Statistical Inference for Misspecified Contextual Bandits

Yongyi Guo
Department of Statistics
University of Wisconsin-Madison
guo98@wisc.edu

Ziping Xu
School of Data Science and Society
University of North Carolina at Chapel Hill
zipingxu@unc.edu

July 31, 2025

Abstract

Contextual bandit algorithms have transformed how modern experiments are conducted by enabling real-time adaptation to data, personalized recommendations, and more efficient use of experimental resources. However, these advantages come with challenges—especially ensuring the validity of statistical inference from the adaptively collected data. A critical requirement is that the behavior policy used to assign treatments becomes stable over time. In this paper, we identify an important and previously unnoticed issue: popular contextual bandit algorithms, such as LinUCB, can become unstable when their underlying models are misspecified. Such model misspecification is not merely theoretical; it frequently arises in practice, for instance, when simplistic linear models are used to handle complex user behaviors for bias-variance trade-off.

Motivated by this insight, we propose and analyze a broad class of algorithms employing Boltzmann exploration strategies that remain stable even when faced with misspecified models. Building on this stability guarantee, we develop a flexible and broadly applicable statistical inference framework based on Z-estimators. Specifically, we introduce an inverse-probability-weighted Z-estimator (IPW-Z) and demonstrate its asymptotic normality under our stability conditions with a consistent estimator of the asymptotic variance. Extensive simulation studies confirm that our proposed inference method provides robust and accurate confidence intervals, outperforming previous approaches such as Contextual Adaptive Doubly Robust (CADR) estimator [8]. Our findings underscore the practical importance of carefully designed adaptive algorithms to ensure stable online experiments that allows for valid after-study statistical inference.

1 Introduction

There is a growing interest in data collected from adaptive experimental designs. As opposed to traditional randomized experiments, adaptive designs sequentially update each arm's allocation probabilities based on accumulated outcomes. This adaptivity allows real-time improvement in treatment strategies and more efficient sample usages. For example, in mobile-health studies, adaptive designs personalize digital interventions—delivering motivational messages or prompts tailored to each user's state—to promote healthy behaviors [27, 59]. In online advertising, adaptive designs adjust email or content recommendations in real time to maximize click-through rates [41]. Other work has focused on dynamically reallocating samples to increase statistical power for hypothesis tests, thereby reducing total sample sizes needed for a given significance level [53].

Such adaptive experimental designs are often implemented via *online contextual bandit algorithms* [41]. Contextual bandit problem can be formulated as an agent interacting with an unknown environment over a sequence of steps. At each step t, the agent observes a context $\boldsymbol{X}_t$ (for example user feature vector) and chooses a behavior (logging) policy π_t that maps contexts to treatment probabilities. The agent then samples a treatment arm $A_t \sim \pi_t(\cdot \mid \boldsymbol{X}_t)$ and observes an outcome Y_t. The goal of the agent is to optimize cumulative outcomes over time. To effectively learn the unknown environment and improve its decision rule over time, the agent must balance exploration (trying actions with uncertain outcomes) and exploitation (favoring actions known to yield high outcomes). This need for continual exploration forces high adaptivity: the agent's policy is continually updated based on the outcomes of previous actions. Common approaches to this problem include LinUCB [41] and Thompson Sampling [54], both of which enjoy strong sample complexity guarantees.

Our focus is on **statistical inference** for data adaptively collected via a contextual bandit algorithm. In the literature, there are two main approaches to inference with adaptive data: the first approach leverages the martingale nature of the data through martingale central limit theorem for asymptotically valid confidence intervals [21, 22, 34]. The second approach employs non-asymptotic concentration inequalities for self-normalized martingales [2]. While the latter approach provides finite-sample guarantees, the resulting confidence intervals are typically much wider, as they rely on conservative high-probability bounds. Therefore, in this paper we focus on the first approach, aiming to derive exact asymptotic distribution results for inferential procedures.

A major challenge in performing inference on adaptively collected data is that the adaptivity breaks the usual i.i.d. assumptions. Adaptive data collection leads to a non-stationary behavior policy and introduces dependence over time, violating the classical assumptions of independent observations. **A key idea to address this issue is to ensure the stability of the behavior policy—namely, that the policy converges to a fixed distribution over actions in the long run.** This notion of stability is conceptually related to conditions ensuring the stability of sample covariance matrices in adaptive designs [39]. Intuitively, data collected under a stable policy become "almost" i.i.d. in the limit, enabling the use of standard asymptotic normality techniques. Moreover, stability is crucial for replicability in adaptive experiments. Replicability means that an adaptive algorithm yields consistent results when an experiment is repeated under the same conditions, which is crucial for validating findings and building confidence in data-driven decisions. Recent work has shown that many online contextual bandit algorithms, when applied to a batched setting where n participants are recruited simultaneously and asymptotics is as $n \to \infty$, are not stable in a misspecified environment [71].

Despite the importance of policy stability, there is a lack of comprehensive understanding of when a behavior policy will stabilize and what the consequences are if it does not—particularly for the validity of statistical inference on adaptively collected data. In the simpler multi-armed bandit (MAB) setting without context, it is well known that many algorithms are stable in a "regular" environment (i.e. when there is a non-zero suboptimality gap between the optimal and suboptimal arms). Recent works have shown that even in an irregular environment with a zero suboptimality gap, existing online MAB algorithms (e.g., UCB [31] and Thompson Sampling [30]) remain stable. In contrast, there has been little discussion of policy stability in the contextual bandit setting. MAB problem aims at mean of each arm, thereby always have the correctly specified outcome model,

when the environemnt is stationary. It is not clear whether common contextual bandit algorithms in misspecified environments are unstable as it is the case in the batched setting [71].

Apart from the stability issue, a second gap in the literature is that most existing statistical inference methods for adaptive experiments target specific parameters of a presumed correctly-specified outcome model. The outcome models considered in the literature include generalize linear/partial linear models [11, 22, 46, 49], structured nested mean models [50, 52, 57], and additive models or factor models [1, 24, 65], and the inference target are parameters of the outcome model. Some infer the mean of arms in multi-armed bandit (MAB) settings [30, 31] or causal effects in longitudinal data and panel data settings [7, 47, 51]. Another related line of works is the offline policy evaluation (OPE), where the inference target is the treatment effects [8, 29]. Overall, there is no general inference framework in the existing literature that is agnostic to the form of the outcome model.

These observations motivate the another main focus of the paper: **model misspecification in both adaptive data collection phase and inference phase**. Model misspecification setting has strong practical relevance. In complex environments with rich heterogeneity, the deployed contextual bandit algorithm often operates with misspecified working models. For instance, real-world deployments tend to use simple linear working reward models with carefully chosen features to balance bias and variance, even though the underlying true model outcome relationship may be far more complex [5, 58]. Model misspecification can also arise from heterogeneity across units that either is not captured by the observed context features or is intentionally left out to reduce variance [59]. Another source of misspecification is high noise (or measurement error) in the context–a true linear context-reward relationship will no longer appear linear once the observed contexts are corrupted by noise.

While model misspecification has been extensively studied in terms of its effect on learning efficiency (e.g. how it impacts regret bounds [25]), there has been little investigation of its impact on the stability of the behavior policy, or further the validity of statistical inference. In this work, we demonstrate that under model misspecification, many standard contextual bandit algorithms (e.g. linear UCB or Thompson Sampling) lack policy stability. We further show that if the behavior policy fails to stabilize, the asymptotic variance of common estimators may not converge. Consequently, asymptotic normality breaks down, invalidating standard inference approaches — even those that use importance weighting to correct for adaptively collected data.

1.1 Our Contributions

A summary of the our main contributions and a brief outline of the paper is as follows:

- **General inference target (Z-estimation framework):** In Section 2, we formulate a general inference target for adaptively collected data using a Z-estimator framework: for any arm a, the inference target θ_a^* is defined as the solution to the following equation:

$$\mathbb{E}\boldsymbol{g}\left(\boldsymbol{X}, Y(a); \boldsymbol{\theta}_a^*\right) = \mathbf{0}, \tag{1.1}$$

  where $\boldsymbol{g}(\cdot)$ is a score function. We discuss three examples of the inference target of practical interest with different score functions: misspecified linear bandits, bandits with noisy contexts, and off-policy evaluation.

- **IPW-Z estimator and asymptotic normality:** In Section 3, we propose an inverse-probability-weighted Z-estimator (IPW-Z) defined in (2.6) for this target. Under the assumption of policy stability, we prove that the proposed estimator is consistent and asymptotically normal, and the asymptotic variance can be consistently estimated.

- **Role of policy stability (sufficient conditions and examples):** In Section 4, we investigate the crucial role of policy stability via detailed simulations. We first demonstrate that LinUCB under misspecified environment can be unstable, and the asymptotic distribution of IPW-Z estimator deviates from Gaussian. We follow by introducing a sufficient condition for policy stability. We further introduce a rich family of policies that satisfy the sufficient condition including MAB algorithms that ignores context, ϵ-greedy exploration w.r.t. IPW-Z estimator, and Boltzmann exploration w.r.t Ridge regression estimator and stochastic gradient descent estimator.

- **Simulation results:** In Section 6, we extensively validate asymptotic normality results and valid inference for all three inference targets through simulations. Additionally, we compare with Contextual Adaptive Doubly Robust (CADR) estimator [8] for the OPE target, and show that our inference methods are more robust and perform better than existing approaches.

1.2 Related work

Statistical inference with adaptively collected data has been a long-standing but challenging problem. Although such data are common in practice, classical inference procedures designed for i.i.d. samples can break down under adaptive collection [48, 69]. A growing body of work has therefore focused on establishing conditions on the data generating process or developing new methodologies to ensure valid inference. Most existing work focuses on estimating parameters in correctly specified outcome models. For example, [4, 19, 39] analyze adaptive linear regression under various conditions; [36, 38] study least squares estimation in adaptive nonlinear regression; and [17] establishes asymptotic properties of maximum quasi-likelihood estimators for generalized linear models with adaptive designs.

With the emergence of modern data collection schemes such as reinforcement learning, recent work has developed inference procedures accommodating more general mechanisms such as contextual bandits, which are not covered by earlier results. [22, 34] propose online debiasing estimators for adaptive linear regression with weaker restrictions on the data collection process. [69] study batched linear regression under general contextual bandit algorithms. [15] develop inference for linear contextual bandits with ϵ-greedy policies, and [16] extend their results to nonlinear reward models with parameter updates via weighted stochastic gradient descent. [70] establish inference for M-estimators under contextual bandit sampling, and [46] consider adaptive inference in generalized partial linear outcome models. [57] study inference for structural parameters in structural nested mean models with data collected via reinforcement learning algorithms.

In the absence of a well-specified reward model, several works have examined statistical inference with adaptively collected data for specific target estimands, primarily various forms of average outcomes. [31, 35] study inference on arm means under data collected by UCB-type algorithms, while [30] consider the same estimand with Thompson sampling variants. [8, 29, 63, 68] analyze off-policy evaluation in bandits with general adaptive behavior policies, targeting the average reward of

a specified evaluation policy. [44, 45] develop inference for the long-run average reward in Markov decision processes with adaptive policies. Additional targets arising in longitudinal and causal panel data settings include mean responses to dynamic treatment regimes and average causal effects; see, for example, [14, 37].

Works on inference for general target parameters beyond average outcomes without a well-specified reward model is comparatively sparse. A result closely related to ours is [15], who study inference in linear contextual bandits under both well-specified and misspecified reward models. In the misspecified case, they propose a weighted least squares estimator for the least false parameter, defined as the best linear projection of the true reward, and establish its asymptotic properties under an ϵ-greedy behavior policy constructed from the same estimator. [71, 72] analyze inference after adaptive sampling in a general longitudinal data setting, but in a different regime where the time horizon is fixed and the number of trajectories diverges.

2 Problem Setup

We consider the problem of statistical inference with an adaptively collected dataset $\mathcal{D} = \{\boldsymbol{X}_t, \pi_t, A_t, Y_t\}_{t=1}^T$ from a contextual bandit environment. The data collection process proceeds at each time t as follows:

- **Context:** The environment reveals a context $\boldsymbol{X}_t \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$.
- **Action Selection:** Based on the current context $\boldsymbol{X}_t$ and past history $\mathcal{H}_{t-1} := \{\boldsymbol{X}_\tau, \pi_\tau, A_\tau, Y_\tau\}_{\tau<t}$, the agent selects an action $A_t \in \mathcal{A}$ according to a stochastic behavior policy $\pi_t(\cdot \mid \boldsymbol{X}_t, \mathcal{H}_{t-1}) \in \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the set of probability distributions over the action space $\mathcal{A}$. The realized selection probability is recorded as $\pi_t := \pi_t(A_t \mid \boldsymbol{X}_t, \mathcal{H}_{t-1})$.
- **Outcome:** After choosing the action, the agent observes outcome $Y_t \in \mathbb{R}$.

We consider a finite action space $\mathcal{A}$ and, without loss of generality, write $\mathcal{A} = \{1, \ldots, K\}$. Adopting the potential outcomes framework [32], we let $\{Y_t(a) : a \in \mathcal{A}\}$ denote the potential outcomes for each action, with the observed outcome satisfying $Y_t = Y_t(A_t)$. We assume a stochastic contextual bandit environment in which $\{\boldsymbol{X}_t, Y_t(a) : a \in \mathcal{A}\} \overset{\text{i.i.d.}}{\sim} \mathcal{P}$, for $t = 1, \ldots, T$. In addition, in this adaptive experimental setting, we assume the following unconfoundedness condition.

Assumption 2.1. $A_t \perp \{Y_t(a)\}_{a\in\mathcal{A}} | (\mathcal{H}_{t-1}, \boldsymbol{X}_t)$, for $t = 1, \ldots, T$.

Note that even though the potential outcomes are i.i.d., the observations in $\mathcal{D}$ are not. This is because each action A_t is selected based on the evolving history $\mathcal{H}_{t-1}$, introducing temporal dependence into the observations. This dependence poses additional challenges for valid estimation and inference.

2.1 Inference Targets

Our goal is to infer the parameter $\boldsymbol{\theta}_a^* \in \mathbb{R}^d$ associated with a treatment arm $a \in \mathcal{A}$, or jointly the collection $\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}$. Each $\boldsymbol{\theta}_a^*$ is defined as the solution to the equation

$$\mathbb{E}\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}_a^*) = \boldsymbol{0} \tag{2.1}$$

for some known score function $\boldsymbol{g} : \mathcal{X} \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$.[1] Here, $(\boldsymbol{X}, Y(a))$ denotes a generic observation drawn from the same distribution as $(\boldsymbol{X}_t, Y_t(a))$. Unlike many prior works on statistical inference with adaptively collected data [16, 39, 46, 70, 71], we do not assume a well-specified outcome model. In particular, the validity of our inference procedure does not rely on correctly modeling the conditional distribution of Y_t given A_t and $\boldsymbol{X}_t$ in $\mathcal{P}$. This feature makes our approach especially well-suited for adaptive experiments conducted in complex environments, where data may be limited, noise is non-negligible, and model misspecification is common. We illustrate this setup with three examples below.

Example 1 (Misspecified linear bandits [15])**.** Consider a target parameter $\boldsymbol{\theta}_a^*$ which solves (2.1) with

$$\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta}) := \boldsymbol{x}(y - \boldsymbol{x}^\top \boldsymbol{\theta}). \tag{2.2}$$

This score function corresponds to the the best linear approximation of $Y_t(a)$ based on $\boldsymbol{X}_t$ for arm $a \in \mathcal{A}$. When the true dependence of $Y_t(a)$ on $\boldsymbol{X}_t$ is linear, (2.2) yields the true linear parameter. Otherwise, (2.2) defines the best linear projection of $Y_t(a)$ onto the covariates $\boldsymbol{X}_t$ in the least squares sense.

Example 2 (Bandits with noisy contexts [28])**.** Suppose the potential outcome $Y_t(a)$ follows a linear model based on the unobserved true covariates $\boldsymbol{S}_t$:

$$Y_t(a) = \boldsymbol{S}_t^\top \boldsymbol{\theta}_a^* + \eta_t,$$

where η_t is a mean-zero noise term. Instead of observing $\boldsymbol{S}_t$, the observed context $\boldsymbol{X}_t$ is a noisy proxy $\boldsymbol{X}_t = \boldsymbol{S}_t + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t$ is mean zero, uncorrelated with η_t, and has covariance matrix $\boldsymbol{\Sigma}_e$. Although the outcome model is linear in $\boldsymbol{S}_t$, we do not assume any parametric form for the distribution of the measurement error $\boldsymbol{\epsilon}_t$, making it difficult to characterize the conditional distribution of $Y_t(a) \mid \boldsymbol{X}_t$.

With a non-adpative data collection process, this model has been well studied in statistics literature and is called the measurement error model [12, 26]. Assuming $\boldsymbol{\Sigma}_e$ is known, then $\boldsymbol{\theta}_a^*$ solves (2.1) with

$$\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta}) := (\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma}_e)\boldsymbol{\theta} - \boldsymbol{x}y. \tag{2.3}$$

The measurement error model is motivated by practice in behavioral psychology. For example, in a mobile health study about reducing negative affect to improve medication adherence [64], the negative affect can only be measured through a short survey, a noisy proxy of the latent negative affect. However, the scientists are truly interested in the relationship between medication adherence and the latent negative affect rather than the noisy proxy.

Example 3 (Off-policy evaluation in contextual bandits)**.** Our goal is to estimate the average outcome under a target policy $\pi^e : \mathcal{X} \to \Delta(\mathcal{A})$, defined as $V^* = \mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{A \sim \pi^e(\cdot|\boldsymbol{X})} Y(A)$. This target can be expressed as $V^* = \sum_{a \in \mathcal{A}} \boldsymbol{\theta}_a^*$, where each $\boldsymbol{\theta}_a^*$ solves (2.1) with an arm-specific score function $\boldsymbol{g}$:

$$\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta}) = \boldsymbol{g}_a(\boldsymbol{x}, y; \boldsymbol{\theta}) := \pi^e(a|\boldsymbol{x})y - \boldsymbol{\theta}. \tag{2.4}$$

These examples will be discussed in more details in Section 5.

[1] For simplicity, we assume a common score function $\boldsymbol{g}$ across actions. Our analysis extends to action-specific score functions; see Appendix B.2 for details

2.2 Challenge

A central challenge in the absence of a well-specified outcome model is that standard Z-estimation approaches [46, 70, 71], which analyze estimators $\widehat{\boldsymbol{\theta}}'_a$ satisfying

$$\frac{1}{T}\sum_{t=1}^{T} 1_{\{A_t=a\}}\boldsymbol{g}\big(\boldsymbol{X}_t, Y_t, \widehat{\boldsymbol{\theta}}'_a\big) = o_p\big(1/\sqrt{T}\big), \tag{2.5}$$

fail to yield valid inference. This failure essentially stems from the interaction between the policy and the complex environment. Specifically, for a general distribution $\mathcal{P}$, the solution $\boldsymbol{\theta}$ to the conditional moment equation $\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}, Y(a); \boldsymbol{\theta}) \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbf{0}$ can vary with $\boldsymbol{x}$. As a result, the solution to (2.5) depends on the behavior policy used to collect the data, and is generally not consistent. This issue, as it arises in Examples 1 and 2, is discussed in detail in [15] and [28], respectively.

In this work, we develop statistical inference of the target parameters $\{\boldsymbol{\theta}^*_a\}_{a\in\mathcal{A}}$ by studying the asymptotic properties of the inverse probability weighted Z-estimators $\{\widehat{\boldsymbol{\theta}}^{(T)}_a\}_{a\in\mathcal{A}}$, where $\widehat{\boldsymbol{\theta}}^{(T)}_a$ satisfies

$$\boldsymbol{G}_T(\boldsymbol{\theta}) := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}}\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) = o_p\big(1/\sqrt{T}\big). \tag{2.6}$$

Here, $\pi_t(a)$ abbreviates the action selection probability $\pi_t(a \mid \mathcal{H}_{t-1}, \boldsymbol{X}_t)$ for $a \in \mathcal{A}$. The use of inverse probability weights $1/\pi_t(A_t)$ is standard in the off-policy evaluation literature (e.g., [42, 61]) and has also been employed in specific adaptive settings by [15] and [28] for various purposes. Our goal is to show that, in our general setting, the inverse probability weights effectively decouple the policy from the underlying environment, enabling valid inference. Specifically, we will establish the joint asymptotic normality of $\{\widehat{\boldsymbol{\theta}}^{(T)}_a\}_{a\in\mathcal{A}}$ under mild conditions on the environment and for a broad class of behavior policies.

3 Statistical Inference Guarantees

The main results of this section establish the joint asymptotic normality of the proposed estimators $\{\widehat{\boldsymbol{\theta}}^{(T)}_a\}_{a\in\mathcal{A}}$, along with consistent estimators of the asymptotic variance, which enable statistical inference for $\{\boldsymbol{\theta}^*_a\}_{a\in\mathcal{A}}$. To achieve these guarantees, we highlight the critical role of policy convergence (Definition 3.1), an important condition on the behavior policy used to collect the data. In Section 3.1, we present a concrete example demonstrating how the failure of this condition can lead to the breakdown of asymptotic normality. We begin by formally stating the condition below.

Definition 3.1 (Policy convergence)**.** The behavior policy $\pi = \{\pi_t(\cdot)\}_{t\geq 1}$ is said to satisfy policy convergence, or to converge, at action $a \in \mathcal{A}$ if there exists a stationary policy $\bar{\pi} : \mathcal{X} \mapsto \Delta(\mathcal{A})$ such that

$$\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}) - \bar{\pi}(a|\boldsymbol{X}_t) \xrightarrow{p} 0 \quad \text{as } t \to \infty. \tag{3.1}$$

We say that π satisfies policy convergence—or simply converges—if (3.1) holds for every action $a \in \mathcal{A}$ with respect to a common stationary policy $\bar{\pi}$.

The policy convergence condition essentially requires that, in the long run, the action-selection probabilities stabilize in probability across the context. Beyond its practical importance, this

condition plays a central theoretical role: it ensures that the conditional variances of each term in the estimating equation (2.6) given past history stabilize, which guarantees the martingale central limit theorems and thereby yields the desired asymptotic normality of the estimators (see Appendix B.2 for details). In Section 4, we further investigate sufficient conditions for policy convergence in general environments without a well-specified reward model. These conditions are straightforward to verify, offering concrete guidance for implementing adaptive policies in online settings.

Remark 3.1. Policy convergence, or related notions of policy stability, is a common assumption in the literature on inference with adaptively collected data—particularly in the absence of a well-specified model. For instance, in misspecified linear bandits (Example 1), [15] study inference under an ϵ-greedy algorithm with a weighted online least squares (LS) estimator—a special case that satisfies policy convergence in their setting (see Section 4). Similar convergence conditions are also assumed in other adaptive inference problems under model misspecification [68, 71, 72]. In settings without model misspecification, prior work has shown that various forms of stability conditions lead to valid statistical inference. These conditions are typically weaker than, but often closely related to, policy convergence. For example, [39] studies stochastic linear regression and derives asymptotic normality of the OLS estimator under a stability condition that requires the design matrix to behave regularly over time. [31, 35] analyze the UCB algorithm in multi-armed bandits and show valid inference under stability conditions where the ratio between the number of arm pulls and a diverging sequence converges to one.

We next introduce the technical assumptions required to establish the asymptotic properties of the estimator $\widehat{\boldsymbol{\theta}}_a^{(T)}$ for a single action $a \in \mathcal{A}$.

Assumption 3.1 (Well-separated solution)**.** $\forall \epsilon > 0$, $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2>\epsilon} \|\mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2 > 0$.

Assumption 3.2 (Bounded moments)**.** There exist constants R_θ, M_2 such that
(i) $\|\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top | \boldsymbol{X}_t]\|_2 \le M_2$, a.e. $\boldsymbol{X}_t$;
(ii) $\|\boldsymbol{\theta}_a^*\|_2 < R_\theta$, $\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \mathbb{E}\|\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2^2 < \infty$; (iii) $\mathbb{E}\|\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\|_2^4 < \infty$.

Assumption 3.3 (Smoothness)**.** (i) The function $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$, with $\mathbb{E}\nabla\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$ nonsingular; (ii) There exists a function $\phi : \mathbb{R}^{d_X} \times \mathbb{R} \mapsto \mathbb{R}$ such that $\forall \boldsymbol{x}, y$, $\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \|\nabla\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})\|_2 \le \phi(\boldsymbol{x}, y)$, and $\mathbb{E}\phi(\boldsymbol{X}_t, Y_t(a))^2 < \infty$; (iii) There exists a constant $\epsilon_0 > 0$ and a function $\Phi : \mathbb{R}^{d_X} \times \mathbb{R} \mapsto \mathbb{R}$ such that $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0, i\in[d]} \|\nabla^2\boldsymbol{g}^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})\|_2 \le \Phi(\boldsymbol{x}, y)$ and $\mathbb{E}\Phi(\boldsymbol{X}_t, Y_t(a)) < \infty$. Here $\boldsymbol{g}^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})$ denotes the i-th entry of $\boldsymbol{g}(\boldsymbol{x}, y; \boldsymbol{\theta})$.

Assumption 3.4 (Minimum sampling probability)**.** $\pi_t(a) \ge \pi_{\min}$ almost surely for some constant $\pi_{\min} \in (0, 1)$.

Remark 3.2. We provide a few comments on these assumptions. First, Assumption 3.1 is a standard condition in Z-estimation that ensures the identifiability of the target parameter $\boldsymbol{\theta}_a^*$ [62]. Assumption 3.2 imposes mild regularity conditions on the boundedness of moments and conditional moments of the score function $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$ evaluated at the true parameter $\boldsymbol{\theta}_a^*$. Notably, it does not require the potential outcomes $Y_t(a)$ themselves to be bounded or to satisfy sub-Gaussian tail conditions. Similar assumptions appear in related works, such as [15, 68, 70, 71]. Assumption 3.3 imposes smoothness conditions on the score function $\boldsymbol{g}$, a standard requirement in classical Z-estimation as well as in recent work on Z- and M-estimation with adaptively collected data [70, 72].

Finally, Assumption 3.4 imposes a minimum sampling probability, which is frequently assumed in adaptive inference without a well-specified outcome model (e.g., [15, 70, 72]). In practice, in adaptive experiments with highly noisy and complex environment, keeping a minimum exploration rate ensures statistical power for flexible post-hoc analysis [40, 67], particularly when the analysis objective is not pre-specified at the time of data collection. It also enables policy updates and re-optimization for future users, accommodating potential non-stationarity across trials [43, 66].

We now state the first main result of this section, which establishes the asymptotic properties of the estimator $\widehat{\boldsymbol{\theta}}_a^{(T)}$ for a single action $a \in \mathcal{A}$. The proof is provided in Appendix B.1.

Theorem 3.2. *Suppose Assumptions 2.1 and 3.1–3.4 hold for a given action $a \in \mathcal{A}$. If the behavior policy π converges to a policy $\bar{\pi}$ at action a in the sense of Definition 3.1, then there exists an estimator sequence $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{T\geq 1}$ satisfying the estimating equation (2.6) with $\|\widehat{\boldsymbol{\theta}}_a^{(T)}\|_2 \leq R_\theta$ for all T. Moreover, for any such sequence, as $T \to \infty$,*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_a^*). \tag{3.2}$$

Here $\boldsymbol{\Sigma}_a^* := \boldsymbol{J}_a^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{J}_a^{-1,\top}$, with $\bar{\boldsymbol{I}}_a := \mathbb{E}\big[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top\big]$,
$\boldsymbol{J}_a := \mathbb{E}\nabla\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$.

The role of inverse probability weights. To build intuition for Theorem 3.2, we present a heuristic argument illustrating how the inverse probability weights in the estimating equation (2.6), which defines our estimator $\widehat{\boldsymbol{\theta}}_a^{(T)}$, contribute to achieving valid inference. Specifically, a key step in the proof of Theorem 3.2 is to show that $\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{Z}_t$ forms a sum of a martingale difference sequence with respect to the filtration $\{\mathcal{H}_t\}_{t=0}^T$, where

$$\boldsymbol{Z}_t := \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}_a^*).$$

Define $w_a(A_t) := \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}}$. Then we have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}_t|\mathcal{H}_{t-1}] &\overset{(a)}{=} \mathbb{E}_{\boldsymbol{X}_t}\Big[\mathbb{E}_{A_t\sim\pi_t(\cdot),Y_t(a)}[\boldsymbol{Z}_t|\mathcal{H}_{t-1},\boldsymbol{X}_t]\Big|\mathcal{H}_{t-1}\Big] \\
&\overset{(b)}{=} \mathbb{E}_{\boldsymbol{X}_t}\Big[\mathbb{E}_{A_t\sim\pi_t(\cdot)}\big[w_a(A_t)\big|\mathcal{H}_{t-1},\boldsymbol{X}_t\big]\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)|\mathcal{H}_{t-1},\boldsymbol{X}_t]\Big|\mathcal{H}_{t-1}\Big] \\
&\overset{(c)}{=} \mathbb{E}_{\boldsymbol{X}_t}\Big[1\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)|\mathcal{H}_{t-1},\boldsymbol{X}_t]\Big|\mathcal{H}_{t-1}\Big] \\
&\overset{(d)}{=} \mathbb{E}\left[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\right] = \mathbf{0}.
\end{aligned}$$

Here, step (a) follows from the law of iterated expectations. Step (b) uses Assumption 2.1, which ensures that A_t and $Y_t(a)$ are independent condition on $(\mathcal{H}_{t-1}, \boldsymbol{X}_t)$. Step (c) follows from a direct computation:

$$\mathbb{E}_{A_t\sim\pi_t(\cdot)}\big[w_a(A_t)\big|\mathcal{H}_{t-1},\boldsymbol{X}_t\big] = \sum_{a'\in\mathcal{A}} \pi_t(a')\cdot\frac{1_{\{a'=a\}}}{\pi_t(a')} = 1.$$

Step (d) again applies the law of iterated expectations along with the assumption that $\{\boldsymbol{X}_t, Y_t(a) : a \in \mathcal{A}\}$ are i.i.d. over time. In contrast, if the estimator is derived without incorporating inverse probability weights—as in (2.5)—we instead have

$$\mathbb{E}[\boldsymbol{Z}_t|\mathcal{H}_{t-1}] = \mathbb{E}_{\boldsymbol{X}_t}\Big[\pi_t(a|\mathcal{H}_{t-1},\boldsymbol{X}_t)\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)|\mathcal{H}_{t-1},\boldsymbol{X}_t]\Big|\mathcal{H}_{t-1}\Big],$$

which clearly depends on the behavior policy $\pi_t(\cdot)$ and is generally nonzero. This in turn prevents the unweighted Z-estimator in (2.5) from being consistent.

