Statistical Inference with M-Estimators on Bandit Data

Kelly W. Zhang KELLYWZHANG@SEAS.HARVARD.EDU
Lucas Janson LJANSON@FAS.HARVARD.EDU
Susan A. Murphy SAMURPHY@FAS.HARVARD.EDU
Harvard University, Cambridge, MA 02139, USA

Abstract

Bandit algorithms are increasingly used in real world sequential decision making problems, from online advertising to mobile health. As a result, there are more datasets collected using bandit algorithms and with that an increased desire to be able to use these datasets to answer scientific questions like: Did one type of ad increase the click-through rate more or lead to more purchases? In which contexts is a mobile health intervention effective? However, it has been shown that classical statistical approaches, like those based on the ordinary least squares estimator, fail to provide reliable confidence intervals when used with bandit data. Recently methods have been developed to conduct statistical inference using simple models fit to data collected with multi-armed bandits. However there is a lack of general methods for conducting statistical inference using more complex models. In this work, we develop theory justifying the use of M-estimation (Van der Vaart, 2000), traditionally used with i.i.d data, to provide inferential methods for a large class of estimators—including least squares and maximum likelihood estimators—but now with data collected with (contextual) bandit algorithms. To do this we generalize the use of adaptive weights pioneered by Hadad et al. (2019) and Deshpande et al. (2018). Specifically, in settings in which the data is collected via a (contextual) bandit algorithm, we prove that certain adaptively weighted M-estimators are uniformly asymptotically normal and demonstrate empirically that we can use their asymptotic distribution to construct reliable confidence regions for a variety of inferential targets.

Keywords: contextual bandits, statistical inference, adaptively collected data

Contents

1. Introduction

Due to the need for interventions that are personalized to users, bandit algorithms are increasingly used to address sequential decision making problems in health-care, online education, and public policy (Caria et al., 2020; Kasy and Sautmann, 2019). For example, in mobile health, bandit algorithms are used to determine when interventions should be delivered to help users manage chronic health problems like obesity (Klasnja et al., 2020), addiction (Barnett et al., 2020; Song et al., 2019; Bell et al., 2020) and mental illness (Neary and Schueller, 2018; Lipschitz et al., 2019). Bandit algorithms are used to personalize teaching strategies to each student in online education problems (Liu et al., 2014; Rafferty et al., 2019; Shaikh et al., 2019) and to learn when to send useful reminders to defendants about their upcoming court dates[1]. In the contextual bandit framework, the algorithm personalizes by learning to choose the best interventions in each context, i.e., the action that leads to the greatest expected reward. How well a bandit algorithm personalizes is measured by its regret, which is the total expected reward under that algorithm compared to the total expected reward of an oracle algorithm.

Besides the goal of regret minimization and personalizing interventions to each user online, another critical goal in these real world problems is to be able to use the resulting data collected by bandit algorithms to advance scientific knowledge (Liu et al., 2014; Erraqabi et al., 2017). By scientific knowledge, we specifically mean information gained by using the data collected to conduct a variety of machine learning and statistical analyses, including confidence interval construction. **While the personalization or regret-minimizing objective is a *within*-experiment learning objective, gaining scientific knowledge from the resulting adaptively collected data is a *between*-experiment learning objective**, which ultimately helps with regret minimization between deployments of bandit algorithms. Here we refer to deployments of bandit algorithms as "experiments" because using a bandit algorithm is a type of experimental design. Note that the data collected by any reinforcement learning algorithm is *adaptively collected* because previous observations are used to inform what actions to select in future timesteps.

While in theoretical problems it is generally assumed that all interesting outcomes in bandit experiments are subsumed in the reward, in real life sequential decision making problems there are often a variety of additional scientifically interesting outcomes that are collected during the bandit experiment. For example, in the online advertising setting, the reward might be whether an ad is clicked on, but one may be interested in the outcome of amount of money spent or the subsequent time spent on the advertiser's website. If it was found an ad had high click-through rate, but low amounts of money was spent after clicking on the ad, one may redesign the reward for use in the next bandit experiment. One type of after-study analysis would be to construct confidence intervals for the relative effect of the actions on multiple outcomes (in addition to the reward) conditional on the context. Furthermore, due to engineering and practical limitations, some of the variables that might be useful as context are often not accessible to the bandit algorithm during the bandit experiment. If after-study analyses find some such contextual variables to have sufficiently strong influence on the relative usefulness of an action, this might lead investigators to ensure these variable are accessible to the bandit algorithm in the next experiment. Similarly,

1. Stanford Computational Public Policy Lab: `https://policylab.stanford.edu/projects/nudge.html`

engineering and practical limitations might mean that the action space included in the current bandit experiment does not subsume the entire space of actions of interest. For example in the online education setting, not all interesting teaching strategies might be included as actions due to time constraints in implementing the bandit experiment. If it is found that no teaching strategies perform well in a particular context (context could include student's background, student's recent progress, etc.), it may make sense to develop new teaching strategies that are specifically designed to do better in that context.

As discussed above, we can gain scientific knowledge from bandit data by constructing confidence intervals and hypothesis tests for unknown quantities such as the effect of the actions on outcomes in different contexts. Unfortunately, standard statistical methods developed for independently collected data fail to provide valid inference when applied to bandit data. For example, on data collected with common bandit algorithms, the sample mean of rewards from an arm fails to converge uniformly in distribution, which precludes use of this distribution in constructing uniformly valid confidence intervals (Hadad et al., 2019; Deshpande et al., 2018; Zhang et al., 2020).

Recently, statistical inference methods for use with simple models have been developed for data collected using bandit algorithms (Hadad et al., 2019; Deshpande et al., 2018; Zhang et al., 2020). However, there is a lack of general statistical inference methods for bandit data in more complex data analytic settings, including parameters in non-linear models for outcomes such as the reward. For example, there are currently no methods for constructing uniformly valid confidence intervals for the parameters of a logistic regression model for binary outcomes or for constructing confidence intervals based on robust estimators like the median or minimizers of the Huber loss function on bandit data.

In this work we show that a wide variety of estimators which are frequently used both in science and industry on independently sampled data, namely, M-estimators (Van der Vaart, 2000), can be used to conduct valid inference on bandit data when adjusted with adaptive weights, i.e., weights that are a function of previously collected data. The use of adaptive weights is common throughout the previously developed methods for conducting statistical inference on bandit data for simple models (Deshpande et al., 2018; Hadad et al., 2019; Zhang et al., 2020). The adaptive weights we use most closely resemble those of Hadad et al. (2019). Hadad et al. (2019) use adaptive weights to stabilize the variance of estimators for both the mean reward for an arm and the difference in mean rewards for two arms. We show that adaptive weights can be used with a much larger class of statistical models to conduct valid inference on bandit data.

An outline of our paper is as follows:

- Section 2 formalizes the problem statement and introduces notation.
- Section 3 introduces adaptively weighted M-estimators and presents how to construct uniformly valid asymptotic confidence intervals with these estimators.
- Section 4 discusses related work.
- Section 5 gives simulation results for weighted versions of common M-estimators on bandit data.

- Section 6 discusses open questions regarding the use of bandit algorithms to collect data when the goal is to both personalize as well as use the resulting data to improve scientific knowledge.

2. Problem Formulation and Notation

We assume that the data we have after running a contextual bandit algorithm is comprised of contexts $\{X_t\}_{t=1}^T$, actions $\{A_t\}_{t=1}^T$, and rewards $\{R_t\}_{t=1}^T$, and outcomes $\{Y_t\}_{t=1}^T$. We assume that rewards are a deterministic function of the outcomes, i.e., $R_t = f(Y_t)$ for some known function f. We are primarily interested in estimating parameters of the data generating distribution of Y_t given (X_t, A_t). T is deterministic. Below we consider $T \to \infty$ in order to derive the asymptotic distributions of estimators and the resulting confidence intervals based on the asymptotic distributions. We use potential outcome notation (Imbens and Rubin, 2015) and assume that the potential outcomes $\{X_t, Y_t(a) : a \in \mathcal{A}\}$ are i.i.d. random variables across $t \in [1\colon T]$. The observed outcomes are given by $Y_t := Y_t(A_t)$ and $R_t := R_t(A_t)$. We define the history $\mathcal{H}_t := \{X_{t'}, A_{t'}, Y_{t'}\}_{t'=1}^t$. Actions $A_t \in \mathcal{A}$ are selected according to action selection probabilities $\pi_{t,a} = P\left(A_t = a | \mathcal{H}_{t-1}, X_t\right)$. We allow the action space $\mathcal{A}$ to be finite or infinite; if the action space is infinite then for each $t \in [1\colon T]$, $\pi_{t,a}$ is a probability density function over $a \in \mathcal{A}$. Note that while we assume i.i.d potential outcomes, since the contextual bandit algorithm uses past data $\mathcal{H}_{t-1}$ to choose action selection probabilities $\pi_{t,a}$, the *observed* data $\{X_t, A_t, Y_t\}$ are *not* independent across $t \in [1\colon T]$.

We are interested in constructing confidence intervals for some unknown $\theta^* \in \Theta \subset \mathbb{R}^d$, which is a parameter of the data generating distribution of Y_t given (X_t, A_t). Specifically, we assume that θ^* is a conditionally maximizing value of criterion m_θ, i.e.,

$$\theta^* \in \operatorname*{argmax}_{\theta \in \Theta} E_{\theta^*}\left[m_\theta(Y_t, X_t, A_t) | X_t, A_t\right] \quad \text{w.p. } 1. \tag{1}$$

Note that θ^* does not depend on X_t and A_t. To estimate θ^*, we use M-estimation (Huber, 1992; Van der Vaart, 2000), which selects the estimator $\hat{\theta}$ to be the $\theta \in \Theta$ that maximizes a criterion function of the following form:

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^T m_\theta(Y_t, X_t, A_t).$$

For example, in a classical linear regression setting with K treatments, a natural choice for m_θ would be the squared loss function:

$$m_\theta(Y_t, X_t, A_t) = -\sum_{k=1}^K \left(Y_t - X_t^\top \theta_{A_t}\right)^2.$$

Other natural choices for m_θ include likelihood functions and other loss functions. In the case that Y_t is binary and $|\mathcal{A}| = K$, we might choose m_θ to be a logistic regression model:

$$m_\theta(Y_t, X_t, A_t) = -\left[Y_t X_t^\top \theta_{A_t} - \log\left(1 + \exp\left(X_t^\top \theta_{A_t}\right)\right)\right].$$

For more details about GLMs see Appendix A.2 and Agresti (2015). When the data, $\{X_t, A_t, Y_t\}_{t=1}^T$, are independently distributed, classical approaches can be used to prove the consistency and asymptotic normality of M-estimators (Van der Vaart, 2000). However standard M-estimators (like the ordinary least squares estimator) fail to be asymptotically normal when used on bandit data and fail to provide valid confidence intervals based on a normal approximation (Hadad et al., 2019; Deshpande et al., 2018; Zhang et al., 2020). In this work, we show that M-estimators can still be used for provide valid statistical inference on bandit data as long as they are adjusted with well-chosen adaptive weights.

3. Adaptively Weighted M-Estimators

3.1 Introducing the Estimator

We consider a weighted M-estimating criteria with adaptive weights $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ given by $W_t = \sqrt{\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}}$. Here $\{\pi_t^{\text{eval}}\}_{t=1}^T \in \sigma(\mathcal{H}_0)$ are evaluation policies of choice and $\pi_{t,a}^{\text{eval}} = P_{\pi_t^{\text{eval}}}(A_t = a | X_t)$. We discuss considerations for the choice of $\{\pi_t^{\text{eval}}\}_{t=1}^T$ in section 3.5, but a default choice for the evaluation policy when the action space is finite and of size K, is just $\pi_{t,a}^{\text{eval}} = 1/K$. We call these weights *square-root importance weights* because they are the square-root of the standard importance weights which are commonly used in reinforcement learning (Thomas and Brunskill, 2016; Jiang and Li, 2016) and Monte Carlo methods (Hammersley, 2013). Our estimator $\hat{\theta}_T$ for θ^* is the maximizer of the weighted M-estimating criterion:

$$\hat{\theta}_T := \operatorname*{argmax}_{\theta \in \Theta} \sum_{t=1}^T W_t m_\theta(Y_t, X_t, A_t) = \operatorname*{argmax}_{\theta \in \Theta} M_T(\theta).$$

Although W_t in this paper will always be square-root importance weights, we note that if instead W_t were set to 1, the above equation would correspond to a standard unweighted M-estimator.

In this work, our goal is to derive asymptotically valid confidence intervals for θ^*. Specifically we will do this by deriving the asymptotic limiting distribution of $\hat{\theta}_T$ as $T \to \infty$ and by demonstrating that the convergence to the limiting distribution is *uniform* over $\theta^* \in \Theta$. Intuitively, uniformity fails if for any sample size, T, there exists a $\theta_T^* \in \Theta$ for which the probability that the $1-\alpha$ confidence interval contains θ_T^* is less than $1-\alpha$. Formally, uniformly asymptotically valid $1-\alpha$ level confidence intervals C_T for θ^* satisfy:

$$\liminf_{T \to \infty} \inf_{\theta^* \in \Theta} P_{\theta^*} \left(\theta^* \in C_T(\alpha) \right) \geq 1 - \alpha. \tag{2}$$

Confidence intervals constructed using asymptotic approximations fail to be reliable when convergences does not hold uniformly over the parameter space. The importance of uniformity for the reliability of confidence intervals in practice is discussed in many areas of statistics including statistical inference via boostrap (Romano et al., 2012), statistical inference after model selection (Leeb and Pötscher, 2005), and econometrics (Kasy, 2019). Note that on i.i.d. data, "uniformity" in convergence typically comes for free, i.e., on i.i.d. data it is generally straightforward to show that estimators that converge in distribution do so uniformly. However, as discussed in Zhang et al. (2020), estimators that converge uniformly on independently collected data can fail to converge uniformly on data collected with bandit algorithms. See Appendix D for information on the importance of uniformity and how uniform convergence of estimators can fail.

To construct uniformly valid confidence intervals for θ^* we prove that $\hat{\theta}_T$ is uniformly asymptotically normal:

$$\Sigma_{\theta^*,T}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\hat{\theta}_T) \right] \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \theta^* \in \Theta,$$

where $\dot{M}_T(\theta) = \frac{\partial}{\partial \theta} M_T(\theta)$, $\ddot{M}_T(\theta) = \frac{\partial^2}{\partial^2 \theta} M_T(\theta)$, and $\Sigma_{\theta^*,T} := \frac{1}{T} \sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}} \left[\dot{m}_{\theta^*}(Y_t, X_t, A_t)^{\otimes 2} \right]$. We define $P_{\theta^*, \pi_t^{\text{eval}}}$ to be the distribution of (Y_t, X_t, A_t) when A_t are selected according to the evaluation policy π_t^{eval}.

3.2 Intuition for Square-Root Importance Weights

To provide some intuition for how weighting M-estimators with square-root importance weights $W_t = \sqrt{\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}}$ ensures uniform asymptotic normality, we consider the weighted least squares estimator as an example. Assume that $E_\theta[Y_t | A_t, X_t] = \phi(X_t, A_t)^\top \theta$ holds with probability 1 and that π_t^{eval} is the same for all $t \in [1 \colon T]$ so that we can drop the t index. We include adaptive weights W_t on the following least squares criterion to estimate θ^*, as defined in equation (1),

$$\hat{\theta}^{\text{AW-LS}} := \operatorname*{argmax}_{\theta \in \Theta} \left\{ -\frac{1}{2} \sum_{t=1}^T W_t \left(Y_t - \phi(X_t, A_t)^\top \theta \right)^2 \right\}. \tag{3}$$

By taking the derivative of the criterion equation (3), we have that

$$0 = \sum_{t=1}^T W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \hat{\theta}^{\text{AW-LS}} \right).$$

By Taylor Series expansion,

$$= \sum_{t=1}^T W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right) - \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}^{\text{AW-LS}} - \theta^* \right).$$

Rearranging terms, we have the following:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}^{\text{AW-LS}} - \theta^* \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right). \tag{4}$$

Note that the right hand side of equation (4) above is a martingale difference sequence with respect to history $\{\mathcal{H}_{t-1}\}_{t=1}^T$, i.e., $E_{\theta^*} \left[W_t \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right) | \mathcal{H}_{t-1} \right] = 0$, since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ and $E_{\theta^*}[Y_t | A_t, X_t] = \phi(X_t, A_t)^\top \theta^*$. We can show that (4) is asymptotically normal by applying a martingale central limit theorem (Dvoretzky, 1972). Specifically for a martingale difference sequence $\{Z_t\}$ with respect to $\{\mathcal{H}_{t-1}\}_{t=1}^T$, $\frac{1}{\sqrt{T}} \sum_{t=1}^T Z_t \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ uniformly over $\theta^* \in \Theta$ under the following two conditions:

- **Conditional Variance:** $\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[Z_t^2 | \mathcal{H}_{t-1}\right] \xrightarrow{P} \sigma_{\theta^*}^2$ uniformly over $\theta^* \in \Theta$ for σ_{θ^*} uniformly bounded above zero for all $\theta^* \in \Theta$
- **Conditional Lindeberg:** For any $\delta > 0$, $\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[Z_t^2 \mathbb{1}_{|Z_t| > \delta\sqrt{T}} \middle| \mathcal{H}_{t-1}\right] \xrightarrow{P} 0$ uniformly over $\theta^* \in \Theta$.

