Adaptive Experimental Design for Policy Learning: Contextual Best Arm Identification

Masahiro Kato[1], Kyohei Okumura[2], Takuya Ishihara[3], and Toru Kitagawa[4]

[1]Mizuho–DL Financial Technology, Co., Ltd.
[2]Northwestern University
[1]Tohoku University
[3]Brown University

Abstract

This study investigates the contextual best arm identification (BAI) problem, aiming to design an adaptive experiment to identify the best treatment arm conditioned on contextual information (covariates). We consider a decision-maker who assigns treatment arms to experimental units during an experiment and recommends the estimated best treatment arm based on the contexts at the end of the experiment. The decision-maker uses a policy for recommendations, which is a function that provides the estimated best treatment arm given the contexts. In our evaluation, we focus on the worst-case *expected regret*, a relative measure between the expected outcomes of an optimal policy and our proposed policy. We derive a lower bound for the expected simple regret and then propose a strategy called *Adaptive Sampling-Policy Learning* (PLAS). We prove that this strategy is minimax rate-optimal in the sense that its leading factor in the regret upper bound matches the lower bound as the number of experimental units increases.

1 Introduction

In this study, we design an *adaptive experiment for policy learning*. We consider the problem of decision-making given multiple *treatment arms*, such as arms in slot machines, diverse therapies, and distinct unemployment assistance programs. The primary objective is to identify the *best treatment arm* for individuals given covariates, often referred to as context, at the end of an experiment. For this purpose, we aim to learn a policy that recommends the conditional best treatment arm by using data adaptively collected via an experiment.

In our setting, at each round of an adaptive experiment, a decision-maker sequentially observes a context (covariate) and assigns one of the treatment arms to a research subject based on past observations and the observed contexts. At the end of the experiment, the decision-maker recommends an estimated best treatment arm conditional on a context.

We design the adaptive experiment by developing a strategy that the decision-maker follows. A strategy is defined as a pair of a sampling rule and a recommendation rule. In the adaptive experiment, the decision-maker assigns treatment arms following the sampling rule during the experiment and recommends a treatment arm following the recommendation rule at the end of the experiment.

We measure the performance of a strategy using the expected simple regret, which is the difference between the maximum expected outcome that could be achieved with full knowledge of the distributions of the treatment arms and the expected outcome of the treatment arm recommended by the decision-maker's strategy. Our goal is to develop a strategy that minimizes the expected simple regret.

The challenge in our problem arises from the need for context-specific recommendations. Unlike prior studies that do not consider contextual information, developing a model that captures the relationship between context and outcomes becomes imperative.

To address this issue, we define a function suggesting a treatment arm given a context as a *policy*. By using a policy, we restrict strategies to ones with recommendation rules using a policy learned from observations obtained from an adaptive experiment. Policy learning has been extensively studied in causal inference and reinforcement learning (Dudík et al., 2011; Swaminathan & Joachims, 2015; Kitagawa & Tetenov, 2018; Athey & Wager, 2021; Zhou et al., 2023), but to the best of our knowledge, adaptive experimental design for policy learning has not been fully explored. Note that there are existing studies that address contextual BAI aiming to find the best treatment arm marginalized over the contextual distribution (Russac et al., 2021; Kato & Ariu, 2021; Simchi-Levi et al., 2024), motivated by the studies of efficient average treatment effect estimation (van der Laan, 2008; Hahn et al., 2011; Kato et al., 2020; Cook et al., 2023).

Our problem corresponds to a generalization of *best arm identification* (BAI, Bubeck et al., 2009, 2011; Audibert et al., 2010), an instance of the stochastic multi-armed bandit (MAB) problem (Thompson, 1933; Lai & Robbins, 1985). Therefore, we refer to the problem as contextual fixed-budget BAI, as well as an adaptive experimental design for policy learning.

Contribution. We propose a strategy that assigns treatment arms following *Adaptive Sampling* (AS) and recommends a treatment arm using *Policy Learning* (PL). In the AS rule, a decision-maker assigns a treatment arm to an experimental unit based on a probability depending on the variances of the experimental units' outcomes. Because the variances are unknown, the decision-maker estimates them during the experiment and continues updating the assignment probability. At the end of the experiment, the decision-maker trains a policy using observations obtained in the experiment and recommends a treatment arm using the trained policy. We refer to our strategy as the *PLAS strategy*.

To design an optimal strategy, we first develop a lower bound (theoretical limit) for the expected simple regret. Subsequently, we design a strategy and evaluate its upper bound (performance) by comparing it to the lower bound.

In the evaluation, given the inherent uncertainties, we use the minimax criterion for performance assessment, which evaluates the worst-case scenario among a set of distributions. The minimax approach has garnered attention in studies about experimental design, including BAI (Bubeck et al., 2009, 2011; Carpentier & Locatelli, 2016; Ariu et al., 2021; Yang & Tan, 2022; Komiyama et al., 2022). A critical quantity in the minimax evaluation is the gap between the expected outcomes of the best treatment and the other suboptimal treatment arms, referred to as the average treatment effects in the literature of causal inference. The worst-case distributions are characterized by gaps approaching zero at a rate of order $\sqrt{T}$ (Bubeck et al., 2009, 2011), where T denotes the sample size (total rounds of an adaptive experiment).

Our research identifies the leading factor in the lower bound as the variances of potential outcomes, also providing a variance-dependent sampling rule. We subsequently show that the PLAS strategy is asymptotically minimax optimal, as its foremost factor of the worst-case expected simple regret aligns with the lower bound.

In summary, our contributions include: (i) a lower bound for the worst-case expected simple regret; (ii) the PLAS strategy with a closed-form target assignment ratio, characterized by the variances of outcomes; and (iii) the asymptotic minimax optimality of the PLAS strategy. These findings contribute to a variety of subjects, including decision theory and causal inference, in addition to BAI.

Organization. The structure of this paper is as follows: Section 2 defines our problem. Section 3 develops lower bounds for the worst-case expected simple regret. Section 4 introduces the PLAS strategy, and Section 5 presents upper bounds for the proposed strategy and its asymptotic minimax optimality. Further related work is introduced in Appendix A.

2 Problem Setting

This study considers an adaptive experiment with a fixed budget (sample size) $T \in \mathbb{N}$, a set of treatment arms $[K] \coloneqq \{1, 2, \ldots, K\}$, and a decision-maker who aims to identify the context-

conditional best treatment arm. In each round $t \in [T]$, experimental units sequentially visit, and the decision-maker can assign treatment arms to them. At the end of the experiment, the decision-maker recommends an estimated context-conditional best treatment arm.

2.1 Potential Outcomes

Following the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974), let $Y^a \in \mathcal{Y}$ be a potential random outcome of treatment arm $a \in [K]$, where $\mathcal{Y} \subset \mathbb{R}$ denotes a set of possible outcomes. We also define $X \in \mathcal{X}$ as a context, also called covariates, that characterizes an experimental unit, where $\mathcal{X} \subset \mathbb{R}^d$ denotes a set of possible d-dimensional covariates. We define a tuple $(Y^1, \ldots, Y^K, X)$ as a $(\mathcal{Y}^K \times \mathcal{X})$-valued random variable following a probability distribution $P \in \mathcal{M}(\mathcal{Y}^K \times \mathcal{X})$, where $\mathcal{M}(\mathcal{T})$ denotes the set of all Borel probability measures on a topological space $\mathcal{T}$. Let $\mathbb{E}_P$ and Var_P be the expectation and variance operators under P, respectively. For each $P \in \mathcal{P}$, let us denote the first and second moments of the potential outcome as $\mu^a(P)(x) := \mathbb{E}_P[Y^a \mid X = x]$ and $\nu^a(P)(x) := \mathbb{E}_P[(Y^a)^2 \mid X = x]$, respectively.

In our setting, a distribution P of $(Y^1, \ldots, Y^K, X)$ belongs to a bandit model $\mathcal{P}$ defined below. We refer to the bandit model as a location-shift bandit model.

Definition 2.1 (Location-shift bandit model). *Let $(\sigma^a)^2 \colon \mathcal{X} \to (0, +\infty)$ be a function that is exogenously given but* unknown *to the decision-maker. Let ζ be a distribution of X that is exogenously given but* unknown *to the decision-maker, which is supported on $\mathcal{X}$. Then, a* location-shift bandit model $\mathcal{P}$ *is defined as follows:*

$$\mathcal{P} := \mathcal{P}_\zeta := \Big\{ P \in \mathcal{M}(\mathcal{Y}^K \times \mathcal{X}) \colon \forall a \in [K] \quad \forall x \in \mathcal{X}, \quad \mathrm{Var}_P(Y^a \mid X = x) = \sigma^a(x),$$

$$Y^a - \mu^a(P)(x) \text{ is zeo-mean subgaussian given } x, \quad \mu^a(P)(x) \in (-\infty, +\infty), \quad \mathrm{marg}_{\mathcal{X}}(P) = \zeta \Big\},$$

where $\mathrm{marg}_{\mathcal{X}}(P)$ denotes the marginal distribution of P on $\mathcal{X}$.

Our lower and upper bounds are characterized by $\sigma^a(x)$, given independently of the experiment. Location-shift models are a commonly employed assumption in statistical analysis (Lehmann & Casella, 1998). A key example is a Gaussian distribution, where for all P, the variances are fixed and only mean parameters shift. Note that we can omit the condition $|\nu^a(P)(x)| < \overline{C}$ from the boundedness of the variance. However, we introduce it to simplify the definition of our strategies.

2.2 Adaptive Experiment

We consider a decision-maker who aims to identify the best treatment arm

$$a^*(P)(x) \in \arg\max_{a \in [K]} \mu^a(P)(x)$$

for each context $x \in \mathcal{X}$ through an adaptive experiment. A fixed number of rounds T, called the budget or sample size, is exogenously given. At each round $t \in [T]$, the decision-maker uses the following procedure:

1. A potential outcome $(Y_t^1, Y_t^2, \ldots, Y_t^K, X_t)$ is generated from P.
2. The decision-maker observes a context X_t.
3. The decision-maker assigns treatment arm A_t to an experimental unit based on past observations $(X_s, A_s, Y_s)_{s=1}^{t-1}$ and the context X_t.
4. The decision-maker observes an outcome of the assigned treatment arm, $Y_t = \sum_{a \in [K]} \mathbb{1}[A_t = a]Y_t^a$.

This process is referred to as the *exploration phase*. Note that outcomes from unassigned treatment arms remain unobserved. This setting is called the bandit feedback or Rubin causal model (Neyman, 1923; Rubin, 1974). Through the exploration phase, the decision-maker obtains observations $(X_t, A_t, Y_t)_{t=1}^{T}$. After round T, the decision-maker recommends an estimated best treatment arm $\widehat{a}_T(x)$ for each context x given the observations $(X_t, A_t, Y_t)_{t=1}^{T}$.

3

2.3 Strategy with Policy Learning

A strategy of the decision-maker defines which treatment arm to assign in each period during the exploration phase and which treatment arm to recommend as an estimated best arm for each context. A strategy is defined as a pair $((A_t)_{t\in[T]}, \widehat{a}_T)$ in which:

- the *sampling rule* $(A_t)_{t\in[T]}$ determines which treatment arm A_t to assign in round t based on the past observations. In other words, A_t is $\mathcal{G}_t$-measurable, where $\mathcal{G}_t \coloneqq \sigma(X_1, A_1, Y_1, \dots, X_{t-1}, A_{t-1}, Y_{t-1}, X_t)$ for each $t \in [T]$.
- the *recommendation rule* $\widehat{a}_T : \mathcal{X} \to [K]$ returns an estimated best treatment arm for each context $\widehat{a}_T$ based on the observations collected during the exploration phase. In other words, for each $x \in \mathcal{X}$, $\widehat{a}_T(x)$ is $\mathcal{F}_T$-measurable, where $\mathcal{F}_T \coloneqq \sigma(X_1, A_1, Y_1, \dots, X_T, A_T, Y_T)$.

Policy. In this study, we impose a restriction on a class of recommendation rules that the decision-maker can use. We assume that there is an *exogenously* given class of policies Π, whose typical element is a policy $\pi \colon [K] \times \mathcal{X} \to [0,1]$ that is measurable and satisfies $\sum_a \pi(a, x) = 1$ for each x. Here, $\pi(a, x)$ denotes a probability that the decision-maker recommends treatment arm $a \in [K]$ as an estimated best treatment arm for context x. With a slight abuse of notation, we write $\pi(a \mid x)$ instead of $\pi(a, x)$. We require the decision-maker to obtain an estimator $\widehat{a}_T(x)$ of $a^*(P)(x)$ as follows: first, the decision-maker constructs a policy $\pi \in \Pi$ based on the observations collected during the exploration phase; then, $\widehat{a}_T(x)$ is drawn from $\pi(\cdot \mid x)$ for each x.

Optimal policy. We evaluate the performance of policies via a simple regret. The *value* of policy $\pi \in \Pi$ under P is the expected outcome when the decision-maker uses a policy π, which is defined as

$$Q(P)(\pi) \coloneqq \mathbb{E}_P \left[\sum_{a\in[K]} \pi(a \mid X)\mu^a(P)(X); \pi \right],$$

and the optimal policy within class Π is defined as $\pi^*(P) \coloneqq \arg\max_{\pi\in\Pi} Q(P)(\pi)$.

Given a strategy with a policy $\widehat{\pi} \in \Pi$ of the decision-maker, we define a simple regret for each context $x \in \mathcal{X}$ under $P \in \mathcal{P}$ as

$$r_T(P)(\widehat{\pi})(x) \coloneqq \sum_{a\in[K]} \pi^*(P)(a \mid x)\mu^a(P)(x) - \mu^{\widehat{a}_T(x)}(P)(x),$$

and the marginalized simple regret is defined as

$$R_T(P)(\widehat{\pi}) \coloneqq \mathbb{E}_{X\sim\zeta}[r_T(P)(\widehat{\pi})(X)] = Q(P)(\pi^*(P)) - Q(P)(\widehat{\pi}).$$

Then, we define the *expected simple regret* as $\mathbb{E}_P[R_T(P)(\widehat{\pi})]$, where the expectation is taken over $\widehat{\pi}$. This expected simple regret is our performance measure of interest. We also refer to the expected simple regret $\mathbb{E}_P[R_T(P)(\widehat{\pi})]$ as the *policy regret*. The decision-maker aims to identify the best treatment arm with a smaller expected simple regret.

Notation. Let $o(g(x))$ be Landau's notation, and $f(x) = o(g(x))$ implies that $\forall \varepsilon > 0 \; \exists x_0 \; \forall x > x_0 \colon |f(x)| < \varepsilon g(x)$ holds. Let let $\text{thre}(A, a, b) \coloneqq \max\{\min\{A, a\}, b\}$ be a truncation function.

3 Regret Lower Bound

This section presents a lower bound on the expected simple regret $\mathbb{E}_P[R_T(P)(\widehat{\pi})]$. The lower bound is provided under weak conditions on the policy. Not only does the lower bound offer insights into the difficulty of the problem, but it also helps argue which sample allocations are optimal.

