Safe Exploration for Efficient Policy Evaluation and Comparison

Runzhe Wan [1] **Branislav Kveton** [2] **Rui Song** [1]

Abstract

High-quality data plays a central role in ensuring the accuracy of policy evaluation. This paper initiates the study of efficient and safe data collection for bandit policy evaluation. We formulate the problem and investigate its several representative variants. For each variant, we analyze its statistical properties, derive the corresponding exploration policy, and design an efficient algorithm for computing it. Both theoretical analysis and experiments support the usefulness of the proposed methods.

1. Introduction

Bandit policies have been widely applied to areas including advertising (Bottou et al., 2013), search (Li et al., 2011) and healthcare (Zhou et al., 2017). Before deploying a target policy, it is typically crucial to have an accurate evaluation of its performance, which can subsequently provide valuable information for deployment decisions or model improvement. This is in general achieved by utilizing logged data, known as the *off-policy evaluation (OPE)* problem.

Although OPE has been studied extensively (Dudík et al., 2014; Li et al., 2015; Swaminathan et al., 2016; Wang et al., 2017; Su et al., 2020; Kallus et al., 2021; Cai et al., 2021), the existing works mainly focus on various estimators with a *fixed* dataset. However, limited attention has been paid to the dataset itself, which plays a vital role in estimation accuracy. For instance, inverse probability weighting (IPW, Li et al., 2015), one popular OPE method, has a key requirement that the logging policy has "full support". This usually translates into requiring the logging policy to be close enough to the target policy (Sachdeva et al., 2020; Tran-The et al., 2021). For direct method (DM, Dudík et al., 2014), it is also common that some feature directions are less explored in the logged dataset, which impacts the accuracy of this regression-based estimator.

[1]Department of Statistics, North Carolina State University [2]Amazon. Correspondence to: Rui Song <rsong@ncsu.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

To improve the accuracy of policy evaluation, a natural idea is to actively collect high-quality data, and ideally, the data collection rule should be tailored to the given evaluation task. This problem is surprisingly unexplored. Moreover, due to the common safety concerns (Wu et al., 2016; Zhu & Kveton, 2022), it is of great practical importance to make sure the data collection rule is safe.

In this paper, we initiate the study of *Safe Exploration for Policy Evaluation and Comparison (SEPEC)*, the design of an efficient and safe exploration policy that collects a high-quality dataset for the given evaluation task. We focus on non-adaptive policies, as they are simple to implement logistically. Indeed, in practice, the challenge of requiring investments (e.g., in infrastructures) for adaptive algorithms has been widely recognized (Zanette et al., 2021; Zhu & Kveton, 2022), especially before the value is proved via OPE. Our proposal is summarized in Figure 1.

Another possible approach to policy evaluation is being on-policy (i.e., following the target policy). However, the safety concern typically does not allow a direct deployment. In addition, perhaps surprisingly, the on-policy evaluation is not optimal under some setups (e.g., when the variances of different arms differ) or for some evaluation tasks (e.g., comparing the target policy with a baseline policy). See Section 3.1 for an example. Therefore, more careful analysis and design are required.

As expected and shown later, our proposed solutions vary across different bandit setups, evaluation tasks, and value estimators. To shed light on this novel problem, we study its three representative variants, including multi-armed bandits (MAB) with IPW, contextual MAB (CMAB) with IPW, and linear bandit with DM. Our results can be extended to several other setups that are introduced as well. Finally, although other evaluation problems (e.g., estimating the value of the target policy in Section 3.4) can be addressed similarly to our work, for concreteness, we focus on *policy comparison*, where we estimate the value difference between the target and baseline policies. This task is closely related to deployment decision making.

Contribution. Our contributions can be summarized as follows. First, motivated by practical needs for improving policy evaluation accuracy and the common safety concerns, we propose and formulate the SEPEC framework. To

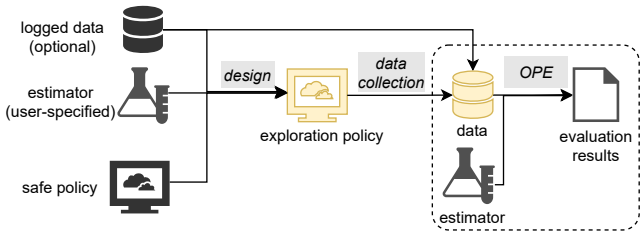


Figure 1: SEPEC: designing a safe and efficient exploration policy to collect high-quality data for a given evaluation task.

the best of our knowledge, this is the *first* work studying how to collect data for efficient policy evaluation, with or without safety constraints. Second, we investigate its three representative variants thoroughly. We analyze their statistical properties, design tractable optimization problems, and provide efficient optimization algorithms. The situations with or without side information are both considered. Third, we theoretically prove the efficiency and optimality of our approach. Lastly, we demonstrate the superior performance of SEPEC through extensive experiments.

2. Objective and Existing Approaches

For concreteness, we first introduce the objective of SEPEC under MAB, and will extend it to other settings later. For any positive integer M, we denote the set $\{1, \dots, M\}$ by $[M]$. Let Δ_{K-1} be the K-dimensional simplex. In a K-armed bandit, we represent a policy by a vector $\pi = (\pi(1), \dots, \pi(K))^T \in \Delta_{K-1}$, where $\pi(a)$ is the probability that arm a is pulled under policy π. After choosing an arm A_t, we receive its stochastic reward R_t. Let $r_a = \mathbb{E}[R_t | A_t = a]$ be the expected reward of arm a and $\sigma_a^2 = \mathrm{var}(R_t | A_t = a)$ be the variance of its rewards. Let $\boldsymbol{r} = (r_1, \dots, r_K)^T$. The value of a policy is $V(\pi) = \sum_{a \in [K]} \pi(a) r_a$. Sometimes we write $V_{\boldsymbol{r}}(\pi)$ to emphasize the dependency on $\boldsymbol{r}$. We make the standard assumptions that $r_a \in [0, 1]$ and $\sigma_a^2 \le \sigma^2$ for some $\sigma > 0$. A problem instance is specified by $(\boldsymbol{r}, \{\sigma_a\})$.

We assume that we are given a *target policy* π_1 and a safe *baseline policy* π_0. We may also have side information, such as an existing dataset $\mathcal{D}_0$. We focus on estimating the value difference $V(\pi_1) - V(\pi_0)$, although several other estimands (e.g., $V(\pi_1)$) can be addressed similarly. In practice, OPE methods are commonly used to estimate this lift and test its significance using existing data, which determines the deployment decision. However, the dataset is typically assumed as well explored, and little attention has been paid to how it arises. We aim to fill this gap. Specifically, given an *exploration budget* T, a *risk tolerance* $\epsilon \in (0, 1)$, and a user-specified estimator that maps a dataset $\mathcal{D}$ to a value estimate of policy π as $\widehat{V}(\pi; \mathcal{D})$, we aim to design an *exploration policy* π_e to collect a dataset $\mathcal{D}_e$ of size T, while achieving the following two objectives simultaneously:

- **Safety**: Exploration should be safe, in the sense that $V_{\boldsymbol{r}}(\pi_e) \ge (1 - \epsilon) V_{\boldsymbol{r}}(\pi_0)$ holds, either for all problem instances $\boldsymbol{r}$ or with a high probability given side information.
- **Efficiency**: The exploration should be efficient, which we define as the minimization of $\mathrm{var}\big(\widehat{V}(\pi_1; \mathcal{D}_0 \cup \mathcal{D}_e) - \widehat{V}(\pi_0; \mathcal{D}_0 \cup \mathcal{D}_e)\big)$, i.e., the maximization of the evaluation accuracy with the given estimator. Specifically, we mainly focus on unbiased value estimators, and in this case, the variance is closely related to many practical metrics, such as the mean squared error and the statistical power of testing $V(\pi_1) > V(\pi_0)$. See Section 4.2 for details.

Two existing approaches. To satisfy the safety constraint while exploring, the most popular practice (Thomas et al., 2015; Jiang & Li, 2016; Slivkins, 2019), arguably, is to allocate an $\epsilon \in (0, 1)$ proportion of the budget to an exploration policy $b' \in \Delta_{K-1}$, and construct a *mixture policy* $\pi_e = \epsilon b' + (1 - \epsilon)\pi_0$. The common choice of b' is π_1 (i.e., on-policy) or the uniform distribution (i.e., random exploration). This approach, albeit being safe, could be inefficient. For example, π_e is confined in a small policy class, and as we will show shortly, we can actually obtain a safe policy directly in a larger class via optimization. Besides, even within this policy class, b' needs to be carefully designed and this problem alone is also underexplored.

Another existing approach is Zhu & Kveton (2022), which also aims to collect high-quality bandit feedback in a safe manner. Our approach to handling the safety constraints is partially inspired by this paper. However, their targeted application is policy optimization, and hence they propose to maximize $\min_{a \in [K]} \pi_e(a)$, or in other words, they aim to collect data for evaluating *all* policies jointly by reducing uncertainty uniformly. Therefore, their exploration policy is not tailored for evaluation, and will be less efficient without utilizing its specific structure. The SEPEC problem requires a more careful task-oriented analysis and design. A concrete example follows below.

Illustrative example. We start with analyzing the efficiency in a toy example without safety constraints. Consider $\pi_0 = (0.4, 0.1, 0.5)^T$ and $\pi_1 = (0.1, 0.4, 0.5)^T$. Then, one immediate observation is that, arm 3 does not affect the estimate of the value difference $V(\pi_1) - V(\pi_0) = (0.1 - 0.4) r_1 + (0.4 - 0.1) r_2 + 0$, and so no budget should be spent on arm 3 when exploring. Therefore, either following π_1, running A/B experiment (allocating $x\%$ budget to π_1 and $(100 - x)\%$ to π_0), or using the uniform exploration

of Zhu & Kveton (2022) is clearly sub-optimal. The problem will become even more interesting when we maximize the efficiency with safety constraints, stochastic contexts, and generalization functions.

3. Methodology

In this section, we discuss our methodology by studying three OPE setups, including MAB with IPW, CMAB with IPW, and linear bandits with DM. We choose these setups as they are representative, covering common features such as stochastic contexts and generalization functions. These discussions provide insights into this novel problem, and a few extensions will be introduced as well. In addition, we study both the case without side information, where we need to ensure safety under all instances, and the case with side information, which will enter the final estimator and also can help us relax the safety constraint.

For each setup, we start by analyzing the variance of the whole exploration and evaluation procedure. This raw objective is typically infeasible to optimize (e.g., it may involve unknown parameters). Therefore, we then design a tractable surrogate optimization problem, by solving which one can obtain an efficient policy. The optimization algorithm will be introduced at the end. We summarize the high-level idea in Figure 2.

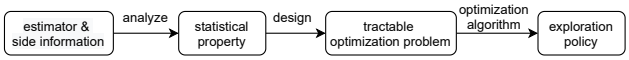


Figure 2: Workflow for designing an exploration policy.

3.1. MAB with IPW

We first apply SEPEC to evaluating an MAB policy using the IPW estimator, without side information. With a dataset $\mathcal{D}_e = \{(A_t, R_t)\}_{t=1}^T$ collected following π_e, the IPW estimator is

$$\widehat{V}_{IPW}(\pi; \mathcal{D}_e) = \frac{1}{T}\sum_{t=1}^{T} \frac{\pi(A_t)}{\pi_e(A_t)} R_t.$$

Our objective can be written as

$$\begin{aligned} &\arg\min_{\pi_e \in \Delta_{K-1}} \operatorname{var}\big(\widehat{V}_{IPW}(\pi_1; \mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0; \mathcal{D}_e)\big) \\ &s.t.\ \ V_{\boldsymbol{r}}(\pi_e) \geq (1-\epsilon) V_{\boldsymbol{r}}(\pi_0), \forall \boldsymbol{r} \in [0,1]^K. \end{aligned}$$

This problem is particularly challenging because (i) the dependency of the objective on π_e is complex, and (ii) the safety constraint involves the unknown function V and hence is not easy to guarantee.

Objective function. We aim to first transform the objective into a tractable form. Let $\pi_\Delta = \pi_1 - \pi_0$. We have

$$\begin{aligned} & T \times \operatorname{var}\Big(\widehat{V}_{IPW}(\pi_1; \mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0; \mathcal{D}_e)\Big) \\ &= T \times \operatorname{var}\Big(T^{-1}\sum_{t=1}^{T} \frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t\Big) \\ &= \operatorname{var}_{A_t \sim \pi_e}\Big(\frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t\Big) \qquad (1) \\ &= \mathbb{E}_{A_t \sim \pi_e}\Big[\operatorname{var}(\frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t | A_t)\Big] \\ &\quad + \operatorname{var}_{A_t \sim \pi_e}\Big[\mathbb{E}(\frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t | A_t)\Big] \\ &= \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} \sigma_a^2 + \operatorname{var}_{A_t \sim \pi_e}\Big[\frac{\pi_\Delta(A_t)}{\pi_e(A_t)} r_{A_t}\Big] \qquad (2) \\ &= \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} \sigma_a^2 + \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} r_a^2 - c, \qquad (3) \end{aligned}$$

where $c = (V(\pi_1) - V(\pi_0))^2$ is independent of π_e. The third equality is due to the law of total variance and the last one is due to $\operatorname{var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$. We note that (3) cannot be optimized directly, as σ_a's and r_a's are in general unknown. In fact, there does not exist a policy π_e that achieves global optimum for all instances (see Appendix D.4 for proof). However, this transformation provides insights into an efficient allocation rule. Without additional information, we consider an upper bound of the objective, by replacing σ_a's and r_a's with their joint upper bounds. Our relaxed objective is

$$\arg\min_{\pi_e \in \Pi_\Delta} \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}, \qquad (4)$$

where the feasible set $\Pi_\Delta = \{\pi \in \Delta_{K-1} : \pi(a) > 0 \text{ if } \pi_\Delta(a) > 0\}$ can be easily verified as convex. This ensures the *positivity assumption* required by IPW (Li et al., 2015). Although commonly assumed, this assumption has been found violated in many OPE applications (Sachdeva et al., 2020; Tran-The et al., 2021), where IPW methods can fail catastrophically. This is particularly severe with contextual bandits or large action space. SEPEC proactively resolves this issue. Finally, without safety constraints, the solution of (4) is $\pi_e(a) \propto |\pi_\Delta(a)|$. This result is intuitive: the larger the difference, the more we explore.