Note that to satisfy the conditional Lindeberg condition, it is sufficient for $E_{\theta^*}[Z_t^4 | \mathcal{H}_{t-1}] < m < \infty$ for some constant m for all t and $\theta^* \in \Theta$. Since it is more straightforward to show the Lindeberg condition holds under sufficient moment assumptions, we focus on ensuring that our choice of adaptive weights ensures that the conditional variance condition holds. To ensure that the martingale difference sequence of equation (4) satisfies the conditional variance condition, we choose adaptive weights W_t such that the following holds for some fixed positive definite matrix Σ_{θ^*}:

$$\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2 \phi(X_t, A_t)^{\otimes 2}\left(Y_t - \phi(X_t, A_t)^\top \theta^*\right)^2 \middle| \mathcal{H}_{t-1}\right] \xrightarrow{P} \Sigma_{\theta^*} \text{ uniformly over } \theta^* \in \Theta. \tag{5}$$

We will now show how our choice of weight $W_t = \sqrt{\frac{\pi^{\text{eval}}_{A_t}}{\pi_{t,A_t}}}$ ensures (5) holds by changing the sampling measure for the conditional variance from the adaptive policy π_{t,A_t} to the non-adaptive evaluation policy π^{eval}. For square-root importance sampling weights, equation (5) equals the following:

$$\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[E_{\theta^*}\left[\frac{\pi^{\text{eval}}_{A_t}}{\pi_{t,A_t}}\phi(X_t, A_t)^{\otimes 2}\left(Y_t - \phi(X_t, A_t)^\top \theta^*\right)^2 \middle| \mathcal{H}_{t-1}, X_t\right] \middle| \mathcal{H}_{t-1}\right] \tag{6}$$

$$\underset{(a)}{=} \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[E_{\theta^*,\pi^{\text{eval}}}\left[\phi(X_t, A_t)^{\otimes 2}\left(Y_t - \phi(X_t, A_t)^\top \theta^*\right)^2 \middle| \mathcal{H}_{t-1}, X_t\right] \middle| \mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{=} E_{\theta^*,\pi^{\text{eval}}}\left[\phi(X_t, A_t)^{\otimes 2}\left(Y_t - \phi(X_t, A_t)^\top \theta^*\right)^2\right].$$

Above, (a) holds because the weights changes the sampling measure from the adaptive policy π_{t,A_t} to the evaluation policy π^{eval}. (b) holds because after changing the sampling measure from the adaptive policy $\pi_t \in \sigma(\mathcal{H}_{t-1})$ to the non-adaptive policy $\pi^{\text{eval}} \in \sigma(\mathcal{H}_0)$, the distribution of (Y_t, X_t, A_t) no longer depends on history $\mathcal{H}_{t-1}$.

Since the weights $W_t = \sqrt{\frac{\pi^{\text{eval}}_{A_t}}{\pi_{t,A_t}}}$ have the effect of changing the sampling measure of the conditional variance from the adaptive policy π_t to evaluation policy π^{eval}, they can be interpreted as square-root importance weights or square-root Radon–Nikodym derivative weights. Note that in the setting that the action space $\mathcal{A}$ is finite, the conditional Lindeberg condition will also hold if we let $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$; this is equivalent to setting the evaluation policy π^{eval} to the uniform policy over $\mathcal{A}$. The square-root importance weights we use are a form of variance stabilizing weights, akin to those introduced in Hadad et al. (2019) for estimating mean rewards and differences in mean rewards on data collected using multi-armed bandits. In fact, in the special case that $|\mathcal{A}| = K$ and $\phi(X_t, A_t) = [\mathbb{1}_{A_t=1}, \mathbb{1}_{A_t=2}, ..., \mathbb{1}_{A_t=K}]^\top$, the adaptively weighted least squares estimator is equivalent to the weighted average estimator of Hadad et al. (2019).

In Figure 1, we show plots of the least squares estimator z-statistic both with and without adaptive weighting for data collected using Thompson Sampling on a two-arm bandit. It is clear that the least squares z-statistic for estimators of θ_1^* is much closer to a normal distribution with adaptive weighting.

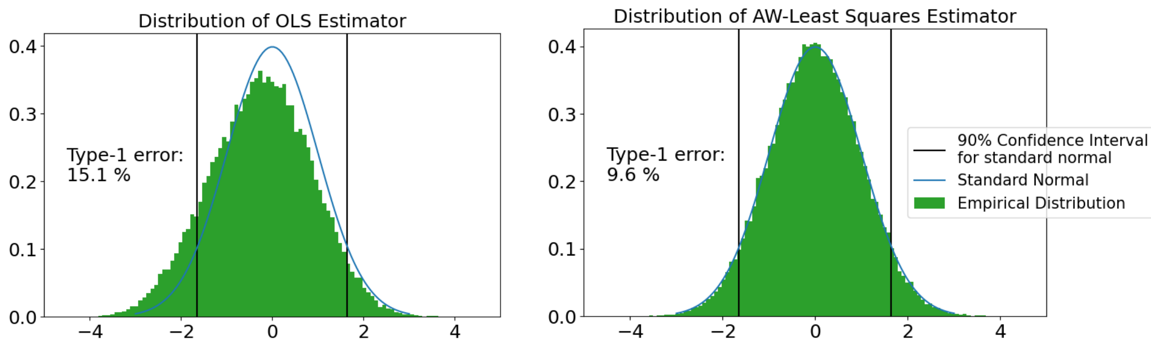

Figure 1: Above we plot the empirical distribution of the OLS and the AW-LS estimator with weights $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$, in a two arm bandit setting. We perform Thompson Sampling with standard normal priors, $\theta^* = [\theta_0^*, \theta_1^*]^\top = [0, 0]^\top$, standard normal errors, and $T = 1000$. Specifically we plot $\sqrt{\sum_{t=1}^T A_t}(\hat{\theta}_{1,T}^{\text{OLS}} - \theta_1^*)$ in the left histogram and $\frac{1}{\sqrt{T}} \sum_{t=1}^T \frac{A_t}{\sqrt{\pi_{t,1}}}(\hat{\theta}_{1,T}^{\text{AW-LS}} - \theta_1^*)$ in the right histogram.

3.3 Asymptotic Normality

In this section, we discuss the conditions under which we prove that

$$\Sigma_{\theta^*,T}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\hat{\theta}_T) \right] \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \theta^* \in \Theta.$$

In general, the conditions we assume differ from those made for standard M-estimators on i.i.d. data because (i) we guarantee convergence in distribution that is uniform over $\theta^* \in \Theta$, which is stronger than the guarantee of asymptotic normality for each fixed $\theta^* \in \Theta$, and (ii) $\pi_{t,a}$ can depend on $\mathcal{H}_{t-1}$ and X_t.

Condition 1 (Stochastic Contextual Bandit Environment) *X_t are i.i.d., Y_t is independent of $\mathcal{H}_{t-1}$ given X_t, A_t, and the conditional distribution $Y_t \mid X_t, A_t$ is in variant over $t \in [1\colon T]$.*

Adopting potential outcome notation, condition 1 is satisfied when $\{X_t, Y_t(a) : a \in \mathcal{A}\}$ are i.i.d. across $t \in [1\colon T]$. Note that action space $\mathcal{A}$ can be finite or infinitely large.

Condition 2 (Differentiable) *The first three derivatives of $m_\theta(Y_t, X_t, A_t)$ with respect to θ exist for all $\theta \in \Theta$ almost surely with respect to $P_{\theta^*, \pi_t^{\text{eval}}}$ for all $t \in [1\colon T]$.*

Condition 3 (Bounded Parameter Space) *$\theta^* \in \Theta$ compact for Θ open subset of $\mathbb{R}^d$.*

Condition 4 (Lipschitz) *There exists some function g such that $E_{\theta^*, \pi_t^{\text{eval}}}[g(Y_t, X_t, A_t)^2]$ is uniformly bounded over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$, which satisfies the following for all $\theta, \theta' \in \Theta$:*

$$|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\|_2.$$

Conditions 3 and 4 together are used to ensure the following functionally uniform law of large numbers result that $\sup_{\theta \in \Theta} \left| \sum_{t=1}^T W_t m_\theta(Y_t, X_t, A_t) - E_{\theta^*}\left[W_t m_\theta(Y_t, X_t, A_t) | \mathcal{H}_{t-1} \right] \right| = o_p(1)$ uniformly over $\theta^* \in \Theta$. This property is used to prove the consistency of $\hat{\theta}_T$. On i.i.d. data, an analogous functional uniform law of large numbers result holds under these conditions. Additionally, on i.i.d. data consistency can be shown without assuming the parameter space is compact when m_θ is assumed to be concave in θ, by showing that $\hat{\theta}_T$ are contained within a bounded space eventually (Van der Vaart, 2000; Engle, 1994; Bura et al., 2018). We expect that a similar result would hold for our adaptively weighted M-estimators.

Condition 5 (Moments) *The fourth moments of $m_{\theta^*}(Y_t, X_t, A_t)$, $\dot{m}_{\theta^*}(Y_t, X_t, A_t)$ and $\ddot{m}_{\theta^*}(Y_t, X_t, A_t)$ with respect to $P_{\theta^*, \pi_t^{\text{eval}}}$ are bounded uniformly over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$. For all sufficiently large T, the minimum eigenvalue of $\Sigma_{T,\theta^*} = \frac{1}{T} \sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}} \left[\dot{m}_{\theta^*}(Y_t, X_t, A_t)^{\otimes 2} \right]$ is bounded above $\delta_{\dot{m}^2} > 0$ for all $\theta^* \in \Theta$.*

Condition 5 is similar to those those made in Van der Vaart (2000, Theorem 5.41). However, in order to guarantee uniform convergence in distribution we assume that moment bounds hold uniformly over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$.

Condition 6 (Third Derivative Domination) *There exists a function $\dddot{m}(Y_t, X_t, A_t)$ such that for some $\epsilon_{\dddot{m}} > 0$, the following holds with probability 1,*

$$\sup_{\theta \in \Theta : \|\theta - \theta^*\| \leq \epsilon_{\dddot{m}}} \|\dddot{m}_\theta(Y_t, X_t, A_t)\|_1 \leq \|\dddot{m}(Y_t, X_t, A_t)\|_1. \tag{7}$$

Above, for any $B \in \mathbb{R}^{d \times d \times d}$, we denote $\|B\|_1 = \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d |B_{i,j,k}|$. Also the second moment of $\dddot{m}(Y_t, X_t, A_t)$ with respect to $P_{\theta^*, \pi_t^{\text{eval}}}$ are bounded uniformly over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$.

Condition 6 equation (7) is similar to assumptions made in classical proofs of the asymptotic normality of M-estimators; see Van der Vaart (2000, Theorem 5.41).

Condition 7 (Maximizing Solution)

- $\theta^* \in \operatorname{argmax}_{\theta \in \Theta} E_{\theta^*} \left[m_\theta(Y_t, X_t, A_t) \middle| X_t, A_t \right]$ *with probability 1 for all $t \in [1\colon T]$. Also* $E_{\theta^*} \left[\dot{m}_{\theta^*}(Y_t, X_t, A_t) \middle| X_t, A_t \right] = 0$ *with probability 1 for all $t \in [1\colon T]$.*
- *The minimum of the absolute values of eigenvalues of $\frac{1}{T} \sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}} \left[\ddot{m}_{\theta^*}(Y_t, X_t, A_t) \right]$ is bounded above $\delta_{\ddot{m}} > 0$ for all $\theta^* \in \Theta$ for all sufficiently large T.*

Condition 7 ensures that θ^* is a conditionally maximizing solution for all contexts X_t and actions A_t. Note that it does not require that θ^* is conditionally always a *unique* optimal solution, which means that our theory allows for there to be no unique optimal θ^* for some contexts and actions. We will use condition 7 to ensure that $\{\dot{m}_{\theta^*}(Y_t, X_t, A_t)\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=1}^T$.

Condition 8 (Well-Separated Solution) *For sufficiently large T, for any $\epsilon > 0$, there exists some $\delta > 0$ such that*

$$\inf_{\theta \in \Theta : \|\theta - \theta^*\|_2 > \epsilon} \left\{ \frac{1}{T} \sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}} \left[m_{\theta^*,t} - m_{\theta,t} \right] \right\} \geq \delta,$$

for all $\theta^* \in \Theta$.

A well-separated solution condition akin to condition 8 is commonly assumed to prove consistency of M-estimators, e.g., see Van der Vaart (2000, Theorem 5.7).

Condition 9 (Adaptive Square-Root Importance Weights) *We set adaptive weights $W_t = \sqrt{\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}}$ for evaluation policies $\{\pi_t^{\text{eval}}\}_{t=1}^T \in \sigma(\mathcal{H}_0)$. Also for all $t \in [1 \colon T]$, $\sqrt{\rho_{\min}} \leq W_t \leq \sqrt{\rho_{\max}}$ w.p. 1 for some constants $0 < \rho_{\min} \leq \rho_{\max} < \infty$.*

Theorem 1 (Consistency of $\hat{\theta}_T$) *Under Conditions 1-5 and 7-9 we have that $\hat{\theta}_T \xrightarrow{P} \theta^*$ uniformly over $\theta^* \in \Theta$.*

Theorem 2 (Asymptotic Normality of Adaptively Weighted M-Estimators) *Let $\hat{\theta}_T \xrightarrow{P} \theta^*$ uniformly over $\theta^* \in \Theta$. Under Conditions 1, 2, 5- 7, and 9,*

$$\Sigma_{\theta^*,T}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\hat{\theta}_T) \right] \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \theta^* \in \Theta. \tag{8}$$

3.4 Confidence Intervals

Standard asymptotic normality results for estimators on i.i.d. data typically are of the form $\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, \Sigma)$. However, our result is not of this form because $\frac{1}{T}\ddot{M}_T(\hat{\theta}_T)$ in equation (8) fails to converge in probability to a constant uniformly under common bandit algorithms. For this reason, we use projected confidence regions when constructing confidence regions for subsets of indices of θ^*; note that these projections are not needed when constructing joint confidence regions for the entire vector θ^*. The asymptotic normality result of equation (8) guarantees that for d-dimensional θ^*,

$$\liminf_{T \to \infty} \inf_{\theta^* \in \Theta} P_{\theta^*} \left(\frac{1}{T} (\hat{\theta}_T - \theta^*)^\top \ddot{M}_T(\hat{\theta}_T)^\top \Sigma_{\theta^*,T}^{-1} \ddot{M}_T(\hat{\theta}_T)(\hat{\theta}_T - \theta^*) \leq \chi^2_{d,(1-\alpha)} \right) = 1 - \alpha.$$

Note that the region

$$\left\{ \theta \in \Theta : \frac{1}{T} (\hat{\theta}_T - \theta)^\top \ddot{M}_T(\hat{\theta}_T)^\top \Sigma_{\theta^*,T}^{-1} \ddot{M}_T(\hat{\theta}_T)(\hat{\theta}_T - \theta) \leq \chi^2_{d,(1-\alpha)} \right\} \tag{9}$$

defines a d-dimensional hyper-ellipsoid confidence region for θ^*. Suppose we are interested in constructing a confidence region for d_1-dimensional θ_1^* where $\theta^* = [\theta_0^*, \theta_1^*]$. We can project the ellipse defined in equation (9) onto a d_1-dimensional space to get a d_1-dimensional ellipsoid, which represents the confidence region for θ_1^*. See Figure 2 for an example of a two-dimensional ellipsoid projected onto one dimension. See Appendix A.3 for more details of how to construct projected confidence regions.

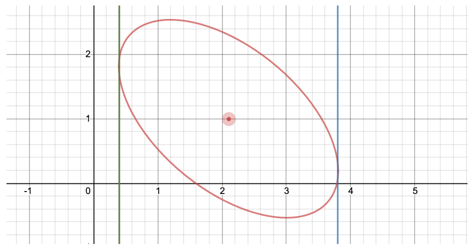

Figure 2: Projection of two-dimensional ellipse onto one dimension (x-axis).

3.5 Choice of Evaluation Policy

When the action space is bounded, a natural default choice for π_t^{eval} is the uniform distribution over $\mathcal{A}$, and since the weights are invariant to scaling by a global constant, this choice corresponds to $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$. In general though, when the action space is not bounded, it is necessary to explicitly choose $\pi_{t,A_t}^{\text{eval}}$.

