Statistical inference on multi-armed bandits with delayed feedback

Anonymous Authors[1]

Abstract

Multi armed bandit (MAB) algorithms have been increasingly used to complement or integrate with A/B tests and randomized clinical trials in e-commerce, healthcare, and policymaking. Recent developments incorporate possible delayed feedback. While existing MAB literature often focuses on maximizing the expected cumulative reward outcomes (or, equivalently, regret minimization), few efforts have been devoted to establish valid statistical inference approaches to quantify the uncertainty of learned policies. We attempt to fill this gap by providing a unified statistical inference framework for policy evaluation where a target policy is allowed to differ from the data collecting policy, and our framework allows delay to be associated with the treatment arms. We present an adaptively weighted estimator that on one hand incorporates the arm-dependent delaying mechanism to achieve consistency, and on the other hand mitigates the variance inflation across stages due to vanishing sampling probability. In particular, our estimator does not critically depend on the ability to estimate the unknown delay mechanism. Under appropriate conditions, we prove that our estimator converges to a normal distribution as the number of time points goes to infinity, which provides guarantees for large-sample statistical inference. We illustrate the finite-sample performance of our approach through Monte Carlo experiments.

1. Introduction

1.1. Motivation and contribution

In recent years, multi armed bandit (MAB) algorithms have been frequently used to complement A/B tests and clinical trials in practice, potentially because MAB algorithms not only aim to identify the best policies but also improve the overall outcomes for participants enrolled in the experiments. Whenever participant outcomes (or feedback) are not immediately observed, an increasing number of recent advancements further expand the practicality of classical MAB algorithms by incorporating such random *delays*. While carrying out MAB algorithms in real world scenarios can be time consuming and labor intensive, there is an increased desire to be able to use those adaptively collected data from MAB algorithms to assist future decision making by answering the following cause and effect questions: Does one content recommendation plan lead to more revenue than others in e-commerce for consumers who are not enrolled in the experiment? Does one medical treatment plan cause better clinical outcomes than other plans in clinical trials?

To answer the above questions, following the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974), we shall first formalize the causal parameters of interest and establish nonparametric identification results with arm-dependent *delayed* feedback (Section 2), meaning that the unobserved causal effect can be written as a function of the observed data. We then build a unified statistical inference framework allowing us to construct consistent point estimates and valid confidence intervals on the causal parameters in the presence of *delayed* outcomes when the number of time point goes to infinity (Section 3). This inference framework thus enables us to answers those raised questions with rigorous statistical guarantees. In what follows, we briefly summarize our contributions:

From a methodological standpoint, on the one hand, we propose a consistent causal effect estimator that converges to a normal distribution under a wide range of unknown delay mechanisms without estimating the arm-delay joint density (Theorem 4.1). As a result, the proposed estimator avoids estimating the delay mechanisms and can be more reliable compared to estimators using estimated arm-delay joint densities with noisy nonparametric approaches. On the other hand, our framework alleviates a well-known tension between the MAB design objective (regret minimization or reward maximization) and statistical inference objectives. This is achieved by simultaneously adaptive reweighting under-sampled arms similar to (Luedtke & Van Der Laan, 2016; Hadad et al., 2021) and self-normalizing the propensity score weights with inflated variance. The tension exists

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

because MAB algorithms are often designed to maximize expected cumulative reward outcomes and tend to assign all participants to a beneficiary arm, which leaves limited evidence to compare between the expected outcomes of a beneficiary arm and a seemingly inferior arm. The two above features of our approach together allows us to construct valid confidence intervals of the desired causal parameters in the presence of delayed feedback and vanishing propensity scores.

From a theoretical point of view, under appropriate conditions, we provide guarantees of the proposed statistical inference framework by proving that when the number of time points goes to infinity: (1) the proposed estimator converges the true causal effect in probability (Theorem 4.1), (2) our estimator converges to a normal distribution (Theorem 4.2), and (3) the variance of our proposed estimator can be consistently estimated (Theorem 4.3). The adopted sufficient conditions reveal the trade-offs among the tails of the delay distribution, the outcome distribution and the vanishing propensity score distribution. Furthermore, to solidate our high level conditions, we use ϵ-greedy as an illustrative sample to demonstrate the feasibility of our framework in Section 4.2. In particular, our theoretical result reveals that our estimator is reliable in the sense that, establishing asymptotic normality of our estimator requires neither the propensity score of any arm to converge to a positive constant, nor the estimated outcome model to converge to the true expected outcome. This is a new contribution of our approach compared to the existing literature that adopts adaptive weighting strategy.

From a practical point of view, combining our propose statistical inference framework with rigorous large sample guarantees, we hope that our approach can be readily used to assist future decision making. Take e-commerce for example, practitioners may adopt our approach to provide an accurate revenue estimate of a content recommendation plan with confidence intervals and conduct hypothesis testing to decide if two plans lead to significant revenue difference.

1.2. Related literature

Our approach is built upon the data collection mechanism in delayed multi-armed bandit algorithms. (Li et al., 2019) studied bandit online learning with unknown delays. (Vernade et al., 2020) proposed learning algorithms for linear Bandits with stochastic delayed feedback. (Gael et al., 2020) investigated the setting of stochastic bandits with arm dependent delays. (Lancewicki et al., 2021) studied stochastic bandits with unrestricted delay distributions. (Zhou et al., 2019) dug into generalized linear contextual bandits in the presence of stochastic delays. (Gyorgy & Joulani, 2021) adapted bandit learning algorithms to accommodate data in adversarial MABs. In MAB algorithms, the data are typically collected following certain learning algorithms, such as ϵ-greedy, upper confidence bound (UCB) methods and Thompson sampling, to name a few (Sutton et al., 1998). These learning programs provide a sequence of running policies that are history-dependent and evolves adaptively with time. Other than multi-armed bandit problems, delays are commonly encountered in the practical RL literature (Schuitema et al., 2010; Liu et al., 2014; Mahmood et al., 2018; Derman et al., 2021), but is less studied in theory. Recently, (Howson et al., 2021) studied regret minimization in episodic Markov decision processes with stochastic delays. (Lancewicki et al., 2022) studied MDPs with adversarial delays under full-information feedback. (Jin et al., 2022) further proposed an online-learning style algorithm for MDPs with adversarial delays under bandit feedback.

Conducting statistical inference on datasets collected from MAB algorithms has attracted attention in the past few years. For example, (Zhang et al., 2020) studied statistical inference for batched bandits. (Hadad et al., 2021) studied online policy evaluation in adaptive experiments. (Zhan et al., 2021; Bibaut et al., 2021) studied off-policy evaluation and inference in contextual bandits. (Zhang et al., 2021) proposed inference strategies based on M-estimation for adaptively collected data. (Shi et al., 2020) studied inference for off-policy evaluation in reinforcement learning settings. (Shi et al., 2022) studied inference for confounded markov decision processes. (Ramprasad et al., 2022) proposed online bootstrap inference for policy evaluation in reinforcement learning setting. (Zhou et al., 2022) studied offline policy learning for contextual bandits.

2. MAB algorithms with delayed feedback: Problem setup

We start with introducing the structure of adaptively collected data from MAB algorithms with delayed feedback. For each time point $t \in [T] \triangleq \{1, \ldots, T\}$, one action A_t is taken according to an underlying policy (see definition in Condition 2.2). Due to the possible delay of the feedback, we may not be able to observe the outcome Y_t immediately. Instead, we need to wait for a certain period of time $D_t \in \mathbb{N}_\infty$ to acquire the outcome Y_t. In other words, if $D_t = 0$, we have access to (A_t, Y_t); otherwise, only A_t is available at time point t, while Y_t will be visible at time $t + D_t$. To clearly characterize the causal parameters of interest, in accordance with the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974), we denote by $Y_t(a)$ the unobserved potential outcome that would be observed if action a was taken at time point t. Given K actions $\mathcal{A} \triangleq a_1, \ldots, a_K$, we denote the vector of potential outcomes indexed at time point t as $\{Y_t(a)\}_{a \in \mathcal{A}} \in \mathbb{R}^K$.

For the observed outcome, we work under a frequently adopted stable unit treatment value assumption (SUTVA) in

the causal inference literature (Imbens & Rubin, 2015):

Condition 2.1 (SUTVA)**.** *The outcome at time point t is determined by action A_t and not impacted by potential outcomes at other time points, suggesting that the observed outcome at t to be formulated as $Y_t = Y_t(A_t)$.*

Define historical data H_t as follows:

$$H_0 = \varnothing;\ H_t = \{(A_{t'}, (Y_{t'})_{a\in\mathcal{A}}, D_{t'})\}_{t'\le t} \text{ for } t \ge 1.$$

For the policy π_t (with subscript t) taken at time point t, we assume:

Condition 2.2 (Historical dependence of policy)**.** *$\pi_t(\cdot)$ is a vector in $\Delta(\mathcal{A})$ and only depends on the historical data. Here $\Delta(\mathcal{A})$ is a probability simplex for an action space $\mathcal{A} = \{a_1, \cdots, a_K\}$*

Following the mainstream literature on causal inference (Imbens & Rubin, 2015), we refer to $\pi_t(a)$ as the propensity score of arm a.

Throughout this manuscript, we assume that collected data are generated following the directed cyclic graph (DAG) in Figure 1. To be more rigorous, we also formalize the causal

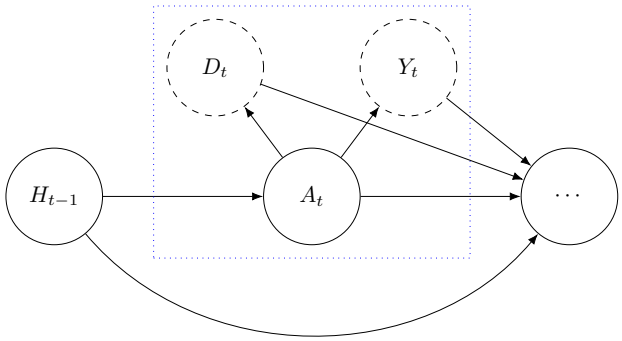

Figure 1. Data generating mechanism for MAB with delay-independent feedbacks

relationships in Figure 1 with mathematical languages in Conditions 2.3 and 2.4. Concretely, for the actions and delays, we assume:

Condition 2.3 (Action-delay mechanism)**.** *The action-delay pairs are generated as follows:*

(i) Each A_t is generated according to distribution π_t.

