: Observe a context $c_t \sim q$
8: Evaluate the black-box $y_t = f(\mathbf{x}_t, c_t) + \eta_t$
9: Augment $D_t = D_{t-1} \cup (\mathbf{x}_t, c_t, y_t)$
10: **end for**

While our method is not restricted to the form of the acquisition function, for convenience in the theoretical analysis, we follow the GP-UCB [52]. Given the GP predictive mean and variance from Eqs. (7,8), we have the acquisition function for the χ^2 in Example 2 as follows:

$$\alpha^{\chi^2}(\mathbf{x}) := \frac{1}{|C|}\sum_c \Big[\mu_t(\mathbf{x}, c) + \sqrt{\beta_t}\sigma_t(\mathbf{x}, c)\Big] - \sqrt{\frac{\varepsilon_t}{|C|}\sum_c \big(\mu_t(\mathbf{x}, c) - \bar{\mu}_t\big)^2} \tag{4}$$

where β_t is a explore-exploit hyperparameter defined in Srinivas et al. [52], $\bar{\mu}_t = \frac{1}{|C|}\sum_c \mu_t(\mathbf{x}, c)$ and $c \sim q$ can be generated in a one dimensional space to approximate the expectation and the variance. In the experiment, we select q as the uniform distribution, but it is not restricted to. Similarly, an acquisition function for Total Variation in Example 3 is written as

$$\alpha^{TV}(\mathbf{x}) := \frac{1}{|C|}\sum_c \Big[\mu_t(\mathbf{x}, c) + \sqrt{\beta_t}\sigma_t(\mathbf{x}, c)\Big] - \frac{\varepsilon_t}{2}\big(\max \mu_t(\mathbf{x}, c) - \min \mu_t(\mathbf{x}, c)\big). \tag{5}$$

We summarize all computational steps in Algorithm 1.

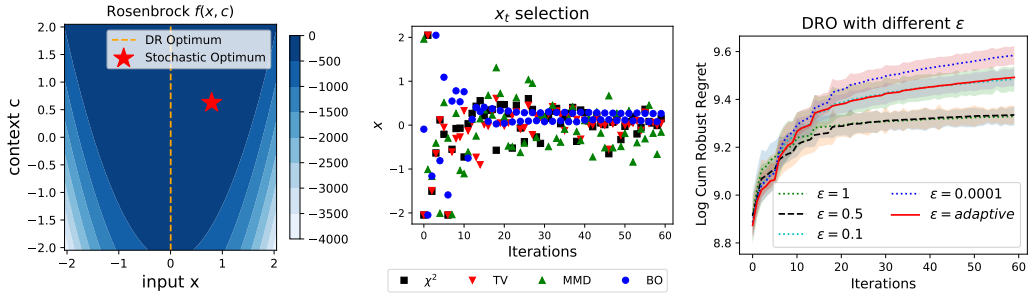
Computational Efficiency against MMD. We make an important remark that since we do not require our context space to be finite, our implementation scales only linearly with the number of context samples $|C|$ drawing from q. This allows us to discretize our space and draw as many context samples as required while only paying a linear price. On the other hand, the MMD [26] at every iteration of t requires solving an $|C|$-dimensional constraint optimization problem that has no closed form solution. We refer to Section 5.2 for the empirical comparison.