We now discuss how to choose evaluation policies $\{\pi_t^{\text{eval}}\}_{t=1}^T$ in order to minimize the asymptotic variance of adaptively weighted estimators. We focus on the adaptively weighted least squares estimator in a multi-armed bandit setting. Let $\vec{A}_t = [\mathbb{1}_{A_t=1}, \mathbb{1}_{A_t=2}, ..., \mathbb{1}_{A_t=|\mathcal{A}|}]^\top$. The least squares estimator has function $m_\theta = -\frac{1}{2}(Y_t - \vec{A}_t^\top \theta)^2$. We assume that $E[Y_t|A_t] = \theta^\top \vec{A}_t$ and that $E[(Y_t - \vec{A}_t^\top \theta)^2 | A_t = a] = \sigma_a^2$ for all $a \in \mathcal{A}$. The adaptively weighted least squares estimator is defined as $\hat{\theta}^{\text{AW-LS}} := \text{argmax}_{\theta \in \Theta} \left\{ \sum_{t=1}^T W_t \left(Y_t - \vec{A}_t^\top \theta \right)^2 \right\}$. The variance of $\hat{\theta}_T^{\text{AW-LS}}$ is as follows:

$$\sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(o_p(1) + \frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}\pi_{t,a}^{\text{eval}}} \right)^2} \tag{10}$$

Above, $\vec{a} = [\mathbb{1}_{a=1}, \mathbb{1}_{a=2}, ..., \mathbb{1}_{a=|\mathcal{A}|}]^\top$. By Cauchy-Schwartz inequality, we can lower bound each term in the summation in (10):

$$\vec{a}^\top \vec{a} \sigma_a^2 \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(\frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}\pi_{t,a}^{\text{eval}}} \right)^2} \succeq \vec{a}^\top \vec{a} \sigma_a^2 \frac{1}{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}}. \tag{11}$$

Note that this lower bound is achieved when $\pi_{t,a}^{\text{eval}} = \pi_{t,a}$. However, since $\pi_{t,a}$ is random and $\{\pi_t^{\text{eval}}\}_{t=1}^T \in \sigma(\mathcal{H}_0)$, setting $\pi_{t,a}^{\text{eval}} = \pi_{t,a}$ is generally an unfeasible choice. We want to choose $\pi_{t,a}^{\text{eval}}$ to match $\pi_{t,a}$ as close as possible subject to the constraint that $\pi_{t,a}^{\text{eval}} \in \mathcal{H}_0$.

Thus to approximately minimize the variance of the AW-LS estimator we propose setting $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$ for all $t \in [1\colon T]$. Note that $E_{\theta^*}[\pi_{t,a}]$ depends on the value of θ^*, which is unknown. Thus, it is natural to choose $\pi_{t,a}^{\text{eval}}$ to be $E_{\theta^*}[\pi_{t,a}]$ weighted by a prior on θ^*. Note that as long as the evaluation policy ensures that weights W_t are bounded, the choice of evaluation policy does not affect the asymptotic validity of the estimator.

In Figure 3, we display the difference in mean squared error for the AW-LS estimator for two different choices of evaluation policy: (1) the uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) the expected $\pi_{t,a}$ evaluation policy for which $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$. We can see in this setting that by setting $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$ we are able to decrease the mean squared error of the AW-LS estimator compared AW-LS with the uniform evaluation policy. Note though that in some cases setting $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$ is equivalent to choosing the uniform evaluation policy. For example, in the two arm bandit case in which $\theta^* = [\theta_0^*, \theta_1^*]^\top$, note that when $\theta_0^* = \theta_1^*$, both arms are optimal so under common bandit algorithms $E_{\theta^*}[\pi_{t,a}] = 0.5$ for all $t \in [1\colon T]$, which will make the evaluation policy $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$ equivalent to the uniform policy.

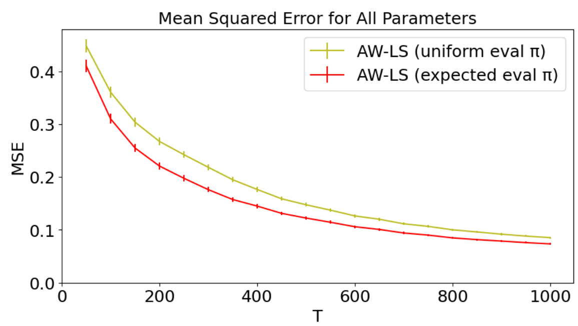

Figure 3: Above we plot the mean squared errors for the adaptively weighted least squares estimator with evaluation policies: (1) uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) expected $\pi_{t,a}$ evaluation policy for which $\pi_{t,a}^{\text{eval}} = E_{\theta^*}[\pi_{t,a}]$ (oracle quantity). In a two arm bandit setting we perform Thompson Sampling with standard normal priors, 0.01 clipping, $\theta^* = [\theta_0^*, \theta_1^*]^\top = [0, 1]^\top$, standard normal errors, and $T = 1000$. Error bars denote standard errors computed over 5,000 Monte Carlo simulations.

4. Related Work

The adaptive clinical trials literature mostly uses classical statistical methods, like ordinary least squares (OLS) and maximum likelihood estimation, and associated asymptotic approximations to the distributions of the estimators and test statistics to conduct statistical inference (Hu and Rosenberger, 2006; Hu et al., 2009). This is possible because algorithms with smooth limiting allocations (i.e., the limiting rate at which arms are selected is a smooth function of the difference in expected rewards) are used. While smooth limiting allocation functions preclude algorithms with sub-linear regret (which includes all regret-minimizing bandit algorithms), they are the focus of that literature because the increased exploration increases the power of tests on the resulting data.

Villar et al. (2015) and Rafferty et al. (2019) empirically illustrate that classical OLS inference methods have inflated Type-1 error when used on data collected with any of a variety of regret-minimizing multi-armed bandit algorithms. There have been a variety of recent works proposing alternative methods for statistical inference using asymptotic approximations on bandit data. The common thread between all of these methods is that they all utilize a form of *adaptive weighting*. The first work in this area was by Deshpande et al. (2018) who construct the W-decorrelated estimator, which adjusts the standard OLS estimator with a sum of adaptively weighted residuals. As discussed in the appendix of Zhang et al. (2020), in the multi-armed bandit setting, the W-decorrelated estimator upweights observations from early in the study and down-weights observations from later in the study. In the batched bandit setting, Zhang et al. (2020) show that the Z-statistics for the OLS estimators computed separately on each batch are jointly asymptotically normal. Standardizing the OLS statistic for each batch effectively adaptively re-weights the observations in each batch.

Hadad et al. (2019) introduce the Adaptively-Weighted Augmented Inverse Propensity Weighted estimator (AW-AIPW) and the Adaptively-Weighted Average estimator (AWA) for inference regarding the expected reward and margin (difference in expected rewards between two arms) for data collected with bandit algorithms. AW-AIPW and AWA are modifications of the standard augmented-inverse propensity weighted (doubly robust) estimator and the sample mean estimator with adaptive weights, respectively. In particular, Hadad et al. (2019) introduce a class of adaptive "variance stabilizing" weights, which guarantee that the variance of a normalized version of their estimators converges in probability to a constant. Note that the AW-AIPW and AWA estimators as presented in Hadad et al. (2019) are limited to only estimating the scalar values (e.g., difference in mean rewards for two arms of a multi-armed bandit) on data collected with bandit algorithms. In the discussion section of Hadad et al. (2019) they discuss three open questions related to their work. Our work addresses two of these questions: 1) "What additional estimators can be used for normal inference with adaptively collected data?" and 2) How do their results generalize to more complex sampling designs, like data collected with contextual bandit algorithms? In this work, we demonstrate that adaptive variance stabilizing weights are not only applicable to estimators for the mean reward and differences in mean rewards. Rather, a large class of M-estimators modified with adaptive square-root importance weights, which are variance stabilizing, can be used to construct uniformly valid confidence intervals using data collected with (contextual) bandit algorithms. This generalization allows us to construct confidence intervals for a much larger class of inferential targets, including multivariate parameters in the contextual bandit setting for a variety of models for the expected outcome.

An alternative to using asymptotic approximations to construct confidence intervals is to use high probability confidence bounds. High probability anytime confidence bounds are used throughout the bandit literature to prove guarantees for algorithms—like regret bounds and best-arm identification probabilities—and for providing anytime-valid confidence intervals and p-values (Abbasi-Yadkori et al., 2011; Howard et al., 2018). High probability anytime bounds are guaranteed to hold for finite samples (not just asymptotically) and hold simultaneously for any number of observations, i.e., for all $T \geq 1$, and thus provide stronger guarantees than confidence intervals constructed using asymptotic approximations. The downside of using anytime high probability confidence bounds is that they are generally wider than those constructed using asymptotic approximations. Additionally, these methods require a bounded parameter space and moreover that the parameter space bounds are known, as they are explicitly incorporated into the confidence bounds. The above is a common reason why much of classical statistics, with a long history of success in science, uses asymptotic approximations to provide valid statistical inference. Here we do the same. In the simulation Section 5, we compare our proposed approach to high probability confidence bounds.

5. Simulation Results

In this section, we let $Y_t = R_t$. We first consider a linear contextual bandit setting, where the expected reward is a linear function of the parameters. We focus on this setting first because other methods for constructing valid confidence intervals on bandit data have been developed for linear contextual bandit, which we compare to. Later, we will discuss more complicated statistical models.

Specifically, we consider a two-arm linear contextual bandit setting with the following linear model for the expected reward:

$$E_\theta[R_t|X_t, A_t] = \tilde{X}_t^\top \theta_0 + A_t \tilde{X}_t^\top \theta_1,$$

where $\tilde{X}_t = [1, X_t]$ (X_t with intercept term) and $A_t \in \{0, 1\}$. We use the weighted least squares estimator (an example of an M-estimator):

$$\hat{\theta}_T^{\text{AW-LS}} = \operatorname*{argmax}_{\theta \in \Theta} \left\{ -\sum_{t=1}^{T} W_t \left(R_t - \tilde{X}_t^\top \theta_0 + A_t \tilde{X}_t^\top \theta_1 \right)^2 \right\}.$$

In the simulations below we generate data using the following model for the expected reward $E_{\theta^*}[R_t|A_t, X_t] = \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$. In our simulations, X_t is two dimensional, which makes θ^* six dimensional and the parameters of the advantage function, θ_1^*, three dimensional. We set $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$. The contexts X_t are drawn i.i.d. from a uniform distribution. We generate the noise on the rewards, $R_t(A_t) - E_{\theta^*}[R_t|A_t, X_t]$, by sampling i.i.d. from a student's t distribution with five degrees of freedom. We collect data using Thompson Sampling with a linear model for the expected reward and normal priors (Agrawal and Goyal, 2013). We constrain the action selection probabilities with *clipping* at a rate of 0.01; this means that while typical Thompson Sampling produces action selection probabilities $\pi_{t,a}^{\text{TS}}$, we instead use action selection probabilities $\pi_{t,a} = 0.01 \vee \left(0.99 \wedge \pi_{t,a}^{\text{TS}}\right)$ to select actions. We constrain the action selection probabilities in order to satisfy condition 9 with the uniform evaluation policy. Also note that increasing the amount the algorithm explores decreases the width of confidence intervals constructed on the resulting data; see section 6.2 for more discussion of the trade-off between regret minimization and statistical inference objectives. As discussed earlier, in order to construct confidence regions for the parameters of the treatment effect or advantage function, θ_1^*, we use a projected confidence region. We also compare to the W-decorrelated estimator of Deshpande et al. (2018) and the high probability self-normalized martingale bound of Abbasi-Yadkori et al. (2011). For all our simulations, we use $5,000$ Monte-Carlo repetitions and the error bars plotted are standard errors. See our Appendix A for more details about our simulations, as well as additional simulation results.

Figure 4 plots the mean squared error of the AW-LS estimator, the unweighted least squares estimator, and the W-decorrelated estimator (Deshpande et al., 2018) computed on the data collected as decribed in the previous paragraph. We can see from the figure that the AW-LS estimator has comparable mean squared error than the standard least squares estimator. The W-decorrelated estimator has much greater MSE than other estimators, which is not surprising because the W-decorrelated estimator decreases the bias of estimators by increasing the variance (Deshpande et al., 2018).

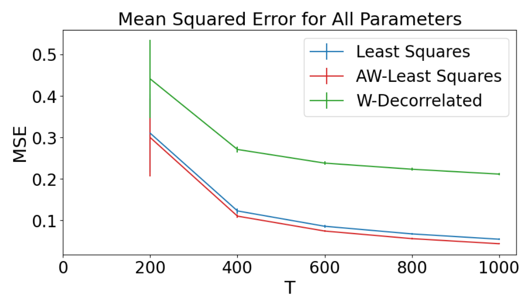

Figure 4: Mean squared error for different estimators of θ^* on data collected under Thompson Sampling.

In Figure 5, we plot the empirical coverage probabilities and volume of 90% confidence regions for θ^*. For convenience, through the remainder of this section we refer to a confidence region constructed using the asymptotic distribution of an estimator as the ADCR (asymptotic distribution confidence region) of an estimator. The least squares estimator ADCRs have significant undercoverage, which does not improve as T increases, while the AW-LS estimator ADCRs have very reliable coverage. The confidence regions constructed using the self-normalized martingale bound have reliable coverage, but are very conservative. Methods for constructing confidence regions which undercover can have much smaller confidence regions than those that provide at least the nominal coverage level. To allow for a fair comparison between methods for constructing confidence regions with varying levels of empirical coverage, when computing the volumes of the confidence regions for all our simulations, if a method for constructing confidence regions under-covers for any value of T, we set the cutoff values for the statistic to that which empirically guarantees 90% coverage. The AW-LS estimator ADCRs have similar volume to the least squares estimator ADCRs and significantly smaller volume than the self-normalized martingale bound confidence regions and W-decorrelated estimator ADCRs. Empirically, we found that the coverage probability of the W-decorrelated estimator ADCRs were very sensitive to the choice of tuning parameters; we follow a similar method for choosing these parameters as Deshpande et al. (2018) as discussed in Appendix A.3.

In Figure 6, we plot the empirical coverage probabilities and volume of 90% confidence regions for the advantage parameters, θ_1^*. Due to the need for using projected confidence regions, the AW-LS estimator ADCRs are significantly more conservative than the unweighted least squares estimator ADCRs in this setting. The volume of the AW-LS estimator ADCRs is still close to that of the unweighted least squares estimator ADCRs, and significantly outperforms both the W-decorrelated estimator ADCRs and the self-normalized martingale bound confidence regions.

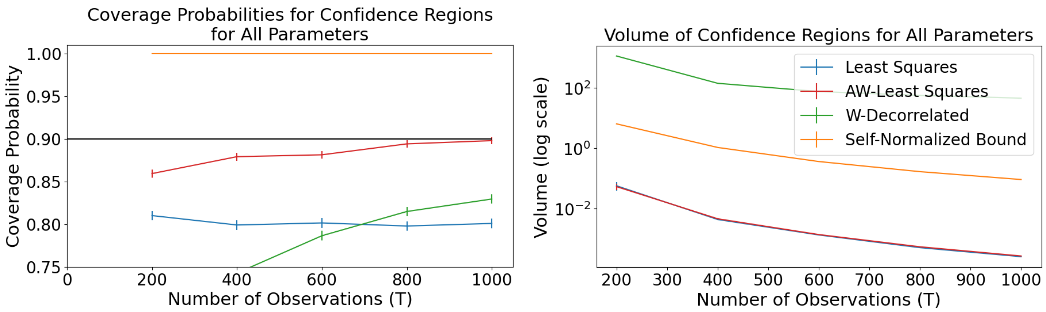

Figure 5: **Parameters θ^*:** Empirical coverage probabilities for 90% confidence ellipsoids for parameters θ^* (left) and volume of 90% confidence ellipsoids for θ^* (right).

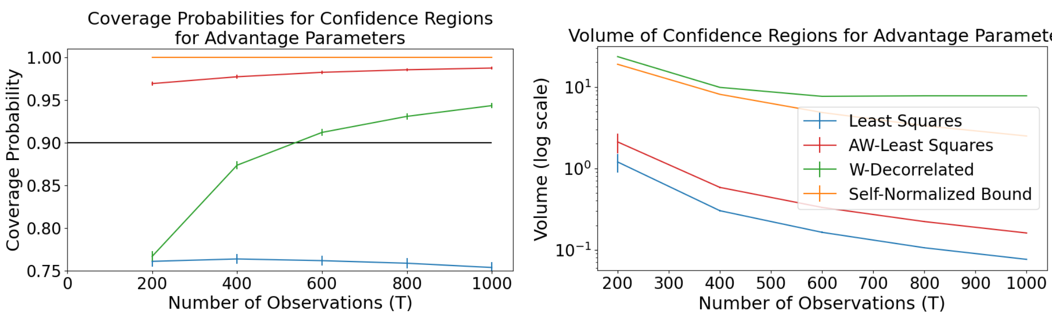

Figure 6: **Advantage Parameters θ_1^*:** Empirical coverage probabilities for 90% confidence ellipsoids for advantage parameters θ_1^* (left) and volume of 90% confidence ellipsoids for θ_1^* (right).