We make two remarks regarding the raw objective function (2). First, perhaps surprisingly, by similar arguments, we can show that the on-policy strategy (i.e., $\pi_e = \pi_1$) is not optimal even in minimizing $\operatorname{var}\big(\widehat{V}(\pi_1; \mathcal{D}_e)\big)$ without safety constraints, when σ_a's or r_a's vary across arms. Instead, $\pi_e(a) \propto \pi_1(a) \times (r_a^2 + \sigma_a^2)$ is the most efficient one.

Second, the two terms of (2) correspond to the *intrinsic uncertainty* from the reward noise after pulling arms and the *extrinsic randomness* from sampling arms. When $T \gg K$,

it is possible to (approximately) remove the second term by considering a deterministic allocation. Specifically, given π_e, instead of sampling the T data points, we can assign $N_a = T \times \pi_e(a)$ data points to the ath arm, which is known as the optimal continuous design in statistics (Lee, 1988). However, in general, it is hard to guarantee that N_a's are all integers, and one has to resort to randomized rounding to obtain an allocation $\{N_a\}$ that is close to π_e. Although a few such techniques (Bouhtou et al., 2010; Allen-Zhu et al., 2017) can yield small variance inflation when $T \gg K$, they are in general hard to generalize to broader setups such as contextual bandits, where the budget for each context is small and also stochastic. This is a setting of sharp difference with the standard experimental design literature. Therefore, we choose to sample from Multinomial(T, π_e), which is a popular randomized rounding approach recently (Azizi et al., 2021; Zhu & Kveton, 2022).

Optimization without side information. With a tractable and appropriate surrogate objective, we are now ready to consider the safety constraint. The first observation is that $V_{\boldsymbol{r}}(\pi_e) \geq (1-\epsilon)V_{\boldsymbol{r}}(\pi_0), \forall \boldsymbol{r} \in [0,1]^K$ is equivalent to $\pi_e(a) \geq (1-\epsilon)\pi_0(a), \forall a \in [K]$. This is because we need to ensure safety even in the worst case. Indeed, if there exists at least one arm a such that $\pi_e(a) < (1-\epsilon)\pi_0$, we can always construct an instance to violate the constraint, with $r_a = 1$ and $r_{a'} = 0, \forall a' \neq a$. Thus our optimization problem is

$$\begin{aligned} &\arg\min_{\pi_e \in \Pi_\Delta} \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} \\ &s.t.\ \ \pi_e(a) \geq (1-\epsilon)\pi_0(a), \forall a \in [K], \end{aligned} \tag{5}$$

and its solution is a safe policy. This problem is convex and hence can be efficiently solved by many optimization algorithms (Boyd et al., 2004).

Utilization of side information. Next, we study how to improve over (5) when there is side information, such as an existing dataset $\mathcal{D}_0$ generated by a known behavior policy π_b. Typically π_b is just π_0, but could be different. We assume π_b satisfies the positivity assumption; otherwise one can use IPW with a mixed propensity (Kallus et al., 2021). Specifically, we consider IPW with multiple logging policies (Kallus et al., 2021), and the original problem becomes

$$\begin{aligned} &\arg\min_{\pi_e \in \Pi_\Delta} \operatorname{var}\Big(\sum_{t=1}^T \frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t + \sum_{(A_i,R_i) \in \mathcal{D}_0} \frac{\pi_\Delta(A_i)}{\pi_b(A_i)} R_i \Big) \\ &s.t.\ \ V_{\boldsymbol{r}}(\pi_e) \geq (1-\epsilon)V_{\boldsymbol{r}}(\pi_0), \forall \boldsymbol{r} \in \mathcal{C}_\delta(\mathcal{D}_0). \end{aligned}$$

Here $\mathcal{C}_\delta(\mathcal{D}_0)$ is an $(1-\delta)$-confidence region on $\boldsymbol{r}$ obtained from $\mathcal{D}_0$, which can be constructed based on concentration inequalities. Alternatively, one can consider a Bayesian viewpoint when there is a prior on $\boldsymbol{r}$ that summarizes domain knowledge, and $\mathcal{C}_\delta(\mathcal{D}_0)$ is the credible region derived from the corresponding posteriors. We assume $\mathcal{C}_\delta(\mathcal{D}_0)$ is a convex region, which is commonly the case following either approach. See Appendix A.1 for details. The convexity allows finding the most violated constraint, for any fixed π_e, efficiently. As such, the side information helps relax the safety constraint. Finally, we notice that for any $\boldsymbol{r}$, $V_{\boldsymbol{r}}(\pi_e) \geq (1-\epsilon)V_{\boldsymbol{r}}(\pi_0)$ is equivalent to $\boldsymbol{r}^T(\pi_e - (1-\epsilon)\pi_0) \geq 0$, a linear constraint in π_e.

In addition, the objective is equal to

$$\operatorname{var}\Big(\sum_{t=1}^T \frac{\pi_\Delta(A_t)}{\pi_e(A_t)} R_t \Big) + \operatorname{var}\Big(\sum_{(A_i,R_i) \in \mathcal{D}_0} \frac{\pi_\Delta(A_i)}{\pi_b(A_i)} R_i \Big),$$

thanks to the independence between $\mathcal{D}_e$ and $\mathcal{D}_0$. The second term does not depend on π_e and hence can be dropped in optimization. In other words, perhaps surprisingly, the additional dataset $\mathcal{D}_0$ does not *directly* enter the objective. The main reason is that, every data point contributes independently to the IPW estimator. This may not hold for other estimators, as shown for DM in Section 3.3.

Optimization with side information. By similar arguments to (5), we propose an optimization problem

$$\begin{aligned} &\arg\min_{\pi_e \in \Pi_\Delta} \sum_{a \in [K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} \\ &s.t.\ \ \boldsymbol{r}^T(\pi_e - (1-\epsilon)\pi_0) \geq 0, \forall \boldsymbol{r} \in \mathcal{C}_\delta(\mathcal{D}_0). \end{aligned} \tag{6}$$

This problem is particularly challenging as the constraint implicitly contains an *infinite* number of constraints. We note that, for any finite set $\Theta \subset \Delta_{k-1}$, the corresponding constraint $\boldsymbol{r}^T(\pi_e - (1-\epsilon)\pi_0) \geq 0, \forall \boldsymbol{r} \in \Theta$ is actually a finite number of linear constraints on π_e. Therefore, we propose to solve (6) by adapting the cutting-plane method (Bertsimas & Tsitsiklis, 1997). Specifically, the cutting-plane method iteratively adds more constraints to a finite set to approximate the convex feasible set, until no constraint in the original problem is violated. The convergence analysis is provided in Boyd & Vandenberghe (2007). We present the algorithm details in Appendix A.3.

Remark 1. *Recall that the raw objective (3) involves the unknown parameters σ_a's and r_a's. We relax it using the upper bounds to guarantee the worst-case efficiency. When side information is available, alternatively, one can try to infer these parameters and consider a different objective function than (6), with a Bayesian or high-probability guarantee. We discuss this more in Appendix B.5.*

3.2. CMAB with IPW

In this section, we apply `SEPEC` to contextual MAB (CMAB). At every round t, we observe a context $\mathbf{x}_t$ from a finite set of contexts $\mathcal{X}$, choose an arm $A_t \in [K]$, and then receive a stochastic reward R_t. We similarly define the mean reward $r(a|\mathbf{x}) = \mathbb{E}[R_t|A_t = a, \mathbf{x}_t = \mathbf{x}]$,

the mean reward vector $\boldsymbol{r}(\cdot|\mathbf{x}) \in [0,1]^K$, and the policy $\pi(\cdot|\mathbf{x}) \in \Delta_{K-1}$. We also use $\boldsymbol{r} = (\boldsymbol{r}(\cdot|\mathbf{x}))_{\mathbf{x}\in\mathcal{X}}$ and $\pi = (\pi(\cdot|\mathbf{x}))_{\mathbf{x}\in\mathcal{X}}$ to denote the corresponding vector concatenated over $\mathcal{X}$. With π_0 and π_1 given, we define π_Δ as their difference. For design purpose, we make a common assumption (Wu et al., 2015; Zhu & Kveton, 2022) that the context distribution $p(\mathbf{x})$ is known, since there is typically a large dataset of contexts. Let $V_{\boldsymbol{r}}(\pi|\mathbf{x}) = \sum_{a\in[K]} \pi(a|\mathbf{x}) r(a|\mathbf{x})$. The policy value is $V_{\boldsymbol{r}}(\pi) = \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x}) V_{\boldsymbol{r}}(\pi|\mathbf{x})$. Define $\Pi_\Delta = \{\pi : \pi(a|\mathbf{x}) > 0 \text{ if } \pi_\Delta(a|\mathbf{x}) > 0, \forall a, \mathbf{x}\}$. Π_Δ is a convex set by definition.

For any dataset $\mathcal{D}_e = \{(\mathbf{x}_t, A_t, R_t)\}_{t=1}^T$ collected following a policy π_e, we consider the IPW estimator $\widehat{V}_{IPW}(\pi; \mathcal{D}_e) = T^{-1} \sum_{t=1}^T \frac{\pi(A_t|\mathbf{x}_t)}{\pi_e(A_t|\mathbf{x}_t)} R_t$. Without side information, our problem is defined as

$$\begin{aligned} &\arg\min_{\pi_e\in\Pi_\Delta} T \times \mathrm{var}\big(\widehat{V}_{IPW}(\pi_1; \mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0; \mathcal{D}_e)\big) \\ &s.t.\ \ V_{\boldsymbol{r}}(\pi_e) \geq (1-\epsilon) V_{\boldsymbol{r}}(\pi_0), \forall \boldsymbol{r} \in [0,1]^{K|\mathcal{X}|}. \end{aligned} \tag{7}$$

We aim to construct an efficient and safe exploration policy as in the MAB (Section 3.1). The main additional challenges come from the stochasticity in the contexts and that we need to solve the problem jointly over all contexts.

No side information. We begin with studying the case without side information. We first note that our objective in (7) can be decomposed as

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_t\sim p(\mathbf{x})}\big\{\mathrm{var}[\big(\frac{\pi_\Delta(A_t|\mathbf{x}_t)}{\pi_e(A_t|\mathbf{x}_t)} R_t\big)|\mathbf{x}_t]\big\} \\ &\quad + \mathrm{var}_{\mathbf{x}_t\sim p(\mathbf{x})}\big\{\mathbb{E}[\big(\frac{\pi_\Delta(A_t|\mathbf{x}_t)}{\pi_e(A_t|\mathbf{x}_t)} R_t\big)|\mathbf{x}_t]\big\}. \end{aligned}$$

The second term is equal to

$$\mathrm{var}_{\mathbf{x}_t\sim p(\mathbf{x})}\big[\sum_{a\in[K]} \pi_\Delta(a|\mathbf{x}_t) r(a|\mathbf{x}_t)\big],$$

does not depend on π_e, and hence can be dropped in optimization. The first term can be regarded as the expectation of the MAB objective (1) over the context distribution.

The second observation is that, the safety constraint is equivalent to $\pi_e(\cdot|\mathbf{x}) \geq (1-\epsilon)\pi_0(\cdot|\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$. Otherwise, one can always construct a counter-example by setting $\boldsymbol{r}(\cdot|\mathbf{x}) = \mathbf{0}$, except for one context $\mathbf{x}'$ where the constraint is violated. Then, we have $V(\pi_e) = p(\mathbf{x}')V(\pi_e|\mathbf{x}') < (1-\epsilon)p(\mathbf{x}')V(\pi_0|\mathbf{x}') = (1-\epsilon)V(\pi_0)$, when $p(\mathbf{x}') > 0$.

Therefore, by similar arguments as in Section 3.1 on deriving the surrogate objective, we propose the following tractable optimization problem

$$\begin{aligned} &\arg\min_{\pi_e\in\Pi_\Delta} \sum_{\mathbf{x}\in\mathcal{X}} \sum_{a\in[K]} p(\mathbf{x}) \frac{\pi_\Delta^2(a|\mathbf{x})}{\pi_e(a|\mathbf{x})} \\ &s.t.\ \ \pi_e(\cdot|\mathbf{x}) \geq (1-\epsilon)\pi_0(\cdot|\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}, \end{aligned}$$

which can be solved by optimizing

$$\begin{aligned} &\arg\min_{\pi_e(\cdot|\mathbf{x})\in\Pi_\Delta(\mathbf{x})} \sum_a \frac{\pi_\Delta^2(a|\mathbf{x})}{\pi_e(a|\mathbf{x})} \\ &s.t.\ \ \pi_e(\cdot|\mathbf{x}) \geq (1-\epsilon)\pi_0(\cdot|\mathbf{x}), \end{aligned} \tag{8}$$

for each context $\mathbf{x} \in \mathcal{X}$ separately. This is due to the additive form of IPW and the worst-case safety constraint. The optimization can be done similarly to (5).

Side information. The scenario considered above is arguably conservative. For instance, one may expect to solve all contexts jointly, which allows us to violate the constraint on some contexts and remedy on the others. We next consider the case with side information, e.g., an existing dataset $\mathcal{D}_0 = \{(\mathbf{x}_i, A_i, R_i)\}$, that allows us to do so.