4.2 Convergence Analysis

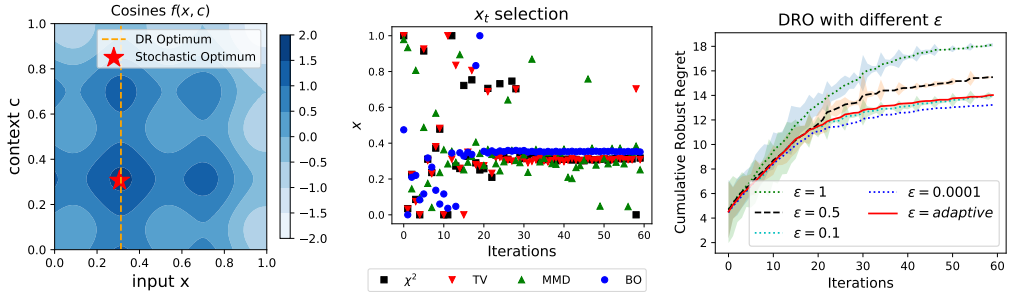
One of the main advantages of Kirschner et al. [26] is the choice of MMD makes the regret analysis simpler due to the nice structure and properties of MMD. In particular, the MMD is well-celebrated for a $O(t^{-1/2})$ convergence where no such results exist for φ-divergences. However, using Theorem 1, we can show a regret bound for the Total Variation with a simple boundedness assumption and show how one can extend this result to other φ-divergences. We begin by defining the *robust regret*, R_T, with φ-divergence balls:

$$R_T(\varphi) = \sum_{t=1}^{T} \inf_{q\in B_\varphi^t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t^*, c)] - \inf_{q\in B_\varphi^t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)], \tag{6}$$

https://doi.org/10.52202/075280-0883



(a) Stochastic and DRO solutions are different. Our method using $\varepsilon = \{0.5, 1\}$ result in the best performance.



(b) Stochastic and DRO solutions are coincide. Our method with $\varepsilon \to 0$ is the best.

Figure 1: Two settings in DRO when the stochastic solution and robust solution are different (*top*) and identical (*bottom*). *Left*: original function $f(\mathbf{x}, c)$. *Middle*: selection of input $\mathbf{x}_t$ over iterations. *Right*: performance with different ε.

where $\mathbf{x}_t^* = \arg\max_{\mathbf{x}\in\mathcal{X}} \inf_{q\in B^t_{\varepsilon,\varphi}} \mathbb{E}_{q(c)}[f(\mathbf{x}, c)]$. We use $\boldsymbol{K}_t$ to denote the generated kernel matrix from dataset $D_t = \{(\mathbf{x}_i, c_i)\}_{i=1}^t \subset \mathcal{X}\times\mathcal{C}$. we now introduce a standard quantity in regret analysis in BO is the *maximum information gain*: $\gamma_t = \max_{D\subset\mathcal{X}\times\mathcal{C}:|D|=t} \log\det\left(\mathbf{I}_t + \sigma_f^{-2}\boldsymbol{K}_t\right)$ where $\boldsymbol{K}_t = [k([\mathbf{x}_i, c_i], [\mathbf{x}_j, c_j])]_{\forall i,j\le t}$ is the covariance matrix and σ_f^2 is the output noise variance.

Theorem 4 (**φ-divergence Regret**) *Suppose the target function is bounded, meaning that $M = \sup_{(\mathbf{x},c)\in\mathcal{X}\times\mathcal{C}} |f(\mathbf{x}, c)| < \infty$ and suppose f has bounded RKHS norm with respect to k. For any lower semicontinuous convex $\varphi : \mathbb{R} \to (-\infty, \infty]$ with $\varphi(1) = 0$, if there exists a monotonic invertible function $\Gamma_\varphi : [0, \infty) \to \mathbb{R}$ such that $\mathrm{TV}(p, q) \le \Gamma_\varphi(\mathsf{D}_\varphi(p, q))$, the following holds*

$$R_T(\varphi) \le \frac{\sqrt{8T\beta_T\gamma_T}}{\log(1+\sigma_f^{-2})} + \left(2M + \sqrt{\beta_T}\right)\sum_{t=1}^T \Gamma_\varphi(\varepsilon_t),$$

with probability $1-\delta$, where $\beta_t = 2||f||_k^2 + 300\gamma_t \ln^3(t/\delta)$, γ_t is the maximum information gain as defined above, and σ_f is the standard deviation of the output noise.

