Efficient Distributionally Robust Bayesian Optimization with Worst-case Sensitivity

Sebastian Shenghong Tay[1 2], Chuan Sheng Foo[2], Daisuke Urano[3], Richalynn Chiu Xian Leong[3], Bryan Kian Hsiang Low[1]

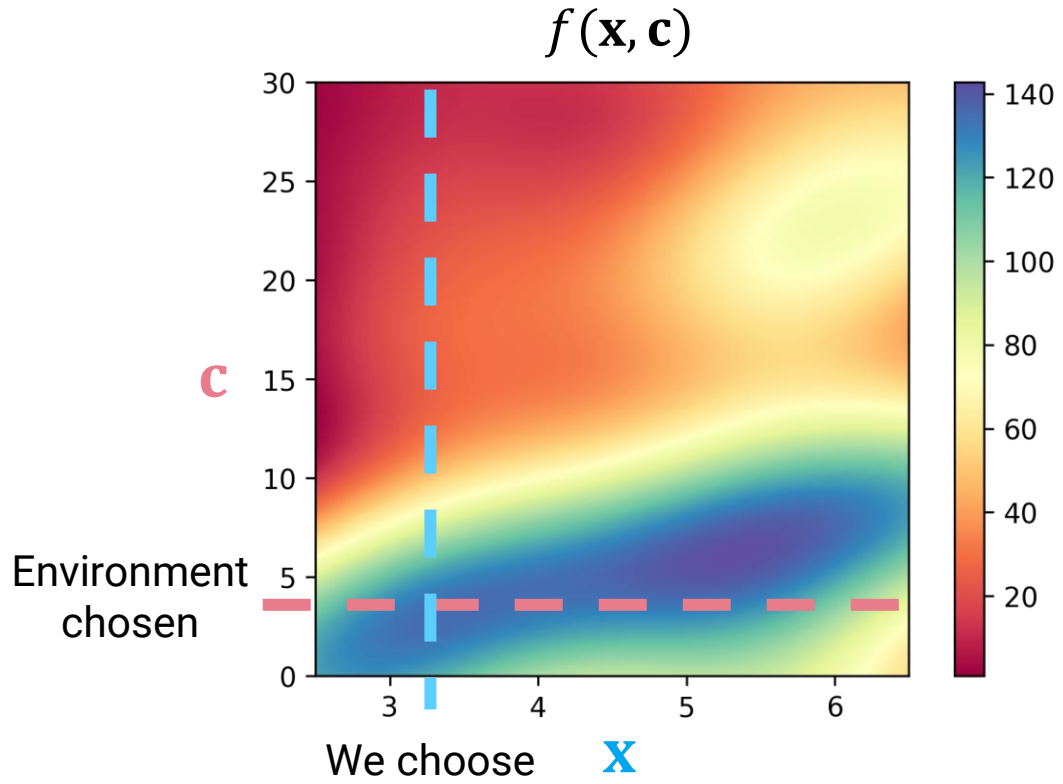
[1]Department of Computer Science, National University of Singapore, Singapore

[2]Institute of Infocomm Research, A*STAR, Singapore

[3]Temasek Life Sciences Laboratory, Singapore

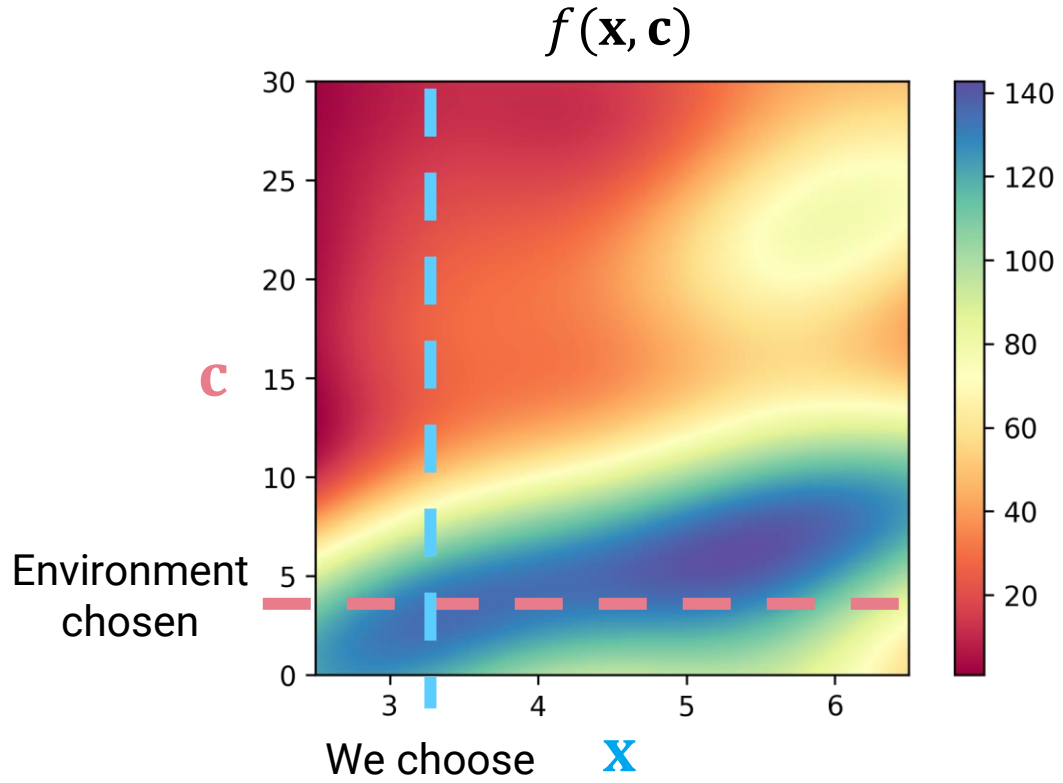
Motivation

- We have a **decision variable x** (within our control, e.g., crop nutrients) and a **context variable c** (uncontrollable, e.g., amount of sunlight in a day).