Next we show simulation results for an adaptively weighted maximum likelihood estimator (AW-MLE) for a logistic regression model in a contextual bandit problem with binary rewards. We assume the following model for expected rewards:

$$E_\theta[R_t|X_t, A_t] = \left[1 + \exp\left(-\tilde{X}_t^\top \theta_0 - A_t \tilde{X}_t^\top \theta_1\right)\right]^{-1}.$$

Our simulation set-up is analogous to that for the AW-LS estimator, but for binary rewards. We collect data using Thompson Sampling with a linear model for the expected reward and Normal priors with constrained action selection probabilities via clipping a rate of 0.01. In our simulations, X_t is two dimensional, which makes θ^* six dimensional and the parameters of the advantage function, θ_1^*, three dimensional. We set $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$. We use a uniform evaluation policy. See more details in Appendix A.2.

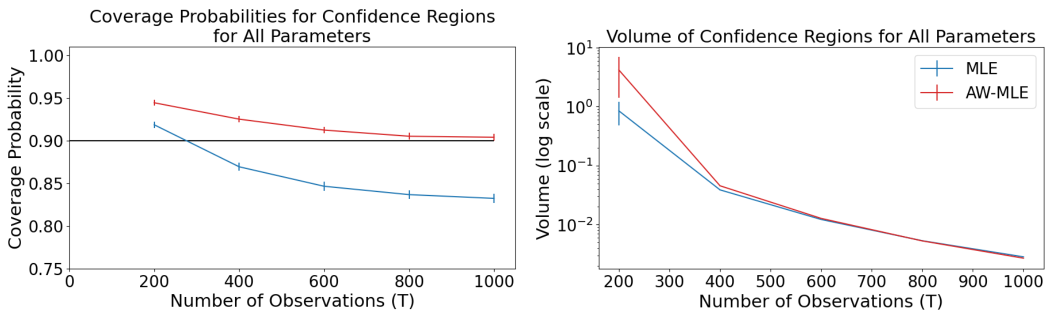

Figure 7: **Bernoulli Rewards, parameters θ^*:** Empirical coverage probabilities for 90% confidence ellipsoids for parameters θ^* (left) and volume of 90% confidence ellipsoids for θ^* (right).

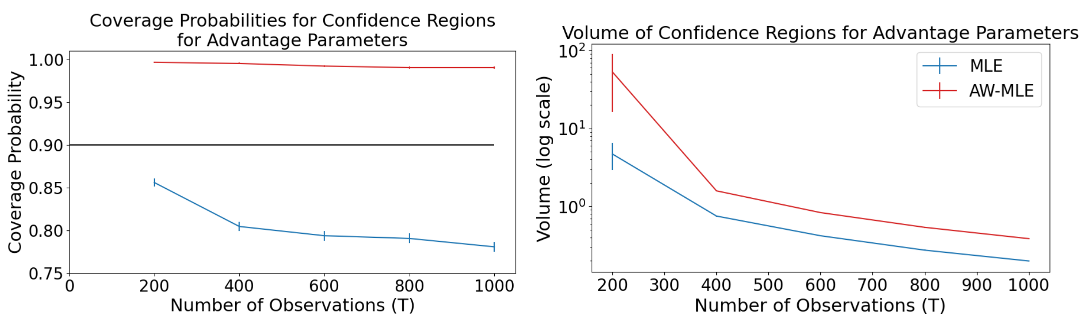

Figure 8: **Bernoulli Rewards, advantage parameters θ_1^*:** Empirical coverage probabilities for 90% confidence ellipsoids for advantage parameters θ_1^* (left) and volume of 90% confidence ellipsoids for θ_1^* (right).

In Figure 7, we plot the empirical coverage probabilities and volume of 90% confidence regions for θ^* for Bernoulli rewards. Note that the W-decorrelated estimator and Self-Normalized Martingale bounds were developed for settings in which the expected reward is a linear function of the context and are thus not applicable in this setting. The unweighted MLE ADCRs have significant undercoverage, which does not improve as T increases, while the AW-MLE ADCRs have very reliable coverage. The ADCRs for both the unweighted and weighted MLEs have similar volumes. In Figures 8, we plot the empirical coverage probabilities and volume of 90% confidence intervals for the advantage parameters, θ_1^*, for Bernoulli rewards. Due to the projected confidence region, the AW-MLE ADCRs are significantly more conservative than the unweighted MLE ADCRs in this setting. However, the AW-MLE ADCRs have much more reliable coverage than the unweighted MLE ADCRs. We also show simulation results for the AW-MLE for a generalized linear model with Poisson distributed rewards in Appendix A.1. Overall, the results for the AW-MLE ADCRs for Bernoulli and Poisson distributed rewards are very similar.

6. Discussion

There are many open research areas regarding statistical inference on data collected with bandit and other adaptive algorithms. We discuss two main areas of open research: (i) more general and robust statistical inference methods and (ii) developing methods to help guarantee power on adaptively collected data.

6.1 Developing More General and Robust Statistical Inference Methods on Bandit Data

Weakening model assumptions In this work we make the conditional optimality assumption that $\theta^* \in \text{argmax}_{\theta \in \Theta} E_{\theta^*} \left[m_\theta(Y_t, X_t, A_t) | X_t, A_t \right]$ w.p. 1. Our theorems heavily rely on the conditional optimality assumption to ensure that $\{W_t \dot{m}_\theta(Y_t, X_t, A_t)\}_{t \geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t \geq 0}$. On i.i.d. data it is common to define θ^* to be the best *projected* solution, i.e., $\theta^* \in \text{argmax}_{\theta \in \Theta} E_{\theta^*} \left[m_\theta(Y_t, X_t, A_t) \right]$. Note that the best projected solution, θ^*, depends on the distribution of (Y_t, X_t, A_t). It would be ideal to also be able to perform inference for a projected solution on adaptively collected data as well.

Statistical inference methods for adaptively collected data in more complex environments One natural question is whether variance stabilizing methods work in more complicated environments, like for Markov Decision Processes (MDPs). It is not straightforward to apply the adaptive weighting method without knowing state transition probabilities. For example, consider the least squares estimator we discuss in Section 3.2. Recall that to apply the uniform martingale central limit theorem, we needed to ensure that the conditional variance condition (as discussed in section 3.2) holds, i.e., $\frac{1}{T} \sum_{t=1}^T E_{\theta^*} \left[W_t^2 \phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right)^2 \middle| \mathcal{H}_{t-1} \right] \xrightarrow{P} \Sigma_{\theta^*}$ uniformly over $\theta^* \in \Theta$. In equation (6) we show for that weights $W_t = \sqrt{\frac{\pi_{A_t}^{\text{eval}}}{\pi_{t,a}}}$ this condition holds when $\{X_t\}_{t \geq 1}$ are i.i.d.; however, these weights are not generally sufficient in the MDP setting in which the distribution of X_t can depend on (X_{t-1}, A_{t-1}). However, suppose we knew transition probabilities $P(X_t | X_{t-1}, A_{t-1})$ and set weights $W_t = \sqrt{\frac{\pi_{A_t}^{\text{eval}} P_{\text{eval}}(X_t)}{\pi_{t,a} P(X_t | X_{t-1}, A_{t-1})}}$, where $P_{\text{eval}}(X_t)$ is some distribution over X_t. Then the conditional variance would be

$$\frac{1}{T} \sum_{t=1}^T E_{\theta^*} \left[W_t^2 \phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right)^2 \middle| \mathcal{H}_{t-1} \right]$$

$$= \frac{1}{T} \sum_{t=1}^T E_{\theta^*} \left[E_{\theta^*} \left[\frac{\pi_{A_t}^{\text{eval}} P_{\text{eval}}(X_t)}{\pi_{t,A_t} P(X_t | X_{t-1}, A_{t-1})} \phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right)^2 \middle| \mathcal{H}_{t-1}, X_t \right] \middle| \mathcal{H}_{t-1} \right]$$

$$= \frac{1}{T} \sum_{t=1}^T E_{\theta^*} \left[\frac{P_{\text{eval}}(X_t)}{P(X_t | X_{t-1}, A_{t-1})} E_{\theta^*, \pi^{\text{eval}}} \left[\phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right)^2 \middle| \mathcal{H}_{t-1}, X_t \right] \middle| \mathcal{H}_{t-1} \right]$$

$$= E_{\theta^*,\text{eval}} \left[E_{\theta^*,\pi^{\text{eval}}} \left[\phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^* \right)^2 \middle| X_t \right] \right].$$

The outer expectation in the last term above is over X_t which follows the distribution P_{eval}. Since the above is a constant, this satisfies the conditional variance condition. However, in general we do not expect to know the transition probabilities $P(X_t|X_{t-1}, A_{t-1})$ and if we tried to estimate them, our theory would require the estimator to have error $o_p(1/\sqrt{T})$, *below* the parametric rate.[2]

6.2 Developing Methods to Guarantee Power on Adaptively Collected Data

As discussed in the literature, in sequential decision making problems there is a fundamental trade-off between minimizing regret and minimizing estimation error for parameters of the environment (e.g., the difference in expected reward for two arms of a multi-armed bandit) using the resulting data collected (Bubeck et al., 2009; Dean et al., 2018; Erraqabi et al., 2017). Given this trade-off there are many open questions regarding how to guarantee power, i.e., the probability of correctly rejecting a null hypothesis and making a scientific discovery. For example, there is the need for algorithms that balance the objectives of regret minimization and power (Offer-Westort et al., 2019; Liu et al., 2014; Erraqabi et al., 2017; Yao et al., 2020) and a need for the development of sample size calculators that output the sample size needed to guarantee a certain amount of power under a particular alternative when using a particular bandit algorithm for data collection.

2. Specifically, if we estimate $P(X_t|X_{t-1}, A_{t-1})$ and use estimated weights $\hat{W}_t$, for $\{\hat{W}_t \dot{m}(Y_t, X_t, A_t)\}_{t\geq 1}$ to be an approximate martingale difference sequence to apply the martingale CLT, we require that $\frac{1}{\sqrt{T}} \sum_{t=1}^T \hat{W}_t \dot{m}(Y_t, X_t, A_t) \xrightarrow{P} 0$, which given our assumptions requires $\hat{W}_t - W_t = o_p(1/\sqrt{T})$.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Alan Agresti. *Foundations of linear and generalized linear models.* John Wiley & Sons, 2015.

Adrian Barnett, Hang Ding, Karen E Hay, Ian A Yang, Rayleen V Bowman, Kwun M Fong, and Henry M Marshall. The effectiveness of smartphone applications to aid smoking cessation: A meta-analysis. *Clinical eHealth*, 2020.

Lauren Bell, Claire Garnett, Tianchen Qian, Olga Perski, Henry WW Potts, and Elizabeth Williamson. Notifications to improve engagement with an alcohol reduction app: protocol for a micro-randomized trial. *JMIR research protocols*, 9(8):e18690, 2020.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Efstathia Bura, Sabrina Duarte, Liliana Forzani, Ezequiel Smucler, and Mariela Sued. Asymptotic theory for maximum likelihood estimates in reduced-rank multivariate generalized linear models. *Statistics*, 52(5):1005–1024, 2018.

Stefano Caria, Maximilian Kasy, Simon Quinn, Soha Shami, Alex Teytelboym, et al. An adaptive targeted field experiment: Job search assistance for refugees in jordan. 2020.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. *arXiv preprint arXiv:1805.09388*, 2018.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1194–1203, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Aryeh Dvoretzky. Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory.* The Regents of the University of California, 1972.

Robert F Engle. *Handbook of econometrics: volume 4.* Number 330.015195 E53 v. 4. 1994.

Akram Erraqabi, Alessandro Lazaric, Michal Valko, Emma Brunskill, and Yun-En Liu. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR, 2017.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019.

John Hammersley. *Monte carlo methods*. Springer Science & Business Media, 2013.

Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

Feifang Hu and William F Rosenberger. *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons, 2006.

Feifang Hu, Li-Xin Zhang, and Xuming He. Efficient randomized-adaptive designs. *The Annals of Statistics*, pages 2543–2560, 2009.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

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

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

Maximilian Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8 (1), 2019.

Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. 2019.

Predrag Klasnja, Dori E Rosenberg, Jing Zhou, Jane Anau, Anirban Gupta, and David E Arterburn. A quality-improvement optimization pilot of barifit, a mobile health intervention to promote physical activity after bariatric surgery. *Translational Behavioral Medicine*, 2020.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, pages 21–59, 2005.

Jessica Lipschitz, Christopher J Miller, Timothy P Hogan, Katherine E Burdick, Rachel Lippin-Foster, Steven R Simon, and James Burgess. Adoption of mobile apps for depression and anxiety: cross-sectional survey study on patient interest and barriers to engagement. *JMIR mental health*, 6(1):e11334, 2019.

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pages 161–168, 2014.

Martha Neary and Stephen M Schueller. State of the field of mental health apps. *Cognitive and Behavioral Practice*, 25(4):531–537, 2018.

Molly Offer-Westort, Alexander Coppock, and Donald P Green. Adaptive experimental design: Prospects and applications in political science. *Available at SSRN 3364402*, 2019.

Anna Rafferty, Huiji Ying, and Joseph Williams. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *JEDM— Journal of Educational Data Mining*, 11(1):47–79, 2019.

Joseph P Romano, Azeem M Shaikh, et al. On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822, 2012.

Hammad Shaikh, Arghavan Modiri, Joseph Jay Williams, and Anna N Rafferty. Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*, 2019.

Ting Song, Siyu Qian, and Ping Yu. Mobile health interventions for self-control of unhealthy alcohol use: systematic review. *JMIR mHealth and uHealth*, 7(1):e10899, 2019.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Power-constrained bandits. *arXiv preprint arXiv:2004.06230*, 2020.

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Inference for batched bandits. *NeurIPS 2020*, 2020.

Appendix A. Simulations

A.1 Additional Simulation Results

We show simulation results for an adaptively weighted maximum likelihood estimator (AW-MLE) for a generalized linear model in a contextual bandit problem with Poisson rewards. We assume the following model for expected rewards:

$$E_\theta[R_t|X_t, A_t] = \exp\left(\tilde{X}_t^\top \theta_0 - A_t \tilde{X}_t^\top \theta_1\right).$$

Our simulation set-up is analogous to that for the AW-MLE for binary rewards, but for Poisson rewards. We collect data using Thompson Sampling with a linear model for the expected reward and Normal priors with constrained action selection probabilities via clipping a rate of 0.01. In our simulations, X_t is two dimensional, which makes θ^* six dimensional and the parameters of the advantage function, θ_1^*, three dimensional. We set $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$. See more Appendix section A.2.

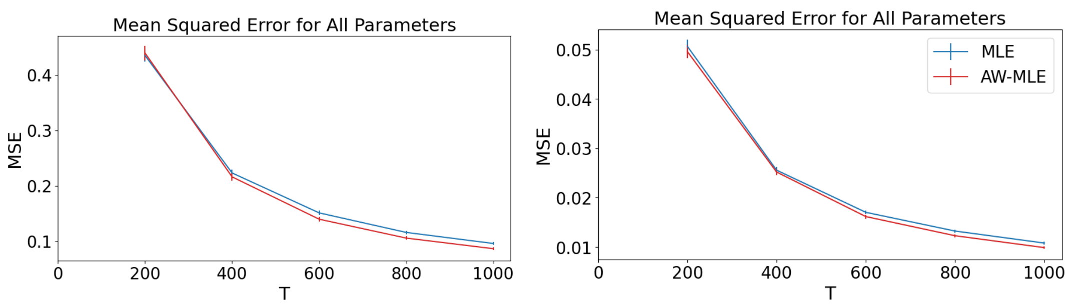

Figure 9: Mean squared error estimators of θ^* for Bernoulli (left) and Poisson (right) rewards.

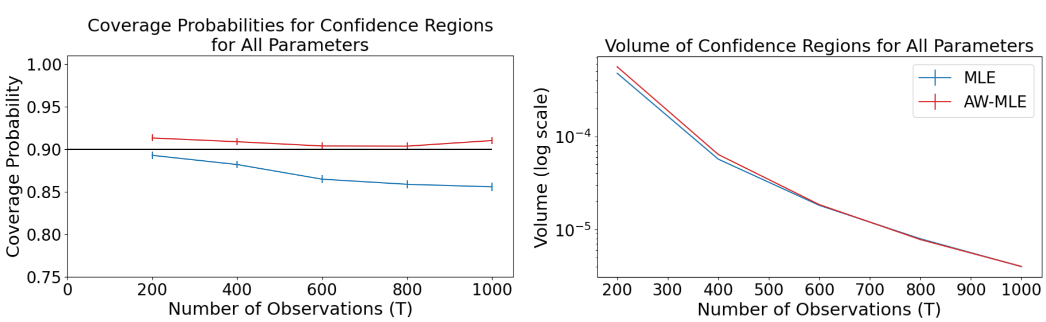

Figure 10: **Poisson Rewards, parameters θ^*:** Empirical coverage probabilities for 90% confidence ellipsoids for parameters θ^* (left) and volume of 90% confidence ellipsoids for θ^* (right).

