Distributionally Robust Bayesian Quadrature Optimization

Thanh Tang Nguyen **Sunil Gupta** **Huong Ha** **Santu Rana** **Svetha Venkatesh**
Applied Artificial Intelligence Institute (A^2I^2)
Deakin University, Geelong, Australia

Abstract

Bayesian quadrature optimization (BQO) maximizes the expectation of an expensive black-box integrand taken over a known probability distribution. In this work, we study BQO under distributional uncertainty in which the underlying probability distribution is unknown except for a limited set of its i.i.d. samples. A standard BQO approach maximizes the Monte Carlo estimate of the true expected objective given the fixed sample set. Though Monte Carlo estimate is unbiased, it has high variance given a small set of samples; thus can result in a spurious objective function. We adopt the distributionally robust optimization perspective to this problem by maximizing the expected objective under the most adversarial distribution. In particular, we propose a novel posterior sampling based algorithm, namely distributionally robust BQO (DRBQO) for this purpose. We demonstrate the empirical effectiveness of our proposed framework in synthetic and real-world problems, and characterize its theoretical convergence via Bayesian regret.

1 INTRODUCTION

Making robust decisions in the face of parameter uncertainty is critical to many real-world decision problems in machine learning, engineering and economics. Besides the uncertainty that is inherent in data, a further difficulty arises due to the uncertainty over contexts. A common example is hyperparameter selection of machine learning algorithms where cross-validation is performed using a small to medium sized validation set. Due to limited size of validation set, the variance across different folds might be high. Ignoring this uncertainty results in sub-optimal and non-robust decisions. This problem in practice can be further exacerbated as the outcome measurements may be noisy and the black-box function itself is expensive to evaluate.

One way to capture the uncertainty in the contexts is through a probability distribution. In this work, we consider stochastic black-box optimization that is distributionally robust to misspecification in the context distribution. We formulate the problem as

$$\max_{x\in\mathcal{X}\subset\mathbb{R}^d} g(x) := \max_{x\in\mathcal{X}} \mathbb{E}_{P_0(w)}[f(x,w)], \tag{1}$$

where f is an expensive black-box function and $P_0(w)$ is a distribution over context w. We assume distributional uncertainty in which the context distribution P_0 is known only through a limited set of its i.i.d samples $S_n = \{w_1, ..., w_n\}$. This is equivalent to the scenario in which we are able to evaluate f only on $\mathcal{X} \times S_n$ during optimization.

In the case that P_0 is known (e.g., P_0 is either available in an analytical form or easy to evaluate), a standard solution to the problem in Equation (1) is based on Bayesian quadrature [O'Hagan, 1991; Rasmussen and Ghahramani, 2002; Oates et al., 2016; Oates and Sullivan, 2019]. The main idea in this approach is that we can build a Gaussian Process (GP) model of f and use the known relationship in the integral to imply a second GP model of g. This is possible because integration is a linear operator.

Given the distributional uncertainty in which P_0 is only known through a limited set of its samples, a naive approach to the problem in Equation (1) is to maximize its Monte Carlo estimate:

$$g_{mc}(x) := \mathbb{E}_{\hat{P}_n(w)}[f(x,w)], \tag{2}$$

where $\hat{P}_n(w) = \frac{1}{n}\sum_{i=1}^{n}\delta(w - w_i)$ and $\delta(.)$ is the Dirac distribution. When n is sufficiently large, $g_{mc}(x)$ approximates $g(x)$ reasonably well as guaranteed by the weak law of large numbers; thus, the optimal solution of $g_{mc}(x)$ represents that of $g(x)$. In contrast, when n

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

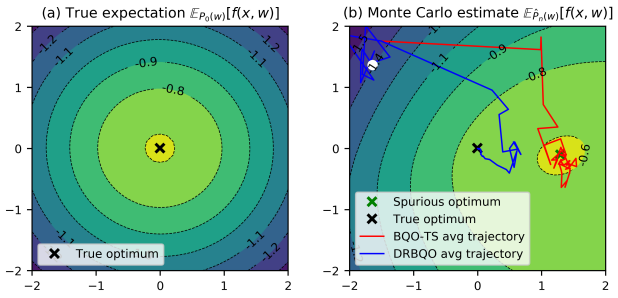

Figure 1: (a) The expected logistic function $g(x) = \mathbb{E}_{\mathcal{N}(w;0,I)}[-\log(1+e^{x^T w})]$ and (b) its Monte Carlo estimate using 10 samples of w, and the averaged trajectories of our proposed algorithm DRBQO (detailed in Section 4) and a standard Bayesian Quadrature Optimization (BQO) baseline. Though being unbiased, Monte Carlo estimates can suffer from high variance given limited samples, resulting in spurious function estimates. Our proposed algorithm DRBQO approaches this mismatch problem by finding the distributionally robust solution under the most adversarial distribution over a χ^2 distributional ball.

is small, the optimal solution of $g_{mc}(x)$ might be suboptimal to $g(x)$. Since we are considering distributional perturbation, we cannot guarantee the Monte Carlo estimate $g_{mc}(x)$ to be a good surrogate objective.

A more conservative approach from statistical learning is to maximize the variance-regularized objective:

$$g_{bv}(x) := \mathbb{E}_{\hat{P}_n}[f(x,w)] - C_1\sqrt{Var_{\hat{P}_n}[f(x,w)]/n}, \quad (3)$$

where $Var_{\hat{P}_n}$ denotes the empirical variance and C_1 is a constant determining the trade-off between bias and variance. Thus, given the context of limited samples, it is logical to use $g_{bv}(x)$ instead of $g_{mc}(x)$ as a surrogate objective for maximizing $g(x)$. However, unlike $g_{mc}(x)$, the variance term in $g_{bv}(x)$ breaks the linear relationship with f. As a result, though f is a Gaussian Process, $g_{bv}(x)$ need not be [O'Hagan, 1991].

