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 [10], 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) [13, 9, 27, 25, 19] allows us to model an 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 [11, 30], 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 [29, 8, 14, 28] and also a staple in machine learning for the crucial problem of hyperparameter tuning [18, 21, 22].

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. A recent approach to solving this has been *Distributionally Robust Bayesian Optimization* (DRBO) where one models against an adversary who shifts the context distribution with respect to a distance over probability measures. [10] 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. However 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. 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.

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

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 use exploit 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 [10]. 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. A practical 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 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 with respect to the φ-divergence as

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

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 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 [10]. 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 [5, 3, 4].

We will exploit various properties of the φ-divergence in order 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. We first define the convex conjugate of φ as $\varphi^\star(u) = \sup_{u' \in \mathrm{dom}_\varphi}(u \cdot u' - \varphi(u'))$, 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*

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

We remark that similar results exist for other areas such as supervised learning [26] and certifying robust radii [6]. However this is, to the best of our knowledge, the first development for BO. The Theorem above 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 5.

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

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

The above example can be very 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 [5, 3]. We remark that this penalty and its relationship to χ^2-divergence has been developed in the similar yet related problem of Bayesian quadrature [16]. 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 2 (Total Variation) *Let $\varphi(u) = |u-1|$, then for any measurable and bounded f we have for any choice of ε_t*

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

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 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.

2.1 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-th, 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:

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

While our method is not restricted to the form of the acquisition function, for convenient in the theoretical analysis, we follow the GP-UCB [27]. Given the GP predictive mean and variance, we have the acquisition function for the χ^2 is as follows:

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

where $\bar{\mu_t} = \frac{1}{|C|}\sum_c \mu_t(\mathbf{x}, c)$ and $c \sim q$ can be generated in a one dimensional space to approximate the expectation and the variance. We make an important remark here that since we do not require our context space to be finite, our implementation scales only linearly with the number of context samples M we draw from q. On the other hand, [10] at every iteration of t solves an M-dimensional constraint optimization problem that is substantially more compuationally expensive. In the experiment, we select q as the uniform distribution, but it is not restricted to. Similarly, an acquisition function for Total Variation is written as

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

2.2 Convergence Analysis

One of the main advantages of [10] 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:

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. (1)
6: Total Variation: $\alpha(\mathbf{x}) := \alpha^{TV}(\mathbf{x})$ from Eq. (2)
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**

$$R_T(\varphi) = \sum_{t=1}^{T} \inf_{q \in B_\varphi^t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t^*, c)] - \inf_{q \in B_\varphi^t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)], \tag{3}$$

where $\mathbf{x}_t^* = \arg\max_{\mathbf{x}\in\mathcal{X}} \inf_{q \in B_{\varepsilon,\varphi}^t} \mathbb{E}_{q(c)}[f(\mathbf{x}, c)]$. 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*: $\gamma_t = \max_{D \subset \mathcal{X}\times\mathcal{C}:|D|=t} \log\det(\mathbf{I}_t + \boldsymbol{K}_t)$ where $\boldsymbol{K}_t = [k([\mathbf{x}_i, c_i], [\mathbf{x}_j, c_j])]_{\forall i,j \le t}$.

Theorem 2 (φ-divergence Regret) *Let $M = \sup_{(\mathbf{x},c)\in\mathcal{X}\times\mathcal{C}} |f(\mathbf{x}, c)| < \infty$ and suppose f lives in an RKHS. 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 $\mathrm{TV}(p, q) \le \Gamma_\varphi(\mathsf{D}_\varphi(p, q))$*

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

where σ_f is the standard deviation of the output noise.

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 $\Gamma_\varphi(0) = 0$. In order to ensure sublinear regret, we can select $\varepsilon_t = \Gamma_\varphi^{-1}\left(\frac{1}{\sqrt{t}+\sqrt{t+1}}\right)$, noting that $\sum_{t=1}^T \varepsilon_t \le \sqrt{T}$ (see Lemma 5). This gives us an actionable item for implementation. To see how one can select Γ_φ, for the Total Variation case, one can select $\Gamma_\varphi(t) = t$ noting that the conditions are met. For other popular choices, $\Gamma_{\chi^2}(t) = 2\sqrt{\frac{t}{1+t}}$ and $\Gamma_{\mathrm{KL}}(t) = 1 - \exp(-t)$. We refer the reader to [24] for more inequalities between the Total Variation and other φ-divergences.

3 Experiments

Due to the space limit, we refer to the Appendix Section 7 for the experimental results.

References

[1] Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012. 7

[2] 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. 13, 14

[3] John Duchi, Peter Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016. 2, 3

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

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

[6] 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. 3

[7] Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953. 8

[8] 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

[9] 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

[10] 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, 10, 13, 14

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

[12] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. 12

[13] 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

[14] 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

[15] Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018. 8

[16] 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. 3, 14

[17] 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. 10

[18] 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

[19] 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

[20] 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. 8

[21] 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, 2020. 1

[22] 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 black-box optimization. *arXiv preprint arXiv:2012.08489*, 2020. 1

[23] Carl E Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Prees, 2006. 7, 13

[24] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016. 4

[25] 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, 7

[26] Alexander Shapiro. Distributionally robust stochastic programming. *SIAM Journal on Optimization*, 27(4):2258–2275, 2017. 3

[27] 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, 3, 9, 10, 12, 13

[28] 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

[29] 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

[30] 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

[31] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 8

[32] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient Bayesian optimization. In *International Conference on Machine Learning*, pages 3627–3635, 2017. 14

[33] Jingzhao Zhang, Aditya Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*, 2020. 12

4 Gaussian Process Modeling with Input x and Context c

Gaussian Processes. We follow a popular choice in BO [25] to use GP as a surrogate model for optimizing f. A GP [23] 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 $p(f_* \mid \boldsymbol{f}) \sim \mathcal{N}\left(\mu(\mathbf{x}_*), \sigma^2(\mathbf{x}_*)\right)$. Its mean and variance are given by:

$$\mu(\mathbf{x}_*) = \mathbf{k}_{*,N}\mathbf{K}_{N,N}^{-1}\mathbf{y}, \quad (4) \qquad \sigma^2(\mathbf{x}_*) = k_{**} - \mathbf{k}_{*,N}\mathbf{K}_{N,N}^{-1}\mathbf{k}_{*,N}^T \quad (5)$$

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 [23] for further details on GPs.