3.1 Restriction and Complexity of a Policy Class

To establish lower bounds, we introduce a moderate precondition related to the strategy space of the decision-maker. Specifically, we require that, in the limit, strategies choose all the arms with an equal probability when, for a given covariate x, the expected outcomes associated with all arms are identical. Strategies adhering to this criterion are termed *null consistent strategies*.

Definition 3.1 (Null consistent strategy). *We say a strategy is* null consistent *if the following condition is satisfied: If $\mu^1(P)(x) = \mu^2(P)(x) = \cdots = \mu^K(P)(x)$, then for any $a, b \in [K]$, we have*

$$\left| \mathbb{P}_P \left(\widehat{a}_T(X) = a \mid X = x \right) - \mathbb{P}_P \left(\widehat{a}_T(X) = b \mid X = x \right) \right| \to 0 \quad (T \to \infty).$$

Under any null consistent strategies, $\left| \mathbb{P}_P \left(\widehat{a}_T(X) = a \mid X = x \right) - 1/K \right| = o(1)$ holds for each $a \in [K]$ as $T \to \infty$ if $\mu^1(P)(x) = \mu^2(P)(x) = \cdots = \mu^K(P)(x)$.

Next, we introduce the Natarajan dimension, a metric that measures the complexity of a policy class Π (Natarajan, 1989). Our lower bounds are characterized by the Natarajan dimension.

Definition 3.2 (Natarajan dimension). *We say that Π shatters M points $\{s_1, s_2, \ldots, s_M\} \subseteq \mathcal{X}$ if there exist $f_1, f_{-1} : \{s_1, s_2, \ldots, s_M\} \to [K]$ such that*

1. for any $j \in [M]$, $f_{-1}(s_j) \neq f_1(s_j)$ holds;

2. for any $\boldsymbol{\sigma} := \{\sigma_1, \sigma_2, \ldots, \sigma_M\} \in \{\pm 1\}^M$, there exists a policy $\pi \in \Pi$ such that for any $j \in [M]$, it holds that $\pi(s_j) = \begin{cases} f_1(s_j) & \text{if} \quad \sigma_j = 1 \\ f_{-1}(s_j) & \text{if} \quad \sigma_j = -1 \end{cases}$.

The Natarajan dimension of Π, denoted by $d_{\mathrm{N}}(\Pi)$, is the maximum cardinality of a set shattered by Π.

Let $d_{\mathrm{VC}}(\Pi)$ be the *Vapnik-Chervonenkis (VC) dimension* of Π. Note that when $K = 2$, the Natarajan dimension is equivalent to the VC dimension; that is, when $K = 2$, $d_{\mathrm{VC}}(\Pi) = d_{\mathrm{N}}(\Pi)$ holds.

3.2 Regret Lower Bounds

We derive the following lower bounds of the expected simple regret, which hold for any null consistent strategies and depend on the complexity of a policy class Π, measured by the Natarajan dimension. The proof is shown in Appendix B.

Theorem 3.3. *There exists a distribution ζ on $\mathcal{X}$ such that for any $K \geq 2$, any null consistent strategy π with a policy class Π such that $d_{\mathrm{N}}(\Pi) = M$ satisfies*

$$\sup_{P \in \mathcal{P}_\zeta} \sqrt{T} \mathbb{E}_P \left[R(P)(\pi) \right] \geq \frac{1}{8} \mathbb{E}_{X \sim \zeta} \left[\sqrt{M \sum_{a \in [K]} (\sigma^a(X))^2} \right] + o(1) \quad \text{as} \ T \to \infty,$$

where $\mathbb{E}_{X \sim \zeta}$ denotes the expectation of a random variable X under the probability distribution ζ.

Our lower bounds also depend on the variances of outcomes Y^a.

When $K = 2$, we can obtain a tighter lower bound than the one in Theorem 3.3. The proof is provided in Section B.4.

Theorem 3.4. *There exists a distribution ζ on $\mathcal{X}$ such that for $K = 2$, any null consistent strategy π with policy class Π such that $d_{\mathrm{VC}}(\Pi) = M$ satisfies*

$$\sup_{P \in \mathcal{P}_\zeta} \sqrt{T} \mathbb{E}_P \left[R(P)(\pi) \right] \geq \frac{1}{8} \mathbb{E}_{X \sim \zeta} \left[\sqrt{M (\sigma^1(X) + \sigma^2(X))^2} \right] + o(1) \quad \text{as} \ T \to \infty.$$

Here, note that $\sum_{a \in \{1,2\}} (\sigma^a(x))^2 \leq (\sigma^1(x) + \sigma^2(x))^2$ holds for each $x \in \mathcal{X}$. Therefore, when $K = 2$, we use the lower bound in Theorem 3.4 and when $K \geq 3$, we use the one in Theorem 3.3.

4 The PLAS Strategy

Our strategy consists of the following sampling and recommendation rules. First, we define a *target assignment ratio*, which is an ideal treatment assignment probability. At each round, $t = 1, 2, \ldots, T$, our sampling rule randomly assigns a unit to a treatment arm with a probability identical to an estimated target assignment ratio. After the final round T, our recommendation rule recommends a treatment arm with the highest value of a policy trained by maximizing empirical policy value. We refer to our strategy as the PLAS strategy. Our strategy depends on hyperparameters $\overline{C} \in (0, \infty)$, which are introduced for technical purpose in the proof and can be set as sufficiently large values. We show a pseudo-code in Algorithm 1.

4.1 Optimal Target Assignment Ratio

We first define a target assignment ratio. The target assignment ratio is the expected value of the sample average of A_t of a strategy ($\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left[\mathbb{1}[A_t = a] \mid X_t = x\right]$) under which a leading factor of its expected simple regret aligns with that of our derived lower bound.

Definition 4.1 (Target assignment ratio). *When $K = 2$, for each $a \in [K] = \{1, 2\}$, we define the target assignment ratio w^* as*

$$w^*(a \mid x) = \frac{\sigma^a(x)}{\sigma^1(x) + \sigma^2(x)}. \tag{1}$$

When $K \geq 3$, for each $a \in [K]$, we define the target assignment ratio as

$$w^*(a \mid x) = \frac{\left(\sigma^a(x)\right)^2}{\sum_{b\in[K]}\left(\sigma^b(x)\right)^2}. \tag{2}$$

This target assignment ratio is given in the course of proving Theorem 3.3, in which we solve $\min_{\boldsymbol{w}\in\mathcal{W}}\left\{\sum_{s\in\mathcal{S}}\max_{a\in[K]}\left\{\sqrt{\frac{(\sigma^a(s))^2}{w(a\mid s)}}\right\}\right\}$, where $\mathcal{S} \coloneqq \{s_1, s_2, \ldots, s_M\} \subseteq \mathcal{X}$, and $\mathcal{W}$ is the set of all measurable functions $w : \mathcal{X} \times [K] \rightarrow (0, 1)$ such that $\sum_{a\in[K]} w(a \mid x) = 1$ for each $x \in \mathcal{X}$. The solutions w^*, whose explicit forms appear in (1) and (2), work as a conjecture for $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_P\left[\mathbb{1}[A_t = a] \mid X_t = x\right]$.

These target assignment ratios are ex-ante unknown to the decision-maker since the variance $(\sigma^a(x))^2$ is unknown. Therefore, the decision-maker needs to estimate it during an adaptive experiment and employ the estimator as a probability of assigning a treatment arm.

Remark (Efficiency gain). *For each $a \in [K]$, let $(\sigma^a)^2$ be the unconditional variance of Y_t^a, and $w : [K] \rightarrow [0, 1]$ be an assignment ratio such that $\sum_{a\in[K]} w(a) = 1$ when we cannot utilize the contextual information. Then, the following inequality holds:*

$$\sup_{P\in\mathcal{P}}\sqrt{T}\mathbb{E}_P\left[R(P)(\widehat{\pi})\right] \geq \frac{1}{8}\inf_{w\in\mathcal{W}}\max_{a\in[K]}\sqrt{M(\sigma^a)^2/w(a)}$$

$$\geq \frac{1}{8}\inf_{w\in\mathcal{W}}\max_{a\in[K]}\sqrt{M\underset{X\sim\zeta}{\mathbb{E}}\left[(\sigma^a(X))^2/w(a \mid X)\right]} \geq \frac{1}{8}\inf_{w\in\mathcal{W}}\max_{a\in[K]}\underset{X\sim\zeta}{\mathbb{E}}\left[\sqrt{M(\sigma^a(X))^2/w(a \mid X)}\right].$$

This result implies that we can minimize a lower bound by using contextual information; that is, strategies utilizing contextual information are more efficient than ones not utilizing contextual information.

4.2 Sampling Rule with Adaptive Sampling (AS)

In this section, we describe our sampling rule, referred to as athe AS rule.

In round $t \leq K$, the strategy chooses $A_t = t$, i.e., each arm is pulled once as initialization. In round $t > K$, given an estimated target assignment ratio $\widehat{w}_t(a \mid x)$, we assign treatment arm a with probability $\widehat{w}_t(a \mid X_t)$. Below, we describe the construction of $\widehat{w}_t(a \mid x)$.

In each round t, we estimate w^* using the past observations. We first construct estimators $\widehat{\mu}_t^a(x)$ and $\widehat{\nu}_t^a(x)$ of the first moment $\mu^a(P)(x)$ and the second moment $\nu^a(P)(x)$ of Y^a. The estimators constructed to converges to the true functions with probability one, as stated in Assumption 5.1, and their absolute values are bounded by $\overline{C}$. Then, given these estimators, we estimate the variances as $(\widehat{\sigma}_t^a(x))^2 = \text{thre}\left((\widehat{\sigma}_t^{\dagger a}(x))^2, \overline{C}, 1/\overline{C}\right)$, where $(\widehat{\sigma}_t^{\dagger a}(x))^2 = \widehat{\nu}_t^a(x) - (\widehat{\mu}_t^a(x))^2$. Lastly, we construct the estimator of the target assignment ratio $\widehat{w}_t(a \mid x)$ by replacing $\sigma^a(x)$ by the estimator $\widehat{\sigma}_t^a(x)$.

For obtaining estimators $\widehat{\mu}_t^a(x)$, $\widehat{\nu}_t^a(x)$, and $\widehat{\sigma}_t^a(x)$, we can use nonparametric estimators, such as the nearest neighbor regression estimator and kernel regression estimator, which have been proven to converge to the true function almost surely under a bounded sampling probability $\widehat{w}_t$ by Yang & Zhu (2002) and Qian & Yang (2016). It should be noted that we do not assume specific convergence rates for estimators for $\mu^a(P)(x)$ and w^* as the asymptotic optimality of the AIPW estimator can be demonstrated without them (van der Laan, 2008; Kato et al., 2020, 2021).

Algorithm 1 PLAS strategy

Parameter: Positive constants C_μ and C_{σ^2}.
Initialization:
for $t = 1$ to K **do**
　　Assign $A_t = t$. For each $a \in [K]$, set $\widehat{w}_t(a \mid x) = 1/K$.
end for
for $t = K + 1$ to T **do**
　　Observe covariate X_t.
　　Construct the estimated target assignment ratio $\widehat{w}_t$ defined in Definition 4.1.
　　Draw ξ_t from the uniform distribution on $[0, 1]$.
　　$A_t = 1$ if $\xi_t \leq \widehat{w}_t(1 \mid X_t)$ and $A_t = a$ for $a \geq 2$ if $\xi_t \in \left(\sum_{b=1}^{a-1} \widehat{w}_t(b \mid X_t), \sum_{b=1}^{a} \widehat{w}_t(b \mid X_t) \right]$.
end for
Construct $\widehat{Q}(\pi)$ following (3).
Train a policy $\widehat{\pi}_T^{\mathrm{PLAS}}$ as $\widehat{\pi}_T^{\mathrm{PLAS}} = \arg\max_{\pi \in \Pi} \widehat{Q}(\pi)$.
Recommend $\widehat{a}_T$ following $\widehat{\pi}_T^{\mathrm{PLAS}}$.

4.3 Recommendation Rule with Policy Learning

The following part presents our recommendation rule. To recommend the conditionally best treatment arm $a^*(P)(x)$, we train a policy $\pi : [K] \times \mathcal{X} \to [0, 1]$ by maximizing the empirically approximated policy value function, which we will describe below.

At the end of an experiment, we estimate the policy value $Q(\pi)$ by using the augmented doubly robust estimator, which is defined as follows:

$$\widehat{Q}_T(\pi) := \frac{1}{T} \sum_{t=1}^{T} \sum_{a \in [K]} \pi(a \mid X_t) \widehat{\Gamma}_t^a, \tag{3}$$

where

$$\widehat{\Gamma}_t^a := \frac{\mathbb{1}[A_t = a] \big(c_T(Y_t) - \widehat{\mu}_t^a(X_t) \big)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t), \quad c_T(Y_t) = \mathrm{thre}\Big(Y_t, U_T, -U_T \Big),$$

and U_T is a positive value that approaches infinity as $T \to \infty$. Then, we train a policy as

$$\widehat{\pi}_T^{\mathrm{PLAS}} := \arg\max_{\pi \in \Pi} \widehat{Q}_T(\pi)$$

By using this trained policy, given $x \in \mathcal{X}$, we recommend $\widehat{a}_T(x) \in [K]$ as the best treatment arm with probability $\widehat{\pi}_T^{\mathrm{PLAS}}(\widehat{a}_T(x) \mid x)$.

The AIPW estimator debiases the sample selection bias resulting from treatment assignment based on contextual information. Additionally, the AIPW estimator possesses the following properties: (i) its components $\{\widehat{\Gamma}_t^a - \mu^a(P)(x)\}_{t=1}^{T}$ are a martingale difference sequence, allowing us to employ the martingale limit theorems in derivation of the upper bound; (ii) it has the minimal asymptotic variance among the possible estimators. For example, other estimators with a martingale property, such as the inverse probability weighting (IPW) estimator, may be employed, yet their asymptotic variance would be greater than that of the AIPW estimator. The t-th element of the sum in the AIPW estimator utilizes nuisance parameters ($\mu^a(P)(x)$ and w^*) estimated from past observations up to round $t - 1$ for constructing a martingale difference sequence (van der Laan, 2008; Hadad et al., 2021; Kato et al., 2020, 2021). For those reasons, this estimator is often used in the context of adaptive experimental design.

5 Regret Upper Bound

This section provides upper bounds for the expected simple regret of the PLAS strategy. First, we assume the following convergence rate for estimators of $\mu^a(P)(x)$ and $w^*(a \mid x)$.

Assumption 5.1. *For any ζ, any $P \in \mathcal{P}_\zeta$, and all $a \in [K]$, it holds that*

$$\sup_{x \in \mathcal{X}} \left| \widehat{w}_T(a \mid x) - w^*(a \mid x) \right| \xrightarrow{\text{a.s.}} 0, \quad \sup_{x \in \mathcal{X}} \left| \widehat{\mu}_T^a(x) - \mu^a(P)(x) \right| \xrightarrow{\text{a.s.}} 0 \qquad \text{as } T \to \infty.$$

Next, we define an entropy integral of a policy class.

