Statistical Inference with M-Estimators on Adaptively Collected Data

Kelly W. Zhang
Department of Computer Science
Harvard University
kellywzhang@seas.harvard.edu

Lucas Janson
Departments of Statistics
Harvard University
ljanson@fas.harvard.edu

Susan A. Murphy
Departments of Statistics and Computer Science
Harvard University
samurphy@fas.harvard.edu

Abstract

Bandit algorithms are increasingly used in real-world sequential decision-making problems. Associated with this is an increased desire to be able to use the resulting datasets to answer scientific questions like: Did one type of ad lead to more purchases? In which contexts is a mobile health intervention effective? However, classical statistical approaches fail to provide valid confidence intervals when used with data collected with bandit algorithms. Alternative methods have recently been developed for simple models (e.g., comparison of means). Yet there is a lack of general methods for conducting statistical inference using more complex models on data collected with (contextual) bandit algorithms; for example, current methods cannot be used for valid inference on parameters in a logistic regression model for a binary reward. In this work, we develop theory justifying the use of M-estimators—which includes estimators based on empirical risk minimization as well as maximum likelihood—on data collected with adaptive algorithms, including (contextual) bandit algorithms. Specifically, we show that M-estimators, modified with particular adaptive weights, can be used to construct asymptotically valid confidence regions for a variety of inferential targets.

1 Introduction

Due to the need for interventions that are personalized to users, (contextual) bandit algorithms are increasingly used to address sequential decision making problems in health-care [Yom-Tov et al., 2017, Liao et al., 2020], online education [Liu et al., 2014, Shaikh et al., 2019], and public policy [Kasy and Sautmann, 2021, Caria et al., 2020]. Contextual bandits personalize, that is, minimize regret, by learning to choose the best intervention in each context, i.e., the action that leads to the greatest expected reward. Besides the goal of regret minimization, 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 mean information gained by using the data to conduct a variety of statistical analyses, including confidence interval construction and hypothesis testing. **While regret minimization 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. Note that the data collected by bandit algorithms are *adaptively collected* because previously observed contexts, actions, and rewards are used to inform what actions to select in future timesteps.

Preprint. Under review.

There are a variety of between-experiment learning questions encountered in real-life applications of bandit algorithms. For example, in real-life sequential decision-making problems there are often a number of additional scientifically interesting outcomes besides the reward that are collected during the experiment. 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 that an ad had high click-through rate, but low amounts of money was spent after clicking on the ad, one may redesign the reward used in the next bandit experiment. One type of statistical 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 online. 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 variables are accessible to the bandit algorithm in the next experiment.

As discussed above, we can gain scientific knowledge from data collected with (contextual) bandit algorithms by constructing confidence intervals and performing hypothesis tests for unknown quantities such as the expected outcome for different actions in various contexts. Unfortunately, standard statistical methods developed for i.i.d. data fail to provide valid inference when applied to data collected with common bandit algorithms. For example, assuming the sample mean of rewards for an arm is approximately normal can lead to unreliable confidence intervals and inflated type-1 error; see Section 3.1 for an illustration. Recently statistical inference methods have been developed for data collected using bandit algorithms [Hadad et al., 2019, Deshpande et al., 2018, Zhang et al., 2020]; however, these methods are limited to inference for parameters of simple models. There is a lack of general statistical inference methods for data collected with (contextual) bandit algorithms in more complex data-analytic settings, including parameters in non-linear models for outcomes; for example, there are currently no methods for constructing valid confidence intervals for the parameters of a logistic regression model for binary outcomes or for constructing confidence intervals based on robust estimators like minimizers of the Huber loss function.

In this work we show that a wide variety of estimators which are frequently used both in science and industry on i.i.d. data, namely, M-estimators [Van der Vaart, 2000], can be used to conduct valid inference on data collected with (contextual) bandit algorithms when adjusted with particular adaptive weights, i.e., weights that are a function of previously collected data. Different forms of adaptive weights are used by existing methods for simple models [Deshpande et al., 2018, Hadad et al., 2019, Zhang et al., 2020]. Our work is a step towards developing a general framework for statistical inference on data collected with adaptive algorithms, including (contextual) bandit algorithms.

2 Problem Formulation

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 primary outcomes $\{Y_t\}_{t=1}^T$. T is deterministic and known. We assume that rewards are a deterministic function of the primary outcomes, i.e., $R_t = f(Y_t)$ for some known function f. We are interested in constructing confidence regions for the parameters of the conditional distribution of Y_t given (X_t, A_t). Below we consider $T \to \infty$ in order to derive the asymptotic distributions of estimators and construct asymptotically valid confidence intervals. We allow the action space $\mathcal{A}$ to be finite or infinite. We use potential outcome notation [Imbens and Rubin, 2015] and let $\{Y_t(a) : a \in \mathcal{A}\}$ denote the potential outcomes of the primary outcome and let $Y_t := Y_t(A_t)$ be the observed outcome. We assume a stochastic contextual bandit environment in which $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P} \in \mathbf{P}$ for $t \in [1 \colon T]$; the contextual bandit environment distribution $\mathcal{P}$ is in a space of possible environment distributions $\mathbf{P}$. We define the history $\mathcal{H}_t := \{X_{t'}, A_{t'}, Y_{t'}\}_{t'=1}^t$ for $t \geq 1$ and $\mathcal{H}_0 := \emptyset$. Actions $A_t \in \mathcal{A}$ are selected according to policies $\pi := \{\pi_t\}_{t \geq 1}$, which define action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) := \mathbb{P}(A_t | \mathcal{H}_{t-1}, X_t)$. Even though the potential outcomes are i.i.d., the *observed* data $\{X_t, A_t, Y_t\}_{t=1}^T$ are *not* because the actions are selected using policies π_t which are a function of past data, $\mathcal{H}_{t-1}$. Non-independence of observations is a key property of adaptively collected data.

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

$$\theta^*(\mathcal{P}) \in \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathbb{E}_{\mathcal{P}} \left[m_\theta(Y_t, X_t, A_t) | X_t, A_t \right] \;\; \text{w.p. } 1. \tag{1}$$

Note that $\theta^*(\mathcal{P})$ does not depend on (X_t, A_t) and it is an implicit modelling assumption that such a $\theta^*(\mathcal{P})$ exists for a given m_θ. To estimate $\theta^*(\mathcal{P})$, we build on M-estimation [Huber, 1992], which classically selects the estimator $f\hat{\theta}$ to be the $\theta \in \Theta$ that maximizes the empirical analogue of Equation (1):

$$\hat{\theta}_T := \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{T} \sum_{t=1}^{T} m_\theta(Y_t, X_t, A_t). \tag{2}$$

For example, in a classical linear regression setting with $|\mathcal{A}| < \infty$ actions, a natural choice for m_θ is the negative of the squared loss function, $m_\theta(Y_t, X_t, A_t) = -(Y_t - X_t^\top \theta_{A_t})^2$. When Y_t is binary, a natural choice is instead the negative log-likelihood function for a logistic regression model, i.e., $m_\theta(Y_t, X_t, A_t) = -[Y_t X_t^\top \theta_{A_t} - \log(1 + \exp(X_t^\top \theta_{A_t}))]$. More generally, m_θ is commonly chosen to be a log-likelihood function or the negative of a robust loss function such as the Huber loss. If the data, $\{X_t, A_t, Y_t\}_{t=1}^T$, were independent across time, classical approaches could be used to prove the consistency and asymptotic normality of M-estimators [Van der Vaart, 2000]. However, on data collected with bandit algorithms, standard M-estimators like the ordinary least-squares estimator fail to provide valid confidence intervals [Hadad et al., 2019, Deshpande et al., 2018, Zhang et al., 2020]. In this work, we show that M-estimators can still be used to provide valid statistical inference on adaptively collected data when adjusted with well-chosen adaptive weights.

3 Adaptively Weighted M-Estimators

We consider a weighted M-estimating criteria with adaptive weights $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ given by $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$. Here $\{\pi_t^{\text{sta}}\}_{t \geq 1}$ are pre-specified *stabilizing policies* that do not depend on data $\{Y_t, X_t, A_t\}_{t \geq 1}$. A default choice for the stabilizing policy when the action space is of size $|\mathcal{A}| < \infty$ is just $\pi_t^{\text{sta}}(a, x) = 1/|\mathcal{A}|$ for all x, a, and t; we discuss considerations for the choice of $\{\pi_t^{\text{sta}}\}_{t=1}^T$ in Section 3.3. We call these weights *square-root importance weights* because they are the square-root of the standard importance weights [Hammersley, 2013, Wang et al., 2017]. Our proposed estimator for $\theta^*(\mathcal{P})$, $\hat{\theta}_T$, is the maximizer of a *weighted* version of the M-estimation criterion of Equation (2):

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

Note that $M_T(\theta)$ defined above depends on both the data $\{X_t, A_t, Y_t\}_{t=1}^T$ and weights $\{W_t\}_{t=1}^T$. We provide asymptotically valid confidence regions for $\theta^*(\mathcal{P})$ by deriving the asymptotic distribution of $\hat{\theta}_T$ as $T \to \infty$ and by proving that the convergence in distribution is *uniform* over $\mathcal{P} \in \mathbf{P}$. Such convergence allows us to construct a uniformly asymptotically valid $1 - \alpha$ level confidence region, $C_T(\alpha)$, for $\theta^*(\mathcal{P})$, which is a confidence region that satisfies

$$\liminf_{T \to \infty} \inf_{\mathcal{P} \in \mathbf{P}} \mathbb{P}_{\mathcal{P}, \pi} \left(\theta^*(\mathcal{P}) \in C_T(\alpha) \right) \geq 1 - \alpha. \tag{3}$$

If $C_T(\alpha)$ were *not* uniformly valid, then there would exist an $\epsilon > 0$ such that for *every* sample size T, $C_T(\alpha)$'s coverage would be below $1 - \alpha - \epsilon$ for some worst-case $P_T \in \mathbf{P}$. Confidence regions which are asymptotically valid, but not *uniformly* asymptotically valid, fail to be reliable in practice [Leeb and Pötscher, 2005, Romano et al., 2012]. Note that 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] and Appendix D, this is not the case on data collected with bandit algorithms.

To construct uniformly valid confidence regions for $\theta^*(\mathcal{P})$ we prove that $\hat{\theta}_T$ is uniformly asymptotically normal in the following sense:

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \sqrt{T} (\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \;\; \text{uniformly over } \mathcal{P} \in \mathbf{P}, \tag{4}$$

where $\ddot{M}_T(\theta) := \frac{\partial^2}{\partial^2 \theta} M_T(\theta)$, $\Sigma_T(\mathcal{P}) := \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^{\otimes 2} \right]$, and $\dot{m}_\theta := \frac{\partial}{\partial \theta} m_\theta$. For any vector z we define $z^{\otimes 2} := z z^\top$.

3.1 Intuition for Square-Root Importance Weights

The critical role of the square-root importance weights $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$ is to adjust for instability in the *variance* of M-estimators due to the bandit algorithm. These weights act akin to standard importance weights when squared and adjust a key term in the variance of M-estimators from depending on adaptive policies $\{\pi_t\}_{t=1}^T$, which can be ill-behaved, to depending on the pre-specified stabilizing policies $\{\pi_t^{\text{sta}}\}_{t=1}^T$. See Zhang et al. [2020] and Deshpande et al. [2018] for more discussion of the ill-behavior of the action selection probabilities for common bandit algorithms, which occurs particularly when there is no unique optimal policy.

As an illustrative example, consider the least-squares estimators in a finite-arm linear contextual bandit setting. Assume that $\mathbb{E}_{\mathcal{P}}[Y_t | X_t, A_t = a] = X_t^\top \theta_a^*(\mathcal{P})$ w.p. 1. We focus on estimating $\theta_a^*(\mathcal{P})$ for some $a \in \mathcal{A}$. The least-squares estimator corresponds to an M-estimator with $m_{\theta_a}(Y_t, X_t, A_t) = -\mathbb{1}_{A_t=a}(Y_t - X_t^\top \theta_a)^2$. The adaptively weighted least-squares (AW-LS) estimator is $\hat{\theta}_{T,a}^{\text{AW-LS}} := \operatorname{argmax}_{\theta_a}\{-\sum_{t=1}^T W_t \mathbb{1}_{A_t=a}(Y_t - X_t^\top \theta_a)^2\}$. For simplicity, suppose that the stabilizing policy does not change with t and drop the index t to get π^{sta}. Taking the derivative of this criterion, we get $0 = \sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t \big(Y_t - X_t^\top \hat{\theta}_{T,a}^{\text{AW-LS}}\big)$, and rearranging terms gives

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t X_t^\top \left(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P})\right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right). \tag{5}$$

Note that the right hand side of Equation (5) is a martingale difference sequence with respect to history $\{\mathcal{H}_t\}_{t=0}^T$ because $\mathbb{E}_{\mathcal{P},\pi}[W_t \mathbb{1}_{A_t=a}(Y_t - X_t^\top \theta_a^*(\mathcal{P})) | \mathcal{H}_{t-1}] = 0$ for all t; by law of iterated expectations and since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$, $\mathbb{E}_{\mathcal{P},\pi}[W_t \mathbb{1}_{A_t=a}(Y_t - X_t^\top \theta_a^*(\mathcal{P})) | \mathcal{H}_{t-1}]$ equals

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\left[W_t \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}\left[Y_t - X_t^\top \theta_a^*(\mathcal{P}) | \mathcal{H}_{t-1}, X_t, A_t = a\right] \big| \mathcal{H}_{t-1}\right] \\
&\underset{(i)}{=} \mathbb{E}_{\mathcal{P}}\left[W_t \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}\left[Y_t - X_t^\top \theta_a^*(\mathcal{P}) | X_t, A_t = a\right] \big| \mathcal{H}_{t-1}\right] \underset{(ii)}{=} 0.
\end{aligned}$$