Comparison to prior work. In the setting of misspecified linear bandits (Example 1), Theorem 4.1 of [15] establishes the asymptotic normality of $\widehat{\boldsymbol{\theta}}_a^{(T)}$ under a specific converging behavior policy: an ϵ-greedy policy paired with a weighted online LS estimator. In contrast, Theorem 3.2 applies to any online decision-making algorithm that satisfies policy convergence and a minimum sampling probability condition. Section 4 demonstrates that a broad class of policies meet these requirements. In practice, statisticians seeking to conduct inference often do not control the algorithm used to collect the data. Therefore, Theorem 3.2 offers a broader and more flexible generalization of Theorem 4.1 in [15], both in terms of the allowable data collection mechanisms and the inferential targets.

In practice, it is often desirable to jointly infer the collection $\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}$, where for each arm $a \in \mathcal{A} = \{1, \ldots, K\}$, $\boldsymbol{\theta}_a^*$ denotes the solution to (2.1). Such joint inference is particularly relevant for tasks like estimating individual treatment effects or evaluating the value of a general target policy. To achieve this goal, Theorem 3.3 below establishes the joint asymptotic normality of the estimators $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}}$, with the proof provided in Appendix B.2. For simplicity, we assume a common score function $\boldsymbol{g}$ across actions, though our analysis extends to action-specific score functions; see Appendix B.2 for details.

Theorem 3.3. *Suppose Assumption 2.1 holds, and Assumptions 3.1–3.4 hold for every action $a \in \mathcal{A}$. If the behavior policy π satisfies policy convergence to a policy $\bar{\pi}$ in the sense of Definition 3.1, then there exist estimators $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A},T\geq 1}$ such that (2.6) holds for each $a \in \mathcal{A}$, and $\|\widehat{\boldsymbol{\theta}}_a^{(T)}\|_2 \leq R_\theta$ for all $a \in \mathcal{A}, T \geq 1$. In addition, any such estimators satisfy*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}^{(T)} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*) \tag{3.3}$$

as $T \to \infty$. Here $\widehat{\boldsymbol{\theta}}^{(T)} = \big((\widehat{\boldsymbol{\theta}}_1^{(T)})^\top, \ldots, (\widehat{\boldsymbol{\theta}}_K^{(T)})^\top\big)^\top$, $\boldsymbol{\theta}^* = \big(\boldsymbol{\theta}_1^{*,\top}, \ldots, \boldsymbol{\theta}_K^{*,\top}\big)^\top$, and $\boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{\Sigma}_1^*, \ldots, \boldsymbol{\Sigma}_K^*)$, with each $\boldsymbol{\Sigma}_a^*$ defined as in Theorem 3.2.

Theorem 3.3 indicates that the estimators $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}}$ are asymptotically uncorrelated. Intuitively, this is because each $\widehat{\boldsymbol{\theta}}_a^{(T)}$ is constructed using data exclusively from time points when action a is selected, and these sets of time points are disjoint across different actions.

The asymptotic variances in Theorems 3.2 and 3.3 can be consistently estimated from data, as shown in the proposition below, thereby enabling valid statistical inference. The proof is provided in Appendix B.3.

Proposition 3.1. Under the same conditions of Theorem 3.2, the asymptotic variance $\boldsymbol{\Sigma}_a^*$ can be consistently estimated by

$$\widehat{\boldsymbol{\Sigma}}_a = \Big[\widehat{\dot{\boldsymbol{G}}}_{a,T}\Big]^{-1} \widehat{\boldsymbol{I}}_{a,T} \Big[\widehat{\dot{\boldsymbol{G}}}_{a,T}\Big]^{-1,\top}, \tag{3.4}$$

where

$$\widehat{\dot{\boldsymbol{G}}}_{a,T} := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_a^{(T)}),$$

$$\widehat{\boldsymbol{I}}_{a,T} := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_t(A_t)^2} 1_{\{A_t=a\}} \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_a^{(T)}) \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_a^{(T)})^\top.$$

Importantly, Proposition 3.1 shows that consistent estimation of the asymptotic variance does not rely on knowledge of the limit policy $\bar{\pi}$.

3.1 Examples of Policy Non-convergence

Previous literature has shown that IPW estimators can be non-normal without any constraint on the behavior policy [8, 29], for example, when the behavior policy has no minimum sampling probability. Such non-normal asymptotic behavior often arises from unbounded inverse probability weights. In this section, we primarily interested in the non-normal asymptotic behavior due to policy non-convergence. Similar failure modes have been observed for Thompson Sampling and UCB [30]. We point out that model misspecification may result in policy non-convergence, which can lead to non-normal asymptotic behavior of IPW estimators.

To illustrate, we consider a two-armed bandit environment with contexts generated as $\boldsymbol{X}_t \sim \text{Uniform}(\{-4, 1\})$ and mean reward $y(\boldsymbol{x}, a) = \mathbb{E}[Y_t(a)|\boldsymbol{X}_t = \boldsymbol{x}]$ given by

$$y(-4, a_0) = y(1, a_0) = 1/2, \quad y(-4, a_1) = y(1, a_1) = 1/12.$$

The rewards are subject to Gaussian noise. We independently run LinUCB and Random algorithms for 2500 times for 10^4 steps using a working linear model $Y_t(a) = \boldsymbol{\theta}_a^\top \boldsymbol{X}_t$, where $\boldsymbol{\theta}_a \in \mathbb{R}$ is an unknown parameter. We clip the LinUCB sampling probability at the level of 0.01 to maintain bounded inverse probability weights.

In Figure 1, we show the distributions of the last-step ridge regression estimator for $\boldsymbol{\theta}_a$, sampling probabilities, the proposed IPW-Z estimator defined in (2.5), and the QQ plot comparing the IPW-Z estimator's empirical distribution to a standard Gaussian distribution. We observe that the ridge regression estimator maintained by LinUCB exhibits two distinct convergence modes, evident from the two peaks in the histogram (panel a) and the bimodal distribution of the sampling probability at context $\boldsymbol{x} = -4$ (panel b). Additionally, the asymptotic distribution of the IPW-Z estimator based on LinUCB dataset substantially deviates from Gaussianity (panels c and d), while the estimator based on Random algorithm dataset remains asymptotically normal.

4 Sufficient Conditions for Policy Convergence

In this section, we first study sufficient conditions under which a policy satisfies the policy convergence condition in Definition 3.1. Then, in Sections 4.1 to **??**, we illustrate several classes of convergent policies in general environments without assuming a well-specified reward model. These results offer a principled foundation for constructing stable bandit policies in practice, particularly in noisy and complex settings.

We focus on behavior policies that belong to a parametric class of the form $\pi(\cdot|\boldsymbol{X}_t, \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^{d_\beta}$. At each time t, the behavior policy $\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1})$ takes the form

$$\pi(a|\boldsymbol{X}_t, \widehat{\boldsymbol{\beta}}_{t-1}) \tag{4.1}$$

for a fixed function $\pi : \mathcal{X} \times \mathbb{R}^{d_\beta} \mapsto \Delta(\mathcal{A})$, where $\widehat{\boldsymbol{\beta}}_{t-1} \in \mathbb{R}^{d_\beta}$ is a $\mathcal{H}_{t-1}$-measurable random vector. This vector can be interpreted as a *summary statistic* that aggregates information from the past $t-1$ rounds and, together with the current context $\boldsymbol{X}_t$, determines the action selection probability at

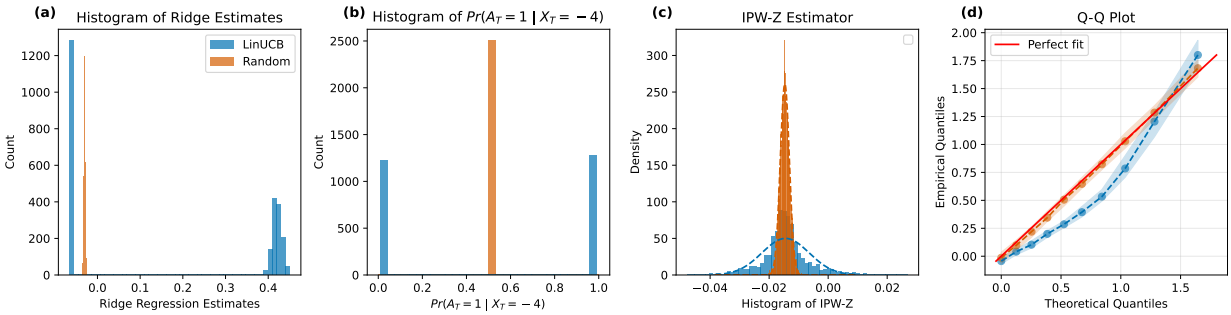

Figure 1: Example of non-convergence. We independently run LinUCB and pur Random algorithm for 10000 steps for 2500 times. **(a)** the last-step ridge regression estimator of $\boldsymbol{\theta}_1$. **(b)** the last-step sampling probability at context $\boldsymbol{x} = -4$. **(c)**, the last-step proposed IPW-Z estimator for inference target 1. **(d)** the QQ plot of the standardized empirical distribution of the last-step IPW-Z estimator compared with standard Gaussian.

time t. This policy class is broad and encompasses many commonly used bandit algorithms—such as ϵ-greedy [56], UCB [6, 20], and Thompson sampling [3, 55]—and is widely adopted in adaptive experimental designs in practice [5, 43, 64].

The following theorem provides a sufficient condition under which policies of the form (4.1) satisfy the policy convergence condition in general environments, without requiring a well-specified reward model. The proof is given in Appendix B.4.

Theorem 4.1. *Suppose the behavior policy $\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1})$ takes the form of (4.1). Then for any $a \in \mathcal{A}$, as long as*

(i) $\widehat{\boldsymbol{\beta}}_t \xrightarrow{p} \boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^{d_\beta}$,

(ii) $\pi(a|\boldsymbol{X}_t, \cdot)$ is continuous at $\boldsymbol{\beta}^*$ a.e. $\boldsymbol{X}_t$,

the behavior policy converges at action a as in Definition 3.1, with the limit policy $\bar{\pi}(a|\boldsymbol{x}) = \pi(a|\boldsymbol{x}, \boldsymbol{\beta}^*)$.

We offer some remarks on conditions (i) and (ii) in Theorem 4.1. Condition (i) is notably general: it only requires $\widehat{\boldsymbol{\beta}}_t$ converges to a deterministic limit $\boldsymbol{\beta}^*$, which need not correspond to any parameter from a correctly specified model, nor be optimal in any sense. In Sections 4.1 to **??**, we present several examples where the summary statistic $\widehat{\boldsymbol{\beta}}_t$ converges to different limits through a variety of mechanisms. Condition (ii) imposes only local continuity of the policy function π at $\boldsymbol{\beta}^*$, rather than requiring global continuity. This requirement is important: even if $\widehat{\boldsymbol{\beta}}_t \xrightarrow{p} \boldsymbol{\beta}^*$, discontinuity of π at $\boldsymbol{\beta}^*$ may result in irregular behavior of the policy and prevent convergence.

In the below, we apply Theorem 4.1 and show several classes of behavior policies satisfy the policy convergence condition in the sense of Definition 3.1. For each policy, we identify suitable summary statistics and verify the continuity of the policy function at the limiting value. Importantly, all these policies converge in general environments under mild regularity conditions, without requiring a well-specified reward model.

4.1 Multi-armed bandit ignoring context

In real-world bandit deployments, simple algorithms like multi-armed bandits (MAB) are often preferred, especially in early trials when prior information is limited. By ignoring context, MAB avoids reward model misspecification, which helps to manage variance and prevent poor decisions in complex systems. In this section, we show that under mild conditions, common MAB algorithms additionally satisfy the policy convergence condition, enabling valid post-study inference.

For any action $a \in \mathcal{A}$ and time t, denote

$$\widehat{\mu}_{a,t} := \frac{\sum_{\tau=1}^{t} 1_{\{A_\tau = a\}} Y_\tau}{N_{a,t}}, \quad N_{a,t} := \sum_{\tau=1}^{t} 1_{\{A_\tau = a\}}. \tag{4.2}$$

Consider the following MAB algorithms with a minimum sampling probability:

- **The ϵ-greedy algorithm:**

$$\pi_t^{\epsilon\text{-greedy}}(a|\mathcal{H}_{t-1}) = \begin{cases} 1 - \frac{K-1}{K}\epsilon, & \text{if } i = \text{argmax}_i\, \widehat{\mu}_{i,t-1}, \\ \frac{1}{K}\epsilon, & \text{otherwise.} \end{cases} \tag{4.3}$$

- **The UCB algorithm:**

$$\pi_t^{\text{UCB}}(a|\mathcal{H}_{t-1}) = \begin{cases} 1 - (K-1)\pi_{\min}, & \text{if } i = \text{argmax}_i \left\{ \widehat{\mu}_{i,t-1} + \sqrt{\frac{C_t}{N_{i,t-1}}} \right\}, \\ \pi_{\min}, & \text{otherwise.} \end{cases} \tag{4.4}$$

  where $\{C_t\}_{t\geq 1}$ is any deterministic sequence such that $\lim_{t\to\infty} \frac{C_t}{t} = 0$.

- **The Thompson Sampling algorithm:**

$$\left(\pi_t^{\text{TS}}(a|\mathcal{H}_{t-1})\right)_{a\in\mathcal{A}} = \text{Clip}\left(\left(\bar{\pi}_t^{\text{TS}}(a|\mathcal{H}_{t-1})\right)_{a\in\mathcal{A}}\right), \tag{4.5}$$

  where

  - $\bar{\pi}_t^{\text{TS}}(a|\mathcal{H}_{t-1}) = \mathbb{E}_{(\mu_i')_{i\in\mathcal{A}} \sim \mathcal{N}(\boldsymbol{\mu}_{t-1}^{\text{post}}, \boldsymbol{\Sigma}_{t-1}^{\text{post}})} 1_{\{\forall i \neq a, \mu_i' < \mu_a'\}}$ is the posterior probability of action a being optimal, with

$$\boldsymbol{\mu}_{t-1}^{\text{post}} = \left(\mu_{a,t-1}^{\text{post}}\right)_{a\in\mathcal{A}}, \quad \boldsymbol{\Sigma}_{t-1}^{\text{post}} = \text{diag}\left((\sigma_{a,t-1}^{\text{post}})^2\right), \tag{4.6}$$

$$\mu_{a,t}^{\text{post}} := \left(\frac{1}{\sigma_0^2} + \frac{N_{a,t}}{\sigma^2}\right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{N_{a,t}\widehat{\mu}_{a,t}}{\sigma^2}\right), \quad (\sigma_{a,t}^{\text{post}})^2 := \left(\frac{1}{\sigma_0^2} + \frac{N_{a,t}}{\sigma^2}\right)^{-1}. \tag{4.7}$$

    Here, $\mu_0 \in \mathbb{R}$ and $\sigma^2, \sigma_0^2 > 0$ are fixed algorithm parameters representing the prior mean, prior variance, and observation noise variance, respectively.

  - The mapping $\text{Clip} : \mathbb{R}^K \to \mathbb{R}^K$ adjusts a probability distribution over K discrete actions to ensure that each coordinate is lower bounded by $\pi_{\min}$ (details in Appendix A.1).

Define the average reward of arm a by $\mu_a^* = \mathbb{E}[R_t \mid A_t = a]$, and let $a^* = \text{argmax}_a\, \mu_a^*$. The following proposition shows that all three algorithms above satisfy the policy convergence condition under a nonzero suboptimality gap. Its proof is in Appendix B.5.

Proposition 4.1. Suppose $\sup_{a\in\mathcal{A}} \mathrm{Var}(Y(a)) \le \sigma_Y^2$ for a universal constant σ_Y^2. Then as long as the suboptimal gap $\Delta = \mu_{a^*}^* - \max_{a'\neq a^*} \mu_{a'}^* > 0$, the policies $\{\pi_t^{\epsilon\text{-greedy}}\}_{t\ge1}$, $\{\pi_t^{\mathrm{UCB}}\}_{t\ge1}$ and $\{\pi_t^{\mathrm{TS}}\}_{t\ge1}$, defined in (4.3), (4.4) and (4.5), respectively, satisfy the policy convergence condition in Definition 3.1.

Choice of summary statistics. Proposition 4.1 follows from Theorem 4.1 by identifying the key summary statistics used by each policy. While the most natural statistics vary across policies, natural choices for the three cases are suitable functions of the arm-specific sample mean $\widehat{\mu}_{a,t}$ and the inverse number of pulls $1/N_{a,t}$. Intuitively, under a minimum sampling probability, each action is chosen infinitely often, and these context-free statistics converge to their population-level values. This ensures that the policies stabilize over time and satisfy the policy convergence condition required for inference. A more rigorous justification for this argument can be found in Lemma B.6 in Appendix B.5.

Beyond minimum sampling probability. Proposition 4.1 focuses on policies with a minimum sampling probability, primarily because our goal is to establish inference guarantees without assuming a well-specified reward model, as in the setting of Theorem 3.3. Technically, a minimum sampling probability is not required for policy convergence in general; some algorithms exhibit stability even without this condition. See, for instance, [30, 35], where the authors analyze the stability properties of the UCB algorithm without requiring a minimum sampling probability.

4.2 Policies based on the IPW-Z estimator

In certain settings, an ideal policy takes the form $\pi(\cdot \mid \boldsymbol{X}_t, \{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}})$, where $\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}$ are our inferential target parameters defined in (2.1). These parameters often inform optimal decision-making within a parametric class, either with respect to maximizing expected reward or achieving other objectives. For instance, in the case of misspecified linear bandits (Example 1), the best linear policy that maximizes expected reward can be parameterized by $\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}$, which encode the best linear approximations of the reward functions [15].

Motivated by this setup, it is natural to consider behavior policies of the form (4.1), where the statistics $\widehat{\boldsymbol{\beta}}_{t-1}$ consist of the IPW-Z estimators defined by (2.6), along with additional converging algorithm parameters. These policies often serve as good approximations to an ideal policy within the specified model class. The following proposition shows that, under mild conditions, such behavior policies converge to a limiting policy parameterized by $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_a^*)_{a\in\mathcal{A}}$.

Proposition 4.2. Under the assumptions of Theorem 3.3, suppose the behavior policy at each time t takes the form

$$\pi(\cdot \mid \boldsymbol{X}_t, \{\widehat{\boldsymbol{\theta}}_a^{(t-1)}\}_{a\in\mathcal{A}}, \boldsymbol{\gamma}_{t-1}), \tag{4.8}$$

where $\{\widehat{\boldsymbol{\theta}}_a^{(t)}\}_{a\in\mathcal{A},t\ge1}$ is the IPW-Z estimator defined by (2.6), and $\{\boldsymbol{\gamma}_t\}_{t\ge1}$ is a deterministic parameter sequence such that $\boldsymbol{\gamma}_t \in \mathbb{R}^{d_\gamma}$, $\lim_{t\to\infty} \boldsymbol{\gamma}_t = \boldsymbol{\gamma}^*$ for some $\boldsymbol{\gamma}^* \in \mathbb{R}^{d_\gamma}$. Assume $\pi : \mathcal{X}\times\mathbb{R}^{Kd}\times\mathbb{R}^{d_\gamma} \to \Delta(\mathcal{A})$ is continuous in its second and third argument at $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ for almost every $\boldsymbol{x} \in \mathcal{X}$ under the distribution of $\boldsymbol{X}_t$. Then the behavior policy satisfies the policy convergence condition in Definition 3.1, with the limit policy $\bar{\pi}(a|\boldsymbol{x}) = \pi(a|\boldsymbol{x}, \boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$.

The proof of Proposition 4.2 is straightforward: Let $\widehat{\boldsymbol{\beta}}_t = (\widehat{\boldsymbol{\theta}}^{(t)}, \boldsymbol{\gamma}_t)$, where $\widehat{\boldsymbol{\theta}}^{(t)} = (\widehat{\boldsymbol{\theta}}_a^{(t)})_{a\in\mathcal{A}}$. From Theorem 3.3, we deduce that $\widehat{\boldsymbol{\beta}}_t \xrightarrow{p} \boldsymbol{\beta}^* := (\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$, which implies that condition (i) of Theorem 4.1

holds in this setting. Under an additional condition of policy continuity, Theorem 4.1 can then be invoked to establish policy convergence.

Remark 4.1. The behavior policy studied in [15]—an ϵ-greedy policy combined with a weighted online LS estimator in the setting of misspecified linear bandits (Example 1)—is a special case of the policies considered in Proposition 4.2, under their assumption of an almost surely nonzero margin. As such, it satisfies the policy convergence condition in Definition 3.1, and the asymptotic normality of the parameter estimates can be derived from Theorem 3.3. See Appendix A.2 for details.

4.3 Boltzmann exploration

Boltzmann exploration—also known as softmax or Gibbs exploration—is a widespread method in bandit and reinforcement learning literature. It chooses actions by sampling from a probability distribution proportional to the exponential of estimated action values with a temperature parameter [13].

In this section, we demonstrate that Boltzmann exploration with different choices of action value estimates has policy convergence with sufficiently large temperature. The Boltzmann exploration policy w.r.t. a given estimator $\boldsymbol{\theta} = (\boldsymbol{\theta}_a)_{a\in\mathcal{A}}$ is defined by

$$\pi^{\gamma}(a \mid \boldsymbol{X}_t, \boldsymbol{\theta}) = \frac{\exp\left(\langle \boldsymbol{\theta}_a, \boldsymbol{X}_t\rangle / \gamma\right)}{\sum_{a'\in\mathcal{A}} \exp\left(\langle \boldsymbol{\theta}_{a'}, \boldsymbol{X}_t\rangle / \gamma\right)}, \tag{4.9}$$

where $\gamma > 0$ is the temperature parameter. Further, we clip the Boltzmann exploration policy by $\pi_{\min} \in (0, 1)$, and we denote the clipped Boltzmann exploration policy by $\widetilde{\pi}^{\gamma}$.

The convergence is conditioned on sufficient large temperature, which provides smoothness in the action sampling probability. By writing the esitmator $\boldsymbol{\theta}$ as an stochastic approximation process, such smoothness gives rise to contraction mapping and enables the application of Theorem 4.1. Specifically, we consider two estimators: ridge regression estimator and stochastic gradient descent estimator.

4.3.1 Ridge regression

The ridge regression estimator at each time t is defined by $\widehat{\boldsymbol{\theta}}_t^{\text{Ridge}} = (\widehat{\boldsymbol{\theta}}_{t,a}^{\text{Ridge}})_{a\in\mathcal{A}}$, where

$$\widehat{\boldsymbol{\theta}}_{t,a}^{\text{Ridge}} = \left(\lambda I + \sum_{i=1}^{t-1} 1_{\{A_i=a\}} \boldsymbol{X}_i \boldsymbol{X}_i^{\top}\right)^{-1} \left(\sum_{i=1}^{t-1} 1_{\{A_i=a\}} \boldsymbol{X}_i Y_i\right). \tag{4.10}$$

Assumption 4.1. We make the following regularity assumptions:

- (A.1) $\|\boldsymbol{X}_t\|_2 \le M$ a.s.
- (A.2) $\mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^{\top}] \succeq \sigma_{\min}^2 I$
- (A.3) $\|\boldsymbol{\theta}_a^*\|_2 \le R_\theta$ for all $a \in \mathcal{A}$ with $\boldsymbol{\theta}_a^*$ being the the solution to (2.1) with $\boldsymbol{g}$ being the score function defined in (2.2).

Theorem 4.2. *Suppose Assumption 4.1 holds. Let $\widehat{\boldsymbol{\theta}}_t^{Ridge}$ be the ridge regression estimator based on the dataset collected by the Boltzmann exploration policy $\widetilde{\pi}^{\gamma}(\cdot \mid \boldsymbol{X}_t, \widehat{\boldsymbol{\theta}}_t^{Ridge})$. We have for all $a \in \mathcal{A}$, $\widehat{\boldsymbol{\theta}}_{t,a}^{Ridge} \xrightarrow{p} \bar{\boldsymbol{\theta}}_a$, for some $\{\bar{\boldsymbol{\theta}}_a\}_{a\in\mathcal{A}}$, if*

$$\gamma \geq 2|\mathcal{A}|(|\mathcal{A}|+1)M^3 \max\left\{\frac{2}{\pi_{\min}\sigma_{\min}^2}, \frac{8M^2R_\theta}{\pi_{\min}^2\sigma_{\min}^4}\right\} \max\{R_\theta, 1\}. \tag{4.11}$$

The proof is given in Appendix B.6.

Many online contextual bandit algorithms, such as LinUCB [20] and Thompson Sampling [55] rely on the ridge regression estimator as a sufficient statistics for their behavior policy. The following theorem states a necessary condition for the ridge regression estimator $\widehat{\boldsymbol{\theta}}_t^{\text{Ridge}}$ to converge under a given behavior policy $\pi(a \mid \boldsymbol{X}_t, \boldsymbol{\theta})$. Equation (4.12) is a critical condition for non-convergence. Intuitively, it states that there is no fixed point of the estimating equation for the ridge regression estimator. We refer to the `NC-Hard1` environment in Section 6 as a simple example that satisfies this condition.

Theorem 4.3. *Suppose that the assumptions in Theorem 4.2 hold. Suppose that the behavior policy $\pi(a \mid \boldsymbol{X}_t, \boldsymbol{\theta})$ satisfies that*

- *for any $\boldsymbol{\theta}_a \in \mathbb{R}^d$,*

$$\|\boldsymbol{\theta}_a - \Sigma_a^{-1}(\boldsymbol{\theta}_a)\varphi_a(\boldsymbol{\theta}_a)\|_2 \geq c, \tag{4.12}$$

  where $\Sigma_a(\boldsymbol{\theta}_a) = \mathbb{E}_{A_t\sim\pi(\cdot|\boldsymbol{X}_t,\boldsymbol{\theta})}[1_{A_t=a}\boldsymbol{X}_t\boldsymbol{X}_t^\top]$ and $\varphi_a(\boldsymbol{\theta}_a) = \mathbb{E}_{A_t\sim\pi(\cdot|\boldsymbol{X}_t,\boldsymbol{\theta})}[1_{A_t=a}\boldsymbol{X}_t Y_t]$.

- *$\pi(a \mid \boldsymbol{x}, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ given any $\boldsymbol{x} \in \mathcal{X}$.*

Then, the ridge regression estimator $\widehat{\boldsymbol{\theta}}_t^{Ridge}$ does not converge to any $\bar{\boldsymbol{\theta}}$ in probability under the behavior policy $\pi(a \mid \boldsymbol{X}_t, \boldsymbol{\theta})$.

The proof is given by Appendix B.8.

4.3.2 Stochastic gradient descent

In addition to ridge regression estimator, we show that a family of stochastic gradient descent (SGD) estimators also has convergence under the Boltzmann exploration policy. Let $\widehat{\boldsymbol{\theta}}_{t,a}^{\text{SGD}}$ be the SGD estimator that admits the following update rule:

$$\widehat{\boldsymbol{\theta}}_{t,a}^{\text{SGD}} = \widehat{\boldsymbol{\theta}}_{t-1,a}^{\text{SGD}} + \eta_t 1\{A_t = a\}\boldsymbol{h}\left(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_{t-1,a}^{\text{SGD}}\right), \tag{4.13}$$

where η_t is the learning rate at time t and $\boldsymbol{h}$ is a function parameterized by $\widehat{\boldsymbol{\theta}}_{t,a}^{\text{SGD}}$. This SGD estimator is widely used in practice especially when complex models such as neural networks are involved [73].

Further, we write $\widehat{\boldsymbol{\theta}}_t^{\text{SGD}} = (\widehat{\boldsymbol{\theta}}_{t,1}^{\text{SGD}}, \ldots, \widehat{\boldsymbol{\theta}}_{t,|\mathcal{A}|}^{\text{SGD}})$, which follows the update rule:

$$\widehat{\boldsymbol{\theta}}_t^{\text{SGD}} = \widehat{\boldsymbol{\theta}}_{t-1}^{\text{SGD}} + \eta_t \begin{pmatrix} 1\{A_t = 1\}\boldsymbol{h}\left(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_{t-1,1}^{\text{SGD}}\right) \\ \vdots \\ 1\{A_t = |\mathcal{A}|\}\boldsymbol{h}\left(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_{t-1,|\mathcal{A}|}^{\text{SGD}}\right) \end{pmatrix}. \tag{4.14}$$

We slightly abuse the notation and let

$$\boldsymbol{\phi}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) \coloneqq \left(1\{A_t = 1\}\boldsymbol{h}^\top(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}), \ldots, 1\{A_t = |\mathcal{A}|\}\boldsymbol{h}^\top(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta})\right)^\top. \tag{4.15}$$

Unlike the ridge regression estimator, where the estimator is guaranteed to be bounded with high probability, SGD estimator does not have boundedness in general. To ensure the stability of the estimator, we make the following assumption (A5 [10]).

Assumption 4.2 (Globally Asymptotically Stable Equilibrium.)**.** Let $\bar{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\phi}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}) \mid \widehat{\boldsymbol{\theta}}_t^{\mathrm{SGD}} = \boldsymbol{\theta}]$. We assume that the function $\bar{\boldsymbol{\phi}}_c(\boldsymbol{\theta}) \coloneqq \boldsymbol{\phi}(c\boldsymbol{\theta})/c, c \geq 1, \boldsymbol{\theta} \in \mathbb{R}^d$, satisfies that $\bar{\boldsymbol{\phi}}_c(\boldsymbol{\theta}) \to \boldsymbol{\phi}_\infty(\boldsymbol{\theta})$ as $c \to \infty$, uniformly on compacts for some $\boldsymbol{\phi}_\infty \in C(\mathbb{R}^d)$. Furthermore, the o.d.e.

$$\dot{\boldsymbol{\theta}}(t) = \boldsymbol{\phi}_\infty(\boldsymbol{\theta}(t))$$

has the origin as its unique globally asymptotically stable equilibrium (Definition A.2).

$\boldsymbol{h}(x, y; \boldsymbol{\theta})$ has Lipschitz constant $L_{\boldsymbol{h}} < \infty$ w.r.t. $\boldsymbol{\theta}$ for all $x, y \in \mathcal{X} \times \mathbb{R}$.

Theorem 4.4. *Let Assumption 4.1 and 4.2 hold. Let $\widehat{\boldsymbol{\theta}}_{t,a}^{SGD}$ be the SGD estimator based on the dataset collected by the Boltzmann exploration policy $\pi_t^\gamma(a \mid \boldsymbol{X}_t, \widehat{\boldsymbol{\theta}}_t^{SGD})$. Assume that $\sum_{t=1}^\infty \eta_t = \infty$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$. For sufficiently large γ, we have*

$$\widehat{\boldsymbol{\theta}}_{t,a}^{SGD} \xrightarrow{p} \bar{\boldsymbol{\theta}}_a \quad \forall a \in \mathcal{A},$$

for some $\{\bar{\boldsymbol{\theta}}_a\}_{a \in \mathcal{A}}$.

The proof is given by Appendix B.7. The following proposition shows that Assumption 4.2 holds for the two estimators corresponding to the two inference targets in Example 1 and 2.