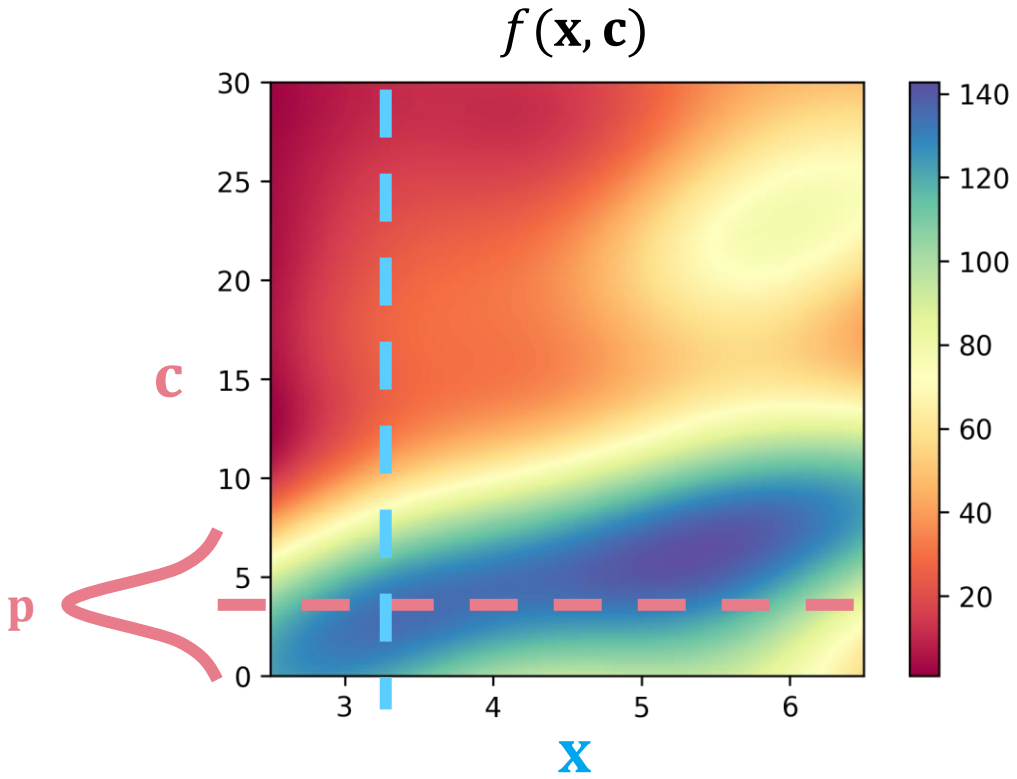
Motivation

- We have a **decision variable $\mathbf{x}$** (within our control, e.g., crop nutrients) and a **context variable $\mathbf{c}$** (uncontrollable, e.g., amount of sunlight in a day).
- We desire large $f(\mathbf{x}, \mathbf{c})$ (e.g., final size of crop). f is **unknown** and **costly to evaluate** (in terms of time, money etc.).



Stochastic BO

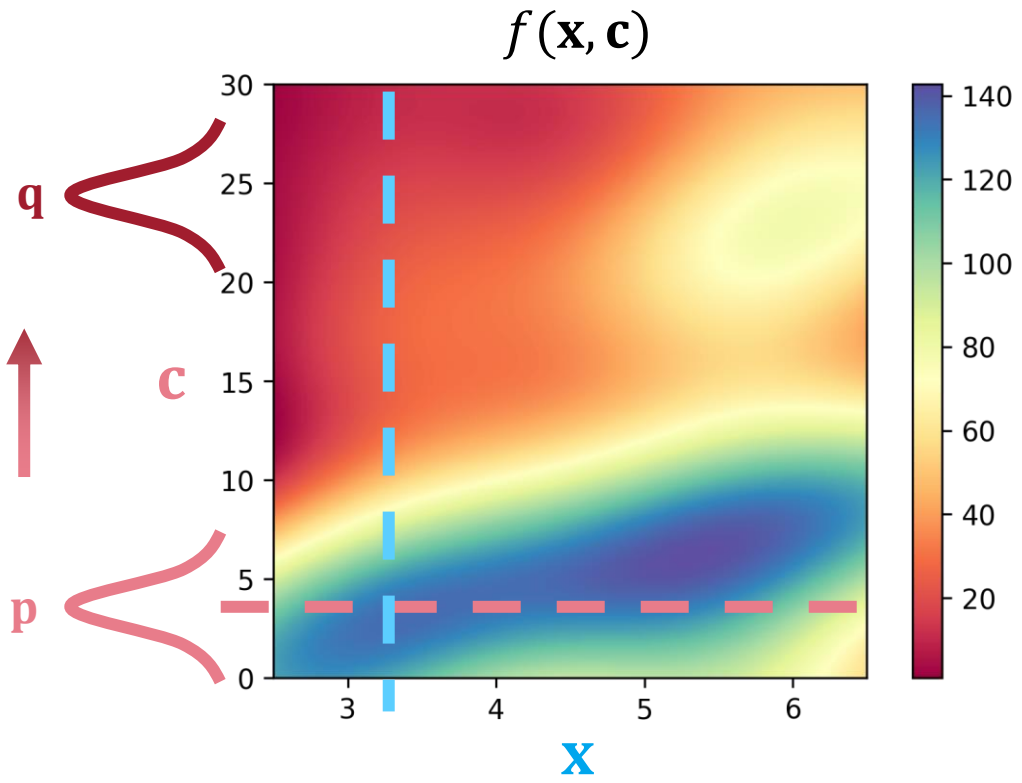
- The same idea applies if $\mathbf{c}$ is a random vector distributed according to **known distribution $\mathbf{p}$**. We may then desire to maximize the expected value $\mathbb{E}_{\mathbf{c}\sim\mathbf{p}}[f(\mathbf{x}, \mathbf{c})]$.



Distributionally robust BO

- Suppose the environment is an adversary that is allowed to choose the true distribution among a set of distributions known as the uncertainty set $\mathcal{U}$:

$$\mathcal{U} \coloneqq \{\mathbf{q}' \mid d(\mathbf{p}, \mathbf{q}') \leq \epsilon\}$$