Definition 5.2. *Given the feature domain $\mathcal{X}$, a policy class Π, a set of n points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, define:*

1. *Hamming distance between any two policies π_1 and π_2 in Π: $H(\pi_1, \pi_2) = \frac{1}{n} \sum_{j=1}^n \mathbb{1}\left[\pi_1(x_j) \neq \pi_2(x_j)\right]$.*

2. *ϵ-Hamming covering number of the set $\{x_1, \ldots, x_n\}$:*

   $\mathbb{N}_H(\epsilon, \Pi, \{x_1, \ldots, x_n\})$ is the smallest number K of policies $\{\pi_1, \ldots, \pi_K\}$ in Π, such that $\forall \pi \in \Pi, \exists \pi_i, H(\pi, \pi_i) \leq \epsilon$.

3. *ϵ-Hamming covering number of Π: $\mathbb{N}_H(\epsilon, \Pi) = \sup\{\mathbb{N}_H(\epsilon, \Pi, \{x_1, \ldots, x_m\}) \mid m \geq 1, x_1, \ldots, x_m \in \mathcal{X}\}$.*

4. *Entropy integral: $\kappa(\Pi) = \int_0^1 \sqrt{\log \mathbb{N}_H(\epsilon^2, \Pi)} d\epsilon$.*

The entropy represents the complexity of a policy class, as well as the Natarajan dimension. Between the entropy integral $\kappa(\Pi)$ and the Natarajan dimension $d_{\mathrm{N}}(\Pi)$, $\kappa(\Pi) \leq C\sqrt{\log(d) d_{\mathrm{N}}(\Pi)}$ holds for some universal constant $C > 0$ (Jin, 2023; Zhan et al., 2022) when $K \geq 3$. When $K = 2$, $d_{\mathrm{N}}(\Pi)$ is equal to the VC dimension, and $\kappa(\Pi) \leq 2.5\sqrt{d_{\mathrm{N}}(\Pi)}$ holds (Haussler, 1995).

Furthermore, we make the following assumption for the ϵ-Hamming covering number.

Assumption 5.3. *For all $\epsilon \in (0, 1)$, $\mathbb{N}_H(\epsilon, \Pi) \leq C \exp(D(\frac{1}{\epsilon})^\omega)$ for some constants $C, D > 0$, $0 < \omega < 0.5$.*

Then, we obtain the following upper bound for the expected simple regret of the PLAS strategy.

Theorem 5.4 (Upper bound). *Suppose that Assumptions 5.1–C.1 hold. Then, for any ζ and any $P \in \mathcal{P}_\zeta$, the expected simple regret of the PLAS strategy satisfies*

$$\sqrt{T} \mathbb{E}_P\left[R(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right]$$

$$\leq \begin{cases} \left(108.8 \kappa(\Pi) + 870.4\right) \mathbb{E}_{X \sim \zeta}\left[\sqrt{\sum_{a=1}^K (\sigma^a(X))^2}\right] + o(1) & \text{if } K = 2 \\ \left(108.8 \kappa(\Pi) + 870.4\right) \mathbb{E}_{X \sim \zeta}\left[\sqrt{(\sigma^a(X) + \sigma^2(X))^2}\right] + o(1) & \text{if } K \geq 3 \end{cases} \quad \text{as } T \to \infty,$$

where $\kappa(\Pi) = \begin{cases} 2.5\sqrt{d_{\mathrm{N}}(\Pi)} & \text{if } K = 2 \\ C\sqrt{\log(d) d_{\mathrm{N}}(\Pi)} & \text{if } K \geq 3 \end{cases}$ holds for a universal constant $C > 0$.

From Lemma 5.4 and the relationship between the entropy integral and the Natarajan dimension, the following theorem holds.

When $K = 2$, for $M = d_{\mathrm{VC}}(\Pi)$, the regret upper bound is given as $\left(272\sqrt{M} + 870.4\right) \mathbb{E}_{X \sim \zeta}\left[\sqrt{\sum_{a=1}^K (\sigma^a(X))^2}\right] + o(1)$ as $T \to \infty$. When $K \geq 3$, for $M = d_{\mathrm{N}}(\Pi)$, the regret upper bound is given as $\left(108.8 C\sqrt{\log(d) M} + 870.4\right) \mathbb{E}_{X \sim \zeta}\left[\sqrt{\sum_{a=1}^K (\sigma^a(X))^2}\right] + o(1)$ as $T \to \infty$.

Here, we note that the leading factor in these upper bounds are $\mathbb{E}_{X \sim \zeta}\left[\sqrt{\log(d) M \sum_{a=1}^K (\sigma^a(X))^2}\right]$ and $\mathbb{E}_{X \sim \zeta}\left[\sqrt{M (\sigma^a(X) + \sigma^2(X))^2}\right]$, a product of the policy complexity (Natarajan dimension $d_{\mathrm{N}}(\Pi)$) and outcome variances. This theorem implies that the leading factors align with the lower bounds with high probability.
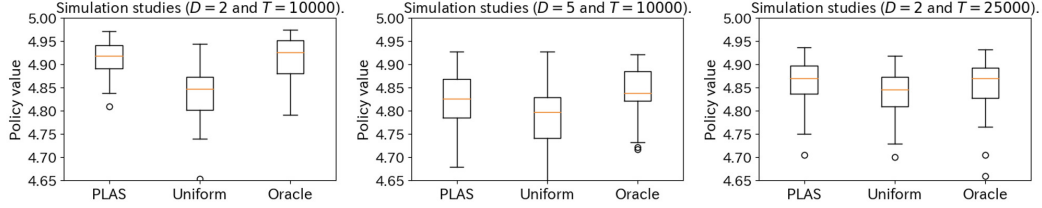
Figure 1: The results of simulation studies. The y-axis is the policy value of the learned policy.

6   Simulation Study

We conduct simulation studies to investigate the empirical performance of our proposed PLAS strategy. We compare the PLAS strategy with a combination of uniform sampling and policy learning, denoted as Uniform. The Uniform strategy assigns treatment arms with an equal ratio of $1/K$ and then applies policy learning. As a baseline method, we use the PLAS strategy with known variances, referred to as Oracle.

We consider a simple scenario with $K = 4$. We examine three cases for d and T: $(d, T) = (2, 10000)$, $(d, T) = (5, 10000)$, and $(d, T) = (5, 25000)$. Let X_i be the i-th dimension of X, and let m_i and v_i be its mean and variance, respectively. The mean m_i is drawn from a uniform distribution with support $[-1, 1]$, and the variance v_i is fixed at 1. If $X_{(1)} > 0.5$ and $X_{(2)} > 0.5$, then $\mu^1(P)(X) = 5.00$ and $\mu^2(P)(X) = \mu^3(P)(X) = \mu^4(P)(X) = 4.50$; if $X_{(1)} < 0.5$ and $X_{(2)} > 0.5$, then $\mu^2(P)(X) = 5.00$ and $\mu^1(P)(X) = \mu^3(P)(X) = \mu^4(P)(X) = 4.50$; if $X_{(1)} > 0.5$ and $X_{(2)} < 0.5$, then $\mu^2(P)(X) = 5.00$ and $\mu^1(P)(X) = \mu^3(P)(X) = \mu^4(P)(X) = 4.50$; if $X_{(1)} < 0.5$ and $X_{(2)} < 0.5$, then $\mu^4(P)(X) = 5.00$ and $\mu^1(P)(X) = \mu^2(P)(X) = \mu^3(P)(X) = 4.50$.

We conduct 50 independent trials to evaluate the performance of the strategies. The results are presented in Figure 1 with three different settings for d and T. The y-axis represents the policy value of the learned policy. From the experimental results, we confirm that our proposed strategy effectively improves the policy value.

7   Conclusion

In this study, we presented an adaptive experiment with policy learning. Our main contributions include the derivation of lower bounds for strategies, the development of the PLAS strategy, and the establishment of its regret upper bound. First, by utilizing the lower bounds developed by Kaufmann et al. (2016), we derived lower bounds for the expected simple regret, which depend on the variances of outcomes. Then, based on these lower bounds, we developed the PLAS strategy, which trains a policy at the end of the experiment. Lastly, we provided upper bounds for the regret of the PLAS strategy.

From a technical perspective, we demonstrated how to use Rademacher complexity for i.i.d. samples in an adaptive experiment with non-i.i.d. samples. We did not employ complexity measures for non-i.i.d. samples as presented by Rakhlin et al. (2015) and Foster et al. (2023). Instead, our technique relies on an approach used by Hahn et al. (2011), which we extended by incorporating sample splitting, also known as double machine learning (van der Laan, 2008; Zheng & van der Laan, 2011; Chernozhukov et al., 2018; Hadad et al., 2021; Kato et al., 2020, 2021).

We also contributed to the literature on policy learning by providing a variance-dependent lower bound, which applies to observational studies with i.i.d. samples, and by discussing matching upper bounds. Our derived lower bound is distinct from that in Athey & Wager (2021) and more tightly depends on variances, necessitating the refinement of existing upper bounds.

Our next step is to tighten both the lower and upper bounds. When $K = 2$, we showed that assigning each treatment arm a proportion based on standard deviations is optimal, consistent with existing works such as Neyman (1934), Glynn & Juneja (2004), and Kaufmann et al. (2016). However, we found that assigning each treatment arm a proportion based on variances is optimal when $K \geq 3$. Other studies on fixed-budget BAI without contextual information, such as Glynn & Juneja (2004), Kaufmann et al. (2016), and Kato (2024a), indicate that strategies with different sampling rules are optimal. Bridging the gap between our study and these existing studies remains an open issue.

References

Adusumilli, K. Risk and optimal policies in bandit experiments, 2021. arXiv:2112.06363.

Adusumilli, K. Minimax policies for best arm identification with two arms, 2022. arXiv:2204.05527.

Ariu, K., Kato, M., Komiyama, J., McAlinn, K., and Qin, C. Policy choice and best arm identification: Asymptotic analysis of exploration sampling, 2021. arXiv:2109.08229.

Armstrong, T. B. Asymptotic efficiency bounds for a class of experimental designs, 2022. arXiv:2205.02726.

Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021. doi: https://doi.org/10.3982/ECTA15732.

Atsidakou, A., Katariya, S., Sanghavi, S., and Kveton, B. Bayesian fixed-budget best-arm identification, 2023. arXiv:2211.08572.

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pp. 41–53, 2010.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–a482, 2003.

Bechhofer, R., Kiefer, J., and Sobel, M. *Sequential Identification and Ranking Procedures: With Special Reference to Koopman-Darmois Populations*. University of Chicago Press, 1968.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory (ALT)*, 2009.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 2011.

Carpentier, A. and Locatelli, A. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory (COLT)*, 2016.

Chen, C.-H., Lin, J., Yücesan, E., and Chick, S. E. Simulation budget allocation for further enhancing theefficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

Cook, T., Mishler, A., and Ramdas, A. Semiparametric efficient inference in adaptive experiments. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*, 2023. URL https://openreview.net/forum?id=xfj5jjpOaL. a]rXiv:2311.18274.

Degenne, R. On the existence of a complexity in fixed budget bandit identification. In *Conference on Learning Theory (COLT)*, 2023.

Deshmukh, A. A., Sharma, S., Cutler, J. W., Moldwin, M., and Scott, C. Simple regret minimization for contextual bandits, 2018. arXiv:1810.07371.

Dominitz, J. and Manski, C. F. Minimax-regret sample design in anticipation of missing data, with application to panel data. *Journal of Econometrics*, 226(1):104–114, 2022. Annals Issue in Honor of Gary Chamberlain.

Dominitz, J. and Manski, F. C. More Data or Better Data? A Statistical Decision Problem. *The Review of Economic Studies*, 84(4):1583–1605, 02 2017.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *International Conference on Machine Learning (ICML)*, 2011.

Even-Dar, E., Mannor, S., Mansour, Y., and Mahadevan, S. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 2006.

Foster, D., Foster, D. J., Golowich, N., and Rakhlin, A. On the complexity of multi-agent decision making: From learning in games to partial monitoring. In *Conference on Learning Theory (COLT)*, 2023.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, 2016.

Glynn, P. and Juneja, S. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference*, volume 1. IEEE, 2004.

Guan, M. and Jiang, H. Nonparametric stochastic contextual bandits. *AAAI Conference on Artificial Intelligence*, 2018.

Gupta, S., Lipton, Z. C., and Childers, D. Efficient online estimation of causal effects by deciding what to observe. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Hahn, J., Hirano, K., and Karlan, D. Adaptive experimental design using the propensity score. *Journal of Business and Economic Statistics*, 2011.

Haussler, D. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

Hirano, K. and Porter, J. R. Asymptotics for statistical treatment rules. *Econometrica*, 77(5): 1683–1701, 2009.

Hirano, K. and Porter, J. R. Asymptotic representations for sequential decisions, adaptive experiments, and batched bandits, 2023. arXiv:2302.03117.

Ito, S., Tsuchiya, T., and Honda, J. Adversarially robust multi-armed bandit algorithm with variance-dependent regret bounds. In *Conference on Learning Theory*, 2022.

Jennison, C., Johnstone, I. M., and Turnbull, B. W. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III*, pp. 55–86. Academic Press, 1982.

Jin, Y. Upper bounds on the natarajan dimensions of some function classes, 2023. arXiv:2209.07015.

Jourdan, M., Rémy, D., and Emilie, K. Dealing with unknown variances in best-arm identification. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201, pp. 776–849, 2023.

Karlan, D. and Wood, D. H. The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. Working Paper 20047, National Bureau of Economic Research, April 2014.

Kasy, M. and Sautmann, A. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

Kato, M. Locally optimal fixed-budget best arm identification in two-armed gaussian bandits with unknown variances, 2024a. arXiv:2312.12741.

Kato, M. Worst-case optimal multi-armed gaussian best arm identification with a fixed budget, 2024b. arXiv:2310.19788.

Kato, M. Adaptive generalized neyman allocation: Local asymptotic minimax optimal best arm identification, 2024c.

Kato, M. and Ariu, K. The role of contextual information in best arm identification, 2021. arXiv:2106.14077.

Kato, M., Ishihara, T., Honda, J., and Narita, Y. Adaptive experimental design for efficient treatment effect estimation: Randomized allocation via contextual bandit algorithm, 2020. arXiv:2002.05308.

Kato, M., McAlinn, K., and Yasui, S. The adaptive doubly robust estimator and a paradox concerning logging policy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kato, M., Imaizumi, M., Ishihara, T., and Kitagawa, T. Asymptotically optimal fixed-budget best arm identification with variance-dependent bounds, 2023. arXiv:2302.02988v2.

Kaufmann, E. *Contributions to the Optimal Solution of Several Bandits Problems*. Habilitation á Diriger des Recherches, Université de Lille, 2020.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Kim, W., Kim, G.-S., and Paik, M. C. Doubly robust thompson sampling with linear payoffs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Kitagawa, T. and Tetenov, A. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Kock, A. B., Preinerstorfer, D., and Veliyev, B. Treatment recommendation with distributional targets. *Journal of Econometrics*, 234(2):624–646, 2023.

Komiyama, J., Tsuchiya, T., and Honda, J. Minimax optimal algorithms for fixed-budget best arm identification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Komiyama, J., Ariu, K., Kato, M., and Qin, C. Rate-optimal bayesian simple regret in best arm identification. *Mathematics of Operations Research*, 2023.

Lai, T. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 1985.