(i) holds by our i.i.d. potential outcomes assumption. (ii) holds since $\mathbb{E}_{\mathcal{P}}[Y_t | X_t, A_t = a] = X_t^\top \theta_a^*(\mathcal{P})$. We prove that (5) is uniformly asymptotically normal by applying a martingale central limit theorem (Appendix B.4). The key condition in this theorem is that the conditional variance converges uniformly, for which it is sufficient to show that the conditional covariance of $W_t \mathbb{1}_{A_t=a}\left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)$ given $\mathcal{H}_{t-1}$ equals some positive-definite matrix $\Sigma(\mathcal{P})$ for every t, i.e.,

$$\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 \mathbb{1}_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \big| \mathcal{H}_{t-1}\right] = \Sigma(\mathcal{P}). \tag{6}$$

By law of iterated expectations, $\mathbb{E}_{\mathcal{P},\pi}[W_t^2 \mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top \theta_a^*(\mathcal{P}))^2 | \mathcal{H}_{t-1}]$ equals

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi}\left[\frac{\pi^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} \mathbb{1}_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \bigg| \mathcal{H}_{t-1}, X_t\right] \bigg| \mathcal{H}_{t-1}\right] \qquad (7) \\
&\underset{(a)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[\mathbb{1}_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \bigg| \mathcal{H}_{t-1}, X_t\right] \bigg| \mathcal{H}_{t-1}\right] \\
&\underset{(b)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[\mathbb{1}_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \bigg| X_t\right] \bigg| \mathcal{H}_{t-1}\right] \\
&\underset{(c)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[\mathbb{1}_{A_t=a} X_t X_t^\top \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)^2 \bigg| X_t\right]\right] \\
&\underset{(d)}{=} \mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}[\mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top \theta_a^*(\mathcal{P}))^2] =: \Sigma(\mathcal{P}).
\end{aligned}$$

Above, (a) holds because the importance weights change the sampling measure from the adaptive policy π_t to the pre-specified stabilizing policy π^{sta}. (b) holds by our i.i.d. potential outcomes assumption and because π^{sta} is a pre-specified policy. (c) holds because X_t does not depend on $\mathcal{H}_{t-1}$ by our i.i.d. potential outcomes assumption. (d) holds by the law of iterated expectations. Note that $\Sigma(\mathcal{P})$ does not depend on t because π^{sta} is not time-varying. In contrast, without the adaptive weighting, i.e., when $W_t = 1$, the conditional covariance of $\mathbb{1}_{A_t=a}\left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right)$ on $\mathcal{H}_{t-1}$ is a random variable, due to the adaptive policy π_t.

In Figure 1 we plot the empirical distributions of the z-statistic for the least-squares estimator both with and without adaptive weighting. We consider a two-armed bandit with $A_t \in \{0,1\}$. Let $\theta_1^*(\mathcal{P}) := \mathbb{E}_{\mathcal{P}}[Y_t(1)]$ and $m_{\theta_1}(Y_t, A_t) := -A_t(Y_t - \theta_1)^2$. The unweighted version, i.e., the ordinary least-squares (OLS) estimator, is $\hat{\theta}_{T,1}^{\text{OLS}} := \text{argmax}_{\theta_1} \frac{1}{T}\sum_{t=1}^{T} m_{\theta_1}(Y_t, A_t)$. The adaptively weighted version is $\hat{\theta}_{T,1}^{\text{AW-LS}} := \text{argmax}_{\theta_1} \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta_1}(Y_t, A_t)$. We collect data using Thompson Sampling and use a uniform stabilizing policy where $\pi^{\text{sta}}(1) = \pi^{\text{sta}}(0) = 0.5$. It is clear that the least-squares estimator with adaptive weighting has a z-statistic that is much closer to a normal distribution.

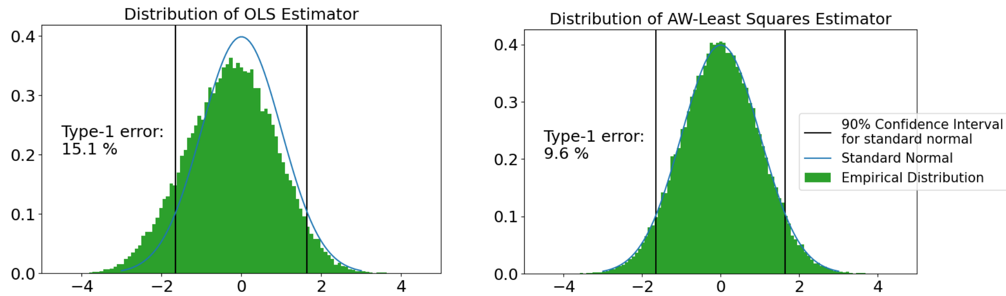

Figure 1: The empirical distributions of the weighted and unweighted least-squares estimators for $\theta_1^*(\mathcal{P}) := \mathbb{E}_{\mathcal{P}}[Y_t(1)]$ in a two arm bandit setting where $\mathbb{E}_{\mathcal{P}}[Y_t(1)] = \mathbb{E}_{\mathcal{P}}[Y_t(0)] = 0$. We perform Thompson Sampling with $\mathcal{N}(0,1)$ priors, $\mathcal{N}(0,1)$ errors, and $T = 1000$. Specifically, we plot $\sqrt{\sum_{t=1}^{T} A_t}(\hat{\theta}_{T,1}^{\text{OLS}} - \theta_1^*(\mathcal{P}))$ on the left and $\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{\frac{0.5}{\pi_t(1)}}A_t\right)(\hat{\theta}_{T,1}^{\text{AW-LS}} - \theta_1^*(\mathcal{P}))$ on the right.

The square-root importance weights are a form of variance stabilizing weights, akin to those introduced in Hadad et al. [2019] for estimating means and differences in means on data collected with multi-armed bandits. In fact, in the special case that $|\mathcal{A}| < \infty$ and $\phi(X_t, A_t) = [\mathbb{1}_{A_t=1}, \mathbb{1}_{A_t=2}, ..., \mathbb{1}_{A_t=|\mathcal{A}|}]^\top$, the adaptively weighted least-squares estimator is equivalent to the weighted average estimator of Hadad et al. [2019]. See Section 4 for more on Hadad et al. [2019].

3.2 Asymptotic Normality and Confidence Regions

We now discuss conditions under which the adaptively weighted M-estimators are asymptotically normal in the sense of Equation (4). In general, our conditions differ from those made for standard M-estimators on i.i.d. data because (i) the data is adaptively collected, i.e., π_t can depend on $\mathcal{H}_{t-1}$ and (ii) we ensure uniform convergence over $\mathcal{P} \in \mathbf{P}$, which is stronger than guaranteeing convergence pointwise for each $\mathcal{P} \in \mathbf{P}$.

Condition 1 (Stochastic Bandit Environment)**.** *Potential outcomes* $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P} \in \boldsymbol{P}$ *over* $t \in [1\colon T]$.

Condition 1 implies that Y_t is independent of $\mathcal{H}_{t-1}$ given X_t and A_t, and the conditional distribution $Y_t \mid X_t, A_t$ is invariant over time. Also note that action space $\mathcal{A}$ can be finite or infinite.

Condition 2 (Differentiable)**.** *The first three derivatives of* $m_\theta(y, x, a)$ *with respect to* θ *exist for every* $\theta \in \Theta$*, every* $a \in \mathcal{A}$*, and every* (x, y) *in the joint support of* $\{\mathcal{P} : \mathcal{P} \in \boldsymbol{P}\}$.

Condition 3 (Bounded Parameter Space)**.** *For all* $\mathcal{P} \in \boldsymbol{P}$*,* $\theta^*(\mathcal{P}) \in \Theta$*, a bounded open subset of* $\mathbb{R}^d$.

Condition 4 (Lipschitz)**.** *There exists some function* g *such that* $\sup_{\mathcal{P}\in\boldsymbol{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2]$ *is bounded and 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 restrict the complexity of the function m in order to ensure a martingale law of large numbers result holds uniformly over functions $\{m_\theta : \theta \in \Theta\}$; this is used to prove the consistency of $\hat{\theta}_T$. Similar conditions are commonly used to prove consistency of M-estimators based on i.i.d. data, although the boundedness of the parameter space can be dropped when m_θ is concave (as it is in many canonical examples such as least squares) [Van der Vaart, 2000, Engle, 1994, Bura et al., 2018]; we expect that a similar result would hold for adaptively weighted M-estimators.

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

Condition 5 is similar to those of Van der Vaart [2000, Theorem 5.41]. However, to guarantee uniform convergence we assume that moment bounds hold uniformly over $\mathcal{P} \in \mathbf{P}$ and $t \geq 1$.

Condition 6 (Third Derivative Domination)**.** *There exists a function $\dddot{m}(Y_t, X_t, A_t) \in \mathbb{R}^{d \times d \times d}$ such that (i) $\sup_{\mathcal{P} \in \boldsymbol{P}, t \geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\|\dddot{m}(Y_t, X_t, A_t)\|_1^2\right]$ is bounded and (ii) for all $\mathcal{P} \in \boldsymbol{P}$ there exists some $\epsilon_{\dddot{m}} > 0$ such that the following holds with probability 1,*

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

For $B \in \mathbb{R}^{d \times d \times d}$, we define $\|B\|_1 := \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} |B_{i,j,k}|$. Condition 6 is again similar to those in classical M-estimator asymptotic normality proofs [Van der Vaart, 2000, Theorem 5.41].

Condition 7 (Maximizing Solution)**.**
(i) For all $\mathcal{P} \in \boldsymbol{P}$, $\theta^*(\mathcal{P}) \in \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}}\left[m_\theta(Y_t, X_t, A_t) \middle| X_t, A_t\right]$ w.p. 1, $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t) \middle| X_t, A_t\right] = 0$ w.p. 1, and $\mathbb{E}_{\mathcal{P}}\left[\ddot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t) \middle| X_t, A_t\right] \preceq 0$ w.p. 1.
(ii) There exists some positive definite matrix H such that $-\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\ddot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\right] \succeq H$ for all $\mathcal{P} \in \boldsymbol{P}$ and all sufficiently large T.

For matrices A, B, we define $A \succeq B$ to mean that $A - B$ is positive semi-definite, as used above. Condition 7 (i) ensures that $\theta^*(\mathcal{P})$ is a conditionally maximizing solution for all contexts X_t and actions A_t; this ensures that $\{\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\}_{t=1}^{T}$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=1}^{T}$. Note it does not require $\theta^*(\mathcal{P})$ to always be a conditionally *unique* optimal solution. Condition 7 (ii) is related to the local curvature at the maximizing solution and the analogous condition in the i.i.d. setting is trivially satisfied; we specifically use this condition to ensure we can replace $\ddot{M}(\theta^*(\mathcal{P}))$ with $\ddot{M}(\hat{\theta}_T)$ in our asymptotic normality result, i.e., that $\ddot{M}(\theta^*(\mathcal{P}))^{-1}\ddot{M}(\hat{\theta}_T) \xrightarrow{P} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$.

Condition 8 (Well-Separated Solution)**.** *For all sufficiently large T, for any $\epsilon > 0$, there exists some $\delta > 0$ such that for all $\mathcal{P} \in \boldsymbol{P}$,*

$$\inf_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\|_2 > \epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t) - m_\theta(Y_t, X_t, A_t)\right] \right\} \geq \delta.$$

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

Condition 9 (Bounded Importance Ratios)**.** *$\{\pi_t^{\text{sta}}\}_{t=1}^{T}$ do not depend on data $\{Y_t, X_t, A_t\}_{t=1}^{T}$. For all $t \geq 1$, $\rho_{\min} \leq \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} \leq \rho_{\max}$ w.p. 1 for some constants $0 < \rho_{\min} \leq \rho_{\max} < \infty$.*

Note that Condition 9 implies that for a stabilizing policy that is not time-varying, the action selection probabilities of the bandit algorithm $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ must be bounded away from zero w.p. 1. Similar boundedness assumptions are also made in the off-policy evaluation literature [Thomas and Brunskill, 2016, Kallus and Uehara, 2020]. We discuss this condition further in Sections 3.3 and 6.

Theorem 1 (Uniform Asymptotic Normality of Adaptively Weighted M-Estimators)**.** *Under Conditions 1-9 we have that $\hat{\theta}_T \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \boldsymbol{P}$. Additionally,*

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \boldsymbol{P}. \tag{8}$$

The asymptotic normality result of equation (8) guarantees that for d-dimensional $\theta^*(\mathcal{P})$,

$$\liminf_{T \to \infty} \inf_{\mathcal{P} \in \mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\left[\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))\right]^{\otimes 2} \leq \chi^2_{d,(1-\alpha)}\right) = 1 - \alpha.$$

Above $\chi^2_{d,(1-\alpha)}$ is the $1 - \alpha$ quantile of the χ^2 distribution with d degrees of freedom. Note that the region $C_T(\alpha) := \left\{\theta \in \Theta : [\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P}))]^{\otimes 2} \leq \chi^2_{d,(1-\alpha)}\right\}$ defines a

d-dimensional hyper-ellipsoid confidence region for $\theta^*(\mathcal{P})$. Also note that since $\ddot{M}_T(\hat{\theta}_T)$ does not concentrate under standard bandit algorithms, we cannot use standard arguments to justify treating $\hat{\theta}_T$ as multivariate normal with covariance $\ddot{M}_T(\hat{\theta}_T)^{-1}\Sigma_T(\mathcal{P})\ddot{M}_T(\hat{\theta}_T)^{-1}$. Nevertheless, Theorem 1 can be used to guarantee valid confidence regions for subset of entries in $\theta^*(\mathcal{P})$ by using projected confidence regions [Nickerson, 1994]. Projected confidence regions take a confidence region for all parameters $\theta^*(\mathcal{P})$ and project it onto the lower dimensional space on which the subset of target parameters lie (Appendix A.2).

3.3 Choice of Stabilizing Policy

When the action space is bounded, using weights $W_t = 1/\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ is equivalent to using square-root importance weights with a stabilizing policy that selects actions uniformly over $\mathcal{A}$; this is because weighted M-estimators are invariant to all weights being scaled by the same constant. It can make sense to choose a non-uniform stabilizing policy in order to prevent the square-root importance weights from growing too large and to ensure Condition 9 holds; disproportionately up-weighting a few observations can lead to unstable estimators. Note that an analogue of our stabilizing policy exists in the causal inference literature, namely, "stabilized weights" use a probability density in the numerator of the weights to prevent them from becoming too large [Robins et al., 2000].