Alternatively, we approach the distributional uncertainty problem above by formulating the distributionally robust Bayesian quadrature optimization. In the face of the uncertainty about P_0, we seek to find a distributionally robust solution under the most adversarial distribution. Our approach is based on solving a surrogate distributionally robust optimization problem generated by posterior sampling at each time step. The surrogate optimization is solved efficiently via bisection search through any optimization. We demonstrate the efficiency of our algorithm in both synthetic and real-world problems. Our contributions are:

- Demonstrating the limitations of standard Bayesian quadrature optimization algorithms under distributional uncertainty (Section 3), and introducing a new algorithm, namely DRBQO, that overcomes these limitations (Section 4);
- Introducing the concept of ρ-regret for measuring algorithmic performance in this formulation (Section 3), and characterizing the theoretical convergence of our proposed algorithm in sublinear Bayesian regret (Section 5);
- Demonstrating the efficiency of DRBQO in finding distributionally robust solutions in both synthetic and real-world problems (Section 6).

2 RELATED WORK

Our work falls in the area of Bayesian quadrature optimization whose goal is to perform black-box global optimization of an expected objective of the form $\int f(x,w)P(w)dw$. This type of problems is known with various names such as optimization of integrated response functions [Williams, 2000], multi-task Bayesian optimization [Swersky et al., 2013], and optimization with expensive integrands [Toscano-Palmerin and Frazier, 2018]. This direction approaches the problem by evaluating $f(x,w)$ at one or several values of w given x. This ameliorates the need of evaluating $f(x,w)$ at all the values of w and can outperform methods that evaluate the full objective via numerical quadrature [Frazier, 2018; Toscano-Palmerin and Frazier, 2018]. While the previous approaches assume the knowledge of the context distribution, we are interested in the distributional uncertainty scenario in which the underlying distribution is unknown except for its empirical estimate.

Our work shares similarity with the distributionally robust optimization (DRO) literature [Rahimian and Mehrotra, 2019]. This problem setup considers the parameter uncertainty in real-world decision making problems. The uncertainty may be due to limited data and noisy measurements. DRO takes into account this uncertainty and approaches the problem by taking the worst-case of the underlying distribution within an uncertainty set of distributions. DRO variants distinguish each other in design choices of the distributional uncertainty set and in problem contexts. Regarding the design of uncertainty sets, common designs specify the set of distributions with respect to the nominal distribution via distributional discrepancy such as χ^2 divergence [Namkoong and Duchi, 2016], Wasserstein distance [Kuhn et al., 2019], and maximum mean discrepancy [Staib and Jegelka, 2019]. Regarding problem contexts, DRO has been studied in various problem settings such as robust optimization [Ben-Tal et al.,

2013], robust risk minimization [Namkoong and Duchi, 2016], sub-modular maximization [Staib et al., 2019], boosting algorithms [Blanchet et al., 2019], graphical models [Fathony et al., 2018], games [Sun and Boyd, 2018; Zhu et al., 2019], fairness in machine learning [Hashimoto et al., 2018], Markov decision process [Xu and Mannor, 2010], reinforcement learning [Smirnova et al., 2019], neural networks [Sagawa* et al., 2020], meta-learning [Collins et al., 2020], and Bayesian optimization [Kirschner et al., 2020]. The distinction of our work is in terms of the problem context where we study DRO in Bayesian quadrature optimization.

3 PROBLEM SETUP

Model. Let $f : \mathcal{X} \times \Omega \to \mathbb{R}$ be an element of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ where k: $\mathcal{X} \times \Omega \times \mathcal{X} \times \Omega \to \mathbb{R}$ is a positive-definite kernel, and $\mathcal{X}$ and Ω are, unless explicitly mentioned otherwise, compact domains in $\mathbb{R}^d$ and $\mathbb{R}^m$ for some dimensions d and m, respectively. We further assume that k is continuous and bounded from above by 1, and that $\|f\|_k = \sqrt{\langle f, f\rangle_k} \leq B$ for some $B > 0$. Two commonly used kernels are Squared Exponential (SE) and Matérn [Rasmussen and Williams, 2006] which are similarly defined on $\mathcal{X} \times \Omega$ as follows:

$$k_{SE}(.,.;.,.) = \exp(-d^2_{\theta,\psi}(.,.;.,.)),$$
$$k_\nu(.,.;.,.) = \frac{2^{1-\nu}}{\Gamma(\nu)}\sqrt{2\nu}d_{\theta,\psi}(.,.;.,.)J_\nu(\sqrt{2\nu}d_{\theta,\psi}(.,.;.,.)),$$

where θ and ψ are the length scales, $\nu > 0$ defines the smoothness in the Matérn kernel, $J(\nu)$ and $\Gamma(\nu)$ define the Bessel function and the gamma function, respectively, and $d^2_{\theta,\psi}(x, w; x', w') = \sum_{i=1}^d (x_i - x'_i)^2/\theta_i^2 + \sum_{j=1}^m (w_j - w'_j)^2/\psi_j^2$.

Let P_0 be a distribution on Ω, and $S_n = \{w_1, ..., w_n\}$ be a fixed set of samples drawn from P_0. Though f is defined on $\mathcal{X} \times \Omega$, we are interested in the distributional uncertainty scenario in which we can query f only on $\mathcal{X} \times S_n$ during optimization. At time t, we query f at $(x_t, w_t) \in \mathcal{X} \times S_n$ and observe a noisy reward $y_t = f(x_t, w_t) + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Our goal is to find a robust solution point $x \in \mathcal{X}$ such that $\mathbb{E}_{P(w)}[f(x, w)]$ remains high even under the most adversarial realization of the unknown distribution P_0.