Gaussian Process with Input x and Context c. Given the additional context variable $c \in \mathbb{R}$, it is natural to consider a GP model in which the input is the concatenation of $[\mathbf{x}, c] \in \mathbb{R}^{d+1}$. In particular, we can write a GP [23] as:

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

where $m([\mathbf{x}, c])$ and $k([\mathbf{x}, c], [\mathbf{x}', c'])$ are the mean and covariance functions. For predicting $f_* = f([\mathbf{x}_*, c_*])$ at a new data point $[\mathbf{x}_*, c_*]$, the conditional probability follows a univariate Gaussian distribution as $p(f_* \mid \boldsymbol{f}) \sim \mathcal{N}\left(\mu([\mathbf{x}_*, c_*]), \sigma^2([\mathbf{x}_*, c_*])\right)$. Its mean and variance are given by:

$$\mu([\mathbf{x}_*, c_*]) = \mathbf{k}_{*,N}\mathbf{K}_{N,N}^{-1}\mathbf{y}, \quad (6) \qquad \sigma^2([\mathbf{x}_*, c_*]) = k_{**} - \mathbf{k}_{*,N}\mathbf{K}_{N,N}^{-1}\mathbf{k}_{*,N}^T \quad (7)$$

where $k_{**} = k([\mathbf{x}_*, c_*], [\mathbf{x}_*, c_*])$, $\mathbf{k}_{*,N} = [k([\mathbf{x}_*, c_*], [\mathbf{x}_i, c_i])]_{\forall i \leq N}$ and $\boldsymbol{K}_{N,N} = [k([\mathbf{x}_i, c_i], [\mathbf{x}_j, c_j])]_{\forall i,j \leq N}$.

5 Proofs of Main Results

In the sequel, when we say a function is measurable, we attribute it with respect to the Borel σ-algebras based on the Polish topologies. We remark that a similar proof to ours has appeared in [1], which is not specific to BO objective yet also we require compactness of the set $\mathcal{C}$.

Theorem 3 *(Theorem 1 in the main paper) Let $\varphi : \mathbb{R} \to (-\infty, \infty]$ be a convex lower semicontinuous mapping such that $\varphi(1) = 0$. For any $\varepsilon > 0$, it holds that*

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

Proof For a fixed $\mathbf{x} \in \mathcal{X}$, we first introduce a Lagrangian variable $\lambda \geq 0$ that acts to enforce the ball constraint $\mathsf{D}_\varphi(p, q) \leq \varepsilon$:

$$\inf_{q \in B_\varphi^t(p)} \mathbb{E}_{q(c)}[f(\mathbf{x}, c)] = \inf_{q \in \Delta(\mathcal{C})} \sup_{\lambda \geq 0} \left(\mathbb{E}_{q(c)}[f(\mathbf{x}, c)] - \lambda\left(\varepsilon_t - \mathsf{D}_\varphi(q, p_t)\right)\right) \quad (8)$$

$$\overset{(1)}{=} \sup_{\lambda \geq 0} \inf_{q \in \Delta(\mathcal{C})} \left(\mathbb{E}_{q(c)}[f(\mathbf{x}, c)] - \lambda\left(\varepsilon_t - \mathsf{D}_\varphi(q, p_t)\right)\right) \quad (9)$$

$$= \sup_{\lambda \geq 0} \left(-\lambda\varepsilon_t - \sup_{q \in \Delta(\mathcal{C})} \left(\mathbb{E}_{q(c)}[-f(\mathbf{x}, c)] - \lambda \mathsf{D}_\varphi(q, p_t)\right)\right) \quad (10)$$

$$\overset{(2)}{=} \sup_{\lambda \geq 0} \left(-\lambda\varepsilon_t - \inf_{b \in \mathbb{R}} \left(\lambda \mathbb{E}_{p_t(c)}\left[\varphi^\star\left(\frac{b - f(\mathbf{x}, c)}{\lambda}\right)\right] - b\right)\right) \quad (11)$$

$$= \sup_{\lambda \geq 0, b \in \mathbb{R}} \left(b - \lambda\varepsilon_t - \lambda \mathbb{E}_{p_t(c)}\left[\varphi^\star\left(\frac{b - f(\mathbf{x}, c)}{\lambda}\right)\right]\right) \quad (12)$$

(1) is due to Fan's minimax Theorem [7, Theorem 2] noting that for any $\mathbf{x} \in \mathcal{X}$, the mapping $q \mapsto \mathbb{E}_{q(c)}[f(\mathbf{x}, c)]$ is linear and for any φ chosen as stated in the Theorem, the mapping $q \mapsto \mathsf{D}_\varphi(q, p_t)$ is convex and lower semi-continuous. Furthermore noting that $\mathcal{C}$ is compact, we also have that $\Delta(\mathcal{C})$ is compact [31]. (2) is due to a standard result due to the Fenchel dual of the φ-divergence, see Eq. (22) of [15] for example. We state the result in the following lemma, which depends on the convex conjugate $\varphi^\star(u) = \sup_{u'} (uu' - \varphi(u'))$.