The full proof can be found in the Appendix Section 8. We first remark that with regularity assumptions on f, sublinear analytical bounds for γ_T are known for a range of kernels, e.g., given $\mathcal{X}\times\mathcal{C} \subset \mathbb{R}^{d+1}$ we have for the RBF kernel, $\gamma_T \le \mathcal{O}\left(\log(T)^{d+2}\right)$ or for the Matérn kernel with $\nu > 1$, $\gamma_T \le \mathcal{O}\left(T^{\frac{(d+1)(d+2)}{2\nu+(d+1)(d+2)}}(\log T)\right)$. The second term in the bound is directly a consequence of DRO and by selecting $\varepsilon_t = 0$, it will vanish since any such Γ_φ will satisfy $\Gamma_\varphi(0) = 0$. To ensure sublinear regret, we can select $\varepsilon_t = \Gamma_\varphi^{-1}\left(\frac{1}{\sqrt{t}+\sqrt{t+1}}\right)$, noting that the second term will reduce to $\sum_{t=1}^T \varepsilon_t \le \sqrt{T}$. Finally, we remark that the existence of Γ_φ is not so stringent since for a wide choices of φ, one can find inequalities between the Total Variation and D_φ, to which we refer the reader to Sason and Verdú [45]. For the examples discussed above, we can select $\Gamma_\varphi(t) = t$ for the TV. For the χ^2 and KL cases, one can choose $\Gamma_{\chi^2}(b) = 2\sqrt{\frac{b}{1+b}}$ and $\Gamma_{\mathrm{KL}}(b) = 1 - \exp(-b)$.

https://doi.org/10.52202/075280-0883

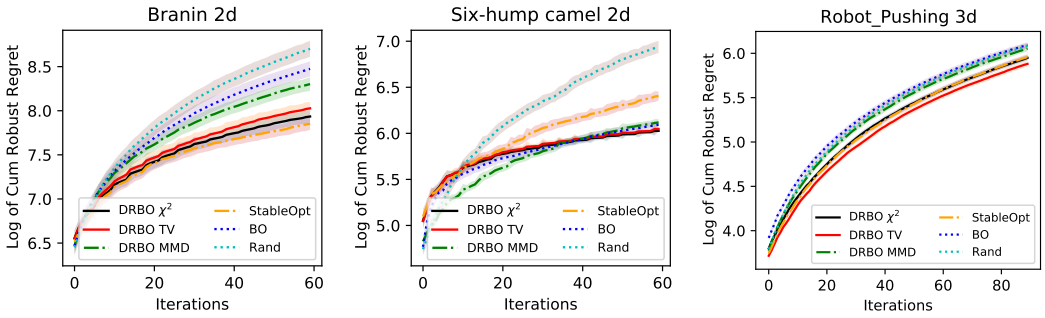


Figure 3: Cumulative robust regret across algorithms. The results show that the proposed χ^2 and TV achieve the best performance across benchmark functions. Random and vanilla BO approaches perform poorly which do not take into account the robustness criteria. Best viewed in color.

5 Experiments

Experimental setting. The experiments are repeated using 30 independent runs. We set $|C| = 30$ which should be sufficient to draw $c \overset{\text{iid}}{\sim} q$ in one-dimensional space to compute Eqs. (4,5). We optimize the GP hyperparameter (e.g., learning rate) by maximizing the GP log marginal likelihood [43]. We will release the Python implementation code in the final version.

Baselines. We consider the following baselines for comparisons. *Rand*: we randomly select $\mathbf{x}_t$ irrespective of c_t. *BO*: we follow the GP-UCB [52] to perform standard Bayesian optimization (ignoring the context c_t). The selection at each iteration is $\mathbf{x}_t = \text{argmax}_{\mathbf{x}}\, \mu(\mathbf{x}) + \beta_t \sigma(\mathbf{x})$. *StableOpt*: we consider the worst-case robust optimization presented in Bogunovic et al. [8]. The selection at each iteration $\mathbf{x}_t = \text{argmax}_{\mathbf{x}} \text{argmin}_c\, \mu(\mathbf{x}, c) + \beta_t \sigma(\mathbf{x}, c)$. *DRBO MMD* [26]: Since there is no official implementation available, we have tried our best to re-implement the algorithm.