We now discuss how to choose stabilizing policies $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ in order to minimize the asymptotic variance of adaptively weighted M-estimators. We focus on the adaptively weighted least-squares estimator when we have a linear outcome model $\mathbb{E}_{\mathcal{P}}[Y_t|X_t, A_t] = X_t^\top \theta_{A_t}$:

$$\hat{\theta}^{\text{AW-LS}} := \underset{\theta\in\Theta}{\operatorname{argmax}} \left\{ \frac{1}{T} \sum_{t=1}^{T} W_t \left(Y_t - X_t^\top \theta_{A_t} \right)^2 \right\}. \tag{9}$$

Recall that our use of adaptive weights is to adjust for instability in the variance of M-estimators induced by the bandit algorithm in order to construct valid confidence regions; note that weighted estimators are not typically used for this reason. On i.i.d. data, the least-squares criterion is weighted like in Equation (9) in order to minimize the variance of estimators under noise heteroskedasticity; in this setting, the best linear unbiased estimator has weights $W_t = 1/\sigma^2(A_t, X_t)$ where $\sigma^2(A_t, X_t) := \mathbb{E}_{\mathcal{P}}[(Y_t - X_t^\top \theta^*_{A_t}(\mathcal{P}))^2|X_t, A_t]$; this up-weights the importance of observations with low noise variance. Intuitively, if we do not need to variance stabilize, $\{W_t\}_{t\geq 1}$ should be determined by the relative importance of minimizing the errors for different observations, i.e., their noise variance.

In light of this observation, we expect that under homoskedastic noise there is no reason to up-weight some observations over others. This would recommend choosing the stabilizing policy to make $W_t = \sqrt{\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ as close to 1 as possible, subject to the constraint that the stabilizing policies are pre-specified, i.e., $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ do not depend on data $\{Y_t, X_t, A_t\}_{t\geq 1}$ (see Appendix C for details). Since adjusting for heteroskedasticity and variance stabilization are distinct uses of weights, under heteroskedasticity, we recommend that the weights are combined in the following sense: $W_t = \left(1/\sigma^2(A_t, X_t)\right) \sqrt{\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$. This would mean that to minimize variance, we still want to choose the stabilizing policies to make $\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ as close to 1 possible, subject to the pre-specified constraint.

4 Related Work

Villar et al. [2015] and Rafferty et al. [2019] empirically illustrate that classical ordinary least squares (OLS) inference methods have inflated Type-1 error when used on data collected with a variety of regret-minimizing multi-armed bandit algorithms. Chen et al. [2020] prove that the OLS estimator is asymptotically normal on data collected with an ϵ-greedy algorithm, but their results do not cover settings in which there is no unique optimal policy, e.g., a multi-arm bandit with two identical arms (Appendix E). Recent work has discussed the non-normality of OLS on data collected with bandit algorithms when there is no unique optimal policy and proposed alternative methods for statistical inference. A common thread between these methods is that they all utilize a form of *adaptive weighting*. Deshpande et al. [2018] introduced the W-decorrelated estimator, which adjusts the OLS estimator with a sum of adaptively weighted residuals. In the multi-armed bandit setting, the W-decorrelated estimator up-weights observations from early in the study and down-weights observations from later in the study [Zhang et al., 2020]. 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 adaptively weighted versions of both the standard augmented-inverse propensity weighted estimator (AW-AIPW) and the sample mean (AWA) for estimating parameters of simple models on data collected with bandit algorithms. They introduce a class of adaptive "variance stabilizing" weights, for which the variance of a normalized version of their estimators converges in probability to a constant. In their discussion section they note open questions, two of which this work addresses: 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? We demonstrate that variance stabilizing adaptive weights can be used to modify a large class of M-estimators to guarantee valid inference. This generalization allows us to perform valid inference for a large class of important inferential targets: parameters of models for expected outcomes that are context dependent.

An alternative to using asymptotic approximations to construct confidence intervals is to use high-probability anytime confidence bounds. These bounds provide stronger guarantees than those based on asymptotic approximations, as they are guaranteed to hold for finite samples and hold simultaneously for any number of observations, i.e., for all $T \geq 1$. The downside is that these bounds are typically much wider, which is why much of classical statistics uses asymptotic approximations. Here we do the same. In Section 5, we empirically compare our to the self-normalized martingale bound [Abbasi-Yadkori et al., 2011], a high-probability bound commonly used in the bandit literature.

5 Simulation Results

In this section, $R_t = Y_t$. We consider two settings: a continuous reward setting and a binary reward setting. In the continuous reward setting, the rewards are generated with mean $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$ and noise drawn from a student's t distribution with five degrees of freedom; here $\tilde{X}_t = [1, X_t] \in \mathbb{R}^3$ (X_t with intercept term), actions $A_t \in \{0, 1\}$, and parameters $\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P}) \in \mathbb{R}^3$. In the binary reward setting, the reward R_t is generated as a Bernoulli with success probability $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = [1 + \exp(-\tilde{X}_t^\top \theta_0^*(\mathcal{P}) - A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P}))]^{-1}$. Furthermore, in both simulation settings we set $\theta_0^*(\mathcal{P}) = [0.1, 0.1, 0.1]$ and $\theta_1^*(\mathcal{P}) = [0, 0, 0]$, so there is no unique optimal arm; we call vector parameter $\theta_1^*(\mathcal{P})$ the *advantage* of selecting $A_t = 1$ over $A_t = 0$. Also in both settings, the contexts X_t are drawn i.i.d. from a uniform distribution.

In both simulation settings we collect data using Thompson Sampling with a linear model for the expected reward and normal priors [Agrawal and Goyal, 2013] (so even when the reward is binary). We constrain the action selection probabilities with *clipping* at a rate of 0.05; this means that while typical Thompson Sampling produces action selection probabilities $\pi_t^{\text{TS}}(A_t, X_t, \mathcal{H}_{t-1})$, we instead use action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) = 0.05 \vee \big(0.95 \wedge \pi_t^{\text{TS}}(A_t, X_t, \mathcal{H}_{t-1})\big)$ to select actions. We constrain the action selection probabilities in order to ensure weights W_t are bounded when using a uniform stabilizing policy; see Sections 3.2 and 6 for more discussion on this boundedness assumption. Also note that increasing the amount the algorithm explores (clipping) decreases the expected width of confidence intervals constructed on the resulting data (see Section 6).

To analyze the data, in the continuous reward setting, we use least-squares estimators with a correctly specified model for the expected reward, i.e., M-estimators with $m_\theta(R_t, X_t, A_t) = -(R_t - \tilde{X}_t^\top \theta_0 - A_t \tilde{X}_t^\top \theta_1)^2$. We consider both the unweighted and adaptively weighted versions. We also compare to the self-normalized martingale bound [Abbasi-Yadkori et al., 2011] and the W-decorrelated estimator [Deshpande et al., 2018], as they were both developed for the linear expected reward setting. For the self-normalized martingale bound, which requires explicit bounds on the parameter space, we set $\Theta = \{\theta \in \mathbb{R}^6 : \|\theta\|_2 \leq 6\}$. In the binary reward setting, we also assume a correctly specified model for the expected reward. We use both unweighted and adaptively weighted maximum likelihood estimators (MLEs), which correspond to an M-estimators with $m_\theta(R_t, X_t, A_t)$ set to the negative log-likelihood of R_t given X_t, A_t. We solve for these estimators using Newton–Raphson optimization and do not put explicit bounds on the parameter space Θ (note in this case m_θ is concave in θ [Agresti, 2015, Chapter 5.4.2]). See Appendix A for additional details and simulation results.

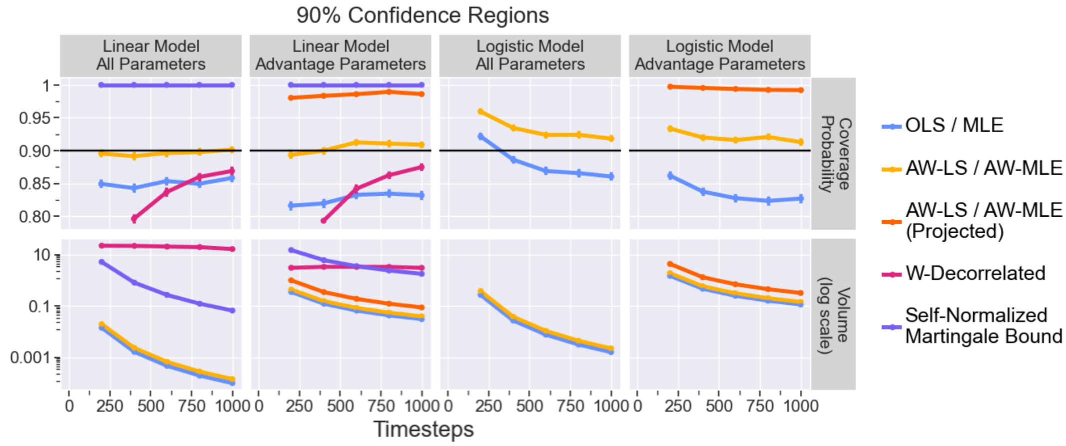

Figure 2: Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. The left two columns are for the linear reward model setting (t-distributed rewards) and the right two columns are for the logistic regression model setting (Bernoulli rewards). We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings.

In Figure 4 we plot the empirical coverage probabilities and volumes of 90% confidence regions for $\theta^*(\mathcal{P}) := [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})]$ and $\theta_1^*(\mathcal{P})$ in both the continuous and binary reward settings. While the confidence regions based on the unweighted least-squares estimator (OLS) and the unweighted MLE have significant undercoverage that does not improve as T increases, the confidence regions based on the adaptively weighted versions, AW-LS and AW-MLE, have very reliable coverage. For the confidence regions for $\theta_1^*(\mathcal{P})$ based on the AW-LS and AW-MLE, we include both projected confidence regions (for which we have theoretical guarantees) and non-projected confidence regions. The confidence regions based on projections are conservative but nevertheless have comparable volume to those based on OLS and MLE respectively. We do not prove theoretical guarantees for the non-projection confidence regions for AW-LS and AW-MLE, however they perform well across in our simulations. Both types of confidence regions based on AW-LS have significantly smaller volumes than those constructed using the self-normalized martingale bound and W-decorrelated estimator. Note that the W-decorrelated estimator and self-normalized martingale bounds are designed for linear contextual bandits and are thus not applicable for the logistic regression model setting. The confidence regions constructed using the self-normalized martingale bound have reliable coverage as well, but are very conservative. Empirically, we found that the coverage probabilities of the confidence regions based on the W-decorrelated estimator were very sensitive to the choice of tuning parameters. We use $5,000$ Monte-Carlo repetitions and the error bars plotted are standard errors.

6 Discussion

Immediate questions We assume that ratios $\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ are bounded for our theoretical results; this precludes $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ from going to zero for a fixed stabilizing policy. For simple models, e.g., the AW-LS estimator, we can let these ratios grow at a certain rate and still guarantee asymptotic normality (Appendix B.5); we conjecture similar results hold more generally.

Generality and robustness This work assumes $\theta^*(\mathcal{P}) \in \text{argmax}_{\theta \in \Theta} \mathbb{E}_{\mathcal{P}}[m_\theta(Y_t, X_t, A_t)|X_t, A_t]$ w.p. 1. Our theorems use this 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 $\theta^*(\mathcal{P})$ to be the best *projected* solution, i.e., $\theta_0(\mathcal{P}) \in \text{argmax}_{\theta \in \Theta} \mathbb{E}_{\mathcal{P},\pi}[m_\theta(Y_t, X_t, A_t)]$. Note that the best projected solution, $\theta^*(\mathcal{P})$, depends on the distribution of the action selection policy π. It would be ideal to also be able to perform inference for a projected solution on adaptively collected data.

Another natural question is whether adaptive weighting methods work in Markov Decision Processes (MDP) environments. Taking the AW-LS estimator introduced in Section 3.1 as an example, our conditional variance derivation in Equation (7) fails to hold in an MDP setting, specifically equality (c). However, the conditional variance condition can be satisfied if we instead use weights

$W_t = \{[\pi_t^{\text{sta}}(A_t, X_t)p^{\text{sta}}(X_t)]/[\pi_t(A_t, X_t, \mathcal{H}_{t-1})\mathbb{P}_{\mathcal{P}}(X_t|X_{t-1}, A_{t-1})]\}^{1/2}$ where $\mathbb{P}_{\mathcal{P}}$ are the state transition probabilities and p^{sta} is a pre-specified distribution over states. In general though we do not expect to know the transition probabilities $\mathbb{P}_{\mathcal{P}}$ 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.

Trading-off regret minimization and statistical inference objectives In sequential decision-making problems there is a fundamental trade-off between minimizing regret and minimizing estimation error for parameters of the environment using the resulting data [Bubeck et al., 2009, Dean et al., 2018]. Given this trade-off there are many open problems regarding how to minimize regret while still guaranteeing a certain amount of power or expected confidence interval width, e.g., developing sample size calculators for use in justifying the number of users in a mobile health trial, and developing new adaptive algorithms [Liu et al., 2014, Erraqabi et al., 2017, Yao et al., 2020].

Acknowledgements and Disclosure of Funding

We thank Yash Nair for feedback on early drafts of this work.

Research reported in this paper was supported by National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health under award number R01AA23187, National Institute on Drug Abuse (NIDA) of the National Institutes of Health under award number P50DA039838, National Cancer Institute (NCI) of the National Institutes of Health under award number U01CA229437, and by NIH/NIBIB and OD award number P41EB028242. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1745303. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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.

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.

Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, pages 1–16, 2020.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, 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.

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.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

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. *Econometrica*, 89(1):113–132, 2021.

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.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

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.