Lalitha, A., Kalantari, K., Ma, Y., Deoras, A., and Kveton, B. Fixed-budget best-arm identification with heterogeneous reward variances. In *Conference on Uncertainty in Artificial Intelligence*, 2023.

Le Cam, L. *Locally Asymptotically Normal Families of Distributions. Certain Approximations to Families of Distributions and Their Use in the Theory of Estimation and Testing Hypotheses*. University of California Publications in Statistics. vol. 3. no. 2. Berkeley & Los Angeles, 1960.

Le Cam, L. Limits of experiments. In *Theory of Statistics*, pp. 245–282. University of California Press, 1972.

Le Cam, L. *Asymptotic Methods in Statistical Decision Theory (Springer Series in Statistics)*. Springer, 1986.

Lehmann, E. L. and Casella, G. *Theory of Point Estimation*. Springer-Verlag, 1998.

Manski, C. Identification problems and decisions under ambiguity: Empirical analysis of treatment response and normative analysis of treatment choice. *Journal of Econometrics*, 95(2):415–442, 2000.

Manski, C. F. Treatment choice under ambiguity induced by inferential problems. *Journal of Statistical Planning and Inference*, 105(1):67–82, 2002.

Manski, C. F. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004.

Manski, C. F. and Tetenov, A. Sufficient trial size to inform clinical practice. *Proceedings of the National Academy of Sciences*, 113(38):10518–10523, 2016.

Masoudian, S. and Seldin, Y. Improved analysis of the tsallis-inf algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory (COLT)*, 2021.

Natarajan, B. K. On learning sets and functions. *Machine Learning*, 4(1):67–97, Oct 1989. ISSN 1573-0565. doi: 10.1007/BF00114804.

Neyman, J. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science*, 5:463–472, 1923.

Neyman, J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97: 123–150, 1934.

Qian, W. and Yang, Y. Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*, 2016.

Qin, C. Open problem: Optimal best arm identification with fixed-budget. In *Conference on Learning Theory*, 2022.

Qin, C. and Russo, D. Adaptivity and confounding in multi-armed bandit experiments, 2022. arXiv:2202.09036.

Qin, C., Klabjan, D., and Russo, D. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Rakhlin, A., Sridharan, K., and Tewari, A. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1):111–153, Feb 2015. ISSN 1432-2064. doi: 10.1007/s00440-013-0545-5.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 1974.

Russac, Y., Katsimerou, C., Bohle, D., Cappé, O., Garivier, A., and Koolen, W. M. A/b/n testing with control in the presence of subpopulations. In *NeurIPS*, 2021.

Russo, D. Simple bayesian algorithms for best arm identification, 2016. arXiv:1602.08448.

Sauro, L. Rapidly finding the best arm using variance. In *European Conference on Artificial Intelligence*, 2020.

Schlag, K. H. Eleven% designing randomized experiments under minimax regret. *Unpublished manuscript, European University Institute, Florence*, 2007.

Shang, X., de Heide, R., Menard, P., Kaufmann, E., and Valko, M. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 1823–1832, 2020.

Simchi-Levi, D., Wang, C., and Xu, J. On experimentation with heterogeneous subgroups: An asymptotic optimal δ-weighted-pac design, 2024. SSRN:4721755.

Stoye, J. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1): 70–81, 2009.

Stoye, J. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 939–941. Association for Computing Machinery, 2015.

Tabord-Meehan, M. Stratification Trees for Adaptive Randomization in Randomized Controlled Trials. *The Review of Economic Studies*, 12 2022.

Tekin, C. and van der Schaar, M. Releaf: An algorithm for learning and exploiting relevance. *IEEE Journal of Selected Topics in Signal Processing*, 2015.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

Tsiatis, A. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007. ISBN 9780387373454. URL https://books.google.co.jp/books?id=xqZFi2EMB40C.

van der Laan, M. J. The construction and analysis of adaptive group sequential designs. https://biostats.bepress.com/ucbbiostat/paper232, 2008.

van der Vaart, A. An asymptotic representation theorem. *International Statistical Review / Revue Internationale de Statistique*, 59(1):97–121, 1991.

van der Vaart, A. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Viviano, D. Experimental design under network interference, 2022. arXiv:2003.08421.

Wager, S. and Xu, K. Diffusion asymptotics for sequential experiments, 2021. arXiv:2101.09855.

Wald, A. Foundations of a general theory of sequential decision functions. *Econometrica*, 15(4):279–313, 1947.

Wald, A. Statistical Decision Functions. *The Annals of Mathematical Statistics*, 20(2):165 – 205, 1949. doi: 10.1214/aoms/1177730030. URL https://doi.org/10.1214/aoms/1177730030.

Wang, P.-A., Ariu, K., and Proutiere, A. On uniformly optimal algorithms for best arm identification in two-armed bandits with fixed budget, 2023. arXiv:2308.12000.

Yang, J. and Tan, V. Minimax optimal fixed-budget best arm identification in linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Yang, Y. and Zhu, D. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Annals of Statistics*, 30(1):100–121, 2002.

Zhan, R., Ren, Z., Athey, S., and Zhou, Z. Policy learning with adaptively collected data, 2022. arXiv:2105.02344.

Zheng, W. and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics. Springer-Verlag New York, 2011.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(1), 2021.

A   Related Work

The MAB problem has been explored as an instance of the sequential decision-making problem (Thompson, 1933; Robbins, 1952; Lai & Robbins, 1985), where BAI is a paradigm within this context (Even-Dar et al., 2006; Audibert et al., 2010; Bubeck et al., 2011).

BAI and ordinal optimization.   The study of BAI can be traced back to sequential testing, ranking, and selection problems in the 1940s (Wald, 1947; Bechhofer et al., 1968). Subsequent studies in operations research, particularly in the realm of ordinal optimization, have garnered considerable attention (Chen et al., 2000; Glynn & Juneja, 2004). These studies focus on devising optimal strategies under the assumption of known distributional parameters. The machine learning community has reframed the problem as the BAI problem, placing a specific emphasis on estimating unknown distributions (Jennison et al., 1982; Even-Dar et al., 2006; Audibert et al., 2010; Bubeck et al., 2011).

Audibert et al. (2010) propose the UCB-E and Successive Rejects (SR) strategies. Bubeck et al. (2011) demonstrates minimax optimal strategies for expected simple regret in a non-asymptotic setting by extending the minimax lower bound of Auer et al. (2002). Carpentier & Locatelli (2016) further enhances the minimax lower bound, showing the optimality of Audibert et al. (2010)'s methods in terms of leading factors in expected simple regret. Based on their lower bound, Yang & Tan (2022) proposes minimax optimal linear fixed-budget BAI.

In addition to minimax evaluation, Komiyama et al. (2023) develop an optimal strategy whose upper bound for simple Bayesian regret lower bound aligns with their derived lower bound. Atsidakou et al. (2023) propose a Bayes optimal strategy for minimizing the probability of misidentification, revealing a surprising result that a $1/\sqrt{T}$-factor dominates the evaluation.

Russo (2016), Qin et al. (2017), and Shang et al. (2020) propose Bayesian BAI strategies that are optimal in terms of posterior convergence rate. Kasy & Sautmann (2021) and Ariu et al. (2021) discuss that such optimality does not necessarily guarantee asymptotic optimality for the probability of misidentification in fixed-budget BAI.

In contrast to the approaches of Bubeck et al. (2011) and Carpentier & Locatelli (2016), regarding asymptotic optimality, Kaufmann et al. (2016) derives distribution-dependent lower bounds for BAI with fixed confidence and a fixed budget, based on change-of-measure arguments and building upon the work of Lai & Robbins (1985). Following their work, Garivier & Kaufmann (2016) proposes an optimal strategy for BAI with fixed confidence; however, in the fixed-budget setting, there is currently a lack of strategies whose upper bound matches the lower bound established by Kaufmann et al. (2016). This issue has been discussed by Kaufmann (2020), Ariu et al. (2021), Qin (2022), Degenne (2023), Kato (2024b), and Wang et al. (2023).

Kock et al. (2023) generalizes the results of Bubeck et al. (2011) for the case where the parameter of interest is a functional of the distribution and finds that, in contrast to the results Bubeck et al. (2011), the target allocation ratio is not uniform.

The problem of BAI with contextual information is still under investigation. For example, Tekin & van der Schaar (2015), Guan & Jiang (2018), Deshmukh et al. (2018), Kato & Ariu (2021), and Qin & Russo (2022) consider this problem, but their analyses and settings differ from those employed in this study.

Our proposed strategy assigns treatment arms with a probability depending on their variances. Variance-dependent BAI has been explored by Chen et al. (2000), Glynn & Juneja (2004), Kaufmann et al. (2016), Sauro (2020), Jourdan et al. (2023), Kato et al. (2023), Kato (2024b,a,c), and Lalitha et al. (2023). Our choice of treatment-assignment probability is also inspired by van der Laan (2008) and Hahn et al. (2011) in adaptive experimental design for efficient treatment effect estimation.

Decision theory and treatment choice.   Beyond BAI, our study is further related to statistical decision theory (Wald, 1949). Manski (2000, 2002, 2004) extend this decision theory and introduce the treatment choice problem from a decision theory perspective, independent of BAI. They focus on recommending the best treatment arm using non-experimental, independently and identically distributed (i.i.d.) observations without adaptive experimental design (Schlag, 2007; Stoye, 2009, 2012; Manski & Tetenov, 2016; Dominitz & Manski, 2017, 2022). Hirano & Porter (2009) employs the limit experiment framework (Le Cam, 1972, 1986; Lehmann & Casella, 1998; van der Vaart,

1991, 1998) for discussing the problem of treatment choice, where the class of alternative hypotheses comprises local models, with parameters of interest converging to the true parameters at a rate of $1/\sqrt{T}$. Armstrong (2022) and Hirano & Porter (2023) apply this framework to adaptive experimental design. Adusumilli (2021, 2022) present an alternative minimax evaluation of bandit strategies for both regret minimization and BAI, based on a formulation utilizing a diffusion process proposed by Wager & Xu (2021) and the limit experiment framework (Le Cam, 1960, 1972, 1986; van der Vaart, 1991, 1998).

Policy learning. Inspired by supervised learning and statistical decision theory, various off-policy learning methods have been proposed. Swaminathan & Joachims (2015) proposes counterfactual risk minimization, an extension of empirical risk minimization for policy learning. Kitagawa & Tetenov (2018) extends counterfactual risk minimization by linking it to the viewpoint of treatment choice and proposes welfare maximization. Athey & Wager (2021) refines the method proposed by Kitagawa & Tetenov (2018). Zhou et al. (2023) shows a tight upper bound for general policy learning methods. Zhan et al. (2022) develops a policy learning method from adaptively collected observations.

For the expected simple regret, Bubeck et al. (2011) shows that the Uniform-Empirical Best Arm (EBA) strategy is minimax optimal for bandit models with bounded supports. Kock et al. (2023) extends the results to cases where parameters of interest are functionals of the distribution and finds that optimal sampling rules are not uniform. Adusumilli (2022, 2021) consider a different minimax evaluation of bandit strategies for both regret minimization and BAI problems, based on a formulation using a diffusion process, as proposed by Wager & Xu (2021).

Other related work. Efficient estimation of ATE via adaptive experiments constitutes another area of related literature. van der Laan (2008) and Hahn et al. (2011) propose experimental design methods to estimate ATE more efficiently by using covariate information in treatment assignments. Karlan & Wood (2014) examine donors' responses to new information by applying the method of Hahn et al. (2011). Subsequently, Tabord-Meehan (2022) and Kato et al. (2020) attempt to improve these studies, and more recently, Gupta et al. (2021) proposes the use of instrumental variables in this context. Viviano (2022) explores experimental designs for network inference.

We employ the Augmented Inverse Probability Weighting (AIPW) estimator in policy learning. The AIPW estimator has been extensively used in the fields of causal inference and semiparametric inference (Tsiatis, 2007; Bang & Robins, 2005; Chernozhukov et al., 2018). More recently, it has also been utilized in other MAB problems, as seen in Kim et al. (2021), Ito et al. (2022), Zimmert & Seldin (2021), and Masoudian & Seldin (2021).

B   Proof of the Minimax Lower Bounds (Theorems 3.3 and 3.4)

In this section, we establish the proofs for Theorems 3.3 and 3.4. Initially, we focus on proving Theorem 3.3, which presents a lower bound applicable for cases where $K \geq 2$. Following this, Section B.4 introduces a refined lower bound specifically for $K = 2$. While the initial lower bound from Theorem 3.3 is valid for $K = 2$, Theorem 3.4 offers a tighter lower bound.

B.1   Transportation Lemma

Let $f_P^a(y^a \mid s)$ be a density of Y^a conditional on $X = s$ under P. Let $\zeta_P(s)$ be a density of X under P.

Kaufmann et al. (2016) derives the following result based on a change-of-measure argument, which is the principal tool in our lower bound. Let us define a density of $(Y^1, Y^2, \ldots, Y^K, X)$ under a bandit model $P \in \mathcal{P}$ as

$$p(y^1, y^2, \ldots, y^K, s) = \prod_{a \in [K]} f_P^a(y^a \mid s)\zeta_P(s).$$

Between two bandit models $P, Q \in \mathcal{P}$, following the proof of Lemma 1 in Kaufmann et al. (2016), referred to as the *transportation* lemma, we define the log-likelihood ratio of a sequence of observations

$(X_t, Y_t, A_t)_{t=1}^T$ under a given strategy as

$$L_T(P, Q) = \sum_{t=1}^T \sum_{a \in [K]} \mathbb{1}[A_t = a] \log \left(\frac{f_P^a(Y_t^a \mid X_t)}{f_Q^a(Y_t^a \mid X_t)} \right).$$

As discussed by Kaufmann et al. (2016), the transportation lemma immediately yields the following lemma.

Lemma B.1 (Lemma 1 and Remark 2 in Kaufmann et al. (2016))**.** *Suppose that for any two bandit model $P, Q \in \mathcal{P}$ with K treatment arms and for all $a \in [K]$, the distributions P^a and Q^a are mutually absolutely continuous, where P^a and Q^a are distributions of (Y^a, X) under P and Q, respectively. Then, for any $a \in [K]$ and $x \in \mathcal{X}$, any strategy satisfies*

$$\left| \mathbb{P}_P(\widehat{a}(x) = a) - \mathbb{P}_Q(\widehat{a}(x) = a) \right| \leq \sqrt{\frac{\mathbb{E}_P\left[L_T(P, Q)\right]}{2}}$$

B.2 Restricted Bandit Models

Fix any policy class Π with Natarajan dimension M. We can find a set of M points $\mathcal{S} := \{s_1, s_2, \ldots, s_M\}$ that are shattered by Π.