We consider the popular benchmark functions[3] with different dimensions d. To create a context variable c, we pick the last dimension of these functions to be the context input while the remaining $d - 1$ dimension becomes the input $\mathbf{x}$.

5.1 Ablation Studies

To gain understanding into how our framework works, we consider two popular settings below.

DRBO solution is different from stochastic solution. In Fig. 1a, the vanilla BO tends to converge greedily toward the stochastic solution (non-distributionally robust) $\text{argmax}_{\mathbf{x}} f(\mathbf{x}, \cdot)$. Thus, BO keeps exploiting in the locality of $\text{argmax}_{\mathbf{x}} f(\mathbf{x}, \cdot)$ from iteration 15. On the other hand, all other DRBO methods will keep exploring to seek for the distributionally robust solutions. Using the high value of $\varepsilon_t \in \{0.5, 1\}$ will result in the best performance.

DRBO solution is identical to stochastic solution. When the stochastic and robust solutions coincide at the same input $\mathbf{x}^*$, the solution of BO will be equivalent to the solution of DRBO methods. This is demonstrated by Fig. 1b. Both stochastic and robust approaches will quickly identify the optimal solution (see the $\mathbf{x}_t$ selection). We learn empirically that setting $\varepsilon_t \to 0$ will lead to the best performance. This is because the DRBO setting will become the standard BO.

The best choice of ε depends on the property of the underlying function, e.g., the gap between the stochastic and DRBO solutions. In practice, we may not be able to identify these scenarios in advance. Therefore, we can use the adaptive value of ε_t presented in Section 4.2. Using this adaptive setting, the performance is stable, as illustrated in the figures.

5.2 Computational efficiency

The key benefit of our framework is simplifying the existing intractable computation by providing the closed-form solution. Additional to improving the quality, we demonstrate this advantage in terms of computational complexity. Our main baseline for comparison is the MMD [26]. As shown in Fig. 2, our DRBO is consistently faster than the constraints linear programming approximation used for

[3] https://www.sfu.ca/ ssurjano/optimization.html

https://doi.org/10.52202/075280-0883

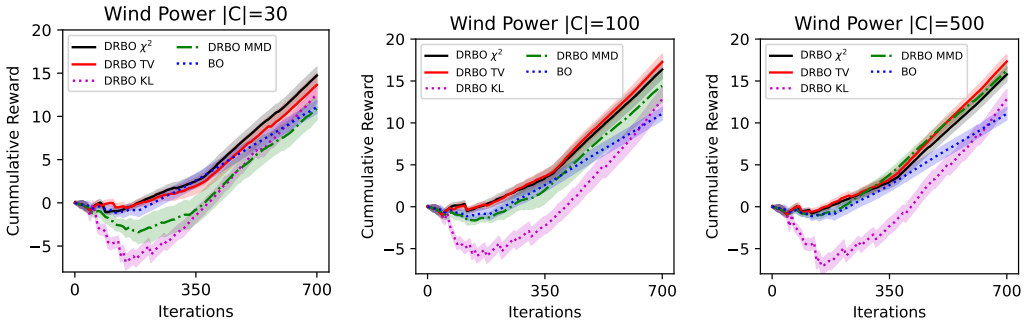


Figure 4: All divergences improve with larger $|C|$. However, MMD comes with the quadratic cost.

MMD. This gap is substantial in higher dimensions. In particular, as compared to Kirschner et al. [26], our DRBO is 5-times faster in *5d* and 10-times faster in *6d*.