Lemma 1 *([15]) For any measurable function $h : \mathcal{C} \to \mathbb{R}$ and convex lower semi-continuous function $\varphi : \mathbb{R} \to (-\infty, \infty]$ with $\varphi(1) = 0$, it holds that*

$$\sup_{q \in \Delta(\mathcal{C})} \left(\mathbb{E}_{q(c)}[h(c)] - \lambda \mathsf{D}_\varphi(q, p) \right) = \inf_{b \in \mathbb{R}} \left(\lambda \mathbb{E}_{p_t(c)} \left[\varphi^\star \left(\frac{b + h(c)}{\lambda} \right) \right] - b \right),$$

where $p \in \Delta(\mathcal{C})$ and $\lambda > 0$.

■

For each of the specific derivations below, we recall the standard derivations for $\varphi^\star$, which can be found in many works, for example in [20].

Example 3 (χ^2-divergence) *(Example 1 in the main paper) If $\varphi(u) = (u-1)^2$, then we have*

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

Proof In this case we have $(\lambda\varphi)^\star(u) = \frac{u^2}{4\lambda} + u$ and so we have

$$\begin{aligned}
\sup_{b \in \mathbb{R}} \left(b - \mathbb{E}_{p_t(c)} \left[\frac{(b - f(\mathbf{x}, c))^2}{4\lambda} + b - f(\mathbf{x}, c) \right] \right) &= \sup_{b \in \mathbb{R}} \left(\mathbb{E}_{p_t(c)}\left[f(\mathbf{x}, c) \right] - \mathbb{E}_{p_t(c)} \left[(b - f(\mathbf{x}, c))^2 \right] \right) \\
&= \mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \frac{1}{4\lambda} \inf_{b \in \mathbb{R}} \mathbb{E}_{p_t(c)} \left[(b - f(\mathbf{x}, c))^2 \right] \\
&= \mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \frac{1}{4\lambda} \operatorname{Var}_{p_t(c)}[f(\mathbf{x}, c)].
\end{aligned}$$

Combining this with the original objective and by Theorem 1 yields

$$\begin{aligned}
\inf_{q \in B^t_\varphi(p)} \mathbb{E}_{c \sim q}[f(\mathbf{x}, c)] &= \sup_{\lambda \geq 0, b \in \mathbb{R}} \left(b - \lambda\varepsilon_t - \lambda \mathbb{E}_{p_t(c)} \left[\varphi^\star \left(\frac{b - f(\mathbf{x}, c)}{\lambda} \right) \right] \right) \\
&= \mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \inf_{\lambda \geq 0} \left(\lambda\varepsilon_t + \frac{1}{4\lambda} \operatorname{Var}_{p_t(c)}[f(\mathbf{x}, c)] \right) \\
&= \mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \sqrt{\varepsilon_t \operatorname{Var}_{p_t(c)}[f(\mathbf{x}, c)]}.
\end{aligned}$$

where the last equation is using the arithmetic and geometric means inequality. ■

Example 4 (Total Variation) *(Example 2 in the main paper) If $\varphi(u) = |u - 1|$, then we have*

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

Proof In this case, the conjugate of φ is $(\lambda\varphi)^\star(u) = 0$ if $|u| \leq \lambda$ and $+\infty$ otherwise. Therefore, when considering the right-hand side of Theorem 1, we will require λ to be larger than $|b - f(\mathbf{x}, c)|$ for all $c \in \mathcal{C}$ for the expression to be finite. In this case, for a fixed $b \in \mathbb{R}$, the optimization over $\lambda \geq 0$ becomes

$$\sup_{b \in \mathbb{R}} \sup_{\lambda \geq 0} \left(b - \lambda\varepsilon_t - \lambda \mathbb{E}_{p_t(c)} \left[\varphi^\star \left(\frac{b - f(\mathbf{x}, c)}{\lambda} \right) \right] \right) \tag{13}$$

$$= \sup_{b \in \mathbb{R}} \left(\mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \sup_{c \in \mathcal{C}} |b - f(\mathbf{x}, c)| \, \varepsilon_t \right) \tag{14}$$

$$= \mathbb{E}_{p_t(c)}[f(\mathbf{x}, c)] - \inf_{b \in \mathbb{R}} \sup_{c \in \mathcal{C}} |b - f(\mathbf{x}, c)| \, \varepsilon_t. \tag{15}$$

We need the following Lemma to proceed.

Lemma 2 *For any set of numbers $A \subset \mathbb{R}$, let $\overline{a}, \underline{a} \in A$ be the maximum and minimal elements. It then holds that*

$$\inf_{b\in\mathbb{R}} \sup_{a\in A} |b-a| = \frac{1}{2}\left|\overline{a}-\underline{a}\right|.$$

Proof First note that for any $b \in \mathbb{R}$, we have that $\sup_{a\in A}|b-a| = \max\left(|\overline{a}-b|, |\underline{a}-b|\right)$. The outer inf can be solved by setting $b = \frac{\overline{a}+\underline{a}}{2}$. ■

The proof then concludes by noting setting $A = \{f(\mathbf{x}, c) : c \in \mathcal{C}\}$. ■

Lemma 3 (Theorem 6 from [27]) *Let $\delta \in (0,1)$. Assume the noise variable ε_t are uniformly bounded by σ_f. Define $\beta_t = 2||f||_k^2 + 300\gamma_t \ln^3(t/\delta)$. Then,*

$$P(\forall T, \forall \mathbf{x}, |\mu_T(\mathbf{x}) - f(\mathbf{x})| \leq \beta_T^{\frac{1}{2}}\sigma_T(\mathbf{x}). \tag{16}$$

We first prove a regret bound for the total variation which will be instrumental in proving the regret bound for any general φ, given the existence of Γ_φ.