Proposition 4.3. Assumption 4.2 holds for the following two estimators under Boltzmann exploration policy with any $\gamma > 0$.

- $h(\boldsymbol{x}, y; \boldsymbol{\theta}_a) = \boldsymbol{x}y - \boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\theta}_a$ for the misspecified linear bandits in Example 1.
- $h(\boldsymbol{x}, y; \boldsymbol{\theta}_a) = \boldsymbol{x}y - (\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a$ for the noisy context model in Example 2.

5 Examples

In this section, we revisit the three examples from Section 2 and establish statistical inference guarantees by applying Theorem 3.2 and 3.3 or suitably adapting them to relevant practical settings.

5.1 Misspecified Linear Bandits

We first consider Example 1, where the target parameter $\boldsymbol{\theta}_a^*$ is defined by (2.1) with the score function $\boldsymbol{g}$ in (2.2). A natural estimator for $\boldsymbol{\theta}_a^*$ that satisfies (2.6) is

$$\widehat{\boldsymbol{\theta}}_a^{(T)} = \left(\frac{1}{T}\sum_{t=1}^T \frac{1_{\{A_t=a\}}}{\pi_t(A_t)}\boldsymbol{X}_t\boldsymbol{X}_t^\top\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \frac{1_{\{A_t=a\}}}{\pi_t(A_t)}\boldsymbol{X}_t Y_t\right). \tag{5.1}$$

In this setting, Assumptions 3.1-3.3, which form part of the conditions for Theorem 3.3, are implied by the following regularity assumption.

Assumption 5.1. There exist constants $R_\theta, M > 0$ such that $\sup_{a\in\mathcal{A}} \|\boldsymbol{\theta}_a^*\|_2 < R_\theta$, $\|\boldsymbol{X}_t\|_2 \le M$. Moreover, $\boldsymbol{\Sigma}_X := \mathbb{E}[\boldsymbol{X}_t\boldsymbol{X}_t^\top]$ is invertible, and $\sup_{a\in\mathcal{A}} \mathbb{E}[Y_t(a)^4] < \infty$.

With Assumption 5.1 in place, together with the remaining conditions of Theorem 3.3, the joint asymptotic normality of $\widehat{\boldsymbol{\theta}}^{(T)} = \big((\widehat{\boldsymbol{\theta}}_1^{(T)})^\top, \ldots, (\widehat{\boldsymbol{\theta}}_K^{(T)})^\top\big)^\top$ and a consistent estimator for its asymptotic variance follow directly from Theorem 3.3 and Proposition 3.1, as stated below.

Corollary 5.1. Suppose Assumption 2.1 and 5.1 holds. Consider any behavior policy that converges to a limiting policy $\bar{\pi}$ as in Definition 3.1 and satisfies Assumption 3.4 for every action $a \in \mathcal{A}$. Then, as $T \to \infty$, (3.3) holds with the joint asymptotic variance $\boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{\Sigma}_1^*, \ldots, \boldsymbol{\Sigma}_K^*)$, where for all $a \in \mathcal{A}$,

$$\boldsymbol{\Sigma}_a^* = \boldsymbol{\Sigma}_X^{-1} \cdot \mathbb{E}\left[\frac{(Y_t(a) - \boldsymbol{X}_t^\top \boldsymbol{\theta}_a^*)^2}{\bar{\pi}(a|\boldsymbol{X}_t)} \boldsymbol{X}_t\boldsymbol{X}_t^\top\right] \cdot \boldsymbol{\Sigma}_X^{-1}. \tag{5.2}$$

A consistent estimator for $\boldsymbol{\Sigma}_a^*$ is given by (3.4), with

$$\widehat{\dot{\boldsymbol{G}}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(A_t)} \boldsymbol{X}_t\boldsymbol{X}_t^\top, \quad \widehat{\boldsymbol{I}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2} (Y_t - \boldsymbol{X}_t^\top \widehat{\boldsymbol{\theta}}_a^{(T)})^2 \boldsymbol{X}_t\boldsymbol{X}_t^\top. \tag{5.3}$$

Remark 5.1. In Corollary 5.1, a simpler consistent estimator for the asymptotic variance is obtained by replacing $\widehat{\dot{\boldsymbol{G}}}_{a,T}$ in (3.4) with $\widehat{\boldsymbol{\Sigma}}_X := \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{X}_t\boldsymbol{X}_t^\top$. This alternative remains consistent since both matrices consistently estimate $\boldsymbol{\Sigma}_X$.

In the special case where the behavior policy is an ϵ-greedy policy combined with the weighted online LS estimator, Corollary 5.1 yields the same asymptotic normality result as [15]; see also the discussion following Theorem 3.2. The variance estimator used in [15] corresponds to the simplified plug-in form described in Remark 5.1. In addition, Corollary 5.1 holds under generally weaker technical assumptions, such as not requiring the context $\boldsymbol{X}_t$ to be a continuous random vector.

5.2 Bandits with Noisy Contexts

We now revisit Example 2. Recall that the potential outcome satisfies $Y_t(a) = \boldsymbol{S}_t^\top \boldsymbol{\theta}_a^* + \eta_t$ with unobserved state $\boldsymbol{S}_t$, while the observed context is the noisy proxy $\boldsymbol{X}_t = \boldsymbol{S}_t + \boldsymbol{\epsilon}_t$. The noise $\boldsymbol{\epsilon}_t$ satisfies $\mathbb{E}[\boldsymbol{\epsilon}_t|\boldsymbol{S}_t] = \boldsymbol{0}$, and $\mathrm{Var}(\boldsymbol{\epsilon}_t|\boldsymbol{S}_t) = \boldsymbol{\Sigma}_e$ for a constant matrix $\boldsymbol{\Sigma}_e \in \mathbb{R}^{d\times d}$. No parametric assumptions are imposed on the distribution of $\boldsymbol{\epsilon}_t$. In addition, assume $\mathbb{E}[\eta_t \mid \boldsymbol{S}_t, \boldsymbol{X}_t] = 0$ and define $\sigma_\eta^2 = \mathrm{Var}(\eta_t \mid \boldsymbol{S}_t, \boldsymbol{X}_t)$.

We first consider the case where the contextual error variance $\boldsymbol{\Sigma}_e$ is known. Then the score function reduces to (2.3). An estimator of $\boldsymbol{\theta}_a^*$ satisfying (2.6) is

$$\widehat{\boldsymbol{\theta}}_a^{(T)} = \left[\frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(A_t)} (\boldsymbol{X}_t\boldsymbol{X}_t^\top - \boldsymbol{\Sigma}_e)\right]^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1_{\{A_t=a\}}}{\pi_t(A_t)} \boldsymbol{X}_t Y_t\right). \tag{5.4}$$

To infer the target parameter $\boldsymbol{\theta}_a^*$ via Theorem 3.3, we impose the following regularity condition, which guarantees Assumptions 3.1–3.3 required for the theorem.

Assumption 5.2. There exist constants $R_\theta, M, M_\eta > 0$ such that $\sup_a \|\boldsymbol{\theta}_a^*\|_2 < R_\theta$, $\max\{\|\boldsymbol{X}_t\|_2, \|\boldsymbol{S}_t\|_2\} \le M$, $\mathbb{E}[\eta_t^4|\boldsymbol{S}_t, \boldsymbol{X}_t] \le M_\eta$. In addition, $\boldsymbol{\Sigma}_S := \mathbb{E}\boldsymbol{S}_t\boldsymbol{S}_t^\top$ is invertible.

The following corollary, obtained from Theorem 3.3 and Proposition 3.1, provides inference for $\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}$ in this setting. The proof establishing the variance expression in the corollary is provided in Appendix B.9.

Corollary 5.2. Suppose Assumption 2.1 and 5.2 holds. Consider any behavior policy that converges to a limiting policy $\bar{\pi}$ as in Definition 3.1 and satisfies Assumption 3.4 for every $a \in \mathcal{A}$. Then, as $T \to \infty$, (3.3) holds with $\boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{\Sigma}_1^*, \ldots, \boldsymbol{\Sigma}_K^*)$, where for all $a \in \mathcal{A}$,

$$\boldsymbol{\Sigma}_a^* = \boldsymbol{\Sigma}_S^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{\Sigma}_S^{-1}, \quad \bar{\boldsymbol{I}}_a := \mathbb{E}\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\big[\boldsymbol{h}_a(\boldsymbol{X}_t,\boldsymbol{S}_t)\boldsymbol{h}_a(\boldsymbol{X}_t,\boldsymbol{S}_t)^\top + \sigma_\eta^2\boldsymbol{X}_t\boldsymbol{X}_t^\top\big], \tag{5.5}$$

and $\boldsymbol{h}_a(\boldsymbol{x},\boldsymbol{s}) := (\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^* - \boldsymbol{x}\boldsymbol{s}^\top\boldsymbol{\theta}_a^*$. A consistent estimator for $\boldsymbol{\Sigma}_a^*$ is in (3.4), with

$$\widehat{\boldsymbol{G}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}(\boldsymbol{X}_t\boldsymbol{X}_t^\top - \boldsymbol{\Sigma}_e), \widehat{\boldsymbol{I}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2}\Big[(\boldsymbol{X}_t\boldsymbol{X}_t^\top - \boldsymbol{\Sigma}_e)\widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{X}_tY_t\Big]^{\otimes 2}. \tag{5.6}$$

In practice, the variance of the contextual error $\boldsymbol{\Sigma}_e$ is often unknown and must be estimated from auxiliary data. In this case, $\widehat{\boldsymbol{\theta}}_a^{(T)}$ is obtained from (2.6) by replacing $\boldsymbol{\Sigma}_e$ with an estimator $\widehat{\boldsymbol{\Sigma}}_e$. The derivation and corresponding inference results are provided in Appendix A.4.

5.3 Off-policy Evaluation with Adaptively Collected Data

We now consider Example 3, where the target parameter $V^* = \sum_{a\in\mathcal{A}}\boldsymbol{\theta}_a^*$, and $\boldsymbol{\theta}_a^* = \mathbb{E}[\pi^e(a|\boldsymbol{X}_t)Y_t]$ solves (2.1) with an arm-specific score function (2.4). We estimate V^* by $\widehat{V}^{(T)} = \sum_{a\in\mathcal{A}}\widehat{\boldsymbol{\theta}}_a^{(T)}$, where

$$\widehat{\boldsymbol{\theta}}_a^{(T)} = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}\pi^e(a|\boldsymbol{X}_t)Y_t\right) \tag{5.7}$$

satisfies (2.6).

The asymptotic distribution of $\widehat{V}^{(T)}$ can be derived from the joint distribution of $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}}$ via Theorem 3.3, enabling statistical inference. Specifically, we impose the following condition, which encompasses Assumptions 3.1–3.3 required for the theorem.

Assumption 5.3. $\mathbb{E}Y_t^4 < \infty$, and there exist a constant $R_\theta > 0$ such that $\sup_a \|\boldsymbol{\theta}_a^*\|_2 < R_\theta$.

The corollary below informs valid inference of the target parameter V^*. It follows from Theorem 3.3 and Proposition 3.1.

Corollary 5.3. Suppose Assumption 2.1 and 5.3 holds. Consider any behavior policy that converges to a limiting policy $\bar{\pi}$ as in Definition 3.1 and satisfies Assumption 3.4 for every $a \in \mathcal{A}$. Then, as $T \to \infty$,

$$\sqrt{T}\big(\widehat{V}^T - V^*\big) \xrightarrow{d} \mathcal{N}\left(0, \sum_{a\in\mathcal{A}}\mathbb{E}\frac{\big(\pi^e(a|\boldsymbol{X}_t)Y_t - \boldsymbol{\theta}_a^*\big)^2}{\bar{\pi}(a|\boldsymbol{X}_t)}\right). \tag{5.8}$$

A consistent variance estimator is

$$\sum_{a\in\mathcal{A}}\Big(\widehat{\dot{\boldsymbol{G}}}_{a,T}\Big)^{-2}\widehat{\boldsymbol{I}}_{a,T}, \tag{5.9}$$

where for all $a \in \mathcal{A}$,

$$\widehat{\dot{\boldsymbol{G}}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}, \quad \widehat{\boldsymbol{I}}_{a,T} = \frac{1}{T}\sum_{t=1}^{T}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2}\big(\pi^e(a|\boldsymbol{X}_t)Y_t - \widehat{\boldsymbol{\theta}}_a^{(T)}\big)^2. \tag{5.10}$$

Similar to Example 1, the variance estimator can be simplified by replacing $\widehat{\dot{\boldsymbol{G}}}_{a,T}$ in (5.9) with 1, since $\widehat{\dot{\boldsymbol{G}}}_{a,T} \xrightarrow{p} 1$ in this setting. See Lemma B.12 for details.

Finally, we note that [8, 68] have proposed various alternative method for the same task based on weighting the AIPW estimator. See Section 6.2 for an empirical comparison between these methods.

6 Simulation Studies

In this section, we conduct simulation studies to evaluate the proposed inference methods under three previously introduced inference targets: Target 1 (misspecified linear bandits with score function $\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{\theta} - \boldsymbol{x}y$), Target 2 (linear bandits with noisy contexts with score function $(\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{\Sigma}_e)\boldsymbol{\theta} - \boldsymbol{x}y$), and Target 3 (offline policy valuation) with score function $\pi^e(a|\boldsymbol{x})y - \boldsymbol{\theta}$. Throughout, we use linear working models, despite the true mean reward function $y(x,a) \coloneqq \mathbb{E}[Y_t \mid A_t = a, \boldsymbol{X}_t = x]$ being non-linear, allowing us to assess the robustness of the inference methods under misspecifications.

We examine five simulation environments: `NC-hard1`, `NC-hard2`, `NC-Gaussian`, `MS-Polynomial`, `MS-Neural`. In the first three environments, the reward function w.r.t. the underlying true context is linear, and the model misspecification arises from noisy context. Specifically, `NC-hard1`, `NC-hard2` are deliberately designed challenging environments so that the ordinary least squares estimator either oscillates or converges to a set of parameters randomly. In `NC-Gaussian`, all variables and parameters are jointly Gaussian. The remaining two environments introduce direct non-linear model misspecification: a polynomial reward function in `MS-Polynomial` and a neural network reward function in `MS-Neural`. More details are provided in Appendix E.1.

We evaluate the inference methods using datasets generated by four distinct algorithms: pure random selection (`Random`), Boltzmann exploration (`BE`), a greedy algorithm with respect to the IPW-Z estimator in (2.6) (`IPW-Z`), and stochastic gradient descent (`SGD`). These algorithms are shown to have policy convergence in Section 4. All algorithms implement clipping of the sampling probabilities to ensure bounded inverse probability weights.

6.1 Results

We first examine the empirical coverage of nominal confidence intervals (95%, 90%, 80%, 70%, 60%, and 50%) for inference target 1 across the five environments. Results are based on 2500 Monte Carlo simulations. Figure 2 shows that the proposed inference method consistently achieves the desired coverage across all algorithms and environments. A slight undercoverage (around 90% for nominal 95%) is observed for non-linear environments (`MS-Polynomial`, `MS-Neural`) when using IPW-Z and Boltzmann exploration. Similar trends appear for Target 2 in noisy context environments (Figure 3, panels a and b).

We further analyze how the temperature parameter γ and the minimum sampling probability p_0 affect inference performance. Figure 3(c) shows that increasing temperature γ reduces the variance of the IPW-Z estimator. Figure 3(d) demonstrates that higher minimum sampling probabilities p_0 significantly improve empirical coverage of the 95% confidence interval. Conversely, low values of γ or p_0 cause excessively large inverse probability weights, hindering convergence of the asymptotic variance estimator.

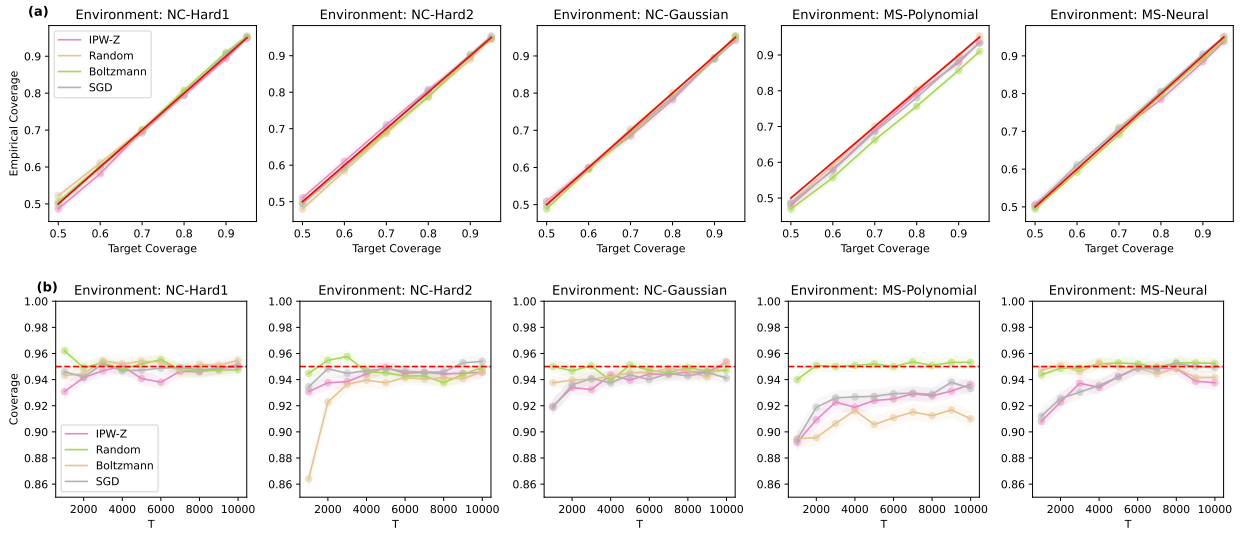

Figure 2: **(a)** Empirical coverage vs. target confidence levels (95%, 90%, ..., 50%) for inference target 1 across five environments. **(b)** Empirical coverage of 95% confidence intervals as a function of time steps (over 10000 steps). Results averaged across 2500 Monte Carlo simulations.

6.2 OPE Results

In the OPE setting, we compare the proposed inference method with the CADR (Contextual Adaptive Doubly Robust) method [8] under various choice of prediction model including linear model, tree-based model, and a dumpy model that always outputs 0. We run CADR on the same dataset collected by Boltzmann exploration w.r.t. Ridge regression in five environments introduced above. In Figure 4 (a), we show that both CADR and our proposed inference method achieve empirical coverage close to the target coverage. In Figure 4 (b), we show that our proposed esimtator has lower variance than CADR under `NC-Hard2` and `MS-Polynomial` environments. This improvement is due to the unstable variance estimator of CADR.

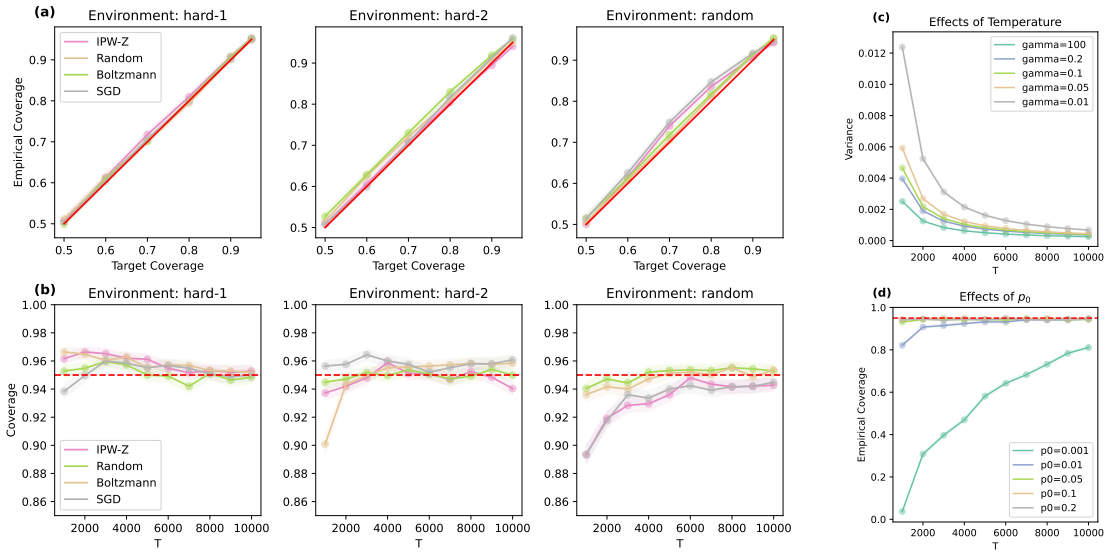

Figure 3: **(a)** Empirical coverages of 95%, 90%, 80%, 70%, 60%, and 50% confidence intervals v.s. the target coverage for Target 2. **(b)** Empirical coverages of 95% confidence interval over 10000 steps under three noisy context environments. **(c)** Variance of the proposed IPW-Z estimator over 10000 steps under Boltzmann exploration with different temperature γ. **(d)** Empirical coverages of 95% confidence interval over 10000 steps under greedy algorithm w.r.t. the IPW-Z estimator with different minimum sampling probabilities p_0.

A Additional Technical Details

A.1 The Clipping Operator

In this section, we present in details a useful transformation,

$$\text{Clip} : \mathbb{R}^K \to \mathbb{R}^K,$$

which maps a probability distribution over K discrete actions to a new distribution to ensure that each coordinate is lower bounded by $\pi_{\min}$. Specifically, for $\boldsymbol{\pi} \in [0,1]^K$, define

$$\text{Clip}(\boldsymbol{\pi}) = \max\{\boldsymbol{\pi} - \nu^*(\boldsymbol{\pi}), \pi_{\min}\}, \tag{A.1}$$

where $\nu^*(\boldsymbol{\pi})$ is defined as the unique value such that

$$q(\nu; \boldsymbol{\pi}) := \sum_{a \in \mathcal{A}} \max\{\boldsymbol{\pi}_a - \nu, \pi_{\min}\} = 1.$$

This transformation adjusts $\boldsymbol{\pi}$ to remain a valid probability distribution while ensuring that each component is at least $\pi_{\min}$. The following lemma shows that this operation is in fact the L_2 projection onto the constrained simplex, providing a principled justification for its use. The proof is in Appendix C.1.

Lemma A.1. *The mapping* $\text{Clip}(\boldsymbol{\pi})$ *defined in (A.1) is the* L_2 *projection of* $\boldsymbol{\pi}$ *onto the set*

$$\left\{ \boldsymbol{\pi} \in [0,1]^{|\mathcal{A}|} \middle| \sum_{a \in \mathcal{A}} \boldsymbol{\pi}_a = 1, \boldsymbol{\pi}_a \geq \pi_{\min} \right\}.$$

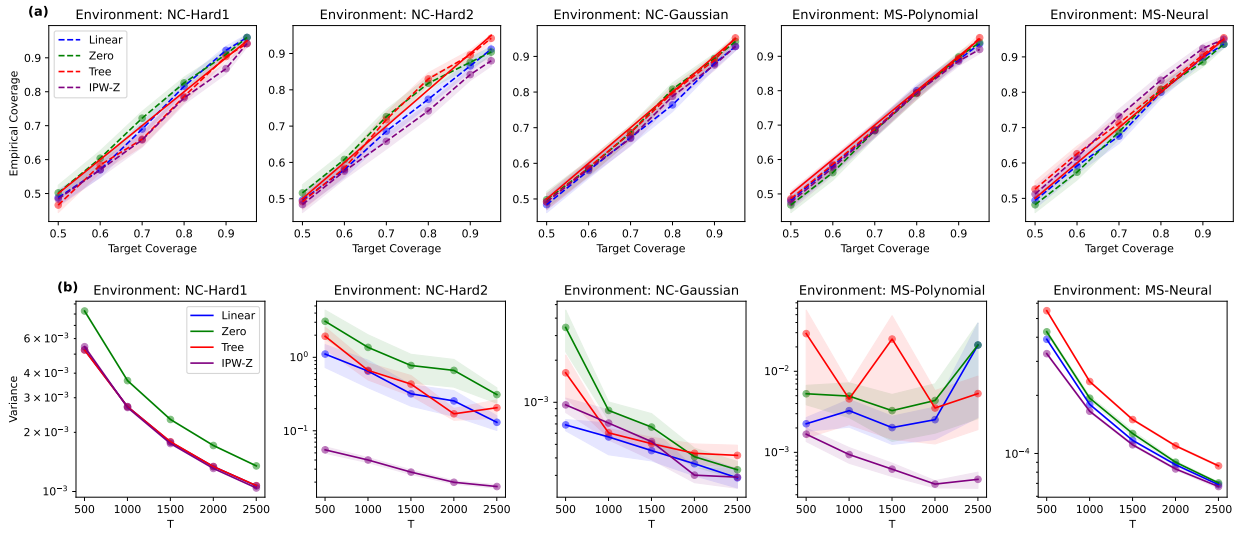

Figure 4: **(a)** Empirical coverage for OPE target from CADR with different prediction models and our proposed inference method across five environments. **(b)** Monte Carlo estimates of the variance of OPE target from CADR and our proposed inference method across five environments.

A.2 Convergence of the Policy in [15]

In this section, we analyze a behavior policy studied in [15], which combines an ϵ-greedy algorithm with a weighted online LS estimator, in the setting of misspecified linear bandits. The policy can be written as

$$\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}) = (1-\epsilon_t)1_{\{(\widehat{\boldsymbol{\beta}}_a^{(t-1)} - \widehat{\boldsymbol{\beta}}_{1-a}^{(t-1)})^\top \boldsymbol{X}_t > 0\}} + \frac{\epsilon_t}{2} \tag{A.2}$$

for $a \in \mathcal{A} = \{0, 1\}$. Here, $\widehat{\boldsymbol{\beta}}_a^{(t-1)}$ denotes the weighted online LS estimate of the reward parameter for arm a using data up to time $t-1$, which is a special case of the IPW-Z estimator $\widehat{\boldsymbol{\theta}}_a^{(t-1)}$ defined by (2.6) with the score function $\boldsymbol{g}$ given by (2.2). $\epsilon_t \in (0, 1)$ is a time-varying exploration parameter such that $\lim_{t\to\infty} \epsilon_t = \epsilon_\infty > 0$.

It is straightforward to verify the above policy is of the form (4.8) with $\widehat{\boldsymbol{\theta}}_a^{(t-1)} = \widehat{\boldsymbol{\beta}}_a^{(t-1)}$, $\boldsymbol{\gamma}_{t-1} = \epsilon_t$, and

$$\pi(a|\boldsymbol{x}, \{\boldsymbol{\theta}_a\}_{a\in\mathcal{A}}, \boldsymbol{\gamma}) = (1-\boldsymbol{\gamma})1_{\{(\boldsymbol{\theta}_a - \boldsymbol{\theta}_{1-a})^\top \boldsymbol{x} > 0\}} + \frac{\boldsymbol{\gamma}}{2}.$$

The convergence of $\{\boldsymbol{\gamma}_t\}_{t\geq 1}$ is guaranteed by the convergence of $\{\epsilon_t\}_{t\geq 1}$. In addition, the function π is continuous in $(\{\boldsymbol{\theta}_a\}_{a\in\mathcal{A}}, \boldsymbol{\gamma})$ at the point $(\{\boldsymbol{\theta}_a^*\}_{a\in\mathcal{A}}, \epsilon_\infty)$ for any $\boldsymbol{x} \in \mathcal{X}_0 := \{\boldsymbol{x} : (\boldsymbol{\theta}_1^* - \boldsymbol{\theta}_0^*)^\top \boldsymbol{x} \neq 0\}$. Here for $a \in \mathcal{A}$, $\boldsymbol{\theta}_a^*$ is defined in (2.1) with the score function $\boldsymbol{g}$ given by (2.2). Note that $\mathcal{X}_0^c$ is a Lebesgue null set, and from Assumption 1 of [15], we deduce that

$$\mathbb{P}(\mathcal{X}_0^c) = 0.$$

Combining the above arguments, we have verified the conditions of Proposition 4.2, which implies that the policy (A.2) satisfies the policy convergence condition in Definition 3.1.

A.3 Asymptotic Stability

Definition A.2 (Asymptotic Stability)**.** The equilibrium point $x = 0$ of $\dot{x} = f(x)$ is

- stable if, for each $\varepsilon > 0$, there is $\delta = \delta(\varepsilon) > 0$ such that

$$\|x(0)\| < \delta \Rightarrow \|x(t)\| < \varepsilon, \quad \forall t \geq 0$$

- unstable if it is not stable.
- asymptotically stable if it is stable and δ can be chosen such that

$$\|x(0)\| < \delta \Rightarrow \lim_{t \to \infty} x(t) = 0$$

A.4 Inference for Bandits with Noisy Contexts under Estimated Contextual Error Variance

In this section, we discuss the statistical inference task in Example 2 where the contextual error variance $\boldsymbol{\Sigma}_e$ is unknown. In addition to the adaptively collected dataset $\mathcal{D}$, suppose we have an auxiliary offline dataset

$$\widetilde{\mathcal{D}} = \{\widetilde{\boldsymbol{X}}_i, \widetilde{\boldsymbol{S}}_i\}_{i=1}^n$$

containing paired noisy and true contexts. Each $(\widetilde{\boldsymbol{X}}_i, \widetilde{\boldsymbol{S}}_i)$ is independent and equal in distribution to $(\boldsymbol{X}_t, \boldsymbol{S}_t)$ in $\mathcal{D}$. Such auxiliary data arise naturally when gold-standard measurements are available for a subset of observations, or when the observed context is generated by a complex prediction algorithm, and its error variance can be estimated using gold-standard labels. In healthcare applications, $\widetilde{\mathcal{D}}$ is often obtained from prior studies, such as pilot trials. Because $\widetilde{\mathcal{D}}$ is collected offline without intervention or adaptive sampling, it is typically accessible in practice.

Under this setting, notice that

$$\boldsymbol{\Sigma}_e = \mathrm{Var}(\boldsymbol{X}_t | \boldsymbol{S}_t) = \mathbb{E}\big[(\boldsymbol{X}_t - \boldsymbol{S}_t)(\boldsymbol{X}_t - \boldsymbol{S}_t)^\top \big| \boldsymbol{S}_t\big] = \mathbb{E}\big[(\boldsymbol{X}_t - \boldsymbol{S}_t)(\boldsymbol{X}_t - \boldsymbol{S}_t)^\top\big],$$

we can estimate $\boldsymbol{\Sigma}_e$ using

$$\widehat{\boldsymbol{\Sigma}}_e = \frac{1}{n} \sum_{i=1}^n \widetilde{\boldsymbol{V}}_i,$$

where $\widetilde{\boldsymbol{V}}_i := (\widetilde{\boldsymbol{X}}_i - \widetilde{\boldsymbol{S}}_i)(\widetilde{\boldsymbol{X}}_i - \widetilde{\boldsymbol{S}}_i)^\top$. A natural estimator for $\boldsymbol{\theta}_a^*$ would be $\widetilde{\boldsymbol{\theta}}_a^{(T)}$ which solves $\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}) := \frac{1}{T} \sum_{t \in [T]} W_t 1_{\{A_t = a\}} \widetilde{\boldsymbol{g}}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}),$$

where $W_t = \frac{1}{\pi_t(A_t)}$, $\widetilde{\boldsymbol{g}}(\boldsymbol{x}, y; \boldsymbol{\theta}) = (\boldsymbol{x}\boldsymbol{x}^\top - \widehat{\boldsymbol{\Sigma}}_e)\boldsymbol{\theta} - \boldsymbol{x}y$.