(ii) Each D_t is generated according to conditional distribution given the action A_t:

$$\mathbb{P}\left\{D_t = d | A_t = a, H_{t-1}\right\} \triangleq \mathbb{P}_a\left\{D = d\right\},\ d \in \mathbb{N}_\infty.$$

We first note that Condition 2.3 does not restrict the delays to be independent of actions, which broadens the applicability of our proposed framework in practice. For example, in clinical trials, our framework allows the delay mechanisms to differ across different treatment plans. Furthermore, we allow the delay to take an infinity value $+\infty$ and hence allow a part of the participants to be censored from the system. This is referred to as "partially observed MAB" in the literature (Chapelle, 2014; Krishnamurthy & Wahlberg, 2009). Lastly, the above assumption also implies that potential outcomes at time point t is independent with arm-delay pair conditional on the historical data in the sense that:

$$\{Y_t(a)\}_{a\in\mathcal{A}} \perp (A_t, D_t) \mid H_{t-1}.$$

For the actions, potential outcomes and delays, we assume:

Condition 2.4 (Distribution of potential outcomes)**.** *For $t \in [T]$, $\{Y_t(a)\}_{a\in\mathcal{A}}$ are assumed to be generated as i.i.d. copies of a random vector $\{Y(a)\}_{a\in\mathcal{A}}$ which follows an unknown distribution $\mathbb{P}$. Besides, $\{Y_t(a_1), \ldots, Y_t(a_K)\}_{t\in[T]}$ are independent of all the actions and delays:*

$$\{Y_t(a_1), \ldots, Y_t(a_K)\}_{t\in[T]} \perp \{A_t, D_t\}_{t\in[T]}.$$

3. Inference target and proposed method

In this section, we formulate our inference target following the problem setup introduced in the previous section and propose a unified statistical inference framework that simultaneously accounts for unknown delay mechanism (i.e., unknown $\mathbb{P}_a\{D = d\}$) and vanishing propensity scores (i.e., $\pi_t(a) \to 0$ as $t \to \infty$).

To unify presentations, we aim to make inference on the following causal parameter:

$$Q(w^\star) = \mathbb{E}\left\{\sum_{a\in\mathcal{A}} w^\star(a)Y(a)\right\} = \sum_{a\in\mathcal{A}} w^\star(a)Q(a), \tag{3.1}$$

where $w^\star(\cdot) \in \mathbb{R}$ is a pre-specified function of $a \in \mathcal{A}$, and $Q(a) = \mathbb{E}[Y(a)]$ measures the expected potential outcome of arm a.

The above general parameter of interest $Q(w^\star)$ is a unified presentation of several causal parameters of interest in practice. For example, when we set $w^\star(a) = 1$ and $w^\star(a') = 0$ for $a' \neq a$, $Q(w^\star) = Q(a)$ which evaluates the causal effect of arm a. When we compare the causal effect sizes between two arms a_1 and a_2, we can set $w^\star(a_1) = 1$, $w^\star(a_2) = -1$ and $w^\star(a) = 0$ for all other a which does not equal to a_1 and a_2. As another example, when we want to conduct policy evaluation where a target policy π is allowed to differ from the data collection policy $\{\pi_t\}_{t=1}^T$, then we can define $w^\star(a) = \pi(a)$. Here, a policy π (without subscript t) is an element of $\Delta(\mathcal{A})$ which is a probability simplex for an action space $\mathcal{A} = \{a_1, \cdots, a_d\}$.

The causal parameter defined in 3.1 can not be "nonparametrically identified" in its current form, because it involves unobserved potential outcomes. Here, by "nonparametric identification" we mean that the causal parameter involving unobserved potential outcomes can be written as a function of observed data. Without the delayed outcome, classical identification approaches such as inverse propensity score weighting (IPW, see (Rosenbaum & Rubin, 1983)) enable us to identify $Q(w^\star)$ with

$$Q(w^\star) = \mathbb{E}\left\{\sum_{a'\in\mathcal{A}} \frac{w^\star(a)Y_t\mathbf{1}\left\{A_t=a\right\}}{\pi_t(a)}\right\}. \tag{3.2}$$

The intuition for the above identification result is to reweight observations using the sampling probability $\pi_t(a)$ of each arm z.

In the presence of delayed feedback, the above classical IPW based identification approach in 3.2 is no longer valid, simply because Y_t might not be available at time point t due to delay. To address this issue, we first provide a new identification result appropriately taking the delay mechanism into account:

$$Q(w^\star) = \mathbb{E}\left\{\sum_{a\in\mathcal{A}} \frac{w^\star(a)Y_t\mathbf{1}\left\{A_t=a, D_t\le T-t\right\}}{\pi_t(a)\mathbb{P}_a\left\{D_t\le T-t\right\}}\right\}.$$

Because the expectation on the right hand side of the above equation only involves observed data, as long as the arm-delay joint density $\mathbb{P}_a\left\{D_t\le T-t\right\}$ is well estimated, it seems that an estimator of $Q(w^\star)$ can be naturally constructed by replacing the unknown expectation with its sample analogue.

Nevertheless, two practical challenges potentially occur when directly using the above plug-in approach. On the one hand, accurately estimation of unknown conditional delayed distribution either requires imposing potentially misspecified parametric models on $\mathbb{P}_a\left\{D_t\le T-t\right\}$ or depends on nonparametric approaches with potentially noisy behaviors in practice. Moreover, estimating $\mathbb{P}_a\left\{D_t\le T-t\right\}$ can be particularly challenging when the delays present a heavy-tailed pattern. As a remedy, we aim to propose an estimator of $Q(w^\star)$ avoids estimating of delaying distribution either parametrically or nonparametrically, and thus free us from the burden of estimating arm-delay distribution.

On the other hand, because MAB algorithms are designed for regret minimization and inferior arms are less likely to be sampled when t is large, this suggests that for those inferior arms the propensity scores $\pi_t(a)$ can be a very small number for large t. In the presence of vanishing propensity scores, IPW based estimators tend to have large variance and no longer converge to a normal distribution as the number of time points goes to infinity (Hadad et al., 2021; Ma & Wang, 2020). This renders statistical inference based on standard normal approximations invalid.

To address the above challenges, we propose a "Delay-adjusted augmented inverse propensity weighting" (DAIPW) estimator that, not only accounts for the arm-dependent delay mechanisms, but also restores the asymptotic normal distribution by self-normalizing and adaptively weighting the arms with vanishing propensity scores:

$$\widehat{Q}_{\texttt{DAIPW}}(w^\star) = \sum_{a\in\mathcal{A}} w^\star(a)\widehat{Q}_{\texttt{DAIPW}}(a), \tag{3.3}$$

where

$$\widehat{Q}_{\texttt{DAIPW}}(a) = \frac{\sum_{t\in[T]} h_t(a)\left\{(Y_t-\widehat{\mu}_t(a))\gamma_t(a)\right\}}{\sum_{t=1}^T h_t(a)\gamma_t(a)} + \frac{\sum_{t\in[T]} h_t(a)\widehat{\mu}_t(a)}{\sum_{t\in[T]} h_t(a)} \tag{3.4}$$

and $\gamma_t(a)$ is the inverse propensity score weighted indicator of whether the outcome is delayed at time t for arm a,

$$\gamma_t(a) = \frac{\mathbf{1}\left\{A_t=a, D_t\le T-t\right\}}{\pi_t(a)}.$$

Here, $\widehat{\mu}_t(a)$ is an estimator for the outcome model $\mu(a) = \mathbb{E}[Y_t|A_t=a]$ at time t, and $h_t(a)$ is a sequence of adaptive weights that, intuitively, down-weighting the role of under-sampled arms in constructing our estimator. Both $\widehat{\mu}_t(a)$ and $h_t(a)$ are constructed only using the historical data H_{t-1} collected before time t. As $h_t(a)$ plays an essential role in constructing $\widehat{Q}_{\texttt{DAIPW}}(w^\star)$, we demonstrate in Section 4.2 how to adaptively choose $h_t(a)$ in commonly adopted MAB algorithms, where we use ϵ-greedy as an illustrative example.

Before proposing a variance estimator for uncertainty quantification, we provide three additional insights hoping to demonstrate the potential merits of our estimator $\widehat{Q}_{\texttt{DAIPW}}(w^\star)$.

First, the proposed estimator takes the presence of delays into account by the introduction of the inverse propensity score weighted observation indicator $\gamma_t(a)$. A critical point is that **we neither need to adjust for the unknown conditional delay distribution $\mathbb{P}_a\left\{D_t\le T-t\right\}$ nor need to worry about situation where the delay mechanism displays a heavy-tailed pattern.** Under appropriate conditions listed in Section 4.1, we can still justify the consistency of $\widehat{Q}_{\texttt{DAIPW}}(w^\star)$ even for heavy-tailed delays.

Second, the proposed estimator can be viewed as a delay-adjusted generalization of the augmented IPW estimator (AIPW) and the classical Hájek estimator (Hájek, 1971), therefore it inherits some instinct strength from these two estimators. On the one hand, due to augmenting the IPW estimator with the estimated outcome model $\widehat{\mu}_t(a)$, our estimator extracts additional information stored in the data

and may further improve our finite sample performance. On the other hand, **the adaptation of Hájek estimation demonstrates two-fold benefits: it is both helpful for variance stabilization due to small propensity scores and necessary for adjusting for delayed, even possibly never observed, outcomes.** In fact, in the extreme case where a part of the outcomes are never observed, the non-Hájek version of our estimator estimator is seriously biased, which is demonstrated by our additional theoretical derivation and simulation results provided in the Appendix.

Third, the proposed estimator **incorporates arm-wise adaptive weights $h_t(a)$ to mitigate the variance inflation** due to possibly vanishing propensity scores (i.e., $\pi_t(a) \to 0$ as $t \to \infty$). Such a consideration is a generalization of the non-delayed policy evaluation schemes provided in (Luedtke & Van Der Laan, 2016; Bibaut et al., 2021; Hadad et al., 2021; Zhan et al., 2021).

Before we end the section, we introduce a variance estimator of $\widehat{Q}_{\texttt{DAIPW}}(w^\star)$,

$$\widehat{V} = \sum_{a\in\mathcal{A}} \{w^\star(a)\}^2 \widehat{V}(a),$$

$$\widehat{V}(a) = \frac{\sum_{t\in[T]} h_t(a)^2 \left\{(Y_t - \widehat{Q}_{\texttt{DAIPW}}(a))\gamma_t(a)\right\}^2}{\left(\widehat{p}(a)\sum_{t\in[T]} h_t(a)\right)^2},$$

where $\widehat{p}(a)$ is an estimator of $p(a) = \mathbb{P}_a\{D < \infty\}$ (i.e., the probability of having finite delays for arm a)

$$\widehat{p}(a) = \frac{\sum_{t\in[T]} h_t(a)\gamma_t(a)}{\sum_{t\in[T]} h_t(a)}.$$

The point estimator $\widehat{Q}_{\texttt{DAIPW}}(w^\star)$ and the variance estimator $\widehat{V}$ together enable us to construct $(1-\alpha)$-level confidence interval for $Q(w^\star)$:

$$\left[\widehat{Q}_{\texttt{DAIPW}}(w^\star) \pm z_{\alpha/2}\widehat{V}^{\frac{1}{2}}\right],$$

where $z_{\alpha/2}$ is the upper $(1-\alpha/2)$ quantile of a standard normal distribution. In our simulation results provided in Section 5, we set $\alpha = 0.05$.