David M Nickerson. Construction of a conservative confidence region from projections of an exact confidence region in multiple linear regression. *The American Statistician*, 48(2):120–124, 1994.

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.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

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.

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.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

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.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

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

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Inference for batched bandits. In *Advances in Neural Information Processing Systems*, 2020.

Appendix

Table of Contents

A Simulations

A.1 Simulation Details

Simulation Environment

- Each dimension of X_t is sampled independently from Uniform$(0, 5)$.
- $\theta^*(\mathcal{P}) = [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0.1, 0.1, 0.1, 0, 0, 0]$, where $\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P}) \in \mathbb{R}^3$.
  Below also include simulations where $[\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$.
- t-Distributed rewards: $R_t | X_t, A_t \sim t_5 + \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$, 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 $\nu_t = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$ and $expit(x) = \frac{1}{1+\exp(-x)}$.
- Poisson rewards: $R_t | X_t, A_t \sim$ Poisson$(\exp(\nu_t))$ for $\nu_t = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$.

Algorithm

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

Compute Time and Resources All simulations run within a few hours on a MacBook Pro.

A.2 Details on Constructing of Confidence Regions

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

A.2.1 Least Squares Estimators

- $\hat{\theta}_T = \left(\sum_{t=1}^T W_t Z_t Z_t^\top \right)^{-1} \sum_{t=1}^T W_t Z_t R_t$
  - For unweighted least squares, $W_t = 1$ and we call the estimator $\hat{\theta}_T^{\text{OLS}}$.
  - For adaptively weighted least squares, $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$; this is equivalent to using square-root importance weights with a uniform stabilizing policy. We call the estimator $\hat{\theta}_T^{\text{AW-LS}}$.
- 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 $\theta^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{ \theta \in \mathbb{R}^d : \left[\hat{\Sigma}_T^{-1/2} \left(\frac{1}{T} \sum_{t=1}^T W_t Z_t Z_t^\top \right) \sqrt{T} (\hat{\theta}_T - \theta) \right]^{\otimes 2} \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha) \right\}. \quad (10)$$

  - For the unweighted least-squares estimator we use the following variance estimator: $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^T Z_t Z_t^\top$.
  - For the AW-Least Squares estimator we use the following variance estimator: $\hat{\Sigma}_T = \hat{\sigma}_T^2 \frac{1}{T} \sum_{t=1}^T \frac{1}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{A_t} \frac{1}{1-\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{1-A_t} Z_t Z_t^\top$.
- To construct (non-projected) confidence regions for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ we treat the unweighted least squares / AW-LS estimators, $\hat{\theta}_{T,1}$, as $\mathcal{N}\left(\theta_1^*(\mathcal{P}), \frac{1}{T} \left(\frac{1}{T} \sum_{t=1}^T W_t Z_t Z_t^\top \right)^{-1} \hat{\Sigma}_T \left(\frac{1}{T} \sum_{t=1}^T W_t Z_t Z_t^\top \right)^{-1} \right)$. We use a Hotelling t-squared test statistic to construct confidence regions for $\theta_1^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{ \theta_1 \in \mathbb{R}^{d_1} : \left[V_{1,T}^{-1/2} \sqrt{T} (\hat{\theta}_{T,1} - \theta_1) \right]^{\otimes 2} \leq \frac{d_1(T-1)}{T-d_1} F_{d_1, T-d_1}(1-\alpha) \right\},$$

  where $V_{1,T}$ is the lower right $d_1 \times d_1$ block of matrix $\left(\frac{1}{T} \sum_{t=1}^T W_t Z_t Z_t^\top \right)^{-1} \hat{\Sigma}_T \left(\frac{1}{T} \sum_{t=1}^T W_t Z_t Z_t^\top \right)^{-1}$. Recall that for the unweighted least squares estimator $W_t = 1$ and for AW-LS $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$.
- For the AW-least squares estimator, we also construct projected confidence regions for $\theta_1^*(\mathcal{P})$ using the confidence region defined in equation (10). See Section A.2.5 below for more details on constructing projected confidence regions.

A.2.2 MLE Estimators

Distribution	ν	$b(\nu)$	$b'(\nu)$	$b''(\nu)$	$b'''(\nu)$
$\mathcal{N}(\mu, 1)$	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)	$\log\left(\frac{p}{1-p}\right)$	$\log(1+e^\nu)$	$\frac{e^\nu}{1+e^\nu} = p$	$\frac{e^\nu}{(1+e^\nu)^2} = p(1-p)$	$p(1-p)(1-2p)$

- $\hat{\theta}_T$ 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, $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$; this is equivalent to using square-root importance weights with a uniform stabilizing policy.

- Second derivative of score function: $-\sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.
- We use a Hotelling t-squared test statistic to construct confidence regions for $\theta^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{ \theta \in \mathbb{R}^d : \left[\hat{\Sigma}_T^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right) \sqrt{T} (\hat{\theta}_T - \theta) \right]^{\otimes 2} \leq \frac{d(T-1)}{T-d} F_{d, T-d}(1-\alpha) \right\}. \quad (11)$$

  - For the MLE variance estimator, we use $\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.
  - For the AW-MLE variance estimator, we use $\hat{\Sigma}_T = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{A_t} \frac{1}{1-\pi_t(A_t, X_t, \mathcal{H}_{t-1})}^{1-A_t} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.

- To construct (non-projected) confidence regions for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ we treat the MLE / AW-MLE estimators, $\hat{\theta}_{T,1}$, as $\mathcal{N}\left(\theta_1^*(\mathcal{P}), \frac{1}{T} \left(\frac{1}{T} \sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right) \hat{\Sigma}_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} W_t b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top \right) \right)$. We use a Hotelling t-squared test statistic to construct confidence regions for $\theta_1^*(\mathcal{P})$:

$$C_T(\alpha) = \left\{ \theta_1 \in \mathbb{R}^{d_1} : \left[V_{1,T}^{-1/2} \sqrt{T} (\hat{\theta}_{T,1} - \theta_1) \right]^{\otimes 2} \leq \frac{d_1(T-1)}{T-d_1} F_{d_1, T-d_1}(1-\alpha) \right\},$$

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

- For the AW-MLE estimator, we also construct projected confidence regions for $\theta_1^*(\mathcal{P})$ using the confidence region defined in equation (11). See Section A.2.5 below for more details on constructing projected confidence regions.

A.2.3 W-Decorrelated

The following is based on Algorithm 1 of Deshpande et al. [2018].

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

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

- The weights are set as follows:
$W_1 = 0 \in \mathbb{R}^d$ and $W_t = (I_d - \sum_{s=1}^{t} \sum_{u=1}^{t} W_s Z_u^\top) Z_t \frac{1}{\lambda_T + \|Z_t\|_2^2}$ for $t > 1$.
- 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 quantile 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:

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

- To construct confidence ellipsoids for $\theta^*(\mathcal{P})$ are constructed using a Hotelling t-squared statistic:

$$C_T(\alpha) = \left\{ \theta \in \mathbb{R}^d : (\hat{\theta}_T^{\text{WD}} - \theta)^\top V_T^{-1} (\hat{\theta}_T^{\text{WD}} - \theta) \leq \frac{d(T-1)}{T-d} F_{d,T-d}(1-\alpha) \right\}$$

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

- To construct confidence ellipsoids for $\theta_1^*(\mathcal{P}) \in \mathbb{R}^{d_1}$ with the following confidence ellipsoid where $V_{T,1}$ is the lower right $d_1 \times d_1$ block of matrix V_T:

$$C_T(\alpha) = \left\{ \theta_1 \in \mathbb{R}^{d_1} : (\hat{\theta}_{T,1}^{\text{WD}} - \theta_1)^\top V_{T,1}^{-1} (\hat{\theta}_{T,1}^{\text{WD}} - \theta_1) \leq \frac{d_1(T-1)}{T-d_1} F_{d_1,T-d_1}(1-\alpha) \right\}.$$

A.2.4 Self-Normalized Martingale Bound

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

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

- $\hat{\theta}_T = \left(\lambda I_d + \sum_{t=1}^T Z_t Z_t^\top \right)^{-1} \sum_{t=1}^T Z_t R_t$.
- $V_T = I_d \lambda + \sum_{t=1}^T Z_t Z_t^\top$.
- $\lambda = 1$ (ridge regression regularization parameter).
- $\sigma = 1$ (assumes rewards are σ-subgaussian).
- $S = 6$, where it is assumed that $\|\theta^*(\mathcal{P})\| \leq S$ (recall that in our simulations $\theta^*(\mathcal{P}) \in \mathbb{R}^6$).
- $\Theta = \{\theta \in \mathbb{R}^6 : \|\theta\|_2 \leq 6\}$.
- For constructing confidence regions for $\theta^*(\mathcal{P})$, we use projected confidence regions.

A.2.5 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, for $p < d$.

- 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

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

  Thus the equation for the final projected ellipsoid is

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

A.3 Additional Simulation Results

In addition to the continuous reward and a binary reward settings, here we also consider a discrete count reward setting. In this discrete reward setting, the reward R_t is generated from a Poisson distribution with expectation $\mathbb{E}_{\mathcal{P}}[R_t | X_t, A_t] = \exp(\tilde{X}_t^\top \theta_0^*(\mathcal{P}) - A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P}))$. All other data generation methods are equivalent to those used for the other simulation settings. Additionally we will consider the setting in which $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ for the continuous reward, binary reward, and discrete count settings.

To analyze the data, in the discrete count reward setting, we assume a correctly specified model for the expected reward. We use both unweighted and adaptively weighted maximum likelihood estimators (MLEs), which correspond to an M-estimators with $m_\theta(R_t, X_t, A_t)$ set to the negative log-likelihood of R_t given X_t, A_t. We solve for these estimators using Newton–Raphson optimization and do not put explicit bounds on the parameter space Θ.

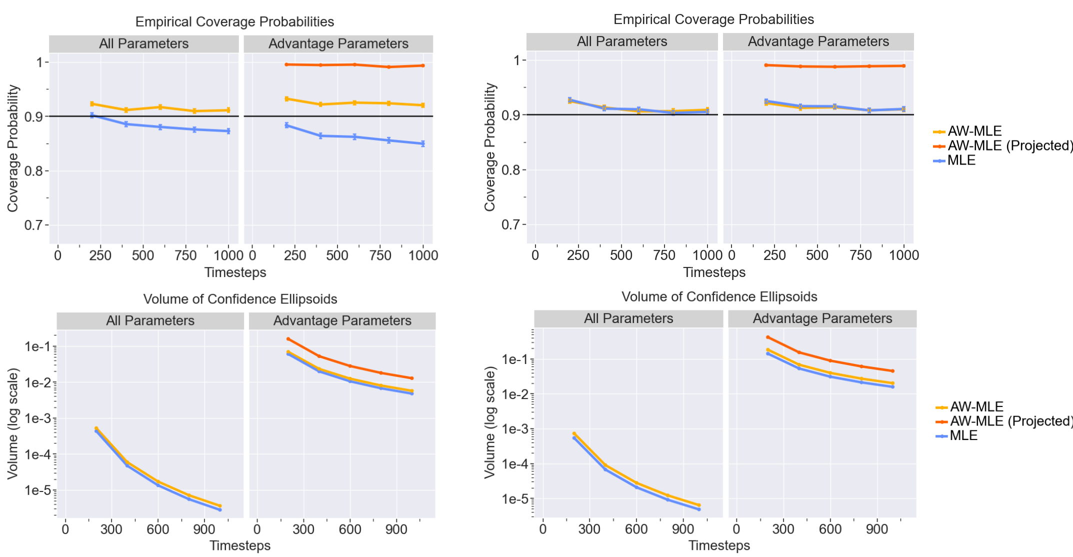

Figure 3: **Poisson Rewards:** Empirical coverage probabilities for 90% confidence ellipsoids for parameters $\theta^*(\mathcal{P})$ and parameters $\theta_1^*(\mathcal{P})$ (top row). We also plot the volumes of these 90% confidence ellipsoids for $\theta^*(\mathcal{P})$ and parameters $\theta_1^*(\mathcal{P})$ (bottom row). We set the true parameters to $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and to $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

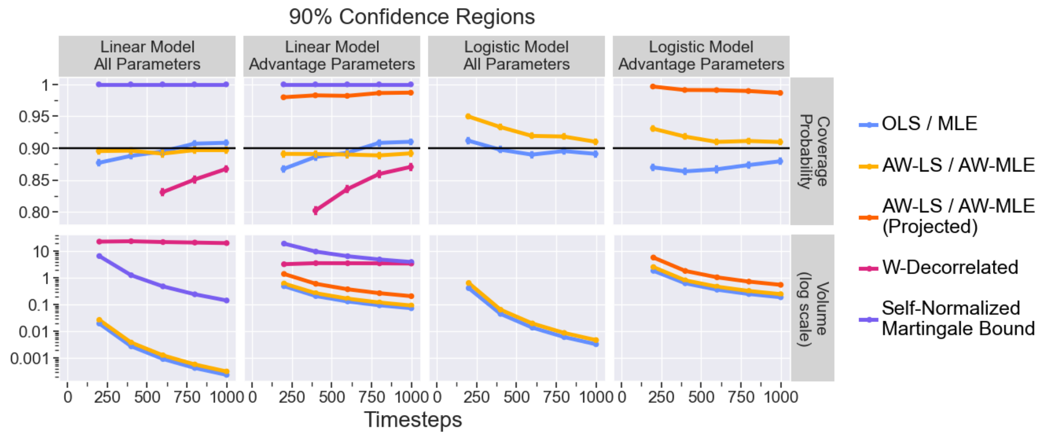

Figure 4: Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. In these simulations, $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$. The left two columns are for the linear reward model setting (t-distributed rewards) and the right two columns are for the logistic regression model setting (Bernoulli rewards). We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings.

In Figure 5, we plot the mean squared errors of all estimators for all three simulation settings (same simulation hyperparameters as described previously for the respective simulation settings).