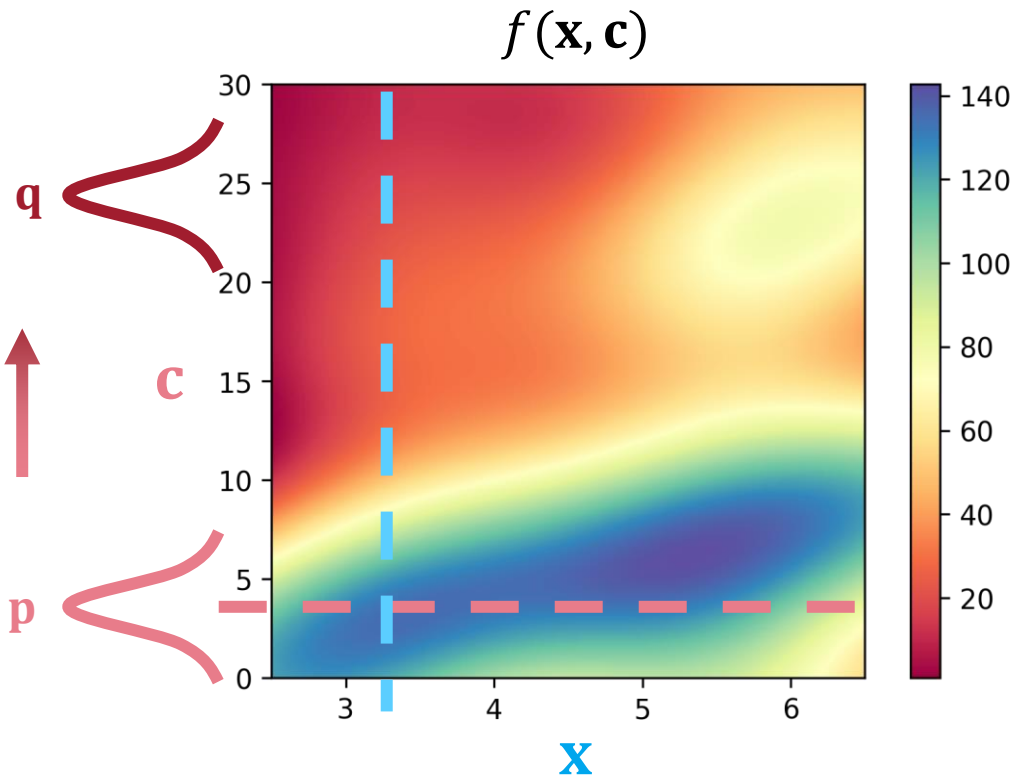


Distributionally robust BO

- Suppose the environment is an adversary that is allowed to choose the true distribution among a set of distributions known as the uncertainty set $\mathcal{U}$:

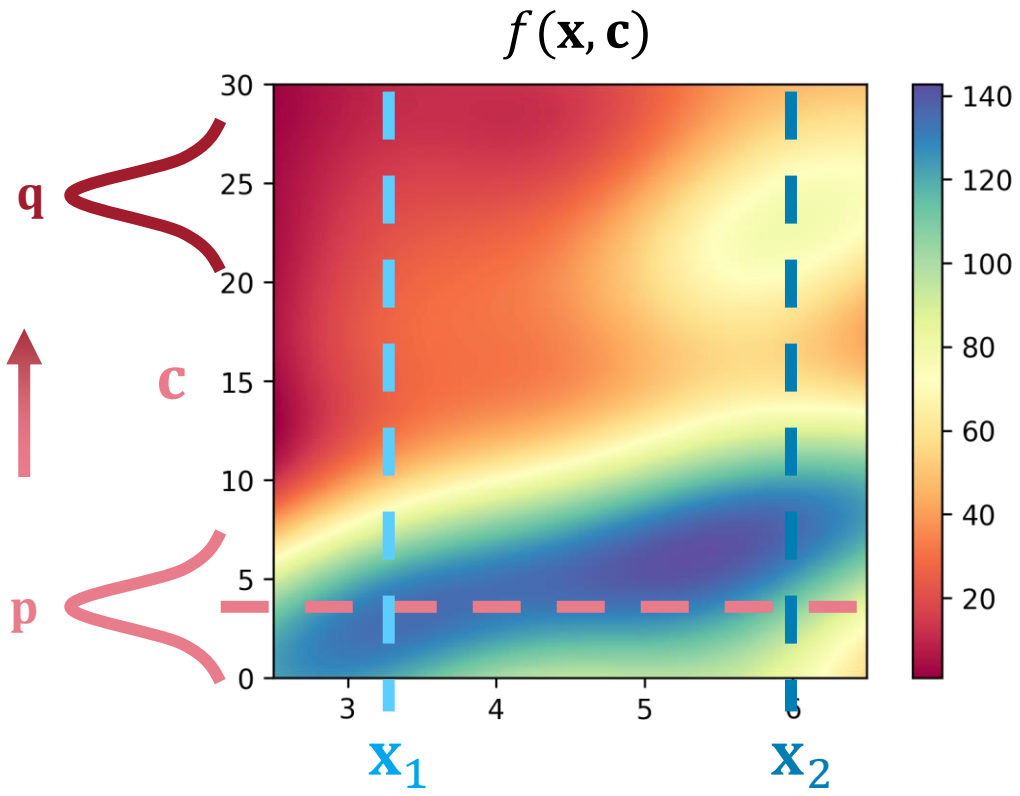
$$\mathcal{U} \coloneqq \{\mathbf{q}' \mid d(\mathbf{p}, \mathbf{q}') \leq \epsilon\}$$

- In the worst case, it chooses the distribution that minimizes our expected value, called the **worst-case distribution q.**



Distributionally robust BO

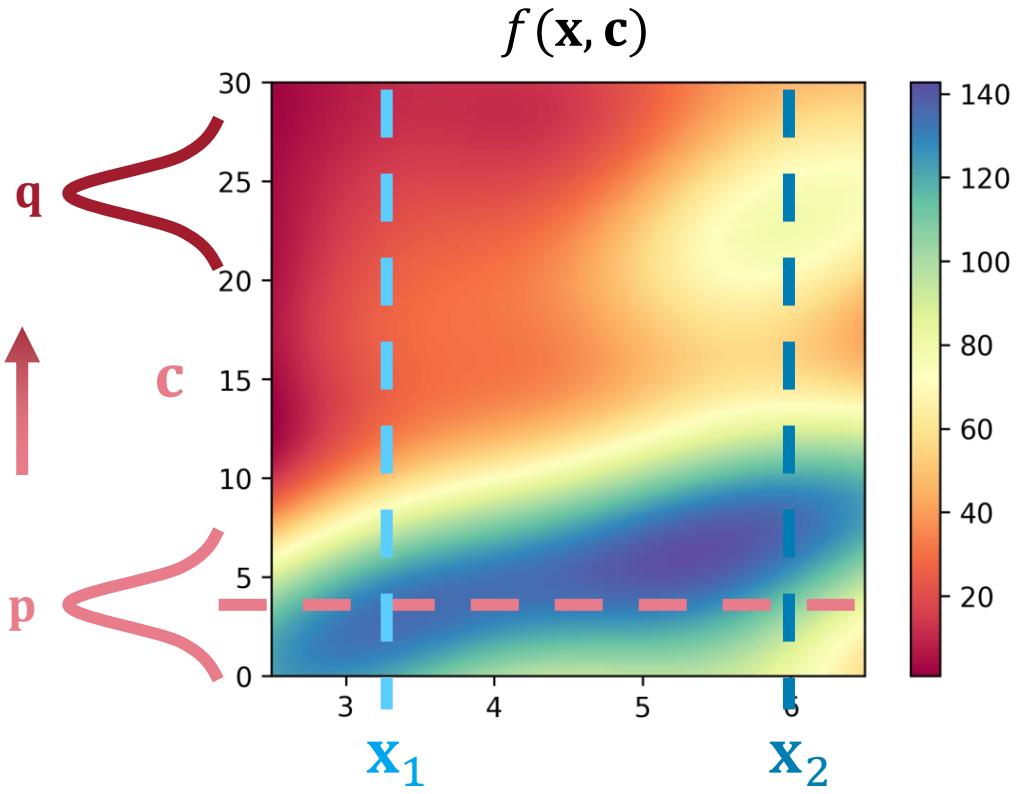
- $\mathbb{E}_{\mathbf{c}\sim\mathbf{p}}[f(\mathbf{x}_1, \mathbf{c})] > \mathbb{E}_{\mathbf{c}\sim\mathbf{p}}[f(\mathbf{x}_2, \mathbf{c})]$, but $\min_{\mathbf{q}\in\mathcal{U}} \mathbb{E}_{\mathbf{c}\sim\mathbf{q}}[f(\mathbf{x}_1, \mathbf{c})] < \min_{\mathbf{q}\in\mathcal{U}} \mathbb{E}_{\mathbf{c}\sim\mathbf{q}}[f(\mathbf{x}_2, \mathbf{c})]$, i.e., $\mathbf{x}_2$ is more distributionally robust.



