Bayesian Regret Minimization in Offline Bandits

Marek Petrik [1] **Guy Tennenholtz** [2] **Mohammad Ghavamzadeh** [3]

Abstract

We study how to make decisions that minimize Bayesian regret in offline linear bandits. Prior work suggests that one must take actions with maximum *lower confidence bound* (LCB) on their reward. We argue that the reliance on LCB is inherently flawed in this setting and propose a new algorithm that directly minimizes upper bounds on the Bayesian regret using efficient conic optimization solvers. Our bounds build heavily on new connections to monetary risk measures. Proving a matching lower bound, we show that our upper bounds are tight, and by minimizing them we are guaranteed to outperform the LCB approach. Our numerical results on synthetic domains confirm that our approach is superior to LCB.

1. Introduction

The problem of offline bandits is an important special case of offline reinforcement learning (RL) in which the model consists of a single state and involves no state transitions (Hong et al., 2023). Offline RL, a challenging research problem with a rich history, is inspired by the need to make reliable decisions when learning from a logged dataset (Lange et al., 2012; Rashidinejad et al., 2022). Practical problems from recommendations to search to ranking can be modeled as offline bandits; see, for example, Hong et al. (2023) and references therein. Moreover, gaining a deeper theoretical understanding of offline bandits is a vital stepping stone in understanding the complete offline RL problem.

We study the problem of minimizing the *Bayesian regret* in the offline linear bandit setting. Bayesian regret differs substantially from its *frequentist* counterpart. While frequentist regret assumes a fixed true model and studies algorithms' response to random datasets, Bayesian regret assumes a fixed dataset and studies algorithms' regret as a function of the true model. When provided with good priors, Bayesian methods offer sufficiently tight bounds to achieve excellent practical results (Lattimore & Szepesvari, 2018; Gelman et al., 2014; Vaart, 2000). As such, the strengths of Bayesian methods complement the scalability and simplicity of the frequentist algorithms.

Most prior work on Bayesian offline RL and bandits has adopted a form of pessimism that chooses actions with the highest *lower confidence bounds* (LCBs). These LCB-style algorithms compute a policy or action with the largest expected return (or reward), penalized by its uncertainty. The uncertainty penalty is computed from credible regions derived from the posterior distribution (Delage & Mannor, 2010; Hong et al., 2023; Brown et al., 2020; Javed et al., 2021; Lobo et al., 2023), and often gives rise to some form of robust optimization (Behzadian et al., 2021; Petrik & Russel, 2019). LCB-style Bayesian algorithms are generally inspired by the success of this approach in frequentist settings, where LCB is typically computed from concentration inequalities (Xie et al., 2021; Rashidinejad et al., 2022; Jin et al., 2022; Ghosh et al., 2022; Cheng et al., 2022).

In this paper, we propose a Bayesian regret minimization algorithm, called `BRMOB`, that takes a new approach to Bayesian offline bandits. Instead of adopting an LCB-style strategy, we directly minimize new regret *upper bounds*. To derive these bounds, we reformulate the usual high-confidence objective as a Value-at-Risk (VaR) of the epistemic uncertainty. Then, we bound the VaR by combining techniques from robust optimization and Chernoff analysis. Our bounds apply to both Gaussian and sub-Gaussian posteriors over the latent reward parameter. `BRMOB` minimizes the regret bounds efficiently using convex conic solvers. Finally, we also establish a matching lower bound that shows our upper bounds are tight.

Compared with prior work in Bayesian offline bandits, `BRMOB` achieves tighter theoretical guarantees and better empirical performance. Two main innovations enable these improvements. First, `BRMOB` computes randomized policies. Our numerical results show that randomizing among actions results in hedging that can significantly reduce regret compared to deterministic policies. In contrast to `BRMOB`, most existing algorithms in Bayesian offline bandits (Hong et al., 2023) and Bayesian offline RL (Delage & Mannor, 2010;

*Equal contribution [1]University of New Hampshire [2]Google Research [3]Amazon AGI. Correspondence to: Marek Petrik <mpetrik@cs.unh.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Petrik & Russel, 2019; Behzadian et al., 2021; Angelotti et al., 2021) are restricted to deterministic policies. Second, BRMOB is the only algorithm that explicitly minimizes Bayesian regret bounds. As discussed above, existing algorithms usually maximize the LCB on returns (Hong et al., 2023; Uehara & Sun, 2023), which does not guarantee to reduce Bayesian regret. Similarly, recent algorithms that maximize the expected return (Steimle et al., 2021; Su & Petrik, 2023) are also not known to reduce Bayesian regret.

We also study the general suitability of LCB algorithms for minimizing Bayesian regret. While BRMOB significantly outperforms a particular LCB algorithm known as FlatOPO (Hong et al., 2023), the more critical question is whether the general LCB approach is viable for Bayesian regret minimization. Using our new regret lower bounds, we answer this question negatively. More precisely, we show that penalizing reward uncertainty, the core of all LCB algorithms, is guaranteed to increase the algorithm's regret even in very simple problems. This is because actions with high uncertainty may also have a high upside and avoiding them increases regret. Therefore, we believe that explicit regret minimization, as in BRMOB, is a more promising future direction than LCB-style algorithms.

Bayesian regret minimization in offline bandits can also be framed as a chance-constrained optimization problem (Ben-Tal et al., 2009). The recent chance-constrained optimization literature is mostly focused on constraints in which the function is concave or linear in the uncertain parameter (Gupta, 2019; Bertsimas et al., 2021). However, the chance constraint in the Bayesian regret minimization problem is convex, preventing us from using these methods. Another non-concave chance-constrained optimization approach is to resort to scenario-based or sample-based methods (Calafiore & Campi, 2005; Nemirovski & Shapiro, 2006; Luedtke & Ahmed, 2008; Brown et al., 2020). We briefly discuss these methods in Section 3.2. Such sample-based formulations are general, simple to implement and work well in practice. However, they scale poorly to large problems, provide no theoretical insights, and struggle to compute randomized policies. The closest result to our work is the Bernstein technique for bounding linear chance-constrained programs (Nemirovski & Shapiro, 2007; Pintér, 1989), which is a special case of one of our bounds.

The paper is organized as follows. After introducing our notations and the popular risk-measure Value-at-Risk (VaR) in Section 2, we formally define the problem of Bayesian regret minimization in offline bandits and connect it to minimizing VaR in Section 3. In Section 4, we derive two new upper bounds on the Bayesian regret and propose our main algorithm, BRMOB, that is based on a simultaneous minimization of these two regret bounds. We also prove a lower bound on the regret that shows our upper bound is tight. In Section 5, we first derive a regret bound for BRMOB in terms of problem parameters and show that it compares favorably with LCB-based algorithms. We then argue that the general LCB approach is unsuitable for minimizing Bayesian regret. Finally, in Section 6, we compare BRMOB's performance with three baseline algorithms on synthetic domains and show that it is preferable to LCB-style algorithms.

2. Preliminaries

We begin by defining the notations we use throughout the paper. We use lower and upper case bold letters to denote vectors and matrices, such as $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, and normal font for the elements of vectors and matrices, e.g., x_i. We define the weighted ℓ_2-norm for any vector $\boldsymbol{x} \in \mathbb{R}^d$ and positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ as $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{x}}$. We denote by $\Delta_k, \forall k \in \mathbb{N}$ the k-dimensional probability simplex, and by $\boldsymbol{I}$, $\boldsymbol{0}$, $\boldsymbol{1}$, and $\boldsymbol{1}_a$ the identity matrix, the zero vector, the one vector, and the one-hot vector all with appropriate dimensions. Random variables are adorned with a tilde and are not capitalized. For example, $\tilde{\boldsymbol{x}}$ represents a vector-valued random variable. Finally, we denote by Ω the probability space of a random variable.

Suppose that $\tilde{x}\colon \Omega \to \mathbb{R}$ is a random variable that represents *costs*. Then, its *value-at-risk* (VaR) at a risk-level $\alpha \in [0, 1)$ is usually defined as the largest lower bound on its α-quantile (e.g., Follmer & Schied 2016, definition 4.45, and remark A.20):

$$\mathrm{VaR}_\alpha[\tilde{x}] = \inf\{t \in \mathbb{R} \mid \mathbb{P}[\tilde{x} > t] \le 1-\alpha\} \tag{1a}$$
$$= \sup\{t \in \mathbb{R} \mid \mathbb{P}[\tilde{x} \ge t] > 1-\alpha\}. \tag{1b}$$

The definition of VaR in the literature depends on whether $\tilde{x}$ represents costs or rewards (Hau et al., 2023). If $\tilde{x}$ represents *rewards*, maximizing $-\mathrm{VaR}_\alpha[-\tilde{x}]$ is equivalent to minimizing $\mathrm{VaR}_\alpha[\tilde{x}]$. For Gaussian random variables, $\tilde{x} \sim \mathcal{N}(\mu, \sigma^2)$, VaR has the following analytical form (Follmer & Schied, 2016):

$$\mathrm{VaR}_\alpha[\tilde{x}] \;=\; \mu + \sigma \cdot z_\alpha\,, \tag{2}$$

where z_α is the α-*quantile* of $\mathcal{N}(0, 1)$.

3. Bayesian Offline Bandits

In this section, we first formally define the problem of Bayesian regret minimization in offline bandits and connect it to monetary risk measures. We then describe two techniques that have been used in solving this problem.

3.1. Problem Definition

Consider a stochastic linear bandit problem with $k \in \mathbb{N}$ arms (actions) from the set $\mathcal{A} = \{a_1, \ldots, a_k\}$. Each arm $a \in \mathcal{A}$ is associated with a d-dimensional feature vector $\boldsymbol{\phi}_a \in \mathbb{R}^d$

and its reward distribution has a mean $r(a;\boldsymbol{\theta}) = \boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\theta}$ for some unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. We define the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d\times k}$ as $\boldsymbol{\Phi} = (\boldsymbol{\phi}_a)_{a\in\mathcal{A}}$. The goal of the agent is to learn a (possibly *randomized*) policy $\boldsymbol{\pi} \in \Delta_k$ to choose its actions accordingly. We denote by π_a the probability according to which policy $\boldsymbol{\pi}$ selects an action $a \in \mathcal{A}$. The mean reward, or *value*, of a policy $\boldsymbol{\pi}$ is defined as

$$\begin{aligned} r(\boldsymbol{\pi};\boldsymbol{\theta}) &= \mathbb{E}[r(\tilde{a};\boldsymbol{\theta}) \mid \tilde{a} \sim \boldsymbol{\pi}] \\ &= \sum_{a\in\mathcal{A}} \pi_a \cdot r(a;\boldsymbol{\theta}) = \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\theta}\,. \end{aligned} \tag{3}$$

An optimal policy $\boldsymbol{\pi}^\star(\boldsymbol{\theta})$ is one that maximizes (3).

In the *offline* bandit setting, the agent only has access to a *logged dataset* $\tilde{\mathcal{D}} = \{(\tilde{a}_i, \tilde{y}_i)\}_{i=1}^n$, and is not capable of interacting further with the environment. Each pair $(\tilde{a}_i, \tilde{y}_i)$ in $\tilde{\mathcal{D}}$ consists of an action $\tilde{a}_i$ selected according to some arbitrary logging policy and a sampled reward $\tilde{y}_i$ from the reward distribution of action $\tilde{a}_i$. We use D to refer to an instantiation of the random dataset $\tilde{\mathcal{D}}$.

We take the *Bayesian* perspective in this paper and model our uncertainty about the reward parameter $\tilde{\boldsymbol{\theta}}\colon \Omega \to \mathbb{R}^d$ by assuming it is a random variable with a known prior $P_{\tilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})$. Therefore, all quantities that depend on $\tilde{\boldsymbol{\theta}}$ are also random. The logged dataset D is used to derive the posterior density $P_{\tilde{\boldsymbol{\theta}}|D}(\boldsymbol{\theta})$ over the reward parameter. To streamline the notation, we denote by $\tilde{\boldsymbol{\theta}}_D := (\tilde{\boldsymbol{\theta}} \mid \tilde{\mathcal{D}} = D)$ the random variable distributed according to this posterior distribution $P_{\tilde{\boldsymbol{\theta}}|D}$. We discuss the derivation of the posterior in Section 5.

As described above, in the Bayesian offline bandit setting we assume that the logged data $\tilde{\mathcal{D}}$ is fixed to some D and the uncertainty is over the reward parameter $\tilde{\boldsymbol{\theta}}$. This is different than the *frequentist* offline setting in which the reward parameter is fixed, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$, and the randomness is over different datasets generated by the logging policy.

In the Bayesian offline bandit setting, our goal is to compute a policy $\boldsymbol{\pi} \in \Delta_k$ that minimizes the *high-confidence Bayesian regret* $\mathfrak{R}_\delta\colon \Delta_k \to \mathbb{R}_+$ defined as

$$\begin{aligned} &\mathfrak{R}_\delta(\boldsymbol{\pi}) := \min\ \epsilon \qquad \text{subject to} \\ &\mathbb{P}\left[\max_{a\in\mathcal{A}} r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D) \le \epsilon\right] \ge 1-\delta, \end{aligned} \tag{4}$$

where $\delta \in (0, \frac{1}{2})$ is the small error tolerance parameter. We also use $\alpha = 1-\delta$ to denote the confidence in the solution.