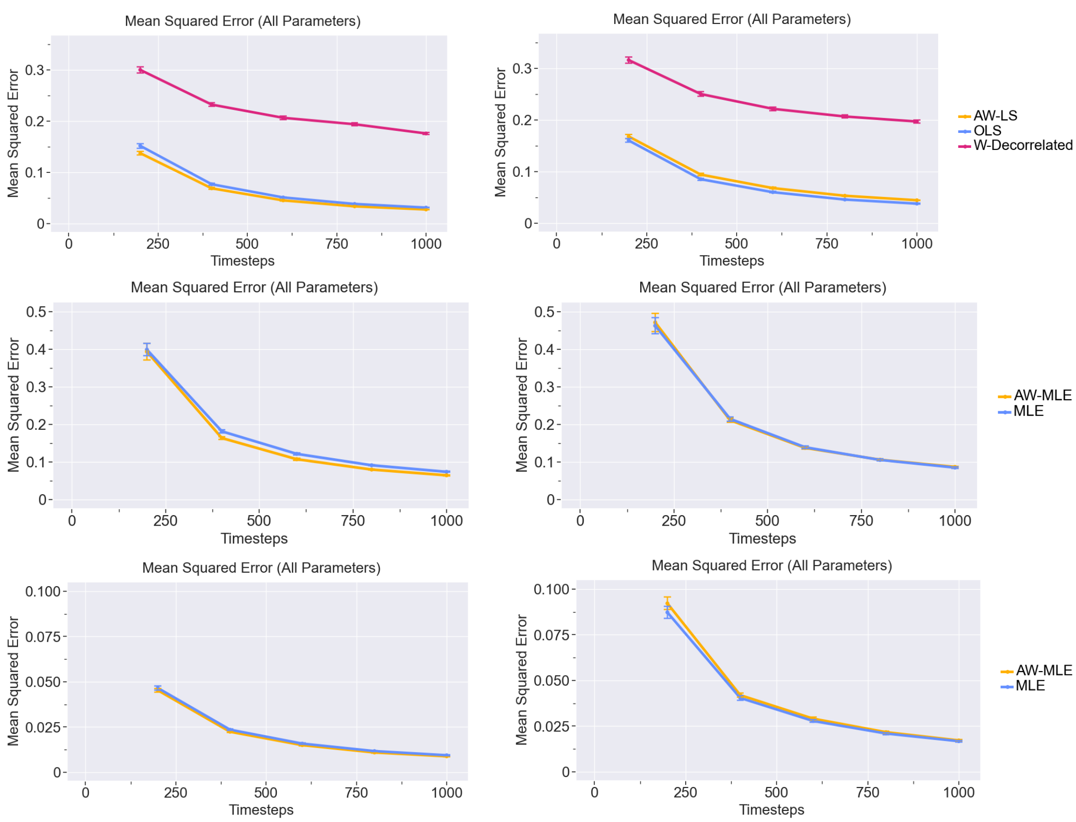

Figure 5: Mean squared error estimators of $\theta^*(\mathcal{P})$ for linear model (top), logistic regression model (middle), and generalized linear model for Poisson rewards (bottom). We consider simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

B Asymptotic Results

Throughout, $\|\cdot\|$ refers to the L_2 norm.

B.1 Definitions

Here we define convergence in probability and distribution that is uniform over the true parameter. We follow the definitions are based on those in Kasy [2019] and Van Der Vaart and Wellner [1996, Chapter 1.12].

Definition 1 (Uniform Convergence in Probability)**.** *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. We say that $Z_T(\mathcal{P}) \xrightarrow{P} c$ uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$ if for any $\epsilon > 0$,*

$$\sup_{\mathcal{P}\in\boldsymbol{P}} \mathbb{P}_{\mathcal{P},\eta}\left(\|Z_T(\mathcal{P}) - c\| > \epsilon\right) \to 0. \tag{12}$$

For simplicity of notation, throughout we denote $Z_T(\mathcal{P}) - c = o_{\mathcal{P}\in\boldsymbol{P}}(1)$ to mean $Z_T(\mathcal{P}) \xrightarrow{P} c$ uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$.

Definition 2 (Uniformly Stochastically Bounded)**.** *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. We say that $Z_T(\mathcal{P})$ is uniformly stochastically bounded over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$ if for any $\epsilon > 0$ there exists some $k < \infty$ such that*

$$\limsup_{T\to\infty} \sup_{\mathcal{P}\in\boldsymbol{P}} \mathbb{P}_{\mathcal{P},\eta}\left(\|Z_T(\mathcal{P})\| > k\right) < \epsilon.$$

Similarly we denote $Z_T(P) = O_{\mathcal{P}\in\boldsymbol{P}}(1)$ to mean $Z_T(\mathcal{P})$ is stochastically bounded uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$.

Definition 3 (Uniform Convergence in Distribution)**.** *Let $Z(\mathcal{P}) \in \mathbb{R}^{d_Z}$ and $\{Z_T(\mathcal{P})\}_{T\geq 1} \in \mathbb{R}^{d_Z}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. We say that $Z_T(\mathcal{P}) \xrightarrow{D} Z(\mathcal{P})$ uniformly over $\mathcal{P} \in \boldsymbol{P}$ as $T \to \infty$ if*

$$\sup_{\mathcal{P}\in\boldsymbol{P}} \sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P},\eta}\left[f\left(Z_T(\mathcal{P})\right)\right] - \mathbb{E}_{\mathcal{P},\eta}\left[f\left(Z(\mathcal{P})\right)\right] \right| \to 0, \tag{13}$$

where BL_1 is the set of functions $f : \mathbb{R}^{d_z} \to \mathbb{R}$ with $\|f(z)\|_\infty \leq 1$ and $|f(z) - f(z')| \leq \|z - z'\|$ for all $z, z' \in \mathbb{R}^{d_Z}$.

As discussed in Kasy [2019], Equation (12) holds if and only if for any $\epsilon > 0$ and any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\mathbb{P}_{\mathcal{P}_T,\eta}\left(\|Z_T(\mathcal{P}_T) - c\| > \epsilon\right) \to 0$.

Similarly, Equation (13) holds if and only if for any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P}_T,\eta}\left[f\left(Z_T(\mathcal{P}_T)\right)\right] - \mathbb{E}_{\mathcal{P}_T,\eta}\left[f\left(Z(\mathcal{P}_T)\right)\right] \right| \to 0$.

B.2 Consistency

We prove the first part of Theorem 1, i.e., that $\hat{\theta}_T \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$. We abbreviate $m_\theta(Y_t, X_t, A_t)$ with $m_{\theta,t}$. By definition of $\hat{\theta}_T$,

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

Note that $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| > \epsilon > 0$ implies that

$$\sup_{\theta\in\Theta : \|\theta-\theta^*(\mathcal{P})\|>\epsilon} \sum_{t=1}^T W_t m_{\theta,t} = \sup_{\theta\in\Theta} \sum_{t=1}^T W_t m_{\theta,t}.$$

Thus, the above two results imply the following inequality:

$$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\|\hat{\theta}_T - \theta^*(\mathcal{P})\| > \epsilon\right) \leq \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta : \|\theta-\theta^*(\mathcal{P})\|>\epsilon} \sum_{t=1}^T W_t m_{\theta,t} \geq \sum_{t=1}^T W_t m_{\theta^*(\mathcal{P}),t}\right)$$

$$= \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} \right\} - \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta^*(\mathcal{P}),t} \geq 0 \right)$$

$$= \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\Bigg(\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] + \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] \right\}$$
$$- \frac{1}{T}\sum_{t=1}^{T} \left\{ W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] + \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] \right\} \geq 0 \Bigg).$$

By triangle inequality,

$$\leq \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\Bigg(\underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] \right) \right\}}_{(a)}$$
$$+ \underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}) \middle| \mathcal{H}_{t-1} \right] \right\}}_{(b)}$$
$$\underbrace{- \frac{1}{T}\sum_{t=1}^{T} \left\{ W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] \right\}}_{(c)} \geq 0 \Bigg) \to 0. \quad (14)$$

We now show that the limit in Equation (14) above holds.

- Regarding term (c), by moment bounds of Condition 5 and Lemma 1, $\frac{1}{T}\sum_{t=1}^{T} \left\{ W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}] \right\} = o_{\mathcal{P}\in\mathbf{P}}(1)$.
- Regarding term (a), by Lemma 2, $\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}] \right) \right\} = o_{\mathcal{P}\in\mathbf{P}}(1)$.

Thus it is sufficient to show that term (b) is such that for some $\delta' > 0$,

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}] \right\} \leq -\delta' \text{ w.p. } 1. \quad (15)$$

By law of iterated expectations,

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

Since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$, we have that $\mathbb{E}_{\mathcal{P}}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}, X_t, A_t = a] = W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}, X_t, A_t = a]$. By Condition 1, we have that $W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}, X_t, A_t = a] = W_t\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a]$. Thus we have,

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

Since for all $\theta \in \Theta$, $\mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t] \leq 0$ with probability 1 by Condition 7 and since $0 < \frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by Condition 9,

$$\leq \sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1}) W_t^2 \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t}|X_t, A_t = a] da \middle| \mathcal{H}_{t-1} \right] \right\}.$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$,

$$= \sup_{\theta \in \Theta: \|\theta - \theta^*(\mathcal{P})\| > \epsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t} | X_t, A_t = a] da \middle| \mathcal{H}_{t-1} \right] \right\}.$$

By Condition 1 and since π_t^{sta} is pre-specified, we can drop the conditioning on $\mathcal{H}_{t-1}$, i.e.,

$$= \sup_{\theta \in \Theta: \|\theta - \theta^*(\mathcal{P})\| > \epsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t} | X_t, A_t = a] da \right] \right\}.$$

By law of iterated expectations,

$$= \sup_{\theta \in \Theta: \|\theta - \theta^*(\mathcal{P})\| > \epsilon} \left\{ \frac{1}{T\sqrt{\rho_{\max}}} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[m_{\theta,t} - m_{\theta^*(\mathcal{P}),t} \right] \right\} \leq -\frac{1}{\sqrt{\rho_{\max}}} \delta.$$

The last inequality above holds for some $\delta > 0$ for all sufficiently large T by Condition 8. Thus Equation (15) holds for $\delta' = \frac{1}{\sqrt{\rho_{\max}}} \delta$.

B.3 Asymptotic Normality

We prove the second part of Theorem 1, i.e., that

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{16}$$

B.3.1 Main Argument

The three results we show to ensure Equation (16) holds are as follows:

$$\Sigma_T(\mathcal{P})^{-1/2} \sqrt{T} \dot{M}_T(\theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{17}$$

For $\dddot{\epsilon}_{\dddot{m}} > 0$ as defined in Condition 6,

$$\sup_{\theta \in \Theta: \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1 = O_{\mathcal{P} \in \mathbf{P}}(1). \tag{18}$$

For matrix H positive definite,

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P} \in \mathbf{P}}(1). \tag{19}$$

For a reminder on the notation of $o_{\mathcal{P} \in \mathbf{P}}(1)$ and $O_{\mathcal{P} \in \mathbf{P}}(1)$ see definitions 12 and 2. For now, we assume that Equations (17), (18), and (19) hold; we will show they hold in Sections B.3.2, B.3.3, and B.3.4 respectively. Our argument is based on Van der Vaart [2000, Theorem of 5.41].

By differentiability Condition 2, since $\hat{\theta}_T$ is the maximizer of criterion $M_T(\theta)$,

$$0 = \dot{M}_T(\hat{\theta}_T).$$

By differentiability Condition 2 again and Taylor's theorem we have that for some random $\tilde{\theta}_T$ on the line segment between $\theta^*(\mathcal{P})$ and $\hat{\theta}_T$,

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

By rearranging terms and multiplying by $\sqrt{T}$,

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

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

Note that by the above equation and Equation (17), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\left[\ddot{M}_T(\theta^*(\mathcal{P}))+\frac{1}{2}(\hat{\theta}_T-\theta^*(\mathcal{P}))^\top\dddot{M}_T(\tilde{\theta}_T)\right]\sqrt{T}(\hat{\theta}_T-\theta^*(\mathcal{P}))$$
$$\xrightarrow{D}\mathcal{N}(0,I_d)\text{ uniformly over }\mathcal{P}\in\mathbf{P}. \quad (20)$$

By Equation (19), the probability that $\ddot{M}_T(\theta^*(\mathcal{P}))$ is invertible goes to 1 uniformly over $\mathcal{P}\in\mathbf{P}$. Thus by Equation (20), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\left[I_d+\frac{1}{2}(\hat{\theta}_T-\theta^*(\mathcal{P}))^\top\dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\right]\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T-\theta^*(\mathcal{P}))$$

$$=\left[I_d+\frac{1}{2}\Sigma_T(\mathcal{P})^{-1/2}(\hat{\theta}_T-\theta^*(\mathcal{P}))^\top\dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2}\right]$$
$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T-\theta^*(\mathcal{P}))\xrightarrow{D}\mathcal{N}(0,I_d)\text{ uniformly over }\mathcal{P}\in\mathbf{P}. \quad (21)$$

We now show that $\frac{1}{2}\Sigma_T(\mathcal{P})^{-1/2}(\hat{\theta}_T-\theta^*(\mathcal{P}))^\top\dddot{M}_T(\tilde{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2}=o_{\mathcal{P}\in\mathbf{P}}(1)$. It is sufficient to show that $\|\Sigma_T(\mathcal{P})^{-1/2}\|\|\hat{\theta}_T-\theta^*(\mathcal{P})\|\|\dddot{M}_T(\tilde{\theta}_T)\|_1\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\|\|\Sigma_T(\mathcal{P})^{1/2}\|=o_{\mathcal{P}\in\mathbf{P}}(1)$.