The following theorem characterize the asymptotic distribution of $\widetilde{\boldsymbol{\theta}}_a^{(T)}$, which can be used to conduct inference on $\boldsymbol{\theta}_a^*$. The proof is in Appendix B.10.

Theorem A.3. *Under Assumptions 2.1, 3.1, 3.4 and 5.2, define $\boldsymbol{\Sigma}_S$ and $\bar{\boldsymbol{I}}_a$ the same as in Corollary 5.2. Then as $n, T \to \infty$,*

- *If $n/T \to \infty$, then $\sqrt{T}(\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_S^{-1} \bar{\boldsymbol{I}}_a \boldsymbol{\Sigma}_S^{-1}\right)$.*

- *If* $n/T \to \kappa$ *for some positive constant* κ*, then* $\sqrt{T}(\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\Sigma}_S^{-1}\big[\bar{\boldsymbol{I}}_a + \frac{\bar{\boldsymbol{H}}_a}{\kappa}\big]\boldsymbol{\Sigma}_S^{-1}\right)$*, where* $\bar{\boldsymbol{H}}_a := \mathbb{E}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*\boldsymbol{\theta}_a^{*,\top}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)$.
- *If* $n/T \to 0$*, then* $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_S^{-1}\bar{\boldsymbol{H}}_a\boldsymbol{\Sigma}_S^{-1}\right)$*, where* $\bar{\boldsymbol{H}}_a := \mathbb{E}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*\boldsymbol{\theta}_a^{*,\top}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)$.

By Theorem A.3, if the sample size of the auxiliary data $\widetilde{\mathcal{D}}$ is sufficiently larger than that of the adaptively collected data $\mathcal{D}$, the distribution of $\widetilde{\boldsymbol{\theta}}_a^{(T)}$ coincides with that of $\widehat{\boldsymbol{\theta}}_a^{(T)}$. This scenario is common in practice since, unlike $\mathcal{D}$, the auxiliary dataset $\widetilde{\mathcal{D}}$ involves no intervention and is typically easier to obtain. In case $\widetilde{\mathcal{D}}$ has comparable or smaller sample size compared to $\mathcal{D}$, valid inference remains possible, but the asymptotic variance of $\widetilde{\boldsymbol{\theta}}_a^{(T)}$ exceeds that of $\widehat{\boldsymbol{\theta}}_a^{(T)}$ due to the additional uncertainty from estimating $\boldsymbol{\Sigma}_e$. The joint asymptotic distribution of $\{\widetilde{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}}$ can be derived similar to Theorem 3.3, and is omitted for brevity.

B Proofs of Main Theorems

We organize the proofs of the main theorems by sections in the main text.

B.1 Proof of Theorem 3.2

We first prove the following lemma. Its proof is in Appendix C.2.

Lemma B.1. *Under the assumptions of Theorem 3.2, there exists a sequence of estimators* $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{T\geq 1}$ *such that (2.6) holds, and* $\|\widehat{\boldsymbol{\theta}}_a^{(T)}\|_2 \leq R_\theta$*,* $\forall T$*. In addition, for any such sequence, as* $T \to \infty$*,* $\widehat{\boldsymbol{\theta}}_a^{(T)} \xrightarrow{p} \boldsymbol{\theta}_a^*$.

For the remainder of this proof, for notational convenience, we omit the dependence of $\widehat{\boldsymbol{\theta}}_a^{(T)}$ on T and write it as $\widehat{\boldsymbol{\theta}}_a$. Let $\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})$ denote the i-th entry of $\boldsymbol{G}_T(\boldsymbol{\theta})$. By Taylor expansion, we have that for any $i \in \{1, \ldots, d\}$, there exists some $\widetilde{\boldsymbol{\theta}}_{a,i}$ on the line segment between $\boldsymbol{\theta}_a^*$ and $\widehat{\boldsymbol{\theta}}_a$ such that

$$\begin{aligned}-\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}_a^*) &= \boldsymbol{G}_T^{(i)}(\widehat{\boldsymbol{\theta}}_a) - \boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}_a^*) + o_p(1/\sqrt{T}) \\ &= \langle \nabla \boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}_a^*), \widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^* \rangle + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*)^\top \nabla^2 \boldsymbol{G}_T^{(i)}(\widetilde{\boldsymbol{\theta}}_{a,i})(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) + o_p(1/\sqrt{T}).\end{aligned}$$

Stacking the above expansions over the entries $i = 1, \ldots, d$, we have

$$-\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*)(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) + \frac{1}{2}\widetilde{\boldsymbol{\delta}}_a(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) + o_p(1/\sqrt{T}),$$

where

$$\widetilde{\boldsymbol{\delta}}_a = \begin{pmatrix} (\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*)^\top \nabla^2 \boldsymbol{G}_T^{(1)}(\widetilde{\boldsymbol{\theta}}_{a,1}) \\ \vdots \\ (\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*)^\top \nabla^2 \boldsymbol{G}_T^{(d)}(\widetilde{\boldsymbol{\theta}}_{a,d}) \end{pmatrix}.$$

By rearranging, we obtain

$$\big[\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + \frac{1}{2}\widetilde{\boldsymbol{\delta}}_a\big] \cdot \sqrt{T}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) = -\sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + o_p(1). \tag{B.1}$$

Below we state the following lemmas, the proof of these lemmas are in Appendix C.3, C.4, and C.5, respectively.

Lemma B.2. *Under the assumptions of Theorem 3.2, as* $T \to \infty$, $\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{p} \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$.

Lemma B.3. *Under the assumptions of Theorem 3.2,* $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0} \|\nabla^2 \boldsymbol{G}_T(\boldsymbol{\theta})\|_1 = \mathcal{O}_p(1)$. *Here, for a tensor* $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, *we define* $\|\boldsymbol{B}\|_1 = \sum_{i\in[d_1], j\in[d_2], k\in[d_3]} |\boldsymbol{B}_{i,j,k}|$.

Lemma B.4. *Under the assumptions of Theorem 3.2, as* $T \to \infty$, $\sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \bar{\boldsymbol{I}}_a)$, *where* $\bar{\boldsymbol{I}}_a = \mathbb{E}\big[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top\big]$.

We now derive the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_a$ from (B.1). First, since $\widetilde{\boldsymbol{\theta}}_{a,i}$ is on the line segment between $\boldsymbol{\theta}_a^*$ and $\widehat{\boldsymbol{\theta}}_a$, from Lemma B.1 and Lemma B.3, we have $\forall i$,

$$\begin{aligned}
\|\nabla^2 \boldsymbol{G}_T^{(i)}(\widetilde{\boldsymbol{\theta}}_{a,i})\|_{1,1} &\le \|\nabla^2 \boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_{a,i})\|_1 \\
&= \|\nabla^2 \boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_{a,i})\|_1 \cdot 1_{\{\|\widetilde{\boldsymbol{\theta}}_{a,i}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0\}} + \|\nabla^2 \boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_{a,i})\|_1 \cdot 1_{\{\|\widetilde{\boldsymbol{\theta}}_{a,i}-\boldsymbol{\theta}_a^*\|_2 > \epsilon_0\}} \\
&\le \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0} \|\nabla^2 \boldsymbol{G}_T(\boldsymbol{\theta})\|_1 + \|\nabla^2 \boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_{a,i})\|_1 \cdot 1_{\{\|\widehat{\boldsymbol{\theta}}_a-\boldsymbol{\theta}_a^*\|_2 > \epsilon_0\}} = \mathcal{O}_p(1).
\end{aligned}$$

Here for a matrix $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$, we define $\|\boldsymbol{B}\|_{1,1} = \sum_{i\in[d_1], j\in[d_2]} |\boldsymbol{B}_{i,j}|$.

Combine the above with Lemma B.1 which implies $\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^* = o_p(1)$ and Lemma B.2 which ensures convergence of $\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*)$, we deduce that

$$\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + \frac{1}{2}\widetilde{\boldsymbol{\delta}}_a \xrightarrow{p} \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*). \tag{B.2}$$

We further combine the above expression with Lemma B.4 and use Slutsky's theorem to obtain (3.2).

B.2 Proof of Theorem 3.3

We prove a more general version of Theorem 3.3 where the score function $\boldsymbol{g}$ in (2.1) for each arm can be different. Suppose that for any $a \in \mathcal{A}$, $\boldsymbol{\theta}_a^*$ satisfies

$$\mathbb{E}\boldsymbol{g}_a(\boldsymbol{X}, Y(a); \boldsymbol{\theta}_a^*) = \boldsymbol{0} \tag{B.3}$$

for a score function $\boldsymbol{g}_a$.

The following conditions extend Assumptions 3.1, 3.2, 3.3, and 3.4 to the more general setting where the conditions apply not only to a single arm but simultaneously across all arms $a \in \mathcal{A}$.

Assumption B.1 (Well-separated solution for all arms)**.** $\forall a \in \mathcal{A}, \forall \epsilon > 0$,
$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\mathbb{E}\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2 > 0$.

Assumption B.2 (Boundedness for all arms)**.** There exist constants R_θ, M_2 such that
(i) $\forall a \in \mathcal{A}$, $\|\mathbb{E}[\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top | \boldsymbol{X}_t]\|_2 \le M_2$, a.e. $\boldsymbol{X}_t$;
(ii) $\forall a \in \mathcal{A}$, $\|\boldsymbol{\theta}_a^*\|_2 < R_\theta$, $\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \mathbb{E}\|\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2^2 < \infty$;
(iii) $\forall a \in \mathcal{A}$, $\mathbb{E}\|\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\|_2^4 < \infty$.

Assumption B.3 (Smoothness for all arms)**.** (i)$\forall a \in \mathcal{A}$, the function $\boldsymbol{g}_a(\boldsymbol{x}, y; \boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$, with $\mathbb{E}\nabla \boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$ nonsingular;
(ii) There exists a function ϕ such that $\forall a \in \mathcal{A}$, $\forall \boldsymbol{x}, y$, $\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \|\nabla \boldsymbol{g}_a(\boldsymbol{x}, y; \boldsymbol{\theta})\|_2 \le \phi(\boldsymbol{x}, y)$, and

$\mathbb{E}\phi(\boldsymbol{X}_t, Y_t(a))^2 < \infty$;
(iii)There exists a constant $\epsilon_0 > 0$ and a function Φ such that $\forall a \in \mathcal{A}$, $\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0, i\in[d]} \|\nabla^2 \boldsymbol{g}_a^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})\|_2 \le \Phi(\boldsymbol{x}, y)$ and $\mathbb{E}\Phi(\boldsymbol{X}_t, Y_t(a)) < \infty$. Here $\boldsymbol{g}_a^{(i)}(\boldsymbol{x}, y; \boldsymbol{\theta})$ denotes the i-th entry of $\boldsymbol{g}_a(\boldsymbol{x}, y; \boldsymbol{\theta})$.

Assumption B.4 (Minimum sampling probability for all arms)**.** $\forall a \in \mathcal{A}$, $\pi_t(a) \ge \pi_{\min}$ almost surely for some constant $\pi_{\min} \in (0, 1)$.

Define $\boldsymbol{G}_{a,T} := \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \boldsymbol{g}_a(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta})$, and similar to (2.6), we look for $\widehat{\boldsymbol{\theta}}_a^{(T)}$ such that $\forall a \in \mathcal{A}$,

$$\boldsymbol{G}_{a,T}(\widehat{\boldsymbol{\theta}}_a) = o_p(1/\sqrt{T}) \tag{B.4}$$

as $T \to \infty$. We prove the following theorem, which is a generalization of Theorem 3.3.

Theorem B.5. *Under Assumptions 2.1, B.1, B.2, B.3, and B.4, if the behavior policy π satisfies policy convergence to a policy $\bar{\pi}$ in the sense of Definition 3.1, then there exist estimators $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}, T\ge 1}$ such that (B.4) holds for any $a \in \mathcal{A}$, and $\|\widehat{\boldsymbol{\theta}}_a^{(T)}\|_2 \le R_\theta$, $\forall a \in \mathcal{A}, T \ge 1$. In addition, any such estimators satisfy*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}^{(T)} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^*) \tag{B.5}$$

as $T \to \infty$. Here $\widehat{\boldsymbol{\theta}}^{(T)} = \big((\widehat{\boldsymbol{\theta}}_1^{(T)})^\top, \ldots, (\widehat{\boldsymbol{\theta}}_K^{(T)})^\top\big)^\top$, $\boldsymbol{\theta}^* = \big(\boldsymbol{\theta}_1^{*,\top}, \ldots, \boldsymbol{\theta}_K^{*,\top}\big)^\top$, and $\boldsymbol{\Sigma}^* = \mathrm{diag}(\boldsymbol{\Sigma}_1^*, \ldots, \boldsymbol{\Sigma}_K^*)$ where $\boldsymbol{\Sigma}_a^* := \boldsymbol{J}_a^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{J}_a^{-1,\top}$, $\bar{\boldsymbol{I}}_a := \mathbb{E}\big[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top\big]$, $\boldsymbol{J}_a := \mathbb{E}\nabla\boldsymbol{g}_a(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$.

Proof of Theorem B.5. First, the existence of $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}, T\ge 1}$ is guaranteed by applying Theorem 3.2 to each individual arm $a \in \mathcal{A}$. Now suppose $\{\widehat{\boldsymbol{\theta}}_a^{(T)}\}_{a\in\mathcal{A}, T\ge 1}$ satisfies (B.4) $\forall a \in \mathcal{A}$, and $\|\widehat{\boldsymbol{\theta}}_a^{(T)}\|_2 \le R_\theta$ $\forall a \in \mathcal{A}, T \ge 1$. From the Cramer-Wold theorem, in order to prove (B.5), we only need to show that for any nonrandom vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K \in \mathbb{R}^d$,

$$\sqrt{T}\sum_{a\in\mathcal{A}} \boldsymbol{\beta}_a^\top(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}\left(0, \sum_{a\in\mathcal{A}} \boldsymbol{\beta}_a^\top \boldsymbol{J}_a^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\right). \tag{B.6}$$

First, for any fixed arm a, we let $\boldsymbol{g} = \boldsymbol{g}_a$ and $\boldsymbol{G}_T = \boldsymbol{G}_{a,T}$ in (B.1), and considering Lemma B.4 as well as (B.2), we obtain that

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) = \big[\boldsymbol{J}_a + \boldsymbol{\delta}_{a,T}\big]^{-1}\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + o_p(1)$$

where $\boldsymbol{\delta}_{a,T} = o_p(1)$. By further analysis, we have

$$\begin{aligned}
\sqrt{T}(\widehat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) &= \boldsymbol{J}_a^{-1}\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + \left(\big[\boldsymbol{J}_a + \boldsymbol{\delta}_{a,T}\big]^{-1} - \boldsymbol{J}_a^{-1}\right)\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + o_p(1)\\
&= \boldsymbol{J}_a^{-1}\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + \left([\boldsymbol{I} + \boldsymbol{J}_a^{-1}\boldsymbol{\delta}_{a,T}]^{-1} - \boldsymbol{I}\right)\boldsymbol{J}_a^{-1}\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + o_p(1)\\
&= \boldsymbol{J}_a^{-1}\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) + o_p(1).
\end{aligned} \tag{B.7}$$

Here in the last equation, we have used the fact that $\boldsymbol{J}_a$ is invertible, $\boldsymbol{\delta}_{a,T} = o_p(1)$, and $\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*) =$

$\mathcal{O}_p(1)$ from Lemma B.4. Applying (B.7) to all $a \in \mathcal{A}$, we deduce that

$$\begin{aligned}\sqrt{T}\sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top(\widehat{\boldsymbol{\theta}}_a-\boldsymbol{\theta}_a^*) &= \sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\cdot\sqrt{T}\boldsymbol{G}_{a,T}(\boldsymbol{\theta}_a^*)+o_p(1)\\ &= \sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\cdot\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)+o_p(1)\\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\bar{Z}_t+o_p(1),\end{aligned} \tag{B.8}$$

where $\bar{Z}_t := \sum_{a\in\mathcal{A}}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)$. Given the above expression, from Theorem 2.2 in [23], (B.6) can be obtained by ensuring

$$\mathbb{E}[\bar{Z}_t|\mathcal{H}_{t-1}]=0 \quad \forall t\in[T], \tag{B.9}$$

$$\frac{1}{T}\sum_{t\in[T]}\mathrm{Var}(\bar{Z}_t|\mathcal{H}_{t-1}) \xrightarrow{p} \sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a, \tag{B.10}$$

$$\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\left[\bar{Z}_t^2 1_{\{|\bar{Z}_t|>\sqrt{T}\delta\}}\middle|\mathcal{H}_{t-1}\right] \xrightarrow{p} 0 \quad \forall\delta>0. \tag{B.11}$$

Below we check these facts one by one.

Check (B.9): We have

$$\begin{aligned}\mathbb{E}[\bar{Z}_t|\mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}[\bar{Z}_t|\mathcal{H}_{t-1},\boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right]\\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}\left[\sum_{a\in\mathcal{A}}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\middle|\mathcal{H}_{t-1},\boldsymbol{X}_t\right]\middle|\mathcal{H}_{t-1}\right]\\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[\sum_{a\in\mathcal{A}}\mathbb{E}_{A_t,Y_t(a)}\left[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\middle|\mathcal{H}_{t-1},\boldsymbol{X}_t\right]\middle|\mathcal{H}_{t-1}\right]\\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[\sum_{a\in\mathcal{A}}\mathbb{E}_{A_t}\left[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\middle|\mathcal{H}_{t-1},\boldsymbol{X}_t\right]\cdot\right.\\ &\qquad\left.\mathbb{E}_{Y_t(a)}\left[\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\middle|\mathcal{H}_{t-1},\boldsymbol{X}_t\right]\middle|\mathcal{H}_{t-1}\right]\\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[\sum_{a\in\mathcal{A}}\mathbb{E}_{Y_t(a)}\left[\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\middle|\mathcal{H}_{t-1},\boldsymbol{X}_t\right]\middle|\mathcal{H}_{t-1}\right]\\ &= \sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\mathbb{E}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)=0.\end{aligned}$$

Here in the fourth equality we use Assumption 2.1, and the last equality is due to (B.3).

Check (B.10): Due to (B.9), we have $\frac{1}{T}\sum_{t\in[T]}\mathrm{Var}(\bar{Z}_t|\mathcal{H}_{t-1}) = \frac{1}{T}\sum_{t\in[T]}\mathbb{E}[\bar{Z}_t^2|\mathcal{H}_{t-1}]$, where

$$
\begin{aligned}
\mathbb{E}[\bar{Z}_t^2|\mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{X}_t}[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}[\bar{Z}_t^2|\mathcal{H}_{t-1},\boldsymbol{X}_t]|\mathcal{H}_{t-1}] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}\Bigg[\Big(\sum_{a\in\mathcal{A}}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\cdot \\
&\qquad\qquad \boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\Big)^2\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}\Bigg[\sum_{a,a'\in\mathcal{A}}\frac{1_{\{A_t=a\}}\cdot 1_{\{A_t=a'\}}}{\pi_t(A_t)^2}\cdot \\
&\qquad \boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_{a'}(\boldsymbol{X}_t,Y_t(a');\boldsymbol{\theta}_{a'}^*)^\top\boldsymbol{J}_{a'}^{-1,\top}\boldsymbol{\beta}_{a'}\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\mathbb{E}_{A_t,\{Y_t(a)\}_{a\in\mathcal{A}}}\Bigg[\sum_{a\in\mathcal{A}}\frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2}\cdot \\
&\qquad \boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)^\top\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\sum_{a\in\mathcal{A}}\mathbb{E}_{A_t,Y_t(a)}\Bigg[\frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2}\cdot \\
&\qquad \boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)^\top\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\sum_{a\in\mathcal{A}}\mathbb{E}_{A_t}\Bigg[\frac{1_{\{A_t=a\}}}{\pi_t(A_t)^2}\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\cdot \\
&\mathbb{E}_{Y_t(a)}\Bigg[\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)^\top\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Bigg]\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\sum_{a\in\mathcal{A}}\frac{1}{\pi_t(a)}\cdot\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\cdot \\
&\qquad \mathbb{E}_{Y_t(a)}\Big[\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)^\top\Big|\mathcal{H}_{t-1},\boldsymbol{X}_t\Big]\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\Bigg[\sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\boldsymbol{I}_{a,t}\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \sum_{a\in\mathcal{A}}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\mathbb{E}_{\boldsymbol{X}_t}\big[\boldsymbol{I}_{a,t}\big|\mathcal{H}_{t-1}\big]\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a. \qquad \text{(B.12)}
\end{aligned}
$$

Here $\boldsymbol{I}_{a,t} := \frac{1}{\pi_t(a)}\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)^\top|\boldsymbol{X}_t]$. This is consistent with the definition in Appendix C.5, where $\boldsymbol{I}_{a,t}$ is defined for a single fixed arm a, by taking $\boldsymbol{g}=\boldsymbol{g}_a$ in the more general setting considered here. In the above, the sixth equality uses Assumption 2.1.

From (C.28) and (C.29) with $\boldsymbol{c}=\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a$, we obtain that $\forall a\in\mathcal{A}$,

$$
\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\mathbb{E}_{\boldsymbol{X}_t}\big[\boldsymbol{I}_{a,t}\big|\mathcal{H}_{t-1}\big]\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a \xrightarrow{p} \boldsymbol{\beta}_a^\top\boldsymbol{J}_a^{-1}\bar{\boldsymbol{I}}_a\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a, \qquad \text{(B.13)}
$$

where $\bar{\boldsymbol{I}}_a$ is defined in Theorem B.5. This is consistent with the definition in Appendix C.5, where $\bar{\boldsymbol{I}}_a$ is defined for a single fixed arm a, by taking $\boldsymbol{g}=\boldsymbol{g}_a$ in the more general setting considered here.

Finally, combining (B.12) and (B.13), we obtain (B.10).

Check (B.11): We have

$$\begin{aligned}
\frac{1}{T}\sum_{t\in[T]}\mathbb{E}\Big[\bar{Z}_t^2 1_{\{|\bar{Z}_t|>\sqrt{T}\delta\}}\Big|\mathcal{H}_{t-1}\Big] &\le \frac{1}{T}\sum_{t=1}^{T}\frac{1}{T\delta^2}\mathbb{E}[\bar{Z}_t^4|\mathcal{H}_{t-1}] \\
&= \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}\Big[\Big(\sum_{a\in\mathcal{A}}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{\beta}_a^\top \boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)\Big)^4\Big|\mathcal{H}_{t-1}\Big] \\
&= \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\mathbb{E}\Big[\sum_{a\in\mathcal{A}}\frac{1}{\pi_t(A_t)^4}1_{\{A_t=a\}}\Big(\boldsymbol{\beta}_a^\top \boldsymbol{J}_a^{-1}\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)\Big)^4\Big|\mathcal{H}_{t-1}\Big] \\
&\le \frac{1}{T^2\delta^2}\sum_{t=1}^{T}\frac{1}{\pi_{\min}^4}\cdot\sum_{a\in\mathcal{A}}\|\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\|_2^4\cdot\mathbb{E}\|\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)\|_2^4 \\
&\le \frac{1}{T\delta^2\pi_{\min}^4}\max_a\|\boldsymbol{J}_a^{-1,\top}\boldsymbol{\beta}_a\|_2^4\cdot\sum_{a\in\mathcal{A}}\mathbb{E}\|\boldsymbol{g}_a(\boldsymbol{X}_t,Y_t;\boldsymbol{\theta}_a^*)\|_2^4\to 0.
\end{aligned}$$

Here the first inequality uses Chebyshev's Inequality. The second equality holds because all cross-product terms vanish when expanding the fourth power of the sum over $a\in\mathcal{A}$. The last convergence uses Assumption B.2, and that $\boldsymbol{J}_a$ is invertible for all $a\in\mathcal{A}$. □

B.3 Proof of Proposition 3.1

We first prove that as $T\to\infty$,

$$\hat{\dot{\boldsymbol{G}}}_{a,T}\xrightarrow{p}\mathbb{E}\nabla\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*). \tag{B.14}$$

From Lemma B.2, we deduce that

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\nabla\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\xrightarrow{p}\mathbb{E}\nabla\boldsymbol{g}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*).$$

In addition, $\forall i$,

$$\begin{aligned}
&\Big\|\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\nabla\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\widehat{\boldsymbol{\theta}}_a^{(T)})-\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\nabla\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\Big\|_2 \\
=&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\|\nabla\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\widehat{\boldsymbol{\theta}}_a^{(T)})-\nabla\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*)\|_2 \\
\le&\frac{1}{\pi_{\min}T}\sum_{t=1}^{T}\Big\|\int_0^1\nabla^2\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*+u(\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*))\mathrm{d}u\cdot(\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*)\Big\|_2 \\
\le&\frac{1}{\pi_{\min}T}\Big[\sum_{t=1}^{T}\Phi(\boldsymbol{X}_t,Y_t(a))\Big]\cdot\|\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*\|_2+ \\
&1_{\{\|\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*\|_2>\epsilon_0\}}\cdot\frac{1}{\pi_{\min}T}\sum_{t=1}^{T}\Big\|\int_0^1\nabla^2\boldsymbol{g}^{(i)}(\boldsymbol{X}_t,Y_t(a);\boldsymbol{\theta}_a^*+u(\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*))\mathrm{d}u\cdot(\widehat{\boldsymbol{\theta}}_a^{(T)}-\boldsymbol{\theta}_a^*)\Big\|_2 \\
=&o_p(1).
\end{aligned}$$

Here we have used Assumption 3.4 and 3.3. Thus,

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \widehat{\boldsymbol{\theta}}_a^{(T)}) - \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) = o_p(1).$$

Combining the above, we obtain (B.14).

Next we prove that as $T \to \infty$,

$$\widehat{\boldsymbol{I}}_{a,T} \xrightarrow{p} \bar{\boldsymbol{I}}_a. \tag{B.15}$$

Denote $\boldsymbol{Z}_t = \frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta}_a^*)$, and $\widehat{\boldsymbol{Z}}_t^{(T)} = \frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{g}(\boldsymbol{X}_t, Y_t; \widehat{\boldsymbol{\theta}}_a^{(T)})$. Then (B.15) is equivalent to

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{Z}}_t^{(T)}\widehat{\boldsymbol{Z}}_t^{(T),\top} \xrightarrow{p} \bar{\boldsymbol{I}}_a,$$

and in order to show this, it suffices to prove

$$\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{Z}}_t^{(T)}\widehat{\boldsymbol{Z}}_t^{(T),\top} - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{Z}_t\boldsymbol{Z}_t^{\top} = o_p(1), \tag{B.16}$$

$$\frac{1}{T}\sum_{t=1}^{T}\Big(\boldsymbol{Z}_t\boldsymbol{Z}_t^{\top} - \mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^{\top}|\mathcal{H}_{t-1}]\Big) = o_p(1), \tag{B.17}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^{\top}|\mathcal{H}_{t-1}] \xrightarrow{p} \bar{\boldsymbol{I}}_a. \tag{B.18}$$

Below we check these facts one by one.

Check (B.16): For convenience, we define $\boldsymbol{g}_t^* = \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$, and $\widehat{\boldsymbol{g}}_t^{(T)} = \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \widehat{\boldsymbol{\theta}}_a^{(T)})$.

Then

$$
\begin{aligned}
&\left\| \frac{1}{T} \sum_{t=1}^{T} \left(\widehat{\boldsymbol{Z}}_t^{(T)} \widehat{\boldsymbol{Z}}_t^{(T),\top} - \boldsymbol{Z}_t \boldsymbol{Z}_t^{\top} \right) \right\|_2 \\
=& \left\| \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\pi_t(A_t)^2} 1_{\{A_t=a\}} \left(\widehat{\boldsymbol{g}}_t^{(T)} \widehat{\boldsymbol{g}}_t^{(T),\top} - \boldsymbol{g}_t^* \boldsymbol{g}_t^{*,\top} \right) \right\|_2 \\
\leq& \frac{1}{T} \cdot \frac{1}{\pi_{\min}^2} \sum_{t=1}^{T} \| \widehat{\boldsymbol{g}}_t^{(T)} \widehat{\boldsymbol{g}}_t^{(T),\top} - \boldsymbol{g}_t^* \boldsymbol{g}_t^{*,\top} \|_2 \\
\leq& \frac{1}{\pi_{\min}^2 T} \sum_{t=1}^{T} \left(\| \widehat{\boldsymbol{g}}_t^{(T)} \widehat{\boldsymbol{g}}_t^{(T),\top} - \widehat{\boldsymbol{g}}_t^{(T)} \boldsymbol{g}_t^{*,\top} \|_2 + \| \widehat{\boldsymbol{g}}_t^{(T)} \boldsymbol{g}_t^{*,\top} - \boldsymbol{g}_t^* \boldsymbol{g}_t^{*,\top} \|_2 \right) \\
\leq& \frac{1}{\pi_{\min}^2 T} \sum_{t=1}^{T} \| \widehat{\boldsymbol{g}}_t^{(T)} - \boldsymbol{g}_t^* \|_2 \cdot \left(\| \widehat{\boldsymbol{g}}_t^{(T)} \|_2 + \| \boldsymbol{g}_t^* \|_2 \right) \\
\leq& \frac{1}{\pi_{\min}^2 T} \sum_{t=1}^{T} \| \widehat{\boldsymbol{g}}_t^{(T)} - \boldsymbol{g}_t^* \|_2 \cdot \left(\| \widehat{\boldsymbol{g}}_t^{(T)} - \boldsymbol{g}_t^* \|_2 + 2 \| \boldsymbol{g}_t^* \|_2 \right) \\
\leq& \frac{1}{\pi_{\min}^2 T} \sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a)) \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2 \left[\phi(\boldsymbol{X}_t, Y_t(a)) \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2 + 2 \| \boldsymbol{g}_t^* \|_2 \right] \\
=& \frac{1}{\pi_{\min}^2 T} \left[\sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a))^2 \right] \cdot \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2^2 + \\
& \quad \frac{2}{\pi_{\min}^2 T} \left[\sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a)) \| \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \|_2 \right] \cdot \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2 \\
\leq& \frac{1}{\pi_{\min}^2 T} \left[\sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a))^2 \right] \cdot \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2^2 + \\
& \quad \frac{2}{\pi_{\min}^2} \sqrt{ \frac{1}{T} \sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a))^2 \cdot \frac{1}{T} \sum_{t=1}^{T} \| \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \|_2^2 } \cdot \| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2 \\
=& o_p(1).
\end{aligned}
\tag{B.19}
$$

Here, the fifth inequality is due to Assumption 3.3. The last inequality uses Cauchy–Schwarz inequality. The final equality is because of $\frac{1}{T} \sum_{t=1}^{T} \phi(\boldsymbol{X}_t, Y_t(a))^2 = \mathcal{O}_p(1)$ and $\frac{1}{T} \sum_{t=1}^{T} \| \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \|_2^2 = \mathcal{O}_p(1)$ (from the law of large numbers), and $\| \widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* \|_2 = o_p(1)$ (from Lemma B.1).