5.3 Optimization performance comparison

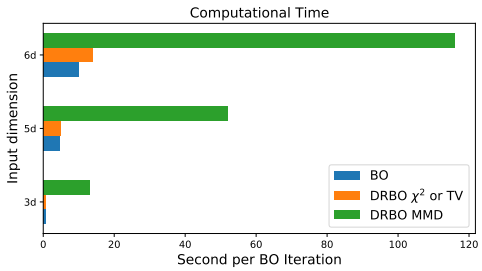


Figure 2: We compare the computational cost across methods. Our proposed DRBO using χ^2 and TV take similar cost per iteration which is significantly lower than the DRBO MMD [26].

We compare the algorithms in Fig. 3 using the robust (cumulative) regret defined in Eq. (6) which is commonly used in DRO literature [26, 33]. The random approach does not make any intelligent information in making decision, thus performs the worst. While BO performs better than random, it is still inferior comparing to other distributionally robust optimization approaches. The reason is that BO does not take into account the context information in making the decision. The StableOpt [8] performs relatively well that considers the worst scenarios in the subset of predefined context. This predefined subset can not cover all possible cases as opposed to the distributional robustness setting.

The MMD approach [26] needs to solve the inner adversary problem using linear programming with convex constraints, additional to the main optimization step. As a result, the performance of MMD is not as strong as our TV and χ^2. Our proposed approach does not suffer this pathology and thus scale well in continuous and high dimensional settings of context input c.

Real-world functions. We consider the deterministic version of the robot pushing objective from Wang and Jegelka [60]. The goal is to find a good pre-image for pushing an object to a target location. The 3-dimensional function takes as input the robot location $(r_x, r_y) \in [-5, 5]^2$ and pushing duration $r_t \in [1, 30]$. We follow Bogunovic et al. [8] to twist this problem in which there is uncertainty regarding the precise target location, so one seeks a set of input parameters that is robust against a number of different potential pushing duration which is a context.

We perform an experiment on Wind Power dataset [8] and vary the context dimensions $|C| \in \{30, 100, 500\}$ in Fig. 4. When $|C|$ enlarges, our DRBO χ^2, TV and KL improves. However, the performances do not improve further when increasing $|C|$ from 100 to 500. Similarly, MMD improves with $|C|$, but it comes with the quadratic cost w.r.t. $|C|$. Overall, our proposed DRBO still performs favourably in terms of optimization quality and computational cost than the MMD.

6 Conclusions, Limitations and Future works

In this work, we showed how one can study the DRBO formulation with respect to φ-divergences and derived a new algorithm that removes much of the computational burden, along with a sublinear regret bound. We compared the performance of our method against others, and showed that our results unveil a deeper connection between regularization and robustness, which serves useful conceptually.

Limitations and Future Works One of the limitations of our framework is in the choice of φ, for which we provide no guidance. For different applications, different choices of φ would prove to be more useful, the study of which we leave for future work.

https://doi.org/10.52202/075280-0883

Acknowledgements

We would like to thank the anonymous reviewers for providing feedback.

References

[1] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pages 1576–1584, 2015. 3

[2] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012. 3, 14

[3] Justin J Beland and Prasanth B Nair. Bayesian optimization under uncertainty. In *NIPS BayesOpt 2017 workshop*, 2017. 2

[4] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. 4, 5

[5] M Bennouna and Bart PG Van Parys. Learning and decision-making with data: Optimal formulations and phase transitions. *arXiv preprint arXiv:2109.06911*, 2021. 4

[6] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. 2, 4

[7] Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019. 2, 4

[8] Ilija Bogunovic, Jonathan Scarlett, Stefanie Jegelka, and Volkan Cevher. Adversarially robust optimization with Gaussian processes. In *Conference on Neural Information Processing Systems (NIPS)*, number CONF, 2018. 8, 9

[9] Zac Cranko, Simon Kornblith, Zhan Shi, and Richard Nock. Lipschitz networks and distributional robustness. *arXiv preprint arXiv:1809.01129*, 2018. 2, 4