Given a sequence of noisy observations $(x_i, w_i, y_i)_{i=1}^t$, the posterior distribution under a GP$(0,\ k(.,.;.,.))$ prior is also also a GP with the following posterior mean and covariance:

$$\mu_t(x, w) = k_t(x, w)^T(K_t + \sigma^2 I)^{-1} y_{1:t},$$
$$C_t(x, w; x', w') = k(x, w; x', w') - k_t(x, w)^T(K_t + \sigma^2 I)^{-1} k_t(x', w'),$$

where $y_{1:t} = (y_1, ..., y_t)$, $k_t(x, w) = [k(x_i, w_i; x, w)]_{i=1}^t$, and $K_t = [k(x_i, w_i; x_j, w_j)]_{1 \leq i,j \leq t}$ is the kernel matrix.

We define the quadrature functional as

$$g(f, x, P) := \int P(w|x) f(x, w) dw, \tag{4}$$

for any conditional distribution $P(.|x)$ on Ω for all $x \in \mathcal{X}$, i.e., $P \in \mathcal{P}_{n,\rho} \times \mathcal{X}$. As an extended result of Bayesian quadrature [O'Hagan, 1991], for any conditional distribution $P \in \mathcal{P}_{n,\rho} \times \mathcal{X}$, $g(f, x, P)$ also follows a GP with the following mean and variance:

$$\mu_t(x, P) := \mathbb{E}_t[g(f, x, P)] = \int P(w|x)\mu_t(x, w)dw \tag{5}$$
$$\sigma_t^2(x, P) := Var_t[g(f, x, P)]$$
$$= \int\int P(w|x)P(w'|x)C_t(x, w; x, w')dwdw'. \tag{6}$$

Optimization goal. We seek to optimize the expected function under the most adversarial distribution over some distributional uncertainty set $\mathcal{P}_{n,\rho} := \{P|D(P, \hat{P}_n) \leq \rho\}$:

$$\max_{x \in \mathcal{X}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)], \tag{7}$$

where $\hat{P}_n(w) = \frac{1}{n}\sum_{i=1}^n \delta(w - w_i)$ is the empirical distribution, $\rho \geq 0$ is the confidence radius around the empirical distribution with respect to a distribution divergence $D(.,.)$ such as Wasserstein distance, maximum mean discrepancy, and ϕ-divergence. We can interpret $\mathcal{P}_{n,\rho}$ as the set of perturbed distributions with respect to the empirical distribution $\hat{P}_n$ within a confidence radius ρ. We then seek a robust solution in the face of adversarial distributional perturbation within $\mathcal{P}_{n,\rho}$.

For any distribution divergence choice D, we define a ρ-robust point to be any x^*_ρ such that

$$x^*_\rho \in \arg\max_{x \in \mathcal{X}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)]. \tag{8}$$

Our goal is to report after time t a distributionally robust point x_t in the sense that it has small ρ-**regret**, which is defined as

$$r_\rho(x) = g(f, x^*_\rho, P^*) - g(f, x, P^*), \tag{9}$$

where $P^*(.|x) = \arg\min_{P \in \mathcal{P}_{n,\rho}} \sum_w P(w|x)f(x, w), \forall x$.

While our framework can be adopted to various distribution divergences, we focus on the specific case when D is χ^2-divergence: $D(P, Q) = \frac{1}{2}\int_\Omega (\frac{dP}{dQ} - 1)^2 dQ, \forall P, Q$. From here on, we refer $\mathcal{P}_{n,\rho}$ as the χ^2 ball with D being χ^2-divergence. In particular, the distributionally robust optimization problem in Equation (7) is equivalent to the variance-regularized optimization in Equation (3) when the variance is sufficiently high, as justified by the following theorem:

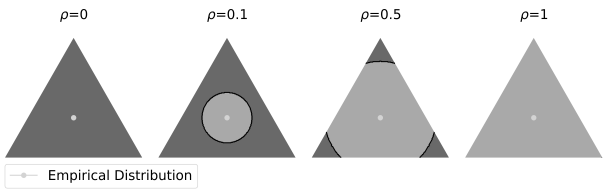

Figure 2: The χ^2 balls with various radii ρ on the n-dimensional simplex ($n = 3$ in this example). The simplex, the χ^2 balls and the empirical distribution are represented in dim gray, dark gray and light gray color, respectively. The χ^2 ball with $\rho = 0$ reduces to a singleton containing only the empirical distribution while the ball becomes the entire simplex for $\rho \geq \frac{n-1}{2}$.

Theorem 1 (Modified from Namkoong and Duchi [2017]). *Let $Z \in [M_0, M_1]$ be a random variable (e.g., $Z = f(x, w)$ for any fixed x), $\rho \geq 0$, $M = M_1 - M_0$, $s_n^2 = Var_{\hat{P}_n}[Z]$, $s^2 = Var[Z]$, and $OPT = \inf_P \{\mathbb{E}_P[Z] : P \in \mathcal{P}_{n,\rho}\}$. Then $\max\left\{\sqrt{2\rho s_n^2} - 2M\rho, 0\right\} \leq \mathbb{E}_{\hat{P}_n}[Z] - OPT \leq \sqrt{2\rho s_n^2}$. Especially if $s^2 \geq \max\{24\rho, \frac{16}{n}, \frac{1}{ns^2}\}M^2$, then $OPT = \mathbb{E}_{\hat{P}_n}[Z] - \sqrt{2\rho s_n^2}$ with probability at least $1 - \exp(-\frac{ns^2}{36M^2})$.*

The intuition for this equivalence is that the χ^2 ball and the variance penalty term in Equation (3) are both quadratic [Staib et al., 2019]. Figure 2 illustrates χ^2 balls with various radii on the 3-dimensional simplex.

Failure of standard methods. Various methods have been developed for achieving small regret in maximizing $g(f, x, P_0) = \mathbb{E}_{P_0(w)}[f(x, w)]$ for some distribution $P_0(w|x) = P_0(w)$ [Williams, 2000; Swersky et al., 2013; Toscano-Palmerin and Frazier, 2018]. These methods leverage the relationships in Equation (5) and (6) to infer the posterior mean and variance of the expected function $g(f, x, P_0)$ from those of f. The inferred posterior mean and variance for $g(f, x, P_0)$ are then used in certain ways to acquire new points. While this is useful in the standard setting when we know P_0, it is not useful when we only have the empirical distribution $\hat{P}_n$. Specifically, an optimal solution found by these methods in the problem associated with the empirical distribution may be sub-optimal to that associated with the true distribution P_0.