Theorem 4 (Total Variation Regret) *Let $M = \sup_{(\mathbf{x},c)\in\mathcal{X}\times\mathcal{C}}|f(\mathbf{x},c)|$ and suppose f lives in an RKHS. If $\varphi(u) = |u-1|$, it then holds that*

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

where σ_f is the standard deviation of the output noise.

Proof We first define the GP predictive mean and variance as

$$\begin{aligned}
\mu_t(\mathbf{x}, c) &= k_t([\mathbf{x}, c])^\intercal \left(\boldsymbol{K}_t + \mathbf{I}_t\right)^{-1} y_t \\
\sigma_t(\mathbf{x}, c)^2 &= k([\mathbf{x}, c], [\mathbf{x}, c]) - k_t([\mathbf{x}, c])^\intercal \left(\boldsymbol{K}_t + \mathbf{I}_t\right)^{-1} k_t([\mathbf{x}, c]),
\end{aligned}$$

which are defined in Eqs. (6,7) where $\boldsymbol{K}_t = \left[k\left([\mathbf{x}_i, c_i], [\mathbf{x}_j, c_j]\right)\right]_{\forall i,j\leq t}$. The proof begins by first bounding the regret at time t using a standard argument with a slight modification that uses our results.

$$\begin{aligned}
r_t &= \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \\
&\overset{(1)}{\leq} \mathbb{E}_{p_t(c)}[f(\mathbf{x}^*, c) - \mu(\mathbf{x}^*, c)] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \\
&\overset{(2)}{\leq} \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)]] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \\
&= \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \\
&\overset{(3)}{=} \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c)] - \mathbb{E}_{p_t(c)}[f(\mathbf{x}_t, c)] + \frac{\varepsilon_t}{2}\left(\sup_{c\in\mathcal{C}} f(\mathbf{x}_t, c) - \inf_{c\in\mathcal{C}} f(\mathbf{x}_t, c)\right) \\
&= \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c) - f(\mathbf{x}_t, c)] + \frac{\varepsilon_t}{2}\left(\sup_{c\in\mathcal{C}} f(\mathbf{x}_t, c) - \inf_{c\in\mathcal{C}} f(\mathbf{x}_t, c)\right) \\
&\overset{(4)}{\leq} \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)] + \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}_t, c) - f(\mathbf{x}_t, c)] + \frac{\varepsilon_t}{2}\left(\sup_{c\in\mathcal{C}} f(\mathbf{x}_t, c) - \inf_{c\in\mathcal{C}} f(\mathbf{x}_t, c)\right) \\
&\overset{(5)}{\leq} 2\sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)] + M\varepsilon_t,
\end{aligned} \tag{17}$$

where (1) holds due to selecting p_t and introducing $\mathbb{E}_{q(c)}[\mu(\mathbf{x}^*, c)]$, (2) is due to the result that $|f(\mathbf{x}, c) - \mu(\mathbf{x}, c)| \leq \sqrt{\beta_t}\sigma_t(\mathbf{x}, c)$ for all $x, c \in \mathcal{X} \times \mathcal{C}$ as stated in Lemma 3. (3) is due to Theorem 4, (4) is due to the choice of $\mathbf{x}_t$ since it satisfies:

$$\mathbb{E}_{p_t(c)}[\mu(\mathbf{x}_t, c)] + \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)] \geq \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}, c)] + \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}, c)],$$

for all $\mathbf{x} \in \mathcal{X}$ and therefore

$$\begin{aligned}\mathbb{E}_{p_t(c)}[\mu(\mathbf{x}^*, c)] + \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)] &\leq \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}_t, c)] + \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)] \\ &\leq \mathbb{E}_{p_t(c)}[\mu(\mathbf{x}_t, c)] + \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)].\end{aligned}$$

Finally, (5) holds due to another application of $|f(\mathbf{x}, c) - \mu(\mathbf{x}, c)| \leq \beta_t\sigma_t(\mathbf{x}, c)$ for all $[\mathbf{x}, c] \in [\mathcal{X} \times \mathcal{C}]$ from Lemma 3. For the final step of our proof, we follow the literature in Bayesian optimization [27] to introduce a sample complexity parameter, namely the maximum information gain:

$$\gamma_T := \max_{\{(\mathbf{x}_t, c_t)\}_{t=1}^T} \log\det(\mathbf{I}_t + \boldsymbol{K}_T).$$

where we use $\boldsymbol{K}_T$ to denote the generated kernel matrix from dataset $D_T = \{[\mathbf{x}_i, c_i]\}_{i=1}^T \subset [\mathcal{X} \times \mathcal{C}]$ at iteration T. The information gain is used in the regret bounds for most of Bayesian optimization research [17, 10].

Lemma 4 *(adapted Lemma 7 in [17]) The sum of the predictive variances is bounded by the maximum information gain γ_T. That is $\forall \mathbf{x}, c \in \mathcal{X} \times \mathcal{C}, \sum_{t=1}^T \sigma_{t-1}^2(\mathbf{x}, c) \leq \frac{2}{\log(1+\sigma_f^{-2})}\gamma_T$ where σ_f is the standard deviation of the output noise.*

Using the above Lemma of maximum information gain, we take the square of the term $2\sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}_t, c)]$ in Eq. (17) to have:

$$\sum_{t=1}^T 4\beta_t\mathbb{E}_{p_t(c)}[\sigma_t^2(\mathbf{x}_t, c)] \leq \frac{8\beta_T\gamma_T}{\log(1+\sigma_f^{-2})} \tag{18}$$

By using Cauchy-Schwarz inequality, we get

$$\sum_{t=1}^T 2\sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}, c)] \leq \sqrt{\frac{8T\beta_T\gamma_T}{\log(1+\sigma_f^{-2})}} \tag{19}$$

in which the term T has been included in the right.