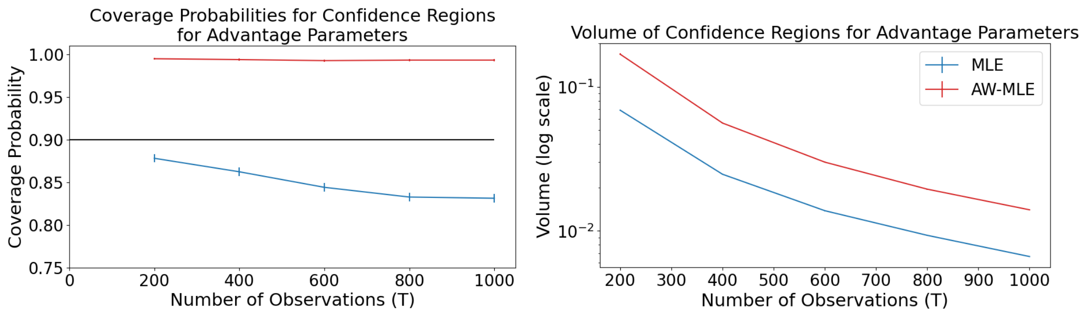

Figure 11: **Poisson Rewards, advantage parameters θ_1^*:** Empirical coverage probabilities for 90% confidence ellipsoids for advantage parameters θ_1^* (left) and volume of 90% confidence ellipsoids for θ_1^* (right).

A.2 Data Generation

Simulation Environment

- Each dimension of X_t is sampled independently from Uniform$(0, 5)$.
- $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$.
- t-Distributed rewards: $R_t | X_t, A_t \sim t_5 + \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$, where t_5 is a t-distribution with 5 degrees of freedom.
- Bernoulli rewards: $R_t | X_t, A_t \sim$ Bernoulli$(expit(\nu_t))$ for $\nu_t = \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$ and $expit(x) = \frac{1}{1+\exp(-x)}$.
- Poisson rewards: $R_t | X_t, A_t \sim$ Poisson$(\exp(\nu_t))$ for $\nu_t = \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$.

Algorithm

- Thompson Sampling with $\mathcal{N}(0, I_d)$ priors on each arm.
- 0.01 clipping
- Pre-processing rewards before received by algorithm:
  - Bernoulli: $R_t * 2 - 1$
  - Poisson: $R_t * 0.6$

A.3 Construction of Confidence Regions

For notational convenience, we define $Z_t = [\tilde{X}_t, A_t \tilde{X}_t]$.

A.3.1 Least Squares / AW-Least Squares

- $\hat{\theta}_T = \left(\sum_{t=1}^T W_t Z_t Z_t^\top\right)^{-1} \sum_{t=1}^T W_t Z_t R_t$
  - For unweighted least squares, $W_t = 1$.
  - For AW-Least Squares, $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$.
- We assume homoskedastic errors and estimate the noise variance σ^2 as follows:

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^T (R_t - Z_t^\top \hat{\theta}_T)^2.$$

- We use a Hotelling t-squared test statistic to construct confidence regions for θ^*:

$$\sqrt{T}(\hat{\theta}_T - \theta)^\top \left(\frac{1}{T}\sum_{t=1}^T W_t Z_t Z_t^\top\right) \hat{\Sigma}_T^{-1} \left(\frac{1}{T}\sum_{t=1}^T W_t Z_t Z_t^\top\right) \sqrt{T}(\hat{\theta}_T - \theta) \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha). \quad (12)$$

  - For unweighted least squares, $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^T Z_t Z_t^\top$ and $W_t = 1$.
  - For AW-Least Squares, $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{\pi_t}^{A_t} \frac{1}{1-\pi_t}^{1-A_t} Z_t Z_t^\top$ and $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$.
- For the unweighted least squares estimator, we construct confidence regions with the following normal approximation: $\sqrt{T}(\hat{\theta}_T^{LS} - \theta^*) \sim \mathcal{N}\left(0, \hat{\sigma}_T^2 (\frac{1}{T}\sum_{t=1}^T Z_t Z_t^\top)^{-1}\right)$. We also use a Hotelling t-squared test statistic to construct confidence regions for θ_1^*:

$$\sqrt{T}(\hat{\theta}_{T,1} - \theta_1)^\top V_{1,T}^{-1} \sqrt{T}(\hat{\theta}_{T,1} - \theta) \leq \frac{d_1(T-1)}{T-d_1} F_{d_1,T-d_1}(1-\alpha),$$

  where $V_{1,T}$ is the lower right $d_1 \times d_1$ block of matrix $\left(\hat{\sigma}_T^2 \frac{1}{T}\sum_{t=1}^T Z_t Z_t^\top\right)^{-1}$.
- For the AW-least squares estimator, we constructed projected construct confidence regions for θ_1^* using the confidence region defined in equation (12). See section A.3.5 below for more details on constructing projected confidence regions.

A.3.2 MLE / AW-MLE Estimator

- $\hat{\theta}_T$ is the root of the score function:

$$0 = \sum_{t=1}^T W_t \left(R_t - b'(\hat{\theta}_T^\top Z_t)\right) Z_t.$$

  We use Newton Raphson optimization to solve for $\hat{\theta}_T$.
  - For unweighted MLE, $W_t = 1$.

- For AW-MLE, $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$.

- Hessian:

$$-\sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$$

- We use a Hotelling t-squared test statistic to construct confidence regions for θ^*:

$$\sqrt{T}(\hat{\theta}_T - \theta)^\top \left(\frac{1}{T} \sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right) \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right) \sqrt{T}(\hat{\theta}_T - \theta) \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha). \quad (13)$$

  - For unweighted least squares, $\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$ and $W_t = 1$.
  - For AW-Least Squares, $\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\pi_t}^{A_t} \frac{1}{1-\pi_t}^{1-A_t} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$ and $W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$.

- For the unweighted MLE estimator, we construct confidence regions with the following normal approximation: $\sqrt{T}(\hat{\theta}_T^{MLE} - \theta^*) \sim \mathcal{N}\left(0, (\frac{1}{T} \sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top)^{-1}\right)$. We also use a Hotelling t-squared test statistic to construct confidence regions for θ_1^*:

$$\sqrt{T}(\hat{\theta}_{T,1} - \theta_1)^\top V_{1,T}^{-1} \sqrt{T}(\hat{\theta}_{T,1} - \theta) \leq \frac{d_1(T-1)}{T-d_1} F_{d_1,T-d_1}(1-\alpha),$$

  where $V_{1,T}$ is the lower right $d_1 \times d_1$ block of matrix $\left(\frac{1}{T} \sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right)^{-1}$.

- For the AW-MLE estimator, we constructed projected construct confidence regions for θ_1^* using the confidence region defined in equation (13). See section A.3.5 below for more details on constructing projected confidence regions.

Distribution	$\boldsymbol{\nu}$	$\boldsymbol{b(\nu)}$	$\boldsymbol{b'(\nu)}$	$\boldsymbol{b''(\nu)}$	$\boldsymbol{b'''(\nu)}$
$\mathcal{N}(\mu, 1)$	μ	$\frac{1}{2}\nu^2$	$\nu = \mu$	1	0
Poisson(λ)	$\log \lambda$	$\exp(\nu)$	$\exp(\nu) = \lambda$	$\exp(\nu) = \lambda$	$\exp(\nu) = \lambda$
Bernoulli(p)	$\log\left(\frac{p}{1-p}\right)$	$\log(1 + e^\nu)$	$\frac{e^\nu}{1+e^\nu} = p$	$\frac{e^\nu}{(1+e^\nu)^2} = p(1-p)$	$p(1-p)(1-2p)$

A.3.3 W-Decorrelated

The following is based on Algorithm 1 of Deshpande et al. (2018).

- The W-decorrelated estimator is constructed as follows with adaptive weights for $W_t \in \mathbb{R}^d$:

$$\hat{\theta}^d = \hat{\theta}_T^{\text{LS}} + \sum_{t=1}^{T} W_t (R_t - \tilde{X}_t^\top \hat{\theta}_T^{\text{LS}}).$$

- The weights are set as follows: $W_t = (I_d - \sum_{s=1}^{t} W_s Z_s^\top) Z_t \frac{1}{\lambda_T + \|Z_t\|_2^2}$.

- We choose $\lambda_T = \text{mineig}_{0.01}(Z_t Z_t^\top)/\log T$ and $\text{mineig}_\alpha(Z_t Z_t^\top)$ represents the α percentile of the minimum eigenvalue of $Z_t Z_t^\top$. This is similar to the procedure used in the simulations of Deshpande et al. (2018) and is guided by Proposition 5 in their paper.

- We assume homoskedastic errors and estimate the noise variance σ^2 as follows:

$$\hat{\sigma}_T^2 = \frac{1}{T} \sum_{t=1}^{T} (R_t - Z_t^\top \hat{\theta}_T^{LS})^2.$$

- To construct confidence ellipsoids for θ^* are constructed using a Hotelling t-squared statistic as follows:

$$(\hat{\theta}^d - \theta)^\top V_T^{-1} (\hat{\theta}^d - \theta) \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha),$$

where $V_T = \hat{\sigma}_T^2 \sum_{t=1}^T W_t W_t^\top$.

- To construct confidence ellipsoids for θ_1^* with the following confidence ellipsoid where $V_{T,1}$ is the lower right $d_1 \times d_1$ block of matrix V_T:

$$(\hat{\theta}_1^d - \theta_1)^\top V_{T,1}^{-1} (\hat{\theta}_1^{d_1} - \theta_1) \leq \frac{d(T-1)}{T-d_1} F_{d_1,T-d_1}(1-\alpha).$$

A.3.4 Self-Normalized Martingale Bound

We construct $1-\alpha$ confidence region using the following equation taken from Theorem 2 of Abbasi-Yadkori et al. (2011):

$$(\hat{\theta}_T - \theta)^\top V_T (\hat{\theta}_T - \theta) \leq \sigma \sqrt{2 \log \left(\frac{\det(V_T)^{1/2} \det(\lambda I_d)^{-1/2}}{\alpha} \right)} + \lambda^{1/2} S.$$

- $\hat{\theta}_T = \left(\lambda I_d + \sum_{t=1}^T Z_t Z_t^\top \right)^{-1} \sum_{t=1}^T Z_t R_t$.

- $V_T = I_d \lambda + \sum_{t=1}^T Z_t Z_t^\top$.

- $\lambda = 1$ (ridge regression regularization parameter).

- $\sigma = 1$ (assumes rewards are σ-subgaussian).

- $S = 6$, where it is assumed that $\|\theta^*\| \leq S$ (recall that in our simulations $\theta^* \in \mathbb{R}^6$).

- For constructing confidence regions for θ_1^*, we use projected confidence regions.

A.3.5 Construction of Projected Confidence Regions

We are interested in getting the confidence ellipsoid of the projection of a d-dimensional ellipsoid onto p-dimensional space.

- Defining the original d-dimensional ellipsoid, for $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{B} \in \mathbb{R}^{d\times d}$:

$$\mathbf{x}^\top \mathbf{B}\mathbf{x} = 1$$

- Partitioning the matrix $\mathbf{B}$ and vector $\mathbf{x}$:
  For $y \in \mathbb{R}^{d-p}$ and $z \in \mathbb{R}^p$.

$$\mathbf{x} = \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}$$

  For $\mathbf{C} \in \mathbb{R}^{d-p\times d-p}$, $\mathbf{E} \in \mathbb{R}^{p\times p}$, and $\mathbf{D} \in \mathbb{R}^{d-p\times p}$.

$$\mathbf{B} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix}$$

- Gradient of $\mathbf{x}^\top \mathbf{B}\mathbf{x}$ with respect to $\mathbf{x}$:

$$(\mathbf{B} + \mathbf{B}^\top)\mathbf{x} = 2\mathbf{B}\mathbf{x} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix}.$$

  Since we are projecting onto the p-dimensional space, our projection is such that the gradient of $\mathbf{x}^\top \mathbf{B}\mathbf{x}$ with respect to $\mathbf{y}$ is zero, which means

$$\mathbf{C}\mathbf{y} + \mathbf{D}\mathbf{z} = 0.$$

  This means in the projection that $\mathbf{y} = -\mathbf{C}^{-1}\mathbf{D}\mathbf{z}$.

- Returning to our definition of the ellipsoid, plugging in $\mathbf{z}$, we have that

$$1 = \mathbf{x}^\top \mathbf{B}\mathbf{x} = \begin{bmatrix} \mathbf{y}^\top & \mathbf{z}^\top \end{bmatrix} \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{E} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \end{bmatrix} = \mathbf{y}^\top \mathbf{C}\mathbf{y} + 2\mathbf{z}^\top \mathbf{D}^\top \mathbf{y} + \mathbf{z}^\top \mathbf{E}\mathbf{z}$$

$$\begin{aligned} &= (\mathbf{C}^{-1}\mathbf{D}\mathbf{z})^\top \mathbf{C}(\mathbf{C}^{-1}\mathbf{D}\mathbf{z}) - 2\mathbf{z}^\top \mathbf{D}^\top (\mathbf{C}^{-1}\mathbf{D}\mathbf{z}) + \mathbf{z}^\top \mathbf{E}\mathbf{z} \\ &= \mathbf{z}^\top \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D}\mathbf{z} - 2\mathbf{z}^\top \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D}\mathbf{z} + \mathbf{z}^\top \mathbf{E}\mathbf{z} \\ &= \mathbf{z}^\top (\mathbf{E} - \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D})\mathbf{z}. \end{aligned}$$

  Thus the equation for the final projected ellipsoid is

$$\mathbf{z}^\top (\mathbf{E} - \mathbf{D}^\top \mathbf{C}^{-1}\mathbf{D})\mathbf{z} = 1.$$

Appendix B. Asymptotic Results

B.1 Definitions

Definition 3 (Uniform Convergence in Probability) *We say that $Z_T \overset{P}{\to} c$ uniformly over Θ as $T \to \infty$ if*

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta \left(|Z_T - c| > \epsilon \right) \to 0. \tag{14}$$

Definition 4 (Uniform Convergence in Distribution) *We say that $Z_T \overset{D}{\to} Z$ uniformly over Θ as $T \to \infty$ if*

$$\sup_{\theta \in \Theta} \sup_{f \in B_1} \left| \mathbb{E}_\theta[f(Z_T)] - \mathbb{E}_\theta[f(Z)] \right| \to 0, \tag{15}$$

where B_1 represents the space of bounded continuous functions with $\|f\|_\infty \leq 1$ for all $f \in B_1$.

As discussed in Kasy (2019), (14) holds if and only if for any sequence $\{\theta_t\}_{t \geq 1}$ where $\theta_t \in \Theta$ for all $t \geq 1$, $\mathbb{P}_{\theta_t} \left(|Z_T - c| > \epsilon \right) \to 0$. Similarly, (15) holds if and only if for any sequence $\{\theta_t\}_{t \geq 1}$ where $\theta_t \in \Theta$ for all $t \geq 1$, $\sup_{f \in B_1} \left| \mathbb{E}_{\theta_t}[f(Z_T)] - \mathbb{E}_{\theta_t}[f(Z)] \right| \to 0$.

B.2 Lemmas

Lemma 5 *Let $f(Y_t, X_t, A_t)$ be a function such that $E_{\theta^*, \pi_t^{\text{eval}}} \left[f_t^2 \right] < m$ for all $\theta^* \in \Theta$, $t \geq 1$, and $\{\pi_t^{\text{eval}} : t \geq 1\}$. Under conditions 1 and 9,*

$$\frac{1}{T} \sum_{t=1}^T \left\{ W_t f(Y_t, X_t, A_t) - E[W_t f(Y_t, X_t, A_t) | \mathcal{H}_{t-1}] \right\} = O_p \left(\frac{1}{\sqrt{T}} \right) \text{ uniformly over } \theta^* \in \Theta. \tag{16}$$

Lemma 5 is a type of martingale weak law of large number result and the proof is similar to the weak law of large numbers proofs for i.i.d. random variables.