An illustrative example is depicted in Figure 1 where the averaged trajectories of our proposed DRBQO (detailed in Section 4) and a standard BQO baseline (detailed in Section 6) are also shown. Due to a limited number of samples of P_0, the Monte Carlo estimate $\mathbb{E}_{\hat{P}_n(w)}[f(x, w)]$ results in a spurious expected objective in this case. By resorting to the empirical distribution $\hat{P}_n$ constructed from the limited set of samples, the standard BQO baseline ignores the distributional uncertainty and converges to the optimum of the spurious expected objective. The same limitation applies to the standard BQO optimization methods, e.g., Williams [2000]; Swersky et al. [2013]; Toscano-Palmerin and Frazier [2018]; Pearce and Branke [2017] whose goal is to find a global non-robust maximum.

4 ALGORITHMIC APPROACH

Our main proposed algorithm is presented in Algorithm 1. In the standard Bayesian quadrature problem in Equation (2), we can easily adopt standard Bayesian optimization algorithms such as expected improvement (EI) [Mockus et al., 1978] and an upper confidence bound (UCB) (e.g., GP-UCB [Srinivas et al., 2010]) using quadrature relationships in Equation (5) and (6) [Swersky et al., 2013]. However, like $g_{bv}(x)$ in Equation (3), $\min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)]$ does not follow a GP if f follows a GP. This difficulty hinders the adoption of EI-like and UCB-like algorithms to our setting. We overcome this problem using posterior sampling [Russo and Roy, 2014].

Algorithm 1: `DRBQO`: Distributionally Robust Bayesian quadrature optimization

Input: Prior $\texttt{GP}(\mu_0, k)$, horizon T, fixed sample set S_n, confidence radius $\rho \geq 0, C_0 = k$.

1 **for** $t = 1$ *to* T **do**

  `/* Posterior sampling */`

2 Sample $\tilde{f}_t \sim \texttt{GP}(\mu_{t-1}, C_{t-1})$.

  `/* A surrogate DR optimization */`

3 Choose $x_t \in \arg\max_{x \in \mathcal{X}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\tilde{f}_t(x, w)]$.

  `/* Highest posterior variance */`

4 Choose $w_t = \arg\max_{w \in S_n} C_{t-1}(x_t, w; x_t, w)$.

5 Observe reward $\hat{y}_t \leftarrow f(x_t, w_t) + \epsilon_t$.

6 Perform update `GP` to get μ_t and C_t.

7 **end**

Output: $\arg\max_{x \in \{x_1, \ldots, x_T\}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_P[\mu_T(x, w)]$.

The main idea of our algorithm is to sample and solve a surrogate distributionally robust optimization problem at each step guided by posterior sampling (lines 2 and 3 in Algorithm 1). In practice, we follow Hernández-Lobato et al. [2014] to perform posterior sampling (line 2 in Algorithm 1). Similar to the way posterior sampling is applied to standard Bayesian optimization problem [Hernández-Lobato et al., 2014], a new point is selected according to the probability it is optimal in the sense of distributional robustness. One of the advantages of posterior sampling is that it avoids the need for confidence bound such as UCB. This is useful

for our setting because the non-Gaussian nature of the distributionally robust objective makes it difficult to construct a deterministic confidence upper bound.

Due to the convexity of the expectation with respect to a distribution, we can efficiently compute the value (therefore the gradients) of the inner minimization in line 8 of Algorithm 1 in an analytical form via Lagrangian multipliers, as presented in Proposition 1.

Proposition 1. *Let* $l = (l_1, ..., l_n) \in \mathbb{R}^n$ *(e.g.,* $l = \left(\tilde{f}_t(x, w_1), ..., \tilde{f}_t(x, w_n)\right)$ *in line 3 of Algorithm 1),* $\hat{P}_n = (\frac{1}{n}, ..., \frac{1}{n})$ *being the weights of the empirical distribution,* Δ_n *being the n-dimensional simplex,* $\mathcal{P}_{n,\rho} = \left\{ P \in \Delta_n \middle| \frac{1}{2}\int_\Omega (\frac{dP}{d\hat{P}_n} - 1)^2 d\hat{P}_n \leq \rho \right\}$ *being the* χ^2*-ball around the empirical distribution with radius* ρ*. Then, the optimal weights* $p = (p_1, ..., p_n) = \arg\min_{q \in \mathcal{P}_{n,\rho}} q^T l$ *satisfy the systems of relations with variables* (p, λ, η)*:*

$$\begin{cases} \lambda p_i = \dfrac{1}{n}\max\{-l_i - \eta, 0\}, \forall 1 \leq i \leq n \\ \eta|A| + n\lambda = -\displaystyle\sum_{i \in A} l_i \ \text{ where } A = \{i : l_i \leq -\eta\} \\ \lambda\left(2\rho + 1 - n\|p\|_2^2\right) = 0 \\ n\|p\|_2^2 \leq 2\rho + 1, \ \text{and } \gamma \geq 0. \end{cases} \tag{10}$$

Proof. The constrained minimization $\min_{p \in \mathcal{P}_{n,\rho}} p^T l$ is a convex optimization problem which forms the Lagrangian: $L(p, \lambda, \eta, \zeta) = p^T l - \lambda\left(\rho - \frac{1}{2n}\sum_{i=0}^n (np_i - 1)^2\right) - \eta(1 - \sum_{i=1}^n p_i) - \sum_{i=1}^n \zeta_i p_i$ where $p \in \mathbb{R}^n, \lambda \geq 0, \eta \in \mathbb{R}$, and $\zeta \in \mathbb{R}^n_+$.