- By Condition 5, the minimum eigenvalue of $\Sigma_T(\mathcal{P})$ is bounded uniformly above some constant greater than zero, so $\sup_{\mathcal{P}\in\mathbf{P}}\|\Sigma_T(\mathcal{P})^{-1/2}\|=O(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T-\theta^*(\mathcal{P})\|=o_{\mathcal{P}\in\mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\mathbb{1}_{\|\tilde{\theta}_T-\theta^*(\mathcal{P})\|\leq\epsilon_{\dddot{m}}}=o_{\mathcal{P}\in\mathbf{P}}(1)$. Thus by Equation (18), $\dddot{M}_T(\tilde{\theta}_T)=O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Equation (19), the minimum eigenvalue of $-\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ is bounded above that of positive definite matrix H. Thus $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\|=O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Condition 5, $\sup_{\mathcal{P}\in\mathbf{P}}\|\Sigma_T(\mathcal{P})^{1/2}\|=O(1)$.

Thus, by Slutsky's Theorem and Equation (21), we have that

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T-\theta^*(\mathcal{P}))\xrightarrow{D}\mathcal{N}(0,I_d)\text{ uniformly over }\mathcal{P}\in\mathbf{P}. \quad (22)$$

Lastly, to show our desired result, that $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T-\theta^*(\mathcal{P}))\xrightarrow{D}\mathcal{N}(0,I_d)$ uniformly over $\mathcal{P}\in\mathbf{P}$, by Equation (22) and Slutsky's Theorem it is sufficient to show that $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2}\xrightarrow{P}I_d$ uniformly over $\mathcal{P}\in\mathbf{P}$. Note if we can show that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\xrightarrow{P}I_d$ uniformly over $\mathcal{P}\in\mathbf{P}$, then $\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\Sigma_T(\mathcal{P})^{1/2}=\Sigma_T(\mathcal{P})^{-1/2}\left[I_d+o_{\mathcal{P}\in\mathbf{P}}(1)\right]\Sigma_T(\mathcal{P})^{1/2}=I_d+\Sigma_T(\mathcal{P})^{-1/2}o_{\mathcal{P}\in\mathbf{P}}(1)\Sigma_T(\mathcal{P})^{1/2}=I_d+o_{\mathcal{P}\in\mathbf{P}}(1)$. The last limit holds since $\|\Sigma_T(\mathcal{P})^{-1/2}\|=O_{\mathcal{P}\in\mathbf{P}}(1)$ and $\|\Sigma_T(\mathcal{P})^{1/2}\|=O_{\mathcal{P}\in\mathbf{P}}(1)$ by Condition 5 (use the same argument as that used in the bullet points below Equation (21)).

Thus it is sufficient to show that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\xrightarrow{P}I_d$ uniformly over $\mathcal{P}\in\mathbf{P}$. By Taylor's Theorem, for some random $\bar{\theta}_T$ on the line segment between $\hat{\theta}_T$ and $\theta^*(\mathcal{P})$,

$$\ddot{M}_T(\hat{\theta}_T)=\ddot{M}_T(\theta^*(\mathcal{P}))+\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T-\theta^*(\mathcal{P})).$$

Recall that the probability the inverse of $\ddot{M}_T(\theta^*(\mathcal{P}))$ exists goes to 1 by Equation (19) (use the same argument as that used in the bullet points below Equation (21)). Thus we have that $\ddot{M}_T(\hat{\theta}_T)\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ equals the following:

$$\left[\ddot{M}_T(\theta^*(\mathcal{P}))+\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T-\theta^*(\mathcal{P}))\right]\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$$
$$=I_d+\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T-\theta^*(\mathcal{P}))\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$$

Note that $\dddot{M}_T(\bar{\theta}_T)(\hat{\theta}_T-\theta^*(\mathcal{P}))\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}=o_{\mathcal{P}\in\mathbf{P}}(1)$ because

- By uniform consistency of $\hat{\theta}_T$, $\mathbb{1}_{\|\tilde{\theta}_T-\theta^*(\mathcal{P})\|\leq\epsilon_{\ddot{m}}} = o_{\mathcal{P}\in\mathbf{P}}(1)$. Thus by Equation (18), $\ddot{M}_T(\tilde{\theta}_T) = O_{\mathcal{P}\in\mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| = o_{\mathcal{P}\in\mathbf{P}}(1)$.
- By Equation (19), $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\| = O_{\mathcal{P}\in\mathbf{P}}(1)$.

B.3.2 Asymptotic Normality of $\Sigma_T(\mathcal{P})^{-1/2}\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P}))$

We will show that Equation (17) holds by applying a martingale central limit theorem. For notational convenience, we let $\dot{m}_{\theta,t} := \dot{m}_\theta(Y_t, X_t, A_t)$. Note that by definition $\Sigma_T(\mathcal{P})^{-1/2}\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P})) = \Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t\dot{m}_{\theta^*(\mathcal{P}),t}$. We first show that $\left\{\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}W_t\dot{m}_{\theta^*(\mathcal{P}),t}\right\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=0}^T$. For any $t \in [1\colon T]$,

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{P},\pi}\left[\frac{1}{\sqrt{T}}\Sigma_T(\mathcal{P})^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*(\mathcal{P}),t}\Big|\mathcal{H}_{t-1}\right] \\
&\underset{(a)}{=} \frac{1}{\sqrt{T}}\mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P}}\left[\Sigma_T(\mathcal{P})^{-1/2}W_t\mathbf{c}^\top\dot{m}_{\theta^*(\mathcal{P}),t}\big|\mathcal{H}_{t-1}, X_t, A_t\right]\Big|\mathcal{H}_{t-1}\right] \\
&\underset{(b)}{=} \frac{1}{\sqrt{T}}\Sigma_T(\mathcal{P})^{-1/2}\mathbb{E}_{\mathcal{P},\pi}\left[W_t\mathbf{c}^\top\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\big|\mathcal{H}_{t-1}, X_t, A_t\right]\Big|\mathcal{H}_{t-1}\right] \underset{(c)}{=} 0
\end{aligned}
$$

- Above, (a) holds by law of iterated expectations.
- (b) holds since $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ and since $\Sigma_T(\mathcal{P})$ are a function of stabilizing policies $\{\pi_t^{\text{sta}}\}_{t\geq 1}$, which are pre-specified.
- By Condition 1, $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\big|\mathcal{H}_{t-1}, X_t, A_t\right] = \mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\big|X_t, A_t\right]$. Equality (c) holds because $\mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^*(\mathcal{P}),t}\big|X_t, A_t\right] = 0$ with probability 1 by Condition 7; note that $\theta^*(\mathcal{P})$ is a critical point of $\mathbb{E}_{\mathcal{P}}[m_{\theta,t}|X_t, A_t]$.

By Cramer-Wold device, to show that Equation (17) holds, it is sufficient to show that for any fixed $\mathbf{c} \in \mathbb{R}^d$ with $\|\mathbf{c}\|_2 = 1$, that $\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t\dot{m}_{\theta^*(\mathcal{P}),t} \xrightarrow{D} \mathcal{N}\left(0, \mathbf{c}^\top I_d\mathbf{c}\right)$ uniformly over $\mathcal{P} \in \mathbf{P}$. We now apply Theorem 2, a uniform version of the martingale central limit theorem of Dvoretzky [1972]; while the original theorem holds for any fixed $\mathcal{P}$, we can show uniform convergence in distribution by ensuring that the conditions of the theorem hold uniformly over $\mathcal{P} \in \mathbf{P}$ (see Definition 3). By Theorem 2, it is sufficient to show that the following two conditions hold:

1. **Conditional Variance:** $\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[\left\{\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\theta^*(\mathcal{P}),t}\right\}^2\Big|\mathcal{H}_{t-1}\right] \xrightarrow{P} \sigma^2$ uniformly over $\mathcal{P} \in \mathbf{P}$.

2. **Conditional Lindeberg:** For any $\delta > 0$,
$\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[\left\{\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\theta^*(\mathcal{P}),t}\right\}^2\mathbb{1}_{|\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}W_t\dot{m}_{\theta^*(\mathcal{P}),t}|>\delta\sqrt{T}}\Big|\mathcal{H}_{t-1}\right] \xrightarrow{P} 0$ uniformly over $\mathcal{P} \in \mathbf{P}$.

1. Conditional Variance

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[\left(\mathbf{c}^\top W_t\Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\theta^*(\mathcal{P}),t}\right)^2\Big|\mathcal{H}_{t-1}\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}\Big|\mathcal{H}_{t-1}\right] \\
&\underset{(a)}{=} \mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\left\{\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2}\Big|\mathcal{H}_{t-1}\right]\right\}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c}
\end{aligned}
$$

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

$$\underset{(c)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \left\{ \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \middle| \mathcal{H}_{t-1}, X_t, A_t = a \right] da \middle| \mathcal{H}_{t-1} \right] \right\} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c}$$

$$\underset{(d)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \left\{ \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \middle| X_t \right] \middle| \mathcal{H}_{t-1} \right] \right\} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c}$$

$$\underset{(e)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \left\{ \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \right] \right\} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c}$$

$$\underset{(f)}{=} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \Sigma_T(P) \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} = \mathbf{c}^\top I_d \mathbf{c}$$

- Above, (a) holds since $\Sigma_T(\mathcal{P})$ are a function of stabilizing policies $\{\pi_t^{\text{sta}}\}_{t \geq 1}$, which are pre-specified.
- Equality (b) holds by law of iterated expectations.
- Equality (c) holds since $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}} \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$.
- Equality (d) holds because by Condition 1, $\mathbb{E}_{\mathcal{P}}[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} | \mathcal{H}_{t-1}, X_t, A_t = a] = \mathbb{E}_{\mathcal{P}}[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} | X_t, A_t = a]$ and by law of iterated expectations.
- Equality (e) holds because by Condition 1, the distribution of X_t does not depend on $\mathcal{H}_{t-1}$, so $\mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \middle| X_t \right] \middle| \mathcal{H}_{t-1} \right] = \mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \middle| X_t \right] \right] = \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \right]$; the last equality holds by law of iterated expectations.
- Equality (f) holds by definition.

2. Conditional Lindeberg

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

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

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

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

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

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

$$\underset{(e)}{=} \frac{\rho_{\max}}{T^2 \delta^2} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}} \left[\mathbb{E}_{\mathcal{P}} \left[\left(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \right)^2 \middle| X_t \right] \middle| \mathcal{H}_{t-1} \right]$$

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

- Above, inequality (a) holds because $\mathbb{1}_{\left|W_t \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}\right| > \sqrt{T}\delta} = 1$ if and only if $W_t^2 \frac{1}{T\delta^2} \mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} > 1$.
- Inequality (b) holds because by Condition 9, $W_t^2 \leq \rho_{\max}$ with probability 1.
- Equality (c) holds by the law of iterated expectations.
- Equality (d) holds since $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}} \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$.
- Equality (e) holds because by Condition 1, $\mathbb{E}_{\mathcal{P}}\left[(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c})^2 \middle| \mathcal{H}_{t-1}, X_t, A_t = a \right] = \mathbb{E}_{\mathcal{P}}\left[(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c})^2 \middle| X_t \right]$ and by law of iterated expectations.
- Equality (f) holds since the distribution of X_t does not depend on $\mathcal{H}_{t-1}$ by Condition 1 and by law of iterated expectations.
- Regarding limit (g), it is sufficient to show that $\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T. By Condition 5, the minimum eigenvalue of $\Sigma_T(P)$ is bounded above zero uniformly over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T; this bounds the maximum eigenvalue of $\Sigma_T(P)^{-1}$. Also by Condition 5 the fourth moment of $\dot{m}_{\theta^*(\mathcal{P}),t}$ with respect to $\mathcal{P}$ and policy π_t^{sta} is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ and $t \geq 1$. With these two properties we have that $\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left(\mathbf{c}^\top \Sigma_T(\mathcal{P})^{-1/2} \dot{m}_{\theta^*(\mathcal{P}),t}^{\otimes 2} \Sigma_T(\mathcal{P})^{-1/2} \mathbf{c} \right)^2 \right]$ is uniformly bounded over $\mathcal{P} \in \mathbf{P}$ for all sufficiently large T.

B.3.3 Showing that $\sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1$ is bounded in probability

Recall that for any $B \in \mathbb{R}^{d \times d \times d}$, we denote $\|B\|_1 = \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d |B_{i,j,k}|$. We abbreviate $\dddot{m}_\theta(Y_t, X_t, A_t)$ with $\dddot{m}_{\theta,t}$.

By triangle inequality, $\left\| \dddot{M}_T(\theta) \right\|_1 = \left\| \frac{1}{T} \sum_{t=1}^T W_t \dddot{m}_{\theta,t} \right\|_1 \leq \frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}_{\theta,t} \right\|_1$. Thus we have that

$$\sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1 \leq \sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}_{\theta,t} \right\|_1 .$$

By Condition 6 (ii), there exists a function $\dddot{m}$ (note it is not indexed by θ) such that for all $\mathcal{P} \in \mathbf{P}$, we have that $\sup_{\theta \in \Theta : \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \left\| \dddot{m}_{\theta,t} \right\|_1 \leq \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1$.

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

Adding and subtracting $\frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi} \left[W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right]$,

$$= \frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right] + \mathbb{E}_{\mathcal{P},\pi} \left[W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right] .$$

By second moment bounds on $\left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1$ from Condition 6 (i), by Lemma 1, we have that $\frac{1}{T} \sum_{t=1}^T W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right] = o_{\mathcal{P} \in \mathbf{P}}(1)$.

$$= o_{\mathcal{P} \in \mathbf{P}}(1) + \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi} \left[W_t \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right]$$

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

$$\leq o_{\mathcal{P} \in \mathbf{P}}(1) + \frac{1}{T \sqrt{\rho_{\min}}} \sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi} \left[W_t^2 \left\| \dddot{m}(Y_t, X_t, A_t) \right\|_1 | \mathcal{H}_{t-1} \right]$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$ and by Condition 1,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) + \frac{1}{T\sqrt{\rho_{\min}}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\| \dddot{m}(Y_t, X_t, A_t) \|_1 \right] = O_{\mathcal{P}\in\mathbf{P}}(1).$$

Note that by Jensen's inequality, $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\| \dddot{m}(Y_t, X_t, A_t) \|_1 \right] \leq \sqrt{\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\| \dddot{m}(Y_t, X_t, A_t) \|_1^2 \right]}$. By Condition 6 (i), $\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\| \dddot{m}(Y_t, X_t, A_t) \|_1^2 \right]$ is bounded, which implies the final limit above.