4. Theoretical investigation

In this section, under appropriate conditions, we provide guarantees of the proposed statistical inference framework by proving that when $T \to \infty$: (1) $\widehat{Q}_{\texttt{DAIPW}}(a)$ converges to $Q(a)$ in probability (Theorem 4.1), (2) $\widehat{Q}_{\texttt{DAIPW}}(a) - Q(a)$ converges to a centered normal distribution with variance $V(a)$ (Theorem 4.2), and (3) $\widehat{V}(a)$ converges to $V(a)$ in probability (Theorem 4.3), for all arms $a \in \mathcal{A}$. As $Q(w^\star)$ is a weighted combination of $Q(a)$, the three results above together justifies the statistical validity of the proposed framework in an asymptotic sense (i.e., T is large).

Furthermore, to solidate our high level conditions in (A1)-4.9, we use ϵ-greedy as a running sample to demonstrate the feasility of these conditions in Section 4.2.

4.1. Large sample guarantees of the proposed estimator

We begin by providing a consistency result of our estimator in 3.3. The core idea is to decompose the error $\widehat{Q}(a) - Q(a)$ into two parts. The first part is a martingale sequence with stable variance and bounded moments under property weighting and mild conditions on the potential outcomes. The second part is an asymptotically vanishing remainder term compared to the first part. Then we can conclude the proof by carefully checking the conditions for martingale limit theorems (Hall & Heyde, 1980).

We list the assumptions used to quantify the scale of the adaptive weights $h_t(a)$ as well as the interplay between $h_t(a)$, the moments of $Y(a)$ and the delay distribution.

(A1) *Negligible adaptive weights.* For all $a \in \mathcal{A}$,

$$\frac{\max_{t\in[T]} h_t(a)}{\sum_{t=1}^T h_t(a)} \xrightarrow{\mathbb{P}} 0. \tag{4.5}$$

(A2) *Appropriate delay tails.* For all $a \in \mathcal{A}$,

$$\frac{\sum_{t\in[T]} h_t(a)\mathbb{P}_a\{T-t<D<\infty\}}{\left(\sum_{t=1}^T \mathbb{E}\left\{\frac{h_t(a)^2}{\pi_t(a)}\mathbb{P}_a\{D\le T-t\}\right\}\right)^{\frac{1}{2}}} = O_{\mathbb{P}}(1). \tag{4.6}$$

(A3) *Infinite sampling.* For all $a \in \mathcal{A}$,

$$\frac{\mathbb{E}\left\{\sum_{t=1}^T h_t(a)^2\pi_t(a)^{-1}\mathbb{P}_{a-1}\{D_t \le T-t\}\right\}}{\left(\sum_{t=1}^T h_t(a)\right)^2} \xrightarrow{\mathbb{P}} 0. \tag{4.7}$$

(A4) *Lyapunov condition.* For all $a \in \mathcal{A}$,

$$\frac{\sum_{t=1}^T h_t(a)^{2+\delta}\pi_t(a)^{-(1+\delta)}\mathbb{P}_a\{D\le T-t\}}{\left(\sum_{t=1}^T \mathbb{E}\left\{\frac{h_t(a)^2}{\pi_t(a)}\mathbb{P}_a\{D\le T-t\}\right\}\right)^{\frac{2+\delta}{2}}} \xrightarrow{\mathbb{P}} 0, \tag{4.8}$$

(A5) *Variance convergence condition.* For all $a \in \mathcal{A}$, these exists some $p > 1$,

$$\frac{\sum_{t=1}^T h_t(a)^2\pi_t(a)^{-1}\mathbb{P}_a\{D\le T-t\}}{\sum_{t=1}^T \mathbb{E}\{h_t(a)^2\pi_t(a)^{-1}\mathbb{P}_a\{D\le T-t\}\}} \xrightarrow{L_p} 1. \tag{4.9}$$

We add some comments on these conditions. Condition (A1) requires each single adaptive weight is negligible compared to the sum of all the weights. Intuitively speaking, this ensures that the weights are relatively balanced

in magnitude and there is no dominating components or outliers. Condition (A2) requires the tail of the delay $\mathbb{P}_a\{T-t<D<\infty\}$ to vanish at some appropriate rate. Later we will show that Condition (A2) allows heavy tailed delay which may even have bounded expectation. Condition (A3) is standard in the adaptive-weighting based policy evaluation frameworks (e.g. Hadad et al., 2021). We generalize it to incorporate delay distributions. Condition (A4) and (A5) are needed for martingale limit theories. In general, Condition (A1) to (A5) can be used as criteria or guidance to construct valid adaptive weights. In Section 4.2 we elaborate more on these conditions using one concrete data collection MAB algorithms.

Under Conditions (A1) to (A5), we next shows the consistency of our estimator:

Theorem 4.1 (Consistency)**.** *Under Conditions (A1) - (A4), we further assume that $Y(a)$ has $(2+\delta)$-th moment for some small positive δ and $\widehat{\mu}_t(a)$ is bounded and converges in probability to some constant. We have*

$$\widehat{p}(a)-p(a)\xrightarrow{\mathbb{P}}0,\quad \widehat{Q}_{\texttt{DAIPW}}(a)-\mu(a)\xrightarrow{\mathbb{P}}0.$$

Theorem 4.1 consists of two results: consistency of the estimated non-censoring probability $\widehat{p}(a)$ and consistency of the DAIPW estimator $\widehat{Q}_{\texttt{DAIPW}}(a)$ for each arm. Then former part works as an intermediate result which provides evidence and insights of the validity of DAIPW even in the presence of neven observed outcomes. It is also a crucial component for justifying the validity of the variance estimation in Theorem 4.3.

We next present the theoretical result demonstrating that our estimator converges to a normal distribution when T goes to infinity:

Theorem 4.2 (Asymptotic normality)**.** *Under Conditions (A1) - (A5), we further assume that $Y(a)$ has $(2+\delta)$-th moment for some small positive δ and $\widehat{\mu}_t(a)$ is bounded and converges in probability to some constant. We have*

$$\left[\frac{\widehat{Q}_{\texttt{DAIPW}}(a_k)-\mu(a_k)}{V(a_k)^{1/2}}\right]^{\top}_{k\in[K]}\to\mathcal{N}(\mathbf{0},\boldsymbol{I}_K),$$

where

$$V(a)=\frac{\sum_{t\in[T]}\mathbb{E}\left\{h_t(a)^2\pi_t(a)^{-1}\sigma_a^2\mathbb{P}_a\{D\le T-t\}\right\}}{\left(p(a)\sum_{t\in[T]}h_t(a)\right)^2}.$$

Theorem 4.2 implies the proposed estimators $[\widehat{Q}(a)]_{a\in\mathcal{A}}$ follow a multivariate normal distribution as T goes to infinity. Besides, their asymptotic between-arm correlation is zero. As $Q(w^\star)$ is a linear combination of $Q(a)$, this suggests that the proposed estimator $\left[\widehat{Q}_{\texttt{DAIPW}}(w^\star)\right.$ also converges to a normal distribution.

A direct implication of Theorem 4.2 is that, **the asymptotic normality of our estimator requires neither the propensity scores $\pi_t(a)$ on any arm to converge to a positive constant, nor the $\widehat{\mu}(a)$ to converge to the true $\mathbb{E}[Y_t|A_t=a]$.** This unique property of our estimator is a result of incorporating Hájek-type strategy in MAB problems, which is a new contribution of our approach compared to the existing literature that adopts adaptive weighting strategy.

Lastly, we justify the validity of the variance estimation and complete the story of statistical inference.

Theorem 4.3 (Variance estimation)**.** *Under the same conditions listed in Theorem 4.2, we have*

$$\frac{\widehat{V}(a)}{V(a)}\xrightarrow{\mathbb{P}}1.$$

4.2. Verification of our framework in ϵ-greedy algorithms

In this section, we discuss the choice of adaptive weights in the ϵ-greedy algorithms and show that the proposed weights satisfy Condition (A1)-(A5).

In ϵ-greedy algorithms, the algorithm picks the arm $a_{t,\max}$ that achieves the highest sample average. Then in the following stage the algorithm pulls $a_{t,\max}$ with probability greater than $1-\epsilon$ (exploitation) and the rest $(d-1)$ arms with probability $\epsilon/(d-1)$ (exploration). In practice, we can set a diminishing ϵ_t to reduce the portion of exploration. In particular, we study policy evaluation problem of the power decaying ϵ_t; i.e., $\epsilon_t=t^{-\alpha}$ for some $\alpha\ge0$. Mathematically, let $\overline{Y}_{a,t}$ denote the averaged observed outcome collected on arm a before time t. Then

$$\pi_t(a)=\begin{cases}1-\epsilon_t, & \overline{Y}_{a,t-1}\ge\overline{Y}_{a',t-1};\\ \frac{\epsilon_t}{d-1}, & \overline{Y}_{a,t-1}<\overline{Y}_{a',t-1}.\end{cases}\tag{4.10}$$

A simple adaptive weighting strategy is available in this algorithm by setting

$$h_t(a)=\sqrt{\pi_t(a)}.\tag{4.11}$$

In the non-delayed setting, this was studied by (Hadad et al., 2021) and termed as "constant allocation strategy". The following corollary justifies this choice in our current setup with delays:

Corollary 4.1. *Assume $\pi_t(a)\ge Ct^{-\alpha}$ for some $\alpha\in[0,1)$. Also assume the delay distribution satisfies:*

$$\mathbb{P}_a\{D=0\}>0,$$

$$\mathbb{P}_a\{t\le D<\infty\}=O(t^{-\beta})\text{ for all }t\ge1\text{ and some }\beta\ge\frac{1}{2}.$$

Then (A1) - (A5) are satisfied.

Remark 4.1. *Define the margin of the bandit as the gap between the largest and second largest expected outcome. The requirement of $\beta \geq 1/2$ can be relaxed to $\alpha + \beta \geq \frac{1}{2}$ if the margin is nonzero. See the proof in the Appendix for more details.*

We can check that for the ϵ-greedy algorithm, the conditions in Corollary 4.1 can be satisfied. More generally, Corollary 4.1 can be applied for a wider class of online learning algorithms, such as Thompson sampling or UCB-based algorithms. The sole requirement is that the algorithm preserves certain level of probability for randomization through the trajectory ($\pi_t(a) \geq Ct^{-\alpha}$ where α is not too small). Such conditions have been discussed similarly in many other sequential policy evaluation frameworks, such as batched bandit learning (Zhang et al., 2020), etc.

5. Simulation study

In this section, we verify our theoretical results in two simulation designs by comparing the performance of the following estimators in Table 1. More concretely, `DAIPW` is the proposed estimator. `Mean` is the simple sample mean estimator $\widehat{Q}(a) = N(a)^{-1} \sum_{t=1}^{T} Y_t \mathbf{1}(A_t = a, D_t \leq T - t)$. `NH0` is the ordinary (non-Hájek) IPW estimator $\widehat{Q}(a) = T^{-1} \sum_{t=1}^{T} Y_t \mathbf{1}(A_t = a, D_t \leq T - t)/\pi_t(a)$. `NH` is the usual AIPW estimator by (Hadad et al., 2021), i.e., `NH0` with outcome adjustment. In the first simulation design, we

Table 1. Estimators Considered in Numerical Studies

Estimator	DAIPW	Mean	NH	NH0
Delay-adjusted	✓	✓	✗	✗
Adaptively-weighted	✓	✗	✓	✓
Hájek-type	✓	✓	✗	✗
Outcome-adjusted	✓	✗	✓	✗

study the performance of varies estimators with/without margins. In the second simulation design, we compare different estimators under multiple delay mechanisms.