The system of linear equations in the proposition emerges from Karush-Kuhn-Tucker (KKT) conditions and simple rearrangements. Note that since the primal problem is convex, the duality gap is zero and the KKT conditions are the sufficient and necessary conditions for the primal problem.

Notice the first two equations that we can compute p_i in terms of λ. These $p_i = p_i(\lambda)$ are then substituted into Equation (10) to solve for λ. In practice, we can use bisection search [Namkoong and Duchi, 2016] to solve for λ satisfying Equation (10). The details of this algorithm and of Proposition 1 are presented in the supplementary material. □

5 THEORETICAL ANALYSIS

For the sake of analysis, we adopt the definition of the T-period regret and Bayesian regret from Russo and Roy [2014] to our setting. In particular, we define a policy π as a mapping from the history $H_t = (x_1, w_1, P_1, ..., x_{t-1}, w_{t-1}, P_{t-1})$ to (x_t, w_t, P_t) where $P_i \in \mathcal{P}_{n,\rho} \times \mathcal{X}, \forall i$.

Definition 1 (T-period regret)**.** The T-period regret of a policy π is defined by

$$Regret(T, \pi, f) = \sum_{t=1}^{T} \mathbb{E}\left[g(f, x^*, P^*) - g(f, x_t, P_t)|f\right],$$

where $P^*(.|x) = \arg\min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)], \forall x \in \mathcal{X}$, $x^* \in \arg\max_{x \in \mathcal{X}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)]$, and for all $T \in \mathbb{N}$.

Definition 2 (T-period Bayesian regret)**.** The T-period Bayesian regret of a policy π is the expectation of the regret with respect to the prior over f,

$$BayesRegret(T, \pi) = \mathbb{E}[Regret(T, \pi, f)]. \tag{11}$$

For simplicity, we focus our analysis on the case where $\mathcal{X}$ is finite and $\mathcal{P}_{n,\rho}$ is a finite subset of the χ^2 ball of radius ρ. Similar to Srinivas et al. [2010], the results can be extended to infinite sets $\mathcal{X}$ and the entire χ^2 ball using discretization trick of Srinivas et al. [2010] as long as a smoothness condition (i.e., the partial derivatives of f are bounded with high probability) is satisfied (see Theorem 2 in Srinivas et al. [2010]).

Theorem 2. *Assume* $\mathcal{X}$ *is a finite subset of* $\mathbb{R}^d$*, and* $\mathcal{P}_{n,\rho}$ *is a finite subset of the* χ^2 *ball of radius* ρ*. Let* π^{DRBQO} *be the DRBQO policy presented in Algorithm 1,* γ_T *be the maximum information gain for* f *defined in Srinivas et al. [2010], then for all* $T \in \mathbb{N}$*,*

$$BayesRegret(T, \pi^{DRBQO}) \leq 1 + \frac{(\sqrt{2\log\frac{(1+T^2)|\mathcal{X}||\mathcal{P}_{n,\rho}|}{\sqrt{2\pi}}} + B)\sqrt{2\pi}}{|\mathcal{X}||\mathcal{P}_{n,\rho}|} + \frac{2\gamma_T\sqrt{(1+2\rho)n}}{1+\sigma^{-2}} + 2\sqrt{T\gamma_T(1+\sigma^{-2})^{-1}\log\frac{(1+T^2)|\mathcal{X}||\mathcal{P}_{n,\rho}|}{\sqrt{2\pi}}}.$$

Note that γ_T can be bounded for three common kernels linear, SE and Matérn kernels [Srinivas et al., 2010]. Using these bounds in Srinivas et al. [2010], Theorem 2 suggests that DRBQO has sublinear Bayesian regret for those common kernels.

Proof. We leverage two proof techniques from Russo and Roy [2014] to derive this bound including posterior sampling regret decomposition and the connection between posterior sampling and UCB. However, an extension from the Bayesian regret bound to our case is non-trivial. The main difficulty is that the ρ-robust quadrature distributions $\arg\min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)]$ are random variables and the resulting quadrature $\min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[f(x, w)]$ does not follow a GP. We overcome this difficulty by decomposing the range $\mathbb{R}$ of $f(x, w)$ into a set of carefully designed disjoint subsets, using several concentration inequalities for Gaussian

distributions, and leveraging the mild assumptions of f from the problem setup. The details are presented in the supplementary material. □

6 EXPERIMENT

In this section, we empirically validate the performance of DRBQO by comparing against several baselines in synthetic and n-fold cross-validation hyperparameter tuning experiments.

We focus on the BQO baselines that directly substitute the inferred posterior mean $\mu_t(x, \hat{P}_n)$ (in Equation (5)) and variance $\sigma_t^2(x, \hat{P}_n)$ (in Equation (6)) of $g(f, x, \hat{P}_n)$ into any standard acquisition (e.g., EI and GP-UCB) to achieve small regret in maximizing $g(f, x, \hat{P}_n)$. More advanced BQO baseline methods, e.g., [Toscano-Palmerin and Frazier, 2018], are expected to perform poorly in the distributional uncertainty setting because they are not set out to account for the robust solutions. There is a distinction between sampled points and report points by each baseline algorithm. A sampled point is a suggested point regarding where to sample next while a report point is chosen from all the sampled points (up to any iteration) based on the objective function that an algorithm aims at optimizing. In standard noiseless Bayesian optimization, sampled points and report points are identical. However, this is not necessarily the case in BQO where the objective function has expectation form and is not directly queried. In particular, we consider the following baselines:

- MTBO: Multi-task Bayesian optimization [Swersky et al., 2013] is a typical BQO algorithm in which the inferred posterior mean and variance are plugged into the standard EI acquisition to select x_t. In addition, each w_t in this case represents a task and MTBO uses multi-task kernels to model the task covariance. Conditioned on x_t, w_t is selected such that the corresponding task yields the highest EI. We include MTBO only in the cross-validation hyperparameter tuning experiments.