Check (B.17): $\forall \boldsymbol{c} \in \mathbb{R}^d$, $\forall \delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}[\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c} - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c}|\mathcal{H}_{t-1}]]\right| > \delta\right) \\
\leq & \frac{1}{\delta^2 T^2}\mathbb{E}\left(\sum_{t=1}^{T}[\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c} - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c}|\mathcal{H}_{t-1}]]\right)^2 \\
= & \frac{1}{\delta^2 T^2}\sum_{t=1}^{T}\mathbb{E}\left(\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c} - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c}|\mathcal{H}_{t-1}]\right)^2 \\
\leq & \frac{1}{\delta^2 T^2}\sum_{t=1}^{T}\mathbb{E}\left(\boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c}\right)^2 \\
\leq & \frac{1}{\delta^2 T^2}\sum_{t=1}^{T}\frac{1}{\pi_{\min}^2}\mathbb{E}\left(\boldsymbol{c}^\top \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\right)^4 \to 0.
\end{aligned}$$

Here the first inequality is because of Chebyshev's Inequality. The second equality is due to the following fact: Let $v_t' = \boldsymbol{c}^\top \boldsymbol{Z}_t \boldsymbol{Z}_t^\top \boldsymbol{c}$. Then for $t_1 < t_2$,

$$\begin{aligned}
&\mathbb{E}(v_{t_1}' - \mathbb{E}[v_{t_1}'|\mathcal{H}_{t_1-1}])(v_{t_2}' - \mathbb{E}[v_{t_2}'|\mathcal{H}_{t_2-1}]) \\
=&\mathbb{E}\big[\mathbb{E}[(v_{t_1}' - \mathbb{E}[v_{t_1}'|\mathcal{H}_{t_1-1}])(v_{t_2}' - \mathbb{E}[v_{t_2}'|\mathcal{H}_{t_2-1}])\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(v_{t_1}' - \mathbb{E}[v_{t_1}'|\mathcal{H}_{t_1-1}]) \cdot \mathbb{E}[v_{t_2}' - \mathbb{E}[v_{t_2}'|\mathcal{H}_{t_2-1}]\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(v_{t_1}' - \mathbb{E}[v_{t_1}'|\mathcal{H}_{t_1-1}]) \cdot 0\big] = 0.
\end{aligned}$$

The last convergence uses Assumption 3.2.

Check (B.18): In fact, (B.18) is a direct consequence of (C.27), (C.28), and (C.29).

B.4 Proof of Theorem 4.1

Fix any $a \in \mathcal{A}$. Denote $f_{\boldsymbol{x}}(\boldsymbol{\beta}) := \pi(a|\boldsymbol{x}, \boldsymbol{\beta})$. Then $\forall \epsilon, \delta > 0$,

$$\begin{aligned}
\mathbb{P}(|\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}) - \bar{\pi}(a|\boldsymbol{X}_t)| > \epsilon) &= \mathbb{P}(|\pi_t(a|\boldsymbol{X}_t, \boldsymbol{\beta}_{t-1}) - \pi(a|\boldsymbol{X}_t, \boldsymbol{\beta}^*)| > \epsilon) \\
&= \mathbb{P}(|f_{\boldsymbol{X}_t}(\widehat{\boldsymbol{\beta}}_{t-1}) - f_{\boldsymbol{X}_t}(\boldsymbol{\beta}^*)| > \epsilon) \\
&\leq \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^*\|_2 \geq \delta) + \mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{D}(f_{\boldsymbol{X}_t})) + \mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{B}_{\epsilon,\delta}(\boldsymbol{X}_t)),
\end{aligned} \tag{B.20}$$

Here $\forall \boldsymbol{x}$, we define

$$\begin{aligned}
\mathcal{D}(f_{\boldsymbol{x}}) &:= \{\boldsymbol{\beta} : f_{\boldsymbol{x}} \text{ is discontinuous at } \boldsymbol{\beta}\}, \\
\mathcal{B}_{\epsilon,\delta}(\boldsymbol{x}) &:= \{\boldsymbol{\beta} \notin \mathcal{D}(f_{\boldsymbol{x}}) : \exists \boldsymbol{\beta}', \text{ s.t.} \|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2 < \delta, |f_{\boldsymbol{x}}(\boldsymbol{\beta}) - f_{\boldsymbol{x}}(\boldsymbol{\beta}')| > \epsilon\}.
\end{aligned}$$

Notice that:

- From condition (i), for the first term on the RHS of (B.20), we have $\limsup_{t\to\infty} \mathbb{P}(\|\widehat{\boldsymbol{\beta}}_{t-1} - \boldsymbol{\beta}^*\|_2 \geq \delta) = 0$,

- From condition (ii), for the second term on the RHS of (B.20), we have $\mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{D}(f_{\boldsymbol{X}_t})) = 0$ $\forall t$,
- Due to $\boldsymbol{X}_t$ are i.i.d., the last two terms on the RHS of (B.20) does not depend on time t.

Plugging in the above into (B.20), we obtain that $\forall \epsilon, \delta > 0$,

$$\limsup_{t\to\infty} \mathbb{P}(|\pi_t(a|\boldsymbol{X}_t, \mathcal{H}_{t-1}) - \bar{\pi}(a|\boldsymbol{X}_t)| > \epsilon) \leq \mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{B}_{\epsilon,\delta}(\boldsymbol{X}_t)). \tag{B.21}$$

Finally, by definition of continuity, we have that $\forall \boldsymbol{x}$,

$$\lim_{\delta\to 0} 1_{\{\boldsymbol{\beta}^* \in \mathcal{B}_{\epsilon,\delta}(\boldsymbol{x})\}} = 0.$$

From the dominated convergence theorem,

$$\lim_{\delta\to 0} \mathbb{P}(\boldsymbol{\beta}^* \in \mathcal{B}_{\epsilon,\delta}(\boldsymbol{X}_t)) = \lim_{\delta\to 0} \mathbb{E}1_{\{\boldsymbol{\beta}^* \in \mathcal{B}_{\epsilon,\delta}(\boldsymbol{X}_t)\}} = 0.$$

We obtain the desired result by plugging the above into (B.21).

B.5 Proof of Proposition 4.1

We first present the following lemma. Its proof is in Appendix C.6.

Lemma B.6. *Consider a setting where the behavior policy does not use the contexts, i.e.,*

$$\pi = \{\pi_t(\cdot|\mathcal{H}^0_{t-1})\}_{t\geq 1},$$

where $\mathcal{H}^0_t := \sigma(\{A_\tau, Y_\tau\}_{\tau=1}^t)$ for $t \geq 1$. Assume that the potential outcomes have uniformly bounded variance, i.e., $\sup_{a\in\mathcal{A}} \mathrm{Var}(Y(a)) \leq \sigma_Y^2$ for some constant σ_Y^2. Suppose Assumption 3.4 is satisfied for all arm $a \in \mathcal{A}$, and we assume unconfoundedness: $\forall t \in [T]$,

$$A_t \perp \{Y_t(a)\}_{a\in\mathcal{A}} | \mathcal{H}^0_{t-1}. \tag{B.22}$$

Then as $t \to \infty$,

(i) For any deterministic sequence $\{C_t\}_{t\geq 1}$ such that $\lim_{t\to\infty} \frac{C_t}{t} = 0$, $\frac{C_t}{N_{a,t}} \xrightarrow{p} 0$ for any $a \in \mathcal{A}$;

(ii) $\widehat{\mu}_{a,t} \xrightarrow{p} \mu^*_a$ for any $a \in \mathcal{A}$.

Below we analyze the three common multi-arm bandits algorithms that ensure a minimum sampling probability.

The ϵ-greedy algorithm. Consider the ϵ-greedy policy defined by (4.3). Define

$$\pi_t^{(1)}(a|\{\widehat{\mu}_{i,t-1}\}_{i\in\mathcal{A}}) := \pi_t^{\epsilon\text{-greedy}}(a|\mathcal{H}_{t-1}) = \begin{cases} 1 - \frac{K-1}{K}\epsilon, & \text{if } i = \mathrm{argmax}_i\, \widehat{\mu}_{i,t-1}, \\ \frac{1}{K}\epsilon, & \text{otherwise.} \end{cases}$$

Then $\pi^{(1)}$ satisfies Assumption 3.4 with $\pi_{\min} = \epsilon/K$. Under the setting of Proposition 4.1, the bounded variance condition and the unconfoundedness condition in Lemma B.6 is also satisfied. Thus, according to part (i) of Lemma B.6, condition (i) of Theorem 4.1 is satisfied with $\boldsymbol{\beta}_t =$

$(\widehat{\mu}_{a,t})_{a\in\mathcal{A}}$, $\boldsymbol{\beta}^* = (\mu_a^*)_{a\in\mathcal{A}}$. In addition, condition (ii) of Theorem 4.1 holds for $\pi^{(1)}$ as long as $\Delta > 0$. From Theorem 4.1, we deduce that the policy convergence condition as in Definition 3.1 holds.

The UCB algorithm. Consider the clipped UCB algorithm defined by (4.4). Let

$$\pi^{(2)}(a|\{\widehat{\mu}_{i,t-1}, C_t/N_{i,t-1}\}_{i\in\mathcal{A}}) := \pi_t^{\mathrm{UCB}}(a|\mathcal{H}_{t-1}) = \begin{cases} 1-(K-1)\pi_{\min}, & \text{if } i = \mathrm{argmax}_i \left\{\widehat{\mu}_{i,t-1} + \sqrt{\frac{C_t}{N_{i,t-1}}}\right\}, \\ \pi_{\min}, & \text{otherwise.} \end{cases}$$

Then $\pi^{(2)}$ satisfies Assumption 3.4, and similar to the previous case, the bounded variance condition and the unconfoundedness condition in Lemma B.6 also hold. According to part (i) and (ii) of Lemma B.6, condition (i) of Theorem 4.1 is satisfied with $\boldsymbol{\beta}_t = (\widehat{\mu}_{a,t}, C_{t+1}/N_{a,t})_{a\in\mathcal{A}}$, $\boldsymbol{\beta}^* = (\mu_a^*, 0)_{a\in\mathcal{A}}$. Also, condition (ii) of Theorem 4.1 holds for $\pi^{(2)}$ as long as $\Delta > 0$. From Theorem 4.1, we deduce that the policy convergence condition as in Definition 3.1 holds.

The TS algorithm. Define

$$\bar{\pi}^{(3)}(a|\boldsymbol{\mu}_{t-1}^{\mathrm{post}}, \boldsymbol{\Sigma}_{t-1}^{\mathrm{post}}) = \bar{\pi}_t^{\mathrm{TS}}(a|\mathcal{H}_{t-1}),$$

where the dependence of $\bar{\pi}^{(3)}$ on $\boldsymbol{\mu}_{t-1}^{\mathrm{post}}$ and $\boldsymbol{\Sigma}_{t-1}^{\mathrm{post}}$ is shown in Section 4.1. In addition, define

$$\pi^{(3)}(a|\boldsymbol{\mu}_{t-1}^{\mathrm{post}}, \boldsymbol{\Sigma}_{t-1}^{\mathrm{post}}) = \left[\mathrm{Clip}\big(\bar{\pi}^{(3)}(i|\boldsymbol{\mu}_{t-1}^{\mathrm{post}}, \boldsymbol{\Sigma}_{t-1}^{\mathrm{post}})_{i\in\mathcal{A}}\big)\right]^{(a)},$$

Here for a vector $\boldsymbol{v} \in \mathbb{R}^K$, $[\boldsymbol{v}]^{(a)}$ denotes its a-th entry. From (4.5), we deduce that

$$\pi^{(3)}(a|\boldsymbol{\mu}_{t-1}^{\mathrm{post}}, \boldsymbol{\Sigma}_{t-1}^{\mathrm{post}}) = \pi_t^{\mathrm{TS}}(a|\mathcal{H}_{t-1}).$$

We now show that conditions (i) and (ii) of Theorem 4.1 holds with $\pi^{(3)}$ as the behavior policy and with statistics $\boldsymbol{\beta}_t = (\boldsymbol{\mu}_t^{\mathrm{post}}, \boldsymbol{\Sigma}_t^{\mathrm{post}})$. First, $\pi^{(3)}$ satisfies Assumption 3.4, the bounded variance condition and the unconfoundedness condition in Lemma B.6. From Lemma B.6, we obtain that $\forall a \in \mathcal{A}$, $1/N_{a,t} \xrightarrow{p} 0$ and $\widehat{\mu}_{a,t} \xrightarrow{p} \mu_a^*$. Combining these results together with (4.7) and the continuous mapping theorem, we deduce that $\forall a \in \mathcal{A}$,

$$\mu_{a,t}^{\mathrm{post}} \xrightarrow{p} \mu_a^*, \quad \sigma_{a,t}^{\mathrm{post}} \xrightarrow{p} 0.$$

Note that entrywise convergence in probability implies joint convergence in probability in a finite dimensional Euclidean space, and thus from (4.6) we have

$$\boldsymbol{\mu}_t^{\mathrm{post}} \xrightarrow{p} \boldsymbol{\mu}^*, \quad \boldsymbol{\Sigma}_t^{\mathrm{post}} \xrightarrow{p} \mathbf{0}_{K\times K},$$

where $\boldsymbol{\mu}^* = (\mu_a^*)_{a\in\mathcal{A}}$. This implies condition (i) of Theorem 4.1 holds with the statistics $\boldsymbol{\beta}_t$ and limit $\boldsymbol{\beta}^* = (\boldsymbol{\mu}^*, \mathbf{0}_{K\times K})$.

In addition, it is not difficult to verify that $\bar{\pi}^{(3)}$ is continuous at $\boldsymbol{\beta}^*$ as long as the suboptimality gap $\Delta > 0$. Also, Clip is a continuous mapping (see Lemma B.9). Therefore, the composite mapping $\pi^{(3)}$ is continuous at $\boldsymbol{\beta}^*$, and condition (ii) of Theorem 4.1 holds.

In summary, we have verified both conditions (i) and (ii) of Theorem 4.1. We deduce from the theorem that the policy convergence condition in Definition 3.1 holds.

B.6 Proof of Theorem 4.2

We adopt the the convergence analysis of the recursive least square fitting based on the stochastic approximation theory. We first rewrite the ridge regression estimator in (4.10) in a recursive form. Define $\boldsymbol{X}_{t,a} = \boldsymbol{X}_t 1_{\{A_t=a\}}$ and $Y_{t,a} = Y_t 1_{\{A_t=a\}}$. Let the sample covariance matrix and the sample cross product matrix be

$$\boldsymbol{\Phi}_{t,a} = \frac{1}{t-1}\left(\lambda I + \sum_{i=1}^{t-1} \boldsymbol{X}_{i,a}\boldsymbol{X}_{i,a}^\top\right), \quad \text{and} \quad \boldsymbol{\varphi}_{t,a} = \frac{1}{t-1}\left(\sum_{i=1}^{t-1} \boldsymbol{X}_{i,a} Y_{i,a}\right). \tag{B.23}$$

Based on the definition of ridge regression estimator, we directly have that

$$\widehat{\boldsymbol{\theta}}_{t,a}^{\text{Ridge}} = \left(\lambda I + \sum_{i=1}^{t-1} \boldsymbol{X}_{i,a}\boldsymbol{X}_{i,a}^\top\right)^{-1}\left(\sum_{i=1}^{t-1} \boldsymbol{X}_{i,a} Y_i\right) = \boldsymbol{\Phi}_{t,a}^{-1}\boldsymbol{\varphi}_{t,a}. \tag{B.24}$$

Thus, it is sufficient to show that both $\boldsymbol{\Phi}_{t,a}$ and $\boldsymbol{\varphi}_{t,a}$ converge as $t \to \infty$. We show this by writing the recursive form for $\boldsymbol{\Phi}_{t,a}$ and $\boldsymbol{\varphi}_{t,a}$. Denote by $\text{vec}(M)$ the vectorized form of a matrix M. We have that

$$\text{vec}(\boldsymbol{\Phi}_{t+1,a}) = \text{vec}(\boldsymbol{\Phi}_{t,a}) + \frac{1}{t}\,\text{vec}\left(\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top - \boldsymbol{\Phi}_{t,a}\right) \tag{B.25}$$

$$\boldsymbol{\varphi}_{t+1,a} = \boldsymbol{\varphi}_{t,a} + \frac{1}{t}(\boldsymbol{X}_{t,a}Y_{t,a} - \boldsymbol{\varphi}_{t,a}). \tag{B.26}$$

We may write the recursive form for the joint vector of $\text{vec}(\Phi_{t,a})$ and $\varphi_{t,a}$ as

$$\begin{pmatrix} \boldsymbol{\varphi}_{t+1,a} \\ \text{vec}(\boldsymbol{\Phi}_{t+1,a}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\varphi}_{t,a} \\ \text{vec}(\boldsymbol{\Phi}_{t,a}) \end{pmatrix} + \frac{1}{t}\begin{pmatrix} \boldsymbol{X}_{t,a}Y_{t,a} - \boldsymbol{\varphi}_{t,a} \\ \text{vec}\left(\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top - \boldsymbol{\Phi}_{t,a}\right) \end{pmatrix} \tag{B.27}$$

We denote by $\boldsymbol{\psi}_{t,a} = \begin{pmatrix} \boldsymbol{\varphi}_{t,a} \\ \text{vec}(\boldsymbol{\Phi}_{t,a}) \end{pmatrix}$ and $\boldsymbol{\psi}_t = \left(\boldsymbol{\psi}_{t,1}^\top, \boldsymbol{\psi}_{t,2}^\top, \ldots, \boldsymbol{\psi}_{t,|\mathcal{A}|}^\top\right)^\top$. Concatenating the above recursive form for all $a \in \mathcal{A}$ into a single expression, we have

$$\boldsymbol{\psi}_{t+1} = \boldsymbol{\psi}_t + w_t\left(\boldsymbol{g}(\boldsymbol{\psi}_t) - \boldsymbol{\psi}_t + \boldsymbol{M}_t\right), \tag{B.28}$$

where for any $\boldsymbol{\psi} = (\boldsymbol{\varphi}_1^\top, \text{vec}(\boldsymbol{\Phi}_1)^\top, \ldots, \boldsymbol{\varphi}_{|\mathcal{A}|}^\top, \text{vec}(\boldsymbol{\Phi}_{|\mathcal{A}|})^\top \in \mathbb{R}^{d+d^2|\mathcal{A}|}$, we have

$$w_t = \frac{1}{t}, \tag{B.29}$$

$$\boldsymbol{g}\left(\boldsymbol{\psi}\right) = \left(g_1(\boldsymbol{\psi})^\top, g_2(\boldsymbol{\psi})^\top, \ldots, g_{|\mathcal{A}|}(\boldsymbol{\psi})^\top\right)^\top, \tag{B.30}$$

$$\boldsymbol{g}_a(\boldsymbol{\psi}) = \begin{pmatrix} \mathbb{E}\left[\boldsymbol{X}_{t,a}Y_{t,a} \mid \boldsymbol{\varphi}_{t,a} = \boldsymbol{\varphi}_a, \text{vec}(\boldsymbol{\Phi}_{t,a}) = \text{vec}(\boldsymbol{\Phi}_a), \forall a \in \mathcal{A}\right] \\ \mathbb{E}\left[\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top \mid \boldsymbol{\varphi}_{t,a} = \boldsymbol{\varphi}_a, \text{vec}(\boldsymbol{\Phi}_{t,a}) = \text{vec}(\boldsymbol{\Phi}_a), \forall a \in \mathcal{A}\right] \end{pmatrix}, \tag{B.31}$$

$$\boldsymbol{M}_t = \left(\boldsymbol{M}_{t,1}^\top, \boldsymbol{M}_{t,2}^\top, \ldots, \boldsymbol{M}_{t,|\mathcal{A}|}^\top\right)^\top, \tag{B.32}$$

$$\boldsymbol{M}_{t,a} = \begin{pmatrix} \boldsymbol{X}_{t,a}Y_{t,a} \\ \text{vec}(\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top) \end{pmatrix} - \boldsymbol{g}_a(\boldsymbol{\psi}_t) \tag{B.33}$$

Typical convergence analysis requires the following three conditions:

- (C.1) $\sup_t \mathbb{E}[\|\boldsymbol{\psi}_t\|_2^2] < \infty$
- (C.2) Contraction mapping condition: $\|\boldsymbol{g}(\boldsymbol{\psi}_1) - \boldsymbol{g}(\boldsymbol{\psi}_2)\| \leq \kappa \|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_2\|$ for some $\kappa < 1$
- (C.3) $\{\boldsymbol{M}_t\}$ is a martingale difference sequence with respect to the increasing family of the σ-field
$$\mathcal{F}_t = \sigma(A_1, \boldsymbol{X}_1, Y_1, \ldots, A_t, \boldsymbol{X}_t, Y_t).$$
That is $\mathbb{E}[\boldsymbol{M}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{0}$. Furthermore, $\{\boldsymbol{M}_t\}$ are square integrable, i.e., $\mathbb{E}[\|\boldsymbol{M}_t\|_2^2 \mid \mathcal{F}_{t-1}] < K(1 + \|\boldsymbol{\psi}_t\|_2^2)$ a.s., $t \geq 1$ for some constant K

For sufficiently large t, we show that the above three conditions hold for all $\tau = t, t+1, \ldots$ under some good event $\mathcal{E}_t$ with a high probability $1 - 2/(t-1)$. The probability is chosen so when t goes infinity, the failure probability of $\mathcal{E}_t$ goes to 0.

B.6.1 Good events

The convergence analysis is based on the following good event $\mathcal{E}_t$ being satisfied with a high probability $1 - 2/(t-1)$. We define $\mathcal{E}_t = \mathcal{E}_{1,t} \cap \mathcal{E}_{2,t}$ where

$$\mathcal{E}_{1,t} = \left\{ \|\Phi_{\tau,a}\|_2 \geq \pi_{\min}\sigma_{\min}^2 - M\sqrt{\frac{16\log(\tau A d)}{\tau - 1}}, \quad \text{for all } \tau \geq t \text{ and } a \in \mathcal{A} \right\} \tag{B.34}$$

$$\mathcal{E}_{2,t} = \left\{ \|\varphi_{\tau,a}\|_2 \leq M^2 R_\theta + 2\sigma_\eta M\sqrt{\frac{2d\log(10A\tau)}{\tau - 1}}, \quad \text{for all } \tau \geq t \text{ and } a \in \mathcal{A} \right\} \tag{B.35}$$

Here we let $1/0 = \infty$.

Lemma B.7. *For any $t = 2, \ldots,$ we have*

$$\mathbb{P}\left(\{\mathcal{E}_t \text{ holds}\}\right) \geq 1 - 2/(t-1). \tag{B.36}$$

Proof. Consider the martingale difference sequence $\{\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top - \mathbb{E}[\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top \mid \mathcal{F}_{t-1}]\}_t$, where $\mathcal{F}_{t-1}$ is the σ-field generated by $\left\{\{X_{\tau,a}, A_{\tau,a}, Y_{\tau,a}\}_{\tau=1}^{t-1}\right\}_{a\in\mathcal{A}}$. Based on Assumption 4.1 (A.1), the martingale difference sequence satisfies

$$\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top - \mathbb{E}[\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top \mid \mathcal{F}_{t-1}] \preceq M^2 I. \tag{B.37}$$

We first apply Corollary D.1 for all $a \in \mathcal{A}$ with a probability of at least $1 - 1/(A(t-1))$ to get the following concentration bound simultaneously for all $a \in \mathcal{A}$ and $\tau \geq t$:

$$\frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1}\left(\boldsymbol{X}_{k,a}\boldsymbol{X}_{k,a}^\top - \mathbb{E}[\boldsymbol{X}_{k,a}\boldsymbol{X}_{k,a}^\top \mid \mathcal{F}_{k-1}]\right)\right\|_2 \leq M\sqrt{\frac{16\log(A\tau d)}{\tau - 1}}. \tag{B.38}$$

Because the sampling probability is clipped below by $\pi_{\min}$, we have that for all $k \in \mathbb{N}$ and $a \in \mathcal{A}$,

$$\pi_{\min}\sigma_{\min}^2 I \preceq \mathbb{E}[\boldsymbol{X}_{k,a}\boldsymbol{X}_{k,a}^\top \mid \mathcal{F}_{k-1}]. \tag{B.39}$$

Therefore, with a probability of at least $1 - 1/(t-1)$, for all $a \in \mathcal{A}$ and $\tau \geq t$,

$$\pi_{\min}\sigma_{\min}^2 - M\sqrt{\frac{16\log(A\tau d)}{\tau - 1}} \leq \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}\boldsymbol{X}_{k,a}^\top\right\|_2. \tag{B.40}$$

Thus, we have with a probability of at least $1 - 1/(t-1)$, for all $a \in \mathcal{A}$ and $\tau \geq t$,

$$\pi_{\min}\sigma_{\min}^2 - M\sqrt{\frac{16\log(A\tau d)}{\tau - 1}} \leq \|\Phi_{\tau,a}\|_2. \tag{B.41}$$

This is our definition of the event $\mathcal{E}_{1,t}$.

We further show that $\mathcal{E}_{2,t}$ holds with a probability of at least $1 - 1/(t-1)$. For any $a \in \mathcal{A}$, and $\tau \geq t$, we have that

$$\begin{aligned}
\|\varphi_{\tau,a}\|_2 &= \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}Y_{k,a}\right\|_2 && \text{(B.42)}\\
&= \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}(\langle \boldsymbol{\theta}_a^*, S_{k,a}\rangle + \eta_k)\right\|_2 && \text{(B.43)}\\
&= \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}S_{k,a}^\top\boldsymbol{\theta}_a^* + \sum_{k=1}^{\tau-1}\boldsymbol{X}_{k,a}\eta_k\right\|_2 && \text{(B.44)}\\
&\leq \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}S_{k,a}^\top\boldsymbol{\theta}_a^*\right\|_2 + \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1}\boldsymbol{X}_{k,a}\eta_k\right\|_2 && \text{(B.45)}\\
&\leq M^2 R_\theta + \frac{1}{\tau - 1}\left\|\sum_{k=1}^{\tau-1}\boldsymbol{X}_{k,a}\eta_k\right\|_2 && \text{(B.46)}
\end{aligned}$$

For any fixed unit vector u, variable $u^\top(\boldsymbol{X}_{k,a}\eta_k)$ is sub-Gaussian with parameter $M\sigma_\eta$. Thus, by standard concentration for sub-Gaussian sums, for any fixed u, and any $\delta > 0$, we have that with a probability of at least $1 - \delta$,

$$\left\|\frac{1}{\tau - 1}\sum_{k=1}^{\tau-1} u^\top\boldsymbol{X}_{k,a}\eta_k\right\|_2 \leq 2\sigma_\eta M\sqrt{\frac{\log(2/\delta)}{\tau - 1}}. \tag{B.47}$$

To obtain a bound on the Euclidean norm, we apply the standard covering argument over the unit sphere in $\mathbb{R}^d$:

$$\left\|\frac{1}{\tau - 1}\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}\eta_k\right\|_2 \leq 2\sigma_\eta M\sqrt{\frac{d\log(10/\delta)}{\tau - 1}}. \tag{B.48}$$

We apply the above bound on each $\tau \geq t$ and $a \in \mathcal{A}$ separately with $\delta = 1/(t^2 A)$. A union bound over all $\tau \geq t$ and $a \in \mathcal{A}$ gives that with probability at least $1 - 1/(t-1)$,

$$\left\|\frac{1}{\tau - 1}\sum_{k=1}^{\tau-1} \boldsymbol{X}_{k,a}\eta_k\right\|_2 \leq 2\sigma_\eta M\sqrt{\frac{d\log(10A\tau)}{\tau - 1}}. \tag{B.49}$$

Combined with (B.46), we have that event $\mathcal{E}_{t,2}$ holds with a probability of at least $1 - 1/(t-1)$. □

B.6.2 Contraction property of g

The following contraction property is shown under the good event $\mathcal{E}_t$.

Lemma B.8. *On event $\mathcal{E}_t$, let $\boldsymbol{\psi}, \boldsymbol{\psi}'$ be two elements in $\{\boldsymbol{\psi}_\tau\}_{\tau \geq t} \subseteq \mathbb{R}^{(d+d^2)|\mathcal{A}|}$. Consider sufficiently large t such that for all $\tau \geq t$,*

$$2\sigma_\eta M \sqrt{\frac{2d \log(10A\tau)}{\tau - 1}} \leq M^2 R_\theta, \tag{B.50}$$

$$\pi_{\min}\sigma_{\min}^2 - M\sqrt{16 \log(Ad\tau)/(\tau-1)} \geq \frac{1}{2}\pi_{\min}\sigma_{\min}^2. \tag{B.51}$$

Then, for sufficiently large γ, for any $\boldsymbol{\psi}, \boldsymbol{\psi}' \in \{\boldsymbol{\psi}_\tau\}_{\tau \geq t}$, we have

$$\|\boldsymbol{g}(\boldsymbol{\psi}) - \boldsymbol{g}(\boldsymbol{\psi}')\|_2 \leq \kappa \|\boldsymbol{\psi} - \boldsymbol{\psi}'\|_2 \text{ for some } \kappa < 1. \tag{B.52}$$

Proof. Define $m_a(\boldsymbol{\psi}) = \mathbb{E}[\boldsymbol{X}_{t,a} Y_{t,a} \mid \boldsymbol{\psi}_t = \boldsymbol{\psi}]$, and $h_a(\boldsymbol{\psi}) = \mathbb{E}[\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top \mid \boldsymbol{\psi}_t = \boldsymbol{\psi}']$.