5.1. Simulation results with different margins in ϵ-greedy algorithms

We run ϵ-greedy on binary bandits and evaluate the impact of margins for the performance of varies estimators in Table 1. The potential outcomes are generated from normal distribution with variance 1. For Arm $a = 1$, there exists a positive censoring probability: $\mathbb{P}_1\{D = \infty\} = 0.5$. Arm 2 does not have never observed outcome, i.e., the censoring probability $\mathbb{P}_2\{D = \infty\} = 0$. We compare two settings with different sizes of margins: (i) zero margin: $\mu(1) - \mu(2) = 0$; (ii) non-zero margin: $\mu(1) - \mu(2) = 0.1$. The results are summarized in Figure 2 and Figure 3.

From Figure 2 and Figure 3 we can see that, the proposed

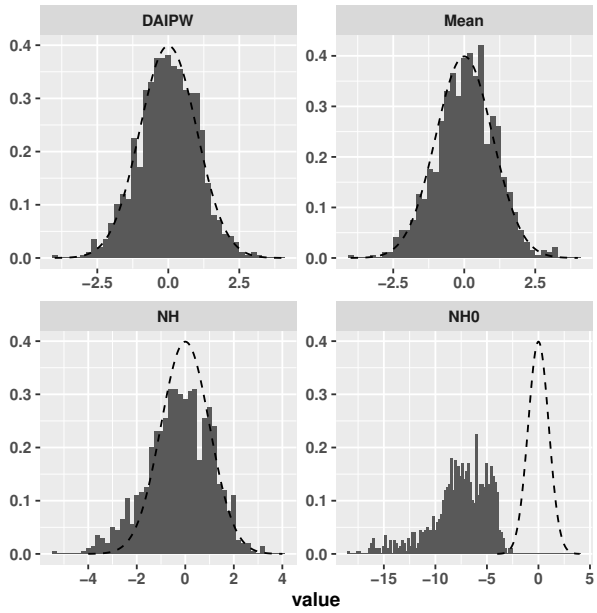

Figure 2. ϵ-greedy over zero margin bandits $\mu(1) - \mu(2) = 0$.

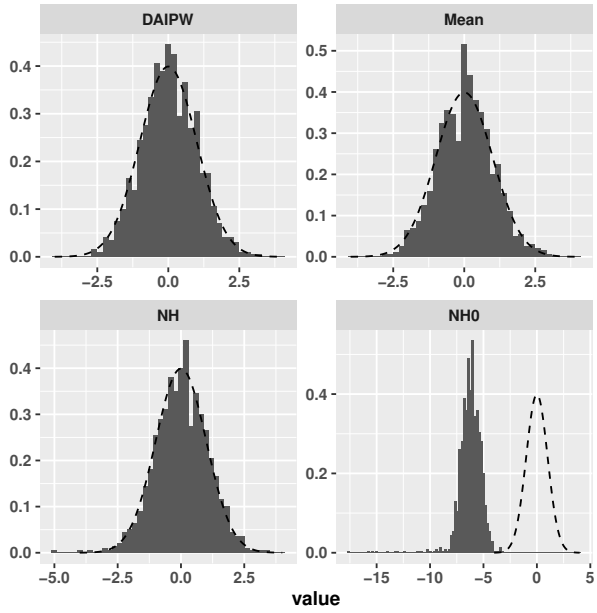

Figure 3. ϵ-greedy over nonzero margin bandits with $\mu(1) - \mu(2) = 0.1$.

`DAIPW` estimator provides a better approximation the normal distribution in both zero margin and nonzero margin settings. The non-adaptively weighted estimator with delay adjustment (`Mean`) slightly skewed to the left for the non-zero margin case. The skewing effect will be more prominent as the margin increases since the propensity score $\pi_t(2)$ converges to 0 faster. The non-Hájek estimator (`NH`)

works poorly when the margin is zero, because the outcome model estimator converges slower in that scenario. `NH0`, which neither accounts for delays nor adjusts for the outcome model, is severely biased regardless of the size of the margin.

5.2. Simulation results under different delay mechanisms

In this section, we run ϵ-greedy on a binary bandit with $\mu(1) = 1.0$ and $\mu(2) = 0.5$. We compare four different delay mechanisms:

- No finite delay. No other source of delay is included except for the censoring on Arm 1.
- Negative binomial delay. The delay distribution on both arms follow Negative Binomial distributions, which gives a subexponential-tailed delay.
- Pareto delay. The delay distribution on both arms follow (rounded) Pareto distributions, which gives a polynomial-type heavy-tailed delay.

Due to space limit, we present the first case with no finite delay, and leave the simulation results for other settings to the Supplementary Material. The results are summarized in Figure 4. From Figure 4, we can see that the unweighted estimator (`Mean`) has a higher absolute bias, and the estimated standard deviation based on `Mean` is also larger, which leads to a longer confidence interval. Using the proposed adaptive weighted Hájek estimator (`DIPW`) or adjusting for the outcome model (`NH`) can mitigate the problem. However, the non-Hájek estimator (`NH`) tends to underestimate the variance because it does not adjust for the censoring probability, leading to under covered confidence intervals. In sum, our approach is the only method that provides accurate point estimate and valid confidence interval (meaning that the coverage probability attains the nominal 95% level).

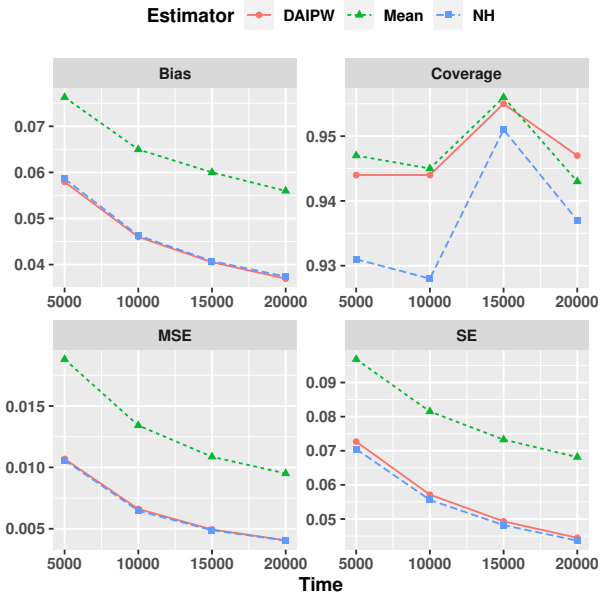

Figure 4. Evaluation of ϵ-greedy with no finite delays

6. Discussion

In this manuscript, we provide a unified statistical inference framework when data are adaptively collected from MABs with delayed feedback. Under appropriate conditions, we prove that our estimator converges to a normal distribution as the number of time points goes to infinity, which provides guarantees for large-sample statistical inference.

We add some discussions on the potential generalizations and challenges of our framework. First, a realistic generalization can potentially include delay-dependent outcome or outcome-dependent delay mechanism. Nonparametric identification and statistical inference can be more challenging in these settings. Second, practitioners may also care about statistical quantification of optimal policies, which is unknown and are typically learned through online algorithms. Therefore, it is of interest to establish methodology that combines policy evaluation and policy learning. We leave these possible extensions as future endeavor.

Software

The source code is attached in the Supplementary Material.

References

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., and van der Laan, M. Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, 34: 28548–28559, 2021.

Chapelle, O. Modeling delayed feedback in display advertising. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1097–1105, 2014.

Derman, E., Dalal, G., and Mannor, S. Acting in delayed environments with non-stationary markov policies. *arXiv preprint arXiv:2101.11992*, 2021.

Gael, M. A., Vernade, C., Carpentier, A., and Valko, M. Stochastic bandits with arm-dependent delays. In *International Conference on Machine Learning*, pp. 3348–3356. PMLR, 2020.

Gyorgy, A. and Joulani, P. Adapting to delays and data in adversarial multi-armed bandits. In *International Con-*

ference on Machine Learning, pp. 3988–3997. PMLR, 2021.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15):e2014602118, 2021.

Hájek, J. Comment on "an essay on the logical foundations of survey sampling, part one". *The foundations of survey sampling*, 236, 1971.

Hall, P. and Heyde, C. C. *Martingale limit theory and its application*. Academic press, 1980.

Howson, B., Pike-Burke, C., and Filippi, S. Delayed feedback in episodic reinforcement learning. *arXiv preprint arXiv:2111.07615*, 2021.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jin, T., Lancewicki, T., Luo, H., Mansour, Y., and Rosenberg, A. Near-optimal regret for adversarial mdp with delayed bandit feedback. *arXiv preprint arXiv:2201.13172*, 2022.

Krishnamurthy, V. and Wahlberg, B. Partially observed markov decision process multiarmed bandits—structural results. *Mathematics of Operations Research*, 34(2):287–302, 2009.

Lancewicki, T., Segal, S., Koren, T., and Mansour, Y. Stochastic multi-armed bandits with unrestricted delay distributions. In *International Conference on Machine Learning*, pp. 5969–5978. PMLR, 2021.

Lancewicki, T., Rosenberg, A., and Mansour, Y. Learning adversarial markov decision processes with delayed feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7281–7289, 2022.

Li, B., Chen, T., and Giannakis, G. B. Bandit online learning with unknown delays. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 993–1002. PMLR, 2019.

Liu, S., Wang, X., and Liu, P. X. Impact of communication delays on secondary frequency control in an islanded microgrid. *IEEE Transactions on Industrial Electronics*, 62(4):2021–2031, 2014.

Luedtke, A. R. and Van Der Laan, M. J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 44(2):713, 2016.

Ma, X. and Wang, J. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Mahmood, A. R., Korenkevych, D., Komer, B. J., and Bergstra, J. Setting up a reinforcement learning task with a real-world robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4635–4640. IEEE, 2018.

Neyman, J. S. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W., and Cheng, G. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, pp. 1–14, 2022.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Schuitema, E., Buşoniu, L., Babuška, R., and Jonker, P. Control delay in reinforcement learning for real-time dynamic systems: A memoryless approach. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3226–3231. IEEE, 2010.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020.

Shi, C., Zhu, J., Ye, S., Luo, S., Zhu, H., and Song, R. Off-policy confidence interval estimation with confounded markov decision process. *Journal of the American Statistical Association*, (just-accepted):1–30, 2022.

Sutton, R. S., Barto, A. G., et al. Introduction to reinforcement learning. 1998.

Vernade, C., Carpentier, A., Lattimore, T., Zappella, G., Ermis, B., and Brueckner, M. Linear bandits with stochastic delayed feedback. In *International Conference on Machine Learning*, pp. 9712–9721. PMLR, 2020.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2125–2135, 2021.

Zhang, K., Janson, L., and Murphy, S. Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829, 2020.

Zhang, K., Janson, L., and Murphy, S. Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34: 7460–7471, 2021.

Zhou, Z., Xu, R., and Blanchet, J. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 2022.