[10] Zac Cranko, Zhan Shi, Xinhua Zhang, Richard Nock, and Simon Kornblith. Generalised lipschitz regularisation equals distributional robustness. In *International Conference on Machine Learning*, pages 2178–2188. PMLR, 2021. 2, 3

[11] John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *The Journal of Machine Learning Research*, 20(1):2450–2504, 2019. 2, 3, 5, 6

[12] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013. 2, 3, 4, 5

[13] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3): 946–969, 2021. 2, 3, 4, 5

[14] KD Dvijotham, J Hayes, B Balle, Z Kolter, C Qin, A Gyorgy, K Xiao, S Gowal, and P Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020. 5

[15] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953. 15

[16] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv e-prints*, pages arXiv–1712, 2017. 2

[17] Ziv Goldfeld, Kristjan Greenewald, Jonathan Niles-Weed, and Yury Polyanskiy. Convergence of smoothed empirical measures with applications to entropy estimation. *IEEE Transactions on Information Theory*, 66(7):4368–4391, 2020. 20

[18] Shivapratap Gopakumar, Sunil Gupta, Santu Rana, Vu Nguyen, and Svetha Venkatesh. Algorithmic assurance: An active approach to algorithmic testing using bayesian optimisation. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[19] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 4

[20] José Miguel Hernández-Lobato, James Requeima, Edward O Pyzer-Knapp, and Alán Aspuru-Guzik. Parallel and distributed Thompson sampling for large-scale accelerated exploration of chemical space. *In International Conference on Machine Learning*, pages 1470–1479, 2017. 1

[21] Linfang Hou, Liang Pang, Xin Hong, Yanyan Lan, Zhiming Ma, and Dawei Yin. Robust reinforcement learning with wasserstein constraint. *arXiv preprint arXiv:2006.00945*, 2020. 2

[22] Hisham Husain. Distributional robustness with ipms and links to regularization and gans. *Advances in Neural Information Processing Systems*, 33:11816–11827, 2020. 2, 3

[23] Hisham Husain and Jeremias Knoblauch. Adversarial interpretation of bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 553–572. PMLR, 2022. 14

[24] Hisham Husain, Richard Nock, and Robert C Williamson. A primal-dual link between gans and autoencoders. In *Advances in Neural Information Processing Systems*, pages 413–422, 2019. 14

[25] 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. 1

[26] Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2174–2184. PMLR, 2020. 1, 2, 4, 5, 6, 8, 9, 17

[27] Andreas Krause and Cheng S Ong. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pages 2447–2455, 2011. 1

[28] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 19

[29] Harold J Kushner. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964. 1

[30] Cheng Li, Rana Santu, Sunil Gupta, Vu Nguyen, Svetha Venkatesh, Alessandra Sutti, David Rubin De Celis Leal, Teo Slezak, Murray Height, Mazher Mohammed, and Ian Gibson. Accelerating experimental design by incorporating experimenter hunches. In *International Conference on Data Mining*, pages 257–266, 2018. 1

[31] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018. 14, 15

[32] Ruben Martinez-Cantin, Kevin Tee, and Michael McCourt. Practical Bayesian optimization in the presence of outliers. In *International Conference on Artificial Intelligence and Statistics*, pages 1722–1731. PMLR, 2018. 2

[33] Thanh Nguyen, Sunil Gupta, Huong Ha, Santu Rana, and Svetha Venkatesh. Distributionally robust Bayesian quadrature optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1921–1931. PMLR, 2020. 2, 5, 9

[34] Vu Nguyen and Michael A Osborne. Knowing the what but not the where in Bayesian optimization. In *International Conference on Machine Learning*, pages 7317–7326, 2020. 1