Distributionally robust BO

- $\mathbb{E}_{c\sim p}[f(x_1, c)] > \mathbb{E}_{c\sim p}[f(x_2, c)]$, but $\min_{q\in\mathcal{U}} \mathbb{E}_{c\sim q}[f(x_1, c)] < \min_{q\in\mathcal{U}} \mathbb{E}_{c\sim q}[f(x_2, c)]$, i.e., x_2 is more distributionally robust.
- The learner is required to learn the optimal distributionally robust point

$$x^* := \max_{x\in\mathcal{X}} \min_{q\in\mathcal{U}} \mathbb{E}_{c\sim q}[f(x, c)].$$



General algorithm

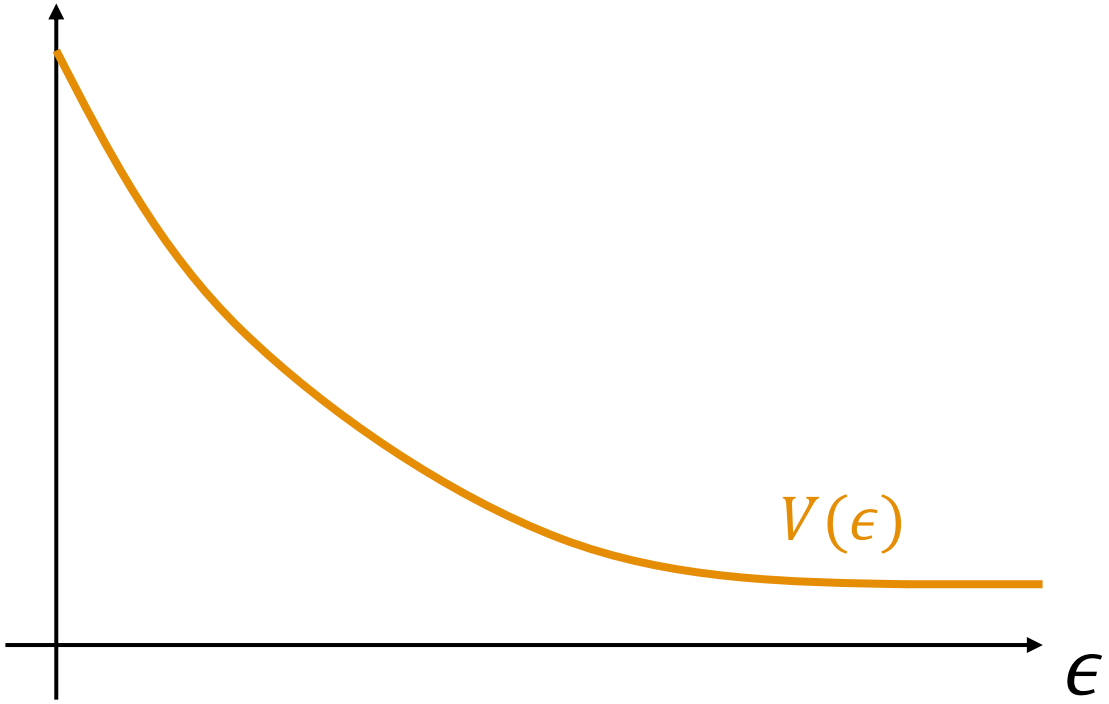
Algorithm 1 Generalized DRBO (Kirschner et al., 2020)

Input: GP with kernel k, score function α
for iteration $t = 1$ **to** T **do**
  Obtain reference distribution p_t and margin ϵ_t
  Compute $\text{ucb}_x^t := (\mu_t(x, \mathcal{C}_j) + \beta_t \sigma_t(x, \mathcal{C}_j))_{j=1,\dots,|\mathcal{C}|}^\top$
  Select action $x_t = \text{argmax}_{x \in \mathcal{X}} \alpha(\text{ucb}_x^t, p_t, \epsilon_t)$
  Observe $c_t \sim p_t^*$ and $y_t = f(x_t, c_t) + \xi_t$
  Update GP posterior with $\mathcal{D}_{t+1} := \{(x_i, c_i, y_i)\}_{i=1}^t$
end for

- In Kirschner et al.[1], $\alpha = \min_{\mathbf{q} \in \mathcal{U}_t} \mathbb{E}_{\mathbf{c} \sim \mathbf{q}}[\text{ucb}^t(\mathbf{x}, \mathbf{c})]$ which requires solving a convex optimization problem with a discretized context set $\mathcal{C}$
- Solving general convex optimization problems with interior-point methods incurs $\mathcal{O}(|\mathcal{C}|^3)$ time. Scales poorly with $|\mathcal{C}|$