Regarding the objective function, by similar arguments to the MAB (Section 3.1), $\mathcal{D}_0$ only adds a π_e-independent term that can be dropped. However, $\mathcal{D}_0$ can help us to relax the safety constraint. We can similarly construct a high-probability region $\mathcal{C}_\delta(\mathcal{D}_0)$ such that $\boldsymbol{r} \in \mathcal{C}_\delta(\mathcal{D}_0)$ holds with probability at least $1-\delta$. An efficient and safe policy can then be obtained from

$$\begin{aligned} &\arg\min_{\pi_e\in\Pi_\Delta} \sum_{\mathbf{x}} p(\mathbf{x}) \sum_a \frac{\pi_\Delta^2(a|\mathbf{x})}{\pi_e(a|\mathbf{x})} \\ &s.t. \sum_{\mathbf{x}\in\mathcal{X}} \Big(\pi_e(\cdot|\mathbf{x}) - (1-\epsilon)\pi_0(\cdot|\mathbf{x})\Big) \\ &\qquad \times p(\mathbf{x}) \boldsymbol{r}(\cdot|\mathbf{x})^T \geq 0, \forall \boldsymbol{r} \in \mathcal{C}_\delta(\mathcal{D}_0), \end{aligned} \tag{9}$$

where we notice that, for every fixed $\boldsymbol{r}$, the constraint is linear in π_e. Therefore, similar to (6), we propose to solve this problem by combining convex optimization algorithms with the cutting-plane method. See Appendix A.3 for algorithm details.

3.3. Linear bandits with DM

Efficiency learning with a large action space typically relies on generalization functions, which also introduce interesting structures to the design of SEPEC algorithms. To shed light on this, we study linear bandits. At round t, we choose an arm $\mathbf{x}_t$ from a set of arms $\mathcal{A}$ with size K, and then receive a stochastic reward R_t. Each arm is represented by a d-dimensional vector. Note that we overload notation and use $\mathbf{x}_t$ to denote the feature vector of the pulled arm, instead of stochastic context. We assume that the expected reward of arm $\mathbf{x}$ is $r(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}^*$ for some unknown parameter vector $\boldsymbol{\theta}^*$. For design purpose, we assume homoscedasticity, i.e., $\mathrm{var}(R_t|A_t) \equiv \sigma^2$. We use most notation introduced in the MAB. We remark that $V_{\boldsymbol{\theta}^*}(\pi) = (\sum_{\mathbf{x}\in\mathcal{A}} \pi(\mathbf{x})\mathbf{x})^T\boldsymbol{\theta}^* = \bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{\theta}^*$, where $\bar{\boldsymbol{\phi}}_\pi = \sum_{\mathbf{x}\in\mathcal{A}} \pi(\mathbf{x})\mathbf{x}$ is the mean feature direction of π. We assume that $V_{\boldsymbol{\theta}^*}(\pi_0) \geq 0$.

To estimate the value of a policy π, we consider the direct method (DM, Dudík et al., 2014). With a dataset

$\mathcal{D} = \{(\mathbf{x}_i, R_i)\}$, DM first estimates $\boldsymbol{\theta}^*$ via least-square regression as $\hat{\boldsymbol{\theta}} = (\sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}} \mathbf{x}_i\mathbf{x}_i^T)^+(\sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}} \mathbf{x}_i R_i)$, and then plugs-in $\hat{\boldsymbol{\theta}}$ in the value definition to construct $\widehat{V}_{DM}(\pi;\mathcal{D}) = \sum_{\mathbf{x}\in\mathcal{A}} \pi(\mathbf{x})(\mathbf{x}^T\hat{\boldsymbol{\theta}}) = \bar{\boldsymbol{\phi}}_\pi^T\hat{\boldsymbol{\theta}}$.

Objective function. We pull T arms $\boldsymbol{\Phi} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)^T$ to collect a dataset $\mathcal{D}_{\boldsymbol{\Phi}}$, which is stochastic and depends on $\boldsymbol{\Phi}$. In this section, we directly consider the case with side information, since it does not make a significant difference in derivations. We assume we have an existing dataset $\mathcal{D}_0$ and denote the corresponding feature matrix as $\boldsymbol{\Phi}_0$, which we assume is full-rank. Its existence simplifies our exposition. Alternatively, one can always use forced exploration to form a basis, or consider regularized least squares. Towards the goal of variance minimization, we note that

$$
\begin{aligned}
&\mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e}\Big(\widehat{V}_{DM}(\pi_1;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}}) - \widehat{V}_{DM}(\pi_0;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}})\Big)\\
&= \mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\mathrm{var}\Big(\widehat{V}_{DM}(\pi_1;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}}) - \widehat{V}_{DM}(\pi_0;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}})|\boldsymbol{\Phi}\Big)\Big]\\
&\quad + \mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\mathbb{E}\Big(\widehat{V}_{DM}(\pi_1;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}}) - \widehat{V}_{DM}(\pi_0;\mathcal{D}_0\cup\mathcal{D}_{\boldsymbol{\Phi}})|\boldsymbol{\Phi}\Big)\Big]\\
&= \sigma^2\mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\bar{\boldsymbol{\phi}}_{\Delta\pi}^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi}\Big]\\
&\qquad + \mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e}[V(\pi_1) - V(\pi_0)]\\
&= \sigma^2\mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\bar{\boldsymbol{\phi}}_{\Delta\pi}^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi}\Big] + 0. \qquad (10)
\end{aligned}
$$

The generalization function introduces interesting structures. For example, unlike in IPW, $\mathcal{D}_0$ also enters our optimization objective, since actions are related through their features. Intuitively, we should spend less budget on those extensively explored directions. Moreover, unlike IPW, the second term of (10) is zero. This is because, conditioned on any set of sampled arms $\boldsymbol{\Phi}$, DM estimator is always unbiased and hence its conditional expectation is independent with $\boldsymbol{\Phi}$.

The optimization of (10) is very challenging, since it involves an expectation over the inverse of a random matrix. We notice that this objective is related to the G-optimal experiment design problem (Shah & Sinha, 2012), where we aim to minimize the maximum uncertainty by $\arg\min_{\boldsymbol{\Phi}} \max_{\tilde{\mathbf{x}}\in\mathcal{X}} \tilde{\mathbf{x}}^T(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\tilde{\mathbf{x}}$. For this problem, since a direct optimization is still NP-hard, it is common to solve its continuous relaxation (Shah & Sinha, 2012; Zhu & Kveton, 2022) $\arg\min_{\pi_e} \max_{\tilde{\mathbf{x}}\in\mathcal{X}} \tilde{\mathbf{x}}^T(T\sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T)^{-1}\tilde{\mathbf{x}}$, and then apply certain randomized rounding methods to construct $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ from the distribution π_e.

Motivated by the good property of this approximation (Shah & Sinha, 2012; Lattimore & Szepesvári, 2020), we consider a similar relaxation for our problem as

$$
\arg\min_{\pi_e\in\Pi'_\Delta} \bar{\boldsymbol{\phi}}_{\Delta\pi}^T(T\sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi},
$$

where $\Pi'_\Delta = \{\pi \in \Delta_{K-1}: \sum_{\mathbf{x}\in\mathcal{A}} \pi(\mathbf{x})\mathbf{x}\mathbf{x}^T \text{ is invertible}\}$ is a convex set by definition. In Section 4, we prove that this objective is actually the *asymptotic variance* of the DM estimator.

Optimization. After obtaining a tractable objective function, we next study solving the exploration policy with the safety constraint. Again, we can construct a high-probability region $\mathcal{C}_\delta(\mathcal{D}_0)$ for $\boldsymbol{\theta}$ in either a frequentist way or a Bayesian way. With either approach, the region is typically an ellipsoid and hence convex, which we assume hereinafter. See Appendix A.1 for some examples. The convexity allows finding the most violated constraint, for any fixed π_e, efficiently. The exploration policy can be obtained from

$$
\begin{aligned}
&\arg\min_{\pi_e\in\Pi'_\Delta} \bar{\boldsymbol{\phi}}_{\Delta\pi}^T(T\sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi} \qquad (11)\\
&s.t.\ V_{\boldsymbol{\theta}}(\pi_e) \geq (1-\epsilon)V_{\boldsymbol{\theta}}(\pi_0), \forall\boldsymbol{\theta}\in\mathcal{C}_\delta(\mathcal{D}_0).
\end{aligned}
$$

The constraint is equivalent to $\big(\pi_e - (1-\epsilon)\pi_0\big)^T\boldsymbol{X}\boldsymbol{\theta} \geq 0, \forall\boldsymbol{\theta}\in\mathcal{C}_\delta(\mathcal{D}_0)$, where $\boldsymbol{X}$ is the feature matrix obtained by stacking vectors in $\mathcal{A}$. Therefore, Problem (11) is convex. For a finite set of constraints, we adapt the popular Frank–Wolfe (FW) algorithm (Frank et al., 1956), which is a projection-free algorithm with good convergence guarantees (Jaggi, 2013). To handle the infinite number of constraints, we combine the FW algorithm with the cutting-plane method. See Appendix A.4 for details.

3.4. Policy evaluation

For concreteness, we choose the policy comparison problem to present our methodology. We would like to emphasize that all discussions are equally applicable to evaluating the value of a single policy, i.e., estimating $V(\pi_1)$. To see this, note that due to linearity, all discussions on the objective functions still hold, e.g., by replacing π_Δ in (1) with π_1 or $\bar{\phi}_{\Delta\pi}$ in (10) with $\bar{\phi}_{\pi_1}$. Moreover, the safety constraint is independent of the estimand. Therefore, the optimization problem can be formulated and solved in almost the same manner as for policy comparison.

3.5. Extensions

As we expect and also observed, the specific solution for `SEPEC` vary across different bandit setups, evaluation tasks and value estimators. To initiate the study of this novel area, we considered three representative variants covering MAB, contextual problems, generalization functions, IPW and DM. Our discussion, derivations and algorithms can be extended to at least the following problems.

First, for MAB, the DM estimator is equivalent to an alternative IPW-form estimator, which we analyze in Appendix B.1. Second, for stochastic contextual linear bandits where the stochastic context $\mathbf{x}_t$ and action A_t together generate a feature vector $\boldsymbol{\phi}_{\mathbf{x}_t,A_t}$ that determines the reward linearly,

our analysis for IPW and linear bandits can be combined and extended. See Appendix B.2 for details. Third, linear bandits with IPW are usually referred to as the pseudo-inverse estimator (Swaminathan et al., 2016), and its contextual version is particularly useful on some structured problems, such as slate recommendation. Our analysis in Section 3.3 can be extended to these problems. See Appendix B.3 for details. Lastly, another popular value estimator is the doubly robust (DR) estimator (Dudík et al., 2014), which combines IPW and DM. Discussions in Section 3.2 can be similarly applied to minimize the asymptotic variance when DR is used as the value estimator. See Appendix B.4 for details.

4. Theoretical Analysis

In this section, we provide theoretical analysis for SEPEC. Since our optimization problems can all be solved *exactly*, the safety constraint can be satisfied as promised. Therefore, we will focus on the efficiency maximization problem in Section 4.1. All proofs are deferred to Appendix D. The connection between our objective and the testing power is discussed in Section 4.2.

4.1. Efficiency

We first study MAB with IPW. The same conclusions can be established for CMAB under similar conditions. To provide insights, we first consider the case without side information and safety constraints to derive an explicit form of the overall variance. Denote $\mathrm{var}(\widehat{V}_{IPW}(\pi_1;\mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0;\mathcal{D}_e))$ as $\nu(\pi_e;\boldsymbol{r},\{\sigma_a\})$ when the true parameters are $\boldsymbol{r}$ and $\{\sigma_a\}$. Denote the solution of (4) as π_e^*.

Lemma 1. *The overall variance* $\nu(\pi_e^*;\boldsymbol{r},\{\sigma_a\})$ *is*

$$\frac{1}{T}\Big[\sum_{a\in[K]}|\pi_\Delta(a)|\times\sum_{a\in[K]}|\pi_\Delta(a)|(\sigma_a^2+r_a^2)-\big(V(\pi_1)-V(\pi_0)\big)^2\Big].$$

As expected, the variance decays at rate T^{-1}. The result is intuitive as it depends on how different the two policies are (i.e., $\sum_a|\pi_\Delta(a)|$) and if the difference is more significant on those more important arms (with larger σ_a^2 or r_a^2).

Next we study the efficiency. As mentioned in Section 3.1, the objective involves unknown parameters and there is no policy that always dominates (see Appendix D.4 for proof). Therefore, to provide insights, we study the minimax performance and the average performance. For the former, we consider the set of instance $\mathcal{I}=\{(\boldsymbol{r},\{\sigma_a\}):\mathbf{0}\le\boldsymbol{r}\le\mathbf{1},0\le\sigma_a\le\sigma,\forall a\}$. For the latter, we consider any instance distribution Q such that $\mathbb{E}_Q(r_a)$, $\mathrm{var}_Q(r_a)$ and $\mathbb{E}_Q(\sigma_a^2)$ all have fixed values across the K arms. For either Problem (4), (5) or (6), denote the solution (i.e., our policy) as π_e^* and the set of feasible policies as Π. Note that Π can be different with Π_Δ, since Π may be under safety constraints. SEPEC enjoys the following nice properties.

Theorem 1. *The policy* π_e^* *is minimax optimal, i.e.,*

$$\pi_e^*\in\arg\min_{\pi_e\in\Pi}\max_{(\boldsymbol{r},\{\sigma_a\})\in\mathcal{I}}\nu(\pi_e;\boldsymbol{r},\{\sigma_a\}).$$

The policy π_e^* *is also the most efficient on average, i.e.,*

$$\pi_e^*\in\arg\min_{\pi_e\in\Pi}\mathbb{E}_{(\boldsymbol{r},\{\sigma_a\})\sim Q}\,\nu(\pi_e;\boldsymbol{r},\{\sigma_a\}).$$

Finally, we study linear bandits with DM. For any policy π_e, with logged data $\mathcal{D}_0$, we denote the ultimate objective (10) as $\nu^*(\pi_e;\mathcal{D}_0)$ and the surrogate objective in (11) as $\nu(\pi_e;\mathcal{D}_0)$. We have the following promised result.