Gaussian bandit models. We choose a specific ζ, a marginal distribution on X, and define a specific subclass $\mathcal{P}^\dagger$ of a class of location-shift bandit models to derive a lower bound. Let ζ be a distribution on $\mathcal{X}$ such that $\operatorname{supp} \zeta = \mathcal{S}$ and $\operatorname{Pr}_\zeta(X = s) = 1/M$ for any $s \in \mathcal{S}$. We focus on Gaussian bandit models, where outcomes follow Gaussian distributions conditional on contexts. A subclass $\mathcal{P}^\dagger \subseteq \mathcal{P}$ is defined as follows:

$$\mathcal{P}^\dagger := \left\{ P \in \mathcal{P} : \forall a \in [K] \ \forall s \in \mathcal{S} \ \left[Y^a \mid X = s \sim \mathcal{N}\left(\mu^a(s), (\sigma^a(s))^2 \right), \ \mu^a(s) \in \mathbb{R} \right], \ \operatorname{marg}_{\mathcal{X}}(P) = \zeta \right\},$$

where $\sigma^a \colon \mathcal{X} \to \mathbb{R}_+$ is given (introduced in Definition 2.1).

Alternative hypothesis The set of alternative hypotheses $\mathcal{Q}^\dagger \subseteq \mathcal{P}^\dagger$ is defined as follows:

$$\mathcal{Q}^\dagger := \left\{ P \in \mathcal{P}^\dagger : \ \forall s \in \mathcal{S}, \ m^s \in \mathbb{R}, \ d(s) \in [K], \ \Delta^{d(s)} > 0, \right.$$

$$\forall s \in \mathcal{S}, \ \mu^{d(s)}(P)(s) = m^s + \Delta^{d(s)}(s),$$

$$\left. \forall s \in \mathcal{S}, \ \forall b \in [K] \backslash \{d(s)\} \ \mu^b(P)(s) = m^s \right\}.$$

Note that a distribution in $\mathcal{Q}^\dagger$ is characterized by a parameter (d, Δ, m), where $d = (d(s))_s$, $\Delta = (\Delta^{d(s)})_s$, and $m = (m^s)_s$. We denote a typical element of $\mathcal{Q}^\dagger$ by $Q_{d,\Delta,m}$.

Next, we will define a distribution $P_{d,\Delta,m}^{\sharp,s} \in \mathcal{P}^\dagger$ by

$$\forall a \in [K], \mu^a \left(P_{d,\Delta,m}^{\sharp,s} \right)(s) = m^s,$$

$$\forall s' \in \mathcal{S} \setminus \{s\} \forall a \in [K], \mu^a \left(P_{d,\Delta,m}^{\sharp,s} \right)(s') = \mu^a \left(Q_{d,\Delta,m} \right)(s').$$

Note that, since $P_{d,\Delta,m}^{\sharp,s} \in \mathcal{P}^\dagger$, it is characterized once we fix the conditional means for each arm-context pair. We write $P_{d,\Delta,m}^{\sharp,s}$ as $P^{\sharp,s}$ when (d, Δ, m) is clear from the context.

Change of measure. For any $Q_{d,\Delta,m} \in \mathcal{Q}^\dagger$ and for each $s \in \mathcal{S}$, the following equation holds:

$$L_T(P^{\sharp,s}, Q_{d,\Delta,m}) = \sum_{t=1}^T \sum_{a \in [K]} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f_{P^{\sharp,s}}^a(Y_t^a \mid X_t)}{f_{Q_{d,\Delta,m}}^a(Y_t^a \mid X_t)} \right) \right\}$$

17

$$
= \sum_{t=1}^{T} \sum_{a \in [K]} \sum_{s' \in \mathcal{S}} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f_{P^{\sharp,s}}^{a}(Y_t^a \mid s')}{f_{Q_{d,\Delta,m}}^{a}(Y_t^a \mid s')} \right) \right\} \mathbb{1}[X_t = s']
$$

$$
\overset{(i)}{=} \sum_{t=1}^{T} \sum_{a \in [K]} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f_{P^{\sharp,s}}^{a}(Y_t^a \mid s)}{f_{Q_{d,\Delta,m}}^{a}(Y_t^a \mid s)} \right) \right\} \mathbb{1}[X_t = s]
$$

$$
\overset{(ii)}{=} \sum_{t=1}^{T} \left\{ \mathbb{1}[A_t = d(s)] \log \left(\frac{f_{P^{\sharp,s}}^{d(s)}(Y_t^{d(s)} \mid s)}{f_{Q_{d,\Delta,m}}^{d(s)}(Y_t^{d(s)} \mid s)} \right) \right\} \mathbb{1}[X_t = s].
$$

In $\overset{(i)}{=}$, we used $\frac{f_{P^{\sharp,s}}^{a}(Y_t^a \mid s')}{f_{Q_{d,\Delta,m}}^{a}(Y_t^a \mid s')} = 1$ for $s' \neq s$ from the definition of $P^{\sharp,s}$. In $\overset{(ii)}{=}$, we used the assumption that for each $b \in [K] \backslash \{d(s)\}$, it holds that $f_{P^{\sharp,s}}^{b}(y \mid s) = f_{Q_{d,\Delta,m}}^{b}(y \mid s)$.

Given a strategy, its *assignment ratio* $w \colon [K] \times \mathcal{X} \to [0,1]$ under P is defined as follows:

$$
w(a \mid x) := \mathbb{E}_P \left[\sum_{t=1}^{T} \mathbb{1}\{A_t = a\} \mid X = x \right].
$$

For submodel $\mathcal{Q}^{\dagger}$, the following lemma holds.

Lemma B.2. *For any $s \in \mathcal{S}$, $P^{\sharp,s} \in \mathcal{P}^{\dagger}$, $Q_{d,\Delta,m} \in \mathcal{Q}^{\dagger}$, and $T \in \mathbb{N}$, the following equality holds:*

$$
\mathbb{E}_{P^{\sharp,s}} \left[L_T(P^{\sharp,s}, Q_{d,\Delta,m}) \right] = \frac{T}{2M} \frac{\left(\Delta^{d(s)}(s) \right)^2}{\frac{\left(\sigma^{d(s)}(s) \right)^2}{w(d(s) \mid s)}}.
$$

Proof. We have

$$
\mathbb{E}_{P^{\sharp,s}} \left[L_T(P^{\sharp,s}, Q_{d,\Delta,m}) \right] = \mathbb{E}_{P^{\sharp,s}} \left[\sum_{t=1}^{T} \left\{ \mathbb{1}[A_t = d(s)] \log \left(\frac{f_{P^{\sharp,s}}^{d(s)}(Y_t^{d(s)} \mid s)}{f_{Q_{d,\Delta,m}}^{d(s)}(Y_t^{d(s)} \mid s)} \right) \right\} \mathbb{1}[X_t = s] \right]
$$

$$
= \sum_{t=1}^{T} \mathbb{E}_{P^{\sharp,s}} \left[\mathbb{E}_{P^{\sharp,s}} \left[\left\{ \mathbb{1}[A_t = d(s)] \log \left(\frac{f_{P^{\sharp,s}}^{d(s)}(Y_t^{d(s)} \mid s)}{f_{Q_{d,\Delta,m}}^{d(s)}(Y_t^{d(s)} \mid s)} \right) \right\} \mid X_t = s \right] \mathbb{1}[X_t = s] \right]
$$

$$
= \sum_{t=1}^{T} \mathbb{E}_{P^{\sharp,s}} \left[w(d(s) \mid s) \mathbb{E}_{P^{\sharp,s}} \left[\log \left(\frac{f_{P^{\sharp,s}}^{d(s)}(Y_t^{d(s)} \mid s)}{f_{Q_{d,\Delta,m}}^{d(s)}(Y_t^{d(s)} \mid s)} \right) \mid X_t = s \right] \mathbb{1}[X_t = s] \right]
$$

$$
= T w(d(s) \mid s) \frac{\left(\mu^{d(s)}(Q_{d,\Delta,m})(s) - \mu^{d(s)}(P^{\sharp,s})(s) \right)^2}{2 \left(\sigma^{d(s)}(s) \right)^2} / M
$$

$$
= \frac{T}{2M} \frac{\left(\Delta^{d(s)}(s) \right)^2}{\left(\sigma^{d(s)}(s) \right)^2 / w(d(s) \mid s)}.
$$

To show the second last equality, we use

$$
\mathbb{E}_{P^{\sharp,s}} \left[\log \left(\frac{f_{P^{\sharp,s}}^{d(s)}(Y_t^{d(s)} \mid s)}{f_{Q_{d,\Delta,m}}^{d(s)}(Y_t^{d(s)} \mid s)} \right) \mid X_t = s \right] = \frac{\left(\mu^{d(s)}(Q_{d,\Delta,m})(s) - \mu^{d(s)}(P^{\sharp,s})(s) \right)^2}{2 \left(\sigma^{d(s)}(s) \right)^2},
$$

which corresponds to a KL divergence between two Gaussian distributions. $\qquad \square$

B.3   Proof of Theorem 3.3: General Minimax Lower Bounds

We derive the lower bound of the expected simple regret.

Proof of Theorem 3.3. Our proof below is built on the following line of reasoning: First, suppose that nature selects a true distribution in a two-step process: Initially, nature determines $(e^a(s))_{s,a} \in$

$\mathbb{R}_+^{M \times K}$, ensuring that $\sum_a e^a(s) = 1$ for each state $s \in \mathcal{S}$. It then selects the optimal arms $d(s)$ with probability $e^{d(s)}(s)$ for each state s. Subsequently, nature chooses some (d, Δ, m), determining $Q_{d,\Delta,m} \in \mathcal{Q}^\dagger$, which represents the true distribution. [1] We then focus on the expected simple regret that the decision-maker could encounter under these strategies of nature, regardless of the strategy employed by the decision-maker. By construction, there must exist at least one distribution $Q_{d,\Delta,m} \in \mathcal{Q}^\dagger \subseteq \mathcal{P}$ that attains this regret. Hence, this value serves as a lower bound for the regret.

Fix any strategy of the decision-maker. First, observe that, for each $P \in \mathcal{P}^\dagger$, the expected simple regret can be simplified as follows:

$$R(P)(\pi) = \mathbb{E}_P \left[\sum_{a \in [K]} \left(\mu^{a^*(P)(X)}(P)(X) - \mu^a(P)(X) \right) \mathbb{P}_P \left(\widehat{a}_T(X) = a \right) \right]$$

$$= \sum_{s \in \mathcal{S}} \sum_{a \in [K]} \left(\mu^{a^*(P)(s)}(P)(s) - \mu^a(P)(s) \right) \mathbb{P}_P \left(\widehat{a}_T(s) = a \right) \mathbb{P}_P \left(X = s \right)$$

$$= \sum_{s \in \mathcal{S}} \sum_{a \in [K]} \left(\mu^{a^*(P)(s)}(P)(s) - \mu^a(P)(s) \right) \mathbb{P}_P \left(\widehat{a}_T(s) = a \right) / M.$$

Next, we decompose the expected simple regret by using the definition of $\mathcal{P}^\dagger$: fix any $(e^a(s))_{s,a}$ and (d, Δ, m). Then, the expected simple regret the decision-maker experiences is:

$$\sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \sum_{b \in [K] \setminus \{d(s)\}} \left(\mu^{d(s)}(Q_{d,\Delta,m})(s) - \mu^b(Q_{d,\Delta,m})(s) \right) \mathbb{P}_{Q_{d,\Delta,m}} \left(\widehat{a}_T(s) = b \right) / M$$

$$= \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \sum_{b \in [K] \setminus \{d(s)\}} \Delta^{d(s)}(s) \mathbb{P}_{Q_{d,\Delta,m}} \left(\widehat{a}_T(s) = b \right) \right\} / M$$

$$= \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \Delta^{d(s)}(s) \mathbb{P}_{Q_{d,\Delta,m}} \left(\widehat{a}_T(s) \neq d(s) \right) \right\} / M$$

$$= \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \Delta^{d(s)}(s) \left(1 - \mathbb{P}_{Q_{d,\Delta,m}} \left(\widehat{a}_T(s) = d(s) \right) \right) \right\} / M, \tag{4}$$

Note that, as we saw in the beginning of the proof, the regret lower bound is bounded from below by (4).

From Lemma B.1. and the definition of null consistent strategies, we have

$$(4) = \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \Delta^{d(s)} \left(1 - \mathbb{P}_{P^{\sharp,s}} \left(\widehat{a}_T(s) = d(s) \right) + \mathbb{P}_{P^{\sharp,s}} \left(\widehat{a}_T(s) = d(s) \right) - \mathbb{P}_{Q_{d,\Delta,m}} \left(\widehat{a}_T(s) = d(s) \right) \right) \right\} / M$$

$$\geq \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \Delta^{d(s)}(s) \left\{ 1 - \mathbb{P}_{P^{\sharp,s}} \left(\widehat{a}_T(s) = d(s) \right) - \sqrt{\frac{\mathbb{E}_{P^{\sharp,s}} \left[L_T^{d(s)}(P^{\sharp,s}, Q_{d,\Delta,m}) \right]}{2}} \right\} \right\} / M. \tag{5}$$

Since we assume that the strategy is null consistent, we have $\mathbb{P}_{P^{\sharp,s}} \left(\widehat{a}_T(s) = d(s) \right) = 1/K + o(1)$. By Lemma B.2, we obtain

$$(5) = \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \Delta^{d(s)}(s) \left\{ 1 - \frac{1}{K} - \sqrt{\frac{T(\Delta^{d(s)}(s))^2}{4M \frac{\left(\sigma^{d(s)}(s) \right)^2}{w(d(s)|s)}}} \right\} / M + o(1).$$

---

[1] Note that the choice of $(m_s)_{s \in \mathcal{S}}$ does not affect any objects in the proof of Theorem 3.3.