A. Proof of the main results

A.1. Some probability results

The following lemma is attributed to Theorem 3.2 and Corollary 3.1 of (Hall & Heyde, 1980), which provides a central limit theorem for martingale difference arrays:

Lemma A.1 (CLT for martingale difference array)**.** *Let $\{S_{ni}, \mathcal{F}_{ni}, 1 \le i \le k_n, n \ge 1\}$ be a zero-mean, square-integrable martingale array with differences X_{ni} and let η^2 be an a.s. finite r.v. Suppose that the following conditions hold:*

(i) Conditional Lindeberg condition:

$$\text{for all } \epsilon > 0, \sum_{i \in [k_n]} \mathbb{E}\left\{X_{ni}^2 \mathbf{1}\left\{|X_{ni}| > \epsilon\right\} \mid \mathcal{F}_{n,i-1}\right\} \xrightarrow{\mathbb{P}} 0,$$

(ii) Convergence of conditional variance:

$$W_{nk_n}^2 = \sum_{i \in [k_n]} \mathbb{E}\left\{X_{ni}^2 \mid \mathcal{F}_{n,i-1}\right\} \xrightarrow{\mathbb{P}} \eta^2.$$

(iii) Nested σ-fields condition:

$$\mathcal{F}_{n,i} \subset \mathcal{F}_{n+1,i} \text{ for } 1 \le i \le k_n, n \ge 1.$$

Two important quantities in the study of a (zero-mean) martingale

$$\{S_{ni} = \sum_{j \in [i]} X_{nj}, \mathcal{F}_{ni}, 1 \le i \le k_n\}$$

is the *conditional variance* and the *squared variation*, which are both estimation of the variance $\mathbb{E}\left\{S_n^2\right\}$ and are defined as follows:

$$W_{ni}^2 = \sum_{j \in [i]} \mathbb{E}\left\{X_{nj}^2 \mid \mathcal{F}_{n,j-1}\right\},$$
$$U_{ni}^2 = \sum_{j \in [i]} X_{nj}^2, \quad \text{for } 1 \le i \le k_n.$$

The following lemma is taken from Theorem 2.23 of (Hall & Heyde, 1980), which gives conditions under which V_{ni}^2 and U_{ni}^2 are asymptotically equivalent:

Lemma A.2 (Asymptotic equivalence of W_{ni}^2 and U_{ni}^2)**.** *Suppose the following conditions hold:*

(i) The conditional variances $W_{nk_n}^2$ are tight:

$$\sup_n \mathbb{P}\left\{W_{nk_n}^2 > \lambda\right\} \to 0 \text{ as } \lambda \to \infty.$$

(ii) The conditional Lindeberg condition holds:

$$\forall \epsilon > 0, \sum_{i \in [k_n]} \mathbb{E}\left\{X_{ni}^2 \mathbf{1}\left\{|X_{ni}| > \epsilon\right\} \mid \mathcal{F}_{n,i-1}\right\} \xrightarrow{\mathbb{P}} 0,$$

Then

$$\max_i |U_{ni}^2 - W_{ni}^2| \xrightarrow{\mathbb{P}} 0.$$

A sufficient condition for the conditional Lindeberg condition is the so called *conditional Lyapunov condition*: for some $\delta > 0$,

$$\sum_{i\in[k_n]} \mathbb{E}\left\{|X_{n,i}|^{2+\delta} \mid \mathcal{F}_{n,i-1}\right\} \xrightarrow{\mathbb{P}} 0.$$

We cite a Lemma from (Hadad et al., 2021) (Lemma 10), which is inherently a probabilistic version of the well-known Topelitz lemma (Hall & Heyde, 1980):

Lemma A.3 (Relaxed version of Lemma 10 of (Hadad et al., 2021))**.** *Let $a_{T,t}$ be a triangular sequence of nonnegative weight vectors satisfying*

$$\max_{1\le t\le T} a_{T,t} \xrightarrow{\mathbb{P}} 0, \quad \sum_{t=1}^{T} a_{T,t} = O_{\mathbb{P}}(1).$$

Let x_t be a sequence of bounded random variables (with bound $B > 0$) satisfying $x_t \xrightarrow{\mathbb{P}} 0$. Then

$$\sum_{t=1}^{T} a_{T,t}x_t \xrightarrow{\mathbb{P}} 0.$$

Remark A.1. *Lemma 10 of (Hadad et al., 2021) assumes* $\operatorname{plim}_{T\to\infty} \sum_{t=1}^{T} a_{T,t} \le C$ *and* $x_t \xrightarrow{a.s.} 0$*, which is not necessary by slightly modifying the proof.*

Proof. For any $\epsilon > 0$ and $\delta > 0$, there exists C_1 and T_1, such that for any $T \ge T_1$,

$$\mathbb{P}\left\{\sum_{t=1}^{T} a_{T,t} < C_1\right\} \ge 1 - \frac{\delta}{3}.$$

There exists T_2, such that for any $T \ge T_2$

$$\mathbb{P}\left\{|x_t| < \frac{\epsilon}{2C_1}\right\} > 1 - \frac{\delta}{3}.$$

There exists T_3, such that for any $T \ge T_3$,

$$\mathbb{P}\left\{\max_{1\le t\le T} a_{T,t} \le \frac{\epsilon}{2BT_2}\right\} > 1 - \frac{\delta}{3}.$$

Therefore, for any $T \ge \max\{T_1, T_2, T_3\}$, over the intersection of the above events,

$$\begin{aligned}\sum_{t=1}^{T} a_{T,t}|x_t| &= \sum_{t=1}^{T_2-1} a_{T,t}|x_t| + \sum_{t=T_2}^{T} a_{T,t}|x_t| \\ &< BT_2 \max_{t\in[T]} a_{T,t} + \Big(\sum_{t=T_2}^{T} a_{T,t}\Big)\frac{\epsilon}{2C_1} < \epsilon.\end{aligned}$$

□

A.2. Proof of Theorem 4.1

Proof of Theorem 4.1. We finish the proof by combining the following three steps: (i) proving consistency of delaying probability estimation to its expectation; (ii) proving consistency of the estimator.

Step (i). We show that

$$\widehat{p}(a) \xrightarrow{\mathbb{P}} p(a). \tag{1.12}$$

(1.12) follows from the fact that

$$h_t(a)(\frac{\mathbf{1}\{A_t = a, D_t \le T-t\}}{\pi_t(a)} - \mathbb{P}_a\{D \le T-t\})$$

is a martingale difference sequence and the assumption (4.6) and (4.7),

$$\widehat{p}(a) - p(a) = \frac{\sum_{t=1}^T h_t(a)(\frac{\mathbf{1}\{A_t=a, D_t \le T-t\}}{\pi_t(a)} - \mathbb{P}_a\{D \le T-t\})}{\sum_{t\in[T]} h_t(a)} \tag{1.13}$$

$$+ \frac{\sum_{t\in[T]} h_t(a)\mathbb{P}_a\{T-t < D < \infty\}}{\sum_{t\in[T]} h_t(a)}$$

$$= O_{\mathbb{P}}\left(\frac{[\sum_{t=1}^T \mathbb{E}\{h_t(a)^2\gamma_t(a)^2\}]^{1/2}}{\sum_{t\in[T]} h_t(a)}\right) = o_{\mathbb{P}}(1). \tag{1.14}$$

Step (ii). We show that

$$\widehat{Q}_T(a) - \mu(a) \xrightarrow{\mathbb{P}} 0.$$

We have

$$\widehat{Q}_T(a) - \mu(a) =$$
$$\frac{\sum_{t\in[T]} h_t(a)\{(Y_t(a) - \mu(a) + \mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\}}{\sum_{t=1}^T h_t(a)\gamma_t(a)}$$
$$+\frac{\sum_{t\in[T]} h_t(a)(\widehat{\mu}_t(a) - \mu_\infty(a))}{\sum_{t\in[T]} h_t(a)}.$$

Introduce the notation

$$\gamma_t(a) = \frac{\mathbf{1}\{A_t = a, D_t \le T-t\}}{\pi_t(a)},$$
$$\Delta_t(a) = (Y_t(a) - \mu(a) + \mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a).$$

Then

$$\frac{\sum_{t\in[T]} h_t(a)}{p(a)^{-1}} \cdot \frac{\widehat{Q}_T(a) - \mu(a)}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}} \tag{1.15}$$

$$= \frac{\sum_{t\in[T]} h_t(a)\{\widehat{p}(a)^{-1}\Delta_t(a) + \widehat{\mu}_t(a) - \mu_\infty(a)\}}{p(a)^{-1}[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}} \tag{1.16}$$

$$= \frac{\widehat{p}(a)^{-1}}{p(a)^{-1}} \underbrace{\frac{\sum_{t\in[T]} h_t(a)\{\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}}}_{\text{Part I}} \tag{1.17}$$

$$+ \frac{\widehat{p}(a)^{-1}}{p(a)^{-1}} \underbrace{\frac{\sum_{t\in[T]} h_t(a)\{\mathbb{E}_{t-1}\{\Delta_t(a)\} + (\widehat{\mu}_t(a) - \mu_\infty(a))\widehat{p}(a)\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}}}_{\text{Part II}} \tag{1.18}$$

where $\mathbb{E}_{t-1}\{\Delta_t(a)\} = (\mu_\infty(a) - \widehat{\mu}_t(a))\mathbb{P}_a\{D \le T-t\}$.

It's clear that Part I $= O_{\mathbb{P}}(1)$. Now we show Part II $= o_{\mathbb{P}}(1)$.