Theorem 2. *For any policy* $\pi_e\in\Pi'_\Delta$, *we have* $T\times|\nu(\pi_e;\mathcal{D}_0)-\nu^*(\pi_e;\mathcal{D}_0)|\to 0$ *as* T *grows. In other words, the surrogate objective* $\nu(\pi_e;\mathcal{D}_0)$ *is the asymptotic variance of* $\widehat{V}_{DM}(\pi_1;\mathcal{D}_0\cup\mathcal{D}_e)-\widehat{V}_{DM}(\pi_0;\mathcal{D}_0\cup\mathcal{D}_e)$.

In this sense, we regard our surrogate problem as a good proxy, which we solve *exactly*. Moreover, notice that ν and ν^* are both continuous. Therefore, suppose the optimum is in a compact set where the convergence is uniform, then we know $T\times|\nu^*(\pi_e^*;\mathcal{D}_0)-\min_{\pi_e}\nu^*(\pi_e;\mathcal{D}_0)|\to 0$, i.e., π_e^* is *asymptotically optimal*. We leave the finite-sample analysis for future research and investigate it experimentally here.

4.2. Connection with hypothesis testing

In real applications, one important task is to test whether or not the target policy π_1 is significantly better than π_0, which impacts deployment of π_1. Thus we investigate the relationship between the power of such a test and our objective $\mathrm{var}\big(\widehat{V}(\pi_1;\mathcal{D}_e)-\widehat{V}(\pi_0;\mathcal{D}_e)\big)$. For simplicity, we discuss MAB with IPW and without side information. A similar connection can be established for other setups considered in this paper (see Appendix A.2). Specifically, we consider the test

$$H_0: V(\pi_1)\le V(\pi_0)\quad\text{v.s.}\quad H_1: V(\pi_1)>V(\pi_0).$$

Denote our variance objective (1) by $\sigma(\pi_e)$ and let $\hat{\sigma}(\pi_e)$ be a consistent estimator with $\mathcal{D}_e$ (typically the plug-in estimator with sample variance). By the central-limit theorem and the Slutsky's lemma, we have

$$\hat{\sigma}(\pi_e)^{-1}\sqrt{T}\Big[\big(\widehat{V}(\pi_1;\mathcal{D}_e)-\widehat{V}(\pi_0;\mathcal{D}_e)\big)-\big(V(\pi_1)-V(\pi_0)\big)\Big]\xrightarrow{d}\mathcal{N}(0,1),$$

when T grows. Therefore, a popular (asymptotically) α-level test (Cai et al., 2020) is to reject the null when

$$\sqrt{T}\hat{\sigma}(\pi_e)^{-1}\big(\widehat{V}(\pi_1;\mathcal{D}_e)-\widehat{V}(\pi_0;\mathcal{D}_e)\big)\ge z_{1-\alpha},$$

where $z_{1-\alpha}$ is the upper αth quantile of the standard normal distribution. For any $\Delta>0$, the (asymptotic) power

under the local alternative $H_1 : V(\pi_1) - V(\pi_0) = \Delta/\sqrt{T}$ is hence $1 - \Phi\big(z_{1-\alpha} - \sigma(\pi_e)^{-1}\Delta\big)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard Gaussian. Since the other terms are all constant, it is clear that $\sigma(\pi_e)$ determines the power. Therefore, with a fixed budget, one can improve the power by designing a better policy π_e to reduce $\sigma(\pi_e)$, which corresponds to our objective.

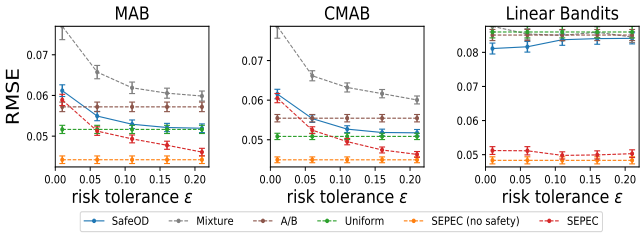
5. Experiments

In this section, we compare the empirical performance of various methods. We focus on three metrics: (i) the root mean square error (RMSE) of estimation, which quantifies the efficiency; (ii) the power of detecting $V(\pi_1) > V(\pi_0)$ when it holds, which quantifies the downstream impact; and (iii) the loss from safety violation $\max(0, T(1-\epsilon)\times V(\pi_0) - \sum_{t=1}^{T} \mathbb{E}(R_t|A_t))$, which quantifies the risk.

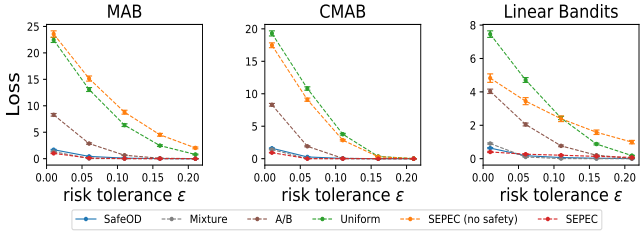
We compare our method SEPEC with several baselines. The first two are introduced in Section 2. To recap, the mixture policy $\pi_e = \epsilon\pi_1 + (1-\epsilon)\pi_0$ is a common way to run safe policy comparison, while the Safe Optimal Design (SafeOD) proposed in Zhu & Kveton (2022) ignores task-specific structure. Besides, we study the performance of *uniform sampling*, which is a common exploration policy; and *A/B test*, which allocates a half of the budget to π_1 and the rest to π_0, and is a common practice for policy comparison. Finally, we consider the variant of SEPEC that does not consider the safety constraint during optimization.

For MAB, we set $K = 10$, $T = 500$, $\delta = 0.05$, $\sigma_a \equiv 3$, sample $r_a \sim \text{Uniform}(0, 1)$, and vary the risk tolerance ϵ. We assume the noise is Gaussian, and construct $\mathcal{C}_\delta(\mathcal{D}_0)$ from the posteriors of $\{r_a\}$, with $|\mathcal{D}_0| = 100$ logged data points and the prior $\mathcal{N}(r_a, 0.2^2)$ for every arm a. To mimic the real applications where risk and opportunity coexist, we generate π_0 and π_1 in the following manner: we first sample from $\text{Ber}(0.5)$ to determine if the null $H_0 : V(\pi_1) \le V(\pi_0)$ is true. If true, we sample $\tilde{r}_a \sim \text{Uniform}(0, 1)$, set π_1 as $\pi_1(a) \propto \tilde{r}_a$, and then set π_0 as $\pi_0(a) \propto 0.5\tilde{r}_a + 0.5r_a$. If not, we set $\pi_0(a) \propto \tilde{r}_a$ and $\pi_1(a) \propto 0.5\tilde{r}_a + 0.5r_a$. The results are used to report the RMSEs and risks. We also run a test of level 5% to detect $V(\pi_1) > V(\pi_0)$, and report the power when H_1 is true. The configurations for CMAB and linear bandits are similar. For CMAB, we consider 30 contexts with $\{p(\mathbf{x}_i)\}$ sampled from $\text{Dirichlet}(\mathbf{1}_{30})$. For linear bandits, we sample $\boldsymbol{\theta}^*$ from the standard multivariate normal and $\mathbf{x}$ from the unit sphere uniformly, with $K = 100$, $d = 5$, $T = 200$, and $|\mathcal{D}_0| = 100$. A detailed description of experiments is in Appendix C.1.

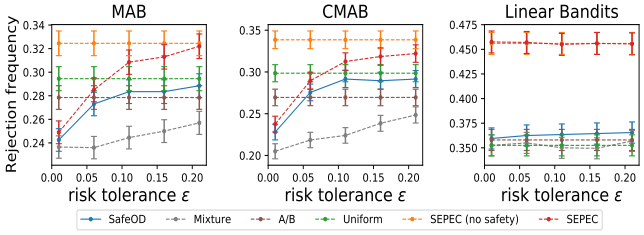
Results. Results aggregated over 2000 runs are reported in Figure 3. Overall, we observe that SEPEC consistently yields negligible loss from violating the safety constraint, and achieves lower RMSE and higher power compared with the other two safe policies (Mixture and SafeOD). The comparison with these two algorithms highlights the importance of carefully designing the exploration policy and utilizing task-specific structures. On the other hand, the A/B test and uniform exploration have slightly higher efficiency when the risk tolerance is very low, but are fairly risky. Moreover, compared with these two algorithms, the variant of SEPEC without safety constraints consistently shows better efficiency, which further supports the usefulness of a carefully designed exploration policy. As expected, when the safety constraint is less tight, the efficiency of SEPEC increases. The cost of satisfying the safety constraint is notably low in linear bandits, which is due to the generalization function.



(a) Estimation efficiency measured by RMSE.



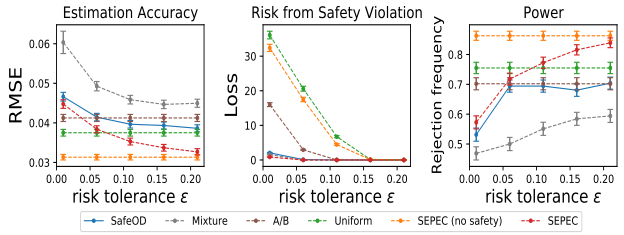
(b) Risk from safety violation.



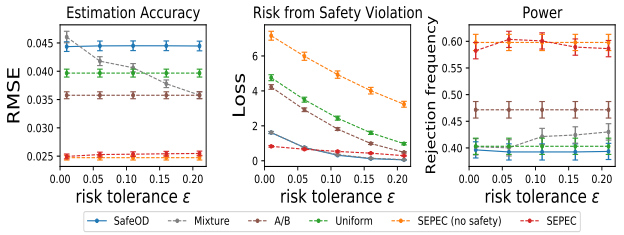
(c) Power of detecting $V(\pi_1) > V(\pi_0)$.

Figure 3: Experiment results. The error bars indicate the standard errors of the averages.

To investigate the robustness of the findings, we repeat the experiment under a wide range of parameters in Appendix C.2. In addition, to study the performance on real datasets, we conduct experiments using the MNIST dataset (Deng, 2012) and present the results in Figure 4, with more details given in Appendix C.3. The findings are consistent.



(a) CMAB with IPW.



(b) Linear bandits with DM.

Figure 4: Experiment results on the MNIST dataset. The errorbars indicate the standard errors of the averages. For methods that consider the safety constraint, the losses from safety violation are around zero and hence may overlap.

6. Related Work

Off-policy evaluation. The OPE literature can be roughly classified as the DM (Dudík et al., 2014), IPW (Horvitz & Thompson, 1952; Li et al., 2015), and doubly robust estimators (Dudík et al., 2014; Kallus et al., 2021; Su et al., 2020). Few of the existing papers pay attention to data quality. Oosterhuis & de Rijke (2020) studies the exploration problem in a very specific ranking task, and only IPW is considered with no theoretical guarantee provided. Tucker & Joachims (2022) is a parallel work that also studies data collection for OPE of bandit policies. The paper focuses on CMAB with IPW (and studies several extensions such as evaluating multiple policies), while we provide a more comprehensive study of various bandit setups and value estimators. Moreover, neither of the two papers studies the safety issue, which is practically important and introduces non-trivial challenges.

Conservative bandits. Safety constraints have been considered in the conservative bandits literature (Wu et al., 2016; Amani et al., 2019; Moradipari et al., 2020), where the key idea is to follow the baseline policy when the next taken action could be risky. This shares similar spirits with the mixture policy. Our objective of designing an exploration policy for OPE is orthogonal to this area which aims to minimize the cumulative regret.

Pure exploration. Similar to our work, the pure exploration literature (Xu et al., 2018; Degenne et al., 2019; Azizi et al., 2021) aims to maximize the information gain, with the objective of learning a good policy. Our work has a different objective in many aspects: we focus on the efficiency of policy evaluation, study non-adaptive policies, and further consider safety constraints.

Optimal experiment design. Our problem is related to optimal experiment design (Allen-Zhu et al., 2017; Baird et al., 2018; Fontaine et al., 2021), where the goal is also to design a high-quality data collection rule for statistical inference. Only a few of these works consider the bandit problem and none of them study OPE. We also consider the safety constraint, which is of great practical importance and makes the problem more challenging.

7. Conclusions

This paper initiates the study of efficient exploration for bandit policy evaluation and additionally considers safety. We call this problem SEPEC and study several of its representative variants. At a high level, our work contributes to disentangling exploration and evaluation/optimization in bandits. We believe that this direction is of huge practical importance, as our exploration strategy is easy to deploy, safe, and enables better statistical inference. Extensions to the Bayesian setup, sequential design, and reinforcement learning are all interesting next steps.

In Section 3.2, we assumed a finite context set for simplicity. A closer look at the derivations shows that, if there is no side information or we do not plan to use it (e.g., when we want worst-case safety guarantee), then this assumption is not required: every time when a new context is observed, we can just solve (8) once. All nice properties are retained. In fact, we empirically observed that the efficiency loss is minimal, unless the risk budget is very low. The same argument applies when one does not need the safety constraint. The action set can be context-dependent as well. If one would like to use side information to guarantee high-probability safety with infinite contexts, then $\pi_e(a|\mathbf{x})$ must be parameterized (e.g., as a neural network). As a trade-off, additional assumptions on the policy class are needed. We leave differentiable policies for future research. Finally, considering a finite context set covers applications where the context variables are discrete (e.g., day of the week, gender, age group, etc.).

We also assumed in Section 3.2 that the context distribution is known. This is a common assumption (Wu et al., 2015; Zhu & Kveton, 2022) that simplifies theory. It is usually reasonable because we typically have a large set of historical contexts. If not, one can either (i) optimize for every context independently as in (8), since possibly there is also not much side information; (ii) or replace $p(x)$ in the objective of (9) by an upper bound and that in the constraint by a convex confidence region, for which we can similarly solve with a high-probability safety guarantee and worst-case efficiency guarantee.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

Allen-Zhu, Z., Li, Y., Singh, A., and Wang, Y. Near-optimal design of experiments via regret minimization. In *International Conference on Machine Learning*, pp. 126–135. PMLR, 2017.