Note that (4) compares the value of a fixed policy $\boldsymbol{\pi}$ with the reward of an action (max action) that depends on the posterior random variable $\tilde{\boldsymbol{\theta}}_D$. Thus, one cannot expect to achieve a regret of zero. By taking a close look at the definition of regret in (4) and using the definition of VaR in (1a), we may equivalently write our objective in (4) as

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) = \mathrm{VaR}_{1-\delta}\left[\max_{a\in\mathcal{A}} r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right]. \tag{5}$$

One could optimize other objectives besides the high-confidence regret in (5). Other objectives, such as maximizing the VaR of the reward, are easier to solve and Appendix E discusses them in greater detail.

3.2. Baseline Algorithms

We now provide a brief description of two methods that have been used to solve Bayesian offline bandits (defined in Section 3.1) and closely related problems.

Lower Confidence Bound (LCB) Pessimism to the uncertainty in the problem's parameter is the most common approach in offline decision-making problems, ranging from offline RL (Uehara & Sun, 2023; Rashidinejad et al., 2022; Xie et al., 2022), to robust RL (Petrik & Russel, 2019; Behzadian et al., 2021; Lobo et al., 2020), and offline bandits (Hong et al., 2023). In the case of offline bandits, this approach is compellingly simple and is known as maximizing a lower confidence bound, or LCB. The general recipe of the LCB algorithm for Gaussian and sub-Gaussian posteriors $\tilde{\boldsymbol{\theta}}_D$ is to simply choose the action $\hat{a} \in \mathcal{A}$ such that

$$\hat{a} \in \arg\max_{a\in\mathcal{A}} \ell_\beta(a) := \left(\boldsymbol{\mu}_n^\mathsf{T}\boldsymbol{\phi}_a - \beta\cdot\sqrt{\boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\phi}_a}\right), \tag{6}$$

for some $\beta > 0$. The terms $\boldsymbol{\mu}_n^\mathsf{T}\boldsymbol{\phi}_a$ and $\sqrt{\boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\phi}_a}$ represent the posterior mean and standard deviation of $r(a,\tilde{\boldsymbol{\theta}}_D) = \boldsymbol{\phi}_a^\mathsf{T}\tilde{\boldsymbol{\theta}}_D$. The parameter β is typically chosen to guarantee that $\ell_\beta(a)$ is a high-probability lower bound on the return of action $a \in \mathcal{A}$:

$$\mathbb{P}\left[\ell_\beta(a) \le r(a,\tilde{\boldsymbol{\theta}}_D)\right] \ \ge\ 1-\delta.$$

The `FlatOPO` algorithm (Hong et al., 2023) is a particular instance of the LCB approach to offline bandits that uses $\beta = \sqrt{5d\log(1/\delta)}$ for Gaussian posteriors. When $\beta = 0$, we refer to an algorithm that implements (6) as `Greedy`.

Scenario-based Methods Another natural approach to minimizing the Bayesian regret in offline bandits is to treat the optimization in (4) as a *chance-constrained optimization* problem. The most general algorithm to solve chance-constrained optimization is to use *scenario-based* techniques to minimize the regret $\mathfrak{R}_\delta(\boldsymbol{\pi})$ (Calafiore & Campi, 2005; Nemirovski & Shapiro, 2006; Luedtke & Ahmed, 2008). A typical scenario-based algorithm first approximates $\tilde{\boldsymbol{\theta}}_D$ with a *discrete* random variable $\tilde{\boldsymbol{q}}$ constructed by sampling from its posterior $P_{\tilde{\boldsymbol{\theta}}|D}(\boldsymbol{\theta})$, and then computes a

deterministic policy by solving

$$\arg\min_{\hat{a}\in\mathcal{A}} \ \mathrm{VaR}_{1-\delta}\left[\max_{a\in\mathcal{A}} r(a;\tilde{\boldsymbol{q}}) - r(\hat{a};\tilde{\boldsymbol{q}})\right]. \qquad (7)$$

The optimization in (7) can be solved by enumerating all the actions and computing the VaR of the discrete random variable within the brackets. The important question that has been extensively studied here is the number of samples needed to obtain a solution with high confidence (Nemirovski & Shapiro, 2007; 2006). The time complexity of this algorithm is a function of the number of samples and the desired confidence to guarantee a certain suboptimality of the solution; we refer the interested reader to Nemirovski & Shapiro (2007) for a detailed analysis discussion.

Despite the generality and simplicity of scenario-based methods, they have several important drawbacks. They require sampling from the posterior, with a sample complexity that scales poorly with the dimension d, number of actions k, and particularly confidence level $1-\delta$. They do not provide theoretical guarantees for the regret of the obtained policy and offer no insights into how the regret scales with the parameters of the problem.

Finally, minimizing the regret in (5) over the space of *randomized* policies is challenging using scenario-based methods because it requires solving a mixed-integer linear program (Lobo et al., 2020). Other ideas have been explored (Calafiore & Campi, 2005; Brown et al., 2020) but a detailed study of such algorithms is beyond our scope.

4. Minimizing Analytical Regret Bounds

In this section, we propose our new approach for minimizing the Bayesian regret, $\mathfrak{R}_\delta(\boldsymbol{\pi})$, defined in (5). In particular, we derive two upper bounds on $\mathfrak{R}_\delta(\boldsymbol{\pi})$ that complement each other depending on the relative sizes of the feature vector d and action space k. We also prove a lower bound on $\mathfrak{R}_\delta(\boldsymbol{\pi})$ that shows our upper bound is tight. Finally, we propose our `BRMOB` algorithm that aims at jointly minimizing our two upper bounds. The proofs of this section are in Appendix B.

4.1. Bayesian Regret Bounds

To avoid unnecessary complexity, we assume in this section that the posterior distribution over the reward parameter is *Gaussian*. We show analogous results for the general *sub-Gaussian* case in Appendix D.

Assumption 4.1. The posterior over the latent reward parameter is distributed as $\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and a *positive definite* covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$.

We begin by showing that under Assumption 4.1, the regret $r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)$ of any policy $\boldsymbol{\pi} \in \Delta_k$ with respect to $a \in \mathcal{A}$ has a Gaussian distribution.

Lemma 4.2. *Suppose that $\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, for any policy $\boldsymbol{\pi} \in \Delta_k$, the Bayesian regret in (5) can be written as*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \ = \ \mathrm{VaR}_{1-\delta}\left[\max_{a\in\mathcal{A}} \tilde{x}_a^{\boldsymbol{\pi}}\right], \qquad (8)$$

where $\tilde{x}_a^{\boldsymbol{\pi}} \sim \mathcal{N}(\mu_a^{\boldsymbol{\pi}}, \sigma_a^{\boldsymbol{\pi}})$ with

$$\mu_a^{\boldsymbol{\pi}} = \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}), \qquad \sigma_a^{\boldsymbol{\pi}} = \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}}. \qquad (9)$$

Lemma 4.2 points to the main challenge in deriving tight bounds on $\mathfrak{R}_\delta(\boldsymbol{\pi})$. Even when $\tilde{\boldsymbol{\theta}}_D$ is normally distributed, the random variable $\max_{a\in\mathcal{A}} \tilde{x}_a^{\boldsymbol{\pi}}$ is unlikely to be Gaussian. The lack of normality prevents us from deriving an *exact* analytical expression for $\mathfrak{R}_\delta(\boldsymbol{\pi})$ using (2). In the remainder of the section, we derive two separate techniques for upper bounding the VaR of the maximum of random variables in (8), thereby also bounding the Bayesian regret $\mathfrak{R}_\delta(\boldsymbol{\pi})$.

Our first bound expresses the overall regret as a maximum over individual action regrets. We refer to it as an *action-set bound*, because it grows with the size of the action space k, and state it in Theorem 4.3.

Theorem 4.3. *The regret for any policy $\boldsymbol{\pi} \in \Delta_k$ satisfies*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \le \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mu_a^{\boldsymbol{\pi}} + \sigma_a^{\boldsymbol{\pi}} \cdot z_{1-\delta\xi_a} \qquad (10a)$$

$$\le \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mu_a^{\boldsymbol{\pi}} + \sigma_a^{\boldsymbol{\pi}} \cdot \sqrt{2\log(1/\delta\xi_a)}, \qquad (10b)$$

where $z_{1-\delta\xi_a}$ is the $(1-\delta\xi_a)$-th standard normal quantile.

A special case of (10) is when $\boldsymbol{\xi} = 1/k \cdot \mathbf{1}$ is uniform, in which case is simplifies to

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \ \le \ \max_{a\in\mathcal{A}} \mu_a^{\boldsymbol{\pi}} + \sigma_a^{\boldsymbol{\pi}} \cdot \sqrt{2\log(k/\delta)}. \qquad (11)$$

This shows that the action-set bound in Theorem 4.3 grows sub-logarithmically with the number of actions k.

Because the bound in Theorem 4.3 is based on a union bound, the question of its tightness is particularly salient. To address this, we prove a lower bound on the regret when the arms are independent (e.g., multi-armed bandits).

Theorem 4.4. *Suppose that $\boldsymbol{\pi} \in \Delta_k$ is a* deterministic *policy such that $\pi_{a_1} = 1$ for $a_1 \in \mathcal{A}$ without loss of generality. When $\mu_2 = \mu_3 = \cdots = \mu_k$, $\boldsymbol{\Sigma}$ is diagonal with $\Sigma_{2,2} = \Sigma_{3,3} = \ldots \Sigma_{k,k}$, and $\boldsymbol{\Phi} = \boldsymbol{I}$, then*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \ \ge \ \mu_{a_2}^{\boldsymbol{\pi}} + \sigma_{a_2}^{\boldsymbol{\pi}} \cdot \kappa_{\mathrm{l}}(k-1),$$

where

$$\kappa_{\mathrm{l}}(k) = -1 + \sqrt{1 - \log(\sqrt{2\pi}) - 2\log\left(1-(1-\delta)^{1/k}\right)}.$$

The lower bound in Theorem 4.4 indicates that Theorem 4.3 is tight. For an ease of reference we use $\kappa_{\mathrm{u}}(k) =$

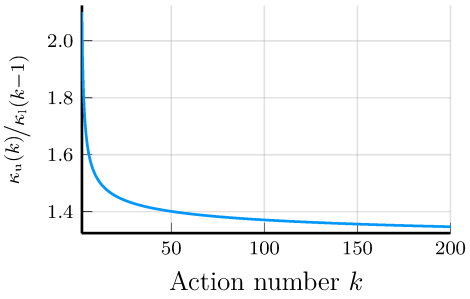

Figure 1. The quotient of the upper bound coefficient $\kappa_u(k)$ and the lower bound coefficient $\kappa_l(k)$.

$\sqrt{2\log(k/\delta)}$ to refer to the coefficient on the RHS of (11). The main difference between the upper and lower bounds are the coefficients $\kappa_u(k)$ and $\kappa_l(k-1)$. One can readily show that $\kappa_u(k) \in O(\kappa_l(k-1))$, since $\kappa_u(1)/\kappa_l(1) < 10$ when $\delta \leq 1/2$ and $\kappa_u(k)/\kappa_l(k-1)$ is a non-increasing function of k. Figure 1 depicts the quotient of the upper and lower bound coefficients as a function of k for $\delta = 0.1$.

We now state our second upper bound on the regret in Theorem 4.5. We refer to it as the *parameter-space bound* because it grows with the dimension d of the parameter.

Theorem 4.5. *The regret for any policy $\boldsymbol{\pi} \in \Delta_k$ satisfies*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \leq \max_{a \in \mathcal{A}} \mu_a^{\boldsymbol{\pi}} + \sigma_a^{\boldsymbol{\pi}} \cdot \sqrt{\chi^2_d(1-\delta)} \tag{12a}$$

$$\leq \max_{a \in \mathcal{A}} \mu_a^{\boldsymbol{\pi}} + \sigma_a^{\boldsymbol{\pi}} \cdot 5d\log(1/\delta)\,, \tag{12b}$$

where $\chi^2_d(1-\delta)$ is the $(1-\delta)$-th quantile of the χ^2 distribution with d degrees of freedom.

Note that a growing body of literature argues that using credible regions in constructing robust approximations of VaR is overly conservative when used with linear or concave functions (Gupta, 2019; Bertsimas et al., 2021; Petrik & Russel, 2019). However, these results do not apply to our setting because the maximum in (8) is *non-concave*.

To compare our two upper bounds in (10) and (12), it is sufficient to compare the terms $z_{1-\delta\xi_a}$ and $\sqrt{\chi^2_d(1-\delta)}$. From Theorems 4.3 and 4.5, we can conclude that the second upper bound is preferable when $d < \log k$.