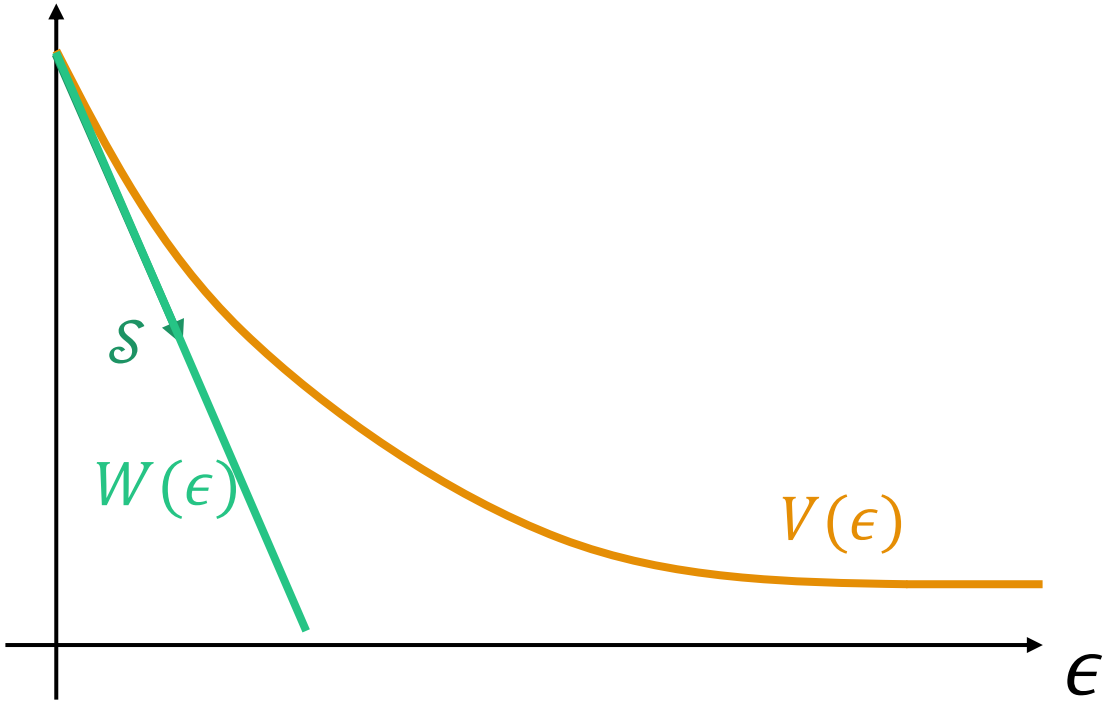
[1] Kirschner, J., Bogunovic, I., Jegelka, S., & Krause, A. (2020, June). Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics* (pp. 2174-2184). PMLR.

Approximating the convex opt. solution



- Define the **worst-case expected value** $V(\epsilon) \coloneqq \min_{\mathbf{q} \in \mathcal{U}} \mathbb{E}_{\mathbf{c} \sim \mathbf{q}}[\mathrm{ucb}(\mathbf{x}, \mathbf{c})]$.
- $V(\epsilon)$ is convex with respect to the margin ϵ (size of the uncertainty set $\mathcal{U}$) when d is convex

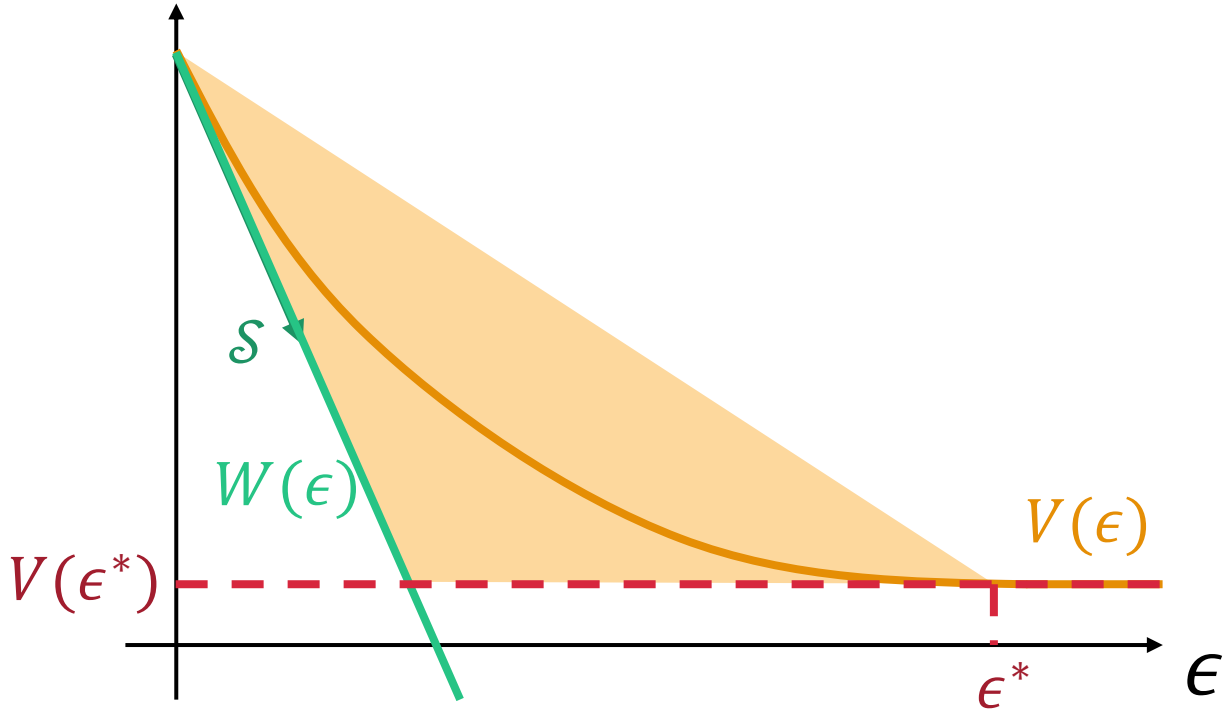
Approximating the convex opt. solution



- Gotoh et al. (2020)[1] defined the **worst-case sensitivity** $\mathcal{S}$ as the gradient of $V(\epsilon)$ as $\epsilon \to 0$, and derived closed forms of $\mathcal{S}$ for many distribution distances.
- Since $V(\epsilon)$ is convex, we can lower bound $V(\epsilon)$ with its linear approximation around $\epsilon = 0$ with the function $W(\epsilon)$ constructed using $\mathcal{S}$.