For Part II, we have

$$\begin{aligned}
&\frac{\sum_{t\in[T]} h_t(a)\{\mathbb{E}_{t-1}\{\Delta_t(a)\} + (\widehat{\mu}_t(a) - \mu_\infty(a))\widehat{p}(a)\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}} \\
=&\frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))(\widehat{p}(a) - \mathbb{P}_a\{D \le T-t\})\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}} \\
=&\frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))(\widehat{p}(a) - p(a))\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}} \\
+&\frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))(p(a) - \mathbb{P}_a\{D \le T-t\})\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}}.
\end{aligned}$$

For the denominator, we first compute

$$\mathbb{E}_{t-1}\left\{|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}|^2\right\} = \mathrm{Var}_{t-1}\{\Delta_t(a)\}.$$

Now using the law of total variance,

$$\mathrm{Var}_{t-1}\{\Delta_t(a)\} = \mathbb{E}_{t-1}\{\mathrm{Var}_{t-1}\{\Delta_t(a) \mid A_t, D_t\}\} + \mathrm{Var}_{t-1}\{\mathbb{E}_{t-1}\{\Delta_t(a) \mid A_t, D_t\}\} = \mathrm{III} + \mathrm{IV}.$$

For III,

$$\begin{aligned}
\mathrm{III} &= \mathbb{E}_{t-1}\left\{\gamma_t(a)^2 \mathrm{Var}_{t-1}\{Y_t(a) - \mu(a) + \mu_\infty(a) - \widehat{\mu}_t(a) \mid A_t, D_t\}\right\} \\
&= \mathbb{E}_{t-1}\left\{\gamma_t(a)^2 \mathrm{Var}_{t-1}\{Y_t(a)\}\right\} = \mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\}.
\end{aligned}$$

For IV,

$$\begin{aligned}
\mathrm{IV} &= \mathrm{Var}_{t-1}\{\mathbb{E}_{t-1}\{\Delta_t(a) \mid A_t, D_t\}\} \\
&= \mathrm{Var}_{t-1}\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{t-1}\left\{|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}|^2\right\} && (1.19)\\
=&\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\} + \mathrm{Var}_{t-1}\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\} && (1.20)\\
\ge&\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\}. && (1.21)
\end{aligned}$$

Hence,

$$\sum_{t\in[T]} \mathbb{E}\left\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\right\} \ge \sum_{t=1}^{T} \mathbb{E}\left\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\right\},$$

which gives

$$\begin{aligned}
\text{Part II} \le &\underbrace{\frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))(\widehat{p}(a) - p(a))\}}{[\sum_{t=1}^{T} \mathbb{E}\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\}]^{1/2}}}_{\text{Term I}} \\
&+ \underbrace{\frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))(p(a) - \mathbb{P}_a\{D \le T-t\})\}}{[\sum_{t=1}^{T} \mathbb{E}\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\}]^{1/2}}}_{\text{Term II}}.
\end{aligned}$$

For Term I, using (1.14), assumption (4.5) and Lemma A.3, we have

$$\text{Term I} = \frac{\sum_{t\in[T]} h_t(a)\{(\widehat{\mu}_t(a) - \mu_\infty(a))\}}{\sum_{t\in[T]} h_t(a)} \frac{(\widehat{p}(a) - p(a))}{[\sum_{t=1}^{T} \mathbb{E}\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\}]^{1/2}/(\sum_{t\in[T]} h_t(a))} = o_{\mathbb{P}}(1).$$

For Term II, using the Lyapunov condition (4.8), we have

$$\begin{aligned}
&\left|\frac{\max_{t\in[T]} h_t(a)^2\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}}{\mathbb{E}\left\{\sum_{t\in[T]} h_t(a)^2\gamma_t(a)^2\right\}}\right|^{\frac{2+\delta}{2}} \\
\leq &\left|\frac{\max_{t\in[T]} h_t(a)^{2+\delta}\mathbb{E}_{t-1}\left\{\gamma_t(a)^{2+\delta}\right\}}{\left(\sum_{t\in[T]}\mathbb{E}\left\{h_t(a)^2\gamma_t(a)^2\right\}\right)^{\frac{2+\delta}{2}}}\right| \\
\leq &\left|\frac{\sum_{t\in[T]} h_t(a)^{2+\delta}\mathbb{E}_{t-1}\left\{\gamma_t(a)^{2+\delta}\right\}}{\left(\sum_{t\in[T]}\mathbb{E}\left\{h_t(a)^2\gamma_t(a)^2\right\}\right)^{\frac{2+\delta}{2}}}\right| \xrightarrow{\mathbb{P}} 0 \text{ (by the Lyapunov Condition).}
\end{aligned} \tag{1.22}$$

using (4.6), (1.22) and Lemma A.3, we can show Term II $= o_{\mathbb{P}}(1)$.

Therefore, Part II vanishes in probability.

Now we combine the above results and conclude

$$\widehat{Q}_T(a) - \mu(a) = O_{\mathbb{P}}\left(\frac{[\sum_{t=1}^T \mathbb{E}\left\{h_t(a)^2\gamma_t(a)^2\right\}]^{1/2}}{\sum_{t\in[T]} h_t(a)}\right) = o_{\mathbb{P}}(1).$$

□

A.3. Proof of Theorem 4.2

Proof of Theorem 4.2. Single arm case: we hope to show

$$\frac{\widehat{Q}_T(a) - \mu(a)}{V_T(a)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$V_T(a) = p(a)^{-2}\sum_{t\in[T]}\mathbb{E}\left\{h_t(a)^2\pi_t(a)^{-1}\sigma_a^2\mathbb{P}_a\left\{D \leq T-t\right\}\right\}.$$

We start from single-arm cases. Recall the two parts decomposition (1.15). We have shown that Part II converge to zero in probability under the given assumptions.

We show that: Part I is a martingale sequence and converge to $\mathcal{N}(0,1)$.

Step I: show that ξ_t is a martingale difference sequence.

Define

$$\xi_{t,T}(a) = \frac{h_t(a)\{\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}\}}{[\sum_{t\in[T]}\mathbb{E}\{h_t(a)^2(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\})^2\}]^{1/2}}.$$

It is not hard to check that the sequence of $\{\xi_t(a), \sigma(H_{t-1})\}_{t=1}^T$ forms a martingale difference sequence.

Step II: Use martingale CLT to prove asymptotic normality.

The crucial step is to verify the Lyapunov condition and the variance convergence condition.

(II-1) Lyapunov condition. We compute

$$\sum_{t=1}^T \mathbb{E}_{t-1}\left\{|\xi_{t,T}(a)|^{2+\delta}\right\} = \frac{\sum_{t=1}^T h_t(a)^{2+\delta}\mathbb{E}_{t-1}\left\{|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}|^{2+\delta}\right\}}{\left[\sum_{t=1}^T \mathbb{E}\left\{h_t(a)^2\left(\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}\right)^2\right\}\right]^{(2+\delta)/2}}. \tag{1.23}$$

(1) For the numerator, we have

$$\|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t\}(a)\|_{L_{t-1}^{2+\delta}} \le \|\Delta_t(a)\|_{L_{t-1}^{2+\delta}} + \|\mathbb{E}_{t-1}\{\Delta_t(a)\}\|_{L_{t-1}^{2+\delta}}. \text{ (By Minkowski's inequality)} \tag{1.24}$$

By Jensen's inequality, we have

$$\|\mathbb{E}_{t-1}\{\Delta_t(a)\}\|_{L_{t-1}^{2+\delta}} \le \mathbb{E}_{t-1}\left\{\|\Delta_t(a)\|_{L_{t-1}^{2+\delta}}\right\} = \|\Delta_t(a)\|_{L_{t-1}^{2+\delta}}.$$

Hence

$$\begin{aligned}
&\|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}\|_{L_{t-1}^{2+\delta}} && (1.25)\\
&\le 2\|\Delta_t(a)\|_{L_{t-1}^{2+\delta}} = 2\|(Y(a) - \mu(a) + \mu_\infty(a) - \widehat{\mu}_t)\gamma_t(a)\|_{L_{t-1}^{2+\delta}} && (1.26)\\
&\le 2\{\underbrace{\|(Y_t(a) - \mu(a))\gamma_t(a)\|_{L_{t-1}^{2+\delta}}}_{\text{Term I}} + \underbrace{\|(\widehat{\mu}_t(a) - \mu_\infty(a))\gamma_t(a)\|_{L_{t-1}^{2+\delta}}}_{\text{Term II}}\}. && (1.27)
\end{aligned}$$

For Term I in (1.27), we have

$$\begin{aligned}
&\|(Y_t(a) - \mu(a))\gamma_t(a)\|_{L_{t-1}^{2+\delta}}^{2+\delta} && (1.28)\\
=&\mathbb{E}_{t-1}\left\{|(Y_t(a) - \mu(a))\gamma_t(a)|^{2+\delta}\right\} && (1.29)\\
=&\mathbb{E}_{t-1}\left\{|\gamma_t(a)|^{2+\delta}\mathbb{E}\left\{|(Y_t(a) - \mu(a))|^{2+\delta}\,\middle|\, H_{t-1}, A_t, D_t\right\}\right\} && (1.30)\\
=&\mathbb{E}_{t-1}\left\{|\gamma_t(a)|^{2+\delta}\mathbb{E}\left\{|(Y_t(a) - \mu(a))|^{2+\delta}\right\}\right\} && (1.31)\\
\le&M_{2+\delta}\mathbb{E}_{t-1}\left\{|\gamma_t(a)|^{2+\delta}\right\}. && (1.32)
\end{aligned}$$

For Term II in (1.27), because we assumed $|\widehat{\mu}(a)|$ is uniformly bounded by some M_μ, Term II is also bounded (up to a constant) by $\mathbb{E}_{t-1}\left\{|\gamma_{t,i}(a)|^{2+\delta}\right\}$. Therefore,

$$\mathbb{E}_{t-1}\left\{|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}|^{2+\delta}\right\} \le M\mathbb{E}_{t-1}\left\{|\gamma_t(a)|^{2+\delta}\right\}.$$

(2) For the denominator, using the variance decomposition (1.20),

$$\begin{aligned}
&\mathbb{E}_{t-1}\left\{|\Delta_t(a) - \mathbb{E}_{t-1}\{\Delta_t(a)\}|^2\right\}\\
=&\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\} + \mathrm{Var}_{t-1}\left\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\right\}\\
\ge&\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\}.
\end{aligned}$$

Therefore, for (1.23),

$$\sum_{t=1}^{T}\mathbb{E}_{t-1}\left\{|\xi_{t,T}(a)|^{2+\delta}\right\} \le \frac{M\sum_{t=1}^{T}h_t(a)^{2+\delta}\mathbb{E}_{t-1}\left\{|\gamma_t(a)|^{2+\delta}\right\}}{\left(\sum_{t=1}^{T}\mathbb{E}\left\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\right\}\right)^{\frac{2+\delta}{2}}} \xrightarrow{\mathbb{P}} 0. \tag{1.33}$$

(II-2) Variance convergence. Now we check variance convergence. Combining (1.20), we need to show

$$\sum_{t=1}^{T}\mathbb{E}_{t-1}\left\{\xi_{t,T}(a)^2\right\} = \frac{\sum_{t=1}^{T}h_t(a)^2\left(\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\} + \mathrm{Var}_{t-1}\left\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\right\}\right)}{\mathbb{E}\left(\sum_{t=1}^{T}h_t(a)^2\left(\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\} + \mathrm{Var}_{t-1}\left\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\right\}\right)\right)} \xrightarrow{\mathbb{P}} 1. \tag{1.34}$$

Define

$$\begin{aligned} Z_T &= \sum_{t=1}^{T} h_t(a)^2 \Bigg(\mathbb{E}_{t-1}\left\{\sigma_a^2 \gamma_t(a)^2\right\} \\ &\quad + \operatorname{Var}_{t-1}\left\{(\mu_\infty(a) - \widehat{\mu}_t(a))\gamma_t(a)\right\} \Bigg) \\ &= \underbrace{\sum_{t=1}^{T} h_t(a)^2 \mathbb{E}_{t-1}\left\{\sigma_a^2 \gamma_t(a)^2\right\}}_{\text{Term I}} \\ &\quad + \underbrace{\sum_{t=1}^{T} h_t(a)^2 \operatorname{Var}_{t-1}\left\{(\widehat{\mu}_t(a) - \mu_\infty(a))\gamma_t(a)\right\}}_{\text{Term II}}. \end{aligned}$$