We have that

$$\|m_a(\boldsymbol{\psi}) - m_a(\boldsymbol{\psi}')\|_2 = \left\|\mathbb{E}[(\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}) - \widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}'))\boldsymbol{X}_t S_t^\top \theta_a^*]\right\|_2 \tag{B.53}$$

$$\leq \left|\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}) - \widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}')\right| \left\|\boldsymbol{X}_t S_t^\top \theta_a^*\right\|_2 \tag{B.54}$$

$$\leq \left|\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}) - \widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}')\right| M^2 R_\theta. \tag{B.55}$$

Similarly, we have that

$$\|h_a(\boldsymbol{\psi}) - h_a(\boldsymbol{\psi}')\|_2 = \left\|\mathbb{E}[(\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}) - \widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}'))\boldsymbol{X}_t \boldsymbol{X}_t^\top]\right\| \tag{B.56}$$

$$\leq \left|\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}) - \widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi}')\right| M^2. \tag{B.57}$$

It suffices to upper bound the Lipschitz constant of $\widetilde{\pi}^\gamma(\cdot \mid \boldsymbol{X}_t, \boldsymbol{\psi})$. We denote by $\widetilde{\boldsymbol{\pi}} = \widetilde{\boldsymbol{\pi}}^\gamma(\cdot \mid \boldsymbol{X}_t, \boldsymbol{\psi}) \in [0,1]^{|\mathcal{A}|}$ and $\widetilde{\boldsymbol{\pi}'} = \widetilde{\boldsymbol{\pi}}^\gamma(\cdot \mid \boldsymbol{X}_t, \boldsymbol{\psi}')$, and similarly define $\boldsymbol{\pi} = \boldsymbol{\pi}^\gamma(\cdot \mid \boldsymbol{X}_t, \boldsymbol{\psi})$ and $\boldsymbol{\pi}' = \boldsymbol{\pi}^\gamma(\cdot \mid \boldsymbol{X}_t, \boldsymbol{\psi}')$.

Lemma B.9 (L2 projection is Lipschitz continuous)**.** *Let* $\mathrm{Clip}(\boldsymbol{\pi})$ *be defined as in (A.1). Then,* $\mathrm{Clip}(\boldsymbol{\pi})$ *is $(|\mathcal{A}|+1)$-Lipschitz continuous in $\boldsymbol{\pi}$.*

Lemma B.9 implies that the clipping operator (defined in **??**) is $(|\mathcal{A}|+1)$-Lipschitz, i.e., $\|\widetilde{\boldsymbol{\pi}} - \widetilde{\boldsymbol{\pi}}'\|_2 \leq (|\mathcal{A}|+1)\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2$. By the fact that the softmax function is $2/\gamma$-Lipschitz, we have that

$$\|\widetilde{\boldsymbol{\pi}} - \widetilde{\boldsymbol{\pi}}'\|_2 \leq (|\mathcal{A}|+1)\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \leq \frac{2(|\mathcal{A}|+1)M}{\gamma}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2, \tag{B.58}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{|\mathcal{A}|})$ and each $\boldsymbol{\theta}_a = \boldsymbol{\Phi}_a^{-1}\boldsymbol{\varphi}_a$.

To bound $\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$, we have that

$$\|\boldsymbol{\theta}_a - \boldsymbol{\theta}_a'\|_2 = \|\boldsymbol{\Phi}_a^{-1}\boldsymbol{\varphi}_a - \boldsymbol{\Phi}_a'^{-1}\boldsymbol{\varphi}_a'\|_2 \tag{B.59}$$

$$\leq \|\boldsymbol{\Phi}_a^{-1}\boldsymbol{\varphi}_a - \boldsymbol{\Phi}_a^{-1}\boldsymbol{\varphi}_a'\|_2 + \|\boldsymbol{\Phi}_a^{-1}\boldsymbol{\varphi}_a' - \boldsymbol{\Phi}_a'^{-1}\boldsymbol{\varphi}_a'\|_2 \tag{B.60}$$

$$\leq \|\boldsymbol{\Phi}_a^{-1}\|_2\|\boldsymbol{\varphi}_a - \boldsymbol{\varphi}_a'\|_2 + \|\boldsymbol{\Phi}_a^{-1} - \boldsymbol{\Phi}_a'^{-1}\|_2\|\boldsymbol{\varphi}_a'\|_2 \tag{B.61}$$

$$\leq \|\boldsymbol{\Phi}_a^{-1}\|_2\|\boldsymbol{\varphi}_a - \boldsymbol{\varphi}_a'\|_2 + \|\boldsymbol{\Phi}_a^{-1}\|_2\|\boldsymbol{\Phi}_a - \boldsymbol{\Phi}_a'\|_2\|\boldsymbol{\Phi}_a'^{-1}\|\|\boldsymbol{\varphi}_a'\|_2 \tag{B.62}$$

Under the good event $\mathcal{E}_t$, we have that both $\|\boldsymbol{\Phi}_a'^{-1}\|_2, \|\boldsymbol{\Phi}_a^{-1}\|_2 \leq 2/(\pi_{\min}\sigma_{\min}^2)$, and $\|\boldsymbol{\varphi}_a\|_2 \leq 2M^2R_\theta$.

Thus, we have that

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq |\mathcal{A}| \left(\frac{2}{\pi_{\min}\sigma_{\min}^2} \|\boldsymbol{\varphi}_a - \boldsymbol{\varphi}_a'\|_2 + \frac{8M^2R_\theta}{\pi_{\min}^2\sigma_{\min}^4} \|\boldsymbol{\Phi}_a - \boldsymbol{\Phi}_a'\|_2 \right) \tag{B.63}$$

Combined with (B.58), we have

$$\|\widetilde{\boldsymbol{\pi}} - \widetilde{\boldsymbol{\pi}}'\|_2 \leq \frac{2(|\mathcal{A}|+1)M}{\gamma} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \frac{2(|\mathcal{A}|+1)M}{\gamma} |\mathcal{A}| \left(\frac{2}{\pi_{\min}\sigma_{\min}^2} \|\boldsymbol{\varphi}_a - \boldsymbol{\varphi}_a'\|_2 + \frac{8M^2R_\theta}{\pi_{\min}^2\sigma_{\min}^4} \|\boldsymbol{\Phi}_a - \boldsymbol{\Phi}_a'\|_2 \right) \tag{B.64}$$

The Lipschitz constant of $\widetilde{\pi}^\gamma(a \mid \boldsymbol{X}_t, \boldsymbol{\psi})$ w.r.t. the parameter $\boldsymbol{\psi}$ is at most

$$\frac{2|\mathcal{A}|(|\mathcal{A}|+1)M}{\gamma} \max\left\{ \frac{2}{\pi_{\min}\sigma_{\min}^2}, \frac{8M^2R_\theta}{\pi_{\min}^2\sigma_{\min}^4} \right\}. \tag{B.65}$$

Combined with (B.55) and (B.57), the Lipschitz constant of function $\boldsymbol{g}$ is

$$\frac{2|\mathcal{A}|(|\mathcal{A}|+1)M^3}{\gamma} \max\left\{ \frac{2}{\pi_{\min}\sigma_{\min}^2}, \frac{8M^2R_\theta}{\pi_{\min}^2\sigma_{\min}^4} \right\} \max\{R_\theta, 1\} \tag{B.66}$$

For sufficiently large γ, this constant is less than 1. □

B.6.3 Boundedness of moments of ψ_t and M_t

Under the good event $\mathcal{E}_t$, for sufficiently large t, we have that for all $a \in \mathcal{A}$, and $\tau \geq t$,

$$\|\boldsymbol{\Phi}_{\tau,a}\|_2 \leq M^2, \text{ and } \|\boldsymbol{\varphi}_a\|_2 \leq 2M^2R_\theta. \tag{B.67}$$

We also have that for all $a \in \mathcal{A}$, and $\tau \geq t$,

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{M}_{\tau,a}\|_2 \mid \mathcal{F}_{t-1}]s &\leq \mathbb{E}[\|\boldsymbol{X}_{t,a}Y_{t,a}\|_2 + \|\operatorname{vec}(\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top)\|_2] && \text{(B.68)}\\
&= \mathbb{E}[\|\boldsymbol{X}_{t,a}(S_t^\top \boldsymbol{\theta}_a^* + \eta_t)\|_2 + \|\operatorname{vec}(\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top)\|_2] && \text{(B.69)}\\
&\leq M^2R_\theta + M\sigma_\eta + M^2. && \text{(B.70)}
\end{aligned}$$

Thus conditions (C.1) and (C.3) are satisfied.

B.6.4 Finishing the proof

Now we are ready to show the convergence of $\boldsymbol{\psi}_t$. The above contraction property guarantees that on event $\mathcal{E}_t$, the sequence $\{\boldsymbol{\psi}_\tau\}_{\tau \geq t}$ converges due to the Robbins-Monro theorem [10]. To show the general convergence, we follow the following argument.

Our goal is to show that for any $\epsilon > 0, \delta > 0$, there exists T_0 such that for all $t \geq T_0$,

$$\mathbb{P}(\|\boldsymbol{\psi}_t - \boldsymbol{\psi}^*\|_2 \geq \epsilon) \leq \delta. \tag{B.71}$$

Now for any $\delta > 0$, let $t_0 = \lceil 4/\delta + 1 \rceil$. Then we have $\mathbb{P}(\mathcal{E}_{t_0}^c) \leq \delta/2$ due to Lemma B.7. Therefore, we have that

$$\mathbb{P}(\|\boldsymbol{\psi}_t - \boldsymbol{\psi}^*\|_2 \geq \epsilon) \leq \mathbb{P}(\{\|\boldsymbol{\psi}_t - \boldsymbol{\psi}^*\|_2 \geq \epsilon\} \cap \mathcal{E}_{t_0}) + \mathbb{P}(\mathcal{E}_{t_0}^c). \tag{B.72}$$

The convergence of $\{\boldsymbol{\psi}_\tau\}_{\tau \geq t_0}$ to $\boldsymbol{\psi}^*$ on event $\mathcal{E}_{t_0}$ implies that there exists a $T_0 \geq t_0$ such that for all $t \geq T_0$, the first term on the right-hand side is less than $\delta/2$. Therefore, we have that

$$\mathbb{P}(\|\boldsymbol{\psi}_t - \boldsymbol{\psi}^*\|_2 \geq \epsilon) \leq \delta \text{ for all } t \geq T_0. \tag{B.73}$$

This shows that $\boldsymbol{\psi}_t$ converges in probability to $\boldsymbol{\psi}^*$ as $t \to \infty$.

B.7 Proof of Theorem 4.4

Proof. Condition 2 combined with Theorem 7 [10] gives that $\sup_{t \geq 0} \|\widehat{\boldsymbol{\theta}}_t^{\mathrm{SGD}}\|_2$ is bounded almost surely. Then, it follows directly from the convergence result of stochastic approximation process. In the following lemma, we show that condition 2 is satisfied for our two examples.

Lemma B.10. *The conditions in Theorem 4.4 hold for our two examples.*

Proof. In example one, we have

$$\boldsymbol{h}(x, y; \boldsymbol{\theta}_a) = (y - \boldsymbol{\theta}_a^\top x)x. \tag{B.74}$$

Thus, we have

$$\begin{aligned}
\bar{\boldsymbol{\phi}}(\boldsymbol{\theta}) &= \mathbb{E}[(Y_t - \boldsymbol{\theta}_{A_t}^\top \boldsymbol{X}_t)\boldsymbol{X}_t \mid \widehat{\boldsymbol{\theta}}_t^{\mathrm{SGD}} = \boldsymbol{\theta}] && \text{(B.75)} \\
&= \sum_a \mathbb{E}[\pi^\gamma(A_t = a \mid \boldsymbol{X}_t, \boldsymbol{\theta})(Y_t - \boldsymbol{\theta}_{A_t}^\top \boldsymbol{X}_t)\boldsymbol{X}_t] && \text{(B.76)} \\
&= \sum_a \mathbb{E}[\pi^\gamma(A_t = a \mid \boldsymbol{X}_t, \boldsymbol{\theta})(f(\boldsymbol{X}_t, a) - \boldsymbol{\theta}_{A_t}^\top \boldsymbol{X}_t)\boldsymbol{X}_t] && \text{(B.77)}
\end{aligned}$$

For any $c \geq 1$, we have

$$\begin{aligned}
\bar{\boldsymbol{\phi}}_c(\boldsymbol{\theta})/c &= \sum_a \mathbb{E}[\pi^\gamma(A_t = a \mid \boldsymbol{X}_t, c\boldsymbol{\theta})(f(\boldsymbol{X}_t, a) - c\boldsymbol{\theta}_a^\top \boldsymbol{X}_t)\boldsymbol{X}_t]/c && \text{(B.78)} \\
&\to -\sum_a \mathbb{E}[1\{a = \operatorname*{argmax}_{a'} \boldsymbol{X}_t^\top \boldsymbol{\theta}_{a'}\}\boldsymbol{X}_t^\top \boldsymbol{X}_t \boldsymbol{\theta}_a] \text{ as } c \to \infty. && \text{(B.79)}
\end{aligned}$$

Define $\Sigma_a(\boldsymbol{\theta}) \coloneqq \mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^\top 1\{a = \operatorname{argmax}_{a'} \boldsymbol{X}_t^\top \boldsymbol{\theta}_{a'}\}]$ for all $a \in \mathcal{A}$, and let $\Sigma(\boldsymbol{\theta}) \coloneqq \operatorname{diag}(\Sigma_1(\boldsymbol{\theta}), \ldots, \Sigma_{|\mathcal{A}|}(\boldsymbol{\theta}))$.

Lemma B.11. *If $\Sigma_X \coloneqq \mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^\top] \succ 0$, then $\Sigma_a(\boldsymbol{\theta}) \succ 0$ for all $a \in \mathcal{A}$, and therefore, $\Sigma(\boldsymbol{\theta}) \succ 0$.*

Then, we have

$$\bar{\boldsymbol{\phi}}_c(\boldsymbol{\theta})/c \to -\Sigma(\boldsymbol{\theta})\boldsymbol{\theta} \text{ as } c \to \infty. \tag{B.80}$$

It suffices to show that the o.d.e.

$$\dot{\boldsymbol{\theta}}(t) = -\Sigma(\boldsymbol{\theta}(t))\boldsymbol{\theta}(t)$$

has the origin as its unique globally asymptotically stable equilibrium.

We first show that the o.d.e. has a unique equilibrium solution. Setting $\dot{\boldsymbol{\theta}}(t) = 0$, we have

$$-\Sigma(\boldsymbol{\theta})\boldsymbol{\theta} = 0. \tag{B.81}$$

Since $\Sigma(\boldsymbol{\theta}) \succ 0$, we have that $\boldsymbol{\theta} = 0$ is the unique equilibrium solution.

Next, we show that the stability of the equilibrium solution. Consider the Lyapunov function:

$$V(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2. \tag{B.82}$$

We have

$$\dot{V}(\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \dot{\boldsymbol{\theta}}(t) = -\boldsymbol{\theta}^\top \Sigma(\boldsymbol{\theta})\boldsymbol{\theta} \leq 0. \tag{B.83}$$

Since $\dot{V}(\boldsymbol{\theta}) \leq 0$, the Lyapunov function strictly decreases along trajectories except at the origin. Because $V(\boldsymbol{\theta})$ is positive and radially unbounded, and $\dot{V}(\boldsymbol{\theta})$ is non-positive, by the standard Lyapunov stability theory (Theorem 4.1 [33]), the origin $\boldsymbol{\theta} = 0$ is globally asymptotically stable.

□

□

B.8 Proof of Theorem 4.3

Suppose that $\widehat{\boldsymbol{\theta}}_t^{\text{Ridge}} \xrightarrow{p} \boldsymbol{\theta}^*$. Recall the definition of $\widehat{\boldsymbol{\theta}}_t^{\text{Ridge}}$ in (4.10):

$$\widehat{\boldsymbol{\theta}}_t^{\text{Ridge}} = \left(\lambda I + \sum_{i=1}^{t-1} \boldsymbol{X}_{i,a}\boldsymbol{X}_{i,a}^\top\right)^{-1} \sum_{i=1}^{t-1} \boldsymbol{X}_{i,a} Y_{i,a}. \tag{B.84}$$

We first note that $\Sigma_a(\boldsymbol{\theta}) = \mathbb{E}[1_{A_t=a}\boldsymbol{X}_t\boldsymbol{X}_t^\top \mid \widehat{\boldsymbol{\theta}}_{t-1}^{\text{Ridge}} = \boldsymbol{\theta}]$ is a continuous function of $\boldsymbol{\theta}$ given any $a \in \mathcal{A}$, because we assume that $\pi(a \mid x, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ given any $x \in \mathcal{X}$.

Therefore, we have that $\Sigma_a(\widehat{\boldsymbol{\theta}}_{t-1}^{\text{Ridge}}) \xrightarrow{p} \Sigma_a(\bar{\boldsymbol{\theta}})$, and

$$\frac{1}{t}\left(\lambda I + \sum_{\tau=1}^{t-1} \Sigma_a(\widehat{\boldsymbol{\theta}}_{\tau,a}^{\text{Ridge}})\right) \xrightarrow{p} \Sigma_a(\bar{\boldsymbol{\theta}}). \tag{B.85}$$

Since $\boldsymbol{X}_{t,a}\boldsymbol{X}_{t,a}^\top - \Sigma_a(\widehat{\boldsymbol{\theta}}_{t-1}^{\text{Ridge}})$ is a zero-mean martingale difference sequence with bounded second moment, by the law of large numbers (Lemma D.2)[18], we have that

$$\frac{1}{t}\sum_{\tau=1}^{t-1}\left(X_{\tau,a}X_{\tau,a}^\top - \Sigma_a(\widehat{\boldsymbol{\theta}}_{\tau,a}^{\text{OLS}})\right) \xrightarrow{p} 0. \tag{B.86}$$

Therefore, we have that

$$\frac{1}{t}\left(\lambda I + \sum_{\tau=1}^{t-1} X_{\tau,a}X_{\tau,a}^\top\right) \xrightarrow{p} \Sigma_a(\bar{\boldsymbol{\theta}}). \tag{B.87}$$

Similar argument gives that

$$\frac{1}{t}\sum_{\tau=1}^{t-1}(X_{\tau,a}Y_{\tau,a}) \xrightarrow{p} \varphi_a(\bar{\boldsymbol{\theta}}). \tag{B.88}$$

Now we are ready to show the contradiction. We first decide on a particular $a \in \mathcal{A}$. Since $\sum_{a\in\mathcal{A}} \Sigma_a(\bar{\boldsymbol{\theta}}) \succeq \sigma_{\min}^2 I$ is positive definite, by the pigeonhole principle, there exists a $\bar{a} \in \mathcal{A}$ such that

$$\Sigma_{\bar{a}}(\bar{\boldsymbol{\theta}}) \succeq \sigma_{\min}^2 I/|\mathcal{A}|. \tag{B.89}$$

Since $M \mapsto M^{-1}$ is a continuous function on the set of positive definite matrices, we have that

$$\left(\frac{1}{t}\left(\lambda I + \sum_{\tau=1}^{t-1} X_{\tau,\bar{a}} X_{\tau,\bar{a}}^\top\right)\right)^{-1} \xrightarrow{p} \left(\Sigma_{\bar{a}}(\bar{\boldsymbol{\theta}})\right)^{-1}, \tag{B.90}$$

by the continuous mapping theorem.

Therefore, we have that

$$\widehat{\boldsymbol{\theta}}_{t,\bar{a}}^{\text{Ridge}} \xrightarrow{p} \left(\Sigma_{\bar{a}}(\bar{\boldsymbol{\theta}})\right)^{-1} \varphi_{\bar{a}}(\bar{\boldsymbol{\theta}}). \tag{B.91}$$

This forms a contradiction to the assumption that $\widehat{\boldsymbol{\theta}}_{t,\bar{a}}^{\text{Ridge}} \xrightarrow{p} \theta_{\bar{a}}$ and the fact that

$$\left\|\left(\Sigma_{\bar{a}}(\bar{\boldsymbol{\theta}})\right)^{-1} \varphi_{\bar{a}}(\bar{\boldsymbol{\theta}}) - \bar{\boldsymbol{\theta}}_{\bar{a}}\right\|_2 \geq c, \text{ for any choice of } \bar{\boldsymbol{\theta}} \in \mathbb{R}^d. \tag{B.92}$$

B.9 Proof of Corollary 5.2

It is straightforward to verify the Assumptions of Theorem 3.3 given Assumption 2.1, 3.1, 3.4 and 5.2. We only need to verify the form of the asymptotic variance in (5.5).

First, we have

$$\mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) = \mathbb{E}[\boldsymbol{X}_t \boldsymbol{X}_t^\top - \boldsymbol{\Sigma}_e] = \mathbb{E}\boldsymbol{S}_t \boldsymbol{S}_t^\top = \boldsymbol{\Sigma}_S. \tag{B.93}$$

Second, by plugging in $\boldsymbol{g}$ to the asymptotic variance in (3.2), we have

$$\begin{aligned}
\bar{\boldsymbol{I}}_a &= \mathbb{E}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top\right] \\
&= \mathbb{E}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \left\{(\boldsymbol{X}_t \boldsymbol{X}_t^\top - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^* - \boldsymbol{X}_t Y_t(a)\right\}^{\otimes 2}\right] \\
&= \mathbb{E}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \left\{\boldsymbol{h}_a(\boldsymbol{X}_t, \boldsymbol{S}_t) - \boldsymbol{X}_t \eta_t\right\}^{\otimes 2}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t, \boldsymbol{S}_t}\left[\mathbb{E}_{\eta_t}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \left\{\boldsymbol{h}_a(\boldsymbol{X}_t, \boldsymbol{S}_t) - \boldsymbol{X}_t \eta_t\right\}^{\otimes 2} | \boldsymbol{X}_t, \boldsymbol{S}_t\right]\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t, \boldsymbol{S}_t}\left[\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)} \left(\boldsymbol{h}_a(\boldsymbol{X}_t, \boldsymbol{S}_t) \boldsymbol{h}_a(\boldsymbol{X}_t, \boldsymbol{S}_t)^\top + \sigma_\eta^2 \boldsymbol{X}_t \boldsymbol{X}_t^\top\right)\right].
\end{aligned} \tag{B.94}$$

Combining the above with (B.93), we verify that the asymptotic variance shown in Corollary 5.2 matches that in Theorem 3.3.

B.10 Proof of Theorem A.3

We first present the following lemma, the proof is in Appendix C.8.

Lemma B.12. *Under the same conditions of Theorem A.3, as* $T \to \infty$, $\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}} \xrightarrow{p} 1$.

Returning to the main proof, we have

$$\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) \cdot \sqrt{T}(\widetilde{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_a^*) = -\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*).$$

We analyze $\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*)$ and $\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*)$ separately. First,

$$\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) = \nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \left(\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}}\right) \cdot (\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)$$

From Lemma B.2, we have $\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{p} \boldsymbol{\Sigma}_S$. From Lemma B.12, we have $\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}} \xrightarrow{p} 1$. In addition, $\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e \xrightarrow{p} 0$. Combining these facts, we obtain that $\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) \xrightarrow{p} \boldsymbol{\Sigma}_S$.

We now analyze $\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*)$. We have

$$\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) = \sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \left(\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}}\right) \cdot \sqrt{T}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*,$$

where from Lemma B.4 and (B.94) we have $\sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{I}}_a)$. Also, $\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}} \xrightarrow{p} 1$.

- If $n \gg T$, $\sqrt{T}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e) = O_P(\sqrt{T/n}) = o_P(1)$, then $\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{I}}_a)$.
- If $n \ll T$, we instead write

  $$\sqrt{n}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) = \sqrt{n}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \left(\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}}\right) \cdot \sqrt{n}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*.$$

  Notice that $\sqrt{n}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \sqrt{n/T} \cdot \sqrt{T}G_T(\boldsymbol{\theta}_a^*) = o_p(1)$, and from the Central Limit Theorem, we have

  $$\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^* \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{H}}_a),$$

  where $\bar{\boldsymbol{H}}_a := \mathbb{E}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*\boldsymbol{\theta}_a^{*,\top}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)$. Thus,

  $$\sqrt{n}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \bar{\boldsymbol{H}}_a.$$

  Combining the limit of $\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*)$, we use Slutsky's Theorem to obtain the desired result.
- If $n = \kappa T$ for some constant κ, we write

  $$\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) = \sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \left(\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}}\right) \cdot \sqrt{\frac{T}{n}} \cdot \sqrt{n}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*.$$

  Note that $\sqrt{T}G_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{I}}_a)$ and $\sqrt{n}(\widehat{\boldsymbol{\Sigma}}_e - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^* \xrightarrow{d} \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{H}}_a)$ are independent, the latter obtained from the Central Limit Theorem. Here $\bar{\boldsymbol{H}}_a := \mathbb{E}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)\boldsymbol{\theta}_a^*\boldsymbol{\theta}_a^{*,\top}(\widetilde{\boldsymbol{V}}_i - \boldsymbol{\Sigma}_e)$. Combining Lemma B.12, we obtain that

  $$\sqrt{T}\widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \bar{\boldsymbol{I}}_a + \frac{1}{\kappa}\bar{\boldsymbol{H}}_a\right).$$

  Combining the limit of $\nabla \widetilde{\boldsymbol{G}}_T(\boldsymbol{\theta}_a^*)$, we use Slutsky's Theorem to obtain the desired result.

C Proofs of Auxiliary Lemmas

C.1 Proof of Lemma A.1

We first show that $\text{Clip}(\boldsymbol{\pi})$ is the L_2 projection of $\boldsymbol{\pi}$ onto the set $\{\boldsymbol{\pi} \in [0,1]^{|\mathcal{A}|} \mid \sum_a \boldsymbol{\pi}_a = 1, \boldsymbol{\pi}_a \geq \pi_{\min}\}$.

The optimization problem is given by

$$\min_{\boldsymbol{\pi}' \in [0,1]^{|\mathcal{A}|}} \frac{1}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2^2 \quad \text{s.t.} \quad \sum_a \boldsymbol{\pi}'_a = 1, \quad \boldsymbol{\pi}'_a \geq \pi_{\min}. \tag{C.1}$$

The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{\pi}', \nu, \mu) = \frac{1}{2}\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2^2 + \nu\left(1 - \sum_a \boldsymbol{\pi}'_a\right) + \boldsymbol{\mu}^\top(\boldsymbol{\pi}' - \pi_{\min}). \tag{C.2}$$

The KKT conditions are given by

$$\nabla_{\boldsymbol{\pi}'}\mathcal{L}(\boldsymbol{\pi}', \nu, \boldsymbol{\mu}) = \boldsymbol{\pi} - \boldsymbol{\pi}' - \nu\mathbf{1} - \boldsymbol{\mu} = 0, \tag{C.3}$$

$$\nu\left(1 - \sum_a \boldsymbol{\pi}'_a\right) = 0, \tag{C.4}$$

$$\boldsymbol{\mu}\left(\boldsymbol{\pi}' - \pi_{\min}\right) = 0, \tag{C.5}$$

$$\boldsymbol{\mu} \geq 0, \quad \boldsymbol{\pi}' \geq \pi_{\min}, \quad \nu\mathbf{1} = 1. \tag{C.6}$$

From the optimality condition, we have:

$$\boldsymbol{\pi}'_a = \boldsymbol{\pi}_a + \nu + \mu_a \tag{C.7}$$

Complementary slackness indicates that for each a :

- if $\boldsymbol{\pi}'_a > \pi_{\min}$, then $\mu_a = 0$, thus:

$$\boldsymbol{\pi}'_a = \boldsymbol{\pi}_a + \nu \tag{C.8}$$

- if $\boldsymbol{\pi}'_a = \pi_{\min}$, then:

$$\pi_{\min} = \boldsymbol{\pi}_a + \nu + \mu_a, \quad \mu_a \geq 0 \Rightarrow \boldsymbol{\pi}_a + \nu \leq \pi_{\min} \tag{C.9}$$

Hence, the solution takes the form:

$$\boldsymbol{\pi}'_a = \max\left(\boldsymbol{\pi}_a + \nu, \pi_{\min}\right) \tag{C.10}$$

with the constraint that $\sum_{a=1}^{|\mathcal{A}|} \boldsymbol{\pi}'_a = 1$.

By taking the derivative of $\mathcal{L}(\boldsymbol{\pi}', \nu, \boldsymbol{\mu})$ w.r.t. ν, we have that ν is the minimum value that satisfies the KKT conditions.

C.2 Proof of Lemma B.1

We first prove that for R_θ stated in Assumption 3.2,

$$\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \|\boldsymbol{G}_T(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 \xrightarrow{p} 0 \tag{C.11}$$

as $T \to \infty$. In fact, we only need to prove

$$\sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \|\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})\|_2 \xrightarrow{p} 0 \tag{C.12}$$

for all $i = 1, \ldots, d$. Here $\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})$ denotes the i-th entry of $\boldsymbol{G}_T(\boldsymbol{\theta})$.

Fix any $\epsilon > 0$. Let $\boldsymbol{\Theta}_\epsilon = \{\boldsymbol{\theta}_j : j = 1, \ldots, N_\epsilon\}$ be an ϵ-net of the set $\boldsymbol{\Theta} := \overline{\mathcal{B}(\mathbf{0}, R_\theta)}$ with finite cardinality N_ϵ. This means that $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon$, $\exists j \in [N_\epsilon]$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \le \epsilon$. Then from Assumption 3.3, we have

$$\begin{aligned}
|\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) - \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j)| &\le \|\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j)\|_2 \\
&\le \left\| \int_0^1 \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j + u(\boldsymbol{\theta} - \boldsymbol{\theta}_j)) \mathrm{d}u \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_j) \right\|_2 \\
&\le \phi(\boldsymbol{X}_t, Y_t(a)) \cdot \|\boldsymbol{\theta} - \boldsymbol{\theta}_j\|_2 \\
&\le \epsilon \phi(\boldsymbol{X}_t, Y_t(a)).
\end{aligned}$$

Here $\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$ denotes the i-th entry of $\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$, for any $\boldsymbol{\theta}$. Thus,

$$l_j(\boldsymbol{X}_t, Y_t(a)) \le \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) \le u_j(\boldsymbol{X}_t, Y_t(a)), \tag{C.13}$$

where we define $l_j(\boldsymbol{X}_t, Y_t(a)) = \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j) - \epsilon\phi(\boldsymbol{X}_t, Y_t(a))$, $u_j(\boldsymbol{X}_t, Y_t(a)) = \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_j) + \epsilon\phi(\boldsymbol{X}_t, Y_t(a))$.