Using the above results we get

$$\begin{aligned}R_T(\varphi) &\leq 2\sum_{t=1}^T \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}, c)] + M\sum_{t=1}^T \varepsilon_t && (20)\\ &\leq 2\sqrt{\beta_T}\sum_{t=1}^T \max_c[\sigma_t(\mathbf{x}, c)] + M\sum_{t=1}^T \varepsilon_t && (21)\\ &\leq 2\sqrt{\beta_T}\max_{q\in\{p_1,\ldots,p_T\}}\sum_{t=1}^T \mathbb{E}_{q(c)}[\sigma_t(\mathbf{x}, c)] + M\sum_{t=1}^T \varepsilon_t && (22)\\ &\leq 2\sqrt{\beta_T}\max_{q\in\{p_1,\ldots,p_T\}}\mathbb{E}_{q(c)}\left[\sum_{t=1}^T \sigma_t(\mathbf{x}, c)\right] + M\sum_{t=1}^T \varepsilon_t && (23)\\ &\leq \frac{\sqrt{8T\beta_T\gamma_T}}{\log(1+\sigma_f^{-2})} + M\sum_{t=1}^T \varepsilon_t && (24)\end{aligned}$$

where we have used Eq. (19) to obtained Eq. (24). ■

Lemma 5 *For any $T > 0$, it holds that*

$$\sum_{t=1}^{T} \left(\frac{1}{\sqrt{t} + \sqrt{t+1}} \right) = \sqrt{T+1} - 1 \leq \sqrt{T},$$

Proof By simply rationalizing the denominator, we have that

$$\frac{1}{\sqrt{t} + \sqrt{t+1}} = \sqrt{t+1} - \sqrt{t},$$

and via a simple telescoping sum, the required result holds. A simple inequality will then yield the final inequality. ■

Theorem 5 (φ Regret) *(Theorem 2 in the main paper) Let $M = \sup_{(\mathbf{x},c)\in\mathcal{X}\times\mathcal{C}} |f(\mathbf{x}, c)|$ and suppose f lives in an RKHS. For any lower semicontinuous convex $\varphi : \mathbb{R} \to (-\infty, \infty]$ with $\varphi(1) = 0$, if there exists a monotonic invertible function $\Gamma_\varphi : \mathbb{R} \to \mathbb{R}$ such that $\mathrm{TV}(p, q) \leq \Gamma_\varphi(\mathsf{D}_\varphi(p, q))$*

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

where σ_f is the standard deviation of the output noise.

Proof The key aspect of this proof is to note that

$$\{q \in \Delta(\mathcal{C}) : \mathsf{D}_\varphi(q, p_t) \leq \varepsilon_t\} \subseteq \{q \in \Delta(\mathcal{C}) : \mathrm{TV}(q, p_t) \leq \Gamma_\varphi(\varepsilon_t)\}, \tag{25}$$

which is due to the inequality and monotonicity of the Γ_φ function. By following similar steps to the Total Variation derivation, we have

$$\inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \tag{26}$$

$$\leq \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)^2 + \mu(\mathbf{x}^*, c)] - \inf_{q:\mathsf{D}_\varphi(q,p_t)\leq\varepsilon_t} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)] \tag{27}$$

$$\leq \sqrt{\beta_t}\mathbb{E}_{p_t(c)}[\sigma_t(\mathbf{x}^*, c)^2 + \mu(\mathbf{x}^*, c)] - \inf_{q:\mathrm{TV}(q,p_t)\leq\Gamma_\varphi(\varepsilon_t)} \mathbb{E}_{q(c)}[f(\mathbf{x}_t, c)]. \tag{28}$$

By following the same decomposition as in the proof for the Total Variation, we achieve the same result except with the radius epsilon replaced with $\Gamma_\varphi(\varepsilon_t)$. ■

6 Extension to KL divergence

Our theoretical result for φ-divergence can also be readily extended to handle KL divergence. However, since the empirical result with KL divergence is inferior to to the result with TV and χ-divergences.

Example 5 (KL-divergence) *Let $\varphi(u) = u \log u$, we have*

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

Proof Using Theorem 1, we derive $\varphi^\star$ for this choice which can easily be verified to be $\varphi^\star(t) = \exp(t - 1)$. For simplicity, let $\lambda = 1$ and note that we have

$$\sup_{b\in\mathbb{R}} \left(b - \mathbb{E}_{p_t(c)} \left[\exp\left(b - f(\mathbf{x}, c) - 1 \right) \right] \right) = \sup_{b\in\mathbb{R}} \left(b - \exp(b) \cdot A \right), \tag{29}$$

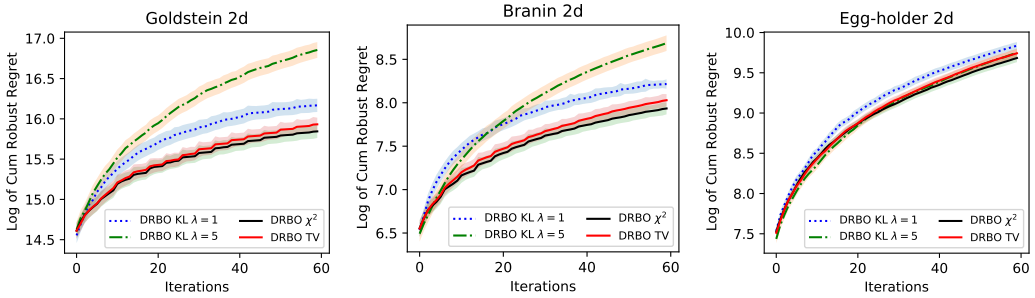


Figure 1: We empirically found that using TV and χ^2 divergences typically obtain better performance than KL divergence. This can be due to the sensitivity of the additional hyperparameter λ in KL. We have considered using $\lambda = 1$ and $\lambda = 5$ in these experiments. The best λ is unknown in advance and depends on the functions.