For Term I, according to the assumption (4.9), we have

$$\frac{\sum_{t=1}^{T} h_t(a)^2 \mathbb{E}_{t-1}\left\{\sigma_a^2 \gamma_t(a)^2\right\}}{\sum_{t=1}^{T} \mathbb{E}\left\{h_t(a)^2 \sigma_a^2 \gamma_t(a)^2\right\}} \xrightarrow{L^p} 1.$$

For Term II, by Hölder's inequality,

$$|\text{Term II}| \le \sum_{t=1}^{T} h_t(a)^2 \mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\} |\widehat{\mu}_t(a) - \mu_\infty(a)|^2.$$

By assumption, $|\widehat{\mu}_t(a) - \mu_\infty(a)| \xrightarrow{\mathbb{P}} 0$. Besides, based on (1.22), the following weights are negligible:

$$\left| \frac{\max_{t\in[T]} h_t(a)^2 \mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}}{\mathbb{E}\left\{\sum_{t\in[T]} h_t^2 \gamma_t(a)^2\right\}} \right| \to 0.$$

Therefore by Lemma A.3, we can conclude

$$\frac{|\text{Term II}|}{\sum_{t=1}^{T} \mathbb{E}\left\{h_t(a)^2 \sigma_a^2 \gamma_t(a)^2\right\}} = o_{\mathbb{P}}(1). \tag{1.35}$$

Combining the fact that

$$\frac{|\text{Term II}|}{\sum_{t=1}^{T} \mathbb{E}\left\{h_t(a)^2 \sigma_a^2 \gamma_t(a)^2\right\}} \le \frac{2M_\mu^2}{\sigma_a^2},$$

(1.35) can be strengthened into L^1-convergence:

$$\frac{\mathbb{E}\left\{|\text{Term II}|\right\}}{\sum_{t=1}^{T} \mathbb{E}\left\{h_t(a)^2 \sigma_a^2 \gamma_t(a)^2\right\}} \to 0.$$

Therefore,

$$\frac{\text{Term I} + \text{Term II}}{\mathbb{E}\left\{\text{Term I} + \text{Term II}\right\}} \xrightarrow{\mathbb{P}} 0.$$

Combining (II-1) and (II-2), we have shown that

$$\sum_{t\in[T]} \xi_t(a) \xrightarrow{d} \mathcal{N}(0,1).$$

Step III: Justify joint asymptotic normality using Cramér-Wold device.

For simplicity we consider two arms. The generalization to a finite number of arms is rather straightforward. Choose a vector $(b_1, b_2)^\top \in \mathbb{R}^2$ with $b_1^2 + b_2^2 = 1$. It suffices to show that under the given conditions,

$$b_1 \xi_{t,T}(a) + b_2 \xi_{t,T}(a') \xrightarrow{d} \mathcal{N}(0,1).$$

The Lyapunov condition is easy to check because we know

$$\begin{aligned}
&|b_1 \xi_{t,T}(a) + b_2 \xi_{t,T}(a')|^{2+\delta} \\
\leq& 2^{1+\delta}(|b_1|^{2+\delta}|\xi_{t,T}(a)|^{2+\delta} + |b_2|^{2+\delta}|\xi_{t,T}(a')|^{2+\delta}).
\end{aligned}$$

Therefore, the Lyapunov condition follows naturally from the results for individual arms.

The tricky part is to check variance convergence. We briefly go through the idea of proof.

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{t-1}\left\{(b_1 \xi_{t,T}(a) + b_2 \xi_{t,T}(a'))^2\right\} \\
=& \sum_{t=1}^{T} b_1^2 \mathbb{E}_{t-1}\left\{\xi_{t,T}(a)^2\right\} + \sum_{t=1}^{T} b_2^2 \mathbb{E}_{t-1}\left\{\xi_{t,T}(a')^2\right\} \\
+& \sum_{t=1}^{T} b_1 b_2 \mathbb{E}_{t-1}\left\{\xi_t(a)\xi_t(a')\right\}.
\end{aligned}$$

According to the proofs in Step 2, (II-2),

$$\mathbb{E}_{t-1}\left\{\xi_{t,T}(a)^2\right\} \xrightarrow{\mathbb{P}} 1, \quad \mathbb{E}_{t-1}\left\{\xi_{t,T}(a')^2\right\} \xrightarrow{\mathbb{P}} 1.$$

Now we show

$$\sum_{t=1}^{T} \mathbb{E}_{t-1}\left\{\xi_{t,T}(a)\xi_{t,T}(a')\right\} \xrightarrow{\mathbb{P}} 0.$$

Write down the expression explicitly,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{t-1}\left\{\xi_{t,T}(a)\xi_{t,T}(a')\right\} \\
=& \frac{\sum_{t\in[T]} h_t(a)h_t(a') \operatorname{Cov}_{t-1}\left\{\Delta_t(a), \Delta_t(a')\right\}}{[\sum_{t\in[T]} \mathbb{E}\{h_t(a)^2 \operatorname{Var}_{t-1}\left\{\Delta_t(a)\right\}\}]^{1/2} \cdot [\sum_{t\in[T]} \mathbb{E}\{h_t(a')^2 \operatorname{Var}_{t-1}\left\{\Delta_t(a')\right\}\}]^{1/2}}.
\end{aligned}$$

For the numerator,

$$\begin{aligned}
&\sum_{t\in[T]} h_t(a)h_t(a') \operatorname{Cov}_{t-1}\left\{\Delta_t(a), \Delta_t(a')\right\} \\
=& \sum_{t\in[T]} h_t(a)h_t(a')(\mathbb{E}_{t-1}\left\{\Delta_t(a)\Delta_t(a')\right\} - \mathbb{E}_{t-1}\left\{\Delta_t(a)\right\}\mathbb{E}_{t-1}\left\{\Delta_t(a')\right\}) \\
=& -\sum_{t\in[T]} h_t(a)h_t(a')\mathbb{E}_{t-1}\left\{\Delta_t(a)\right\}\mathbb{E}_{t-1}\left\{\Delta_t(a')\right\} \\
& \text{(because } \gamma_t(a)\gamma_t(a') = 0\text{)}.
\end{aligned}$$

Therefore, by Cauchy-Schwarz inequality and the variance decomposition (1.20),

$$\begin{aligned}
&\left|\sum_{t=1}^{T}\mathbb{E}_{t-1}\left\{\xi_t(a)\xi_t(a')\right\}\right| \\
\leq &\left[\frac{\sum_{t\in[T]}h_t(a)^2(\mathbb{E}_{t-1}\left\{\Delta_t(a)\right\})^2}{\sum_{t\in[T]}\mathbb{E}\{h_t(a)^2\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\}\}}\right]^{1/2} \\
&\cdot\left[\frac{\sum_{t\in[T]}h_t(a')^2(\mathbb{E}_{t-1}\left\{\Delta_t(a')\right\})^2}{\sum_{t\in[T]}\mathbb{E}\{h_t(a')^2\mathbb{E}_{t-1}\left\{\sigma_{a'}^2\gamma_t(a')^2\right\}\}}\right]^{1/2}.
\end{aligned}$$

Now we have

$$\begin{aligned}
&\sum_{t\in[T]}h_t(a)^2(\mathbb{E}_{t-1}\left\{\Delta_t(a)\right\})^2 \\
\leq &\sum_{t\in[T]}h_t(a)^2\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}\frac{(\mathbb{E}_{t-1}\left\{\Delta_t(a)\right\})^2}{\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}} \\
\leq &\sum_{t\in[T]}h_t(a)^2\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}\frac{(\mathbb{E}_{t-1}\left\{(\mu_\infty(a)-\widehat{\mu}_t(a))\gamma_t(a)\right\})^2}{\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}} \\
\leq &\sum_{t\in[T]}h_t(a)^2\mathbb{E}_{t-1}\left\{\gamma_t(a)^2\right\}(\mu_\infty(a)-\widehat{\mu}_t(a))^2.
\end{aligned}$$

Under the condition $\widehat{\mu}(a)\xrightarrow{\mathbb{P}}\mu_\infty(a)$, the variance convergence assumption (4.9), negligible weights result (1.22) and Lemma A.3, we can conclude:

$$\frac{\sum_{t\in[T]}h_t(a)^2(\mathbb{E}_{t-1}\left\{\Delta_t(a)\right\})^2}{\sum_{t\in[T]}\mathbb{E}\{h_t(a)^2\mathbb{E}_{t-1}\left\{\sigma_a^2\gamma_t(a)^2\right\}\}}\xrightarrow{\mathbb{P}}0. \tag{1.36}$$

□

A.4. Proof of Theorem 4.3

Proof of Theorem 4.3. We mainly need to show

$$\frac{\sum_{t\in[T]}h_t(a)^2\left\{(Y_t-\widehat{Q}_T(a))\gamma_t(a)\right\}^2}{\sum_{t\in[T]}\mathbb{E}\left\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\right\}}\xrightarrow{\mathbb{P}}1. \tag{1.37}$$

We have the decomposition

$$\begin{aligned}
&\sum_{t\in[T]}h_t(a)^2\left\{(Y_t-\widehat{Q}_T(a))\gamma_t(a)\right\}^2 \\
=&\underbrace{\sum_{t\in[T]}h_t(a)^2\left\{(Y_t(a)-\mu(a))\gamma_t(a)\right\}^2}_{\text{Term I}} \\
&+\underbrace{\sum_{t\in[T]}h_t(a)^2\left\{(\mu(a)-\widehat{Q}_T(a))\gamma_t(a)\right\}^2}_{\text{Term II}} \\
&+2\underbrace{\sum_{t\in[T]}h_t(a)^2\left\{(Y_t(a)-\mu(a))(\mu(a)-\widehat{Q}_T(a))\gamma_t(a)^2\right\}}_{\text{Term III}}.
\end{aligned}$$

Term I: Term I is a square variation of the martingale difference sequence

$$\phi_{t,T}(a) = \frac{h_t(a)(Y_t(a) - \mu(a))\gamma_t(a)}{\left[\mathbb{E}\left\{\sum_{t\in[T]} \phi_{t,T}^2\right\}\right]^{1/2}}.$$

By the variance convergence assumption (4.9), the conditional square variation satisfies:

$$\sum_{t\in[T]} \mathbb{E}_{t-1}\left\{\phi_{t,T}^2\right\} \xrightarrow{\mathbb{P}} 1.$$

By the Lyapunov condition (4.8) and Lemma A.2, we have

$$\sum_{t\in[T]} \phi_{t,T}^2 \xrightarrow{\mathbb{P}} 1.$$

Term II: for Term II, we have

$$\begin{aligned}
&\frac{\sum_{t\in[T]} h_t(a)^2 \left\{(\mu(a) - \widehat{Q}_T(a))\gamma_t(a)\right\}^2}{\sum_{t\in[T]} \mathbb{E}\left\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\right\}} \\
=&(\mu(a) - \widehat{Q}_T(a))^2 \frac{\sum_{t\in[T]} h_t(a)^2 \left\{\gamma_t(a)\right\}^2}{\sum_{t\in[T]} \mathbb{E}\left\{h_t(a)^2\sigma_a^2\gamma_t(a)^2\right\}} \\
=&o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1).
\end{aligned}$$

Term III: for Term III, using Cauchy-Schwarz inequality, we have

$$\text{Term III} \leq (\text{Term I})^{1/2}(\text{Term II})^{1/2}.$$

Therefore, based on the results on Term I and Term II, it's not hard to show Term III $= o_{\mathbb{P}}(1)$.