Proof: For notational convenience, let $f_t := f(Y_t, X_t, A_t)$.

$$\sup_{\theta^* \in \Theta} P_{\theta^*} \left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^T \left\{ W_t f_t - E[W_t f_t | \mathcal{H}_{t-1}] \right\} \right| > \epsilon \right)$$

$$\underset{(a)}{\leq} \frac{1}{T \epsilon^2} \sup_{\theta^* \in \Theta} E_{\theta^*} \left[\left(\sum_{t=1}^T \left\{ W_t f_t - E[W_t f_t | \mathcal{H}_{t-1}] \right\} \right)^2 \right]$$

$$\underset{(b)}{=} \frac{1}{T \epsilon^2} \sup_{\theta^* \in \Theta} \sum_{t=1}^T E_{\theta^*} \left[\left\{ W_t f_t - E[W_t f_t | \mathcal{H}_{t-1}] \right\}^2 \right]$$

$$\underset{(c)}{\leq} \frac{1}{T \epsilon^2} \sup_{\theta^* \in \Theta} \sum_{t=1}^T E_{\theta^*} \left[W_t^2 f_t^2 \right] \underset{(d)}{=} \frac{1}{T \epsilon^2} \sup_{\theta^* \in \Theta} \sum_{t=1}^T E_{\theta^*} \left[\int_{a \in \mathcal{A}} W_t^2 \pi_{t,a} E[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]$$

$$\underset{(e)}{=} \frac{1}{T\epsilon^2} \sup_{\theta^* \in \Theta} \sum_{t=1}^{T} E_{\theta^*} \left[\int_{a \in \mathcal{A}} \pi_{t,a}^{\text{eval}} E[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]$$

$$\underset{(f)}{=} \frac{1}{T\epsilon^2} \sup_{\theta^* \in \Theta} \sum_{t=1}^{T} E_{\theta^*, \pi_t^{\text{eval}}} \left[f_t^2 \right] \underset{(g)}{\leq} \frac{4m}{\epsilon^2}$$

- Above (a) holds by Chebyshev's inequality.
- (b) holds because the above terms form a martingale difference sequence with respect to $\mathcal{H}_{t-1}$, i.e., $E[W_t f_t - E[W_t f_t | \mathcal{H}_{t-1}] | \mathcal{H}_{t-1}] = 0$.
- (c) holds because $E\left[\{W_t f_t - E[W_t f_t|\mathcal{H}_{t-1}]\}^2\right] = E\left[W_t^2 f_t^2\right] - E\left[E[W_t f_t|\mathcal{H}_{t-1}]^2\right] \leq E\left[W_t^2 f_t^2\right]$.
- (d) holds by law of iterated expectation.
- (e) holds because $W_t = \sqrt{\frac{\pi_{t,a}^{\text{eval}}}{\pi_{t,a}}}$ by condition 9.
- (f) holds since $E_{\pi_t^{\text{eval}}}\left[f_t^2\right] = E\left[\int_{a \in \mathcal{A}} E[f_t^2|\mathcal{H}_{t-1}, X_t, A_t = a] \pi_{t,a}^{\text{eval}} da\right]$.
- (g) holds since $E_{\theta^*, \pi_t^{\text{eval}}}\left[f_t^2\right] < m < \infty$ for all $\theta^* \in \Theta$, $t \geq 1$, and $\{\pi_t^{\text{eval}} : t \geq 1\}$.

Lemma 6 *Let $m_{\theta,t} := m_\theta(Y_t, X_t, A_t)$. Under Conditions 1, 3, 4, 7, and 9 for any $\delta > 0$,*

$$\sup_{\theta^* \in \Theta} P_{\theta^*} \left(\sup_{\theta \in \Theta} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(W_t m_{\theta,t} - E[W_t m_{\theta,t} | \mathcal{H}_{t-1}] \right) \right\} > \delta \right) \to 0. \tag{17}$$

Lemma 5 is a type of martingale functionally uniform law of large number result and the proof is similar to the functionally uniform law of large numbers proofs for i.i.d. random variables.

Proof:

Finite Bracketing Number: First we show that for any $\epsilon > 0$, there exists bracketing functions $\mathcal{F}_\epsilon$ such that

1. $|\mathcal{F}_\epsilon| < \infty$.
2. For any $f \in \mathcal{F}_\epsilon$, $\sup_{\theta^* \in \Theta, t \geq 1} E_{\theta^*, \pi_t^{\text{eval}}}[f_{\theta^*}(Y_t, X_t, A_t)^2] < m$ for some m which does not depend on ϵ.
3. For any $\theta \in \Theta$, we can find a bracket $l_i, u_i \in \mathcal{F}_\epsilon$ such that (1) $l_i \leq m_\theta \leq u_i$ and (2) $\sup_{t \geq 1, \theta^* \in \Theta} E_{\theta^*, \pi_t^{\text{eval}}}[u_i - l_i] < \epsilon$.

We follow a similar argument to Example 19.7 of Van der Vaart (2000) (page 271).

Make a grid over Θ with meshwidth $\epsilon'/2 > 0$ and let the points in this grid be in set $G_{\epsilon'/2} \subseteq \Theta$; we will specify $\epsilon' > 0$ later. We now show that we can choose $\mathcal{F}_\epsilon = \left\{m_\theta : \theta \in G_{\epsilon'/2}\right\}$. Property 1 above that $|\mathcal{F}_\epsilon| < \infty$ holds by condition 3.

We now show that property 2 above holds. By condition 5 we know that the second moment of $\sup_{\theta^* \in \Theta, t \geq 1} E_{\theta^*, \pi_t^{\text{eval}}} \left[\| m_{\theta^*}(Y_t, X_t, A_t) \|_2^2 \right]$ is bounded. By our Lipschitz condition 4, we have that for any $\theta \in \Theta$, $|m_\theta(Y_t, X_t, A_t) - m_{\theta^*}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t) \| \theta - \theta^* \|$ for g such that $\sup_{\theta^* \in \Theta, t \geq 1} E_{\theta^*, \pi_t^{\text{eval}}} [g(Y_t, X_t, A_t)^2]$ is bounded. By condition $\sup_{\theta \in \Theta} \| \theta - \theta^* \|$ is bounded by condition 3, property 2 above holds.

For property 3 above, first note that for any $\theta \in \Theta$ we can find a $\theta' \in G_{\epsilon'/2}$ in this grid such that $\| \theta' - \theta \| \leq \epsilon'$. Also, by the Lipschitz condition (condition 4), $|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t) \| \theta - \theta' \| \leq g(Y_t, X_t, A_t) \epsilon'$. Thus we have that

$$m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\epsilon' \leq m_\theta(Y_t, X_t, A_t) \leq m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\epsilon'$$

Also note that the brackets $[m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\epsilon, m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\epsilon]$ are such that

$$\sup_{\theta^* \in \Theta, t \geq 1} E_{\theta^*, \pi_t^{\text{eval}}} \left[m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\epsilon' - \left(m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\epsilon' \right) \right]$$

Also since by condition 4 g such that $E_{\theta^*, \pi_t^{\text{eval}}}[g(Y_t, X_t, A_t)^2]$ is uniformly bounded over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$, for some constant m_g,

$$= 2\epsilon' \sup_{\theta^* \in \Theta; t \geq 1} E_{\theta^*, \pi_t^{\text{eval}}} [g(Y_t, X_t, A_t)] < 2\epsilon' m_g < \infty$$

Property 3 above holds by choosing $\epsilon' = \epsilon/(2m_g)$.

Main Argument: Let $\epsilon > 0$; we will choose ϵ later. Let $\mathcal{F}_\epsilon$ be the bracketing functions for $\{m_\theta : \theta \in \Theta\}$ as constructed previously. By property 3 of the bracketing functions, for any $\theta \in \Theta$, we can find a bracket $l_i, u_i \in \mathcal{F}_\epsilon$ such that (1) $l_i \leq m_\theta \leq u_i$ and (2) $\sup_{t \geq 1, \theta^* \in \Theta} E_{\theta^*, \pi_t^{\text{eval}}} [u_i - l_i] < \epsilon$. We define $u_{t,i} := u_i(Y_t, X_t, A_t)$ and $l_{t,i} := l_i(Y_t, X_t, A_t)$. By these bracketing functions, we have that

$$\sup_{\theta \in \Theta} \left\{ \frac{1}{T} \sum_{t=1}^T \left(W_t m_{\theta,t} - E_{\theta^*}[W_t m_{\theta,t} | \mathcal{H}_{t-1}] \right) \right\}$$
$$\leq \max_{i \in [1\colon |\mathcal{F}_\epsilon|]} \left\{ \frac{1}{T} \sum_{t=1}^T \left(W_t u_{t,i} - E_{\theta^*}[W_t l_{t,i} | \mathcal{H}_{t-1}] \right) \right\}$$

$$\leq \max_{i \in [1\colon |\mathcal{F}_\epsilon|]} \left\{ \frac{1}{T} \sum_{t=1}^T E_{\theta^*}[W_t(u_{t,i} - l_{t,i}) | \mathcal{H}_{t-1}] \right\}$$
$$+ \max_{i \in [1\colon |\mathcal{F}_\epsilon|]} \left\{ \frac{1}{T} \sum_{t=1}^T \left(W_t u_{t,i} - E_{\theta^*}[W_t u_{t,i} | \mathcal{H}_{t-1}] \right) \right\}$$

Note that $E_{\theta^*}[W_t(u_{t,i} - l_{t,i}) | \mathcal{H}_{t-1}] = \sqrt{\frac{\pi_{t,a}}{\pi_{t,a}^{\text{eval}}}} E_{\theta^*, \pi_t^{\text{eval}}}[u_{t,i} - l_{t,i}] \leq \frac{1}{\sqrt{\rho_{\min}}} E_{\theta^*, \pi_t^{\text{eval}}}[u_{t,i} - l_{t,i}]$. By construction of bracketing functions $E_{\theta^*, \pi_t^{\text{eval}}}[u_{t,i} - l_{t,i}] \leq \epsilon$ for all $i \in [1\colon |\mathcal{F}_\epsilon|]$. Also since $\max_{i \in [1\colon n]} \{a_i\} \leq \sum_{i=1}^n |a_i|$,

$$\leq \epsilon + \sum_{i=1}^{|\mathcal{F}_\epsilon|} \left| \frac{1}{T} \sum_{t=1}^T \left(W_t u_{t,i} - E_{\theta^*}[W_t u_{t,i} | \mathcal{H}_{t-1}] \right) \right|$$

By Lemma 5, $\frac{1}{T}\sum_{t=1}^{T}\left(W_t u_{t,i} - E_{\theta^*}[W_t u_{t,i}|\mathcal{H}_{t-1}]\right) = o_p(1)$ uniformly over $\theta^* \in \Theta$ for any fixed $\theta \in \Theta$. Since $|\mathcal{F}_\epsilon| < \infty$,

$$= \epsilon + o_p(1)$$

The final result holds by choosing $\epsilon = \frac{\delta}{2}$.

B.3 Consistency

We prove Theorem 1, i.e., that $\hat{\theta}_T \xrightarrow{P} \theta^*$ uniformly over $\theta^* \in \Theta$.

We abbreviate $m_\theta(Y_t, X_t, A_t)$ with $m_{\theta,t}$. By definition of $\hat{\theta}_T$,

$$\sum_{t=1}^{T} W_t m_{\hat{\theta}_T,t} = \sup_{\theta\in\Theta} \sum_{t=1}^{T} W_t m_{\theta,t} \geq \sum_{t=1}^{T} W_t m_{\theta^*,t}$$

Note that $\|\hat{\theta}_T - \theta^*\|_2 > \epsilon > 0$ implies that

$$\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \sum_{t=1}^{T} W_t m_{\theta,t} = \sup_{\theta\in\Theta} \sum_{t=1}^{T} W_t m_{\theta,t} \geq \sum_{t=1}^{T} W_t m_{\theta^*,t}$$

Thus,

$$\sup_{\theta^*\in\Theta} P_{\theta^*}\left(\|\hat{\theta}_T - \theta^*\|_2 > \epsilon\right) \leq P_{\theta^*}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \sum_{t=1}^{T} W_t m_{\theta,t} \geq \sum_{t=1}^{T} W_t m_{\theta^*,t}\right)$$

$$= \sup_{\theta^*\in\Theta} P_{\theta^*}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t}\right\} - \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta^*,t} \geq 0\right)$$

By adding zero,

$$= \sup_{\theta^*\in\Theta} P_{\theta^*}\Bigg(\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} \pm E_{\theta^*}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\}$$
$$- \frac{1}{T}\sum_{t=1}^{T}\left\{W_t m_{\theta^*,t} \pm E_{\theta^*}[W_t m_{\theta^*,t}|\mathcal{H}_{t-1}]\right\} \geq 0\Bigg)$$

By triangle inequality,

$$\leq \sup_{\theta^*\in\Theta} P_{\theta^*}\Bigg(\underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \left\{\frac{1}{T}\sum_{t=1}^{T}\left(W_t m_{\theta,t} - E_{\theta^*}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right)\right\}}_{(a)}$$
$$+ \underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon} \left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t(m_{\theta,t} - m_{\theta^*,t})\big|\mathcal{H}_{t-1}\right]\right\}}_{(b)}$$
$$\underbrace{- \frac{1}{T}\sum_{t=1}^{T}\left\{W_t m_{\theta^*,t} - E_{\theta^*}[W_t m_{\theta^*,t}|\mathcal{H}_{t-1}]\right\}}_{(c)} \geq 0\Bigg) \to 0 \quad (18)$$

We now show that the limit in equation (18) above holds.

- Regarding term (c), by moment bounds of condition 5 and Lemma 5, $\frac{1}{T}\sum_{t=1}^{T}\left\{W_t m_{\theta^*,t} - E_{\theta^*}[W_t m_{\theta^*,t}|\mathcal{H}_{t-1}]\right\} = o_p(1)$ uniformly over $\theta^* \in \Theta$.
- Regarding term (a), by Lemma 6, $\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(W_t m_{\theta,t} - E_{\theta^*}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right)\right\} = o_p(1)$ uniformly over $\theta^* \in \Theta$.

Thus it is sufficient to show that term (b) is such that

$$\sup_{\theta^*\in\Theta} P_{\theta^*}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}[W_t(m_{\theta^*,t} - m_{\theta,t})|\mathcal{H}_{t-1}]\right\} > 0\right) \to 0$$

Below we rewrite term (b),

$$\sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}[W_t(m_{\theta,t} - m_{\theta^*,t})|\mathcal{H}_{t-1}]\right\}$$

By law of iterated expectations,

$$= \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}} \pi_{t,a} E_{\theta^*}[W_t(m_{\theta,t} - m_{\theta^*,t})|\mathcal{H}_{t-1}, X_t, A_t = a] da \middle| \mathcal{H}_{t-1}\right]\right\}$$

Using condition 1 and the fact that $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ by condition 9,

$$= \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}} \pi_{t,a} W_t E_{\theta^*}[m_{\theta,t} - m_{\theta^*,t}|X_t, A_t = a] da \middle| \mathcal{H}_{t-1}\right]\right\}$$

Since for all $\theta \in \Theta$, $E_{\theta^*}[m_{\theta,t} - m_{\theta^*,t}|X_t, A_t] \leq 0$ with probability 1 by condition 7 and since $0 < \frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by condition 9,

$$\leq \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}} \pi_{t,a} W_t^2 E_{\theta^*}[m_{\theta,t} - m_{\theta^*,t}|X_t, A_t = a] da \middle| \mathcal{H}_{t-1}\right]\right\}$$

Since $W_t^2 = \frac{\pi^{\text{eval}}_{t,A_t}}{\pi_{t,A_t}}$,

$$= \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}} \pi^{\text{eval}}_{t,a} E_{\theta^*}[m_{\theta,t} - m_{\theta^*,t}|X_t, A_t = a] da \middle| \mathcal{H}_{t-1}\right]\right\}$$

By condition 1 and since $\pi_t^{\text{eval}} \in \sigma(\mathcal{H}_0)$,

$$= \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}} \pi^{\text{eval}}_{t,a} E_{\theta^*}[m_{\theta,t} - m_{\theta^*,t}|X_t, A_t = a] da\right]\right\}$$

By law of iterated expectations,

$$= \sup_{\theta\in\bar{\Theta}^*:\|\theta-\theta^*\|_2>\epsilon}\left\{\frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}[m_{\theta,t} - m_{\theta^*,t}]\right\} \leq -\frac{1}{\sqrt{\rho_{\max}}}\delta$$

The last inequality above holds by condition 8.

B.4 Asymptotic Normality

We prove Theorem 2, i.e., that

$$\Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T-\theta^*)\xrightarrow{D}\mathcal{N}\left(0,I_d\right) \text{ uniformly over } \theta^*\in\Theta \tag{19}$$

B.4.1 Main Argument

The two results we show to ensure equation (19) holds are as follows:

$$\Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}\dot{M}_T(\theta^*)\xrightarrow{D}\mathcal{N}\left(0,I_d\right) \text{ uniformly over } \theta^*\in\Theta \tag{20}$$

$$\sup_{\theta\in\Theta:\|\theta-\theta^*\|\leq\epsilon_{\dddot{m}}}\frac{1}{T}\sum_{t=1}^{T}W_t\left\|\dddot{m}_{\theta,t}\right\|_1=O_p(1) \text{ uniformly over } \theta^*\in\Theta \tag{21}$$

Above $\ddot{\epsilon}_{\dddot{m}}>0$ is defined in condition 6. For now, we assume that equation (20) and 21 hold; we will show they hold in sections B.4.2 and B.4.3 respectively. Our argument is based on Theorem of 5.41 of (Van der Vaart, 2000).