4.2. Optimization Algorithm

We now describe our main algorithm, *Bayesian Regret Minimization for Offline Bandits* (BRMOB), whose pseudo-code is reported in Algorithm 1. Before describing BRMOB in greater detail, it is important to note that it returns a *randomized* policy. Unlike in online bandits, here the goal of randomization among the actions is not to explore, but rather to reduce the risk of incurring high regret. The numerical results in Section 6 show that the ability to randomize over actions significantly reduces Bayesian regret in many situations.

BRMOB's strategy is to compute a policy with the *minimum* regret guarantee. In Line 2, it computes a policy $\boldsymbol{\pi}^0$ that simultaneously minimizes our two proposed upper bounds: the one in Theorem 4.3 with a uniform $\boldsymbol{\xi}$ as given in (11), and the one in Theorem 4.5 as given in (12). The bounds can be optimized jointly because they differ only in constant ν. The optimization in (13) is a second-order conic program (SOCP), because $\boldsymbol{\nu} \geq \mathbf{0}$ and can be solved very efficiently (ApS, 2022; Lubin et al., 2023). The actual time complexity depends on the particular SOCP solver used, but most interior-point algorithms run in $O(k^6)$ complexity or faster (Kitahara & Tsuchiya, 2018).

After completing Line 2, BRMOB proceeds with m iterations of tightening the regret bound and improving the policy. In each iteration i, it tightens the regret bound in Theorem 4.3 by optimizing $\boldsymbol{\xi}^i$ in (14) for the incumbent policy $\boldsymbol{\pi}^{i-1}$. The minimum in (14) can be computed efficiently using exponential and second-order cones (ApS, 2022; Lubin et al., 2023). Exponential conic optimization is hypothesized to be polynomial time, but this fact has not been established yet to the best of our knowledge. The algorithm then minimizes the tightened bound by solving (13) and obtains an improved policy $\boldsymbol{\pi}^i$.

The tightening steps in Algorithm 1 can be seen as a coordinate descent procedure for joint minimization of $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$ in (10). It would be preferable to minimize the bound simultaneously over $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$, but such optimization appears to be intractable.

Finally, Algorithm 1 returns a policy in the set $\{\boldsymbol{\pi}^i\}_{i=0}^m$ with the smallest regret bound ρ^i in Line 6. Although ρ^i will be generally non-increasing with an increasing i, this is not guaranteed. This is because the tightening step in (14) minimizes the bounds in (10b) and (12b). These bounds are generally looser than the bounds in (10a) and (12a) optimized by (13).

We provide a worst-case error bound on the regret of BRMOB in Section 5. Our regret bound holds for any number of tightening steps, including $m = 0$. We focus on bounds that are independent of m for the sake of simplicity, since the improvements that arise from the tightening procedure can be difficult to quantify cleanly.

We conclude this section with the following result that shows BRMOB indeed minimizes the regret upper bounds in Theorems 4.3 and 4.5 as intended.

Proposition 4.6. *Suppose that* BRMOB *returns a policy $\hat{\boldsymbol{\pi}} \in \Delta_k$ and a regret bound $\hat{\rho}$. Then*

$$\mathfrak{R}_\delta(\hat{\boldsymbol{\pi}}) \;\leq\; \hat{\rho} \;\leq\; \min_{\boldsymbol{\pi} \in \Delta_k} \max_{a \in \mathcal{A}} \mu_a^{\boldsymbol{\pi}}(n) + \sigma_a^{\boldsymbol{\pi}}(n) \cdot \eta\,, \tag{15}$$

Algorithm 1: BRMOB: Bayesian Regret Minimization for Offline Bandits

Input: Posterior parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, risk tolerance $\delta \in (0, 1/2)$, feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d \times k}$, # of iterations m

Initialize $\nu_a^0 \leftarrow \min\left\{\sqrt{\chi^2_d(1-\delta)}, z_{1-\delta/k}\right\}, \ \forall a \in \mathcal{A}$; $\quad i \leftarrow 0$;

Minimize regret bounds: Let ρ^i and $\boldsymbol{\pi}^i$ be the optimizers of

$$\underset{\boldsymbol{\pi},\, \boldsymbol{s}\in\mathbb{R}^k_+,\, \rho\in\mathbb{R}}{\text{minimize}} \ \rho \quad \text{subject to} \quad \mathbf{1}^\mathsf{T}\boldsymbol{\pi} = 1, \quad \rho \geq \mu_a^{\boldsymbol{\pi}} + s_a \cdot \nu_a^i, \quad s_a^2 \geq (\sigma_a^{\boldsymbol{\pi}})^2, \ \forall a \in \mathcal{A}. \tag{13}$$

for $i = 1, \ldots, m$ **do**

Tighten regret bounds: Let $\boldsymbol{\xi}^i$ be an optimizer of

$$\underset{\boldsymbol{\xi},\, \boldsymbol{s}\in\mathbb{R}^k_+,\, \boldsymbol{l}\in\mathbb{R}^k,\, \rho\in\mathbb{R}}{\text{minimize}} \ \rho \quad \text{subject to} \quad \mathbf{1}^\mathsf{T}\boldsymbol{\xi} = \delta, \quad \rho \geq \mu_a^{\boldsymbol{\pi}^{i-1}} + \sigma_a^{\boldsymbol{\pi}^{i-1}} \cdot s_a, \quad s_a^2 \geq -2l_a, \quad l_a \leq \log \xi_a, \ \forall a \in \mathcal{A}. \tag{14}$$

Set $\nu_a^i \leftarrow z_{1-\delta\xi_a^i}, \ \forall a \in \mathcal{A}$;

Solve (13) and let ρ^i and $\boldsymbol{\pi}^i$ be its optimizers ;

$i^\star \leftarrow \arg\min_{i=0,\ldots,m} \rho^i$; `// Choose policy with the best regret guarantee`

return randomized policy $\boldsymbol{\pi}^{i^\star}$, regret upper bound $\rho^{i^\star}$;

where $\eta = \min\left\{\sqrt{2\log(k/\delta)}, \sqrt{5d\log(1/\delta)}\right\}$.

5. Regret Analysis

In this section, we derive a regret bound for BRMOB and compare it with that of FlatOPO (Hong et al., 2023), an LCB-style algorithm. We use a frequentist analysis to bound the Bayesian regret of BRMOB as a function of k, d, number of samples n, and coverage of the dataset D. Section 5.3 concludes by arguing that the general LCB approach in (6) is unsuitable for minimizing Bayesian regret; see Appendix E for other objectives that can be optimized using LCB-style algorithms. Our lower bound shows that LCB can match BRMOB's regret only if the confidence penalty β is very small and *decreases* with k and d. All the proofs of this section are reported in Appendix C.

5.1. Sample-Based Regret Bound

As in prior work (Hong et al., 2023), we assume a Gaussian prior distribution over the reward parameter $P_{\tilde{\boldsymbol{\theta}}} = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with an invertible $\boldsymbol{\Sigma}_0$, and Gaussian rewards $\tilde{y} \sim \mathcal{N}\big(r(a;\tilde{\boldsymbol{\theta}}) = \boldsymbol{\phi}_a^\mathsf{T}\tilde{\boldsymbol{\theta}}, \bar{\sigma}^2\big)$ for each action $a \in \mathcal{A}$. As a result, the posterior distribution over the parameter given a dataset $D = \{(a_i, y_i)\}_{i=1}^n$ is also Gaussian $\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ with

$$\begin{aligned} \boldsymbol{\Sigma}_n &= (\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\boldsymbol{G}_n)^{-1}, \\ \boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \bar{\sigma}^{-2}\boldsymbol{B}_n\boldsymbol{y}_n), \end{aligned} \tag{16}$$

and where $\boldsymbol{B}_n = (\phi_{a_i})_{i=1}^n$ is the matrix with observed features in its columns, $\boldsymbol{y}_n = (y_i)_{i=1}^n$ is the vector of observed rewards, and $\boldsymbol{G}_n = \boldsymbol{B}_n^\mathsf{T}\boldsymbol{B}_n$ is the empirical covariance matrix (see Bayesian linear regression for example in Rasmussen & Williams 2006; Deisenroth et al. 2021).

To express the regret bound as a function of the dataset D, we make the following standard quality assumption.

Assumption 5.1. The feature vectors satisfy $\|\boldsymbol{\phi}_a\|_2 \leq 1, \forall a \in \mathcal{A}$, and there exists a $\gamma > 0$ such that

$$\boldsymbol{G}_n \ \succeq \ \gamma n \cdot \boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T}, \qquad \forall a \in \mathcal{A}, \ \forall n \geq 1\,.$$

Intuitively, Assumption 5.1 states that the dataset provides sufficient information such that the norm of the covariance matrix $\boldsymbol{\Sigma}_n$ of the posterior distribution over $\tilde{\boldsymbol{\theta}}_D$ decreases linearly with n. From a frequentist perspective, this assumption holds with high probability by the Bernstein-Von-Mises theorem under mild conditions (Vaart, 2000).

We are now ready to bound the Bayesian regret of BRMOB. We state the bound for the general case and then tighten it when $\boldsymbol{\mu}_n = \mathbf{0}$ (only the variance of actions matters).

Theorem 5.2. *Suppose that the parameter has a Gaussian posterior* $\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ *and* BRMOB *returns a policy* $\hat{\boldsymbol{\pi}}$. *Then, the regret of* BRMOB *is bounded as* $\mathfrak{R}_\delta(\hat{\boldsymbol{\pi}}) \leq 2\eta$, *where*

$$\eta = \sqrt{\frac{\min\left\{2\log(k/\delta), 5d\log(1/\delta)\right\}}{\lambda_{\max}(\boldsymbol{\Sigma}_0)^{-1} + \gamma n\bar{\sigma}^{-2}}}\,. \tag{17}$$

Moreover, if $\boldsymbol{\mu}_n = 0$ *then* $\mathfrak{R}_\delta(\hat{\boldsymbol{\pi}}) \leq 2\left(1 - \max_{a'\in\mathcal{A}} \hat{\pi}_{a'}\right)\eta$ *with* $\max_{a'\in\mathcal{A}} \hat{\pi}_{a'} \geq 1/d+1$.

5.2. Comparison with FlatOPO

We now compare the regret bound of BRMOB with that of FlatOPO (Hong et al., 2023), an LCB-based algorithm for re-

gret minimization in Bayesian offline bandits. As discussed in Section 3.2, using the LCB principle is the most common approach to regret minimization in offline decision-making. Hong et al. (2023) derived the following regret bound for FlatOPO under Assumption 5.1:

$$\mathfrak{R}_\delta(\hat{\boldsymbol{\pi}}) \;\le\; 2\sqrt{\frac{5d^2\log(1/\delta)}{\lambda_{\max}(\boldsymbol{\Sigma}_0)^{-1}+\gamma n\bar{\sigma}^{-2}}}\,, \qquad (18)$$

where $\hat{\boldsymbol{\pi}}$ is a *deterministic* policy returned by the algorithm.

Comparing our regret bound for BRMOB in (17) with FlatOPO's in (18), we notice two main improvements. The first one is that the BRMOB's regret is bounded by $\sqrt{\log k}$. Thus, when the number of actions k satisfies $k \ll \exp(d)$, the regret guarantee of BRMOB can be dramatically lower than that achieved by FlatOPO. It is unclear how one could extend the existing analysis in Hong et al. (2023) to bound its regret in terms of k. Its design and analysis rely on a robust set, which is difficult to restrict using k.

The second improvement is that the regret bound of BRMOB grows $\sqrt{d}$ slower than FlatOPO's, which is a significant reduction in regret. This improvement is probably a consequence of our tighter analysis rather than better algorithmic design. The analysis in Hong et al. (2023) uses a general upper bound on the trace of a rank one matrix, which introduces an unnecessary $\sqrt{d}$ term. Yet, applying our techniques to FlatOPO yields additional constant terms missing in (18).

5.3. Limitation of LCB

In this section, we argue that the popular LCB approach is inherently unsuitable for minimizing Bayesian *regret* in offline bandits. As we discussed in Section 5.2, BRMOB achieves significantly better regret guarantees than FlatOPO. Our numerical results in Section 6 also show that BRMOB outperforms FLatOPO. However, these results are obtained for a particular value of β in (6). Our theoretical analysis suggests that even a simple Greedy algorithm, which uses $\beta = 0$ in (6), can significantly outperform LCB. The intuition behind the LCB approach is that one should prefer actions with low uncertainty, and thus, limited downside. This intuition is correct when the goal is to maximize the VaR of *reward* as shown in Appendix E.1. However, this intuition does not apply when the objective is *regret* minimization. In fact, actions with low uncertainty also have limited upside and high regret, and thus, as we show, penalizing high variance actions is counterproductive.

We now construct a simple class of offline bandit problems to illustrate LCB's limitations. For this class of problems, we show that the *lower bound* on the regret of LCB can be far greater than the upper bound on BRMOB, or even Greedy, policy. In what follows, we assume that LCB computes the high-confidence lower bound as in (6) for some value of β.