[35] Vu Nguyen, Sunil Gupta, Santu Rana, Cheng Li, and Svetha Venkatesh. Regret for expected improvement over the best-observed value and stopping condition. In *Proceedings of The 9th Asian Conference on Machine Learning (ACML)*, pages 279–294, 2017. 17

https://doi.org/10.52202/075280-0883

[36] Vu Nguyen, Vaden Masrani, Rob Brekelmans, Michael Osborne, and Frank Wood. Gaussian process bandit optimization of the thermodynamic variational objective. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[37] José Nogueira, Ruben Martinez-Cantin, Alexandre Bernardino, and Lorenzo Jamone. Unscented Bayesian optimization for safe robot grasping. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1967–1972. IEEE, 2016. 2

[38] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016. 15

[39] Rafael Oliveira, Lionel Ott, and Fabio Ramos. Bayesian optimisation under uncertain inputs. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1177–1184. PMLR, 2019. 2

[40] Jack Parker-Holder, Vu Nguyen, and Stephen J Roberts. Provably efficient online hyperparameter optimization with population-based bandits. *Advances in neural information processing systems*, 33:17200–17211, 2020. 1

[41] Valerio Perrone, Huibin Shen, Aida Zolic, Iaroslav Shcherbatyi, Amr Ahmed, Tanya Bansal, Michele Donini, Fela Winkelmolen, Rodolphe Jenatton, Jean Baptiste Faddoul, et al. Amazon sagemaker automatic model tuning: Scalable gradient-free optimization. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3463–3471, 2021. 1

[42] Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019. 4

[43] Carl E Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Prees, 2006. 3, 8, 14

[44] Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A Osborne, and Stephen Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In *International Conference on Machine Learning*, pages 8276–8285. PMLR, 2020. 1

[45] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. 7

[46] Herbert E Scarf. A min-max solution of an inventory problem. Technical report, RAND CORP SANTA MONICA CALIF, 1957. 1

[47] Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*, 32:13624–13633, 2019. 2

[48] Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019. 2, 4

[49] 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. 1, 3, 14

[50] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017. 5

[51] Alexander Shapiro, Enlu Zhou, and Yifan Lin. Bayesian distributionally robust optimization. *SIAM Journal on Optimization*, 33(2):1279–1304, 2023. 2

[52] 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*, pages 1015–1022, 2010. 1, 6, 8, 16, 17, 20

[53] Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pages 9131–9141, 2019. 2, 4

[54] Sebastian Shenghong Tay, Chuan Sheng Foo, Urano Daisuke, Richalynn Leong, and Bryan Kian Hsiang Low. Efficient distributionally robust bayesian optimization with worst-case sensitivity. In *International Conference on Machine Learning*, pages 21180–21204. PMLR, 2022. 3, 4

[55] MC Tran, V Nguyen, Richard Bruce, DC Crockett, Federico Formenti, PA Phan, SJ Payne, and AD Farmery. Simulation-based optimisation to quantify heterogeneity of specific ventilation and perfusion in the lung by the inspired sinewave test. *Scientific reports*, 11(1):1–10, 2021. 1

[56] Tsuyoshi Ueno, Trevor David Rhone, Zhufeng Hou, Teruyasu Mizoguchi, and Koji Tsuda. Combo: an efficient Bayesian optimization library for materials science. *Materials discovery*, 4: 18–21, 2016. 1

[57] Nienke ER van Bueren, Thomas L Reed, Vu Nguyen, James G Sheffield, Sanne HG van der Ven, Michael A Osborne, Evelyn H Kroesbergen, and Roi Cohen Kadosh. Personalized brain stimulation for effective neurointervention across participants. *PLOS Computational Biology*, 17(9):e1008886, 2021. 1

[58] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 15

[59] Xingchen Wan, Cong Lu, Jack Parker-Holder, Philip J Ball, Vu Nguyen, Binxin Ru, and Michael Osborne. Bayesian generational population-based training. In *International Conference on Automated Machine Learning*, pages 14–1. PMLR, 2022. 1

[60] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635, 2017. 9

[61] Jingzhao Zhang, Aditya Krishna Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. In *International Conference on Learning Representations 2021*. 19

https://doi.org/10.52202/075280-0883