Amani, S., Alizadeh, M., and Thrampoulidis, C. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*, 2019.

Azizi, M., Kveton, B., and Ghavamzadeh, M. Fixed-budget best-arm identification in contextual bandits: A static-adaptive algorithm. *arXiv preprint arXiv:2106.04763*, 2021.

Baird, S., Bohren, J. A., McIntosh, C., and Özler, B. Optimal design of experiments in the presence of interference. *Review of Economics and Statistics*, 100(5):844–860, 2018.

Bertsimas, D. and Tsitsiklis, J. N. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Bouhtou, M., Gaubert, S., and Sagnol, G. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679–686, 2010.

Boyd, S. and Vandenberghe, L. Localization and cutting-plane methods. *From Stanford EE 364b lecture notes*, 2007.

Boyd, S., Boyd, S. P., and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

Brown, L. D. and Gajek, L. Information inequalities for the bayes risk. *The Annals of Statistics*, 18(4):1578–1594, 1990.

Brumback, B. A. A note on using the estimated versus the known propensity score to estimate the average treatment effect. *Statistics & Probability Letters*, 79(4):537–542, 2009.

Cai, H., Lu, W., and Song, R. On validation and planning of an optimal decision rule with application in healthcare studies. In *International Conference on Machine Learning*, pp. 1262–1270. PMLR, 2020.

Cai, H., Shi, C., Song, R., and Lu, W. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34, 2021.

Degenne, R., Koolen, W. M., and Ménard, P. Non-asymptotic pure exploration by solving games. *arXiv preprint arXiv:1906.10431*, 2019.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Fontaine, X., Perrault, P., Valko, M., and Perchet, V. Online a-optimal design and active linear regression. In *International Conference on Machine Learning*, pp. 3374–3383. PMLR, 2021.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2): 95–110, 1956.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952.

Hsu, D., Kakade, S. M., and Zhang, T. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 2011.

Jaggi, M. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pp. 427–435. PMLR, 2013.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Kallus, N., Saito, Y., and Uehara, M. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pp. 5247–5256. PMLR, 2021.

Kazerouni, A., Ghavamzadeh, M., Abbasi-Yadkori, Y., and Van Roy, B. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. Meta-thompson sampling. In *International Conference on Machine Learning*, pp. 5884–5893. PMLR, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, C. M.-S. Constrained optimal designs. *Journal of Statistical Planning and Inference*, 18(3):377–389, 1988.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pp. 608–616. PMLR, 2015.

Moradipari, A., Thrampoulidis, C., and Alizadeh, M. Stage-wise conservative linear bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Oosterhuis, H. and de Rijke, M. Taking the counterfactual online: Efficient and unbiased online evaluation for ranking. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pp. 137–144, 2020.

Sachdeva, N., Su, Y., and Joachims, T. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 965–975, 2020.

Shah, K. R. and Sinha, B. *Theory of optimal designs*, volume 54. Springer Science & Business Media, 2012.

Shi, C., Wan, R., Chernozhukov, V., and Song, R. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pp. 9580–9591. PMLR, 2021.

Slivkins, A. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudík, M. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pp. 9167–9176. PMLR, 2020.

Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudík, M., Langford, J., Jose, D., and Zitouni, I. Off-policy evaluation for slate recommendation. *arXiv preprint arXiv:1605.04812*, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388. PMLR, 2015.

Tran-The, H., Gupta, S., Nguyen-Tang, T., Rana, S., and Venkatesh, S. Combining online learning and offline learning for contextual bandits with deficient support. *arXiv preprint arXiv:2107.11533*, 2021.

Tsiatis, A. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

Tucker, A. D. and Joachims, T. Variance-optimal augmentation logging for counterfactual evaluation in contextual bandits. *arXiv preprint arXiv:2202.01721*, 2022.

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

Wan, R., Ge, L., and Song, R. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34, 2021.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

Wu, H., Srikant, R., Liu, X., and Jiang, C. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *arXiv preprint arXiv:1504.06937*, 2015.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262. PMLR, 2016.

Xu, L., Honda, J., and Sugiyama, M. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 843–851. PMLR, 2018.

Zanette, A., Dong, K., Lee, J. N., and Brunskill, E. Design of experiments for stochastic contextual linear bandits. *Advances in Neural Information Processing Systems*, 34: 22720–22731, 2021.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association*, 112(517):169–187, 2017.

Zhu, R. and Kveton, B. Safe optimal design with applications in off-policy learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 2436–2447, 2022.

A. Additional Details of the Proposed Methods

A.1. Safety constraint with side information

In this section, we give a few examples on how to construct the high-probability confidence (or credible) region $\mathcal{C}_\delta(\mathcal{D}_0)$ and show why it is convex in these cases. For any vector $\mathbf{v}$ and matrix V, we denote $||\mathbf{v}||_V = \mathbf{v}^T V^{-1}\mathbf{v}$.

For MAB, we can construct a confidence region by using an $(1-\delta/K)$-confidence interval for each arm separately. For example, suppose the noise is sub-Gaussian and there is at least one data point for each arm. We have $\mathbb{P}\big[|\hat{r}_a - r_a|/\sigma_a \geq \sqrt{2/n_a log(2/\delta)}\big] \leq \delta$, where n_a and $\hat{r}_a$ are the count and sample mean for arm a, respectively. Therefore,

$$\mathcal{C}_\delta(\mathcal{D}_0) = \{\boldsymbol{r} : max(\hat{r}_a - \sigma_a\sqrt{2/n_a log(2K/\delta)}, 0) \leq r_a \leq min(\hat{r}_a + \sigma_a\sqrt{2/n_a log(2K/\delta)}, 1), \forall a \in [K]\}$$

is a valid confidence region, based on the Bonferroni correction. We remark that, other choice is also possible than splitting the δ equally over arms.

Alternatively, we can consider the Bayesian perspective. We additionally assume the existence of a prior over r_a as $\mathcal{N}(\mu_{a,0}, \sigma^2_{a,0})$ for every arm a independently. By assuming that the reward from arm a follows $\mathcal{N}(0, \sigma_a^2)$, we can derive $r_a|\mathcal{D}_0 \sim \mathcal{N}\big(\hat{r}'_a, \hat{\sigma}^2_a\big)$, where $\hat{\sigma}^2_a = (1/\sigma^2_{a,0} + n_a/\sigma^2_a)^{-1}$ and $\hat{r}'_a = \hat{\sigma}^2_a(\mu_{a,0}/\sigma^2_{a,0} + n_a\hat{r}_a/\sigma^2_a)$. Therefore, one valid credible region is

$$\mathcal{C}_\delta(\mathcal{D}_0) = \{\boldsymbol{r} : max(\hat{r}'_a - z_{\delta/2K}\hat{\sigma}_a, 0) \leq r_a \leq min(\hat{r}'_a + z_{\delta/2K}\hat{\sigma}_a, 1), \forall a \in [K]\},$$

where we construct an interval for each arm separately. Alternatively, utilizing the fact that these Gaussian variables are independent, we can construct a joint region.

Notably, all examples above give us a convex set. The high-probability region in the CMAB case can be similarly derived.

For linear bandits, based on $\mathcal{D}_0$, with a given regularization parameter λ, we can construct a covariance matrix $\widehat{V}_0 = \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}_0}\mathbf{x}_i\mathbf{x}_i^T + \lambda\mathbf{I}$, an initial estimate $\hat{\boldsymbol{\theta}}_0 = \widehat{V}_0^-1(\sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}_0}\mathbf{x}_i R_i)$, and a corresponding high-confidence region for $\boldsymbol{\theta}^*$ as $\mathcal{C}_\delta(\mathcal{D}_0) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0||_{\widehat{V}_0} \leq S_\delta\}$. By carefully choosing S_δ as given in Theorem 2 of Abbasi-Yadkori et al. (2011), we can guarantee that $\mathbb{P}\big(\boldsymbol{\theta}^* \in \mathcal{C}_\delta(\mathcal{D}_0)\big) \geq 1-\delta$.

For the Bayesian perspective, we additionally assume the existence of a prior over $\boldsymbol{\theta}$ as $\mathcal{N}\big(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}\big)$, which can be learned either from domain knowledge or via meta-learning (Kveton et al., 2021; Wan et al., 2021). Assume that the noise follows $\mathcal{N}(0,\sigma^2)$. The standard Bayesian results on Gaussian linear regression tell us that

$$\boldsymbol{\theta}|\mathcal{D}_0 \sim \mathcal{N}\Big(\widehat{\Sigma}\big(\Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}_0}\mathbf{x}_i R_i/\sigma^2\big), \widehat{\Sigma}\Big), \widehat{\Sigma} = \big(\Sigma_{\boldsymbol{\theta}}^{-1} + \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}_0}\mathbf{x}_i\mathbf{x}_i^T/\sigma^2\big)^{-1}$$

Let $\hat{\boldsymbol{\theta}} = \widehat{\Sigma}\big(\Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\theta}} + \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}_0}\mathbf{x}_i R_i/\sigma^2\big)$. We can then define the ellipsoid as $\mathcal{C}_\delta(\mathcal{D}_0) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_{\widehat{V}} \leq \chi^2_d(1-\delta)\}$.

A.2. Connections with hypothesis testing

For linear bandits, the problem reduces to the random design analysis of linear models (Hsu et al., 2011). We assume homoscedasticity, i.e., $\mathrm{var}(\epsilon_t) \equiv \sigma^2$. We focus on that $\mathcal{D}_0$ is empty for simplicity, and it is straightforward to extend to the case with $\mathcal{D}_0$. By the central limit theorem and the Slutsky's lemma, we can establish that

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \sigma^2\big(\sum_{\mathbf{x}}\pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T\big)^{-1}\Big), T\to\infty.$$

which implies

$$\sqrt{T}\Big[\big(\widehat{V}_{DM}(\pi_1;\mathcal{D}) - \widehat{V}_{DM}(\pi_0;\mathcal{D})\big) - \big(V(\pi_1;\mathcal{D}) - V(\pi_0;\mathcal{D})\big)\Big] \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}, \sigma^2\bar{\boldsymbol{\phi}}_{\Delta\pi}^T\big(\sum_{\mathbf{x}}\pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T\big)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi}\Big), T\to\infty.$$

Therefore, by similar derivations as in the MAB case, we can design a valid testing procedure and identify that the testing power is determined by $\bar{\boldsymbol{\phi}}_{\Delta\pi}^T\big(\sum_{\mathbf{x}}\pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T\big)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi}$, i.e., our optimization objective.

Regarding CMAB with IPW, note the fact the the estimator, as in the MAB case, is also an U-statistic (i.e., the average of i.i.d samples). Therefore, we can similar derive its asymptotic distribution and relate the asymptotic variance (and hence the power) with our optimization objective. For example, let $\sigma(\pi_e) = \sqrt{\mathrm{var}(\pi_e(A_t|\mathbf{x}_t)^{-1}\pi_\Delta(A_t|\mathbf{x}_t)R_t)}$ and $\hat{\sigma}(\pi_e)$ be a

consistent estimator (typically the plug-in estimator with sample variance). By the central-limit theorem and the Slutsky's lemma, we again have

$$\frac{\sqrt{T}}{\hat{\sigma}(\pi_e)}\Big[\big(\widehat{V}(\pi_1;\mathcal{D}_e)-\widehat{V}(\pi_0;\mathcal{D}_e)\big)-\big(V(\pi_1)-V(\pi_0)\big)\Big] \xrightarrow{d} \mathcal{N}(0,1),$$

when T grows. The remaining discussions are exactly the same with the MAB case.

A.3. Optimization algorithm for MAB and CMAB with IPW

As we mentioned in the main text, at the first glance, it is challenging to solve (6) as it involves infinite number of linear constraints. Fortunately, both the objective and the feasible set are convex, and so we can adapt the cutting-plane method (Boyd & Vandenberghe, 2007) to solve.

As an iterative algorithm, there are two main steps in the cutting-plane method. At every iteration, we need to (i) proposed a *query point* $\pi_e^{(i)}$, and then (ii) check with an oracle on whether or not $\pi_e^{(i)}$ is in a target convex set (the feasible set in our case): if true, then we can terminate; otherwise, we find a plane to separate $\pi_e^{(i)}$ and the target convex set (i.e., the cutting plane).

The choices of $\pi_e^{(i)}$ and the cutting plane are both central to the computational efficiency. Following (Zhu & Kveton, 2022), we design the algorithm as in Algorithm 1, where both are obtained by a solving a simple subproblem. Specifically, the one for $\pi_e^{(i)}$ has a convex objective function and a set of linear constraints, and the one for $\boldsymbol{r}^{(i)}$ is a linear programming problem. The optimization algorithm for the CMAB case can be designed similarly. The convergence analysis is provided in Boyd & Vandenberghe (2007).