Let

$$\mathcal{E} := \left\{ (e^d(s))_{d,s} \in \mathbb{R}^{[K] \times \mathcal{S}} : e^d(s) \in [0,1], \ \forall s \in \mathcal{S}, \ \sum_{d \in [K]} e^d(s) = 1 \right\},$$

and denote its typical element by e. In principle, if we take the supremum with respect to $e \in \mathcal{E}$ and $(\Delta^{d(s)}(s))_s \in \mathbb{R}_{>0}^M$ in (5), that will be a regret lower bound. By substituting $\Delta^{d(s)}(s) = \sqrt{M \frac{(\sigma^{d(s)}(s))^2}{w(d(s)|s)}/2T}$, taking the infimum with respect to w, and taking the supremum with respect to $(e^a(s))_{a,s}$, we obtain the following regret lower bound, which is lower than the lower bound obtained by taking the supremum with respect to e and Δ:

$$\frac{1}{8} \frac{1}{\sqrt{MT}} \sum_{s \in \mathcal{S}} \sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \sqrt{\frac{(\sigma^{d(s)}(s))^2}{w(d(s) \mid s)}} \right\} + o(1). \tag{6}$$

Fix any $s \in \mathcal{S}$. Let $h(e, w) := \sum_{d \in [K]} e^d(s) \sqrt{\frac{(\sigma^d(s))^2}{w(d|s)}}$ and consider $\sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} h(e, w)$. Note that h is concave in e, convex in w, and continuous in (e, w). Denote the closure of $\mathcal{W}$ by $\overline{\mathcal{W}}$. First, observe that

$$\sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} h(e, w) \geq \sup_{e \in \mathcal{E}} \inf_{w \in \overline{\mathcal{W}}} h(e, w).$$

Since $\overline{\mathcal{W}}$ is compact and $h(e, \cdot)$ is continuous for each e, we have

$$\sup_{e \in \mathcal{E}} \inf_{w \in \overline{\mathcal{W}}} h(e, w) = \sup_{e \in \mathcal{E}} \min_{w \in \overline{\mathcal{W}}} h(e, w).$$

By Berge's Maximum Theorem, $\min_{w \in \overline{\mathcal{W}}} h(e, w)$ is continuous in w. Since $\mathcal{E}$ is compact, we have

$$\sup_{e \in \mathcal{E}} \min_{w \in \overline{\mathcal{W}}} h(e, w) = \max_{e \in \mathcal{E}} \min_{w \in \overline{\mathcal{W}}} h(e, w).$$

Since $\mathcal{E}$ and $\overline{\mathcal{W}}$ are both compact and convex and h is concave-convex, by the minimax theorem, we have

$$\max_{e \in \mathcal{E}} \min_{w \in \overline{\mathcal{W}}} h(e, w) = \min_{w \in \overline{\mathcal{W}}} \max_{e \in \mathcal{E}} h(e, w).$$

Combining these results, we have

$$\sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} h(e, w) \geq \min_{w \in \overline{\mathcal{W}}} \max_{e \in \mathcal{E}} h(e, w).$$

Let us consider $\max_{e \in \mathcal{E}} h(e, w)$ for a fixed w. At the optimum, we have $e^d(s) = 1$ iff $d \in \arg\max_{d \in [K]} \sqrt{\frac{(\sigma^d(s))^2}{w(d|s)}}$. Thus, we have

$$\min_{w \in \overline{\mathcal{W}}} \max_{e \in \mathcal{E}} h(e, w) = \min_{w \in \overline{\mathcal{W}}} \max_{d \in [K]} \sqrt{\frac{(\sigma^d(s))^2}{w(d \mid s)}}. \tag{7}$$

For each s, we consider the following constrained optimization:

$$\min_{R \geq 0, w \in \overline{\mathcal{W}}} R$$

$$\text{s.t.} \quad R \geq \sqrt{\frac{(\sigma^d(s))^2}{w(d \mid s)}} \quad \forall d \in [K]$$

Note that if (R^*, w^*) is the optimal solution to this problem, w^* is also the optimal solution to (7). After some algebra, we can show that

$$w^*(d \mid s) = \frac{(\sigma^d(s))^2}{\sum_{a \in [K]} (\sigma^a(s))^2}$$

for each $d \in [K]$.

Thus, we have

$$(7) = \sqrt{\sum_{a \in [K]} (\sigma^a(s))^2},$$

and hence

$$(6) \geq \frac{1}{8} \frac{1}{\sqrt{TM}} \sum_{s \in \mathcal{S}} \sqrt{\sum_{a \in [K]} (\sigma^a(s))^2} = \frac{1}{8} \frac{1}{\sqrt{T}} \sum_{s \in \mathcal{S}} \sqrt{M \sum_{a \in [K]} (\sigma^a(s))^2} \frac{1}{M} + o(1)$$

$$= \frac{1}{8} \frac{1}{\sqrt{T}} \mathbb{E}_{X \sim \varsigma} \left[\sqrt{M \sum_{a \in [K]} (\sigma^a(s))^2} \right] + o(1). \tag{8}$$

The last equality follows since we choose ξ so that ξ puts equal probabilities on $\mathcal{S}$. This implies that, for any strategy of the decision-maker, there exists a distribution in $\mathcal{P}$ under which the simple regret is lower bounded by (8).

$\square$

Although this lower bound is applicable to a case with $K = 2$, we can tighten the lower bound by changing the definiton of the parametric submodel.

B.4   Refined Minimax Lower Bounds for Two-armed Bandits (Proof of Theorem 3.4)

When $K = 2$, we can derive a tighter lower bound. As in previous sections, we choose $\xi \in \mathcal{M}(\mathcal{X})$ such that $\Pr_{\xi}(X = s) = 1/M$ for each $s \in \mathcal{S}$.

Change of measure.   For any $Q_{d,\Delta,m} \in \mathcal{Q}^{\dagger}$ and $s \in \mathcal{S}$, the following equation holds:

$$L_T(P^{\sharp,s}, Q_{d,\Delta,m}) = \sum_{t=1}^{T} \sum_{a \in [K]} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f^a_{P^{\sharp,s}}(Y^a_t \mid X_t)}{f^a_{Q_{d,\Delta,m}}(Y^a_t \mid X_t)} \right) \right\}$$

$$= \sum_{t=1}^{T} \sum_{a \in [K]} \sum_{s' \in \mathcal{S}} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f^a_{P^{\sharp,s}}(Y^a_t \mid s')}{f^a_{Q_{d,\Delta,m}}(Y^a_t \mid s')} \right) \right\} \mathbb{1}[X_t = s']$$

$$\overset{(i)}{=} \sum_{t=1}^{T} \sum_{a \in [K]} \left\{ \mathbb{1}[A_t = a] \log \left(\frac{f^a_{P^{\sharp,s}}(Y^a_t \mid s)}{f^a_{Q_{d,\Delta,m}}(Y^a_t \mid s)} \right) \right\} \mathbb{1}[X_t = s].$$

In $\overset{(i)}{=}$, we used $\frac{f^a_{P^{\sharp,s}}(Y^a_t \mid s')}{f^a_{Q_{d,\Delta,m}}(Y^a_t \mid s')} = 1$ for $s' \neq s$ from the definition of $P^{\sharp,s}$.

Proof of Theorem 3.4.   By the same argument as in Section B.3, we have the following lower bound for the simple regret (cf. equation (5)):

$$\sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \Delta^{d(s)}(s) \left\{ 1 - \mathbb{P}_{P^{\sharp,s}}\left(\widehat{a}_T(s) = d(s)\right) - \sqrt{\frac{\mathbb{E}_{P^{\sharp,s}}\left[L_T^{d(s)}(P^{\sharp,s}, Q_{d,\Delta,m}) \right]}{2}} \right\} \right\} / M. \tag{9}$$

By the same argument as in the proof of Lemma B.2, we have

$$\mathbb{E}_{P^{\sharp,s}}\left[L_T(P^{\sharp,s}, Q_{d,\Delta,m}) \right] = \frac{T}{2M} \left\{ w(1 \mid s) \frac{\left(\mu^1(Q_{d,\Delta,m})(s) - m^s\right)^2}{(\sigma^1(s))^2} + w(2 \mid s) \frac{\left(\mu^2(Q_{d,\Delta,m})(s) - m^s\right)^2}{(\sigma^2(s))^2} \right\}.$$

21

We consider the following optimization problem with respect to m^s:

$$\min_{m_s \in \mathbb{R}} \left\{ w(1 \mid s) \frac{\left(\mu^1(Q_{d,\Delta,m})(s) - m^s\right)^2}{(\sigma^1(s))^2} + w(2 \mid s) \frac{\left(\mu^2(Q_{d,\Delta,m})(s) - m^s\right)^2}{(\sigma^2(s))^2} \right\}.$$

The solution is

$$m^s = \frac{c^1 \mu^1(Q_{d,\Delta,m})(s) + c^2 \mu^2(Q_{d,\Delta,m})(s)}{c^1 + c^2},$$

where

$$c^a = \frac{w(a \mid s)}{(\sigma^a(s))^2},$$

and, the optimal value is

$$\left(\Delta^{d(s)}\right)^2 \left[\frac{(\sigma^1(s))^2}{w(1 \mid s)} + \frac{(\sigma^2(s))^2}{w(2 \mid s)}\right]^{-1}. \tag{10}$$

Since we assume a null consistent strategy, we have $\mathbb{P}_{P^{\sharp,s}}\left(\hat{a}_T(s) = d(s)\right) = 1/K + o(1)$. By (9) and (10), we have the following regret lower bound:

$$\inf_{w \in \mathcal{W}} \sup_{\substack{\Delta^{d(s)} \in (0,\infty) \\ \text{for } s \in \mathcal{S}}} \sum_{s \in \mathcal{S}} \sum_{d(s) \in [K]} e^{d(s)}(s) \Delta^{d(s)} \left\{ 1 - \frac{1}{K} - \sqrt{\frac{T\Delta^2(s)}{4M\left\{\frac{(\sigma^1(s))^2}{w(1|s)} + \frac{(\sigma^2(s))^2}{w(2|s)}\right\}}} \right\} /M + o(1).$$

By substituting $\Delta^{d(s)} = \sqrt{M\left\{\frac{(\sigma^1(s))^2}{w(1|s)} + \frac{(\sigma^2(s))^2}{w(2|s)}\right\}/2T}$, we obtain the following regret lower bound:

$$\frac{1}{8} \frac{1}{\sqrt{MT}} \sum_{s \in \mathcal{S}} \sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} \sum_{d(s) \in [K]} e^{d(s)}(s) \left\{ \sqrt{\frac{(\sigma^1(s))^2}{w(1 \mid s)} + \frac{(\sigma^2(s))^2}{w(2 \mid s)}} \right\} + o(1)$$

$$= \frac{1}{8} \frac{1}{\sqrt{MT}} \sum_{s \in \mathcal{S}} \sup_{e \in \mathcal{E}} \inf_{w \in \mathcal{W}} \left\{ \sqrt{\frac{(\sigma^1(s))^2}{w(1 \mid s)} + \frac{(\sigma^2(s))^2}{w(2 \mid s)}} \right\} \sum_{d(s) \in [K]} e^{d(s)}(s) + o(1)$$

$$= \frac{1}{8} \frac{1}{\sqrt{MT}} \sum_{s \in \mathcal{S}} \inf_{w \in \mathcal{W}} \left\{ \sqrt{\frac{(\sigma^1(s))^2}{w(1 \mid s)} + \frac{(\sigma^2(s))^2}{w(2 \mid s)}} \right\} + o(1).$$

Consider the following optimization problem:

$$\min_{w \in \mathcal{W}} \sqrt{\frac{(\sigma^1(s))^2}{w(1 \mid s)} + \frac{(\sigma^2(s))^2}{w(2 \mid s)}}.$$

The solution is

$$w(a \mid s) = \frac{\sigma^a(s)}{\sigma^1(s) + \sigma^2(s)},$$

and the optimal value is

$$\sqrt{\left(\sigma^1(s) + \sigma^2(s)\right)^2}.$$

Therefore, we obtain the following lower bound:

$$\frac{1}{8} \frac{1}{\sqrt{MT}} \sum_{s \in \mathcal{S}} \sqrt{\left(\sigma^1(s) + \sigma^2(s)\right)^2} + o(1) = \frac{1}{8} \frac{1}{\sqrt{T}} \mathbb{E}_{X \sim \xi} \sqrt{M\left(\sigma^1(X) + \sigma^2(X)\right)^2} + o(1).$$

This completes the proof.

$\square$

22

C Proof of Theorem 5.4

In the following sections, we prove the upper bound. To show the upper bound, we aim to use the result of Zhou et al. (2023), which provides an upper bound for expected simple regret in the problem of policy learning with multiple treatment arms. Here, note that we cannot directly apply the results of Zhou et al. (2023) because they assume that the observations are i.i.d. in their study, but observations are non-i.i.d. in our study.

Zhou et al. (2023) assumes that their outcomes are bounded. Therefore, to apply their result, for analysis, let us define the following quantities:

$$\mu_c^a(P)(X) := \mathbb{E}_P[c_T(Y^a)], \quad Q_c(P)(\pi) := \mathbb{E}_{X \sim \zeta}\left[\sum_{a \in [K]} \pi(a \mid X)\mu_c^a(P)(X)\right],$$

$$\pi_c^*(P) := \arg\max_{\pi \in \Pi} Q_c(P)(\pi), \quad R_c(P)(\pi) := Q_c(P)(\pi_c^*) - Q_c(P)(\pi).$$

These quantities are used in the conditions in the main theorem and the following proof. For Q and Q_c, we state the following lemma. We omit the proof.

Lemma C.1. *There exist $0 < \alpha < 1/2$ and $U_T = T^\alpha$ such that and for any ζ, and any $P \in \mathcal{P}_\zeta$, $\sup_{\pi \in \Pi} \left|\sqrt{T}Q(P)(\pi) - \sqrt{T}Q_c(P)(\pi)\right| \to 0$ holds as $T \to \infty$.*

Because we set U_T in $c_T(\cdot)$ as $U_T \to \infty$ when $T \to \infty$, this assumptions implies that when a clipping $c_T(Y_t) = \text{thre}\left(Y_t, U_T, -U_T\right)$ asymptotically vanishes as $T \to \infty$, π_c^* approaches π^*, an optimal policy under outcomes without the clipping. For example, this assumption is satisfied by assuming sub-Gaussianity about $Y(a)$.

Thus, by clipping the outcomes using $c_T(\cdot)$ in Q_c, our problem satisfies the boundedness in Zhou et al. (2023). To circumvent the issue of non-i.i.d. observations, we show an asymptotic equivalence between our empirical policy value $\widehat{Q}_T(\pi)$ and a hypothetical empirical policy value constructed from hypothetical i.i.d. observations in Section C.1. Then, using the results of Zhou et al. (2023), we upper bound a regret for $Q_c(P)(\pi_c^*) - Q_c(P)(\widehat{\pi})$ with the hypothetical policy value in Section C.2. Finally, in Section C.3, we derive the upper bounds for $R(\pi^*) = Q(P)(\pi^*) - Q(P)(\widehat{\pi})$ from the upper bounds for $Q_c(P)(\pi_c^*) - Q_c(P)(\widehat{\pi})$.

C.1 Asymptotic Equivalence

Let $w_t(0 \mid x)$ be 0. We write $\widehat{\Gamma}_t^a$ as

$$\widehat{\Gamma}_t^a = \widehat{\Gamma}^a((Y_t^a)_{a \in [K]}, \xi_t, X_t)$$

$$= \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} \widehat{w}_t(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a} \widehat{w}_t(a \mid X_t)\right]\left(c_T(Y_t^a) - \widehat{\mu}_t^a(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t).$$

This expression equals the original definition of $\widehat{\Gamma}_t^a$. Let $w^*(0 \mid x)$ be 0. Then, we define

$$\Gamma_t^{*a}((Y_t^a)_{a \in [K]}, \xi_t, X_t)$$

$$:= \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a} w^*(a \mid X_t)\right]\left(c_T(Y_t^a) - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t).$$

Here. we show that $\widehat{\Gamma}_t^a$ and Γ_t^{*a} are asymptotically equivalent. The proof is shown in Appendix D.

Lemma C.2. *Suppose that Assumption 5.1 holds. Then,*

$$\sqrt{T}\widehat{Q}_T\left(\widehat{\pi}_T^{\text{PLAS}}\right) = \sqrt{T}\sum_{t=1}^{T}\sum_{a \in [K]} \pi(a \mid X_t)\Gamma_c^{*a}((Y_t^a)_{a \in [K]}, \xi_t, X_t) + o_P(1).$$

holds as $T \to \infty$.

Denote $\sqrt{T} \sum_{t=1}^{T} \sum_{a\in[K]} \pi(a \mid X_t)\Gamma_t^{*a}$ by $\widehat{Q}_T\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$. Note that $\widehat{Q}_T\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$ consists of only i.i.d. observations; therefore, we can directly apply the results of Zhou et al. (2023) to derive the upper bound for the policy regret.