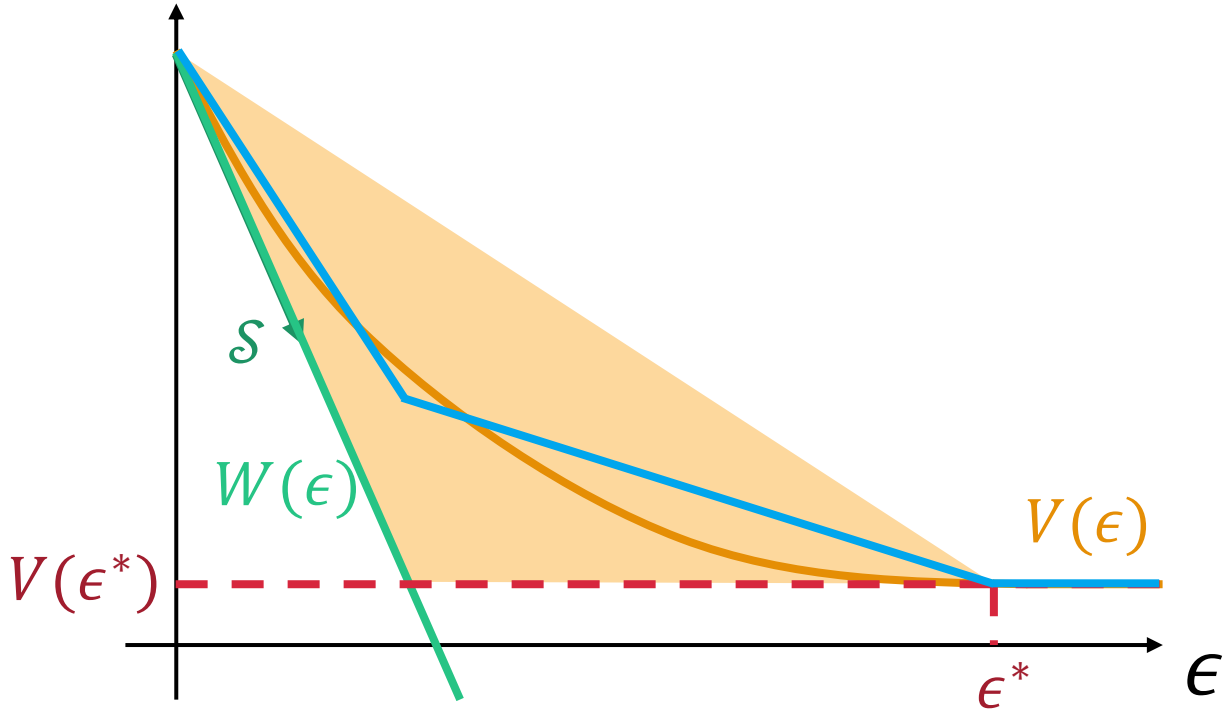
[1]Gotoh, J., Kim, M. J., and Lim, A. E. B. Worst-case sensitivity. arXiv:2010.10794, 2020.

Approximating the convex opt. solution



- However, the approximation $W(\epsilon)$ gets arbitrarily worse as ϵ increases. We can refine it further by computing (cheaply) ϵ^* and $V(\epsilon^*)$ (worst possible value of $V(\epsilon)$). By the convexity of $V(\epsilon)$, we can then upper and lower bound $V(\epsilon)$ into a region termed the **valid region**.

Approximating the convex opt. solution



- Our approximation titled **MinimaxApprox** is then a piece-wise linear bisection of the valid region. This minimizes the maximum possible approximation error incurred.

Computational efficiency

Table 1. Comparing time complexity of DRBO algorithms utilizing the EXACT worst-case expected value (Kirschner et al., 2020) vs. our fast approximation called MINIMAXAPPROX with various distribution distances d. The EXACT worst-case expected value is obtained by solving a general convex optimization problem with $|\mathcal{C}|$ variables using interior point methods which, we assume, incur $\mathcal{O}(|\mathcal{C}|^3)$ time.

Distribution distance d	EXACT	MINIMAXAPPROX
Maximum mean discrepancy (MMD)	$\mathcal{O}(\|\mathcal{C}\|^3)$	$\mathcal{O}(\|\mathcal{C}\|^2)$
Total variation (TV)	$\mathcal{O}(\|\mathcal{C}\|^3)$	$\mathcal{O}(\|\mathcal{C}\|)$
Modified χ^2-divergence (χ^2)	$\mathcal{O}(\|\mathcal{C}\|^3)$	$\mathcal{O}(\|\mathcal{C}\|)$
Wasserstein metric ($\mathcal{W}$)	$\mathcal{O}(\|\mathcal{C}\|^6)$	$\mathcal{O}(\|\mathcal{C}\|^2)$

Approximation quality

- Robust regret is bounded by

$$R_T \leq 4\beta_T \sqrt{T\left(\gamma_T + 4\log\frac{12}{\delta}\right)} + \sum_{t=1}^{T}(2B'_{d,t}\epsilon_{d,t} + 2A^{\max}_{d,t})$$

- Scales linearly in T, however so does the robust regret of the previous work in the same setting. Our approximation does no worse than the exact solution in terms of dependence on T.
- Confirms intuition that better approximation ($A^{\max}_{d,t}$) ultimately leads to better robust regret.

Choice of distribution distance

- Worst-case sensitivity has interpretable meanings: for e.g., worst-case sensitivity when d is χ^2-divergence is the **variance** of outcome values $\mathbf{g}$, while that when using total variation (TV) is the **range**.

Choice of distribution distance

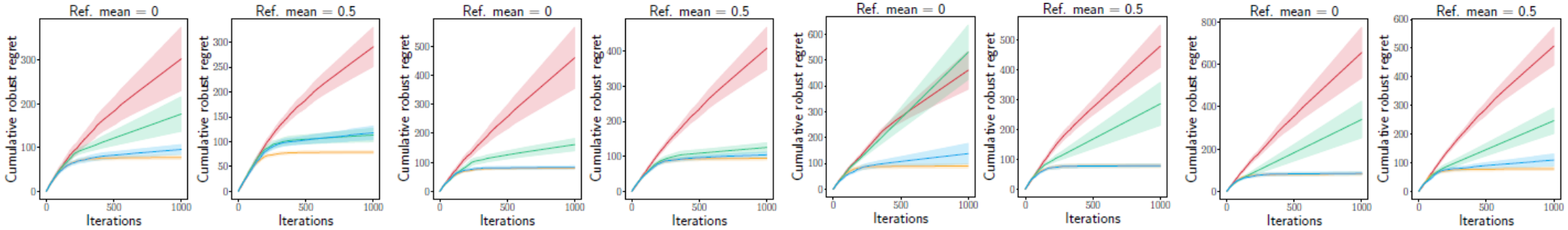
- Worst-case sensitivity has interpretable meanings: for e.g., worst-case sensitivity when d is χ^2-divergence is the **variance** of outcome values $\mathbf{g}$, while that when using total variation (TV) is the **range**.
- Denoting MinimaxApprox as $\hat{V}$, in some regimes of ϵ, $\hat{V}$ can be re-written as a linear combination of the **expected value**, **worst-case sensitivity**, and the **worst value**. Presence of worst-case sensitivity term provides interpretability and guides choice of d.

$$\widehat{V}_d(\epsilon_d, g) = \left(1 - \frac{\epsilon_d}{2\epsilon_d^*}\right) \underbrace{\mathbb{E}_p[g]}_{\text{Expected value}} + \frac{\epsilon_d}{2} \underbrace{\mathcal{S}_d(g)}_{\text{Worst-case sensitivity}} + \frac{\epsilon_d}{2\epsilon_d^*} \underbrace{\min_i [g]_i}_{\text{Worst value}}$$