By Taylor Expansion we have that for some random $\tilde{\theta}_T$ on the line segment between θ^* and $\hat{\theta}_T$,

$$0=\dot{M}_T(\hat{\theta}_T)=\dot{M}_T(\theta^*)+\ddot{M}_T(\theta^*)(\hat{\theta}_T-\theta^*)+\frac{1}{2}(\hat{\theta}_T-\theta^*)^\top\dddot{M}_T(\tilde{\theta}_T)(\hat{\theta}_T-\theta^*)$$

Rearranging terms,

$$-\frac{1}{\sqrt{T}}\dot{M}_T(\theta^*)=\frac{1}{T}\ddot{M}_T(\theta^*)\sqrt{T}(\hat{\theta}_T-\theta^*)+\frac{1}{2}(\hat{\theta}_T-\theta^*)^\top\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)\sqrt{T}(\hat{\theta}_T-\theta^*)$$

$$=\left[\frac{1}{T}\ddot{M}_T(\theta^*)+\frac{1}{2}(\hat{\theta}_T-\theta^*)^\top\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)\right]\sqrt{T}(\hat{\theta}_T-\theta^*) \tag{22}$$

Showing that $\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)$ is bounded in probability We now show that $\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)=O_p(1)$ uniformly over $\theta^*\in\Theta$. Note that $\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)\in\mathbb{R}^{d\times d\times d}$. Recall that for any $B\in\mathbb{R}^{d\times d\times d}$, we denote $\|B\|_1=\sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d|B_{i,j,k}|$. Note that by triangle inequality, $\left\|\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)\right\|_1=\left\|\frac{1}{T}\sum_{t=1}^T W_t\dddot{m}_{\tilde{\theta}_T,t}\right\|_1\leq\frac{1}{T}\sum_{t=1}^T W_t\left\|\dddot{m}_{\tilde{\theta}_T,t}\right\|_1$. By uniform consistency of $\hat{\theta}_T$, we have $\sup_{\theta^*\in\Theta}P_{\theta^*}\left(\left\|\hat{\theta}_T-\theta^*\right\|_2>\epsilon_{\dddot{m}}\right)\to 0$, for $\epsilon_{\dddot{m}}>0$ defined in condition 6. Thus we have that

$$\left\|\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T)\right\|_1\leq\frac{1}{T}\sum_{t=1}^{T}W_t\left\|\dddot{m}_{\tilde{\theta}_T,t}\right\|_1$$

$$=\left(\mathbb{1}_{\|\hat{\theta}_T-\theta^*\|>\epsilon_{\dddot{m}}}+\mathbb{1}_{\|\hat{\theta}_T-\theta^*\|\leq\epsilon_{\dddot{m}}}\right)\frac{1}{T}\sum_{t=1}^{T}W_t\left\|\dddot{m}_{\tilde{\theta}_T,t}\right\|_1$$

$$=o_p(1)+\mathbb{1}_{\|\hat{\theta}_T-\theta^*\|\leq\epsilon_{\dddot{m}}}\frac{1}{T}\sum_{t=1}^{T}W_t\left\|\dddot{m}_{\tilde{\theta}_T,t}\right\|_1$$

$$\leq o_p(1) + \sup_{\theta \in \Theta : \|\theta - \theta^*\| \leq \epsilon_{\dddot{m}}} \frac{1}{T} \sum_{t=1}^{T} W_t \left\| \dddot{m}_{\theta,t} \right\|_1 = O_p(1) \text{ uniformly over } \theta^* \in \Theta \tag{23}$$

The last equality above holds by equation (21).

Asymptotic normality of $\Sigma_{T,\theta^*}^{-1/2} \frac{1}{\sqrt{T}} \dot{M}_T(\theta^*)$ By uniform consistency of $\hat{\theta}_T$, we have that $(\hat{\theta}_T - \theta^*)^\top \frac{1}{T} \dddot{M}_T(\tilde{\theta}_T) = o_p(1) O_p(1) = o_p(1)$ uniformly over $\theta^* \in \Theta$. Equation (22) can be rewritten as follows:

$$-\frac{1}{\sqrt{T}} \dot{M}_T(\theta^*) = \left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right] \sqrt{T} (\hat{\theta}_T - \theta^*).$$

Multiplying both sides of the above equation by $\Sigma_{T,\theta^*}^{-1/2}$ we have that

$$-\Sigma_{T,\theta^*}^{-1/2} \frac{1}{\sqrt{T}} \dot{M}_T(\theta^*) = \Sigma_{T,\theta^*}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right] \sqrt{T} (\hat{\theta}_T - \theta^*).$$

By equation (20),

$$\begin{aligned} -\Sigma_{T,\theta^*}^{-1/2} \frac{1}{\sqrt{T}} \dot{M}_T(\theta^*) = \Sigma_{T,\theta^*}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right] \sqrt{T} (\hat{\theta}_T - \theta^*) \\ \xrightarrow{D} \mathcal{N}(0, I_p) \text{ uniformly over } \theta^* \in \Theta \end{aligned} \tag{24}$$

Showing that $\sqrt{T}(\hat{\theta}_T - \theta^*)$ is bounded in probability We now show that $\sqrt{T}(\hat{\theta}_T - \theta^*) = O_p(1)$ uniformly over $\theta^* \in \Theta$. Note by the above asymptotic normality result, $\Sigma_{T,\theta^*}^{-1/2} \left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right] \sqrt{T} (\hat{\theta}_T - \theta^*) = O_p(1)$ uniformly over $\theta^* \in \Theta$. By moment bounds of condition 5, $\Sigma_{T,\theta^*}^{1/2}$ is uniformly bounded over $\theta^* \in \Theta$, so $\left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right] \sqrt{T} (\hat{\theta}_T - \theta^*) = O_p(1)$ uniformly over $\theta^* \in \Theta$. Multiplying both sides by $\left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right]^{-1}$, we have $\sqrt{T}(\hat{\theta}_T - \theta^*) = \left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right]^{-1} O_p(1)$ uniformly over $\theta^* \in \Theta$. To show that $\left[\frac{1}{T} \ddot{M}_T(\theta^*) + o_p(1) \right]^{-1} = O_p(1)$ uniformly over $\theta^* \in \Theta$, it is sufficient to show that the minimum eigenvalue of $\frac{1}{T} \ddot{M}_T(\theta^*)$ is bounded above zero uniformly over $\theta^* \in \Theta$ with probability going to 1.

By condition 5 and Lemma 5, $\frac{1}{T} \sum_{t=1}^{T} W_t \ddot{m}_{\theta^*,t} - E_{\theta^*} \left[W_t \ddot{m}_{\theta^*,t} | \mathcal{H}_{t-1} \right] = o_p(1)$ uniformly over $\theta^* \in \Theta$, so

$$\frac{1}{T} \ddot{M}_T(\theta^*) = \frac{1}{T} \sum_{t=1}^{T} W_t \ddot{m}_{\theta^*,t} = o_p(1) + \frac{1}{T} \sum_{t=1}^{T} E_{\theta^*} \left[W_t \ddot{m}_{\theta^*,t} | \mathcal{H}_{t-1} \right]$$

By law of iterated expectations,

$$= o_p(1) + \frac{1}{T} \sum_{t=1}^{T} E_{\theta^*} \left[W_t E_{\theta^*} \left[\ddot{m}_{\theta^*,t} | \mathcal{H}_{t-1}, X_t, A_t \right] | \mathcal{H}_{t-1} \right]$$

By condition 1,

$$= o_p(1) + \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t E_{\theta^*}\left[\ddot{m}_{\theta^*,t}|X_t, A_t\right]|\mathcal{H}_{t-1}\right]$$

Since θ^* is a maximizing value of $E_{\theta^*}[m_{\theta^*,t}|X_t, A_t]$ by condition 7, we have that $E_{\theta^*}[\ddot{m}_{\theta^*,t}|X_t, A_t] \preceq 0$. Also since $\frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by condition 9,

$$\preceq o_p(1) + \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2 E_{\theta^*}\left[\ddot{m}_{\theta^*,t}|X_t, A_t\right]|\mathcal{H}_{t-1}\right]$$

Since $W_t^2 = \frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}$,

$$= o_p(1) + \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}\left[\ddot{m}_{\theta^*,t}|\mathcal{H}_{t-1}\right]$$

Since $\pi_t^{\text{eval}} \in \sigma(\mathcal{H}_0)$,

$$= o_p(1) + \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}\left[\ddot{m}_{\theta^*,t}\right]$$

By condition 7, the minimum of the absolute value of eigenvalues of $\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}[\ddot{m}_{\theta^*,t}]$ is bounded above $\delta_{\ddot{m}} > 0$ for all $\theta^* \in \Theta$ for sufficiently large T. Thus we have shown that $\sqrt{T}(\hat{\theta}_T - \theta^*) = O_p(1)$ uniformly over $\theta^* \in \Theta$.

Final result Now we will show that $\Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_p)$ uniformly over $\theta^* \in \Theta$ [note this differs from (20) since we have $\ddot{M}_T(\hat{\theta}_T)$ rather than $\ddot{M}_T(\theta^*)$].

$$\Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*)$$

$$= \Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right]\sqrt{T}(\hat{\theta}_T - \theta^*) + \Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\theta^*)\sqrt{T}(\hat{\theta}_T - \theta^*)$$

Since by condition 7 the the minimum eigenvalue of Σ_{T,θ^*} is bounded above $\delta_{\dot{m}^2} > 0$ for all $\theta^* \in \Theta$ for sufficiently large T, we have that $\Sigma_{T,\theta^*}^{-1/2} = O(1)$. And since we showed earlier that $\sqrt{T}(\hat{\theta}_T - \theta^*) = O_p(1)$ uniformly over $\theta^* \in \Theta$, we have that $\Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right]\sqrt{T}(\hat{\theta}_T - \theta^*) = O_p(1)\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right]O_p(1)$ uniformly over $\theta^* \in \Theta$. If we can show that $\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right] = o_p(1)$ uniformly over $\theta^* \in \Theta$, this will mean that

$$= o_p(1) + \Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\theta^*)\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_p) \text{ uniformly over } \theta^* \in \Theta$$

where the limit holds by (24). By Slutsky's theorem, the above means that $\Sigma_{T,\theta^*}^{-1/2}\frac{1}{T}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_p)$ uniformly over $\theta^* \in \Theta$.

We now show that $\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right] = o_p(1)$ uniformly over $\theta^* \in \Theta$. By Taylor expansion for random some $\bar{\theta}_T$ on the line segment between θ^* and $\hat{\theta}_T$, we have that

$\frac{1}{T}\left[\ddot{M}_T(\hat{\theta}_T) - \ddot{M}_T(\theta^*)\right] = \frac{1}{T}\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T - \theta^*)$. By the same argument that $\frac{1}{T}\dddot{M}_T(\tilde{\theta}_T) = O_p(1)$ uniformly over $\theta^* \in \Theta$ [see equation (23)], $\frac{1}{T}\dddot{M}_T(\hat{\theta}_T) = O_p(1)$ uniformly over $\theta^* \in \Theta$. The result holds by uniform consistency of $\hat{\theta}_T$, which means that we have that $\hat{\theta}_T - \theta^* = o_p(1)$ uniformly over $\theta^* \in \Theta$.

B.4.2 Asymptotic Normality of $\Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}\dot{M}_T(\theta^*)$

We will show that equation (20) holds by applying the martingale central limit theorem of Dvoretzky (1972). For notational convenience, we let $\dot{m}_{\theta^*,t} := \dot{m}_{\theta^*}(Y_t, X_t, A_t)$. Note that $\Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}\dot{M}_T(\theta^*) = \Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t \dot{m}_{\theta^*,t}$. We first show that $\left\{\Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}W_t\dot{m}_{\theta^*,t}\right\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=0}^T$. For any $t \in \{1\colon T\}$,

$$E_{\theta^*}\left[\frac{1}{\sqrt{T}}\Sigma_{T,\theta^*}^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*,t}\middle|\mathcal{H}_{t-1}\right] \underset{(a)}{=} \frac{1}{\sqrt{T}}E_{\theta^*}\left[E_{\theta^*}\left[\Sigma_{T,\theta^*}^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*,t}\middle|\mathcal{H}_{t-1}, X_t, A_t\right]\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{=} \frac{1}{\sqrt{T}}\Sigma_{T,\theta^*}^{-1/2}E_{\theta^*}\left[W_t\mathbf{c}^\top E_{\theta^*}\left[\dot{m}_{\theta^*,t}\middle|\mathcal{H}_{t-1}, X_t, A_t\right]\middle|\mathcal{H}_{t-1}\right] \underset{(c)}{=} 0$$

- Above, equality (a) holds by law of iterated expectations.
- Equality (b) holds since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ and since $\Sigma_{T,\theta^*} \in \sigma(\mathcal{H}_0)$ because $\{\pi_t^{\text{eval}}\}_{t\geq 1} \in \sigma(\mathcal{H}_0)$.
- By condition 1, $E_{\theta^*}\left[\dot{m}_{\theta^*,t}\middle|\mathcal{H}_{t-1}, X_t, A_t\right] = E_{\theta^*}\left[\dot{m}_{\theta^*,t}\middle|X_t, A_t\right]$. Equality (c) holds because $E_{\theta^*}\left[\dot{m}_{\theta^*,t}\middle|X_t, A_t\right] = 0$ with probability 1 since θ^* is a critical point of $E_{\theta^*}[m_{\theta,t}|X_t, A_t]$ by Condition 7.

We now apply the martingale central limit theorem of Dvoretzky (1972); while the original theorem holds for any fixed parameter, we can show uniform convergence in distribution by ensuring that the conditions of the theorem hold uniformly over $\theta^* \in \Theta$ (see definition 4). We now state the uniform version of the martingale central limit theorem of Dvoretzky (1972). For a martingale difference sequence $\{Z_t\}$ with respect to $\{\mathcal{H}_{t-1}\}_{t=1}^T$, $\frac{1}{\sqrt{T}}\sum_{t=1}^T Z_t \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ uniformly over $\theta^* \in \Theta$ under the following two conditions:

1. **Conditional Variance:** $\frac{1}{T}\sum_{t=1}^T E_{\theta^*}\left[Z_t^2|\mathcal{H}_{t-1}\right] \xrightarrow{P} \sigma^2$ uniformly over $\theta^* \in \Theta$.
2. **Conditional Lindeberg:** For any $\delta > 0$, $\frac{1}{T}\sum_{t=1}^T E_{\theta^*}\left[Z_t^2 \mathbb{1}_{|Z_t|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right] \xrightarrow{P} 0$ uniformly over $\theta^* \in \Theta$.

By Cramer Wold device, to show that (20) holds, it is sufficient to show that for any fixed $\mathbf{c} \in \mathbb{R}^d$ with $\|\mathbf{c}\|_2 = 1$, that $\mathbf{c}^\top\Sigma_{T,\theta^*}^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t\dot{m}_{\theta^*,t} \xrightarrow{D} \mathcal{N}\left(0, \mathbf{c}^\top I_d\mathbf{c}\right)$ uniformly over $\theta^* \in \Theta$.