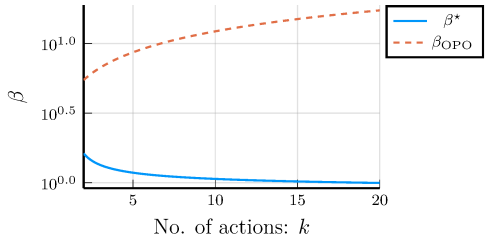

Figure 2. The value of β used by FlatOPO in Example 1, β_{OPO}, and the upper bound $\beta^\star$ that may avoid the under-performance of LCB, defined in (22), as functions of the number of actions k.

Example 1. Consider a class of offline bandit problems parametrized with the β used in (6). The bandit has $k \ge 2$ arms, feature dimension $d = k$, and a feature matrix $\boldsymbol{\Phi} = \boldsymbol{I}$. Suppose that the posterior covariance over the reward parameter $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ is diagonal with the diagonal elements $\sigma_1 = 0$ and $\sigma_2 = \cdots = \sigma_k$, and the posterior mean has the following form: $\mu_1 = 0$ and $\mu_2 = \cdots = \mu_k = \beta \cdot \sigma_2 \ge 0$.

The intuition underlying the bandit problems in Example 1 is as follows. It has one action, a_1, with low reward and low variance. The other $k-1$ arms are i.i.d. with higher mean and variance. LCB prefers to take action a_1 because of its low variance and forgoing the higher mean of the other actions. The next theorem shows that taking any of the other actions with a higher mean, as would be chosen by BRMOB, or even Greedy that selects an action with the largest posterior mean, leads to a far lower regret.

Theorem 5.3. *Consider the bandit problems in Example 1 and assume a realization of LCB with a coefficient $\beta > 0$ that breaks ties by choosing an a_i with the smallest i. Then, LCB returns $\boldsymbol{\pi}_{\mathrm{LCB}} \in \Delta_k$ with $\boldsymbol{\pi}_{\mathrm{LCB}}(a_1) = 1$ and*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{LCB}}) \;\ge\; (\beta + \kappa_{\mathrm{l}}(k)) \cdot \sigma_{a_2}. \qquad (19)$$

Moreover, Greedy *with the same tie-breaks will return a policy $\boldsymbol{\pi}_{\mathrm{G}} \in \Delta_k$ with $\boldsymbol{\pi}_{\mathrm{G}}(a_2) = 1$ and*

$$\mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{G}}) \;\le\; \sqrt{2} \cdot \sigma_{a_2} \cdot \kappa_{\mathrm{u}}(k). \qquad (20)$$

Finally, BRMOB*'s regret also satisfies the bound in (20).*

Theorem 5.3 shows that even in a simple class of problems, Greedy (or BRMOB) computes a policy that outperforms LCB significantly. The increase in regret of LCB versus Greedy (or BRMOB) can be bounded from below as

$$\mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{LCB}}) - \mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{G}}) \ge \left(\beta + \kappa_{\mathrm{l}}(k) - \sqrt{2}\kappa_{\mathrm{u}}(k)\right) \sigma_{a_2}. \qquad (21)$$

Note that the bound, when positive, can be made arbitrarily large by scaling σ_{a_2}.

Using algebraic manipulation of the bound in (21), we can show that β should satisfy the following condition for LCB

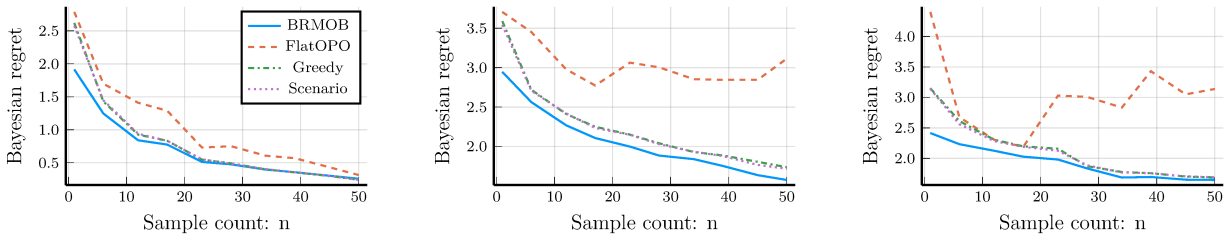

Figure 3. Bayesian regret with $k = d = 5$ *(left)*, $k = d = 50$ *(middle and right)*. The prior mean is $\boldsymbol{\mu}_0 = \mathbf{0}$ *(left and middle)* and $(\boldsymbol{\mu}_0)_a = \sqrt{a}$ for $a = 1, \dots 50$ *(right)*.

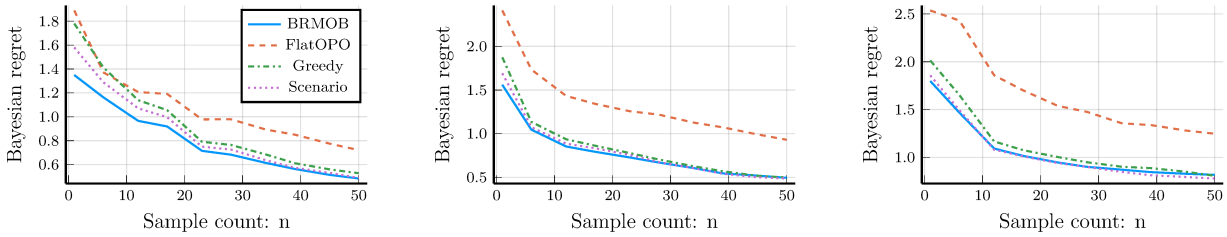

Figure 4. Bayesian regret with $d = 4$ and $k = 10$ *(left)*, $k = 50$ *(middle)*, and $k = 100$ *(right)*.

to perform better than `Greedy` and `BRMOB`:

$$\beta \leq \beta^\star = \frac{\sqrt{2}\kappa_{\mathrm{u}}(k)}{\kappa_{\mathrm{l}}(k-1)} - 1. \tag{22}$$

The inequality in (22) indicates that penalizing an action's uncertainty with a β greater than $\beta^\star$ *increases* the regret of LCB. For comparison, the β used by `FlatOPO` for the class of problems in Example 1 in which $d = k$ is $\beta_{\mathrm{OPO}} = \sqrt{5k \log(1/\delta)}$. Figure 2 shows that for $\delta = 0.1$, β_{OPO} exceeds $\beta^\star$, and thus, violates the condition in (22) and performs worse than `Greedy` and `BRMOB` for all values of k.

6. Numerical Results

In this section, we compare `BRMOB` to several baselines on synthetic domains. Here, we evaluate the basic version of `BRMOB` with $m = 0$ iterations and defer results that demonstrate the improvement from the tightening step to Appendix F. Particularly, we compare it to (i) `FlatOPO` (Hong et al., 2023) that is based on the LCB principle, (ii) `Greedy` method which selects an action a with the largest value of μ_a, and finally, (iii) `Scenario`, the scenario-based method described in (7) in Section 3.2. We execute `Scenario` with 4000 samples from the posterior. Increasing this number did not improve our results.

Our experiments use synthetic domains, each defined by a normal prior $(\boldsymbol{\mu}_0, \boldsymbol{I})$ and a feature matrix $\boldsymbol{\Phi}$. To evaluate the Bayesian regret as a function of data size n, we first sample a single large dataset and then vary the number of data points n from this set used to estimate the posterior distributions. The regret for each policy is computed by a scenario-based algorithm that samples from the posterior and computes the empirical VaR. We use the error tolerance of $\delta = 0.1$ throughout. Results are averaged over 100 runs of this process to reduce variance. As confidence intervals were negligible for all algorithms except `FlatOPO`, to avoid clutter, we do not plot them here and refer the reader to Appendix F for additional details.

We evaluate the algorithms on three domains. The first one uses $k = d$ actions, an identity feature matrix $\boldsymbol{\Phi} = \boldsymbol{I}$, and zero prior mean $\boldsymbol{\mu}_0 = \mathbf{0}$. The second one is the same, except $(\boldsymbol{\mu}_0)_a = \sqrt{a}$ to simulate varying rewards for actions. Finally, the third one fixes dimension $d = 4$ and varies k while using randomly generated features from the ℓ_∞-ball.

Figures 3 and 4 summarize our numerical results. They consistently show across all domains that `BRMOB` significantly outperforms all the other algorithms, particularly when the posterior uncertainty is large. The only challenging setting for `BRMOB` is when $k \gg d$. Note that `FlatOPO`'s performance is noisy in Figure 3 because its β coefficient grows fast with d. A common practice is to tune β, but we did not find any value of β for which `FlatOPO` performs better than `Greedy`, which is consistent with our theoretical analysis in Section 5.3. It is also notable that `Greedy` outperforms LCB significantly, furnishing further evidence that LCB is unsuitable for minimizing regret. We provide additional results, including confidence bounds and runtime, in Appendix F.

7. Conclusion

We proposed BRMOB, a new approach for Bayesian regret minimization in offline bandits, that is based on jointly minimizing two analytical upper bounds on the Bayesian regret. We proved a regret bound for BRMOB and showed that it compares favorably with an existing LCB-style algorithm FlatOPO (Hong et al., 2023). Finally, we showed theoretically and empirically that the popular LCB approach is unsuitable for minimizing Bayesian regret.

Our approach can be extended to several more general settings. The algorithm and bounds can generalize to sub-Gaussian posterior distributions as described in Appendix D. The algorithm can also be extended to *contextual* linear bandits by computing a separate policy π for each context individually or by assuming a random context. Another important future direction is understanding the implications of our results to frequentist settings where similar concerns about the value of the LCB approach have been raised (Xie et al., 2022; Xiao et al., 2021).

Acknowledgments

We thank the anonymous referees for their helpful comments and suggestions. The work in the paper was supported, in part, by NSF grants 2144601 and 2218063. In addition, this work was performed in part while Mohammad Ghavamzadeh and Marek Petrik were at Google Research.

Impact Statement

This paper aims to advance the theory field of Machine Learning. There are many potential societal consequences of our work, but it is difficult to speculate what they may be. Of course, we sincerely hope that the impacts will only be positive.

References

Ahmadi-Javid, A. Entropic Value-at-Risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Angelotti, G., Drougard, N., and Chanel, C. P. C. Exploitation vs caution: Risk-sensitive policies for offline learning. *arXiv:2105.13431 [cs, eess]*, 2021.

ApS, M. *The MOSEK optimization toolbox for MATLAB manual. Version 10.0.*, 2022.

Behzadian, B., Russel, R., Ho, C. P., and Petrik, M. Optimizing percentile criterion using robust MDPs. In *International Conference on Artificial Intelligence and Statistics (AIStats)*, 2021.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust Optimization*. Princeton University Press, 2009.

Bertsimas, D., den Hertog, D., and Pauphilet, J. Probabilistic Guarantees in Robust Optimization. *SIAM Journal on Optimization*, 31(4):2893–2920, 2021.

Brown, D. S., Niekum, S., and Petrik, M. Bayesian robust optimization for imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Calafiore, G. and Campi, M. C. Uncertain convex programs: Randomized solutions and confidence levels. *Mathematical Programming, Series A*, 102:25–46, 2005.

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning, 2022.

David, H. A. and Nagaraja, H. N. *Order Statistics*. John Wiley & Sons, Ltd, 3 edition, 2003.

Deisenroth, M. P., Faisal, A. A., and Ong, C. S. *Mathematics for Machine Learning*. Cambridge University Press, 2021.

Delage, E. and Mannor, S. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

Follmer, H. and Schied, A. *Stochastic Finance: Introduction in Discrete Time*. De Gruyter Graduate, fourth edition, 2016.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd edition, 2014.

Ghosh, D., Ajay, A., Agrawal, P., and Levine, S. Offline RL policies should be trained to be adaptive, 2022.

Gupta, V. Near-optimal bayesian ambiguity sets for distributionally robust optimization. *Management Science*, 65 (9):4242–4260, 2019.

Hau, J. L., Petrik, M., and Ghavamzadeh, M. Entropic risk optimization in discounted MDPs. In *Artificial Intelligence and Statistics (AISTATS)*, 2023.

Hong, J., Kveton, B., Katariya, S., Zaheer, M., and Ghavamzadeh, M. Multi-task off-policy learning from bandit feedback. In *International Conference on Machine Learning*, 2023.

Horn, R. A. and Johnson, C. A. *Matrix Analysis*. Cambridge University Press, second edition, 2013.

Javed, Z., Brown, D., Sharma, S., Zhu, J., Balakrishna, A., Petrik, M., Dragan, A., and Goldberg, K. Policy gradient Bayesian robust optimization for imitation learning. In *International Conference on Machine Learning (ICML)*, 2021.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline RL?, 2022.

Kitahara, T. and Tsuchiya, T. An extension of Chubanov's polynomial-time linear programming algorithm to second-order cone programming. *Optimization Methods and Software*, 33(1):1–25, January 2018. ISSN 1055-6788. doi: 10.1080/10556788.2017.1382495.