Now, notice that $\boldsymbol{G}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{Z}_t(\boldsymbol{\theta})$, where $\boldsymbol{Z}_t(\boldsymbol{\theta}) = \frac{1}{\pi_t(A_t)} 1_{\{A_t = a\}} \boldsymbol{g}(\boldsymbol{X}_t, Y_t; \boldsymbol{\theta})$. We have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}_t(\boldsymbol{\theta})|\mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}[\boldsymbol{Z}_t(\boldsymbol{\theta})|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t \sim \pi_t}\left[\frac{1}{\pi_t(A_t)} 1_{\{A_t = a\}}\middle|\mathcal{H}_{t-1}, \boldsymbol{X}_t\right] \cdot \mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right] \\
&= \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}).
\end{aligned}$$

Here in the second equation, we use Assumption 2.1. This implies that

$$\mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{Z}_t(\boldsymbol{\theta})|\mathcal{H}_{t-1}], \tag{C.14}$$

$\forall t \in [T]$. Combining (C.13), we deduce that

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})\} &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t;\boldsymbol{\theta}) \\
&\qquad - \mathbb{E}\bigg[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t;\boldsymbol{\theta})\bigg|\mathcal{H}_{t-1}\bigg]\Bigg\} \\
&\le \max_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}u_j(\boldsymbol{X}_t, Y_t(a)) \\
&\qquad - \mathbb{E}\bigg[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}l_j(\boldsymbol{X}_t, Y_t(a))\bigg|\mathcal{H}_{t-1}\bigg]\Bigg\} \\
&\le \Delta_1 + \Delta_2, \qquad \text{(C.15)}
\end{aligned}$$

where

$$\Delta_1 := \max_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}u_j(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\bigg[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}u_j(\boldsymbol{X}_t, Y_t(a))\bigg|\mathcal{H}_{t-1}\bigg]\Bigg\},$$

$$\Delta_2 := \max_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\bigg[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\big(u_j(\boldsymbol{X}_t, Y_t(a)) - l_j(\boldsymbol{X}_t, Y_t(a))\big)\bigg|\mathcal{H}_{t-1}\bigg].$$

We first analyze Δ_1. In fact we have

$$\Delta_1 \le \sum_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T}\Bigg\{\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}u_j(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\bigg[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}u_j(\boldsymbol{X}_t, Y_t(a))\bigg|\mathcal{H}_{t-1}\bigg]\Bigg\}. \quad \text{(C.16)}$$

For any $\epsilon' > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a))\middle|\mathcal{H}_{t-1}\right]\right\}\right| > \epsilon'\right) \\
\le&\frac{1}{T^2\epsilon'^2}\mathbb{E}\left(\sum_{t=1}^{T}\left\{\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a))\middle|\mathcal{H}_{t-1}\right]\right\}\right)^2 \\
=&\frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\mathbb{E}\left(\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a)) - \mathbb{E}\left[\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a))\middle|\mathcal{H}_{t-1}\right]\right)^2 \\
\le&\frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\mathbb{E}\left(\frac{1_{\{A_t=a\}}}{\pi_t(A_t)}u_j(\boldsymbol{X}_t, Y_t(a))\right)^2 \\
\le&\frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\frac{1}{\pi_{\min}^2}\mathbb{E}u_j^2(\boldsymbol{X}_t, Y_t(a)) \\
=&\frac{1}{T^2\epsilon'^2}\sum_{t=1}^{T}\frac{1}{\pi_{\min}^2}\mathbb{E}\big(\boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a);\boldsymbol{\theta}_j) + \epsilon\phi(\boldsymbol{X}_t, Y_t(a))\big)^2 \\
\le&\frac{1}{\pi_{\min}^2 T^2\epsilon'^2}\cdot T\mathbb{E}\big[2(\boldsymbol{g}^{(i)}(\boldsymbol{X}_1, Y_1(a);\boldsymbol{\theta}_j))^2 + 2\epsilon^2\phi^2(\boldsymbol{X}_1, Y_1(a))\big] \\
\le&\frac{2M_2' + \epsilon^2 M_\phi}{\pi_{\min}^2 T\epsilon'^2} \to 0. \qquad \text{(C.17)}
\end{aligned}$$

Here $M_2' := \sup_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \mathbb{E}\|\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2^2$, $M_\phi := \mathbb{E}\phi(\boldsymbol{X}_t, Y_t(a))^2$. Above, the first inequality is due to Chebyshev's inequality. The first equality is due to the following fact: Denote $u_{j,t} = \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} u_j(\boldsymbol{X}_t, Y_t(a))$. Then for $t_1 < t_2$,

$$\begin{aligned}
&\mathbb{E}(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{H}_{t_1-1}])(u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{H}_{t_2-1}]) \\
=&\mathbb{E}\big[\mathbb{E}[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{H}_{t_1-1}])(u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{H}_{t_2-1}])\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{H}_{t_1-1}]) \cdot \mathbb{E}[u_{j,t_2} - \mathbb{E}[u_{j,t_2}|\mathcal{H}_{t_2-1}]\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(u_{j,t_1} - \mathbb{E}[u_{j,t_1}|\mathcal{H}_{t_1-1}]) \cdot 0\big] = 0.
\end{aligned}$$

The second to last inequality is due to the fact that $(a+b)^2 \le 2(a^2+b^2)$ for any $a, b \in \mathbb{R}$. The final inequality uses Assumption 3.2 and Assumption 3.3, which implies $M_2' < \infty$ and $M_\phi < \infty$.

Combining (C.16) and (C.17), we obtain that $\Delta_1 \xrightarrow{p} 0$.

Next, we analyze Δ_2. We have

$$\begin{aligned}
\Delta_2 &\le \max_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_{\min}} \mathbb{E}\left[u_j(\boldsymbol{X}_t, Y_t(a)) - l_j(\boldsymbol{X}_t, Y_t(a))\middle|\mathcal{H}_{t-1}\right] \\
&\le \max_{j\in[N_\epsilon]} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\pi_{\min}} \mathbb{E}\left[2\epsilon\phi(\boldsymbol{X}_t, Y_t(a))\middle|\mathcal{H}_{t-1}\right] \\
&= \frac{2\epsilon}{\pi_{\min}} \mathbb{E}\phi(\boldsymbol{X}_1, Y_1(a)) \\
&\le \frac{2\epsilon}{\pi_{\min}} \cdot \sqrt{\mathbb{E}\phi(\boldsymbol{X}_1, Y_1(a))^2} \\
&\le \frac{2\epsilon\sqrt{M_\phi}}{\pi_{\min}}.
\end{aligned}$$

Plugging in the analysis of Δ_1 and Δ_2 into (C.15), we obtain that $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})\} > \frac{2\epsilon\sqrt{M_\phi}}{\pi_{\min}}\right) \to 0.$$

This implies that $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})\} > \epsilon\right) \to 0.$$

Similarly, we have

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\{-\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta}) + \mathbb{E}\boldsymbol{G}_T^{(i)}(\boldsymbol{\theta})\} > \epsilon\right) \to 0.$$

Therefore, we have proved (C.12). Further, (C.11) is proved.

Now we return to prove the main results. Define

$$\widehat{\boldsymbol{\theta}}_a^{(T)} = \operatorname*{argmin}_{\|\boldsymbol{\theta}\|_2 \le R_\theta} \|\boldsymbol{G}_T(\boldsymbol{\theta})\|_2,$$

and

$$\widetilde{\boldsymbol{\theta}}_a^{(T)} = \boldsymbol{\theta}_a^* - [\nabla\bar{\boldsymbol{G}}(\boldsymbol{\theta}_a^*)]^{-1}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*).$$

Here $\bar{\boldsymbol{G}}(\boldsymbol{\theta}) := \mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta}) = \mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$ due to (C.14). From Assumption 3.3, $\nabla\bar{\boldsymbol{G}}(\boldsymbol{\theta}_a^*)$ is invertible. Combining Lemma B.4, we have

$$\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^* = O_p(1/\sqrt{T}). \tag{C.18}$$

In addition, from Taylor expansion, we have

$$\begin{aligned}
\boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_a^{(T)}) &= \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + \nabla\boldsymbol{G}_T(\boldsymbol{\theta}_a^*)(\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*) + o_p(\|\widetilde{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*\|_2) \\
&= \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \nabla\boldsymbol{G}_T(\boldsymbol{\theta}_a^*)[\nabla\bar{\boldsymbol{G}}(\boldsymbol{\theta}_a^*)]^{-1}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + o_p(1/\sqrt{T}) \\
&= \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - (1 + o_p(1))\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) + o_p(1/\sqrt{T}) \\
&= o_p(1/\sqrt{T}).
\end{aligned} \tag{C.19}$$

Here the second equality is from the definition of $\widetilde{\boldsymbol{\theta}}_a^{(T)}$ and due to (C.18). The third equality is obtained from the following two facts: From Lemma B.2, $\nabla\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) \xrightarrow{p} \mathbb{E}\nabla\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$; From (C.14) and Assumption 3.3, $\nabla\bar{\boldsymbol{G}}(\boldsymbol{\theta}_a^*) = \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_a^*} = \frac{\partial}{\partial\boldsymbol{\theta}}\mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_a^*} = \mathbb{E}\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_a^*}$ nonsingular. Here the exchangeability between expectation and differentiation can be obtained by standard arguments using dominated convergence theorem, see e.g. Theorem 16.8 in [9]. The last equality is obtained from Lemma B.4.

Combining (C.18) (which implies $\mathbb{P}(\|\widetilde{\boldsymbol{\theta}}_a^{(T)}\|_2 > R_\theta) \to 0$) and (C.19), we deduce that

$$\begin{aligned}
\|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 &= \|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 1_{\{\|\widetilde{\boldsymbol{\theta}}_a^{(T)}\|_2 \le R_\theta\}} + \|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 1_{\{\|\widetilde{\boldsymbol{\theta}}_a^{(T)}\|_2 > R_\theta\}} \\
&\le \|\boldsymbol{G}_T(\widetilde{\boldsymbol{\theta}}_a^{(T)})\|_2 + o_p(1/\sqrt{T}) = o_p(1/\sqrt{T}).
\end{aligned}$$

Thus we have proved the first part of Lemma B.1.

Next, for any sequence $\widehat{\boldsymbol{\theta}}_a^{(T)}$ such that $\widehat{\boldsymbol{\theta}}_a^{(T)} \le R_\theta$ and (2.6) holds, we proceed to prove $\widehat{\boldsymbol{\theta}}_a^{(T)} \xrightarrow{p} \boldsymbol{\theta}_a^*$. According to Assumption 3.1, for any $\epsilon > 0$, we have

$$\delta_\epsilon := \inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2 > 0.$$

From (C.14), we deduce that

$$\inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 = \inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\mathbb{E}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2 \ge \delta_\epsilon.$$

Combine (C.11), we have

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 \ge \inf_{\|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{G}_T(\boldsymbol{\theta}) - \mathbb{E}\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 \ge \delta_\epsilon - o_p(1).$$

Here $\boldsymbol{\Theta} = \overline{\mathcal{B}(\mathbf{0}, R_\theta)}$. This implies

$$\lim_{T\to\infty} \mathbb{P}\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta} - \boldsymbol{\theta}_a^*\|_2 > \epsilon} \|\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 \le \frac{1}{2}\delta_\epsilon \right) = 0. \tag{C.20}$$

At the same time, from the property of $\widehat{\boldsymbol{\theta}}_a^{(T)}$, we have for any $\epsilon' > 0$,

$$\lim_{T\to\infty} \mathbb{P}\left(\|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 > \epsilon'/\sqrt{T} \right) = 0.$$

Thus,

$$\lim_{T\to\infty} \mathbb{P}\left(\|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 > \frac{1}{2}\delta_\epsilon\right) = 0. \tag{C.21}$$

We combine (C.20) and (C.21) and get

$$\begin{aligned}
\mathbb{P}\big(\|\widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*\|_2 > \epsilon\big) &= \mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*\|_2 > \epsilon, \|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 \le \frac{1}{2}\delta_\epsilon\right) \\
&\quad + \mathbb{P}\left(\|\widehat{\boldsymbol{\theta}}_a^{(T)} - \boldsymbol{\theta}_a^*\|_2 > \epsilon, \|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 > \frac{1}{2}\delta_\epsilon\right) \\
&\le \mathbb{P}\left(\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2>\epsilon} \|\boldsymbol{G}_T(\boldsymbol{\theta})\|_2 \le \frac{1}{2}\delta_\epsilon\right) + \mathbb{P}\left(\|\boldsymbol{G}_T(\widehat{\boldsymbol{\theta}}_a^{(T)})\|_2 > \frac{1}{2}\delta_\epsilon\right) \\
&\to 0
\end{aligned}$$

as $T \to \infty$. As the above holds for any $\epsilon > 0$, the consistency of $\widehat{\boldsymbol{\theta}}_a^{(T)}$ is proved.

C.3 Proof of Lemma B.2

Recall that $\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{V}_t$, where
$\boldsymbol{V}_t := \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$. We have for any nonrandom vectors $\boldsymbol{c}, \boldsymbol{c}' \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}[c^\top \boldsymbol{V}_t c' | \mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t, Y_t(a)}[\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' | \mathcal{H}_{t-1}, \boldsymbol{X}_t] \middle| \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t}\left[\frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \middle| \mathcal{H}_{t-1}, \boldsymbol{X}_t\right] \cdot \right. \\
&\qquad \left. \mathbb{E}_{Y_t(a)}[\boldsymbol{c}^\top \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \boldsymbol{c}' | \mathcal{H}_{t-1}, \boldsymbol{X}_t] \middle| \mathcal{H}_{t-1}\right] \\
&= \mathbb{E}_{\boldsymbol{X}_t}\left[1 \cdot \mathbb{E}_{Y_t(a)}[\boldsymbol{c}^\top \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \boldsymbol{c}' | \mathcal{H}_{t-1}, \boldsymbol{X}_t] \middle| \mathcal{H}_{t-1}\right] \\
&= \boldsymbol{c}^\top \mathbb{E}[\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)] \boldsymbol{c}'.
\end{aligned}$$

Here the second inequality uses Assumption 2.1. From the above we deduce that $\forall \delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(|\boldsymbol{c}^\top [\nabla \boldsymbol{G}_T(\boldsymbol{\theta}_a^*) - \mathbb{E}\nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)] \boldsymbol{c}'| > \delta\right) \\
=&\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T [\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' | \mathcal{H}_{t-1}]]\right| > \delta\right) \\
\le&\frac{1}{\delta^2 T^2}\mathbb{E}\left(\sum_{t=1}^T [\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' | \mathcal{H}_{t-1}]]\right)^2 \\
=&\frac{1}{\delta^2 T^2}\sum_{t=1}^T \mathbb{E}\left(\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' - \mathbb{E}[\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}' | \mathcal{H}_{t-1}]\right)^2 \\
\le&\frac{1}{\delta^2 T^2}\sum_{t=1}^T \mathbb{E}\left(\boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}'\right)^2 \\
\le&\frac{1}{\delta^2 T^2}\sum_{t=1}^T \frac{1}{\pi_{\min}^2}\mathbb{E}\left(\boldsymbol{c}^\top \nabla \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*) \boldsymbol{c}'\right)^2 \to 0.
\end{aligned} \tag{C.22}$$

Here the first inequality is because of Chebyshev's inequality. The second equality is due to the following fact: Let $v_t = \boldsymbol{c}^\top \boldsymbol{V}_t \boldsymbol{c}'$. Then for $t_1 < t_2$,

$$\begin{aligned}
&\mathbb{E}(v_{t_1} - \mathbb{E}[v_{t_1}|\mathcal{H}_{t_1-1}])(v_{t_2} - \mathbb{E}[v_{t_2}|\mathcal{H}_{t_2-1}]) \\
=&\mathbb{E}\big[\mathbb{E}[(v_{t_1} - \mathbb{E}[v_{t_1}|\mathcal{H}_{t_1-1}])(v_{t_2} - \mathbb{E}[v_{t_2}|\mathcal{H}_{t_2-1}])\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(v_{t_1} - \mathbb{E}[v_{t_1}|\mathcal{H}_{t_1-1}]) \cdot \mathbb{E}[v_{t_2} - \mathbb{E}[v_{t_2}|\mathcal{H}_{t_2-1}]\big|\mathcal{H}_{t_1-1}]\big] \\
=&\mathbb{E}\big[(v_{t_1} - \mathbb{E}[v_{t_1}|\mathcal{H}_{t_1-1}]) \cdot 0\big] = 0.
\end{aligned}$$

The last convergence uses Assumption 3.3.

Finally, because (C.22) holds for any $\boldsymbol{c}, \boldsymbol{c}'$, we conclude our proof.

C.4 Proof of Lemma B.3

Note that $\nabla^2 \boldsymbol{G}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^T \frac{1_{\{A_t=a\}}}{\pi_t(A_t)} \nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})$. According to Assumption 3.3, $\forall \boldsymbol{\theta} \in \overline{\mathcal{B}(\boldsymbol{\theta}_a^*, \epsilon_0)}$,

$$\begin{aligned}
\|\nabla^2 \boldsymbol{G}_T(\boldsymbol{\theta})\|_1 &\le \frac{1}{\pi_{\min}} \cdot \frac{1}{T}\sum_{t=1}^T \|\nabla^2 \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_1 \\
&\le \frac{1}{\pi_{\min} T}\sum_{t=1}^T d^2 \sup_{i\in[d]} \|\nabla^2 \boldsymbol{g}^{(i)}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta})\|_2 \\
&\le \frac{d^2}{\pi_{\min} T}\sum_{t=1}^T \Phi(\boldsymbol{X}_t, Y_t(a)).
\end{aligned}$$

Thus,

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_a^*\|_2 \le \epsilon_0} \|\nabla^2 \boldsymbol{G}_T(\boldsymbol{\theta})\|_1 \le \frac{d^2}{\pi_{\min}} \cdot \frac{1}{T}\sum_{t=1}^T \Phi(\boldsymbol{X}_t, Y_t(a)) = \mathcal{O}_p(1)$$

as $T \to \infty$.

C.5 Proof of Lemma B.4

We have $\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \frac{1}{T}\sum_{t=1}^T \boldsymbol{Z}_t$, where we define $\boldsymbol{Z}_t := \frac{1}{\pi_t(A_t)} 1_{\{A_t=a\}} \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)$. From the Cramer-Wold theorem, in order to show the desired asymptotic normality, it suffices to show that for any $\boldsymbol{c} \in \mathbb{R}^d$, $\boldsymbol{c}^\top \cdot \sqrt{T}\boldsymbol{G}_T(\boldsymbol{\theta}_a^*) = \boldsymbol{c}^\top \cdot \frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{Z}_t \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{c}^\top \bar{\boldsymbol{I}}_a \boldsymbol{c})$.

From [23], Theorem 2.2, the above asymptotic result can be obtained by ensuring

$$\mathbb{E}[\boldsymbol{Z}_t|\mathcal{H}_{t-1}] = \boldsymbol{0} \quad \forall t \in [T], \tag{C.23}$$

$$\frac{1}{T}\sum_{t\in[T]} \mathrm{Var}(\boldsymbol{c}^\top \boldsymbol{Z}_t|\mathcal{H}_{t-1}) \xrightarrow{p} \boldsymbol{c}^\top \bar{\boldsymbol{I}}_a \boldsymbol{c}, \tag{C.24}$$

$$\frac{1}{T}\sum_{t\in[T]} \mathbb{E}\left[(\boldsymbol{c}^\top \boldsymbol{Z}_t)^2 1_{\{|\boldsymbol{c}^\top \boldsymbol{Z}_t| > \sqrt{T}\delta\}} \middle| \mathcal{H}_{t-1}\right] \xrightarrow{p} 0 \quad \forall \delta > 0. \tag{C.25}$$

Below we check these facts one by one.

Check (C.23): We have

$$\begin{aligned}\mathbb{E}[\boldsymbol{Z}_t|\mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t, Y_t(a)}[\boldsymbol{Z}_t|\mathcal{H}_{t-1}, \boldsymbol{X}_t]|\mathcal{H}_{t-1}\right] \\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t}\left[\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\middle|\mathcal{H}_{t-1}, \boldsymbol{X}_t\right]\right. \\ &\qquad\left.\cdot\, \mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right] \\ &= \mathbb{E}_{\boldsymbol{X}_t}\left[1\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right] \\ &= \mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)] = \mathbf{0}.\end{aligned}$$

Here, the second equality is because of Assumption 2.1.

Check (C.24): Based on (C.23),

$$\begin{aligned}\frac{1}{T}\sum_{t\in[T]}\mathrm{Var}(\boldsymbol{c}^\top\boldsymbol{Z}_t|\mathcal{H}_{t-1}) &= \frac{1}{T}\sum_{t\in[T]}\mathbb{E}[\boldsymbol{c}^\top\boldsymbol{Z}_t\boldsymbol{Z}_t^\top\boldsymbol{c}|\mathcal{H}_{t-1}] \\ &= \frac{1}{T}\sum_{t\in[T]}\boldsymbol{c}^\top\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top|\mathcal{H}_{t-1}]\boldsymbol{c} \qquad \text{(C.26)}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{c}^\top\mathbb{E}[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top|\mathcal{H}_{t-1}]\boldsymbol{c} &= \boldsymbol{c}^\top\mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t, Y_t(a)}[\boldsymbol{Z}_t\boldsymbol{Z}_t^\top|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right]\boldsymbol{c} \\ &= \boldsymbol{c}^\top\mathbb{E}_{\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t}\left[\frac{1}{\pi_t^2(A_t)}1_{\{A_t=a\}}\middle|\mathcal{H}_{t-1}, \boldsymbol{X}_t\right]\cdot\right. \\ &\qquad\left.\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top|\mathcal{H}_{t-1}, \boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right]\boldsymbol{c} \\ &= \boldsymbol{c}^\top\mathbb{E}_{\boldsymbol{X}_t}\left[\frac{1}{\pi_t(a)}\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top|\boldsymbol{X}_t]\middle|\mathcal{H}_{t-1}\right]\boldsymbol{c} \\ &= \boldsymbol{c}^\top\mathbb{E}_{\boldsymbol{X}_t}[\boldsymbol{I}_{a,t}|\mathcal{H}_{t-1}]\boldsymbol{c}. \qquad \text{(C.27)}\end{aligned}$$

Here we define $\boldsymbol{I}_{a,t} := \frac{1}{\pi_t(a)}\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top|\boldsymbol{X}_t]$. We also define $\bar{\boldsymbol{I}}_{a,t} := \frac{1}{\bar{\pi}(a)}\cdot\mathbb{E}_{Y_t(a)}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top|\boldsymbol{X}_t]$. Then

$$\begin{aligned}|\boldsymbol{c}^\top\boldsymbol{I}_{a,t}\boldsymbol{c} - \boldsymbol{c}^\top\bar{\boldsymbol{I}}_{a,t}\boldsymbol{c}| &= |\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)|\cdot\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)\pi_t(a)}\boldsymbol{c}^\top\mathbb{E}[\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}_a^*)^\top|\boldsymbol{X}_t]\boldsymbol{c} \\ &\le M_2|\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)|,\end{aligned}$$

where we have used Assumption 3.2.

Note that the random variables $\{\pi_t(a) - \bar{\pi}(a|\boldsymbol{X}_t)\}_{t\ge1}$ are uniformly integrable. Thus, $\pi_t(a) - \bar{\pi}(a|\boldsymbol{X}_t) \xrightarrow{p} 0$ implies

$$\lim_{t\to\infty}\mathbb{E}|\bar{\pi}(a|\boldsymbol{X}_t) - \pi_t(a)| = 0.$$

Combining the above facts, if we denote $U_{t,\boldsymbol{c}} := \boldsymbol{c}^\top\boldsymbol{I}_{a,t}\boldsymbol{c} - \boldsymbol{c}^\top\bar{\boldsymbol{I}}_{a,t}\boldsymbol{c}$, then

$$\lim_{t\to\infty}\mathbb{E}|U_{t,\boldsymbol{c}}| = 0.$$

Also, noticing that

$$\mathbb{E}\big|\mathbb{E}[U_{t,\boldsymbol{c}}|\mathcal{H}_{t-1}]\big| \le \mathbb{E}\mathbb{E}\big[|U_{t,\boldsymbol{c}}|\big|\mathcal{H}_{t-1}\big] = \mathbb{E}|U_{t,\boldsymbol{c}}|,$$

We deduce that

$$\lim_{t\to\infty} \mathbb{E}\big|\mathbb{E}[U_{t,\boldsymbol{c}}|\mathcal{H}_{t-1}]\big| = 0.$$

Note that the following property about L_1 convergence is true: For a sequence of random variables U'_t, if $U'_t \xrightarrow{L_1} 0$, then its running average sequence $\bar{U}'_t = \frac{1}{t}\sum_{\tau\in[t]} U'_\tau$ satisfies $\bar{U}'_t \xrightarrow{L_1} 0$. Let $U'_t = U_{t,\boldsymbol{c}}$, then we have

$$\frac{1}{t}\sum_{\tau\le t} \mathbb{E}[U_{\tau,\boldsymbol{c}}|\mathcal{H}_{\tau-1}] \xrightarrow{L_1} 0.$$

Plugging in the expression of $U_{\tau,\boldsymbol{c}}$, we have

$$\frac{1}{T}\sum_{t\le T} \mathbb{E}\big[\boldsymbol{c}^\top \boldsymbol{I}_{a,t}\boldsymbol{c}|\mathcal{H}_{t-1}\big] - \frac{1}{T}\sum_{t\le T} \mathbb{E}\big[\boldsymbol{c}^\top \bar{\boldsymbol{I}}_{a,t}\boldsymbol{c}|\mathcal{H}_{t-1}\big] \xrightarrow{L_1} 0, \tag{C.28}$$

and

$$\frac{1}{T}\sum_{t\le T} \mathbb{E}\big[\boldsymbol{c}^\top \bar{\boldsymbol{I}}_{a,t}\boldsymbol{c}|\mathcal{H}_{t-1}\big] = \frac{1}{T}\sum_{t\le T} \boldsymbol{c}^\top \mathbb{E}\big[\bar{\boldsymbol{I}}_{a,t}\big]\boldsymbol{c} = \boldsymbol{c}^\top \bar{\boldsymbol{I}}_a \boldsymbol{c}, \tag{C.29}$$

where

$$\bar{\boldsymbol{I}}_a := \mathbb{E}\bar{\boldsymbol{I}}_{a,t} = \mathbb{E}\frac{1}{\bar{\pi}(a|\boldsymbol{X}_t)}\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*_a)\boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*_a)^\top. \tag{C.30}$$

Combining (C.26), (C.27), (C.28) and (C.29), we obtain that

$$\frac{1}{T}\sum_{t\in[T]} \mathrm{Var}(\boldsymbol{c}^\top \boldsymbol{Z}_t|\mathcal{H}_{t-1}) \xrightarrow{L_1} \boldsymbol{c}^\top \bar{\boldsymbol{I}}_a \boldsymbol{c}. \tag{C.31}$$

Because L_1 convergence implies convergence in probability, we have verified (C.24).

Check (C.25): Using Chebyshev's inequality,

$$\begin{aligned}
&\frac{1}{T}\sum_{t\in[T]} \mathbb{E}\left[(\boldsymbol{c}^\top \boldsymbol{Z}_t)^2 1_{\{|\boldsymbol{c}^\top \boldsymbol{Z}_t| > \sqrt{T}\delta\}}\Big|\mathcal{H}_{t-1}\right] \\
\le &\frac{1}{T}\cdot\frac{1}{T\delta^2}\sum_{t\in[T]} \mathbb{E}\left[(\boldsymbol{c}^\top \boldsymbol{Z}_t)^4\Big|\mathcal{H}_{t-1}\right] \\
= &\frac{1}{T^2\delta^2}\sum_{t\in[T]} \mathbb{E}\left[\Big(\frac{1}{\pi_t(A_t)}1_{\{A_t=a\}}\boldsymbol{c}^\top \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*_a)\Big)^4\Big|\mathcal{H}_{t-1}\right] \\
\le &\frac{1}{T^2\delta^2}\sum_{t\in[T]} \frac{1}{\pi^4_{\min}}\mathbb{E}[\boldsymbol{c}^\top \boldsymbol{g}(\boldsymbol{X}_t, Y_t(a); \boldsymbol{\theta}^*_a)]^4 \\
= &\frac{1}{T\delta^2}\cdot\frac{1}{\pi^4_{\min}}\mathbb{E}[\boldsymbol{c}^\top \boldsymbol{g}(\boldsymbol{X}_1, Y_1(a); \boldsymbol{\theta}^*_a)]^4 \to 0
\end{aligned}$$

Here we have used Assumptions 3.2 and 3.4.

C.6 Proof of Lemma B.6

By Chebyshev's inequality, $\forall \delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|N_{i,t} - \sum_{i=1}^{t} \mathbb{E}[1_{\{A_\tau = i\}} | \mathcal{H}_{\tau-1}^0]\right| \geq \delta\right) \\
\leq &\frac{1}{\delta^2} \mathbb{E}\left(\sum_{\tau=1}^{t} \left(1_{\{A_\tau = i\}} - \mathbb{E}[1_{\{A_\tau = i\}} | \mathcal{H}_{\tau-1}^0]\right)\right)^2 \\
= &\frac{1}{\delta^2} \sum_{\tau=1}^{t} \mathbb{E}\left(1_{\{A_\tau = i\}} - \mathbb{E}[1_{\{A_\tau = i\}} | \mathcal{H}_{\tau-1}^0]\right)^2 \leq \frac{t}{\delta^2}.
\end{aligned} \tag{C.32}$$

In addition, given the minimum sampling probability $\pi_{\min}$, we have

$$\sum_{\tau=1}^{t} \mathbb{E}[1_{\{A_\tau = i\}} | \mathcal{H}_{\tau-1}^0] \geq \pi_{\min} t. \tag{C.33}$$

Thus, by setting $\delta = \pi_{\min} t/2$ in (C.32), we combine with (C.33) and deduce that

$$\mathbb{P}\left(N_{i,t} \leq \frac{\pi_{\min} t}{2}\right) \leq \frac{4}{\pi_{\min}^2 t}. \tag{C.34}$$

This implies that

$$\mathbb{P}\left(\frac{C_t}{N_{i,t}} \geq \frac{2C_t}{\pi_{\min} t}\right) \leq \frac{4}{\pi_{\min}^2 t},$$

which proves statement (i).

At the same time, $\forall \delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(N_{i,t}\left|\widehat{\mu}_{i,t} - \mu_i^*\right| \geq \delta\right) \\
= &\mathbb{P}\left(\left|\sum_{i=1}^{t} 1_{\{A_\tau = i\}} Y_\tau - \sum_{i=1}^{t} \mathbb{E}[1_{\{A_\tau = i\}} Y_\tau | \mathcal{H}_{\tau-1}^0, A_\tau]\right| \geq \delta\right) \\
\leq &\frac{1}{\delta^2} \mathbb{E}\left(\sum_{i=1}^{t} \left(1_{\{A_\tau = i\}} Y_\tau - \mathbb{E}[1_{\{A_\tau = i\}} Y_\tau | \mathcal{H}_{\tau-1}^0, A_\tau]\right)\right)^2 \\
= &\frac{1}{\delta^2} \sum_{\tau=1}^{t} \mathbb{E} 1_{\{A_\tau = i\}} \left(Y_\tau(i) - \mu_i^*\right)^2 \leq \frac{\sigma_Y^2 t}{\delta^2}.
\end{aligned} \tag{C.35}$$

Here we have used the definition of $N_{i,t}$ and $\widehat{\mu}_{i,t}$, as well as the fact that due to the unconfoundedness assumption,

$$\mathbb{E}[1_{\{A_\tau = i\}} Y_\tau | \mathcal{H}_{\tau-1}^0, A_\tau] = 1_{\{A_\tau = i\}} \mu_i^*.$$

Combining (C.34) and (C.35), we obtain that $\forall \delta > 0$,

$$\mathbb{P}\left(|\widehat{\mu}_{i,t} - \mu_i^*| \geq \frac{2\delta}{\pi_{\min} t}\right) \leq \frac{\sigma_Y^2 t}{\delta^2} + \frac{4}{\pi_{\min}^2 t}.$$

Let $\delta' = \frac{2\delta}{\pi_{\min} t}$, and we obtain that

$$\mathbb{P}\left(|\widehat{\mu}_{i,t} - \mu_i^*| \geq \delta'\right) \leq \frac{4\sigma_Y^2}{\delta'^2 t} + \frac{4}{\pi_{\min}^2 t} \to 0$$

as $t \to \infty$. Thus, statement (ii) is proved.