where $A = \mathbb{E}_{p_t(c)}\left[\exp\left(-f(\mathbf{x}, c) - 1\right)\right]$ is a constant. The above can easily be solved as it admits a differentiable one-dimensional objective and we get the largest value when $b = -\log A$ which then yields

$$\sup_{b \in \mathbb{R}} \left(b - \mathbb{E}_{p_t(c)}\left[\exp\left(b - f(\mathbf{x}, c) - 1\right)\right]\right) = \log \mathbb{E}_{p_t(c)}\left[\exp(-f(\mathbf{x}, c))\right]. \tag{30}$$

The proof concludes noting that $(\lambda\varphi)^\star(u) = \lambda\varphi^\star(u/\lambda)$ for any $\lambda > 0$. ■

The Kullback-Leibler (KL) divergence [12] is a popular choice when quantifying information shift, due to its link with entropy. There exists work that studied distributional shifts with respect to the KL divergence for label shifts [33]. Compared to the general theorem, the KL-divergence variant allows us to find $b \in \mathbb{R}$ in closed form. We remark that we place the KL divergence derivation here for its information-theoretic importance however in our experiments, we find that other choices of φ-divergence outperform the KL-divergence. We now show such examples and particularly illustrate that we can even solve for $\lambda \geq 0$ in closed form for these cases, yielding only a single maximization over $\mathcal{X}$.

The regret bound for the case with KL divergence is presented in Theorem 2 using $\Gamma_{\mathrm{KL}}(t) = 1 - \exp(-t)$.

6.1 Optimization with the GP Surrogate for KL-divergence

To handle the distributional robustness, we have rewritten the objective function using φ divergences in Theorem 1 with the KL in **Example** 5.

Similar to the cases of TV and χ^2, we model the GP surrogate model using the observed data $\{\mathbf{x}_i, y_i\}_{i=1}^{t-1}$ at each iteration t-th, . Then, we select a next point $\mathbf{x}_t$ to query a black-box function by maximizing the acquisition function which is build on top of the GP surrogate:

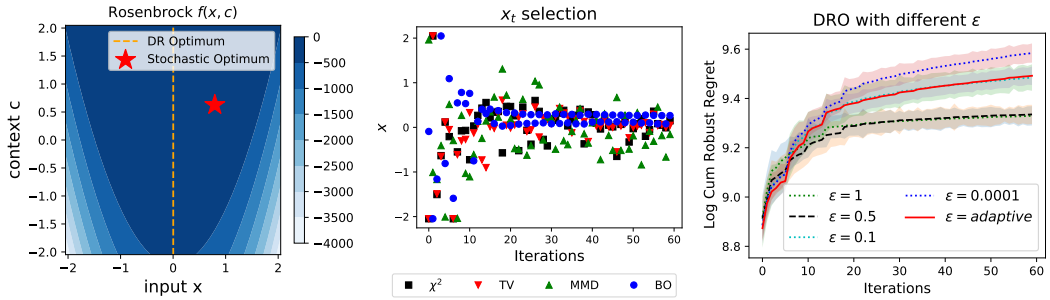
$$\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{X}} \alpha(\mathbf{x}).$$

While our method is not restricted to the form of the acquisition function, for convenient in the theoretical analysis, we follow the GP-UCB [27]. Given the GP predictive mean and variance from Eqs. (6,7), we have the acquisition function for the KL in Example 5 as follows:

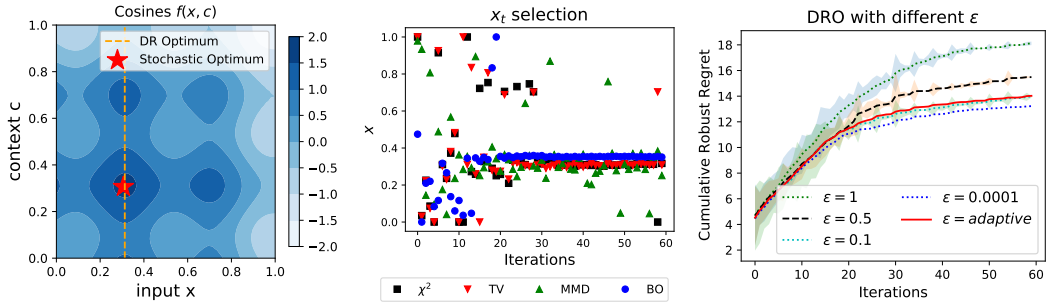
$$\alpha^{KL}(\mathbf{x}) := -\lambda\varepsilon_t - \lambda \log \mathbb{E}_{p_t(c)}\left[\exp\left(\frac{-\frac{1}{|C|}\sum_c \left[\mu_t(\mathbf{x}, c) + \sqrt{\beta_t}\sigma_t(\mathbf{x}, c)\right]}{\lambda}\right)\right] \tag{31}$$

where $c \sim q$ can be generated in a one dimensional space to approximate the expectation and the variance. In the experiment, we define q as the uniform distribution to draw $c \sim q$.