Lange, S., Gabel, T., and Riedmiller, M. Batch reinforcement learning. In *Reinforcement Learning*, pp. 45–73. Springer, 2012.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge University Press, 2018.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

Lobo, E., Cousins, C., Petrik, M., and Zick, Y. Percentile criterion optimization in offline reinforcement learning. In *Neural Information Processing Systems (NeurIPS)*, 2023.

Lobo, E. A., Ghavamzadeh, M., and Petrik, M. Soft-Robust Algorithms for Handling Model Misspecification, 2020.

Lubin, M., Dowson, O., Garcia, J. D., Huchette, J., Legat, B., and Vielma, J. P. Jump 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*, 2023.

Luedtke, J. and Ahmed, S. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

Nemirovski, A. and Shapiro, A. Scenario Approximations of Chance Constraints. In Calafiore, G. and Dabbene, F. (eds.), *Probabilistic and Randomized Methods for Design under Uncertainty*, pp. 3–47. Springer, 2006.

Nemirovski, A. and Shapiro, A. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

Petrik, M. and Russel, R. H. Beyond confidence regions: Tight Bayesian ambiguity sets for robust MDPs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Pintér, J. Deterministic approximations of probability inequalities. *Zeitschrift für Operations-Research*, 33(4): 219–239, 1989.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196, 2022.

Rasmussen, C. E. and Williams, C. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Rockafellar, R. T. and Wets, R. J. *Variational Analysis*. Springer, 2009.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2014.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. Multi-model Markov decision processes. *IISE Transactions*, 53, 2021.

Su, X. and Petrik, M. Solving multi-model MDPs by coordinate ascent and dynamic programming,. In *Uncertainty in Artificial Intelligence (UAI)*, 2023.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. In *International Conference on Learning Representations (ICLR)*, 2023.

Vaart, A. W. V. D. *Asymptotic Statistics*. Cambridge University Press, 2000.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Xiao, C., Wu, Y., Littlemore, T., Dai, B., Mei, J., Li, L., Szepesvari, C., and Schuurmans, D. On the optimality of batch policy optimization algorithms. In *International Conference of Machine Learning (ICML)*, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021.

Xie, T., Bhardwaj, M., Jiang, N., and Cheng, C.-A. ARMOR: A model-based framework for improving arbitrary baseline policies with offline data, 2022.

A. Technical Background and Lemmas

A scalar random variable $\tilde{x}\colon \Omega \to \mathbb{R}$ with mean $\mu = \mathbb{E}[\tilde{x}]$ is sub-Gaussian with a variance factor $\sigma^2 \geq 0$ when

$$\mathbb{E}\big[\exp\big(\lambda(\tilde{x}-\mu)\big)\big] \;\leq\; \exp\big(\lambda^2\sigma^2/2\big), \qquad \forall \lambda \in \mathbb{R}\,. \tag{23}$$

A *multivariate* random variable $\tilde{\boldsymbol{x}}\colon \Omega \to \mathbb{R}^d$ with a mean $\boldsymbol{\mu} = \mathbb{E}[\tilde{\boldsymbol{x}}]$ is *sub-Gaussian* with a covariance factor $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ when (Vershynin, 2010; Jin et al., 2019)

$$\mathbb{E}\big[\exp\big(\lambda \boldsymbol{w}^\mathsf{T}(\tilde{\boldsymbol{x}}-\boldsymbol{\mu})\big)\big] \;\leq\; \exp\big(\lambda^2 \boldsymbol{w}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{w}/2\big), \qquad \forall \lambda \in \mathbb{R},\ \forall \boldsymbol{w} \in \Delta_d\,. \tag{24}$$

The Entropic Value at Risk (EVaR) is a risk measure related to VaR, defined as (Ahmadi-Javid, 2012)

$$\mathrm{EVaR}_\alpha\left[\tilde{x}\right] \;=\; \inf_{\beta>0}\, \beta^{-1}\big(\mathbb{E}[\exp(\beta\tilde{x})] - \log(1-\alpha)\big)\,, \qquad \forall \alpha \in [0,1). \tag{25}$$

The following lemma shows that EVaR is an upper bound on VaR. This is a property that will be useful in our proofs later on.

Lemma A.1. *For any random variable $\tilde{x}\colon \Omega \to \mathbb{R}$, we have that*

$$\mathrm{VaR}_\alpha\left[\tilde{x}\right] \;\leq\; \mathrm{EVaR}_\alpha\left[\tilde{x}\right], \qquad \forall \alpha \in [0,1).$$

Proof. This is a consequence of Proposition 3.2 in Ahmadi-Javid (2012) and the fact that CVaR upper bounds VaR. □

Similar to (2) for VaR, we can show that for Gaussian random variables, $\tilde{x} \sim \mathcal{N}(\mu, \sigma^2)$, EVaR has the following analytical form (Ahmadi-Javid, 2012):

$$\mathrm{EVaR}_\alpha\left[\tilde{x}\right] \;=\; \mu + \sigma\cdot\sqrt{-2\log(1-\alpha)}. \tag{26}$$

One advantage of EVaR over VaR is that we can bound it in the more general case of sub-Gaussian random variables by the same bound as for a normal random variable in (26) (see the following lemma).

Lemma A.2. *Let $\tilde{x}\colon \Omega \to \mathbb{R}$ be a sub-Gaussian random variable defined according to (23). Then, we have*

$$\mathrm{EVaR}_\alpha\left[\tilde{x}\right] \;\leq\; \mu + \sigma\cdot\sqrt{-2\log(1-\alpha)}\,, \qquad \forall \alpha \in [0,1)\,.$$

Proof. From the translation invariance of EVaR (Ahmadi-Javid, 2012, Theorem 3.1) and the definitions in (23) and (25), we have

$$\begin{aligned}\mathrm{EVaR}_\alpha\left[\tilde{x}\right] = \mu + \mathrm{EVaR}_\alpha\left[\tilde{x}-\mu\right] &= \mu + \inf_{\beta>0}\, \beta^{-1}\cdot\Big(\mathbb{E}\big[\exp\big(\beta\cdot(\tilde{x}-\mu)\big)\big] - \log(1-\alpha)\Big)\\ &\leq \mu + \inf_{\beta>0}\, \beta^{-1}\cdot\left(\frac{\beta^2\sigma^2}{2} - \log(1-\alpha)\right) = \mu + \sigma\cdot\sqrt{-2\log(1-\alpha)}\,.\end{aligned}$$

The last step follows by solving for the optimal $\beta^\star = \sigma^{-1}\sqrt{-2\log(1-\alpha)}$ from the first-order optimality conditions of the convex objective function. □

B. Proofs of Section 4

Proof of Lemma 4.2. We obtain by algebraic manipulation that

$$\begin{aligned}\max_{a\in\mathcal{A}}\; r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D) &= \max_{a\in\mathcal{A}}\; \mathbf{1}_a^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D - \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D = \max_{a\in\mathcal{A}}\; \mathbf{1}_a^\mathsf{T}\Big(\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D - \mathbf{1}\boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D\Big)\\ &= \max_{a\in\mathcal{A}}\; \mathbf{1}_a^\mathsf{T}\left(\boldsymbol{I} - \mathbf{1}\boldsymbol{\pi}^\mathsf{T}\right)\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D\,.\end{aligned}$$

Let $\tilde{\boldsymbol{x}}^{\boldsymbol{\pi}} = \left(\boldsymbol{I} - \mathbf{1}\boldsymbol{\pi}^\mathsf{T}\right)\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D$, which is a linear transformation of the normal random variable $\tilde{\boldsymbol{\theta}}_D$. The result follows because linear transformations preserve normality (Deisenroth et al., 2021). □

B.1. Proof of Theorem 4.3

We first report a result in the following lemma which we later use to prove Theorem 4.3.

Lemma B.1. *Suppose $\tilde{\boldsymbol{x}}\colon \Omega \to \mathbb{R}^k$ is a random variable such that all α-quantiles, $\forall \alpha \in [0,1)$, for each $\tilde{x}_a, a \in \mathcal{A}$ are unique. Then, the following inequality holds for each $\alpha \in [0,1)$:*

$$\mathrm{VaR}_\alpha\left[\max_{a\in\mathcal{A}} \tilde{x}_a\right] \;\leq\; \inf\left\{\max_{a\in\mathcal{A}} \mathrm{VaR}_{1-\xi_a}\left[\tilde{x}_a\right] \mid \boldsymbol{\xi} \in \mathbb{R}^k_+,\, \mathbf{1}^\mathsf{T}\boldsymbol{\xi} = 1-\alpha\right\}.$$

We interpret the maximum of all $-\infty$ as $-\infty$.

Proof. The result develops as

$$\begin{aligned}
\mathrm{VaR}_\alpha\left[\max_{a\in\mathcal{A}} \tilde{x}_a\right] &\overset{\text{(a)}}{=} \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\max_{a\in\mathcal{A}} \tilde{x}_a \geq t\right] > 1-\alpha\right\} \\
&\overset{\text{(b)}}{\leq} \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\max_{a\in\mathcal{A}} \tilde{x}_a \geq t\right] \geq 1-\alpha\right\} \\
&\overset{\text{(c)}}{\leq} \sup\left\{t \in \mathbb{R} \mid \sum_{a\in\mathcal{A}} \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq 1-\alpha\right\} \\
&\overset{\text{(d)}}{=} \inf\left\{\sup\left\{t \in \mathbb{R} \mid \sum_{a\in\mathcal{A}} \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \sum_{a\in\mathcal{A}} \xi_a\right\} \mid \boldsymbol{\xi} \in \mathbb{R}^k_+, \mathbf{1}^\mathsf{T}\boldsymbol{\xi} = 1-\alpha\right\} \\
&\overset{\text{(e)}}{\leq} \inf\left\{\max_{a\in\mathcal{A}} \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \xi_a\right\} \mid \boldsymbol{\xi} \in \mathbb{R}^k_+, \mathbf{1}^\mathsf{T}\boldsymbol{\xi} = 1-\alpha\right\} \\
&\overset{\text{(f)}}{\leq} \inf\left\{\max_{a\in\mathcal{A}} \mathrm{VaR}_{1-\xi_a}\left[\tilde{x}_a\right] \mid \boldsymbol{\xi} \in \mathbb{R}^k_+, \mathbf{1}^\mathsf{T}\boldsymbol{\xi} = 1-\alpha\right\}.
\end{aligned}$$

(a) is from the definition of VaR. *(b)* follows by relaxing the set by replacing the strict inequality with a non-strict one. *(c)* follows by relaxing the constraint further using the union bound. *(d)* follows from algebraic manipulation because the objective is constant in the choice of $\boldsymbol{\xi}$. *(e)* holds by relaxing the sum constraints and then representing the supremum over a union of sets by a maximum of the suprema of the sets as

$$\begin{aligned}
\sup\left\{t \in \mathbb{R} \mid \sum_{a\in\mathcal{A}} \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \sum_{a\in\mathcal{A}} \xi_a\right\} &\leq \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \xi_a,\, \exists a \in \mathcal{A}\right\} \\
&= \max_{a\in\mathcal{A}} \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \xi_a\right\}.
\end{aligned}$$

Finally, *(f)* follows from the definition of VaR and because then the quantiles are unique (Follmer & Schied, 2016)

$$\mathrm{VaR}_{1-\xi_a}\left[\tilde{x}_a\right] = \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_a \geq t\right] \geq \xi_a\right\} = \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_a \geq t\right] > \xi_a\right\}.$$

The first equality is the definition of the upper quantile q^+ and the second equality is the definition of the lower quantile q^-, which are equal by the uniqueness assumption. □

We are now ready to prove Theorem 4.3.

Proof of Theorem 4.3. The first inequality in (10) follows from Lemma 4.2 and Lemma B.1 by some algebraic manipulation. The second inequality in (10) follows from upper bounding the VaR of a Gaussian random variable using (2) and the fact that $\tilde{x}^{\boldsymbol{\pi}}_a$ is a Gaussian random variable with mean $\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})$ and standard deviation $\|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}}$.

The inequality $z_{1-\delta\xi_a} \leq \sqrt{2\log 1/\delta\xi_a}$ holds because for a standard normal random variable $\tilde{y}$, we have that

$$z_{1-\delta\xi_a} = \mathrm{VaR}_{1-\delta\xi_a}\left[\tilde{y}\right] \overset{\text{(a)}}{\leq} \mathrm{EVaR}_{1-\delta\xi_a}\left[\tilde{y}\right] \overset{\text{(b)}}{=} \sqrt{2\log(1/\delta\xi_a)}.$$

(a) follows from Lemma A.1 and *(b)* is by (26). □

B.2. Proof of Theorem 4.4

First, we prove a lower bound on the VaR of a single Gaussian random variable.