- BQO-EI: This algorithm is similar to MTBO except for two distinctions. First, w_t is selected such that it yields the highest posterior variance on f, similar to our algorithm (see line 4 in Algorithm 1). Second, this uses kernels defined on the Cartesian product space $\mathcal{X} \times \Omega$ instead of the multi-task kernels as in MTBO. In addition, the report point at time t is $\arg\max_{x \in x_{1:t}} \mathbb{E}_{\hat{P}_n(w)}[\mu_t(x, w)]$.

- Maximin-BQO-EI: This method is the same as BQO-EI except that the report point is $\arg\max_{x \in x_{1:t}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[\mu_t(x, w)]$.

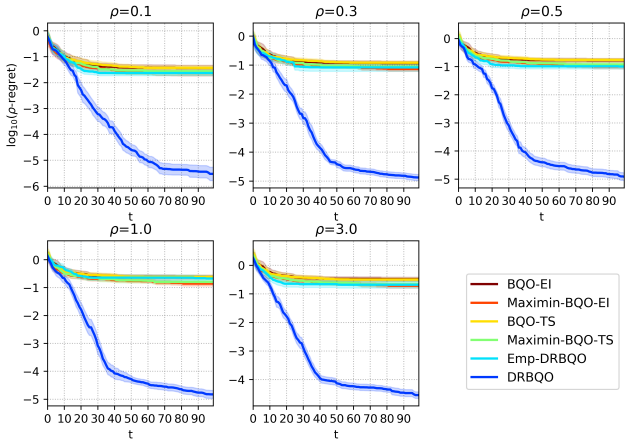

Figure 3: The best so-far ρ-regret values (plotted on the $\log_{10}$ scale) of the baseline BQO methods and our proposed method DRBQO for the synthetic function in Section 6. DRBQO significantly outperforms the baselines with respect to the ρ-regret in this experiment. The larger the value of ρ (i.e., the more conservative against the adversarial distributional perturbation), the higher is the ρ-regret of the non-robust baselines.

- BQO-TS: This method is a non-robust version of our proposed DRBQO. The only distinctions between BQO-TS and DRBQO are in the way x_t is selected (line 3 of Algorithm 1) and the way a report point is chosen. In BQO-TS, x_t is selected with respect to the empirical distribution as follows: $x_t \in \arg\max_{x \in \mathcal{X}} \mathbb{E}_{\hat{P}_n(w)}[\tilde{f}_t(x, w)]$, and the report point at time t is chosen as $\arg\max_{x \in x_{1:t}} \mathbb{E}_{\hat{P}_n(w)}[\mu_t(x, w)]$.

- Maximin-BQO-TS: This is the same as BQO-TS except that the final report point is $\arg\max_{x \in x_{1:t}} \min_{P \in \mathcal{P}_{n,\rho}} \mathbb{E}_{P(w)}[\mu_t(x, w)]$.

- Emp-DRBQO: This is the same as DRBQO except that the report point is chosen as $\arg\max_{x \in x_{1:t}} \mathbb{E}_{\hat{P}_n(w)}[\mu_t(x, w)]$.

Synthetic function. The distributional uncertainty problem is more pronounced when $f(x, w)$ is more significantly distinct across different values of $w \in S_n$, i.e., $f(x, w)$ experiences high variance along the dimension of w. Inspired by the logistic regression and the experimental evaluation from the original variance-based regularization work [Namkoong and Duchi, 2016], we use a logistic form for synthetic function: $f(x, w) = -\log(1 + \exp(x^T w))$, where $x, w \in \mathbb{R}^d$. The true distribution P_0 is the standard Gaussian $\mathcal{N}(0, I)$. In this example, we use $d = 2$ for better visualization. We sample $n = 10$ values of w from $\mathcal{N}(w; 0, I)$ and fix this set for the empirical distribution

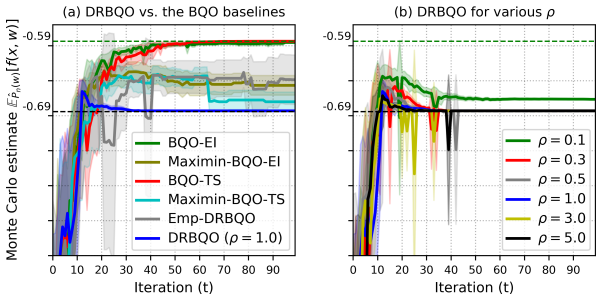

Figure 4: The empirical expected function $\mathbb{E}_{\hat{P}_n(w)}[f(x,w)]$ evaluated at each point x reported at time t by DRBQO and the standard BQO baselines (a), and by DRBQO for various values of ρ (b). In this example, $\mathbb{E}_{\hat{P}_n(w)}[f(x,w)]$ has a maximum value of -0.59 while it has a value of -0.69 evaluated at the optimum of the true expected function $\mathbb{E}_{P_0(w)}[f(x,w)]$. The BQO baselines achieve higher values of the empirical expected function than DRBQO but DRBQO converges to the distributionally robust solutions.

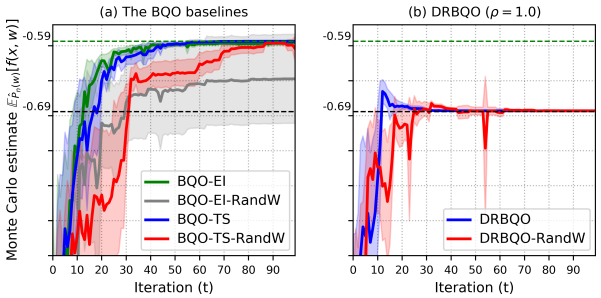

Figure 5: The effect of different methods of selecting w_t on the performance of the BQO baselines (a) and DRBQO (b). We observe that random selection of w_t can either slow down or prevent the convergence of both the standard BQO baselines and DRBQO in this experiment.