Comparison of KL, TV and χ^2 divergences. In Fig. 1, we present the additional experiments showing the comparison of using different divergences including KL, TV and χ^2. We show empirically that the KL divergence performs generally inferior than the TV and χ^2.



(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 2: 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. Best viewed in color.

7 Experiments

Experimental setting. The experiments are repeated using 30 independent runs. We optimize the GP hyperparameter (e.g., learning rate) by maximizing the GP log marginal likelihood [23]. We will release the Python implementation code in the final version.

Baselines and benchmark functions. We consider the following baselines for comparisons.

- Rand: we randomly select $\mathbf{x}_t$ irrespective of c_t.
- BO: we follow the GP-UCB [27] to perform standard Bayesian optimization (ignoring the context c_t). The selection at each iteration is $\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x}} \mu(\mathbf{x}) + \beta_t \sigma(\mathbf{x})$.
- Stable-Opt: we consider the worst-case robust optimization presented in [2]. The selection at each iteration $\mathbf{x}_t = \operatorname{argmax}_{\mathbf{x}} \operatorname{argmin}_c \mu(\mathbf{x}, c) + \beta_t \sigma(\mathbf{x}, c)$.
- DRBO using MMD [10]: There is no official implementation available. We have tried our best to re-implement the algorithm.

We consider the popular benchmark functions[1] 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 $\mathbf{x}$.

7.1 Ablation Studies

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

DRO solution is different from stochastic solution. In Fig. 2a, the vanilla BO tends to converge greedily toward the stochastic solution (non-distributionally robust) $\operatorname{argmax}_{\mathbf{x}} f(\mathbf{x}, \cdot)$. Thus, BO keeps exploiting in the locality of $\operatorname{argmax}_{\mathbf{x}} f(\mathbf{x}, \cdot)$ from iteration 15. On the other hand, all other

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

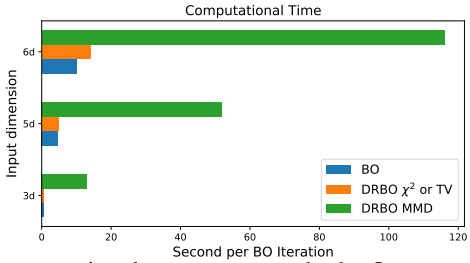


Figure 3: We compare the computational cost across methods. Our proposed DRO using χ^2 and total variation (TV) take similar cost per iteration which is significantly lower than the DRO MMD [10].

DRO 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.

DRO 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 DRO methods. This is demonstrated by Fig. 2b. 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 DRO setting will become the standard BO.

The best choice of ε depends on the property of the underlying function, e.g., how difference between the stochastic and DRO 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 Sec. 2.2. Using this adaptive setting, the performance is stable, see the red curves in Right Fig. 2a.

7.2 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 [10]. As shown in Fig. 3, our DRBO is consistently faster than the constraints linear programming approximation used for MMD. This gap is substantial in higher dimensions. In particular, as compared to [10], our DRBO is 5-times faster in 5d and 10-times faster in 6d.

7.3 Optimization performance comparison

We compare the algorithms in Fig. 4 using the robust (cumulative) regret defined in Eq. (3) which is commonly used in DRO literature [10, 16]. 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 [2] performs relatively well that considers the worst scenarios in the subset of predefined context. This predefined subset is not covered all possible cases as opposed to the distributional robustness setting.

The MMD approach [10] needs to solve the inner adversary problem using linear programming with convex constraints. This program can be solved efficiently but is of size of context discretization $|C|$, which currently limits the method to relatively small context sets. 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 [32]. 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 [2] 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 locations. We present the comparison in Bottom Right Fig. 4. Our proposed DRBO approaches outperform the baselines, especially the Total Variation (TV).

Adaptive value of ε_t. We show the adaptive value of ε_t by iterations for TV, χ^2 and KL in Fig. 5.

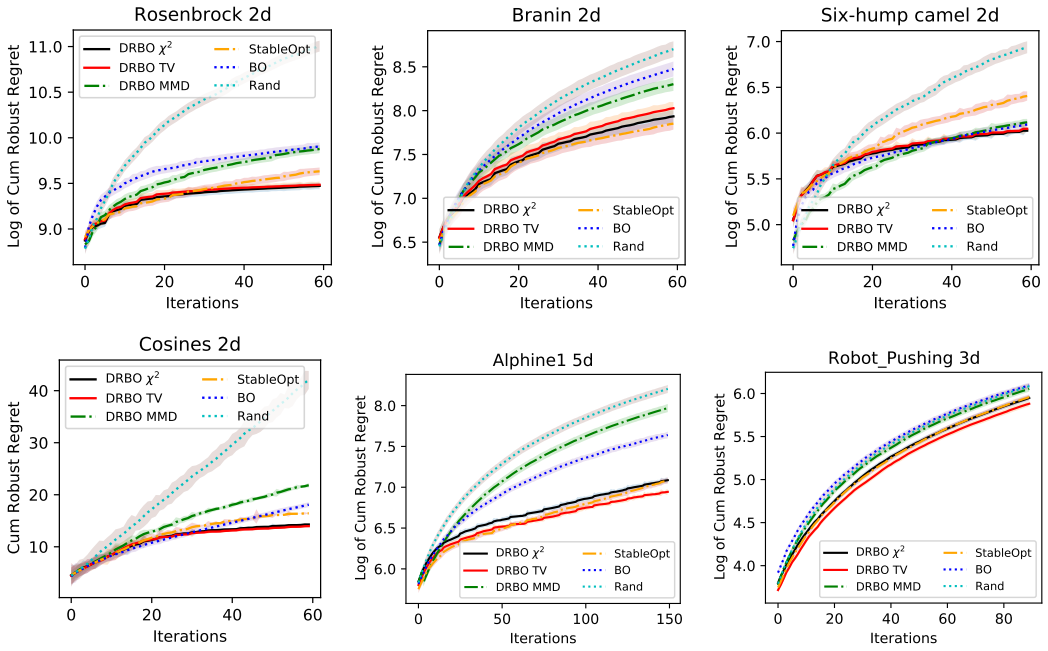


Figure 4: We compare the cumulative robust regret across algorithms. The results show that the proposed χ^2 and Total Variation (TV) achieves the best performance across benchmark functions. Random and vanilla BO approaches performs poorly which do not take into account the robustness criteria.