Lemma B.2. *Suppose that $\tilde{x} \sim \mathcal{N}(0,1)$ and $\alpha \geq \frac{1}{2}$. Then*

$$\mathrm{VaR}_\alpha\left[\tilde{x}\right] \;\geq\; -1 + \sqrt{1 - \log(\sqrt{2\pi}) - 2\log(1-\alpha)}.$$

Proof. To establish this lower bound on VaR, we use the known bounds on the cumulative distribution function of a Gaussian random variable as stated, for example, in eq. (13.1) in Lattimore & Szepesvari (2018). For any $t \in \mathbb{R}$ we have that

$$\mathbb{P}\left[\tilde{x} \geq t\right] \;\geq\; \frac{\sqrt{8\pi^{-1}}}{2|t| + \sqrt{4t^2+16}} \exp\left(-\frac{t^2}{2}\right).$$

From the definition of VaR in (1b) we get that

$$\begin{aligned}
\mathrm{VaR}_\alpha\left[\tilde{x}\right] &= \sup\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x} \geq t\right] > 1-\alpha\right\} \\
&= \sup\left\{t \in \mathbb{R}_+ \mid \mathbb{P}\left[\tilde{x} \geq t\right] > 1-\alpha\right\} \\
&\geq \sup\left\{t \in \mathbb{R}_+ \mid \frac{\sqrt{8\pi^{-1}}}{2t + \sqrt{4t^2+16}} \exp\left(-\frac{t^2}{2}\right) > 1-\alpha\right\} \\
&\geq \sup\left\{t \in \mathbb{R}_+ \mid \frac{\sqrt{8\pi^{-1}}}{4(t+1)} \exp\left(-\frac{t^2}{2}\right) > 1-\alpha\right\}.
\end{aligned}$$

Here, we restricted t to be non-negative, which does not impact the VaR value because for $\alpha \geq 0.5$ we have that $\mathrm{VaR}_\alpha\left[\tilde{x}\right] \geq 0$. The first inequality is a lower bound that follows by tightening the feasible set in the supremum. The final inequality follows since $\sqrt{4t^2+16} \leq 2t+4$ from the triangle inequality.

Then, algebraic manipulation of the right-hand side above gives us that

$$\mathrm{VaR}_\alpha\left[\tilde{x}\right] \geq \sup\left\{t \in \mathbb{R}_+ \mid -t^2 - 2t > 1\log(1-\alpha) + 2\log\sqrt{2\pi}\right\}.$$

Then, using the fact that the constraint is concave in t, we get the final lower bound on VaR by solving the quadratic equation. □

The following lemma bounds the VaR of a maximum of independent random variables. This is possible because the maximum is the first *order statistic* which has an easy-to-represent CDF (David & Nagaraja, 2003).

Lemma B.3. *Suppose that $\tilde{x}_i \colon \Omega \to \mathbb{R}, i = 1, \ldots, n$ are i.i.d. random variables. Then*

$$\mathrm{VaR}_\alpha\left[\max_{i=1,\ldots,n} \tilde{x}_i\right] \;=\; \mathrm{VaR}_{\alpha^{1/n}}\left[\tilde{x}_1\right].$$

Proof. Recall i.i.d. random variables satisfy that

$$\mathbb{P}\left[\max_{i=1,\ldots,n} \tilde{x}_i\right] = \prod_{i=1,\ldots,n} \mathbb{P}\left[\tilde{x}_i\right] = \mathbb{P}\left[\tilde{x}_1\right]^n.$$

The result then follows from the definition of VaR in (1a) and from algebraic manipulation as

$$\begin{aligned}
\mathrm{VaR}_\alpha\left[\max_{i=1,\ldots,n} \tilde{x}_i\right] &= \inf\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\max_{i=1,\ldots,n} \tilde{x}_i > t\right] \leq 1-\alpha\right\} = \inf\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\max_{i=1,\ldots,n} \tilde{x}_i \leq t\right] \geq \alpha\right\} \\
&= \inf\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_1 \leq t\right]^n \geq \alpha\right\} = \inf\left\{t \in \mathbb{R} \mid \mathbb{P}\left[\tilde{x}_1 \leq t\right] \geq \alpha^{1/n}\right\} = \mathrm{VaR}_{\alpha^{1/n}}\left[\tilde{x}_1\right].
\end{aligned}$$

□

Proof of Theorem 4.4. Define a restricted set of actions $\mathcal{A}_2 = \mathcal{A} \setminus \{a_1\}$. As in the remainder of the paper, we use $\alpha = 1 - \delta$ to simplify the notation in this proof.

From the definition of regret in (5) and the monotonicity of VaR (Shapiro et al., 2014) we get that the regret of the $\boldsymbol{\pi}$ can be lower bounded as the maximum regret compared only to actions in $\mathcal{A}_2$:

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) = \mathrm{VaR}_\alpha\left[\max_{a\in\mathcal{A}} r(\tilde{\boldsymbol{\theta}}, a) - r(\tilde{\boldsymbol{\theta}}, a_1)\right] \geq \mathrm{VaR}_\alpha\left[\max_{a\in\mathcal{A}_2} r(\tilde{\boldsymbol{\theta}}, a) - r(\tilde{\boldsymbol{\theta}}, a_1)\right].$$

From the theorem's assumptions, the random variables $\tilde{z}_a = r(\tilde{\boldsymbol{\theta}}, a) - r(\tilde{\boldsymbol{\theta}}, a_1)$ for $a \in \mathcal{A}_2$ are independent and identically distributed as $\mathcal{N}(\mu_2 - \mu_1, \sigma_2^2 + \sigma_1^2)$ where $\sigma_i = \Sigma_{i,i}$ for $i = 1, \ldots, k$. Then, using the inequality above and Lemma B.3 we get that

$$\begin{aligned}\mathfrak{R}_\delta(\boldsymbol{\pi}) &= \mathrm{VaR}_\alpha\left[\max_{a\in\mathcal{A}_2} r(\tilde{\boldsymbol{\theta}}, a) - r(\tilde{\boldsymbol{\theta}}, a_1)\right] \geq \mathrm{VaR}_{1-\alpha^{1/k}}\left[\tilde{z}\right] \\ &= (\mu_2 - \mu_1) + \sqrt{\sigma_1^2 + \sigma_2^2}\cdot \mathrm{VaR}_{1-\alpha^{1/k}}\left[\frac{\tilde{z}_{a_2} - (\mu_2 - \mu_1)}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right].\end{aligned}$$

Here, we used the fact that VaR is positively homogenous and translation equivariant. The result follows by Lemma B.2 since the random variable inside of the VaR above is distributed as $\mathcal{N}(0, 1)$. □

B.3. Proof of Theorem 4.5

This result follows from standard robust optimization techniques (see, for example, Gupta (2019); Petrik & Russel (2019)) as well as bandit analysis. In fact, similar or perhaps almost identical analysis has been used to analyze the regret of `FlatOPO` in Hong et al. (2023). We provide an independent proof for the sake of completeness.

The following two auxiliary lemmas are used to show that a robust optimization over a credible region can be used to upper bound the VaR of any random variable. The first auxiliary lemma establishes a sufficient condition for a robust optimization being an overestimate of VaR.

Lemma B.4. *Suppose that we are given an ambiguity set $\mathcal{P} \subseteq \mathcal{X}$, a function $g\colon \mathcal{X} \to \mathbb{R}$, and a random variable $\tilde{\boldsymbol{x}} : \Omega \to \mathcal{X}$. If $\mathcal{P} \cap \mathcal{Z} \neq \emptyset$ for $\mathcal{Z} = \left\{\boldsymbol{x} \in \mathcal{X} \mid g(\boldsymbol{x}) \geq \mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right]\right\}$, then*

$$\mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right] \;\leq\; \sup_{\boldsymbol{x}\in\mathcal{P}} g(\boldsymbol{x}).$$

Proof. By the hopothesis, there exists some $\hat{\boldsymbol{x}} \in \mathcal{P} \cap \mathcal{Z}$. Then, we have $\sup_{\boldsymbol{x}\in\mathcal{P}} g(\boldsymbol{x}) \;\geq\; g(\hat{\boldsymbol{x}}) \;\geq\; \mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right]$ that concludes the proof, where the first inequality is by definition and the second one is from the definition of the set $\mathcal{Z}$. □

The second auxiliary lemma shows that a credible region is sufficient to upper bound VaR using a robust optimization problem.

Lemma B.5. *Suppose that we are given an ambiguity set $\mathcal{P} \subseteq \mathcal{X}$, a function $g\colon \mathcal{X} \to \mathbb{R}$, and a random variable $\tilde{\boldsymbol{x}} : \Omega \to \mathcal{X}$. Then, we have*

$$\mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P}] \geq \alpha \qquad \Longrightarrow \qquad \mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right] \;\leq\; \sup_{\boldsymbol{x}\in\mathcal{P}} g(\boldsymbol{x}).$$

Proof. Our proof is by contradiction using Lemma B.4. We start by assuming that $\mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P}]$. Define $\mathcal{Z} = \left\{\boldsymbol{x} \in \mathcal{X} \mid g(\boldsymbol{x}) \geq \mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right]\right\}$ as in Lemma B.4. From Lemma B.4, we know that if $\sup_{\boldsymbol{x}\in\mathcal{P}} g(\boldsymbol{x}) \;\geq\; \mathrm{VaR}_\alpha\left[g(\tilde{\boldsymbol{x}})\right]$ is false, then we should have $\mathcal{P} \cap \mathcal{Z} = \emptyset$. By the definition of VaR, we have that $\mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{Z}] > 1 - \alpha$. Then, we get a contradiction with $\mathcal{P} \cap \mathcal{Z} = \emptyset$ as follows

$$1 \geq \mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P} \cup \mathcal{Z}] = \mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P}] + \mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{Z}] > \alpha + 1 - \alpha > 1.$$

□

The following lemma uses a standard technique for constructing a credible region for a multivariate normal distribution (Hong et al., 2023; Gupta, 2019).

Lemma B.6. *Suppose that $\tilde{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multi-variate normal random variable with a mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. Then the set $\mathcal{P} \subseteq \mathbb{R}^d$, defined as*

$$\mathcal{P} = \left\{\boldsymbol{x} \in \mathbb{R}^d \mid \|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \leq \chi^2_d(\alpha)\right\},$$

with $\chi^2_d(\alpha)$ being the α-quantile of the χ^2_d distribution, satisfies that $\mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P}] = \alpha$.

Proof. One can readily verify that $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\tilde{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ is a standard multivariate normal distribution. The norm of this value is a sum of i.i.d. standard normal variables, and thus, is distributed according to the χ^2_d distribution with d degrees of freedom:

$$\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})\right)^{\mathsf{T}} \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}(\tilde{\boldsymbol{x}} - \boldsymbol{\mu})\right) = \|\tilde{\boldsymbol{x}} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \sim \chi^2_d.$$

Therefore, by algebraic manipulation and the definition of a quantile, we obtain that

$$\mathbb{P}[\tilde{\boldsymbol{x}} \in \mathcal{P}] = \mathbb{P}\left[\|\tilde{\boldsymbol{x}} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \leq \chi^2_d(\alpha)\right] = \alpha.$$

□

Finally, the following lemma derives the optimal solution of a quadratic optimization problem that arises in the formulation.

Lemma B.7. *The equality*

$$\max_{\boldsymbol{p}\in\mathbb{R}^d} \left\{\boldsymbol{x}^{\mathsf{T}}\boldsymbol{p} \mid \|\boldsymbol{p} - \hat{\boldsymbol{p}}\|^2_{\boldsymbol{C}} \leq b,\, \boldsymbol{p} \in \mathbb{R}^k\right\} \;=\; \boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{p}} + \sqrt{b} \cdot \|\boldsymbol{x}\|_{\boldsymbol{C}^{-1}} \tag{27}$$

holds for any given vectors $\boldsymbol{x}, \hat{\boldsymbol{p}} \in \mathbb{R}^d$ and a matrix $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ that is positive definite: $\boldsymbol{C} \succ \boldsymbol{0}$.

Proof. From the convexity of the optimization problem in (27), we can construct the optimizer $\boldsymbol{p}^\star$ using KKT conditions as

$$\boldsymbol{p}^\star = \hat{\boldsymbol{p}} + \sqrt{b} \cdot \|\boldsymbol{x}\|_{\boldsymbol{C}^{-1}} \cdot \boldsymbol{C}^{-1}\boldsymbol{x}.$$

The result then follows by substituting $\boldsymbol{p}^\star$ into the maximization problem in the lemma. □

We are now ready to prove the main theorem.

Proof of Theorem 4.5. We derive the bound in (12) using the robust representation of VaR (Ben-Tal et al., 2009). We first construct the set $\mathcal{P}_\delta \subseteq \mathbb{R}^d$ as

$$\mathcal{P}_\delta \;=\; \left\{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\theta} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}} \leq \chi^2_d(1-\delta)\right\}. \tag{28}$$

Using Lemma B.6 and the definition of $\mathcal{P}_\delta$ in (28), we can see that $\mathcal{P}_\delta$ is indeed a credible region:

$$\mathbb{P}\left[\tilde{\boldsymbol{\theta}} \in \mathcal{P}_\delta\right] = 1 - \delta.$$

Then, Lemma B.5 gives us the first inequality in (12):

$$\mathfrak{R}_\delta(\boldsymbol{\pi}) \leq \max_{\boldsymbol{\theta}\in\mathcal{P}_\delta} \max_{a\in\mathcal{A}} \left(r(a;\boldsymbol{\theta}) - r(\boldsymbol{\pi};\boldsymbol{\theta})\right).$$

The second inequality in (12) is a consequence of Lemma B.7 with $\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{1}_a - \boldsymbol{\pi})$, $\hat{\boldsymbol{p}} = \boldsymbol{\mu}$, $\boldsymbol{p} = \boldsymbol{\theta}$, $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$, and $b = \sqrt{\chi^2(1-\delta)}$.