C.7 Proof of Lemma B.9

Proof. We show the Lipschitz continuity of $\mathrm{Clip}(\boldsymbol{\pi})$. Altough q function is not smooth, it is continuous and piecewise differentiable in both ν and $\boldsymbol{\pi}$, so we may analyze the slope of each piece.

We first note that for any ν, the function $q(\nu;\cdot)$ is 1-Lipschitz continuous in $\boldsymbol{\pi}$. To see this, for any two vectors $\boldsymbol{\pi}, \boldsymbol{\pi}' \in [0,1]^{|\mathcal{A}|}$, we have

$$|q(\nu;\boldsymbol{\pi}) - q(\nu;\boldsymbol{\pi}')| = \left|\sum_a \max\{\boldsymbol{\pi}_a - \nu, \pi_{\min}\} - \sum_a \max\{\boldsymbol{\pi}'_a - \nu, \pi_{\min}\}\right| \le \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1. \tag{C.36}$$

Since $q(\nu^*(\boldsymbol{\pi});\boldsymbol{\pi}) = 1 = q(\nu^*(\boldsymbol{\pi}');\boldsymbol{\pi}')$, we have that

$$|q(\nu(\boldsymbol{\pi});\boldsymbol{\pi}) - q(\nu(\boldsymbol{\pi});\boldsymbol{\pi}')| = |q(\nu(\boldsymbol{\pi}');\boldsymbol{\pi}') - q(\nu(\boldsymbol{\pi});\boldsymbol{\pi}')| \le \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1. \tag{C.37}$$

Since $q(\nu;\boldsymbol{\pi}')$ is piecewise linear and decreasing in ν, we can lower bound its slope. In particular, over intervals where some entries of $\boldsymbol{\pi}'_a - \nu > \pi_{\min}$, the derivative of q w.r.t. ν is:

$$\frac{\partial q(\nu;\boldsymbol{\pi}')}{\partial \nu} = -\left|\left\{a : \boldsymbol{\pi}'_a - \nu > \pi_{\min}\right\}\right| \le -1. \tag{C.38}$$

Because at least one action must be unclipped, so the slope is at most -1. Thus:

$$|q(\nu(\boldsymbol{\pi});\boldsymbol{\pi}') - q(\nu(\boldsymbol{\pi}');\boldsymbol{\pi}')| \ge |\nu(\boldsymbol{\pi}) - \nu(\boldsymbol{\pi}')| \cdot 1. \tag{C.39}$$

Therefore, we have that

$$|\nu(\boldsymbol{\pi}) - \nu(\boldsymbol{\pi}')| \le \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_1 \le |\mathcal{A}|\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2. \tag{C.40}$$

The second inequality follows from the Cauchy-Schwarz inequality.

To complete the proof, we see that

$$\|\mathrm{Clip}(\boldsymbol{\pi}) - \mathrm{Clip}(\boldsymbol{\pi}')\|_2 \le \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 + |\nu^*(\boldsymbol{\pi}) - \nu^*(\boldsymbol{\pi}')| \le \|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 + |\mathcal{A}|\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 = (|\mathcal{A}| + 1)\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2. \tag{C.41}$$

□

C.8 Proof of Lemma B.12

Note that

$$\begin{aligned}\mathbb{E}[W_t 1_{\{A_t=a\}} - 1|\mathcal{H}_{t-1}] &= \mathbb{E}_{\boldsymbol{S}_t,\boldsymbol{X}_t}\left[\mathbb{E}_{A_t\sim\pi_t}\left[\frac{1}{\pi_t(a)}1_{\{A_t=a\}}\middle|\mathcal{H}_{t-1},\boldsymbol{S}_t,\boldsymbol{X}_t\right]\middle|\mathcal{H}_{t-1}\right] - 1 \\ &= 1 - 1 = 0.\end{aligned}$$

Thus, for any constant $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} W_t 1_{\{A_t=a\}} - 1\right| > \delta\right) \\
\leq &\frac{1}{\delta^2 T^2}\mathbb{E}\left[\sum_{t=1}^{T}(W_t 1_{\{A_t=a\}} - 1)\right]^2 \\
= &\frac{1}{\delta^2 T^2}\mathbb{E}\left[\sum_{t=1}^{T}(W_t 1_{\{A_t=a\}} - \mathbb{E}[W_t 1_{\{A_t=a\}}|\mathcal{H}_{t-1}])\right]^2 \\
= &\frac{1}{\delta^2 T^2}\sum_{t=1}^{T}\mathbb{E}\left(W_t 1_{\{A_t=a\}} - \mathbb{E}[W_t 1_{\{A_t=a\}}|\mathcal{H}_{t-1}]\right)^2 \\
\leq &\frac{1}{\delta^2 T^2}\mathbb{E}W_t^2 \leq \frac{1}{\delta^2 T^2}\cdot\frac{T}{\pi_{\min}^2} \to 0.
\end{aligned}$$

Here the first inequality is due to Chebyshev's Inequality, the second equality is because the cross terms has zero expectation after expansion due to martingale properties. The lemma follows.

D Additional Technical Lemmas

Lemma D.1 (Matrix Azuma [60])**.** *Consider a finite adapted sequence $\{X_k\}_{k=1}^t$ of self-adjoint $d \times d$ matrices with respect to the filtration $\{\mathcal{F}_k\}_{k=1}^t$, and a fixed sequence $\{A_k\}_{k=1}^t$ of self-adjoint matrices that satisfy*

$$\mathbb{E}[A_k \mid \mathcal{F}_{k-1}] = 0, \quad \textit{and} \quad X_k^2 \preceq A_k^2 \ \textit{a.s.} \tag{D.1}$$

Compute the variance parameter

$$\sigma_t^2 = \left\|\sum_{i=1}^{t} A_i^2\right\|_2. \tag{D.2}$$

Then, for all $\epsilon \geq 0$,

$$\mathbb{P}\left(\left\|\sum_{k=1}^{t} X_k\right\|_2 \geq \epsilon\right) \leq d\exp\left(-\frac{\epsilon^2}{8\sigma_t^2}\right). \tag{D.3}$$

Applying the union bound on all $t' \in [t]$, we have the following corollary.

Corollary D.1. Under the same conditions as in Lemma D.1, we have that with a probability at least $1 - \delta/(t-1)$, for all $\tau \in t, t+1, \ldots,$

$$\frac{1}{\tau}\left\|\sum_{k=1}^{\tau} X_k\right\|_2 \leq \sqrt{\frac{16\sigma_\tau^2}{\tau^2}\log\left(\frac{\tau d}{\delta}\right)}. \tag{D.4}$$

Proof. The proof is to apply the union bound on all $\tau \in t, t+1, \ldots$ with each event having a probability of at least $1 - \delta/\tau^2$. The total failure probability is at most $\sum_{\tau=t}^{\infty}\frac{1}{\tau^2} \leq \frac{\delta}{t-1}$. □

Lemma D.2 (Law of large numbers for martingale difference sequence [18]). *Let* $Y_n = \sum_{t=1}^{n} \boldsymbol{X}_t$ *be a martingale difference sequence, such that*

$$\sum_{t=1}^{\infty} \mathbb{E}\left[|\boldsymbol{X}_t|^{2\alpha}\right] / k^{1+\alpha} < \infty. \tag{D.5}$$

Then,

$$\frac{1}{n}\sum_{t=1}^{n} Y_t \xrightarrow{a.s.} 0. \tag{D.6}$$

E Additional Details on Simulation Studies

E.1 Environment Settings

We give the details of the simulation environment settings.

The first type of environments is the noisy contextual linear bandit environment including `NC-Hard1`, `NC-Hard2`, `NC-Gaussian`. Each environment has a ground-truth parameter $\boldsymbol{\theta}_a^*$. At each time t, the following variables are generated:

$$\begin{aligned} \text{True context: } & \boldsymbol{S}_t \sim \mathcal{D}_S, \\ \text{Predicted context: } & f(\boldsymbol{X}_t) = \boldsymbol{S}_t + \boldsymbol{\epsilon}_t, \text{ where } \boldsymbol{\epsilon}_t \sim \mathcal{D}_\epsilon(\cdot \mid \boldsymbol{S}_t),. \\ \text{Reward: } & Y_t = \langle \boldsymbol{\theta}_{A_t}^*, \boldsymbol{S}_t \rangle + \eta_t, \text{ where } \eta_t \sim \mathcal{D}_\eta, \end{aligned} \tag{E.1}$$

where A_t is the algorithm-chosen action. Recall that $\Sigma_S = \mathbb{E}[\boldsymbol{S}_t \boldsymbol{S}_t^\top]$, Σ_e is the covariance matrix of $\boldsymbol{\epsilon}_t$ (assumed to be independent of $\boldsymbol{S}_t$), and Σ_η is the covariance matrix of η_t.

Both hard environments have one-dimensional context and true parameters ($d = 1$), two actions $|\mathcal{A}| = 2$, and two contexts $\mathcal{S} = \{0, -1\}$. The context distribution $\mathcal{D}_S$ is uniform over $\mathcal{S}$. `hard-1` and `hard-2` have the true parameters $\boldsymbol{\theta}_0^* = (3, 1)$ and $\boldsymbol{\theta}_1^* = (-3, -1)$ respectively. They further share the same prediction error distribution $\mathcal{D}_\epsilon(\cdot \mid \boldsymbol{S}_t)$, given by (E.2). We set the reward noise $\mathcal{D}_\eta = \mathcal{N}(0, \sigma_\eta^2)$.

$$\begin{aligned} &\mathbb{P}(f(\boldsymbol{X}_t) = 1 \mid \boldsymbol{S}_t = 0) = 2/3 \qquad &&\mathbb{P}(f(\boldsymbol{X}_t) = -2 \mid \boldsymbol{S}_t = 0) = 1/3 \\ &\mathbb{P}(f(\boldsymbol{X}_t) = -2 \mid \boldsymbol{S}_t = -1) = 2/3 \qquad &&\mathbb{P}(f(\boldsymbol{X}_t) = 1 \mid \boldsymbol{S}_t = -1) = 1/3. \end{aligned} \tag{E.2}$$

For the `NC-Gaussian` environment, we randomly sample the true parameters $\boldsymbol{\theta}_a^*$ from $\mathcal{N}(0, \Sigma_\theta)$ for each $a \in \mathcal{A}$ independently. We choose $\mathcal{D}_S = \mathcal{N}(0, \Sigma_S)$, $\mathcal{D}_\epsilon(\cdot \mid \boldsymbol{S}_t) = \mathcal{N}(0, \Sigma_e)$, $\mathcal{D}_\eta = \mathcal{N}(0, \sigma_\eta^2)$, respectively.

The second type of environments is the misspecified contextual linear bandit environment including `MC-Polynomial`, `MC-Neural`. In these two environments, we have $|\mathcal{A}| = 2$, $d = 1$. The context is sampled from $\mathcal{D}_X = \mathcal{N}(0, \Sigma_X)$. In the `MC-Polynomial` environment, the true reward function is given by

$$y(\boldsymbol{x}, a) = \langle \boldsymbol{\theta}_{a,1}^*, \boldsymbol{x} \rangle + \langle \boldsymbol{\theta}_{a,2}^*, \boldsymbol{x}^2 \rangle + \cdots + \langle \boldsymbol{\theta}_{a,d}^*, \boldsymbol{x}^d \rangle,$$

where d is the degree of the polynomial. The true parameters $\boldsymbol{\theta}_{a,i}^*$ are randomly sampled from $\mathcal{N}(0, \Sigma_\theta)$ for each $a \in \mathcal{A}$ and $i \in \{1, 2, \ldots, d\}$ independently. In the `MC-Neural` environment, the true reward function is given by a two layer neural network with one hidden layer of size d.

$$y(\boldsymbol{x}, a) = \mathrm{ReLU}(\langle \boldsymbol{\theta}_a^*, \boldsymbol{x} \rangle),$$

where $\mathrm{ReLU}(x) = \max(0, x)$ is the ReLU activation function.

The true parameters $\boldsymbol{\theta}_a^*$ are randomly sampled from $\mathcal{N}(0, \Sigma_\theta)$ for each $a \in \mathcal{A}$ independently. We choose $\mathcal{D}_X = \mathcal{N}(0, \Sigma_X)$, $\mathcal{D}_\eta = \mathcal{N}(0, \sigma_\eta^2)$, respectively.

E.2 Additional Information on OPE

In the OPE setting, we compare the proposed inference method with the CADR (Contextual Adaptive Doubly Robust) method (Equation **??**) [8] under various choice of prediction model including linear model, tree-based model, and a dumpy model that always outputs 0. We run CADR on the same dataset collected by Boltzmann exploration w.r.t. Ridge regression in five environments introduced above.

To implement the CADR method, we define the following functions:

$$\Psi\left(g, Q_Y\right) := \mathbb{E}_{A_t \sim g(A_t \mid \boldsymbol{X}_t)}[Q_Y(A_t, \boldsymbol{X}_t)]. \tag{E.3}$$

$$D'(g, \bar{Q})(x, a, y) := \frac{g^*(a \mid x)}{g(a \mid x)}(y - \bar{Q}(a, x)) + \int \bar{Q}\left(a', x\right) g^*\left(a' \mid x\right) d\mu_{\mathcal{A}}\left(a'\right). \tag{E.4}$$

$$D(g, \bar{Q})(x, a, y) = D'(g, \bar{Q})(x, a, y) - \Psi\left(g, \bar{Q}\right). \tag{E.5}$$

Let $g_1, \ldots, g_T$ be the logging policy that collects the data, and g^* be the target policy that we aim to evaluate.

For each step $t = 1, \ldots, T$, the CADR method computes the following quantities:

- Train $\widehat{Q}_{t-1} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ on the dataset $((\boldsymbol{X}_s, A_s, Y_s))_{s=1}^{t-1}$ using the outcome regression estimator.
- Set $D'_{t,s} = D(g_s, \widehat{Q}_{t-1})(\boldsymbol{X}_t, A_t, Y_t)$ for each $s = t, \ldots, T$.
- Set

$$\widehat{\sigma}_t^2 = \frac{1}{t-1} \sum_{s=1}^{t-1} \frac{g_t(A_s \mid \boldsymbol{X}_s)}{g_s(A_s \mid \boldsymbol{X}_s)} (D'_{t,s})^2 - \left(\frac{1}{t-1} \sum_{s=1}^{t-1} \frac{g_t(A_s \mid \boldsymbol{X}_s)}{g_s(A_s \mid \boldsymbol{X}_s)} D'_{t,s} \right)^2. \tag{E.6}$$

In the end, the CADR method outputs the following estimate:

$$\widehat{\Psi}_T = \frac{\Gamma_T}{T} \sum_{t=1}^{T} \widehat{\sigma}_t^{-1} D'_{t,t}, \text{ where } \Gamma_T = \left(\frac{1}{T} \sum_{t=1}^{T} \widehat{\sigma}_t^{-1} \right)^{-1}. \tag{E.7}$$

and confidence interval

$$\mathrm{CI}_\alpha = [\widehat{\Psi}_T \pm \xi_{1-\alpha/2} \Gamma_T / \sqrt{T}]. \tag{E.8}$$

[Acknowledgments] The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

This work is partly supported by NSF Grant DMS-2515285.

Title of Supplement A Short description of Supplement A. Title of Supplement B Short description of Supplement B.

References

[1] [author] Abadie, AlbertoA., Agarwal, AnishA., Dwivedi, RaazR. Shah, AbhinA. (2024). Doubly robust inference in causal latent factor models. arXiv preprint arXiv:2402.11652.

[2] [author] Abbasi-Yadkori, YasinY., Pál, DávidD. Szepesvári, CsabaC. (2011). Improved algorithms for linear stochastic bandits. Advances in neural information processing systems 24.

[3] [author] Agrawal, ShipraS. Goyal, NavinN. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In Conference on learning theory 39–1. JMLR Workshop and Conference Proceedings.

[4] [author] Anderson, TWT. Taylor, John BJ. B. (1979). Strong consistency of least squares estimates in dynamic models. The annals of Statistics 7 484–489.

[5] [author] Athey, SusanS., Byambadalai, UndralU., Hadad, VitorV., Krishnamurthy, Sanath KumarS. K., Leung, WeiwenW. Williams, Joseph JayJ. J. (2022). Contextual bandits in a survey experiment on charitable giving: Within-experiment outcomes versus policy learning. arXiv preprint arXiv:2211.12004.

[6] [author] Auer, PP. (2002). Finite-time Analysis of the Multiarmed Bandit Problem.

[7] [author] Bang, HeejungH. Robins, James MJ. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics 61 962–973.

[8] [author] Bibaut, AurélienA., Dimakopoulou, MariaM., Kallus, NathanN., Chambaz, AntoineA. van Der Laan, MarkM. (2021). Post-contextual-bandit inference. Advances in neural information processing systems 34 28548–28559.

[9] [author] Billingsley, PatrickP. (1995). Probability and Measure, 3rd ed. Wiley, New York.

[10] [author] Borkar, Vivek SV. S. Borkar, Vivek SV. S. (2008). Stochastic approximation: a dynamical systems viewpoint 9. Springer.

[11] [author] Boruvka, AudreyA., Almirall, DanielD., Witkiewitz, KatieK. Murphy, Susan AS. A. (2018). Assessing time-varying causal effect moderation in mobile health. Journal of the American Statistical Association 113 1112–1121.

[12] [author] Carroll, Raymond JR. J., Ruppert, DavidD. Stefanski, Leonard AL. A. (1995). Measurement error in nonlinear models 105. CRC press.

[13] [author] Cesa-Bianchi, NicolòN., Gentile, ClaudioC., Lugosi, GáborG. Neu, GergelyG. (2017). Boltzmann exploration done right. Advances in neural information processing systems 30.

[14] [author] Chakraborty, BibhasB. Moodie, Erica EME. E. (2013). Statistical methods for dynamic treatment regimes 2. Springer.

[15] [author] Chen, HaoyuH., Lu, WenbinW. Song, RuiR. (2021). Statistical inference for online decision making: In a contextual bandit setting. Journal of the American Statistical Association 116 240–255.

[16] [author] Chen, HaoyuH., Lu, WenbinW. Song, RuiR. (2021). Statistical inference for online decision making via stochastic gradient descent. Journal of the American Statistical Association 116 708–719.

[17] [author] Chen, KaniK., Hu, InchiI. Ying, ZhiliangZ. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. The Annals of Statistics 27 1155–1163.

[18] [author] Chow, Yuan ShihY. S. (1967). On a strong law of large numbers for martingales. Ann. Math. Statist. 38.

[19] [author] Christopeit, NorbertN. Helmes, KurtK. (1980). Strong consistency of least squares estimators in linear regression models. The Annals of Statistics 8 778–788.

[20] [author] Chu, WeiW., Li, LihongL., Reyzin, LevL. Schapire, RobertR. (2011). Contextual bandits with linear payoff functions. In Proceedings of the fourteenth international conference on artificial intelligence and statistics 208–214. JMLR Workshop and Conference Proceedings.

[21] [author] Deshpande, YashY., Javanmard, AdelA. Mehrabi, MohammadM. (2023). Online debiasing for adaptively collected high-dimensional data with applications to time series analysis. Journal of the American Statistical Association 118 1126–1139.

[22] [author] Deshpande, YashY., Mackey, LesterL., Syrgkanis, VasilisV. Taddy, MattM. (2018). Accurate inference for adaptive linear models. In International Conference on Machine Learning 1194–1203. PMLR.

[23] [author] Dvoretzky, AryehA. (1972). Asymptotic normality for sums of dependent random variables. In Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory 6 513–536. University of California Press.

[24] [author] Dwivedi, RaazR., Tian, KatherineK., Tomkins, SabinaS., Klasnja, PredragP., Murphy, SusanS. Shah, DevavratD. (2022). Counterfactual inference for sequential experiments. arXiv preprint arXiv:2202.06891.

[25] [author] Foster, Dylan JD. J., Gentile, ClaudioC., Mohri, MehryarM. Zimmert, JulianJ. (2020). Adapting to misspecification in contextual bandits. Advances in Neural Information Processing Systems 33 11478–11489.

[26] [author] Fuller, Wayne AW. A. (2009). Measurement error models. John Wiley & Sons.

[27] [author] Ghosh, SusobhanS., Guo, YongyiY., Hung, Pei-YaoP.-Y., Coughlin, LaraL., Bonar, ErinE., Nahum-Shani, InbalI., Walton, MaureenM. Murphy, SusanS. (2024). Miwaves reinforcement learning algorithm. arXiv preprint arXiv:2408.15076.

[28] [author] Guo, YongyiY., Xu, ZipingZ. Murphy, SusanS. (2024). Online learning in bandits with predicted context. In International Conference on Artificial Intelligence and Statistics 2215–2223. PMLR.

[29] [author] Hadad, VitorV., Hirshberg, David AD. A., Zhan, RuohanR., Wager, StefanS. Athey, SusanS. (2021). Confidence intervals for policy evaluation in adaptive experiments. Proceedings of the national academy of sciences 118 e2014602118.

[30] [author] Halder, BudhadityaB., Pan, ShubhayanS. Khamaru, KoulikK. (2025). Stable Thompson Sampling: Valid Inference via Variance Inflation. arXiv preprint arXiv:2505.23260.

[31] [author] Han, QiyangQ., Khamaru, KoulikK. Zhang, Cun-HuiC.-H. (2024). UCB Algorithms for Multi-Armed Bandits: Precise Regret and Adaptive Inference. arXiv preprint arXiv:2412.06126.

[32] [author] Imbens, Guido WG. W. Rubin, Donald BD. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge university press.

[33] [author] Khalil, Hassan KH. K. Grizzle, Jessy WJ. W. (2002). Nonlinear systems 3. Prentice hall Upper Saddle River, NJ.

[34] [author] Khamaru, KoulikK., Deshpande, YashY., Lattimore, TorT., Mackey, LesterL. Wainwright, Martin JM. J. (2021). Near-optimal inference in adaptive linear regression. arXiv preprint arXiv:2107.02266.

[35] [author] Khamaru, KoulikK. Zhang, Cun-HuiC.-H. (2024). Inference with the upper confidence bound algorithm. arXiv preprint arXiv:2408.04595.

[36] [author] Klimko, Lawrence AL. A. Nelson, Paul IP. I. (1978). On conditional least squares estimation for stochastic processes. The Annals of statistics 629–642.

[37] [author] Laan, Mark JM. J. Robins, James MJ. M. (2003). Unified methods for censored longitudinal data and causality. Springer.

[38] [author] Lai, Tze LeungT. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. The Annals of Statistics 1917–1930.

[39] [author] Lai, Tze LeungT. L. Wei, Ching ZongC. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. The Annals of Statistics 154–166.

[40] [author] Lauffenburger, Julie CJ. C., Yom-Tov, EladE., Keller, Punam AP. A., McDonnell, Marie EM. E., Crum, Katherine LK. L., Bhatkhande, GauriG., Sears, Ellen SE. S., Hanken, KaitlinK., Bessette, Lily GL. G., Fontanet, Constance PC. P. et al. (2024). The impact of using reinforcement learning to personalize communication on medication adherence: findings from the REINFORCE trial. npj Digital Medicine 7 39.

[41] [author] Li, LihongL., Chu, WeiW., Langford, JohnJ. Schapire, Robert ER. E. (2010). A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web 661–670.

[42] [author] Li, LihongL., Chu, WeiW., Langford, JohnJ. Wang, XuanhuiX. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In Proceedings of the fourth ACM international conference on Web search and data mining 297–306.

[43] [author] Liao, PengP., Greenewald, KristjanK., Klasnja, PredragP. Murphy, SusanS. (2020). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 4 1–22.

[44] [author] Liao, PengP., Klasnja, PredragP. Murphy, SusanS. (2021). Off-policy estimation of long-term average outcomes with applications to mobile health. Journal of the American Statistical Association 116 382–391.

[45] [author] Liao, PengP., Qi, ZhenglingZ., Wan, RunzheR., Klasnja, PredragP. Murphy, Susan AS. A. (2022). Batch policy learning in average reward markov decision processes. Annals of statistics 50 3364.

[46] [author] Lin, LicongL., Khamaru, KoulikK. Wainwright, Martin JM. J. (2023). Semiparametric inference based on adaptively collected data. arXiv preprint arXiv:2303.02534.

[47] [author] Murphy, Susan AS. A., van der Laan, Mark JM. J., Robins, James MJ. M. Group, Conduct Problems Prevention ResearchC. P. P. R. (2001). Marginal mean models for dynamic regimes. Journal of the American Statistical Association 96 1410–1423.

[48] [author] Nie, XinkunX., Tian, XiaoyingX., Taylor, JonathanJ. Zou, JamesJ. (2018). Why adaptively collected data have negative bias and how to correct for it. In International Conference on Artificial Intelligence and Statistics 1261–1269. PMLR.

[49] [author] Qian, TianchenT., Yoo, HyesunH., Klasnja, PredragP., Almirall, DanielD. Murphy, Susan AS. A. (2021). Estimating time-varying causal excursion effects in mobile health with binary outcomes. Biometrika 108 507–527.

[50] [author] Robins, James MJ. M. (1997). Causal inference from complex longitudinal data. In Latent variable modeling and applications to causality 69–117. Springer.

[51] [author] Robins, James MJ. M. (2004). Optimal structural nested models for optimal sequential decisions. In Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data 189–326. Springer.

[52] [author] Robins, James MJ. M. Wasserman, Larry AL. A. (2013). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. arXiv preprint arXiv:1302.1566.

[53] [author] Russo, DanielD. (2016). Simple bayesian algorithms for best arm identification. In Conference on learning theory 1417–1418. PMLR.

[54] [author] Russo, DanielD. Van Roy, BenjaminB. (2014). Learning to optimize via posterior sampling. Mathematics of Operations Research 39 1221–1243.

[55] [author] Russo, Daniel JD. J., Van Roy, BenjaminB., Kazerouni, AbbasA., Osband, IanI., Wen, ZhengZ. et al. (2018). A tutorial on thompson sampling. Foundations and Trends® in Machine Learning 11 1–96.

[56] [author] Sutton, Richard SR. S. Barto, Andrew GA. G. (1998). Reinforcement learning: an introduction MIT Press. Cambridge, MA 22447 10.

[57] [author] Syrgkanis, VasilisV. Zhan, RuohanR. (2023). Post-episodic reinforcement learning inference. arXiv e-prints arXiv–2302.

[58] [author] Tewari, AmbujA. Murphy, Susan AS. A. (2017). From ads to interventions: Contextual bandits in mobile health. In Mobile health: sensors, analytic methods, and applications 495–517. Springer.

[59] [author] Trella, Anna LA. L., Zhang, Kelly WK. W., Jajal, HinalH., Nahum-Shani, InbalI., Shetty, VivekV., Doshi-Velez, FinaleF. Murphy, Susan AS. A. (2025). A deployed online reinforcement learning algorithm in an oral health clinical trial. In Proceedings of the AAAI Conference on Artificial Intelligence 39 28792–28800.

[60] [author] Tropp, Joel AJ. A. (2012). User-friendly tail bounds for sums of random matrices. Foundations of computational mathematics 12 389–434.

[61] [author] Uehara, MasatoshiM., Shi, ChengchunC. Kallus, NathanN. (2022). A review of off-policy evaluation in reinforcement learning. arXiv preprint arXiv:2212.06355.

[62] [author] Van der Vaart, Aad WA. W. (2000). Asymptotic statistics 3. Cambridge university press.

[63] [author] Waudby-Smith, IanI., Wu, LiliL., Ramdas, AadityaA., Karampatziakis, NikosN. Mineiro, PaulP. (2024). Anytime-valid off-policy inference for contextual bandits. ACM/IMS Journal of Data Science 1 1–42.

[64] [author] Xu, ZipingZ., Jajal, HinalH., Choi, Sung WonS. W., Nahum-Shani, InbalI., Shani, GuyG., Psihogios, Alexandra MA. M., Hung, Pei-YaoP.-Y. Murphy, Susan AS. A. (2025). Reinforcement Learning on Dyads to Enhance Medication Adherence. In International Conference on Artificial Intelligence in Medicine 490–499. Springer.

[65] [author] Yan, YulingY. Wainwright, Martin JM. J. (2024). Entrywise Inference for Missing Panel Data: A Simple and Instance-Optimal Approach. arXiv preprint arXiv:2401.13665.

[66] [author] Yang, JeremyJ., Eckles, DeanD., Dhillon, ParamveerP. Aral, SinanS. (2024). Targeting for long-term outcomes. Management Science 70 3841–3855.

[67] [author] Yao, JiayuJ., Brunskill, EmmaE., Pan, WeiweiW., Murphy, SusanS. Doshi-Velez, FinaleF. (2021). Power constrained bandits. In Machine Learning for Healthcare Conference 209–259. PMLR.

[68] [author] Zhan, RuohanR., Hadad, VitorV., Hirshberg, David AD. A. Athey, SusanS. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining 2125–2135.

[69] [author] Zhang, KellyK., Janson, LucasL. Murphy, SusanS. (2020). Inference for batched bandits. Advances in neural information processing systems 33 9818–9829.

[70] [author] Zhang, KellyK., Janson, LucasL. Murphy, SusanS. (2021). Statistical inference with m-estimators on adaptively collected data. Advances in neural information processing systems 34 7460–7471.

[71] [author] Zhang, Kelly WK. W., Closser, NowellN., Trella, Anna LA. L. Murphy, Susan AS. A. (2024). Replicable Bandits for Digital Health Interventions. arXiv preprint arXiv:2407.15377.

[72] [author] Zhang, Kelly WK. W., Janson, LucasL. Murphy, Susan AS. A. (2022). Statistical inference after adaptive sampling for longitudinal data. arXiv preprint arXiv:2202.07098.

[73] [author] Zhou, DongruoD., Li, LihongL. Gu, QuanquanQ. (2020). Neural contextual bandits with ucb-based exploration. In International Conference on Machine Learning 11492–11502. PMLR.