1. Conditional Variance

$$\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\left(\mathbf{c}^\top W_t \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}\right)^2 \middle| \mathcal{H}_{t-1}\right] = \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2 \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}^{\otimes 2}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(a)}{=} \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1}\right]\right\}\Sigma_{T,\theta^*}^{-1}\mathbf{c}$$

$$\underset{(b)}{=} \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\int_{a\in\mathcal{A}}\frac{\pi_{t,a}^{\text{eval}}}{\pi_{t,a}}\pi_{t,a}E\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1},X_t,A_t=a\right]\middle|\mathcal{H}_{t-1}\right]\right\}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}$$

$$\underset{(c)}{=} \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1}\right]\right\}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c} \underset{(d)}{=} \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\Sigma_{T,\theta^*}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c} = \mathbf{c}^\top I_d \mathbf{c}$$

- Above, equality (a) holds since $W_t = \sqrt{\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}}$ since $\Sigma_{T,\theta^*} \in \sigma(\mathcal{H}_0)$ because $\{\pi_t^{\text{eval}}\}_{t\geq 1} \in \sigma(\mathcal{H}_0)$.
- Equalities (b) and (c) hold by law of iterated expectations.
- Equality (d) holds because for all $t \in [1\colon T]$,

$$E_{\theta^*,\pi_t^{\text{eval}}}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1}\right] = E_{\theta^*,\pi_t^{\text{eval}}}\left[E_{\theta^*}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1},X_t,A_t\right]\middle|\mathcal{H}_{t-1}\right]$$

By condition 1, $E_{\theta^*}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|\mathcal{H}_{t-1},X_t,A_t\right] = E_{\theta^*}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|X_t,A_t\right]$,

$$= E_{\theta^*,\pi_t^{\text{eval}}}\left[E_{\theta^*}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|X_t,A_t\right]\middle|\mathcal{H}_{t-1}\right] = E_{\theta^*,\pi_t^{\text{eval}}}\left[E_{\theta^*}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\middle|X_t,A_t\right]\right] = E_{\theta^*,\pi_t^{\text{eval}}}\left[\dot{m}_{\theta^*,t}^{\otimes 2}\right].$$

2. Lindeberg Condition

$$\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[\left(\mathbf{c}^\top W_t \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}\right)^2 \mathbb{1}_{\left|\mathbf{c}^\top W_t \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}\right|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right]$$

$$= \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2 \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}^{\otimes 2}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}\mathbb{1}_{\left|\mathbf{c}^\top W_t \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}\right|>\delta\sqrt{T}}\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(a)}{\leq} \frac{1}{T^2\delta^2}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^4\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}^{\otimes 2}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(b)}{\leq} \frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}^{\otimes 2}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(c)}{=} \frac{\rho_{\max}}{T^2\delta^2}\sum_{t=1}^{T} E\left[\int_{a\in\mathcal{A}}\pi_{t,a}E\left[W_t^2\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2}\dot{m}_{\theta^*,t}^{\otimes 2}\Sigma_{T,\theta^*}^{-1/2}\mathbf{c}\right)^2\middle|\mathcal{H}_{t-1},X_t,A_t=a\right]da\middle|\mathcal{H}_{t-1}\right]$$

$$\underset{(d)}{=} \frac{\rho_{\max}}{T^2\delta^2} \sum_{t=1}^T E\left[\int_{a \in \mathcal{A}} \pi_{t,a} \left(\frac{\pi_{t,a}^{\text{eval}}}{\pi_{t,a}} \right) E\left[\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} \right)^2 \middle| \mathcal{H}_{t-1}, X_t, A_t = a \right] da \middle| \mathcal{H}_{t-1} \right]$$

$$\underset{(e)}{=} \frac{\rho_{\max}}{T^2\delta^2} \sum_{t=1}^T E_{\theta^*,\pi_t^{\text{eval}}} \left[\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} \right)^2 \middle| \mathcal{H}_{t-1} \right]$$

$$\underset{(f)}{=} \frac{\rho_{\max}}{T^2\delta^2} \sum_{t=1}^T E_{\theta^*,\pi_t^{\text{eval}}} \left[\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} \right)^2 \right] \underset{(g)}{\to} 0$$

- Above, inequality (a) holds because $\mathbb{1}_{\left| W_t \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t} \right| > \sqrt{T}\delta} = 1$ if and only if $W_t^2 \frac{1}{T\delta^2} \mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} > 1$.
- Inequality (b) holds because by condition 9, $W_t \leq \rho_{\max}$ with probability 1.
- Equality (c) holds by law of iterated expectations.
- Equality (d) holds since $W_t = \sqrt{\frac{\pi_{t,A_t}^{\text{eval}}}{\pi_{t,A_t}}} \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ by condition 9.
- Equality (e) holds by law of iterated expectations.
- Equality (f) holds by condition 1 (see argument for equality (d) under Conditional Variance condition for more details).
- Regarding limit (g), it is sufficient to show that $\frac{1}{T} \sum_{t=1}^T E_{\theta^*,\pi_t^{\text{eval}}} \left[\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$ for sufficiently large T. By condition 5, the minimum eigenvalue of Σ_{T,θ^*}^2 is bounded above zero uniformly over $\theta^* \in \Theta$ for sufficiently large T. Also by condition 5 the fourth moment of $\dot{m}_{\theta^*,t}$ with respect to $P_{\theta^*,\pi_t^{\text{eval}}}$ is uniformly bounded over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$. With these two properties we have that $\frac{1}{T} \sum_{t=1}^T E_{\theta^*,\pi_t^{\text{eval}}} \left[\left(\mathbf{c}^\top \Sigma_{T,\theta^*}^{-1/2} \dot{m}_{\theta^*,t}^{\otimes 2} \Sigma_{T,\theta^*}^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\theta^* \in \Theta$ and $\{\pi_t^{\text{eval}} : t \geq 1\}$ for sufficiently large T.

B.4.3 Third Derivative Term

We now show that (21) holds.

$$\sup_{\theta \in \Theta : \|\theta - \theta^*\|_2 \leq \epsilon_{\dddot{m}}} \frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}_{\theta,t} \right\|_1$$

By condition 6, for all $\theta \in \Theta$ such that $\|\theta - \theta^*\|_2 \leq \epsilon_{\dddot{m}}$, $\|\dddot{m}_{\theta,t}\|_1 \leq \|\dddot{m}(Y_t, X_t, A_t)\|_1$ with probability 1.

$$\leq \frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1$$

$$= \frac{1}{T}\sum_{t=1}^{T} W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 \pm \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 |\mathcal{H}_{t-1}\right]$$

By second moment bounds on $\|\dddot{m}(Y_t, X_t, A_t)\|_1$ by condition 6, by Lemma 5, we have that $\frac{1}{T}\sum_{t=1}^{T} W_t \|\dddot{m}(Y_t, X_t, A_t)\|_1 - E_{\theta^*}\left[W_t \|\dddot{m}(Y_t, X_t, A_t)\|_1 |\mathcal{H}_{t-1}\right] = o_p(1)$ uniformly over $\theta^* \in \Theta$.

$$= o_p(1) + \frac{1}{T}\sum_{t=1}^{T} E_{\theta^*}\left[W_t \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 |\mathcal{H}_{t-1}\right]$$

Since $\frac{W_t}{\sqrt{\rho_{\min}}} \geq 1$ with probability 1, by condition 9,

$$\leq o_p(1) + \frac{1}{T\sqrt{\rho_{\min}}}\sum_{t=1}^{T} E_{\theta^*}\left[W_t^2 \left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1 |\mathcal{H}_{t-1}\right]$$

Since $W_t^2 = \frac{\pi^{\text{eval}}_{t,A_t}}{\pi_{t,A_t}}$,

$$= o_p(1) + \frac{1}{T\sqrt{\rho_{\min}}}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}\left[\left\|\dddot{m}(Y_t, X_t, A_t)\right\|_1\right] = O_p(1) \text{ uniformly over } \theta^* \in \Theta$$

The final limit holds because $\frac{1}{T}\sum_{t=1}^{T} E_{\theta^*,\pi_t^{\text{eval}}}\left[\|\dddot{m}(Y_t, X_t, A_t)\|_1\right]$ is bounded by condition 6.

Appendix C. Derivations Regarding Choice of Evaluation Policy

The least squares estimator has function $m_\theta = -\frac{1}{2}(Y_t - \vec{A}_t^\top \theta)^2$, $\dot{m}_\theta = (Y_t - \vec{A}_t^\top \theta)\vec{A}_t$, and $\ddot{m}_\theta = -\vec{A}_t^\top \vec{A}_t$. Thus by Theorem 2 the asymptotic variance of $\hat{\theta}^{\text{AW-LS}}$ is as follows:

$$\left[\frac{1}{T}\sum_{t=1}^T W_t \ddot{m}_{\theta^*}(Y_t, X_t, A_t)\right]^{-1} \frac{1}{T}\sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}}\left[\dot{m}_{\theta^*}(Y_t, X_t, A_t)^{\otimes 2}\right] \left[\frac{1}{T}\sum_{t=1}^T W_t \ddot{m}_{\theta^*}(Y_t, X_t, A_t)\right]^{-1}$$

$$\underset{(a)}{=} \left[\frac{1}{T}\sum_{t=1}^T W_t \vec{A}_t^\top \vec{A}_t\right]^{-1} \frac{1}{T}\sum_{t=1}^T E_{\theta^*, \pi_t^{\text{eval}}}\left[\vec{A}_t^\top \vec{A}_t (Y_t - \vec{A}_t^\top \theta^*)^2\right] \left[\frac{1}{T}\sum_{t=1}^T W_t \vec{A}_t^\top \vec{A}_t\right]^{-1}.$$

(a) holds by replacing $\dot{m}_\theta = (Y_t - \vec{A}_t^\top \theta)\vec{A}_t$ and $\ddot{m}_\theta = -\vec{A}_t^\top \vec{A}_t$. Recall that $E[(Y_t - \vec{A}_t^\top \theta)^2 | A_t = a] = \sigma_a^2$ and define $\vec{a} = [\mathbb{1}_{a=1}, \mathbb{1}_{a=2}, ..., \mathbb{1}_{a=|\mathcal{A}|}]^\top$.

$$= \left[\sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \frac{1}{T}\sum_{t=1}^T W_t \mathbb{1}_{A_t=a}\right]^{-1} \sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}} \left[\sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \frac{1}{T}\sum_{t=1}^T W_t \mathbb{1}_{A_t=a}\right]^{-1}.$$

Since $\vec{a}^\top \vec{a}$ is a matrix with one diagonal term non-zero,

$$= \sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(\frac{1}{T}\sum_{t=1}^T W_t \mathbb{1}_{A_t=a}\right)^2} \underset{(b)}{=} \sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(o_p(1) + \frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}\pi_{t,a}^{\text{eval}}}\right)^2}.$$

(b) holds because by weighted martingale weak law of large numbers (see Lemma 5), $\frac{1}{T}\sum_{t=1}^T \left(W_t \mathbb{1}_{A_t=a} - \sqrt{\pi_{t,a}^{\text{eval}}\pi_{t,a}}\right) = o_p(1)$. By Cauchy-Schwartz inequality, $\left(\frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}^{\text{eval}}\pi_{t,a}}\right)^2 \leq \left(\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}\right)\left(\frac{1}{T}\sum_{t=1}^T \pi_{t,a}\right)$. Thus, $\frac{1}{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}} \leq \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(\frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}^{\text{eval}}\pi_{t,a}}\right)^2}$. Thus we have that

$$\sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}^{\text{eval}}}{\left(\frac{1}{T}\sum_{t=1}^T \sqrt{\pi_{t,a}\pi_{t,a}^{\text{eval}}}\right)^2} \succeq \sum_{a=1}^{|\mathcal{A}|} \vec{a}^\top \vec{a} \sigma_a^2 \frac{1}{\frac{1}{T}\sum_{t=1}^T \pi_{t,a}}.$$

Appendix D. Need for Uniformly Valid Inference on Bandit Data

For the unweighted least squares estimator in a two-arm bandit setting, the least squares estimator is asymptotically normal on adaptively collected data under the following condition of Lai and Wei (1982),

$$b_T \cdot \sum_{t=1}^{T} A_t \xrightarrow{P} 1. \tag{25}$$

Specifically, by Theorem 3 of Lai and Wei (1982), assuming i.i.d. potential outcomes $\{R_t(1), R_t(0)\}$ such that $E_{\theta^*}[R_t|A_t] = A_t\theta_1^* + (1 - A_t)\theta_0^*$, $\text{Var}(R_t(a)) = \sigma_a$, and $E[R_t(a)^4] < c < \infty$, under (25),

$$\sqrt{\sum_{t=1}^{T} A_t}(\hat{\theta}_1 - \theta_1^*) = \frac{\sum_{t=1}^{T} A_t(R_t - \theta_1^*)}{\sqrt{\sum_{t=1}^{T} A_t}} \xrightarrow{D} \mathcal{N}(0, \sigma^2).$$

However, as discussed in Deshpande et al. (2018) and Zhang et al. (2020), (25) can fail to to hold for common bandit algorithms when there is no unique optimal policy, particularly when $\theta_1^* - \theta_1^* = \Delta^* = 0$. For example, in Figure 12 we plot $\frac{1}{T}\sum_{t=1}^{T} A_t$ for Thompson Sampling and epsilon greedy under zero margin, $\Delta^* = 0$.

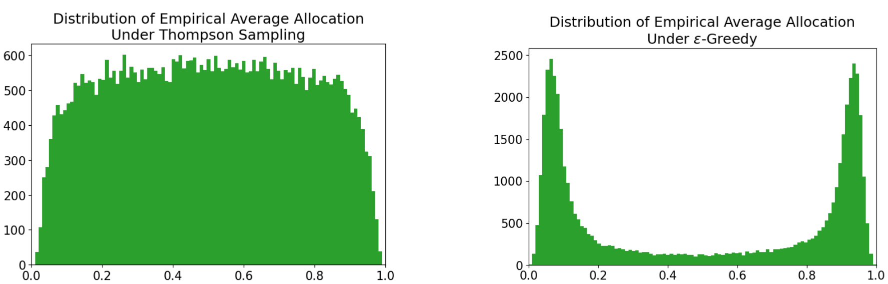

Figure 12: Above we plot empirical allocations $\frac{1}{T}\sum_{t=1}^{T} A_t$ under both Thompson Sampling (standard normal priors, 0.01 clipping) and ϵ-greedy ($\epsilon = 0.1$) under zero margin $\theta_1^* = \theta_1^* = \Delta^* = 0$. For our simulations $T = 100$, errors are standard normal, and we use $50k$ Monte Carlo repetitions.

In order to construct reliable confidence intervals using asymptotic approximations, it is crucial that they converge uniformly in distribution. To illustrate the importance of uniformity, consider the following example. We can modify Thompson Sampling to ensure that $\frac{1}{T}\sum_{t=1}^{T} A_t \xrightarrow{P} 0.5$ when $\Delta^* = 0$. For example, we could do this by using an algorithm we call Thompson Sampling Hodges (inspired by the Hodges estimator; see Van der Vaart (2000) page 109), defined below:

$$\pi_t = P(\tilde{\theta}_1 > \tilde{\theta}_0|\mathcal{H}_{t-1})\mathbb{1}_{|\mu_{1,t}-\mu_{0,t}|>t^{-4}} + 0.5\mathbb{1}_{|\mu_{1,t}-\mu_{0,t}|\leq t^{-4}}$$

Under standard Thompson Sampling arm one is chosen according to the posterior probability that is optimal, so $\pi_t = P(\tilde{\theta}_1 > \tilde{\theta}_0|\mathcal{H}_{t-1})$. Above, $\mu_{a,t}$ denotes the posterior mean for θ_a at time t. Under TS-Hodges, if difference between the posterior means, $|\mu_{1,t} - \mu_{0,t}|$, is less than t^{-4}, π_t is set to 0.5. Additionally, we clip the action selection probabilities to bound them strictly away from 0 and 1 for some constant $\pi_{\min}$ in the following sense $\text{clip}(\pi_t) = (1 - \pi_{\min}) \wedge (\pi_t \vee \pi_{\min})$. Under TS-Hodges with clipping, we can show that

$$\frac{1}{T} \sum_{t=1}^{T} A_t \xrightarrow{P} \begin{cases} 1 - \pi_{\min} & \text{if } \Delta^* > 0 \\ \pi_{\min} & \text{if } \Delta^* < 0 \\ 0.5 & \text{if } \Delta^* = 0 \end{cases} \tag{26}$$

By equation (26), we satisfy (25) pointwise for every fixed θ^* and we have that the OLS estimator is asymptotically normal *pointwise* (Lai and Wei, 1982). However, equation (26) fails to hold uniformly over θ^*. Specifically, it fails to hold for any sequence of $\{\theta_t^*\}_{t=1}^{\infty}$ such that $\Delta_t^* = t^{-4}$. In Figure 13, we show that confidence intervals constructed using normal approximations fail to provide reliable confidence intervals, even for very large sample sizes for the worst case values of Δ^*.

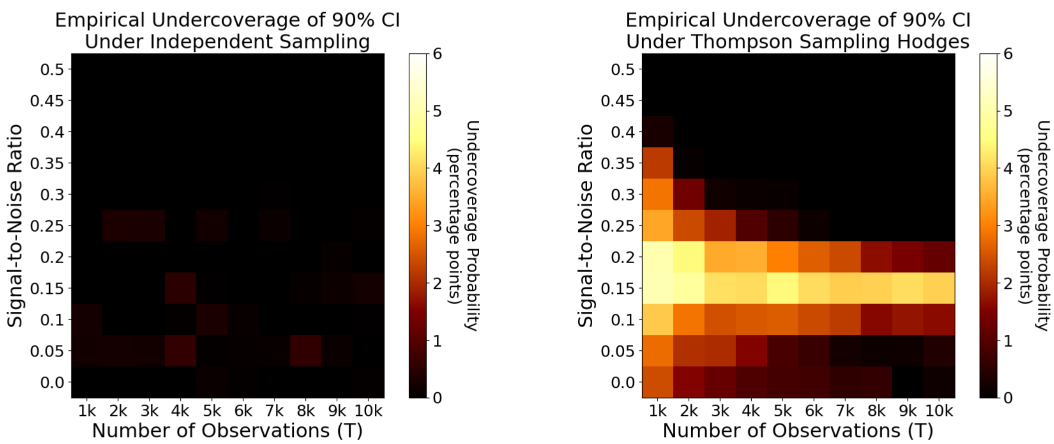

Figure 13: Above we construct confidence intervals for Δ^* using a normal approximation for the OLS estimator. We compare independent sampling ($\pi_t = 0.5$) and TS Hodges, both with standard normal priors, 0.01 clipping, standard normal errors, and $T = 10,000$. We vary the value of Δ^* in the simulations to demonstrate the non-uniformity of the confidence intervals.