This technique of the asymptotic equivalence is inspired by Hahn et al. (2011) and is important because, for dependent observations (even if they are martingales), we cannot apply the tools for upper-bounding regrets or risks, such as the Rademacher complexity. For instance, Zhan et al. (2022) addresses this issue and establishes an off-policy learning method from adaptively collected observations by utilizing the Rademacher complexity for martingales developed by Rakhlin et al. (2015). However, our interest lies in developing upper bounds depending on variances because our lower bounds also depend on variances, and such upper bounds are considered to be tight. Although Zhou et al. (2023) derives such upper bounds for policy learning from i.i.d. observations using the local Rademacher complexity, it is unclear whether we can use the results of Zhou et al. (2023) with the Rademacher complexity for martingales developed by Rakhlin et al. (2015). In contrast, in this study, if we restrict the problem to BAI and the evaluation metric to the worst-case expected simple regret, we show that we can apply the results of Zhou et al. (2023) and avoid the use of the Rademacher complexity for martingales by bypassing a hypothetical policy value that only depends on i.i.d. observations.

C.2 Upper Bound under I.I.D. Observations

Because $\widehat{Q}_T(\widehat{\pi}_T^{\mathrm{PLAS}})$ is asymptotically equivalent to a policy value that consists of i.i.d. observations, we can apply the results of policy learning with i.i.d. observations to bound the policy regret. Specifically, we modify a regret upper bound shown by Zhou et al. (2023), given as the following lemma.

Lemma C.3 (Modified upper bound). *Suppose that Assumptions 5.1 and 5.3 hold. Then, for any ζ and any $P \in \mathcal{P}_\zeta$,*

$$\mathbb{E}_P\left[R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right] = \mathbb{E}_P\left[Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right]$$

$$= \mathbb{E}_P\left[\left(54.4\kappa(\Pi) + 435.2\right)\frac{\Upsilon_*}{\sqrt{T}} + O\left(\frac{\sqrt{U_T}}{T^{3/4}}\right)\right]$$

holds, where

$$\Upsilon_* = \mathbb{E}\left[\sqrt{\sup_{\pi_1,\pi_2\in\Pi}\sum_{t\in[T]}\left\{\sum_{a\in[K]}\left(\pi_1(a\mid Z_i) - \pi_2(a\mid Z_i)\right)\Gamma^{*a}(Y_t,\xi_t,X_t)\right\}^2}\right].$$

The proof is shown in Appendix E.

C.3 Asymptotic Optimality

Finally, we derive the upper bounds of $R(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) = Q(P)(\pi^*) - Q(P)(\widehat{\pi}_T^{\mathrm{PLAS}})$ from the upper bounds of $R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) = Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$.

By using Lemma C.1, we can evaluate the regret $R(\pi^*)$ as follows:

$$R(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) = R(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) - R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) + R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$$

$$= Q(P)(\pi^*) - Q(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$$

$$- \left\{Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}^{\mathrm{PLAS}}\right)\right\} + \left\{Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right\}$$

$$= \left\{Q(P)(\pi^*) - Q(P)(\pi_c^*)\right\}$$

$$+ \left(\left\{Q(P)(\pi_c^*) - Q(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right\} - \left\{Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right\}\right)$$

$$+ \left\{Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right\}$$

$$= \left\{ Q_c(P)\left(\pi_c^*\right) - Q_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) \right\} + o(1/\sqrt{T}) = R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right) + o(1/\sqrt{T}).$$

Let U_T be a value that satisfies Lemma C.1. From the Cauchy-Schwarz inequality, we have

$$\Upsilon_* = \mathbb{E}\left[\sqrt{ \sup_{\pi_1,\pi_2 \in \Pi} \sum_{t \in [T]} \left\{ \sum_{a \in [K]} \left(\pi_1(a \mid X_t) - \pi_2(a \mid X_t) \right) \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 } \right]$$

$$\leq \mathbb{E}\left[\sqrt{ \sup_{\pi_1,\pi_2 \in \Pi} \sum_{t \in [T]} \left\{ \sum_{a \in [K]} \left(\pi_1(a \mid X_t) - \pi_2(a \mid X_t) \right) \right\}^2 \left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 } \right]$$

$$\leq \mathbb{E}\left[\sqrt{ \sup_{\pi_1,\pi_2 \in \Pi} \sum_{t \in [T]} 4 \left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 } \right]$$

$$= 2\mathbb{E}\left[\sqrt{ \sum_{t \in [T]} \left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 } \right].$$

From the law of iterated expectations and the Jensen inequality, we have

$$\Upsilon_* \leq 2\mathbb{E}\left[\sqrt{ \sum_{t \in [T]} \left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 } \right]$$

$$= 2 \mathbb{E}_{X \sim \zeta}\left[\sqrt{ \mathbb{E}\left[\sum_{t \in [T]} \left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right] } \right]$$

$$= 2 \mathbb{E}_{X \sim \zeta}\left[\sqrt{ T\mathbb{E}\left[\left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right] } \right].$$

In conclusion, we obtain

$$\Upsilon_*/\sqrt{T} \leq 2 \mathbb{E}_{X \sim \zeta}\left[\sqrt{ \mathbb{E}\left[\left\{ \sum_{a \in [K]} \Gamma_c^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right] } \right]$$

$$= 2 \mathbb{E}_{X \sim \zeta}\left[\sqrt{ \mathbb{E}\left[\left\{ \sum_{a \in [K]} \Gamma^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right] } \right] + o(1)$$

$$= 2 \mathbb{E}_{X \sim \zeta}\left[\sqrt{ \sum_{a \in [K]} \frac{(\sigma^a(X))^2}{w^*(a \mid X)} } \right] + o(1),$$

as $T \to \infty$ ($c_T \to \infty$).

Similarly, we obtain

$$\widetilde{\Upsilon}_*/\sqrt{T} \leq 2 \sqrt{ \sum_{a \in [K]} \mathbb{E}_{X \sim \zeta}\left[\frac{(\sigma^a(X))^2}{w^*(a \mid X)} \right] } + o(1).$$

By substituting w^* for each case with $K = 2$ and $K \geq 3$, we obtain Theorem 5.4.

D   Proof of Lemma C.2

In this section, we show

$$\sqrt{T}\widehat{Q}_T(\pi) = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\sum_{a\in[K]}\pi(a \mid X_t)\widehat{\Gamma}_t^a = \sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\sum_{a\in[K]}\pi(a \mid X_t)\Gamma_t^{*a}((Y_t^a)_{a\in[K]}, \xi_t, X_t) + o_P(1).$$

$$(11)$$

Here, recall that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(a \mid X_t)\widehat{\Gamma}_t^a((Y_t^a)_{a\in[K]}, \xi_t, X_t)$$

$$= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right](Y_t^a - \widehat{\mu}_t^a(X_t))}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\}.$$

Therefore, to show (11), we show

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right](Y_t^a - \widehat{\mu}_t^a(X_t))}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\}$$

$$= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a}w^*(b \mid X_t)\right](Y_t^a - \mu^a(P)(X_t))}{w^*(a \mid X_t)} + \mu^a(P)(X_t)\right\} + o_P(1).$$

$$(12)$$

Proof. Let us define

$$G_t\big(Y_t^a, X_t, \xi_t; \{\widehat{w}_t(b \mid X_t)\}_{b\in[K]}, \widehat{\mu}_T^a(X_t)\big)$$

$$:= \pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right](Y_t^a - \widehat{\mu}_t^a(X_t))}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\}.$$

Then, we obtain

$$G_t\big(Y_t^a, X_t, \xi_t; \{\widehat{w}_t(b \mid X_t)\}_{b\in[K]}, \widehat{\mu}_T^a(X_t)\big)$$

$$= G_t\big(Y_t^a, X_t, \xi_t; \{w^*(b \mid X_t)\}_{b\in[K]}, \mu^a(P)(X_t)\big)$$

$$\quad - G_t\big(Y_t^a, X_t, \xi_t; \{w^*(b \mid X_t)\}_{b\in[K]}, \mu^a(P)(X_t)\big) + G_t\big(Y_t^a, X_t, \xi_t; \{\widehat{w}_t(b \mid X_t)\}_{b\in[K]}, \widehat{\mu}_T^a(X_t)\big)$$

$$= G_t\big(Y_t^a, X_t, \xi_t; \{w^*(b \mid X_t)\}_{b\in[K]}, \mu^a(P)(X_t)\big) + B_t,$$

where

$$B_t := G_t\big(Y_t^a, X_t, \xi_t; \{\widehat{w}_t(b \mid X_t)\}_{b\in[K]}, \widehat{\mu}_T^a(X_t)\big) - G_t\big(Y_t^a, X_t, \xi_t; \{w^*(b \mid X_t)\}_{b\in[K]}, \mu^a(P)(X_t)\big).$$

To show (12), we consider showing $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} B_t \to 0$ as $T \to \infty$ in probability.

We show $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} B_t \to 0$ as $T \to \infty$ in probability by using the properties of martingales. First, we have

$$\mathbb{E}\left[B_t \mid X_t, \mathcal{F}_{t-1}\right]$$

26

$$= \mathbb{E}\left[\pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right]\left(Y_t^a - \widehat{\mu}_t^a(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\} \mid X_t, \mathcal{F}_{t-1}\right]$$

$$- \mathbb{E}\left[\pi(a \mid X_t)\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}w^*(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}w^*(b \mid X_t)\right]\left(Y_t^a - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t)\right\} \mid X_t, \mathcal{F}_{t-1}\right]$$

$$= \mathbb{E}\left[\pi(a \mid X_t)\left\{\frac{\widehat{w}_t(a \mid X_t)\left(Y_t^a - \widehat{\mu}_t^a(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\} \mid X_t, \mathcal{F}_{t-1}\right]$$

$$- \mathbb{E}\left[\pi(a \mid X_t)\left\{\frac{w^*(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t)\right\} \mid X_t, \mathcal{F}_{t-1}\right]$$

$$= \pi(a \mid X_t)\mu^a(P)(X_t) - \pi(a \mid X_t)\mu^a(P)(X_t)$$

$$= 0.$$

This result implies that $\{B_t\}_{t=1}^T$ is a martingale difference sequence (MDS) because $\mathbb{E}[B_t \mid X_t, \mathcal{F}_{t-1}] = \mathbb{E}[\mathbb{E}[B_t \mid X_t, \mathcal{F}_{t-1}]] = 0$ holds.

Besides, from $\widetilde{w}_t(a \mid X_t) - w^*(a \mid X_t) \xrightarrow{\text{a.s.}} 0$ and $\widehat{\mu}_t^a(X_t) - \mu^a(P)(X_t) \xrightarrow{\text{a.s.}} 0$, $\mathbb{E}\left[B_t^2 \mid X_t, \mathcal{F}_{t-1}\right]$ also converges to zero almost surely as

$$\mathbb{E}[B_t^2 \mid X_t, \mathcal{F}_{t-1}]$$

$$= \mathbb{E}\left[\pi^2(a \mid X_t)\left(\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right]\left(Y_t^a - \widehat{\mu}_t^a(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\}\right.\right.$$

$$\left.\left.- \left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}w^*(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}w^*(b \mid X_t)\right]\left(Y_t^a - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t)\right\}\right)^2 \mid X_t, \mathcal{F}_{t-1}\right]$$

$$\xrightarrow{\text{a.s.}} 0.$$

This is because from $\widehat{\mu}_t^a(X_t) - \mu^a(P)(X_t) \xrightarrow{\text{a.s.}} 0$,

$$\mathbb{E}\left[\pi^2(a \mid X_t)\left(\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right]\left(Y_t^a - \widehat{\mu}_t^a(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \widehat{\mu}_t^a(X_t)\right\}\right.\right.$$

$$\left.\left.- \left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}\widehat{w}(b \mid X_t)\right]\left(Y_t^a - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t)\right\}\right)^2 \mid X_t, \mathcal{F}_{t-1}\right]$$

$$\xrightarrow{\text{a.s.}} 0.$$

holds. Additionally, from $\widetilde{w}_t(a \mid X_t) - w^*(a \mid X_t) \xrightarrow{\text{a.s.}} 0$, for any $\varepsilon > 0$, there exists $T(\varepsilon) > 0$ such that for any $t > T(\varepsilon)$, $|\widetilde{w}_t(a \mid x) - w^*(a \mid x)| < \varepsilon$ holds for all $a \in [K]$ with probability one; that is,

$$\mathbb{E}\left[\pi^2(a \mid X_t)\left(\left\{\frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\widehat{w}_t(b \mid X_t) \le \xi_t \le \sum_{b=0}^{a}\widehat{w}_t(b \mid X_t)\right]\left(Y_t^a - \mu^a(P)(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \mu^a(P)(X_t)\right\}\right.\right.$$

$$- \left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a} w^*(b \mid X_t)\right] \left(Y_t^a - \mu^a(P)(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \mu^a(P)(X_t) \right\} \right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$$\leq \mathbb{E}\left[\left\{ \frac{\mathbb{1}\left[\min\left\{\sum_{b=0}^{a-1} \widehat{w}_t(b \mid X_t), \ \sum_{b=0}^{a-1} w^*(b \mid X_t)\right\} \leq \xi_t \leq \max\left\{\sum_{b=0}^{a-1} \widehat{w}_t(b \mid X_t), \ \sum_{b=0}^{a-1} w^*(b \mid X_t)\right\}\right]}{\widehat{w}_t^2(a \mid X_t)} \right\} \right.$$

$$\left. \times \pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$$+ \mathbb{E}\left[\left\{ \frac{\mathbb{1}\left[\min\left\{\sum_{b=0}^{a} \widehat{w}_t(b \mid X_t), \ \sum_{b=0}^{a} w^*(b \mid X_t)\right\} \leq \xi_t \leq \max\left\{\sum_{b=0}^{a} \widehat{w}_t(b \mid X_t), \ \sum_{b=0}^{a} w^*(b \mid X_t)\right\}\right]}{\widehat{w}_t^2(a \mid X_t)} \right\} \right.$$

$$\left. \times \pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$$\leq \mathbb{E}\left[\left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1}\left\{w^*(b \mid X_t) - \varepsilon\right\} \leq \xi_t \leq \sum_{b=0}^{a-1}\left\{w^*(b \mid X_t) + \varepsilon\right\}\right]}{\widehat{w}_t^2(a \mid X_t)} \right\} \pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$$+ \mathbb{E}\left[\left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a}\left\{w^*(b \mid X_t) - \varepsilon\right\} \leq \xi_t \leq \sum_{b=0}^{a}\left\{w^*(b \mid X_t) + \varepsilon\right\}\right]}{\widehat{w}_t^2(a \mid X_t)} \right\} \pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$$\leq \frac{2(a-1)\varepsilon}{\widehat{w}_t^2(a \mid X_t)} \mathbb{E}\left[\pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid \mathcal{F}_{t-1}\right] + \frac{2a\varepsilon}{\widehat{w}_t^2(a \mid X_t)} \mathbb{E}\left[\pi^2(a \mid X_t)\left(Y_t^a - \mu^a(P)(X_t)\right)^2 \mid X_t, \mathcal{F}_{t-1}\right]$$

$$\leq C\varepsilon$$

holds with probability one, where $C > 0$ is a constant independent from t. This implies that for any $\varepsilon' > 0$, there exists $T'(\varepsilon') > 0$ such that for any $t > T'(\varepsilon')$,

$$\left| \mathbb{E}\left[\pi^2(a \mid X_t)\left(\left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} \widehat{w}_t(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a} \widehat{w}_t(b \mid X_t)\right] \left(Y_t^a - \mu^a(P)(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \mu^a(P)(X_t) \right\} \right.\right.$$

$$\left.\left. - \left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a} w^*(b \mid X_t)\right] \left(Y_t^a - \mu^a(P)(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \mu^a(P)(X_t) \right\} \right)^2 \mid X_t, \mathcal{F}_{t-1} \right] \right| < \varepsilon'$$

with probability one. We also have

$$\mathbb{E}\left[\pi^2(a \mid X_t)\left(\left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a} w^*(b \mid X_t)\right] \left(Y_t^a - \mu^a(P)(X_t)\right)}{\widehat{w}_t(a \mid X_t)} + \mu^a(P)(X_t) \right\} \right.\right.$$

$$\left.\left. - \left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^{a} w^*(b \mid X_t)\right] \left(Y_t^a - \mu^a(P)(X_t)\right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t) \right\} \right)^2 \mid X_t, \mathcal{F}_{t-1} \right]$$

$\xrightarrow{\text{a.s.}} 0.$

Thus, $\mathbb{E}[B_t^2 \mid X_t, \mathcal{F}_{t-1}] \xrightarrow{\text{a.s.}} 0$ holds.