$\hat{P}_n(w) = \frac{1}{n}\sum_{i=1}^{n}\delta(w - w_i)$. The true expected function $\mathbb{E}_{P_0(w)}[f(x,w)]$ and the empirical (Monte Carlo) estimate function $\mathbb{E}_{\hat{P}_n(w)}[f(x,w)]$ are illustrated in Figure 1 (a) and (b), respectively. In this illustration, the Monte Carlo estimate function catastrophically shifts the true optimum to a spurious point due to the limited data in estimating P_0.

We initialize the comparative algorithms by selecting 12 uniformly random inputs $(x,w) \in \mathcal{X} \times S_n$, and we keep these initial points the same for all the algorithms. We use the squared exponential kernel k_{SE} defined on the Cartesian product space of x and w. We normalize the input and output values to the unit cube, and resort to marginal maximum likelihood to learn the GP hyperparameters [Rasmussen and Williams, 2006] every time we acquire a new observation. The time horizon for all the algorithms is $T = 100$. We report the results using two evaluation metrics: the ρ-regret as defined in Equation (9) and the value of the empirical expected function $\mathbb{E}_{\hat{P}_n(w)}[f(x,w)]$ evaluated at point x reported by an algorithm at time t. The former metric quantifies how close a certain point is to the distributionally robust solution while the latter measures the performance of each algorithm from a perspective of the empirical distribution. We repeat the experiment 30 times and report the average mean and the 96% confidence interval for each evaluation metric.

The first results are presented in Figure 3. We report over a range of ρ values $\{0.1, 0.3, 0.5, 1.0, 3.0\}$ capturing the degree of conservativeness against the distributional uncertainty. Note that if $\rho > \frac{n-1}{2} = 4.5$, it represents the most conservative case as the χ^2 ball covers the entire n-dimensional simplex. We observe from Figure 3 that DRBQO significantly outperforms the baselines in this experiment. Also notice that when we increase the conservativeness requirement (i.e., increasing the values of ρ), the standard BQO baselines have higher ρ-regret. This is because the standard BQO baselines are rigid and do not allow for any conservativeness in the optimization. Therefore, these algorithms converge to the optimum of the spurious Monte Carlo estimate function.

We highlight the comparative algorithms in the second metric in Figure 4 where we report the value of the empirical expected function $\mathbb{E}_{\hat{P}_n(w)}[f(x,w)]$ at each point x reported by each algorithm at time t. Since the BQO baselines are set out to maximize the Monte Carlo estimate function, they achieve higher values in this metric than DRBQO. However, the non-robust solutions returned by the BQO baselines are sub-optimal with respect to the ρ-regret in this case, as seen from the corresponding results in Figure 3.

In addition, we evaluate the effectiveness of the selection of w at line 4 in Algorithm 1. Currently, w_t is selected such that it yields the highest posterior mean given x_t. This is to improve exploration in f. We compare this selection strategy with the random strategy in which w_t is uniformly selected from S_n regardless of x_t. The result is reported in Figure 5. In this figure, the post-fix RandW denotes the random selection of w_t. We observe that random selection of w_t can hurt the convergence of both the standard BQO baselines and DRBQO. Furthermore, the selection of w_t for the highest posterior variance (line 4 of Algorithm 1) in DRBQO is also meaningful in proving Theorem 2. More empirical evaluations in other synthetic functions are presented in the supplementary material.

Cross-validation hyperparameter tuning. A typical real-world problem that possesses the quadrature structure of Equation (1) is n-fold cross-validation hyperparameter tuning. The n-fold cross-validation performance can be thought of as a Monte Carlo approximate of the true model performance. Given a fixed learning algorithm associated with a set of hyperparameter x, let $f(x, w)$ be an approximate model performance trained on $\mathcal{D}\backslash w$ and evaluated on the validation set w where $\mathcal{D}$ denotes the training data set, w denotes a subset of training points sampled from $\mathcal{D}$, and $\mathcal{D}\backslash w$ denotes everything in $\mathcal{D}$ but not in w. Increasing the number of folds reduces the variance in the model performance estimate, but it is expensive to evaluate the cross-validation performance for a large value of n. Therefore, a class of Bayesian quadrature optimization methods is beneficial in this case in which we actively select both the algorithm's hyperparameters x_t and a fold w_t to evaluate without the need of training the model in all n folds [Swersky et al., 2013].

However, the standard BQO methods assume the empirical distribution for each fold and are set out to maximize the average n-fold values. In practice, the average n-fold value can be a spurious measure for model performance when there is sufficient discrepancy of the model performance across different folds. This scenario fits well into our distributional uncertainty problem in Equation (1) where the fold distribution $P_0(w)$ is unknown in practice. In addition, we use a one-hot n-dimensional vector to represent each of the n folds. This offers two main advantages: (i) it allows us to leverage the standard kernel such as k_{SE} on the product space $\mathcal{X} \times \Omega$; (ii) it is able to model different covariance between different pairs of folds. For example, the covariance between fold 1 and fold 3 is not necessary the same as that between fold 8 and fold 10 though the fold indicator difference are the same.

We evaluate this experiment on two common machine learning models using the MNIST dataset [Yann et al., 1998]: ElasticNet and Convolutional Neural Network (CNN). For ElasticNet, we tune the l_1 and l_2 regularization hyperparamters, and use the SGDClassificer implementation from the scikit-learn package [Pedregosa et al., 2012]. For CNN, we use the standard architecture with 2 convolutional layers. In CNN, we optimize over three following hyperparamters: the learning rate l and the dropout rates in the first and second pooling layers. We used Adam optimizer [Kingma and Ba, 2015] in 20 epochs with the batch size of 128.

In addition to the previous baselines in the synthetic experiment, we also consider the multi-task Bayesian optimization (MTBO) [Swersky et al., 2013] baseline for this application. MTBO is a standard method for cross-validation hyperparameter tuning.

Table 1: Classification error (%) of ElasticNet and CNN on the MNIST test set tuned by different algorithms. Each bold number in the DRBQO group denotes the classification error that is smaller than any corresponding number in the baseline group.