Finally, the inequality $\sqrt{\chi^2_d(1-\delta)} \leq \sqrt{5d\log(1/\delta)}$ follows from Lemma 1 in Laurent & Massart (2000) as in the proof of Lemma 3 in Hong et al. (2023). □

B.4. Proof of Proposition 4.6

Proof. The corollary is an immediate consequence of Theorems 4.3 and 4.5 and the construction of Algorithm 1. By construction, $\boldsymbol{\pi}^0$ is the solution to

$$\boldsymbol{\pi}^0 \in \arg\min_{\boldsymbol{\pi}\in\Delta_k} \max_{a\in\mathcal{A}} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}} \cdot \nu_a^0\,,$$

where ν_a^0 is defined in Algorithm 1. Therefore, using Theorems 4.3 and 4.5 to upper bound ν_a, we obtain

$$\mathfrak{R}_\delta(\boldsymbol{\pi}^0) \le \min_{\boldsymbol{\pi}\in\Delta_k} \max_{a\in\mathcal{A}} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}} \cdot \min\left\{\sqrt{2\log(k/\delta)}, \sqrt{5d\log(1/\delta)}\right\}.$$

This proves the corollary when $i^\star = 0$ in Algorithm 1. Then, using Theorem 4.3 with general $\boldsymbol{\xi}$, we observe that the algorithm selects $i^\star > 0$ only when $\mathfrak{R}_\delta(\boldsymbol{\pi}^{i^\star}) \le \rho^{i^\star} \le \rho^0$, which means that the corollary also holds. □

C. Proofs of Section 5

C.1. Proof of Theorem 5.2

Proof. To prove the first claim of the theorem, let $\bar{\boldsymbol{\pi}}$ be a policy that minimizes the linear component of the regret:

$$\bar{\boldsymbol{\pi}} \in \arg\min_{\boldsymbol{\pi}\in\Delta_k} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\,.$$

Note that the minimum above is upper-bounded by 0. Next we use Proposition 4.6 to bound the regret:

$$\begin{aligned}
\mathfrak{R}_\delta(\hat{\boldsymbol{\pi}}) &\le \min_{\boldsymbol{\pi}\in\Delta_k} \max_{a\in\mathcal{A}} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}_n} \cdot \min\left\{\sqrt{2\log(k/\delta)}, \sqrt{5d\log(1/\delta)}\right\} \\
&\le \max_{a\in\mathcal{A}} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}})\|_{\boldsymbol{\Sigma}_n} \cdot \min\left\{\sqrt{2\log(k/\delta)}, \sqrt{5d\log(1/\delta)}\right\} \\
&\le \max_{a\in\mathcal{A}} \|\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}})\|_{\boldsymbol{\Sigma}_n} \cdot \min\left\{\sqrt{2\log(k/\delta)}, \sqrt{5d\log(1/\delta)}\right\}.
\end{aligned}$$

Now, we bound the term $\|\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}})\|_{\boldsymbol{\Sigma}_n}$. Recall that $\|\bar{\boldsymbol{\pi}}\|_2 \le \|\bar{\boldsymbol{\pi}}\|_1 \le 1$, since $\bar{\boldsymbol{\pi}} \in \Delta_k$. Then, for each $a \in \mathcal{A}$, we have by algebraic manipulation that

$$\begin{aligned}
\|\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}})\|_{\boldsymbol{\Sigma}_n}^2 &= (\mathbf{1}_a - \bar{\boldsymbol{\pi}})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}(\mathbf{1}_a - \bar{\boldsymbol{\pi}}) \\
&= \mathbf{1}_a^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\mathbf{1}_a + \bar{\boldsymbol{\pi}}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\bar{\boldsymbol{\pi}} - 2\cdot\mathbf{1}_a^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\bar{\boldsymbol{\pi}} \\
&\overset{(a)}{\le} 4\max_{a'\in\mathcal{A}} \mathbf{1}_{a'}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\mathbf{1}_{a'} = 4\max_{a'\in\mathcal{A}} \boldsymbol{\phi}_{a'}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\phi}_{a'}\,.
\end{aligned}$$

(a) holds by the Cauchy-Schwartz inequality because

$$-\mathbf{1}_a^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\bar{\boldsymbol{\pi}} \le \|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\mathbf{1}_a\|_2\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\bar{\boldsymbol{\pi}}\|_2 \le \max_{a'\in\mathcal{A}}\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\mathbf{1}_{a'}\|_2^2\,.$$

The last inequality in the above equation is satisfied because $\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\bar{\boldsymbol{\pi}}\|_2 \le \sum_{a'\in\mathcal{A}}\bar{\pi}_{a'}\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\mathbf{1}_{a'}\|_2 \le \max_{a'\in\mathcal{A}}\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\Phi}\mathbf{1}_{a'}\|_2$, which in turn follows by Jensen's inequality from the convexity of the ℓ_2-norm and the fact that $\bar{\boldsymbol{\pi}} \in \Delta_k$. The term $\bar{\boldsymbol{\pi}}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\Phi}\bar{\boldsymbol{\pi}}$ is upper bounded by an analogous argument.

Now Assumption 5.1 implies the following for each $a \in \mathcal{A}$:

$$\begin{aligned}
\boldsymbol{G}_n &\succeq \gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T} \\
\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\boldsymbol{G}_n &\succeq \boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\cdot\gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T} \succ \mathbf{0} \\
(\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\boldsymbol{G}_n)^{-1} &\preceq (\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\cdot\gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T})^{-1} \\
\boldsymbol{\phi}_a^\mathsf{T}(\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\boldsymbol{G}_n)^{-1}\boldsymbol{\phi}_a &\le \boldsymbol{\phi}_a^\mathsf{T}(\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\cdot\gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T})^{-1}\boldsymbol{\phi}_a \\
\boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\phi}_a &\le \boldsymbol{\phi}_a^\mathsf{T}(\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\cdot\gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T})^{-1}\boldsymbol{\phi}_a.
\end{aligned} \tag{29}$$

The second line holds because we assumed $\boldsymbol{\Sigma}_0 \succ 0$, and thus, $\boldsymbol{\Sigma}_0^{-1} \succ 0$, and adding a positive definite matrix preserves definiteness. The third line holds from the definiteness in the second line and Horn & Johnson (2013, corollary 7.7.4(a)). Finally, the fourth line holds from the definition of positive semi-definiteness.

We continue by applying the Woodbury matrix identity to (29), which give us the following inequality for each $a \in \mathcal{A}$:

$$\begin{aligned}\boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_n\boldsymbol{\phi}_a &\leq \boldsymbol{\phi}_a^\mathsf{T}(\boldsymbol{\Sigma}_0^{-1} + \bar{\sigma}^{-2}\cdot\gamma n\cdot\boldsymbol{\phi}_a\boldsymbol{\phi}_a^\mathsf{T})^{-1}\boldsymbol{\phi}_a &= \frac{1}{(\boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_a)^{-1} + \bar{\sigma}^{-2}\cdot\gamma n}\\ &\leq \frac{1}{\lambda_{\max}(\boldsymbol{\Sigma}_0)^{-1} + \bar{\sigma}^{-2}\cdot\gamma n}\,,\end{aligned}$$

where $\lambda_{\max}$ computes the maximum eigenvalues of the matrix. The inequality above holds because

$$0 \leq \boldsymbol{\phi}_a^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_a \leq \lambda_{\max}(\boldsymbol{\Sigma}_0)\|\boldsymbol{\phi}_a\|,$$

which can be seen from the eigendecomposition of the symmetric matrix. Substituting the inequality above proves the theorem.

To prove the special case of the theorem with $\boldsymbol{\mu}_n = \mathbf{0}$, let $\boldsymbol{\pi}^0$ be the solution in the first iteration of Algorithm 1. Given the posterior distribution of $\tilde{\boldsymbol{\theta}}_D$, the policy $\boldsymbol{\pi}^0$ is chosen as

$$\begin{aligned}\boldsymbol{\pi}^0 &\in \arg\min_{\boldsymbol{\pi}\in\Delta_k}\max_{a\in\mathcal{A}}\ \mathbf{0}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}}\cdot\nu_a^0\\ &= \arg\min_{\boldsymbol{\pi}\in\Delta_k}\max_{a\in\mathcal{A}}\ \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}}\,.\end{aligned}$$

The square of this minimization problem can be formulated as a convex quadratic program

$$\min_{t\in\mathbb{R},\ \boldsymbol{\pi}\in\Delta_k}\left\{t \mid t \geq \|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_a - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Phi}\boldsymbol{\pi}\|_2^2\,,\ \forall a\in\mathcal{A}\right\}. \tag{30}$$

Because $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Phi}\boldsymbol{\pi} \in \mathbb{R}^d$ and is a convex combination of points in $\mathbb{R}^d$, there exists an optimal $\boldsymbol{\pi}^0$ such that $l = |\{a\in\mathcal{A} \mid \pi_a^0 > 0\}| \leq d+1$ (Rockafellar & Wets, 2009). Then, let $\hat{a} \in \arg\max_{a'\in\mathcal{A}}\pi_{a'}^0$. We have that $\pi_{\hat{a}}^0 \geq \frac{1}{l}$ because l actions are positive, and the constraint $t \geq \|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_a - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Phi}\boldsymbol{\pi}\|_2^2$ is active (holds with equality). If the constraint were not active, this would be a contradiction with the optimality of $\boldsymbol{\pi}^0$ because decreasing $\pi_{\hat{a}}^0$ would reduce the objective. Then, using the inequalities above and the triangle inequality, we get that the optimal $t^\star$ in (30) satisfies

$$\begin{aligned}\sqrt{t^\star} &= \|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{\hat{a}} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Phi}\boldsymbol{\pi}^0\|_2 = \left(1 - \max_{a''\in\mathcal{A}}\pi_{a''}^0\right)\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{\hat{a}} - \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{a'}\|_2\\ &\leq \left(1 - \max_{a''\in\mathcal{A}}\pi_{a''}^0\right)\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{\hat{a}} - \boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{a'}\|_2 \leq 2\left(1 - \max_{a''\in\mathcal{A}}\pi_{a''}^0\right)\|\boldsymbol{\Sigma}_n^{1/2}\boldsymbol{\phi}_{a'}\|_2.\end{aligned}$$

The remainder of the proof follows from the same steps as the proof of Theorem 5.2. The lower bound on $\max_{a'\in\mathcal{A}}\hat{\pi}_{a'}$ holds from the existence of π^0 with at most $d+1$ positive elements, as discussed above. □

C.2. Proof of Theorem 5.3

Proof of Theorem 5.3. First, from the construction of Example 1, we have that

$$a_1 \in \arg\min_{a\in\mathcal{A}}\mu_a - \beta\cdot\sigma_a = \arg\min_{a\in\mathcal{A}}\beta\cdot\sigma_a - \beta\cdot\sigma_a = \mathcal{A},$$

and therefore $\boldsymbol{\pi}_{\mathrm{LCB}}$ is the policy returned by LCB that breaks ties as specified. Then, using Theorem 4.4, we bound the regret of LCB as

$$\mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{LCB}}) \ \geq\ \mu_{a_2} + \sigma_{a_2}\cdot\kappa_1(k-1) = \beta\cdot\sigma_{a_2} + \sigma_{a_2}\cdot\kappa_1(k-1) = (\beta + \kappa_1(k-1))\cdot\sigma_{a_2}.$$

In contrast, `Greedy` selects a_2 deterministically since

$$a_2 \in \arg\min_{a\in\mathcal{A}}\mu_a = \{a_2,\ldots,a_k\}\,.$$

Then, using Theorem 4.3 and (11) in particular, we upper bound the regret of $\boldsymbol{\pi}_{\mathrm{G}}$ as

$$\begin{aligned}\mathfrak{R}_\delta(\boldsymbol{\pi}_{\mathrm{G}}) &\leq \max_{a\in\mathcal{A}} \mu_a^{\boldsymbol{\pi}_{\mathrm{G}}} + \sigma_a^{\boldsymbol{\pi}_{\mathrm{G}}} \cdot \kappa_{\mathrm{u}}(k) \\ &= \max_{a\in\{a_2,\dots,a_k\}} \mu_a^{\boldsymbol{\pi}_{\mathrm{G}}} + \sigma_a^{\boldsymbol{\pi}_{\mathrm{G}}} \cdot \kappa_{\mathrm{u}}(k) \\ &= \max_{a\in\{a_2,\dots,a_k\}} \sqrt{\sigma_a^2 + \sigma_{a_2}^2} \cdot \kappa_{\mathrm{u}}(k) \\ &= \max_{a\in\{a_2,\dots,a_k\}} \sqrt{2}\cdot\sigma_{a_2} \cdot \kappa_{\mathrm{u}}(k).\end{aligned}$$

The equalities follow from substituting the definitions of relative means and variances and from algebraic manipulation. □

D. Sub-Gaussian Posterior

We discuss here how our results can extend to $\tilde{\boldsymbol{\theta}}_D$ with sub-Gaussian distributions. The modifications necessary are quite minor. The key to the approach is to generalize Theorem 4.3 to a sub-Gaussian distribution as the following theorem states.