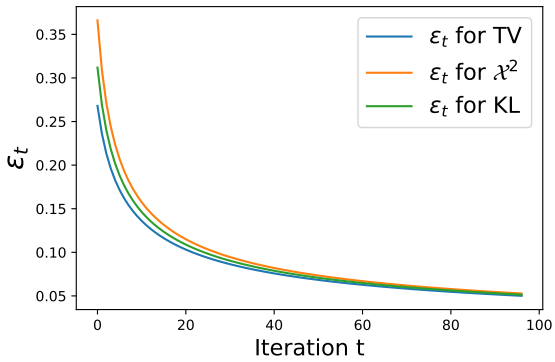
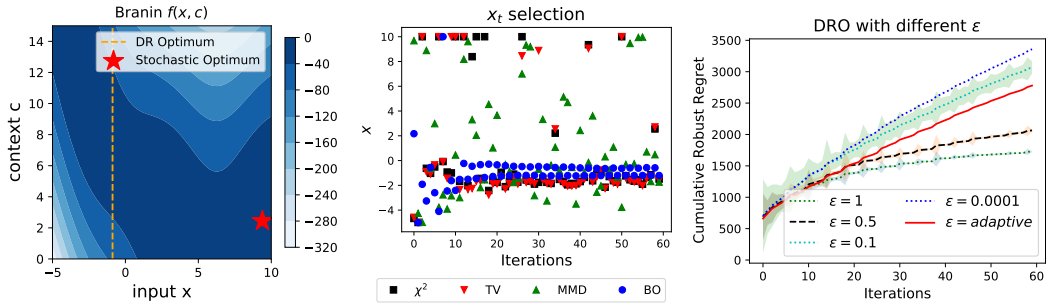
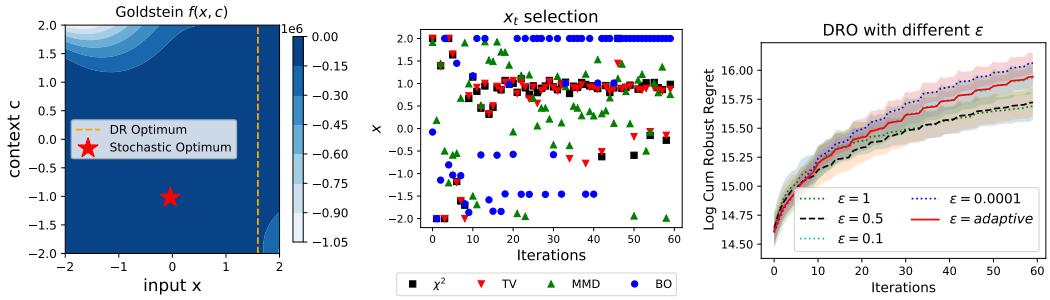


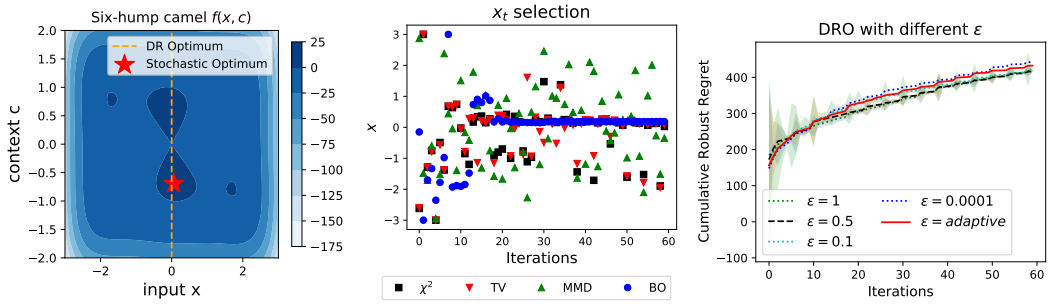
Figure 5: The adaptive value of ε_t over iterations, $\lim_{t\to\infty}, \varepsilon_t = 0$.



(a) Stochastic and DRO solution are different. The choices of $\varepsilon = \{0.5, 1\}$ result in the best performance.



(b) Stochastic and DRO solution are different. The choices of $\varepsilon = \{0.5, 1\}$ result in the best performance.



(c) Stochastic and DRO solution are coincide. $\varepsilon \to 0$ is the best.

Figure 6: We complement the result presented in Fig. 2 using three additional functions. There are two settings in DRO when the stochastic solution and robust solution are different (*top*) and identical (*bottom*). *Left*: the original function $f(\mathbf{x}, c)$. *Middle*: the selection of input $\mathbf{x}_t$ over iterations. *Right*: optimization performance with different ε. The adaptive choice of ε_t (in red) always produces stable performance across various choices of ε_t. This is especially useful in unknown functions where we do not have prior assumption on the underlying structure to decide on which large or small values of ε_t to be specified.

In Fig. 6, we present additional visualization to complement the analysis in Fig. 2. In particular, we illustrate three other functions including branin, goldstein and six-hump camel. The additional results are consistent with the finding presented in the main paper (Section 7.1 and Fig. 2).