B.3.4 Lower bounding $-\ddot{M}_T(\theta^*(\mathcal{P}))$

We now show that $-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1)$, for positive definite matrix H introduced in Condition 7 (ii).

By Condition 5 and Lemma 1, $\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right] = o_{\mathcal{P}\in\mathbf{P}}(1)$, so

$$-\ddot{M}_T(\theta^*(\mathcal{P})) = -\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\theta^*(\mathcal{P}),t} = o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t \ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right]$$

By law of iterated expectations,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t \mathbb{E}_{\mathcal{P}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1}, X_t, A_t \right] | \mathcal{H}_{t-1} \right]$$

By Condition 1,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t \mathbb{E}_{\mathcal{P}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | X_t, A_t \right] | \mathcal{H}_{t-1} \right]$$

By Condition 7, we have that $\mathbb{E}_{\mathcal{P}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | X_t, A_t \right] \preceq 0$; recall that $\theta^*(\mathcal{P})$ is a maximizing value of $\mathbb{E}_{\mathcal{P},\pi} \left[m_{\theta,t} | X_t, A_t \right]$. Also since $\frac{W_t}{\sqrt{\rho_{\max}}} \leq 1$ with probability 1 by Condition 9,

$$\succeq o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t^2 \mathbb{E}_{\mathcal{P},\pi} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | X_t, A_t \right] | \mathcal{H}_{t-1} \right]$$

Since $W_t^2 = \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}$,

$$= o_{\mathcal{P}\in\mathbf{P}}(1) - \frac{1}{T\sqrt{\rho_{\max}}}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right]$$

Note that for any $t \geq 1$, $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} | \mathcal{H}_{t-1} \right] = \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} \right]$ because $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ are pre-specified. Recall that by Condition 7 for all sufficiently large T, $-\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\ddot{m}_{\theta^*(\mathcal{P}),t} \right] \succeq H$ for all $\mathcal{P} \in \mathbf{P}$. Thus our final result is that

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1). \tag{23}$$

B.4 Lemmas and Other Helpful Results

Theorem 2 (Uniform Martingale Central Limit Theorem)**.** *Let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a sequence of random variables whose distributions are defined by some $\mathcal{P} \in \boldsymbol{P}$ and some nuisance component η. Moreover, let $\{Z_T(\mathcal{P})\}_{T\geq 1}$ be a martingale difference sequence with respect to $\mathcal{F}_t$, meaning $\mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})|\mathcal{F}_{t-1}] = 0$ for all $t \geq 1$ and $\mathcal{P} \in \boldsymbol{P}$.*

(a) $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})^2|\mathcal{F}_{t-1}] \xrightarrow{P} \sigma^2$ uniformly over $\mathcal{P} \in \boldsymbol{P}$, where σ^2 is a constant $0 < \sigma^2 < \infty$.

(b) For any $\epsilon > 0$, $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\eta}[Z_t(\mathcal{P})^2 \mathbb{1}_{|Z_t(\mathcal{P})|>\epsilon} | \mathcal{F}_{t-1}] \xrightarrow{P} 0$ uniformly over $\mathcal{P} \in \boldsymbol{P}$.

Under the above conditions,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t(\mathcal{P}) \xrightarrow{D} \mathcal{N}(0, \sigma^2) \text{ uniformly over } \mathcal{P} \in \boldsymbol{P}.$$

Proof: By by Kasy [2019, Lemma 1], it is sufficient to show that for any sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$ with $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t(\mathcal{P}_T) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$. In this setting, since $\mathcal{P}_T$ depends on T, we consider triangular array asymptotics and additionally index by T, e.g., $\mathcal{F}_{T,t}$.

Note that $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}_T,\eta}[Z_t(\mathcal{P}_T)^2 | \mathcal{F}_{T,t-1}] \xrightarrow{P} \sigma^2$, by Kasy [2019, Lemma 1] and condition (a) above.

Also, for any $\epsilon > 0$, $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}_T,\eta} \left[Z_t(\mathcal{P}_T)^2 \mathbb{1}_{|Z_t(\mathcal{P}_T)|>\epsilon} \middle| \mathcal{F}_{T,t-1} \right] \xrightarrow{P} 0$, by Kasy [2019, Lemma 1] and condition (b) above.

Thus by the martingale central limit theorem of Dvoretzky [1972], we have that for the sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} Z_t(\mathcal{P}_T) \xrightarrow{D} \mathcal{N}(0, 1).$$

Since the sequence $\{\mathcal{P}_T\}_{T=1}^{\infty}$ were chosen arbitrarily from $\mathbf{P}$, the desired result is implied again by Kasy [2019, Lemma 1].

Lemma 1. *Let $f(Y_t, X_t, A_t) \in \mathbb{R}^{d_f}$ be a function such that $\sup_{\mathcal{P} \in \boldsymbol{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\| f(Y_t, X_t, A_t) \|^2 \right] < m$ for some $m < \infty$. Under Conditions 1 and 9,*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ W_t f(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t f(Y_t, X_t, A_t) | \mathcal{H}_{t-1}] \right\} = O_{\mathcal{P} \in \boldsymbol{P}}(1). \tag{24}$$

Note that the above equation implies that

$$\frac{1}{T} \sum_{t=1}^{T} \left\{ W_t f(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t f(Y_t, X_t, A_t) | \mathcal{H}_{t-1}] \right\} = o_{\mathcal{P} \in \boldsymbol{P}}(1).$$

Lemma 1 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: We denote the $k^{\text{th}} \in [1\colon d_f]$ dimension of vector $f(Y_t, X_t, A_t)$ as $f^k(Y_t, X_t, A_t)$. It is sufficient to show the result for any dimension of vector $f(Y_t, X_t, A_t)$. For notational convenience, let $f_t := f^k(Y_t, X_t, A_t)$. Let $\epsilon > 0$.

$$\begin{aligned}
&\sup_{\mathcal{P} \in \mathbf{P}} \mathbb{P}_{\mathcal{P},\pi} \left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right\} \right| > \epsilon \right) \\
&\underset{(a)}{\leq} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \mathbb{E}_{\mathcal{P},\pi} \left[\left(\sum_{t=1}^{T} \left\{ W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right\} \right)^2 \right] \\
&\underset{(b)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[\left\{ W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right\}^2 \right] \\
&\underset{(c)}{\leq} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi} \left[W_t^2 f_t^2 \right] \\
&\underset{(d)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} W_t^2 \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]
\end{aligned}$$

$$\underset{(e)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P}\in\mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right]$$

$$\underset{(f)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P}\in\mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[f_t^2 \right] \underset{(g)}{\leq} \frac{4m}{\epsilon^2}$$

- 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., $\mathbb{E}_{\mathcal{P},\pi}\big[W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \big| \mathcal{H}_{t-1}\big] = 0$; this implies that cross terms disappear, i.e., for $t > s$,

$$\mathbb{E}_{\mathcal{P},\pi}\left[\left(W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right) \left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \right]$$

$$= \mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P},\pi}\left[\left(W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \right) \left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \middle| \mathcal{H}_{t-1} \right] \right]$$

Since $s > t$,

$$= \mathbb{E}_{\mathcal{P},\pi}\left[\left(W_s f_s - \mathbb{E}_{\mathcal{P},\pi}[W_s f_s | \mathcal{H}_{s-1}] \right) \mathbb{E}_{\mathcal{P},\pi}\left[W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \middle| \mathcal{H}_{t-1} \right] \right] = 0.$$

- (c) holds because $\mathbb{E}_{\mathcal{P},\pi}\left[\{ W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}] \}^2 \right] = \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2 \right] - \mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P},\pi}[W_t f_t | \mathcal{H}_{t-1}]^2 \right] \leq \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2 \right]$.
- (d) holds by law of iterated expectations.
- (e) holds because $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$.
- (f) holds since by Condition 1, $\mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t] = \mathbb{E}_{\mathcal{P}}[f_t^2 | X_t, A_t]$ and by law of iterated expectations $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[f_t^2 \right] = \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[f_t^2 | X_t, A_t = a] da \right]$.
- (g) holds since $\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[f_t^2 \right] < m < \infty$.

Lemma 2. *Let $m_{\theta,t} := m_\theta(Y_t, X_t, A_t)$. Under Conditions 1, 3, 4, 5, 7, and 9,*

$$\sup_{\theta\in\Theta} \left\{ \frac{1}{T} \sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t} | \mathcal{H}_{t-1}] \right\} = O_{\mathcal{P}\in\boldsymbol{P}}(1). \tag{25}$$

Lemma 1 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 Van Der Vaart and Wellner [1996, Theorem 2.4.1].

Proof:

Finite Bracketing Number: Let $\delta > 0$. We construct a set B_δ which is made up of pairs of functions (l, u). We show that we can find B_δ that satisfies the following:

(a) For any $\theta \in \Theta$, we can find $(l, u) \in B_\delta$ such that
(i) $l(y, x, a) \leq m_\theta(y, x, a) \leq u(y, x, a)$ for all (x, y) in the joint support of $\{\mathcal{P} \in \mathbf{P}\}$ and all $a \in \mathcal{A}$.
(ii) $\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[|u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)| \right] \leq \delta$.

(b) The number of pairs in this set is finite, i.e., $|B_\delta| < \infty$.

(c) For any $(l, u) \in B_\delta$, for some $m < \infty$ which does no depend on δ,
$\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[u(Y_t, X_t, A_t)^2 \right] \leq m$ and $\sup_{\mathcal{P}\in\mathbf{P}, t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[l(Y_t, X_t, A_t)^2 \right] \leq m$.

Showing that we can find B_δ that satisfy (a), means that $|B_\delta|$ is an upper bound on the bracketing number of $\{m_\theta : \theta \in \Theta\}$. For more information on bracketing functions, see Van Der Vaart and Wellner [1996] and Van der Vaart [2000].

To construct B_δ, we follow a similar argument to Example 19.7 of Van der Vaart [2000] (page 271). Make a grid over Θ with meshwidth $\lambda/2 > 0$ and let the points in this grid be the set $G_{\lambda/2} \subseteq \Theta$; we will specify λ later. Note that by construction, for any $\theta \in \Theta$ we can find a $\theta \in G_{\lambda/2}$ such that $\|\theta' - \theta\| \leq \lambda$.

By our Lipschitz Condition 4, we have that for any $\theta, \theta' \in \Theta$, $|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\|$ for function g such that for some $m_g < \infty$,

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2] \leq m_g. \tag{26}$$

We now show that we can choose $B_\delta = \left\{(m_\theta - g(Y_t, X_t, A_t), m_\theta + g(Y_t, X_t, A_t)) : \theta \in G_{\lambda/2}\right\}$. Note that by compactness of Θ, Condition 3, the number of points in $G_{\lambda/2}$ is finite, so (b) above holds.

To show that (a) holds for our choice of B_δ, recall that for any $\theta \in \Theta$ we can find a $\theta' \in G_{\lambda/2}$ such that $\|\theta' - \theta\| \leq \lambda$. Also, by the Lipschitz Condition 4, $|m_\theta(Y_t, X_t, A_t) - m_{\theta'}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta'\| \leq g(Y_t, X_t, A_t)\lambda$. Thus we have that

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

Note that

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[m_{\theta'}(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\lambda - \{m_{\theta'}(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\lambda\}\right]$$
$$= 2\lambda \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)\right] \leq 2\lambda\sqrt{m_g} < \infty.$$

The inequalities above hold by Equation (26) and since $\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)\right] \leq \sqrt{\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right]}$ by Jensen's inequality. (a) above holds for our choice of B_δ by letting meshwidth $\lambda = \delta/(2\sqrt{m_g})$.

We now show that (c) above holds. Note that

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\right\}^2\right]$$
$$\leq 3 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[m_\theta(Y_t, X_t, A_t)^2\right] + 3 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right]. \tag{27}$$

Note that the above upper bound, Equation (27), also holds for $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\right\}^2\right]$.

Since, $m_\theta(Y_t, X_t, A_t) = m_\theta(Y_t, X_t, A_t) - m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t) + m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)$,

$$\leq 9 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t, X_t, A_t) - m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\right\}^2\right]$$
$$+ 9 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^2\right]$$
$$+ 3 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right].$$

Note that $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^2\right]$ is bounded by our moment Condition 5 and that $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right]$ is bounded by Equation (26).

By our Lipschitz Condition 4, for any $\theta \in \Theta$, $|m_\theta(Y_t, X_t, A_t) - m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)| \leq g(Y_t, X_t, A_t)\|\theta - \theta^*(\mathcal{P})\|$. Thus,

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[\left\{m_\theta(Y_t, X_t, A_t) - m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\right\}^2\right]$$
$$\leq \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}\left[g(Y_t, X_t, A_t)^2\right] \|\theta - \theta^*(\mathcal{P})\|^2.$$

The above is bounded by Equation (26) and by compactness of Θ, Condition 3. Thus (c) above holds for our choice of B_δ.