Theorem D.1. *Suppose that $\tilde{\boldsymbol{\theta}}_D$ is a random variable with an atomless distribution that is sub-Gaussian with mean $\boldsymbol{\mu}$ and covariance factor $\boldsymbol{\Sigma}$. Then the regret for each $\boldsymbol{\pi} \in \Delta_k$ satisfies that*

$$\begin{aligned}\mathfrak{R}_\delta(\boldsymbol{\pi}) &\leq \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mathrm{VaR}_{1-\delta\xi_a}\left[r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right] \\ &\leq \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}} \cdot \sqrt{2\log(1/\delta\xi_a)}.\end{aligned} \tag{31}$$

Proof. The first inequality in (31) holds by Theorem 4.3 since this inequality does not require that the posterior is normal. That is, we have that

$$\begin{aligned}\mathfrak{R}_\delta(\boldsymbol{\pi}) &\leq \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mathrm{VaR}_{1-\delta\xi_a}\left[r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right] \\ &= \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mathrm{VaR}_{1-\delta\xi_a}\left[(\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D\right] \\ &\leq \min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mathrm{EVaR}_{1-\delta\xi_a}\left[(\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D\right].\end{aligned}$$

The last inequality follows from Lemma A.1. For each $a \in \mathcal{A}$, the definition of a multi-variate sub-Gaussian random variable in (24) with $\boldsymbol{w}^\mathsf{T} = (\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}$ implies that that $(\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D$ is sub-Gaussian with mean $\mu = (\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\mu}$ and a variance factor $\sigma^2 = (\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})$. Therefore, from Lemma A.2 we have

$$\min_{\boldsymbol{\xi}\in\Delta_k} \max_{a\in\mathcal{A}} \mathrm{EVaR}_{1-\delta\xi_a}\left[(\mathbf{1}_a - \boldsymbol{\pi})^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D\right] \leq \boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}) + \|\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi})\|_{\boldsymbol{\Sigma}} \cdot \sqrt{2\log(1/\delta\xi_a)}\,,$$

which proves the result. □

Theorem 4.5 can also be extended to the sub-Gaussian setting but seems to require an additional assumption that $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}$ is a sub-gamma random variable, and we leave it for future work.

Armed with Theorem D.1, we can adapt Algorithm 1 to the sub-Gaussian setting simply by setting $\nu_a^0 = \sqrt{2\log(k/\delta)}$. Note that (14) already uses the correct inequality for a sub-Gaussian distribution.

E. Other Objectives

We now briefly discuss two other related objectives as alternatives to minimizing the *high-confidence Bayesian regret*, defined in (4) and (5). These objectives may be preferable in some settings because they can be solved optimally using simple and tractable techniques.

E.1. Expected Bayes Regret

The first objective we discuss is *expected Bayes regret*, which is obtained by simply replacing the VaR by expectation in (5). In this case, the goal of the agent is to minimizes the expected regret, defined as

$$\min_{\boldsymbol{\pi}\in\Delta_k} \mathbb{E}\left[\max_{a\in\mathcal{A}} r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right].$$

Using the linearity property of the expectation operator and the reward function r, we have

$$\arg\min_{\boldsymbol{\pi}\in\Delta_k} \mathbb{E}\left[\max_{a\in\mathcal{A}} r(a;\tilde{\boldsymbol{\theta}}_D) - r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right] = \arg\max_{\boldsymbol{\pi}\in\Delta_k} \mathbb{E}\left[r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right] = \arg\max_{\boldsymbol{\pi}\in\Delta_k} r\left(\boldsymbol{\pi};\mathbb{E}\left[\tilde{\boldsymbol{\theta}}_D\right]\right).$$

This means it is sufficient to maximize the return for the mean posterior parameter value. In most case, such as when the posterior over $\tilde{\boldsymbol{\theta}}_D$ is normal, this is an easy optimization problem to solve optimally.

E.2. High-confidence Return

The second objective we discuss is *high-confidence return*, which is obtained by simply replacing the regret with return in (5). In this case, the goal of the agent is to minimizes the VaR of the return random variable as

$$\min_{\boldsymbol{\pi}\in\Delta_k} \mathrm{VaR}_{1-\delta}\left[-r(\boldsymbol{\pi};\tilde{\boldsymbol{\theta}}_D)\right] = \min_{\boldsymbol{\pi}\in\Delta_k} \mathrm{VaR}_{1-\delta}\left[-\boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D)\right]. \tag{32}$$

One may think of this objective as minimizing the regret with respect to 0. The reward inside is negated because we use VaR which measures costs rather than rewards. Note that $-\mathrm{VaR}_{1-\delta}[-\tilde{x}] \approx \mathrm{VaR}_\delta[\tilde{x}]$ with an equality for atomless (continuous) distributions.

When $\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the optimization in (32) can be solved *optimally* using an LCB-style algorithm. Then, using the properties of linear transformation of normal distributions, for each $\boldsymbol{\pi} \in \Delta_k$, we obtain

$$\boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\tilde{\boldsymbol{\theta}}_D \sim \mathcal{N}(\boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\mu},\, \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Phi}\boldsymbol{\pi}).$$

Combining the objective in (32) with (2), we get that the objective is

$$\max_{\boldsymbol{\pi}\in\Delta_k} \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\mu} - \sqrt{\boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Phi}\boldsymbol{\pi}} \cdot z_{1-\delta}. \tag{33}$$

Recall that $z_{1-\delta}$ is the $1-\delta$-th quantile of the standard normal distribution. We can reformulate (33) as the following second-order conic program (for $\delta \leq 1/2$)

$$\begin{aligned} \underset{\boldsymbol{\pi}\in\mathbb{R}^k,\, s\in\mathbb{R}}{\text{maximize}} \quad & \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\mu} - z_{1-\delta}\cdot s \\ \text{subject to} \quad & s^2 \leq \boldsymbol{\pi}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{\Phi}\boldsymbol{\pi}, \\ & \mathbf{1}^\mathsf{T}\boldsymbol{\pi} = 1, \quad \boldsymbol{\pi} \geq \mathbf{0}. \end{aligned}$$

When restricted to deterministic policies, the optimization in (33) reduces to a plain deterministic LCB algorithm. The `FlatOPO` algorithm can be seen as an approximation of (33) in which $z_{1-\delta}$ is replaced by its upper bound.

F. Additional Experimental Details

In this section, we provide some additional experimental results. First, Figures 5 and 6 report the same results as Figures 3 and 4 but also report the 95% confidence interval for the average regret over the 100 runs. Second, we report the effect of the tightening step on the quality of the bounds in Figure 7 compared to a scenario-based estimation. In this simplified example, we fix some policy $\boldsymbol{\pi}$ and *assume* the particular parameters of the distribution of $\tilde{x}_a = \tilde{\boldsymbol{\theta}}_D^\mathsf{T}\boldsymbol{\Phi}(\mathbf{1}_a - \boldsymbol{\pi}), a \in \mathcal{A}$, which is normal by Lemma 4.2. The results in the figure show that when the distribution $\tilde{\boldsymbol{x}}$ is close to i.i.d. the tightening step does not improve the bound. This is expected since the optimal $\boldsymbol{\xi}$ in (14) is nearly uniform. However, when the means or variances of the $\tilde{x}_a$ vary across actions $a \in \mathcal{A}$, then the tightening step can significantly reduce the error bound.

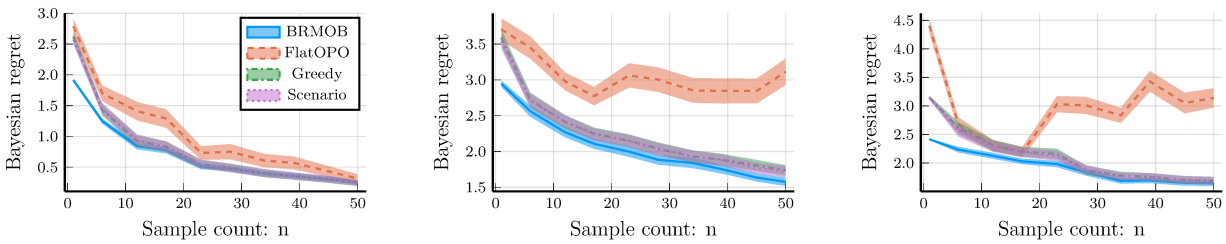

Figure 5. Bayesian regret with $k = d = 5$ *(left)*, $k = d = 50$ *(middle and right)*. The prior mean is $\boldsymbol{\mu}_0 = \mathbf{0}$ *(left and middle)* and $(\boldsymbol{\mu}_0)_a = \sqrt{a}$ for $a = 1, \dots 50$ *(right)*.

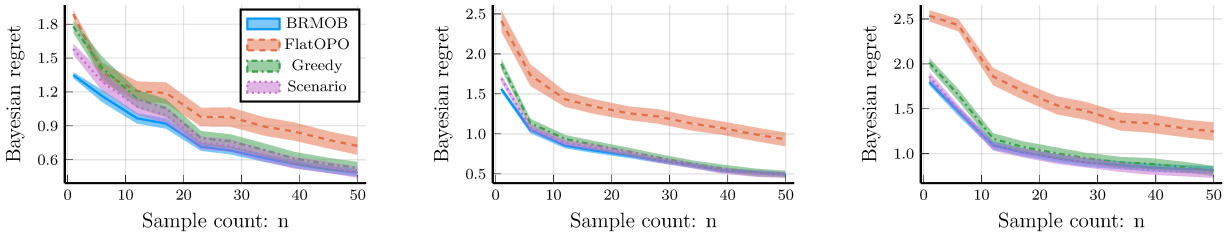

Figure 6. Bayesian regret with $d = 4$ and $k = 10$ *(left)*, $k = 50$ *(middle)*, and $k = 100$ *(right)*.

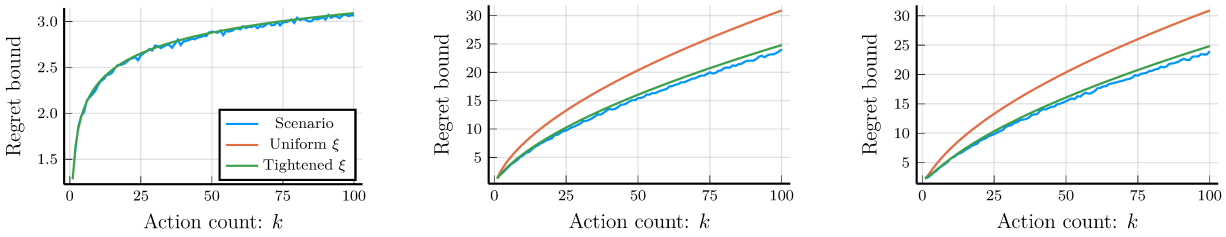

Figure 7. Regret bounds in Theorem 4.3 for different choices of $\boldsymbol{\xi}$ as a function of k. The posterior distribution of $\tilde{\boldsymbol{x}}$ is normal with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ *(left)*, $\boldsymbol{\mu} = \mathbf{0}$, $\Sigma_{aa} = a^2/k$ *(middle)*, and $\mu_a = a/k$, $\Sigma_{aa} = a^2/k$ *(right)* with $a = 1, \dots, k$.

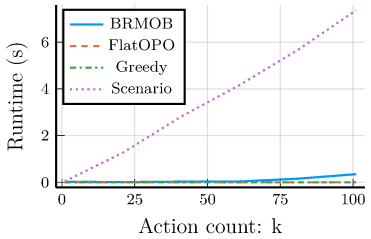

Figure 8. Runtime comparison of algorithms in seconds for a problem with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$.

Figure 8 compares the runtime of the algorithms considered as a function of the number of arms. The runtime excludes the time to compute the posterior distribution which is independent of the particular method considered. We use MOSEK to compute the SOCP optimization and do not run any tightening steps. The number of samples m needed for the Scenario algorithm was derived from the Dvoretzky-Kiefer-Wolfowitz bound as

$$m \;=\; \frac{100}{(1-0.95)^2} \log\left(\frac{2 \cdot k}{0.05}\right).$$

This number of samples guarantees a small sub-optimality gap with probability 95%. We suspect, however, that this number of samples can be reduced with more careful assumptions and algorithmic design (Calafiore & Campi, 2005; Nemirovski & Shapiro, 2007; 2006). Such analysis is beyond the scope of this work.