Methods	**ElasticNet**	**CNN**
MTBO	8.576 ± 0.080	1.712 ± 0.263
BQO-EI	9.166 ± 0.433	1.634 ± 0.157
BQO-TS	8.625 ± 0.116	1.820 ± 0.227
DRBQO($\rho = 0.1$)	**8.450** ± 0.022	1.968 ± 0.310
DRBQO($\rho = 0.3$)	**8.505** ± 0.082	**1.495** ± 0.106
DRBQO($\rho = 0.5$)	**8.515** ± 0.075	1.869 ± 0.232
DRBQO($\rho = 1$)	**8.526** ± 0.065	**1.444** ± 0.071
DRBQO($\rho = 3$)	**8.387** ± 0.013	**1.374** ± 0.066
DRBQO($\rho = 5$)	**8.380** ± 0.022	**1.321** ± 0.061

In this experiment, we also use k_{SE} kernel defined on the Cartesian product space $\mathcal{X} \times \Omega$ of x and w for all the methods except for MTBO which uses task kernel on the domain of w. We initialize 6 (respectively 9) initial points and keep these initial points the same for all the algorithms in ElasticNet (respectively CNN). Each of the algorithms are run for $T = 60$ (respectively $T = 90$) iterations in ElasticNet (respectively CNN). We repeat the experiment 20 times and report the average and standard deviation values of an evaluation metric. We split the training data into $n = 10$ folds and keep these folds the same for all algorithms. We compare DRBQO against the baselines via a practical metric: the classification error in the test set evaluated at the final set of hyperparameters reported by each algorithm at the final step T. This metric is a simple but practical measure of the robustness of the hyperparameters over the unknown data distribution P_0. The result is reported in Table 1. We observe that DRBQO outperforms the baselines for most of the considered values of ρ, especially for large values of ρ (i.e., $\rho \in \{1, 3, 5\}$ in this case). More results for the case of Support Vector Machine (SVM) are presented in the supplementary material.

7 DISCUSSION

In this work, we have proposed a posterior sampling based algorithm, namely DRBQO, that efficiently seeks for the robust solutions under the distributional uncertainty in Bayesian quadrature optimization. Compared to the standard BQO algorithms, DRBQO provides a flexibility to control the conservativeness against distributional perturbation. We have demonstrated the empirical effectiveness and characterized the theoretical convergence of DRBQO in sublinear Bayesian regret.

Acknowledgements

This research was partially funded by the Australian Government through the Australian Research Council (ARC). Prof Venkatesh is the recipient of an ARC Australian Laureate Fellowship (FL170100006).

References

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.

Jose H. Blanchet, Yang Kang, Fan Zhang, and Zhangyi Hu. A distributionally robust boosting algorithm. *CoRR*, abs/1905.07845, 2019.

Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Distribution-agnostic model-agnostic meta-learning, 2020.

Rizal Fathony, Ashkan Rezaei, Mohammad Ali Bashiri, Xinhua Zhang, and Brian D. Ziebart. Distributionally robust graphical models. In *NeurIPS*, pages 8354–8365, 2018.

Peter I. Frazier. A tutorial on bayesian optimization. *CoRR*, abs/1807.02811, 2018.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1934–1943. PMLR, 2018.

José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. In *NIPS*, pages 918–926, 2014.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Johannes Kirschner, Ilija Bogunovic, Stefanie Jegelka, and Andreas Krause. Distributionally robust bayesian optimization, 2020.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *CoRR*, abs/1908.08729, 2019.

Jonas Mockus, Vytautas Tiesis, and Antanas Zilinskas. The application of bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129): 2, 1978.

Hongseok Namkoong and John C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *NIPS*, pages 2208–2216, 2016.

Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *NIPS*, pages 2971–2980, 2017.

Chris J. Oates and Timothy John Sullivan. A modern retrospective on probabilistic numerics. *Statistics and Computing*, 29(6):1335–1351, 2019.

Chris. J. Oates, Steven Niederer, Angela Lee, François-Xavier Briol, and Mark Girolami. Probabilistic models for integration error in the assessment of functional cardiac models, 2016.

Anthony O'Hagan. Bayes–hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

Michael Pearce and Jürgen Branke. Bayesian simulation optimization with input uncertainty. In *WSC*, pages 2268–2278. IEEE, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2012.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *CoRR*, abs/1908.05659, 2019.

Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian monte carlo. In *NIPS*, pages 489–496. MIT Press, 2002.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Math. Oper. Res.*, 39(4):1221–1243, 2014.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *CoRR*, abs/1902.08708, 2019.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, pages 1015–1022. Omnipress, 2010.

Matthew Staib and Stefanie Jegelka. Distributionally robust optimization and generalization in kernel methods. In *NeurIPS*, pages 9131–9141, 2019.

Matthew Staib, Bryan Wilder, and Stefanie Jegelka. Distributionally robust submodular maximization. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 506–516. PMLR, 2019.

Qingyun Sun and Stephen Boyd. Distributional robust kelly gambling, 2018.

Kevin Swersky, Jasper Snoek, and Ryan Prescott Adams. Multi-task bayesian optimization. In *NIPS*, pages 2004–2012, 2013.

Saul Toscano-Palmerin and Peter I. Frazier. Bayesian optimization with expensive integrands. *CoRR*, abs/1803.08661, 2018.

Brian Jonathan Williams. *Sequential Design of Computer Experiments to Minimize Integrated Response Functions*. PhD thesis, 2000. AAI9983009.

Huan Xu and Shie Mannor. Distributionally robust markov decision processes. In *NIPS*, pages 2505–2513. Curran Associates, Inc., 2010.

LeCun Yann, Bottou Leon, Bengio Yoshua, and Haffner Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998.

Dixian Zhu, Zhe Li, Xiaoyu Wang, Boqing Gong, and Tianbao Yang. A robust zero-sum game framework for pool-based active learning. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 517–526. PMLR, 2019.