Algorithm 1: Safe MAB exploration with the cutting-plane method

Data: π_0, π_1 , $\mathcal{C}_\delta(\mathcal{D}_0)$, ϵ
Set $i=0$, $\pi_e^{(0)}=\pi_0$ and $\Theta=\varnothing$.
while *Not converge* **do**
  i = i + 1
  Apply convex optimization algorithms to obtain $\pi_e^{(i)}$ by solving

$$\begin{aligned}&\arg\min_{\pi_e\in\Pi_\Delta}\sum_{a\in[K]}\frac{\pi_\Delta^2(a)}{\pi_e(a)}\\ &s.t.\ \ \boldsymbol{r}^T(\pi_e-(1-\epsilon)\pi_0)\geq 0, \forall \boldsymbol{r}\in\Theta.\end{aligned}$$

  if $min_{\boldsymbol{r}\in\mathcal{C}_\delta(\mathcal{D}_0)}\boldsymbol{r}^T(\pi_e^{(i)}-(1-\epsilon)\pi_0)\geq 0$ **then**
    Terminate the iteration
  else
    Solve $\boldsymbol{r}^{(i)}=\arg\min_{\boldsymbol{r}\in\mathcal{C}_\delta(\mathcal{D}_0)}\boldsymbol{r}^T(\pi_e^{(i)}-(1-\epsilon)\pi_0)$; `// Constraint is maximally violated`
    Update $\Theta=\Theta\cup\{\boldsymbol{r}^{(i)}\}$
  end
end
Output: $\pi_e=\pi_e^{(i)}$

A.4. Optimization algorithm for linear bandits with DM

We first introduce how to solve for a finite set of constraints Θ:

$$\begin{aligned}&\arg\min_{\pi_e\in\Delta_{K-1}}\ \mathcal{J}(\pi_e)=\bar{\boldsymbol{\phi}}_{\Delta\pi}^T(T\sum_{\mathbf{x}\in\mathcal{A}}\pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T+\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^+\bar{\boldsymbol{\phi}}_{\Delta\pi}\\ &s.t.\ \max_{\boldsymbol{\theta}\in\Theta}\big((1-\epsilon)\pi_0-\pi_e\big)^T\boldsymbol{X}\boldsymbol{\theta}\leq 0.\end{aligned}$$

We propose to adapt the Frank–Wolfe (FW) algorithm (Frank et al., 1956), which is a projection-free algorithm for convex optimization and it enjoys nice convergence property (Frank et al., 1956; Lattimore & Szepesvári, 2020; Jaggi, 2013). We

summarize in Algorithm 2. For Step 1, we note that the partial derivative can be derived as

$$\frac{\partial \mathcal{J}(\pi_e)}{\partial \pi_e(\mathbf{x}_i)} = -T\Big\{\bar{\boldsymbol{\phi}}_{\Delta\pi}^T(T\boldsymbol{G}(\pi_e) + \boldsymbol{G}_0)^{-1}\mathbf{x}_i\Big\}^2$$

where $\boldsymbol{G}_0 = \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0$ and $\boldsymbol{G}(\pi_e) = \sum_{\mathbf{x}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T$. For Step 2, we notice that it is a linear programming problem.

Algorithm 2: Optimization for linear bandits with the FW algorithm

Data: π_0 , $\mathcal{C}_\delta(\mathcal{D}_0)$, ϵ , $\boldsymbol{X}$, $\boldsymbol{\Phi}_0$, $\bar{\boldsymbol{\phi}}_{\Delta\pi}$
Set $i = 0$, $\pi_e^{(0)} = \pi_0$ and $\Theta = \varnothing$.
while *TRUE* **do**
i = i + 1
1. Compute the gradient $d^{(i)} = \partial\mathcal{J}(\pi_e)/\partial\pi_e|_{\pi_e=\pi_e^{(i)}}$
2. Compute a feasible direction as $\Delta\pi_e$ by solving

$$\begin{aligned}&\arg\min_{w\in\Delta_{K-1}} \; d^{(i)}w\\ &s.t.\ \max_{\boldsymbol{\theta}\in\Theta}\big((1-\epsilon)\pi_0 - w\big)^T\boldsymbol{X}\boldsymbol{\theta} \le 0.\end{aligned}$$

3. Run line search to find the optimal step size $\lambda^{(i)}$ by solving $\arg\min_{\lambda\in[0,1]} \mathcal{J}\Big(\pi_e^{(i)} + \lambda(\Delta\pi_e - \pi_e^{(i)})\Big)$
4. Update $\pi_e^{(i+1)} = \pi_e^{(i)} + \lambda^{(i)}(\Delta\pi_e - \pi_e^{(i)})$
end
Output: $\pi_e = \pi_e^{(i)}$

Algorithm 3: Safe linear bandits exploration with the FW algorithm and the cutting-plane method

Data: π_0 , $\mathcal{C}_\delta(\mathcal{D}_0)$, ϵ , $\boldsymbol{X}$, $\boldsymbol{\Phi}_0$, $\bar{\boldsymbol{\phi}}_{\Delta\pi}$
Set $i = 0$, $\pi_e^{(0)} = \pi_0$ and $\Theta = \varnothing$.
while *TRUE* **do**
i = i + 1
Run the FW algorithm to obtain $\pi_e^{(i)}$ by solving

$$\begin{aligned}&\arg\min_{\pi_e\in\Delta_{K-1}} \; \mathcal{J}(\pi_e) = \bar{\boldsymbol{\phi}}_{\Delta\pi}^T(T\sum_{\mathbf{x}}\pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi}\\ &s.t.\ \big((1-\epsilon)\pi_0 - \pi_e\big)^T\boldsymbol{X}\boldsymbol{\theta} \le 0, \forall\boldsymbol{\theta}\in\Theta\end{aligned}$$

if $max_{\boldsymbol{\theta}\in C_\delta}\big((1-\epsilon)\pi_0 - \pi_e^{(i)}\big)^T\boldsymbol{X}\boldsymbol{\theta} \le 0$ **then**
Terminate the iteration
else
Solve $\boldsymbol{\theta}^{(i)} = \arg\max_{\boldsymbol{\theta}\in C_\delta}\big((1-\epsilon)\pi_0 - \pi_e^{(i)}\big)^T\boldsymbol{X}\boldsymbol{\theta}$; `// Constraint is maximally violated`
Update $\Theta = \Theta \cup \{\boldsymbol{\theta}^{(i)}\}$
end
end
Output: $\pi_e = \pi_e^{(i)}$

Next, we integrate the cutting-plane method and the FW algorithm to solve the most challenging problem (11), where there exists a infinite number of constraints. Refer to Appendix A.3 for an introduction to the cutting-plane method. We note that, since the cutting-plane method only requires a feasible query point in every iteration instead of running the whole FW algorithm every time, one can stop the FW after a few iterations. This does not affect the convergence over the whole algorithm.

In Algorithm 3, a key step is to find a good cutting plane, or equivalently, $\boldsymbol{\theta} \in C_\delta$ on which the constraint is maximally violated by the current policy $\pi^{(i)}$. Assume $\mathcal{C}_\delta(\mathcal{D}_0) = \{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_0||_{\widehat{V}_0} \le S_\delta\}$ for some $\hat{\boldsymbol{\theta}}_0$ and $\widehat{V}_0$. Let $\mathbf{v}_{\pi^{(i)}} =$

$\Big[\big((1-\epsilon)\pi_0 - \pi^{(i)}\big)^T \boldsymbol{X}\Big]^T$. We need to maximize a linear function on an ellipsoid. Fortunately, the solution can be obtained explicitly (Zhu & Kveton, 2022) as

$$\boldsymbol{\theta}^{(i)} = \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\mathcal{C}_\delta(\mathcal{D}_0)} \Big[\big((1-\epsilon)\pi_0 - \pi^{(i)}\big)^T \boldsymbol{X}\Big]\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0 + \frac{\widehat{V}_0 \mathbf{v}_{\pi^{(i)}}}{\sqrt{\mathbf{v}_{\pi^{(i)}}^T \widehat{V}_0 \mathbf{v}_{\pi^{(i)}}/S_\delta}}.$$

B. Extensions

B.1. MAB with DM

An alternative estimator for MAB is the direct method (DM). In MAB, DM essentially plugs-in the estimated mean of every arm as $\widehat{V}(\pi;\mathcal{D}) = \sum_a \Big[\pi(a)\frac{\sum_{(A_t,R_t)\in\mathcal{D}} R_t\mathbb{I}(A_t=a)}{\sum_{(A_t,R_t)\in\mathcal{D}}\mathbb{I}(A_t=a)}\Big]$. We note this estimator is actually equivalent to $|\mathcal{D}|^{-1}\sum_{(A_i,R_i)\in\mathcal{D}}\frac{\pi(A_i)}{\hat{\pi}_e^*(A_i)}R_i$, where $\hat{\pi}_e^*(a) = |\mathcal{D}|^{-1}\sum_{(A_i,R_i)\in\mathcal{D}}\mathbb{I}(A_t=a)$, i.e., the IPW-form estimator with the estimated propensity $\hat{\pi}_e^*(a)$. A well-known but perhaps counter-intuitive result (Brumback, 2009) is that, even if $\mathcal{D}$ is collected by following π_e and π_e is known, $\widehat{V}_{IPW}(\pi;\mathcal{D})$ is still more efficient than plugging-in the known function π_e.

The following result is standard (Li et al., 2015) and we just recap here for completeness:

$$\begin{aligned}
&\widehat{V}(\pi;\mathcal{D}) - V(\pi)\\
&= \sum_a \Big[\pi(a)\frac{\sum_{(A_t,R_t)\in\mathcal{D}} R_t\mathbb{I}(A_t=a)}{\sum_{(A_t,R_t)\in\mathcal{D}}\mathbb{I}(A_t=a)}\Big] - V(\pi)\\
&= \sum_a \Big[\pi(a)\big(\frac{\sum_{(A_t,R_t)\in\mathcal{D}} R_t\mathbb{I}(A_t=a)}{\sum_{(A_t,R_t)\in\mathcal{D}}\mathbb{I}(A_t=a)} - r_a\big)\Big]\\
&= \sum_a \Big[\pi(a)\big(\frac{\sum_{(A_t,R_t)\in\mathcal{D}} (R_t-r_a)\mathbb{I}(A_t=a)}{\sum_{(A_t,R_t)\in\mathcal{D}}\mathbb{I}(A_t=a)}\big)\Big]\\
&= \sum_a \Big[\pi(a)\big(\frac{|\mathcal{D}|^{-1}\sum_{(A_t,R_t)\in\mathcal{D}} (R_t-r_a)\mathbb{I}(A_t=a)}{|\mathcal{D}|^{-1}\sum_{(A_t,R_t)\in\mathcal{D}}\mathbb{I}(A_t=a)}\big)\Big]\\
&\xrightarrow{d} \mathcal{N}\Big(0, \sum_a \frac{\pi^2(a)}{\pi_e(a)}\sigma_a^2\Big).
\end{aligned}$$

By the additivity, we can see this relationship holds for the value difference $V(\pi_1) - V(\pi_0)$ as well. All of our discussions on the safety constraints and optimization algorithms can be directly extended to this case. Finally, similar extensions can be constructed for CMAB as well.

B.2. Stochastic contextual linear bandits

Next, we consider the contextual bandits setting with a linear generalization function, i.e., the stochastic contextual linear bandits problem. At every round t, a stochastic context $\mathbf{x}_t$ will be sampled i.i.d. and can not be known *a priori*. The agent will then choose an arm A_t, which together with $\mathbf{x}_t$ gives us the transformed feature vector $\boldsymbol{\phi}_t = \boldsymbol{\phi}_{\mathbf{x}_t,A_t}$. Let $\bar{\boldsymbol{\phi}}_\pi = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),a\sim\pi(a|\mathbf{x})}\boldsymbol{\phi}_{\mathbf{x},a}$. Assume $p(\mathbf{x})$ is known and hence $\bar{\boldsymbol{\phi}}_\pi$ is also known. Given an estimate $\hat{\boldsymbol{\theta}}$ obtained via least-square regression, recall our value estimator via DM is $\widehat{V}_{\hat{\boldsymbol{\theta}}}(\pi) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),a\sim\pi(a|\mathbf{x})}\boldsymbol{\phi}_{\mathbf{x},a}^T\hat{\boldsymbol{\theta}} = \big[\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}),a\sim\pi(a|\mathbf{x})}\boldsymbol{\phi}_{\mathbf{x},a}^T\big]\hat{\boldsymbol{\theta}} = \bar{\boldsymbol{\phi}}_\pi^T\hat{\boldsymbol{\theta}}$. Let $\bar{\boldsymbol{\phi}}_{\Delta\pi} = \bar{\boldsymbol{\phi}}_{\pi_1} - \bar{\boldsymbol{\phi}}_{\pi_0}$. W.l.o.g., we assume the noise variance is upper bounded by 1. Without other information or constraints, by similar arguments as in Section 3.3, our objective can be transformed as follows

$$\begin{aligned}
&\operatorname{var}(\widehat{V}(\pi_1;\mathcal{D}_e\cup\mathcal{D}_0) - \widehat{V}(\pi_0;\mathcal{D}_e\cup\mathcal{D}_0))\\
&= \bar{\boldsymbol{\phi}}_{\Delta\pi}^T cov(\hat{\boldsymbol{\theta}})\bar{\boldsymbol{\phi}}_{\Delta\pi}\\
&= \bar{\boldsymbol{\phi}}_{\Delta\pi}^T\Big\{\mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e,\mathbf{x}}\Big[cov(\hat{\boldsymbol{\theta}}\mid\boldsymbol{\Phi})\Big] + cov_{\boldsymbol{\Phi}\sim\pi_e,\mathbf{x}}\Big[\mathbb{E}(\hat{\boldsymbol{\theta}}\mid\boldsymbol{\Phi})\Big]\Big\}\bar{\boldsymbol{\phi}}_{\Delta\pi}\\
&= \bar{\boldsymbol{\phi}}_{\Delta\pi}^T\Big\{\mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e,\mathbf{x}}\Big[(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\Big]\Big\}\bar{\boldsymbol{\phi}}_{\Delta\pi}
\end{aligned} \tag{12}$$

Here, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}_0$ is the feature matrix stacked over points in $\mathcal{D}$ and $\mathcal{D}_0$, respectively. More specifically, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_{\mathbf{x}_1,A_1},\ldots,\boldsymbol{\phi}_{\mathbf{x}_T,A_T})^T$. We use $\boldsymbol{\Phi}\sim\pi_e,\mathbf{x}$ to emphasize the expectation is taken over $\boldsymbol{\Phi}$, which depends on both the

policy π_e and the stochastic context $\mathbf{x}$.