Experiments: Robust regret

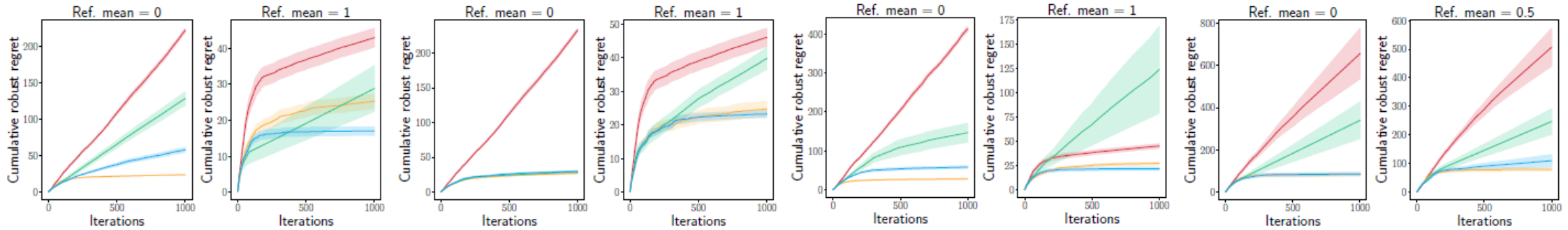
Comparing the robust regret of **stochastic GP-UCB**, ***W***, **Exact** (previous work) and **MinimaxApprox** (ours)

Synthetic random functions:



(a) $d = \text{MMD}$ (b) $d = \text{TV}$ (c) $d = \chi^2$ (d) $d = \mathcal{W}$

Plant maximum leaf area:

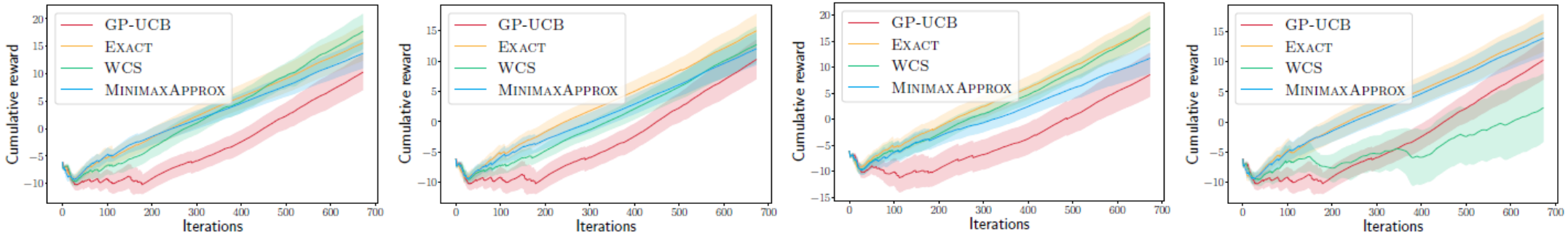


(a) $d = \text{MMD}$ (b) $d = \text{TV}$ (c) $d = \chi^2$ (d) $d = \mathcal{W}$

Experiments: Robust regret

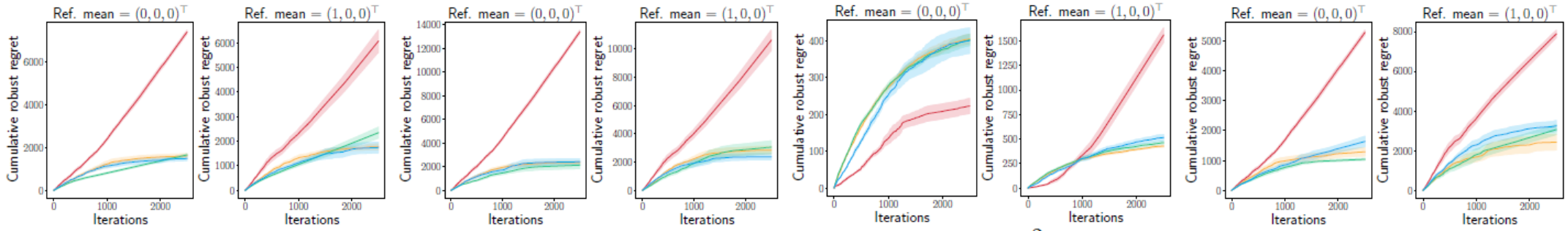
Comparing the robust regret of **stochastic GP-UCB**, ***W***, **Exact** (previous work) and **MinimaxApprox** (ours)

Wind power dataset:



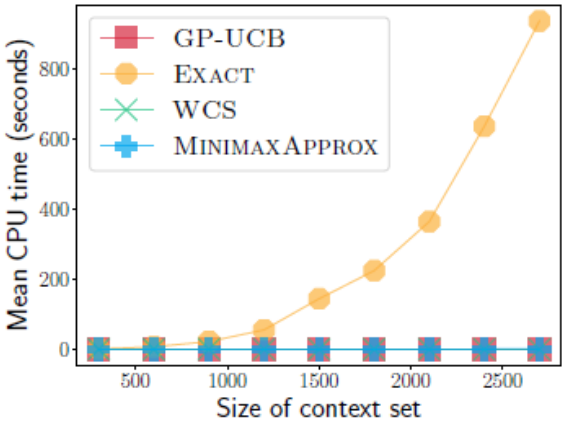
(a) $d = \text{MMD}$ (b) $d = \text{TV}$ (c) $d = \chi^2$ (d) $d = \mathcal{W}$

COVID-19 test allocation:

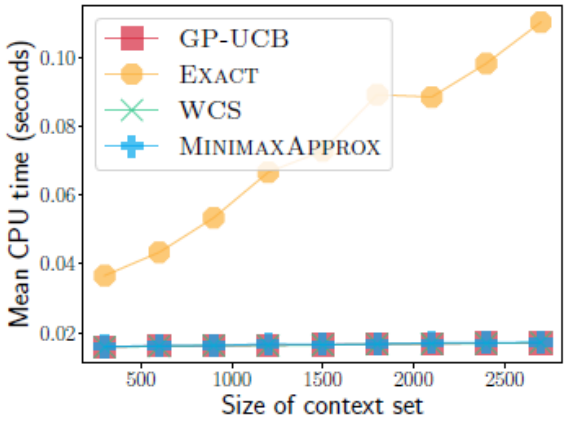


(a) $d = \text{MMD}$ (b) $d = \text{TV}$ (c) $d = \chi^2$ (d) $d = \mathcal{W}$

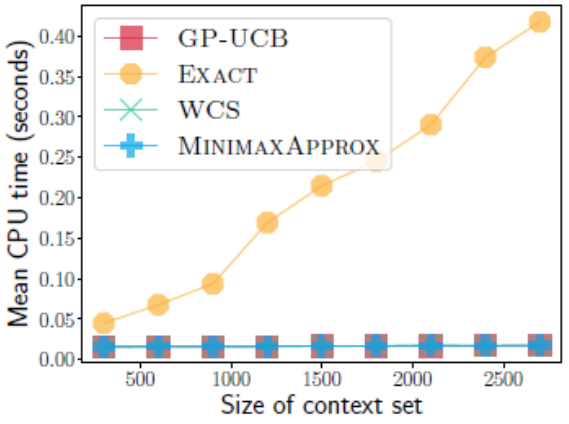
Experiments: Computation time



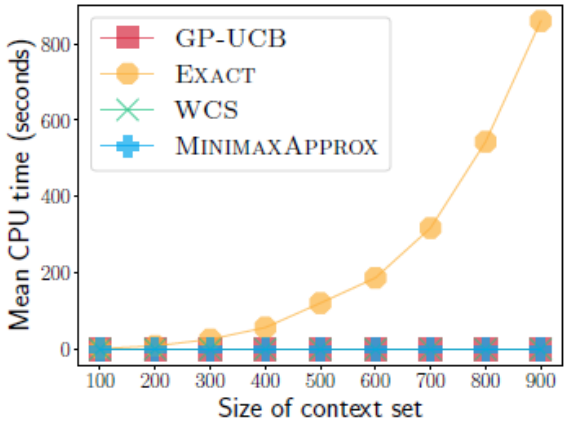
(a) $d = \text{MMD}$



(b) $d = \text{TV}$

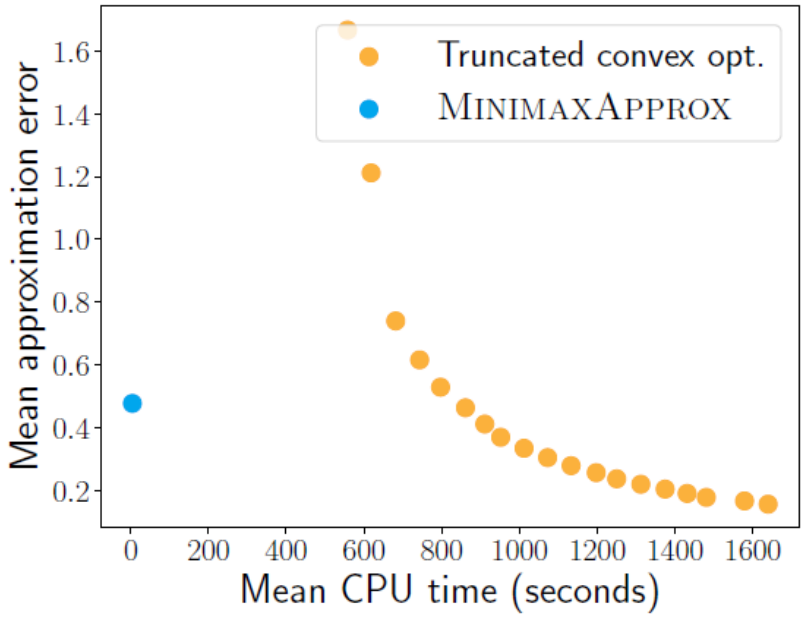


(c) $d = \chi^2$

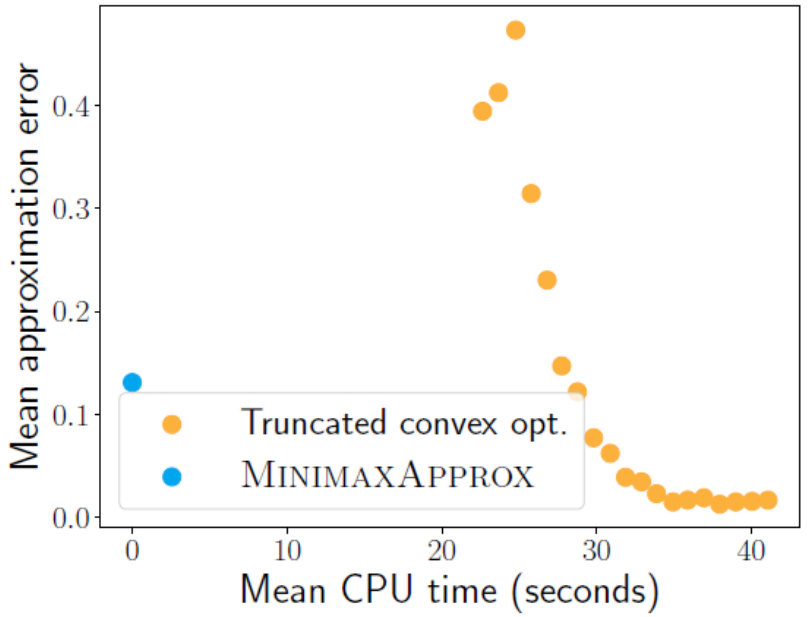


(d) $d = \mathcal{W}$

Experiments: Error-time trade-off



(a) $d = \text{MMD}$



(b) $d = \mathcal{W}$

Summary

- Distributionally robust Bayesian optimization (DRBO) is a novel setting for Bayesian optimization (BO) with stochastic context variables.

Summary

- Distributionally robust Bayesian optimization (DRBO) is a novel setting for Bayesian optimization (BO) with stochastic context variables.
- We borrow a concept from the distributionally robust optimization (DRO) literature known as **worst-case sensitivity** to formulate a fast algorithm.
  - Theoretical bounds
  - Empirically competitive with the previous method[1] while incurring **significantly less computation time**

[1] Kirschner, J., Bogunovic, I., Jegelka, S., & Krause, A. (2020, June). Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics* (pp. 2174-2184). PMLR.

Summary

- Distributionally robust Bayesian optimization (DRBO) is a novel setting for Bayesian optimization (BO) with stochastic context variables.
- We borrow a concept from the distributionally robust optimization (DRO) literature known as **worst-case sensitivity** to formulate a fast algorithm.
  - Theoretical bounds
  - Empirically competitive with the previous method[1] while incurring **significantly less computation time**
- To guide the choice of distribution distance in DRBO (model selection problem), we show that our approximation implicitly optimizes an objective close to an **interpretable risk-sensitive value**.

[1] Kirschner, J., Bogunovic, I., Jegelka, S., & Krause, A. (2020, June). Distributionally robust Bayesian optimization. In *International Conference on Artificial Intelligence and Statistics* (pp. 2174-2184). PMLR.

Thank You