Based on these results, from Chebyshev's inequality, $\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t$ converges to zero in probability. This is because for any $v > 0$,

$$\mathbb{P}_P \left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T B_t - \mathbb{E}\left[\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \right] \right| \geq v \right) = \mathbb{P}_P \left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \right| \geq v \right) \leq \frac{\text{Var}_P \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \right)}{v^2} \tag{13}$$

Because B_t is an MDS, the covariance between B_t and B_s for $t \neq s$ is zero; that is, if $s < t$, $\text{Cov}(B_t, B_s) = \mathbb{E}[B_t B_s] = \mathbb{E}[B_s \mathbb{E}[B_t \mid \mathcal{F}_{t-1}]] = 0$.

Therefore, we can show $\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \xrightarrow{\text{P}} 0$ by showing

$$\text{Var}_P \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \right) = \frac{1}{T} \sum_{t=1}^T \text{Var}(B_t) \to 0, \tag{14}$$

as $T \to \infty$, where we used that the covariance between B_t and B_s for $t \neq s$ is zero.

To show (14), we show

$$\frac{1}{T} \sum_{t=1}^T \text{Var}_P(B_t \mid X_t, \mathcal{F}_{t-1}) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_P[B_t^2 \mid X_t, \mathcal{F}_{t-1}] \xrightarrow{\text{P}} 0,$$

by using $\mathbb{E}[B_t^2 \mid X_t, \mathcal{F}_{t-1}] \xrightarrow{\text{a.s.}} 0$.

Let u_t be $u_t = \mathbb{E}_P[B_t^2 \mid X_t, \mathcal{F}_{t-1}]$. Fix some positive $\epsilon > 0$ and $\delta > 0$. Almost-sure convergence of u_t to zero as $t \to \infty$ implies that we can find a large enough $t(\epsilon)$ such that $|u_t| < \epsilon$ for all $t \geq t(\epsilon)$ with probability at least $1 - \delta$. Let $\mathcal{E}(\epsilon)$ denote the event in which this happens; that is, $\mathcal{E}(\epsilon) = \{|u_t| < \epsilon \quad \forall\, t \geq t(\epsilon)\}$. Under this event, for $T > t(\epsilon)$,

$$\sum_{t=1}^T |u_t| \leq \sum_{t=1}^{t(\epsilon)} C + \sum_{t=t(\epsilon)n+1}^T \epsilon = t(\epsilon)C + T\epsilon.$$

Therefore, we obtain

$$\mathbb{P} \left(\frac{1}{T} \sum_{t=1}^T |u_t| > 2\epsilon \right) = \mathbb{P} \left(\left\{ \frac{1}{T} \sum_{t=1}^T |u_t| > 2\epsilon \right\} \cap \mathcal{E}(\epsilon) \right) + \mathbb{P} \left(\left\{ \frac{1}{T} \sum_{t=1}^T |u_t| > 2\epsilon \right\} \cap \mathcal{E}^c(\epsilon) \right)$$

$$\leq \mathbb{P} \left(\frac{t(\epsilon)}{T} C + \epsilon > 2\epsilon \right) + \mathbb{P}(\mathcal{E}^c(\epsilon)) = \mathbb{P} \left(\frac{t(\epsilon)}{T} C > \epsilon \right) + \mathbb{P}(\mathcal{E}^c(\epsilon)).$$

Letting $T \to \infty$, for arbitrarily small $\delta > 0$, , $\frac{1}{T} \sum_{t=1}^T \mathbb{E}_P[B_t^2 \mid X_t, \mathcal{F}_{t-1}] \xrightarrow{\text{P}} 0$ as $T \to \infty$ holds.

Then, from the dominated convergence theorem and the boundedness of B_t^2, $\frac{1}{T} \sum_{t=1}^T \text{Var}_P(B_t) \xrightarrow{\text{P}} 0$ as $T \to \infty$ holds[2].

Therefore, from (13), $\frac{1}{\sqrt{T}} \sum_{t=1}^T B_t \xrightarrow{\text{P}} 0$ holds, which implies

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \pi(a \mid X_t) \widehat{\Gamma}_t^a(Y_t^a, \xi_t, X_t)$$

$$= \frac{1}{\sqrt{T}} \sum_{t=1}^T \pi(a \mid X_t) \left\{ \frac{\mathbb{1}\left[\sum_{b=0}^{a-1} w^*(b \mid X_t) \leq \xi_t \leq \sum_{b=0}^a w^*(a \mid X_t) \right] \left(Y_t^a - \mu^a(P)(X_t) \right)}{w^*(a \mid X_t)} + \mu^a(P)(X_t) \right\} + o_P(1)$$

as $T \to \infty$. $\qquad\square$

---

[2]Our proof for this part refers to the proof of Lemma 10 in Hadad et al. (2021).

E Proof of Lemma C.3

This section provides the proof of Lemma C.3. In Appendix E.1, we introduce an upper bound of the Rademacher complexity shown by Zhou et al. (2023). Then, in Appendix E.2, we prove a new lemma (Lemma E.3), which directly yields Lemma C.3.

E.1 Upper bound of the Rademacher complexity.

Let us define a policy class

$$\Pi^D := \left\{ h : \mathcal{X} \times \prod_{a=1}^{K} \mathbb{R} \to \mathbb{R} \mid h\big(x, (\Gamma^a)_{a\in[K]}\big) := \sum_{a\in[K]} \big(\pi_1(a \mid x) - \pi_2(a \mid x)\big)\Gamma^a, \pi_1, \pi_2 \in \Pi \right\}$$

Then, let us define the Rademacher complexity as follows.

Definition E.1. *Let $\{Z_t\}_{t=1}^{T}$ be a sequence of i.i.d. Rademacher random variables $Z_t \in \{-1, +1\}$: $\mathbb{P}[Z_t = +1] = \mathbb{P}[Z_t = -1] = \frac{1}{2}$.*

- *The empirical Rademacher complexity $\mathfrak{R}_T\left(\Pi^D; \{X_t, \Gamma_t^{*a}\}_{t=1}^{T}\right)$ of a function class Π^D is defined as*

$$\mathfrak{R}_T\left(\Pi^D; \{X_t, \Gamma_t^{*a}\}_{t=1}^{T}\right)$$
$$= \mathbb{E}\left[\sup_{\pi_1, \pi_2 \in \Pi} \frac{1}{T} \left| \sum_{t=1}^{T} Z_t \sum_{a\in[K]} \big(\pi_1(a \mid X_t) - \pi_2(a \mid X_t)\big)\Gamma_t^{*a} \right| \mid \{X_t, \Gamma_t^{*a}\}_{t=1}^{T} \right],$$

where the expectation is taken with respect to $Z_1, \dots, Z_t$.

- *The Rademacher complexity $\mathfrak{R}_T\left(\Pi^D\right)$ of the function class Π^D is the expected value taken with respect to the observations $\{X_t, \Gamma_t^{*a}\}_{t=1}^{T}$ of the empirical Rademacher complexity:*

$$\mathfrak{R}_T\left(\Pi^D\right) := \mathbb{E}\left[\mathfrak{R}_T\left(\Pi^D; \{X_t, \Gamma_t^{*a}\}_{t=1}^{T}\right) \right].$$

Our proof starts from the following result about the Rademacher complexity shown by Zhou et al. (2023).

Lemma E.2 (From the inequality above (C.14) and the inequality in (C.19) of Zhou et al. (2023))**.** *Suppose that Assumptions 5.1 and 5.3 hold. Then, for any ζ and $P \in \mathcal{P}_\zeta$,*

$$\mathfrak{R}_T\left(\Pi^D\right) \leq 13.6\sqrt{2}\left\{\kappa(\Pi) + 8\right\} \frac{\Upsilon_*}{\sqrt{T}} + +O\left(\frac{\sqrt{U_T}}{T^{3/4}}\right).$$

holds, where

$$\Upsilon_* = \mathbb{E}\left[\sqrt{\sup_{\pi_1, \pi_2 \in \Pi} \sum_{t\in[T]} \left\{ \sum_{a\in[K]} \big(\pi_1(a \mid Z_i) - \pi_2(a \mid Z_i)\big)\Gamma^{*a}(Y_t, \xi_t, X_t) \right\}^2} \right].$$

Following this upper bound, Zhou et al. (2023) applies the Jensen inequality to bound Υ_* by

$$\Upsilon_* \leq \sqrt{\mathbb{E}\left[\sup_{\pi_1, \pi_2 \in \Pi} \left\{ \sum_{a\in[K]} \big(\pi_1(a \mid Z_i) - \pi_2(a \mid Z_i)\big)\Gamma^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right]}.$$

Then, they apply the Talagrand inequality to bound the regret as

$$\Upsilon_* \leq \sqrt{\sup_{\pi_1, \pi_2 \in \Pi} \mathbb{E}\left[\left\{ \sum_{a\in[K]} \big(\pi_1(a \mid Z_i) - \pi_2(a \mid Z_i)\big)\Gamma^{*a}(Y_t, \xi_t, X_t) \right\}^2 \right]}.$$

However, the use of the Jensen inequality yields a loose upper bound, which results in a mismatch between the upper bound and our derived lower bound. Therefore, in our proof, we consider bounding Υ_* without using Jensen's inequality.

E.2 Proof of Lemma C.3

Based on this proof strategy, we show the following lemma.

Lemma E.3 (From the equation above (C.14) in Zhou et al. (2023)). *Suppose that Assumptions 5.1 and 5.3 hold. Then, for each $P \in \mathcal{P}$,*

$$\mathbb{E}_P\left[R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right] \leq 54.4\sqrt{2}\left\{\kappa(\Pi) + 8\right\}\frac{\Upsilon_*}{\sqrt{T}} + +O\left(\frac{\sqrt{U_T}}{T^{3/4}}\right).$$

holds, where Υ_* is defined in Lemma E.2.

Lemma C.3 directly follows from Lemma E.3 by multiplying both sides by $\sqrt{T}$ and letting $T \to \infty$.

Proof of Lemma E.3. Recall that

$$\widehat{Q}_T(\pi) := \frac{1}{T}\sum_{t=1}^{T}\sum_{a\in[K]}\pi(a \mid X_t)\widehat{\Gamma}_t^a,$$

Let us define the following quantities:

$$\widetilde{Q}_T(\pi) := \frac{1}{T}\sum_{t=1}^{T}\sum_{a\in[K]}\pi(a \mid X_t)\Gamma_c^{*a}((Y_t^a)_{a\in[K]}, \xi_t, X_t),$$

$$\widehat{\Delta}(\pi_1, \pi_2) := \widehat{Q}_T(\pi_1) - \widehat{Q}_T(\pi_2),$$

$$\widetilde{\Delta}(\pi_1, \pi_2) := \widetilde{Q}_T(\pi_1) - \widetilde{Q}_T(\pi_2),$$

$$\Delta(\pi_1, \pi_2) := Q_c(P)(\pi_1) - Q_c(P)(\pi_2).$$

Then, we have

$$R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)$$
$$= Q_c(P)(\pi_c^*) - Q_c(P)\left(\widehat{\pi}^{\mathrm{PLAS}}\right)$$
$$= \widehat{Q}_T(\pi_c^*) - \widehat{Q}_T\left(\pi^{\mathrm{PLAS}}\right) + \Delta\left(\pi_c^*, \pi^{\mathrm{PLAS}}\right) - \widehat{\Delta}\left(\pi_c^*, \pi^{\mathrm{PLAS}}\right)$$
$$\leq \left|\Delta\left(\pi_c^*, \pi^{\mathrm{PLAS}}\right) - \widehat{\Delta}\left(\pi_c^*, \pi^{\mathrm{PLAS}}\right)\right|$$
$$\leq \sup_{\pi_1, \pi_2 \in \Pi}\left|\Delta(\pi_1, \pi_2) - \widehat{\Delta}(\pi_1, \pi_2)\right|$$
$$\leq \sup_{\pi_1, \pi_2 \in \Pi}\left|\widehat{\Delta}(\pi_1, \pi_2) - \widehat{\Delta}(\pi_1, \pi_2)\right| + \sup_{\pi_1, \pi_2 \in \Pi}\left|\Delta(\pi_1, \pi_2) - \widetilde{\Delta}(\pi_1, \pi_2)\right|.$$

From Lemma C.2, we have

$$\sqrt{T}R_c(P) \leq \sup_{\pi_1, \pi_2 \in \Pi}\left|\Delta(\pi_1, \pi_2) - \widetilde{\Delta}(\pi_1, \pi_2)\right| + o_P(1).$$

Therefore, we obtain

$$\mathbb{E}_P\left[R_c(P)\left(\widehat{\pi}_T^{\mathrm{PLAS}}\right)\right] \leq \mathbb{E}_P\left[\sup_{\pi_1, \pi_2 \in \Pi}\left|\Delta(\pi_1, \pi_2) - \widetilde{\Delta}(\pi_1, \pi_2)\right|\right],$$

where we used that $\sup_{\pi_1, \pi_2 \in \Pi}\left|\widehat{\Delta}(\pi_1, \pi_2) - \widehat{\Delta}(\pi_1, \pi_2)\right|$ is a bounded random variable.

Then, from the property of the Rademacher complexity (Bartlett & Mendelson, 2003) and Lemma E.2, we have

$$\mathbb{E}_P\left[\sup_{\pi_1, \pi_2 \in \Pi}\left|\Delta(\pi_1, \pi_2) - \widetilde{\Delta}(\pi_1, \pi_2)\right|\right]$$

$$\leq 4\Re_T(\Pi)$$

$$\leq 54.4\sqrt{2}\left\{\kappa(\Pi) + 8\right\}\frac{\Upsilon_*}{\sqrt{T}} + +O\left(\frac{\sqrt{U_T}}{T^{3/4}}\right).$$

This completes the proof. □