Besides, the safety constraint also needs to be satisfied. Therefore, our problem can be written as

$$\begin{aligned} &\arg\min_{\pi_e} \bar{\boldsymbol{\phi}}_{\Delta\pi}^T \Big\{ \mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e,\mathbf{x}} \Big[(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1} \Big] \Big\} \bar{\boldsymbol{\phi}}_{\Delta\pi} \\ &s.t.\ \big[\bar{\boldsymbol{\phi}}_{\pi_e^*}^T - (1-\epsilon)\bar{\boldsymbol{\phi}}_{\pi_0}^T\big]\boldsymbol{\theta} \geq 0, \forall \boldsymbol{\theta} \in \mathcal{C}_\delta(\mathcal{D}_0), \\ &\quad \pi_e(\cdot \mid \mathbf{x}) \in \Delta_{K-1}, \forall \mathbf{x} \in \mathcal{X} \end{aligned} \tag{13}$$

The challenges on optimization include that (i) both the objective and $\bar{\boldsymbol{\phi}}_{\pi_e}$ in the constraint can have complex dependency on π_e, (ii) the constraint implicitly includes infinite linear constraints, and (iii) the stochasticity in the context is hard to handle, unlike in Section 3.3.

When there is a finite number of contexts, following similar arguments as in Section 3.3, we can consider the following surrogate objective instead, which can be similarly proved as the asymptotic variance of our procedure:

$$\begin{aligned} &\arg\min_{\pi_e} \bar{\boldsymbol{\phi}}_{\Delta\pi}^T \Big\{ \Big[(T\sum_{\mathbf{x},a} p(\mathbf{x})\pi_e(a|\mathbf{x})\boldsymbol{\phi}_{\mathbf{x},a}\boldsymbol{\phi}_{\mathbf{x},a}^T + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1} \Big] \Big\} \bar{\boldsymbol{\phi}}_{\Delta\pi} \\ &s.t. \min_{\boldsymbol{\theta}\in\mathcal{C}_\delta(\mathcal{D}_0)} \quad \big[\sum_{\mathbf{x},a} p(\mathbf{x})\pi_e(a|\mathbf{x})\boldsymbol{\phi}_{\mathbf{x},a} - (1-\epsilon)\bar{\boldsymbol{\phi}}_{\pi_0}\big]^T\boldsymbol{\theta} \geq 0. \end{aligned} \tag{14}$$

The problem can be solved similarly as in (11) by utilizing the convexity. See Appendix A.4 for details.

B.3. Pseudo-inverse estimator for linear bandits and its contextual version

Under the linear generalization assumption, besides the DM estimator, we can also consider the Pseudo-Inverse (PI) estimator (Swaminathan et al., 2016). Instead of inferring $\boldsymbol{\theta}$, PI directly constructs a weighted average of the observed rewards as in IPW, but it also utilizes the linear structure. PI is particular useful when being applied to some structured problem such as slate recommendation (Swaminathan et al., 2016). Let $\boldsymbol{G}(\pi) = \sum_{\mathbf{x}} \pi(\mathbf{x})\mathbf{x}\mathbf{x}^T$. We first note that

$$V(\pi;\mathcal{D}) = \bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{\theta} = \bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1}\sum_{\mathbf{x}} \pi(\mathbf{x})\mathbf{x}(\mathbf{x}^T\boldsymbol{\theta})$$

Motivated by this relationship, the PI estimator is constructed as

$$\widehat{V}_{PI}(\pi;\mathcal{D}) = |\mathcal{D}|^{-1} \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}} \bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1}(R_i\mathbf{x}_i) = |\mathcal{D}|^{-1}\bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1} \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}} R_i\mathbf{x}_i \tag{15}$$

Let $\boldsymbol{\Phi}$ be the feature matrix stacked over points in $\mathcal{D}$. We first derive the finite-sample conditional variance $\mathrm{var}(\widehat{V}_{PI}(\pi_1;\mathcal{D})|\boldsymbol{\Phi})$ as

$$\mathrm{var}(\widehat{V}_{PI}(\pi_1;\mathcal{D})|\boldsymbol{\Phi}) = |\mathcal{D}|^{-2} \sum_{(\mathbf{x}_i,R_i)\in\mathcal{D}} \Big[\bar{\boldsymbol{\phi}}_{\pi_1}^T\boldsymbol{G}(\pi_1)^{-1}\mathbf{x}_i\Big]^2 \mathrm{var}(R_i \mid \mathbf{x}_i)$$

For simplicity of notations, w.l.o.g., we assume the noise variance is 1. By the law of total variance, we can obtain that

$$\begin{aligned} &T \times \mathrm{var}(\widehat{V}_{PI}(\pi_1;\mathcal{D}_e)) \\ &= T \times \mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e} \mathrm{var}(\widehat{V}_{PI}(\pi_1;\mathcal{D}_e)|\boldsymbol{\Phi}) + T \times \mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e} \mathbb{E}(\widehat{V}_{PI}(\pi_1;\mathcal{D}_e)|\boldsymbol{\Phi}) \\ &= \mathbb{E}_{\mathbf{x}_i\sim\pi_e}\Big[\bar{\boldsymbol{\phi}}_{\pi_1}^T\boldsymbol{G}(\pi_1)^{-1}\mathbf{x}_i\Big]^2 + T \times \mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e}\Big\{T^{-1}\bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\theta}\Big\} \\ &\to \bar{\boldsymbol{\phi}}_{\pi_1}^T\boldsymbol{G}(\pi_1)^{-1}\boldsymbol{G}(\pi_e)\boldsymbol{G}(\pi_1)^{-1}\bar{\boldsymbol{\phi}}_{\pi_1} + \mathrm{var}_{\mathbf{x}\sim\pi_e}\Big\{\bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1}\mathbf{x}\mathbf{x}^T\boldsymbol{\theta}\Big\}. \end{aligned} \tag{16}$$

The proof for the asymptotic variance statement (i.e., the last row) is by similar arguments as in Appendix D.3.

By similar arguments as in the IPW for MAB case, the raw objective (16) is infeasible to solve, as it involves the unknown parameter $\boldsymbol{\theta}$. Therefore, we relax with the upper bound of the unknown parameters and focus on minimizing the *intrinsic uncertainty* term from the reward noise, i.e., $\bar{\boldsymbol{\phi}}_{\pi_1}^T\boldsymbol{G}(\pi_1)^{-1}\boldsymbol{G}(\pi_e)\boldsymbol{G}(\pi_1)^{-1}\bar{\boldsymbol{\phi}}_{\pi_1}$. Notice that, under the linear generalization assumption, all discussions in Section 3.3 regarding the safety constraint still apply here. Therefore, this problem can be similarly solved as in Section 3.3, via combining the cutting-plane method with the FW algorithm.

Besides, by noting that (15) is linear in $\bar{\boldsymbol{\phi}}_\pi^T\boldsymbol{G}(\pi)^{-1}$, it is straightforward to extend to studying the value difference $V(\pi_1) - V(\pi_0)$. Moreover, as in Swaminathan et al. (2016) and Zhu & Kveton (2022), the PI estimator and the discussions above can be extended to the contextual setup in a straightforward manner.

B.4. Doubly robust estimator for MAB and CMAB

Besides IPW and DM, the doubly robust (DR) estimator is another popular OPE estimator, both in bandits (Dudík et al., 2014) and in reinforcement learning (Shi et al., 2021). We analyze the DR estimator in the MAB setup, and the CMAB case is similar. Let $\widehat{V}_{DM}$ be a consistent DM-type estimator. The DR value estimator (Tsiatis, 2007) is define as

$$\widehat{V}_{DR}(\pi_1;\mathcal{D}_e) = \sum_t \frac{\pi_1(A_t)}{\pi_e(A_t)}(R_t - \widehat{V}_{DM}(\pi_1;\mathcal{D}_e)) + \widehat{V}_{DM}(\pi_1;\mathcal{D}_e).$$

Due to the complex structure, it is well-known that its finite-sample variance does not yield a tractable form as IPW and DM do, and therefore people commonly focus on its asymptotic variance (Tsiatis, 2007)

$$\begin{aligned}
&T \times \text{AsmypVar}\Big(\widehat{V}_{DR}(\pi_1;\mathcal{D}_e))\Big) \\
&= \sum_{a\in[K]} \frac{\pi_1^2(a)}{\pi_e(a)}\sigma_a^2 + \text{var}_{A_t\sim\pi_e}\Big[\frac{\pi_1(A_t)}{\pi_e(A_t)}\big(r_{A_t} - V(\pi_1)\big)\Big] \\
&= \sum_{a\in[K]} \frac{\pi_1^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_1^2(a)}{\pi_e(a)}\big(r_a - V(\pi_1)\big)^2 - 0.
\end{aligned}$$

Compared with the form of IPW, one can find that the only difference is that the r_a^2 in the second term is replaced by $\big(r_a - V(\pi_1)\big)^2$, which is also related with the unknown parameters. Therefore, a tractable optimization objective can be formed by considering the upper bounds as

$$\arg\min_{\pi_e} \sum_{a\in[K]} \frac{\pi_1^2(a)}{\pi_e(a)}.$$

The other discussions on the safety constraint and the optimization for IPW can then be directly applied here. Similar results can be established for $V(\pi_1) - V(\pi_0)$, by noting the additivity.

B.5. Alternative objective function with side information for IPW

For MAB with IPW (similar arguments below apply for CMAB), recall that we derived the following form for the overall variance:

$$T \times \text{var}\Big(\widehat{V}_{IPW}(\pi_1;\mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0;\mathcal{D}_e)\Big) = \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}r_a^2 - \big(V(\pi_1) - V(\pi_0)\big)^2,$$

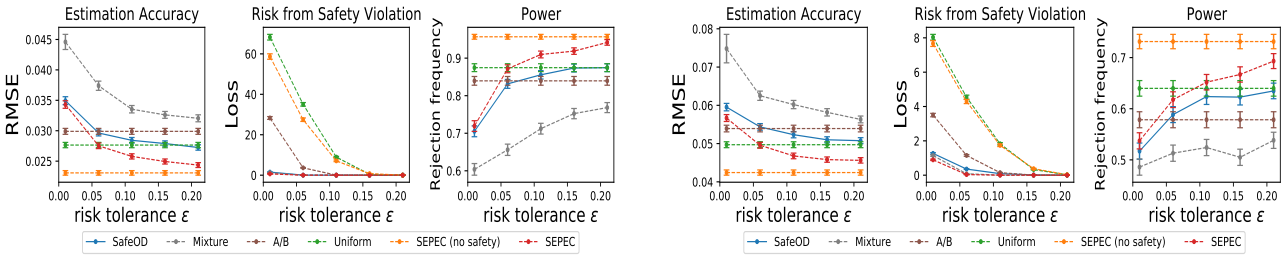
Since either σ_a, r_a, or $V(\pi_1) - V(\pi_0)$ is known, this objective is intractable to directly optimize. From the worst-case point of view, we relax them using the upper bounds and optimize the surrogate objective

$$\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}.$$

The advantage of this surrogate objective is that, for any instances, we can provide decent performance guarantee. However, the downside is such an objective might be conservative. In particular, even with side information (dataset $\mathcal{D}_0$ or posterior distribution Q over instance) available, our utilization of this information is limited (since the corresponding guarantee would be high-probability or Bayesian).

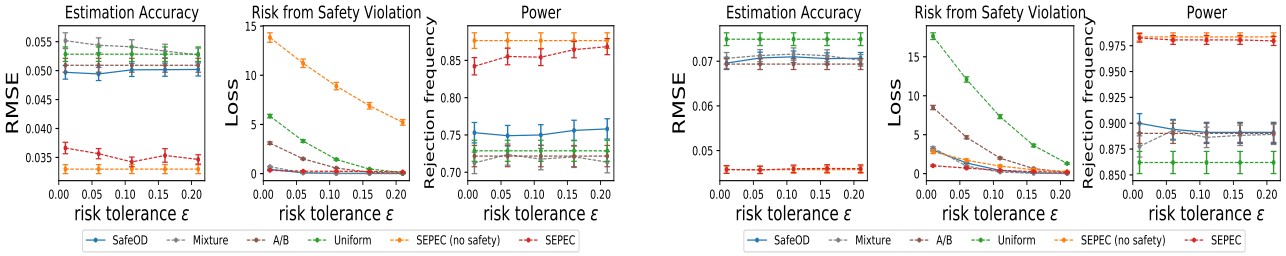
We discuss alternative surrogate optimization objectives in this section. With the posterior distribution Q, it is easy to see that the objective, which is known as the Bayes risk (Brown & Gajek, 1990), can be formulated as

$$\begin{aligned}
&\mathbb{E}_Q\Big\{ \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}r_a^2 - \big(V(\pi_1) - V(\pi_0)\big)^2\Big\} \\
&= \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\mathbb{E}_Q\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\mathbb{E}_Q r_a^2 - \mathbb{E}_Q\big(V(\pi_1) - V(\pi_0)\big)^2 \\
&= \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}(\mathbb{E}_Q\sigma_a^2 + \mathbb{E}_Q r_a^2) - \mathbb{E}_Q\big(V(\pi_1) - V(\pi_0)\big)^2,
\end{aligned}$$



(a) $T = 2000, K = 5, \sigma = 3$. **(b)** $T = 200, K = 20, \sigma = 2$.

Figure 5: Experiment results for CMAB under other combinations of setting parameters.



(a) $T = 100, K = 100, \sigma = 2, d = 10$. **(b)** $T = 200, K = 100, \sigma = 2, d = 3$.

Figure 6: Experiment results for LB under other combinations of setting parameters.

which shares similar form with the one in the main text. Therefore, all discussions can be extended.

Regarding the frequentist way, we can utilize $\mathcal{D}_0$ to construct upper confidence bounds for the unknown terms, and the final performance guarantee would be a high-probability statement.