Now we can combine all parts above to conclude the proof.

□

A.5. Proof of Corollary 4.1

Proof of Corollary 4.1. We check (A1) to ((A5)).

- For (A1), we have

$$\frac{\max_{t\in[T]} h_t(a)}{\sum_{t=1}^T h_t(a)} \leq \frac{1}{C\sum_{t\in[T]} t^{-\alpha/2}} \asymp \frac{1}{T^{1-\alpha/2}} \to 0.$$

- For (A2), we have

$$\begin{aligned}
&\frac{\sum_{t\in[T]} h_t(a)\mathbb{P}_a\left\{T - t < D < \infty\right\}}{[\sum_{t=1}^T \mathbb{E}\left\{h_t^2\pi_t(a)^{-1}\mathbb{P}_a\left\{D \leq T - t\right\}\right\}]^{1/2}} \\
=&\frac{\sum_{t\in[T]} h_t(a)\mathbb{P}_a\left\{T - t < D < \infty\right\}}{[\sum_{t=1}^T \mathbb{P}_a\left\{D \leq T - t\right\}]^{1/2}}.
\end{aligned}$$

  For the denominator, we have

$$\begin{aligned}
\sum_{t=1}^T \mathbb{P}_a\left\{D \leq T - t\right\} &= p(a)T - \sum_{t\in[T]} \mathbb{P}_a\left\{T - t < D < \infty\right\} \\
&\asymp T - \max\{T^{1-\beta}, 1\} \asymp T.
\end{aligned}$$

For the numerator, in general we have

$$\sum_{t\in[T]} h_t(a)\mathbb{P}_a\left\{T-t<D<\infty\right\}$$
$$\leq \sum_{t\in[T]} \mathbb{P}_a\left\{T-t<D<\infty\right\} \asymp T^{1-\beta}.$$

Hence we need $\beta \geq 1/2$. If $\pi_t(a) \asymp t^{-\alpha}$, then the numerator can be bounded by

$$\sum_{t\in[T]} h_t(a)\mathbb{P}_a\left\{T-t<D<\infty\right\}$$
$$\leq \sum_{t\in[T]} t^{-\frac{\alpha}{2}}(T-t+1)^{-\beta}$$
$$= T^{1-\frac{\alpha}{2}-\beta} \sum_{t\in[T]} (\frac{t}{T})^{-\frac{\alpha}{2}} (\frac{T-t+1}{T})^{-\beta} \cdot \frac{1}{T}.$$

When $\beta < 1/2$,

$$\sum_{t\in[T]} (\frac{t}{T})^{-\frac{\alpha}{2}} (\frac{T-t+1}{T})^{-\beta} \cdot \frac{1}{T}$$
$$\to \int_0^1 x^{-\frac{\alpha}{2}}(1-x)^{-\beta}dx = \mathrm{Beta}(1-\frac{\alpha}{2}, 1-\beta).$$

Hence we can allow $\frac{\alpha}{2}+\beta \geq \frac{1}{2}$.

- For (A3), we have

$$\frac{\mathbb{E}\left\{\sum_{t=1}^T h_t(a)^2\pi_t(a)^{-1}\mathbb{P}_{a-1}\left\{D_t \leq T-t\right\}\right\}}{\left(\sum_{t=1}^T h_t(a)\right)^2} \lesssim \frac{T}{T^{2-\alpha}} \asymp T^{-(1-\alpha)} \to 0.$$

- For (A4), the numerator satisfies

$$\sum_{t\in[T]} h_t^{2+\delta}\pi_t(a)^{-(1+\delta)}\mathbb{P}_a\left\{D \leq T-t\right\}$$
$$\leq \sum_{t\in[T]} t^{\frac{\alpha\delta}{2}} \asymp T^{1+\frac{\alpha\delta}{2}}.$$

For the denominator,

$$\left(\sum_{t=1}^T \mathbb{E}\left\{h_t^2\pi_t(a)^{-1}\mathbb{P}_a\left\{D \leq T-t\right\}\right\}\right)^{\frac{2+\delta}{2}} \asymp T^{1+\frac{\delta}{2}}.$$

Therefore,

$$\frac{\sum_{t\in[T]} h_t^{2+\delta}\pi_t(a)^{-(1+\delta)}\mathbb{P}_a\left\{D \leq T-t\right\}}{\left(\sum_{t=1}^T \mathbb{E}\left\{h_t^2\pi_t(a)^{-1}\mathbb{P}_a\left\{D \leq T-t\right\}\right\}\right)^{\frac{2+\delta}{2}}}$$
$$\leq \sum_{t\in[T]} t^{\frac{\alpha\delta}{2}} \asymp T^{\frac{-(1-\alpha)\delta}{2}} \to 0.$$

- For (A5), it is easy to check that with $h_t(a) = \sqrt{\pi_t(a)}$, we always have

$$\frac{\sum_{t=1}^T h_t^2\pi_t(a)^{-1}\mathbb{P}_a\left\{D \leq T-t\right\}}{\sum_{t=1}^T \mathbb{E}\left\{h_t^2\pi_t(a)^{-1}\mathbb{P}_a\left\{D \leq T-t\right\}\right\}} = 1.$$

□

B. Additional numerical experiments

B.1. Simulation results under different delay mechanisms

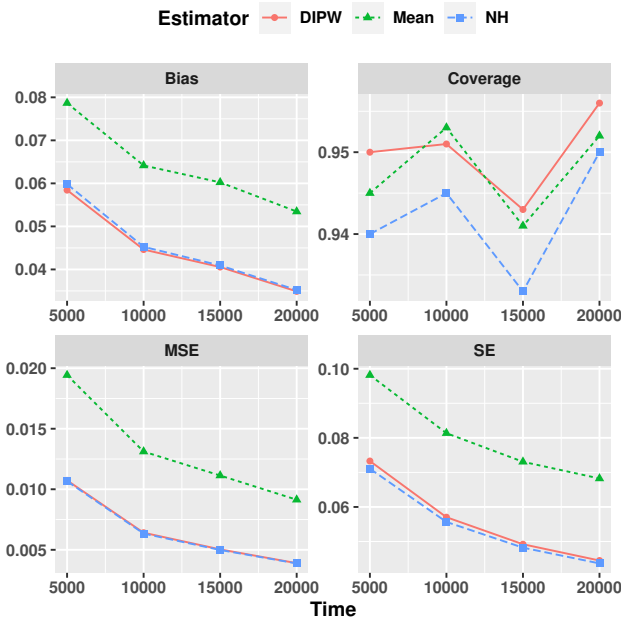

Figure 5. Evaluation of ϵ-greedy with Negative Binomial delays

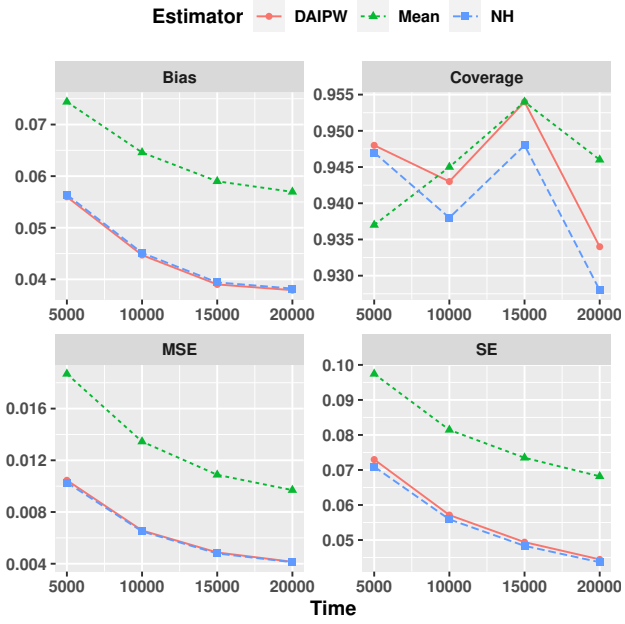

Figure 6. Evaluation of ϵ-greedy with PARETO delays

B.2. Simulation results with different margins in ϵ-greedy algorithms

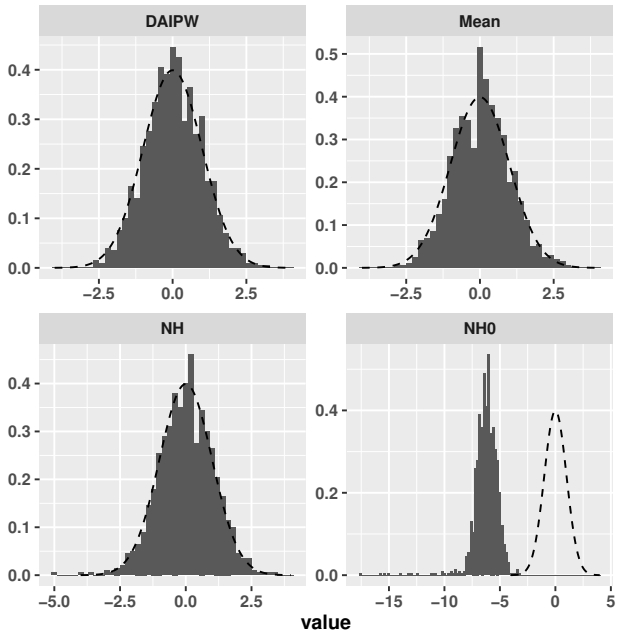

Figure 7. ϵ-greedy over weak margin bandits $\mu(1) - \mu(2) = 0.1$.

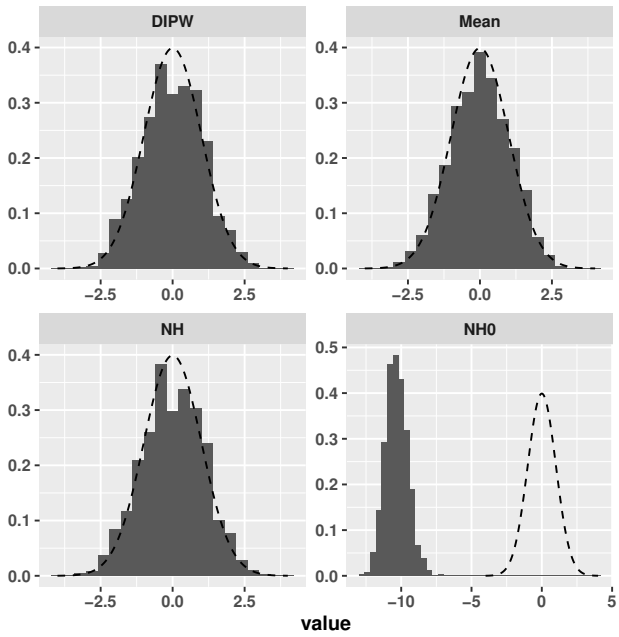

Figure 8. ϵ-greedy over strong margin bandits $\mu(1) - \mu(2) = 0.5$.