C. Additional Experiment Details and Results

C.1. Additional experiment details

In this section, we introduce more details of the experiment setup used in our simulation study. The configuration for CMAB is almost the same with that for MAB introduced in the main text, except for that we generate for every context independently. For linear bandits, we following Kazerouni et al. (2016) to make sure $\mathbf{x}_i^T \boldsymbol{\theta}^*$ is positive for all arms when sampling the feature vectors. To generate the policies π_0 and π_1, we follow a similar design with MAB, except for that we first sample $\tilde{\boldsymbol{\theta}}$ from the multivariate normal and then generate $\pi_1(a)$ (or $\pi_0(a)$) proportional to $\mathbf{x}_a^T \tilde{\boldsymbol{\theta}}$.

C.2. Additional simulation experiments

In this section, we repeat the simulation experiments in the main text under a variety of parameter combinations to study the robustness of our findings. See Figure 5 for results on CMAB and Figure 6 for results on LB. Overall, the main findings are consistent with those in the main text.

C.3. Experiments on the MNIST dataset

In this section, we conduct two experiments on the MNIST dataset (Deng, 2012). In this dataset, every arm is an (vectorized) image of a digit between 0 and 9,

CMAB. The first experiment focuses on contextual bandits. As standard in the literature (Dudík et al., 2014), we adapt a classification task to a CMAB problem. Specifically, in every trail, we randomly pick 100 images as our contexts, with their probability sampled from Dirichlet($\mathbf{1}_{100}$). The $K = 10$ arms correspond to the 10 digits, and only the arm corresponding to the one on the image returns reward 1. All the other arms return reward 0. The two policies π_0 and π_1 are two classifier trained with 1000 randomly sampled data points, one using the decision tree and the other using multi-class

logistic regression. We define the one with higher accuracy as π_1 when H_1 is true, and as π_0 when H_0 is true. We aim to collect $T = 1000$ data points.

This setup aims to mimic a common application: we have two policies trained using logged data and with different functional form assumptions, so a direct comparison is unfair and may suffer from model misspecification. Therefore, one typically utilizes IPW (which is guaranteed to be unbiased under minimal assumptions) to compare these two policies.

Linear bandits. In the second experiment, we closely follow Zhu & Kveton (2022) to study the performance on linear bandits. In every trail, we randomly pick one digit as the correct answer, and we assign reward 1 to the images corresponding to this correct digit. $K = 100$ arms are randomly chosen. We set $\sigma = 1$, $|\mathcal{D}_0| = 100$ and $T = 100$. $\boldsymbol{\theta}^*$ is fitted from the whole dataset.

Results. Results aggregated over 2000 random seeds are presented in Figure 4. The superior performance of SEPEC and other findings are largely consistent with our simulation experiments. In particularly, for linear bandits, the cost of satisfying the safety constraint is negligible, which is consistent with the findings in Zhu & Kveton (2022). Besides, regarding the power under linear bandits, we observe that SEPEC is even slightly better than its non-safe version, though the difference is not statistically significant. A closer look into the intermediate results tell that this is mainly due to that the FW is an iterative algorithm and the convergence situation might slightly vary due to numerical reasons.

D. Proofs

D.1. Proof of Lemma 1

Proof. First, recall that

$$\begin{aligned} T \times \nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) &= T \times \mathrm{var}\Big(\widehat{V}_{IPW}(\pi_1; \mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0; \mathcal{D}_e)\Big) \\ &= \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} r_a^2 - \Big(V(\pi_1) - V(\pi_0)\Big)^2. \end{aligned}$$

Let $W = \sum_a |\pi_\Delta(a)|$. We have $\pi_e^*(a) = |\pi_\Delta(a)|/W$. Therefore

$$\begin{aligned} &T \times \nu(\pi_e^*; \boldsymbol{r}, \{\sigma_a\}) \\ &= \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{|\pi_\Delta(a)|/W}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{|\pi_\Delta(a)|/W} r_a^2 - c \\ &= \sum_{a\in[K]} |\pi_\Delta(a)| W(\sigma_a^2 + r_a^2) - \Big(V(\pi_1) - V(\pi_0)\Big)^2, \end{aligned}$$

which implies

$$\nu(\pi_e^*; \boldsymbol{r}, \{\sigma_a\}) = \frac{1}{T}\Big[\sum_a |\pi_\Delta(a)| \times \sum_{a\in[K]} |\pi_\Delta(a)|(\sigma_a^2 + r_a^2) - \big(V(\pi_1) - V(\pi_0)\big)^2\Big]$$

□

D.2. Proof of Theorem 1

Proof. First, recall that

$$T \times \nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) = T \times \mathrm{var}\Big(\widehat{V}_{IPW}(\pi_1; \mathcal{D}_e) - \widehat{V}_{IPW}(\pi_0; \mathcal{D}_e)\Big) = \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} r_a^2 - c,$$

where $c = \Big(V(\pi_1) - V(\pi_0)\Big)^2$ is independent with π_e.

Denote $\mathbb{E}_Q(r_a^2) \equiv r_Q^2$ and $\mathbb{E}_Q(\sigma^2(a)) \equiv \sigma_Q^2$. Therefore, for any $\pi_e \in \Pi$, we have

$$\begin{aligned}
&T \times \mathbb{E}_Q\Big[\nu(\pi_e^*; \boldsymbol{r}, \{\sigma_a\}) - \nu(\pi_e; \boldsymbol{r}, \{\sigma_a\})\Big] \\
&= \mathbb{E}_Q\Big[\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e^*(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e^*(a)} r_a^2 - \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 - \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} r_a^2\Big] \\
&= \Big[\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e^*(a)} - \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\Big] \times \big(\sigma_a^2 + r_a^2\big) \\
&\le 0,
\end{aligned}$$

where the last inequality is due to that, by design, π_e^* is the minimizer of $\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}$ within the class of safe policies.

To prove the minimax optimality, we first show that, for any policy π_e, $\max_{(\boldsymbol{r},\sigma_a)\in\mathcal{I}} \nu(\pi_e; \boldsymbol{r}, \sigma_a)$ is achieved when $\boldsymbol{r} = \mathbf{1}$ and $\sigma_a \equiv \sigma$. To see this, we exam

$$\begin{aligned}
&T \times \nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) - T \times \nu(\pi_e; \mathbf{1}, \{\sigma, \ldots, \sigma\}) \\
&= \Big[\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} r_a^2 - (V(\pi_1) - V(\pi_0))^2\Big] \\
&\quad - \Big[\sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}\sigma^2 + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)} 1^2 - 0^2\Big] \\
&= \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}(\sigma_a^2 - \sigma^2) + \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}(r_a^2 - 1) - (V(\pi_1) - V(\pi_0))^2 \\
&\le 0.
\end{aligned}$$

Therefore, we have

$$T \times \max_{(\boldsymbol{r},\sigma_a)\in\mathcal{I}} \nu(\pi_e; \boldsymbol{r}, \sigma_a) = \sum_{a\in[K]} \frac{\pi_\Delta^2(a)}{\pi_e(a)}(\sigma^2 + 1)$$

Notice that π_e^*, by design, is the minimizer of this objective. Therefore, we have that π_e^* is minimax optimal, i.e., $\max_{(\boldsymbol{r},\sigma_a)\in\mathcal{I}} \nu(\pi_e^*; \boldsymbol{r}, \sigma_a) = \min_{\pi_e\in\Pi} \max_{(\boldsymbol{r},\sigma_a)\in\mathcal{I}} \nu(\pi_e; \boldsymbol{r}, \sigma_a)$.

□

D.3. Proof of Theorem 2

Proof. Denote $\boldsymbol{G}(\pi_e) = \sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T$. We first note that

$$T \times \nu\pi_e; \mathcal{D}_0) = T \times \bar{\phi}_{\Delta\pi}^T(T\boldsymbol{G}(\pi_e) + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\phi}_{\Delta\pi} = \bar{\phi}_{\Delta\pi}^T(\boldsymbol{G}(\pi_e) + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\phi}_{\Delta\pi} \to \bar{\phi}_{\Delta\pi}^T\boldsymbol{G}(\pi_e)^{-1}\bar{\phi}_{\Delta\pi},$$

when T goes to infinity. Besides, we note the following relationship

$$\begin{aligned}
T \times \nu^*(\pi_e; \mathcal{D}_0) &= T \times \mathrm{var}_{\boldsymbol{\Phi}\sim\pi_e}\Big(\widehat{V}_{DM}(\pi_1; \mathcal{D}_{\boldsymbol{\Phi}} \cup \mathcal{D}_0) - \widehat{V}_{DM}(\pi_0; \mathcal{D}_{\boldsymbol{\Phi}} \cup \mathcal{D}_0)\Big) \\
&= \mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\bar{\phi}_{\Delta\pi}^T(T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\phi}_{\Delta\pi}\Big].
\end{aligned}$$

Since the rows of $\boldsymbol{\Phi}$ are *i.i.d.*, by the strong law of large numbers, we know $T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} \overset{a.s.}{\to} \boldsymbol{G}(\pi_e)$, and hence $T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0 \overset{a.s.}{\to} \boldsymbol{G}(\pi_e)$. Therefore, by the continuous mapping theorem, we have $(T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1} \overset{a.s.}{\to} \boldsymbol{G}(\pi_e)^{-1}$ and also $\bar{\phi}_{\Delta\pi}^T\big[T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0\big]^{-1}\bar{\phi}_{\Delta\pi} \overset{a.s.}{\to} \bar{\phi}_{\Delta\pi}^T\boldsymbol{G}(\pi_e)^{-1}\bar{\phi}_{\Delta\pi}$. Finally, from the random matrix theory for the tail bounds on the eignevalues of random matrix with $i.i.d.$ rows (Vershynin, 2010), we know the uniformly integrable condition can be satisfied and hence we conclude with

$$\mathbb{E}_{\boldsymbol{\Phi}\sim\pi_e}\Big[\bar{\phi}_{\Delta\pi}^T\big[T^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + T^{-1}\boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0\big]^{-1}\bar{\phi}_{\Delta\pi}\Big] \to \bar{\phi}_{\Delta\pi}^T\boldsymbol{G}(\pi_e)^{-1}\bar{\phi}_{\Delta\pi}$$

□

D.4. Counter-example

In this section, we argue that there does not exist a policy π_e that dominates all other policies across all problem instances, when we use the IPW estimator.

Suppose $\pi_1 \neq \pi_0$. For any policy π_e, we can always construct an instance $(\boldsymbol{r}, \{\sigma_a\})$ and a policy b', such that

$$\nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) > \nu(\pi'_e; \boldsymbol{r}, \{\sigma_a\}).$$

To see this, note that

$$T \times \Big[\nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) - \nu(\pi'_e; \boldsymbol{r}, \{\sigma_a\})\Big] = \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{\pi_e(a)}\sigma_a^2 + \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{\pi_e(a)} r_a^2 - \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{b'(a)}\sigma_a^2 - \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{b'(a)} r_a^2.$$

By setting $r_a \equiv 0$, we get

$$T \times \Big[\nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) - \nu(\pi'_e; \boldsymbol{r}, \{\sigma_a\})\Big] = \sum_{a\in[K]} \big(\frac{1}{\pi_e(a)} - \frac{1}{b'(a)}\big)\pi^2_\Delta(a)\sigma_a^2.$$

Since $\pi_1 \neq \pi_0$, there are at least two arms a' and a'' where the two policies differ. By design, we know $\pi_e(a')$ and $\pi_e(a'')$ are both positive. A counter-example can hence be designed by setting $\sigma'_a = 0$, $\sigma_{a''} = 1$, $b'(a') = 0.5\pi_e(a')$, $b'(a'') = 0.5\pi_e(a') + \pi_e(a'')$ and keeping $\pi_e(a) = b'(a)$ on the other arms. In other words, we move more budgets to the arm of high variance. More precisely, we have

$$\begin{aligned}
&T \times \Big[\nu(\pi_e; \boldsymbol{r}, \{\sigma_a\}) - \nu(\pi'_e; \boldsymbol{r}, \{\sigma_a\})\Big] \\
&= 0 + \big(\frac{1}{\pi_e(a')} - \frac{1}{b'(a')}\big)\pi^2_\Delta(a')\sigma^2_{a'} + \big(\frac{1}{\pi_e(a'')} - \frac{1}{b'(a'')}\big)\pi^2_\Delta(a'')\sigma^2_{a''} \\
&= \big(\frac{1}{\pi_e(a'')} - \frac{1}{0.5\pi_e(a') + \pi_e(a'')}\big)\pi^2_\Delta(a'')\sigma^2_{a''} \\
&> 0.
\end{aligned}$$

D.5. Convexity of the objectives and constraints

For the sake of completeness, we show in this section that the objectives and constraints considered in this paper are all convex. Regarding the objective of IPW, we note the relationship that

$$\alpha \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{\pi_e(a)} + (1-\alpha) \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{\pi'_e(a)} \geq \sum_{a\in[K]} \frac{\pi^2_\Delta(a)}{\alpha\pi_e(a) + (1-\alpha)\pi'_e(a)},$$

due to the inequality

$$\alpha\frac{1}{\pi_e(a)} + (1-\alpha)\frac{1}{\pi'_e(a)} \geq \frac{1}{\alpha\pi_e(a) + (1-\alpha)\pi'_e(a)}, \forall a \in [K],$$

which stems from the convexity of $f(x) = 1/x$. The same arguments hold for CMAB with IPW.

Regarding the objective of linear bandits with DM,

$$\bar{\boldsymbol{\phi}}^T_{\Delta\pi}(T \sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T + \boldsymbol{\Phi}_0^T\boldsymbol{\Phi}_0)^{-1}\bar{\boldsymbol{\phi}}_{\Delta\pi},$$

we notice that $\sum_{\mathbf{x}\in\mathcal{A}} \pi_e(\mathbf{x})\mathbf{x}\mathbf{x}^T$ is linear in π_e and the matrix inverse operator is a convex function. Therefore, their composition is still convex.

Regarding the constraint, notice that, for every single instance, the constraint is a linear one and hence convex. The overall feasible set is the intersection of these convex sets, and hence is convex.