Main Argument: We now show that for any $\epsilon > 0$,

$$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} > \epsilon\right) \to 0. \tag{28}$$

An analogous argument can be made to show that
$\sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\sup_{\theta\in\Theta}\left\{-\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} > \epsilon\right) \to 0.$

Let $\delta > 0$; we will choose δ later. Let B_δ be the set of pairs of functions as constructed earlier.

$$\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\}$$

Note that by (a), we get the following upper bound:

$$\leq \max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t u(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}[W_t l(Y_t, X_t, A_t)|\mathcal{H}_{t-1}]\right\}.$$

By adding and subtracting $\mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t, X_t, A_t)\middle|\mathcal{H}_{t-1}\right]$ and triangle inequality,

$$\leq \max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t\left\{u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)\right\}\middle|\mathcal{H}_{t-1}\right]\right\}$$
$$+ \max_{(l,u)\in B_\delta}\left\{\frac{1}{T}\sum_{t=1}^{T} W_t u(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t, X_t, A_t)\middle|\mathcal{H}_{t-1}\right]\right\}.$$

Note that by Condition 9, $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}} \leq \sqrt{\rho_{\max}}$ with probability 1, so
$\mathbb{E}_{\mathcal{P},\pi}\left[W_t\left\{u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)\right\}\middle|\mathcal{H}_{t-1}\right] \leq \frac{1}{\sqrt{\rho_{\max}}}\mathbb{E}_{\mathcal{P},\pi}\left[W_t^2\left\{u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)\right\}\middle|\mathcal{H}_{t-1}\right]$
$= \frac{1}{\sqrt{\rho_{\max}}}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[u(Y_t, X_t, A_t) - l(Y_t, X_t, A_t)\right] \leq \frac{1}{\sqrt{\rho_{\max}}}\delta$; the last equality holds by Condition 1 and the last inequality holds by (a). And since $\max_{i\in[1:\,n]}\{a_i\} \leq \sum_{i=1}^{n}|a_i|$,

$$\leq \frac{1}{\sqrt{\rho_{\max}}}\delta + \sum_{(l,u)\in B_\delta}\left|\frac{1}{T}\sum_{t=1}^{T} W_t u(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t, X_t, A_t)|\mathcal{H}_{t-1}\right]\right|$$

By Lemma 1 and (c), for any $(l,u) \in B_\delta$, $\frac{1}{T}\sum_{t=1}^{T} W_t u(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P},\pi}\left[W_t u(Y_t, X_t, A_t)\middle|\mathcal{H}_{t-1}\right] = o_{\mathcal{P}\in\mathbf{P}}(1)$. Since $|B_\delta| < \infty$ by (b), the convergence holds for all $(l,u) \in B_\delta$ simultaneously, so

$$= \frac{1}{\sqrt{\rho_{\max}}}\delta + o_{\mathcal{P}\in\mathbf{P}}(1).$$

Equation (28) holds by choosing $\delta = \sqrt{\rho_{\max}}\epsilon/2$.

B.5 Least-Squares Estimator

We use $\phi(X_t, A_t)$ to denote a feature vector that constructed using context X_t and action A_t.

Condition 10 (Linear Expected Outcome)**.** *For all $\mathcal{P} \in \mathbf{P}$, the following holds w.p. 1,*

$$\mathbb{E}_{\mathcal{P}}\left[Y_t|X_t, A_t\right] = \phi(X_t, A_t)^\top\theta^*(\mathcal{P}).$$

Condition 11 (Moment Conditions for Least Squares)**.** *The fourth moments of $\phi(X_t, A_t)\left(Y_t - \phi(X_t, A_t)^\top\theta^*(\mathcal{P})\right)$ and $\phi(X_t, A_t)$ with respect to $\mathcal{P}$ and policy π_t^{sta} are respectively bounded uniformly over $\mathcal{P} \in \mathbf{P}$ and $t \geq 1$.*

Also the minimum eigenvalue of $\Sigma_T(\mathcal{P}) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\phi(Y_t, X_t, A_t)^{\otimes 2}\left(Y_t - \phi(Y_t, X_t, A_t)^\top\theta^*(\mathcal{P})\right)^2\right]$ and $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\phi(X_t, A_t)^{\otimes 2}\right]$ respectively are both bounded above constant some constant greater than zero for all $\mathcal{P} \in \mathbf{P}$.

Condition 12 (Importance Ratios for Least Squares)**.** *Let $\rho_{\min} > 0$ and $\rho_{\max,T} > 0$ be a non-random sequence such that $\frac{\rho_{\max,T}}{T} \to 0$. $\{\pi_t^{\text{sta}}\}_{t=1}^T$ are pre-specified and do not depend on data $\{Y_t, X_t, A_t\}_{t=1}^T$. For all $\mathcal{P} \in \boldsymbol{P}$, the following holds w.p. 1,*

$$\rho_{\min} \leq \frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})} \leq \rho_{\max,T}.$$

Note that Condition 12 allows $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ to go to zero at some rate for stabilizing policies $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ that are strictly bounded away from 0 and 1.

We now define the AW-LS estimator for $\theta^*(\mathcal{P}) \in \mathbb{R}^d$:

$$\hat{\theta}_T^{\text{AW-LS}} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \left\{ -\sum_{t=1}^T W_t \left(Y_t - \phi(X_t, A_t)^\top \theta \right)^2 \right\}. \tag{29}$$

Theorem 3 (Consistency and Asymptotic Normality of Adaptively-Weighted Least Squares Estimator)**.** *Under Conditions 1, 10, 11, and 12,*

$$\Sigma_T(\mathcal{P})^{-1/2} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2} \right) \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P}) \right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \boldsymbol{P},$$

where $\Sigma_T(\mathcal{P}) := \frac{1}{T} \sum_{t=1}^T \phi(X_t, A_t)^{\otimes 2} \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right)^2$.

Proof: By taking the derivative of Equation (29) with respect to the parameters, we have that

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

By rearranging terms, we have that

$$-\frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P}) \right). \tag{30}$$

We first show that the following holds:

$$\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{31}$$

Equation (31) holds by a similar argument as that used in Section B.3.2, for $\dot{m}_\theta(Y_t, X_t, A_t) = \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right)$ by showing that the conditions of Theorem 2 hold. It can be checked that all the arguments hold even when we allow $\rho_{\max,T}$ to grow at a rate such that $\frac{\rho_{\max,T}}{T} \to 0$.

By Equations (30) and (31),

$$\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P}) \right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{32}$$

By Equation (32), to ensure that $\hat{\theta}_T^{\text{AW-LS}} \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$, it is sufficient to show that the minimum eigenvalue of $\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^T W_t \phi(X_t, A_t)^{\otimes 2}$ goes to infinity uniformly over $\mathcal{P} \in \mathbf{P}$ as $T \to \infty$.

By Condition 11, the maximum eigenvalue of $\Sigma_T(\mathcal{P})$ is bounded uniformly over $\mathcal{P} \in \mathbf{P}$, so the minimum eigenvalue of $\Sigma_T(\mathcal{P})^{-1/2}$ is bounded uniformly above 0. Thus it is sufficient to show that the minimum eigenvalue of $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2}$ goes to infinity uniformly over $\mathcal{P} \in \mathbf{P}$ as $T \to \infty$.

Note that by Lemma 1 and Condition 11,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2} - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \phi(X_t, A_t)^{\otimes 2} \middle| \mathcal{H}_{t-1} \right] = O_{\mathcal{P} \in \mathbf{P}}(1). \tag{33}$$

Note that by law of iterated expectations,

$$\mathbb{E}_{\mathcal{P},\pi} \left[W_t \phi(X_t, A_t)^{\otimes 2} \middle| \mathcal{H}_{t-1} \right]$$

$$= \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}} \left[W_t \phi(X_t, A_t)^{\otimes 2} | \mathcal{H}_{t,1}, X_t, a \right] da \middle| \mathcal{H}_{t-1} \right].$$

By Condition 1 and since $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$,

$$= \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \sqrt{\frac{\pi_t(a, X_t, \mathcal{H}_{t-1})}{\pi_t^{\text{sta}}(a, X_t)}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}} \left[\phi(X_t, A_t)^{\otimes 2} | X_t, a \right] da \middle| \mathcal{H}_{t-1} \right]$$

Since by Condition 12, $\frac{\pi_t(a, X_t, \mathcal{H}_{t-1})}{\pi_t^{\text{sta}}(a, X_t)} \geq \frac{1}{\sqrt{\rho_{\max,T}}}$ and $\phi(X_t, A_t)^{\otimes 2} \succeq 0$,

$$\succeq \frac{1}{\sqrt{\rho_{\max,T}}} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}} \left[\phi(X_t, A_t)^{\otimes 2} | X_t, a \right] da \middle| \mathcal{H}_{t-1} \right].$$

Since π_t^{sta} are pre-specified and since by our i.i.d. potential outcomes assumption (Condition 1) X_t do not depend on $\mathcal{H}_{t-1}$,

$$= \frac{1}{\sqrt{\rho_{\max,T}}} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}} \left[\phi(X_t, A_t)^{\otimes 2} | X_t, a \right] da \right].$$

By law of iterated expectations,

$$= \frac{1}{\sqrt{\rho_{\max,T}}} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\phi(X_t, A_t)^{\otimes 2} \right].$$

The above result and Equation (33) implies that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2} \succeq O_{\mathcal{P} \in \mathbf{P}}(1) + \sqrt{\frac{T}{\rho_{\max,T}}} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\phi(X_t, A_t)^{\otimes 2} \right]. \tag{34}$$

By Condition 11, the minimum eigenvalue of $\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\phi(X_t, A_t)^{\otimes 2} \right]$ is bounded above some constant greater than zero for all $\mathcal{P} \in \mathbf{P}$. By Condition 12, $\sqrt{\frac{T}{\rho_{\max,T}}} \to \infty$. Thus by Equation (32) and Equation (34), we have that $\hat{\theta}_T^{\text{AW-LS}} \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$.

C Choice of Stabilizing Policy

C.1 Optimal Stabilizing Policy in Multi-Arm Bandit Setting

Here we consider the multi-armed bandit setting where $\mathbb{E}_{\mathcal{P}}[Y_t(a)] = \theta_a^*(\mathcal{P})$ and $\text{Var}_{\mathcal{P}}(Y_t(a)) = \sigma^2$. We consider the adaptively-weighted least-squares estimator where $m_\theta(Y_t, A_t) = -\mathbb{1}_{A_t=a}(Y_t - \theta_a^*(\mathcal{P}))^2$. By Theorem 1, we have that

$$\left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\mathbb{1}_{A_t=a}(Y_t - \theta_a^*(\mathcal{P}))^2 \right] \right)^{-1/2} \left(\frac{1}{T} \sum_{t=1}^{T} W_t \mathbb{1}_{A_t=a} \right) \sqrt{T}(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0,1).$$

While the asymptotic variance of $\sqrt{T}(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P}))$ does not necessarily concentrate we can examine the following:

$$\left(\frac{1}{T} \sum_{t=1}^{T} W_t \mathbb{1}_{A_t=a} \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}} \left[\mathbb{1}_{A_t=a}(Y_t - \theta_a^*(\mathcal{P}))^2 \right] \right) \left(\frac{1}{T} \sum_{t=1}^{T} W_t \mathbb{1}_{A_t=a} \right)^{-1}$$

By Lemma 1, we have that $\frac{1}{T} \sum_{t=1}^{T} W_t \mathbb{1}_{A_t=a} - \sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t, \mathcal{H}_{t-1})} \xrightarrow{P} 0$. Thus we have

$$= \left(\frac{1}{T} \sum_{t=1}^{T} \pi_t^{\text{sta}}(a)\sigma^2 \right) \left(o_p(1) + \frac{1}{T} \sum_{t=1}^{T} \sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t, \mathcal{H}_{t-1})} \right)^{-2}.$$

As long as $\pi_t^{\text{sta}}(a), \pi_t(A_t, \mathcal{H}_{t-1})$ are bounded away from zero w.p. 1, the $o_p(1)$ term is asymptotically negligible and we can just consider $\left(\frac{1}{T} \sum_{t=1}^{T} \pi_t^{\text{sta}}(a)\sigma^2 \right) \left(\frac{1}{T} \sum_{t=1}^{T} \sqrt{\pi_t^{\text{sta}}(a)\pi_t(A_t, \mathcal{H}_{t-1})} \right)^{-2}$.

By Cauchy-Schwartz inequality,
$\left(\frac{1}{T} \sum_{t=1}^{T} \sqrt{\pi_t^{\text{sta}}(a)\pi_t(a, \mathcal{H}_{t-1})} \right)^2 \leq \left(\frac{1}{T} \sum_{t=1}^{T} \pi_t^{\text{sta}}(a) \right) \left(\frac{1}{T} \sum_{t=1}^{T} \pi_t(a, \mathcal{H}_{t-1}) \right)$.

Thus, $\frac{1}{\frac{1}{T} \sum_{t=1}^{T} \pi_t(a, \mathcal{H}_{t-1})} \leq \frac{\frac{1}{T} \sum_{t=1}^{T} \pi_t^{\text{sta}}(a)}{\left(\frac{1}{T} \sum_{t=1}^{T} \sqrt{\pi_t^{\text{sta}}(a)\pi_t(a, \mathcal{H}_{t-1})} \right)^2}$, so

$$\frac{\frac{1}{T} \sum_{t=1}^{T} \pi_t^{\text{sta}}(a)}{\left(\frac{1}{T} \sum_{t=1}^{T} \sqrt{\pi_t(a, \mathcal{H}_{t-1})\pi_t^{\text{sta}}(a)} \right)^2} \geq \frac{1}{\frac{1}{T} \sum_{t=1}^{T} \pi_t(a, \mathcal{H}_{t-1})}.$$

Note that this lower bound is achieved when $\pi_t^{\text{sta}}(a) = \pi_t(a)$. However, since π_t is a function of $\mathcal{H}_{t-1}$ and stabilizing policies$\{\pi_t^{\text{sta}}\}_{t=1}^T$ are pre-specified, setting $\pi_t^{\text{sta}}(A_t) = \pi_{t,a}$ is generally an unfeasible choice. Thus we want to choose π_t^{sta} to be as close to π_t as possible, subject to the constraint that the stabilizing policies are pre-specified, i.e., not a function of the data $\{Y_t, X_t, A_t\}_{t \geq 1}$.

C.2 Approximating the Optimal Stabilizing Policy

One way to approximately choose the optimal evaluation policy is to select $\pi_t^{\text{sta}}(a,x) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$. Note that $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ depends on the $\mathcal{P}$, which is unknown. Thus it is natural to choose $\pi_t^{\text{sta}}(a,x)$ to be $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ weighted by a prior on $\mathcal{P}$. 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 6, we display the difference in mean squared error for the AW-LS estimator in a two-armed bandit setting for two different choices of evaluation policy: (1) the uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) the expected $\pi_t(a, \mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$. We can see in this setting that by setting $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$ 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 $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$ is equivalent to choosing the uniform evaluation policy. For example, a two-armed bandit with identical arms so under common bandit algorithms $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})] = 0.5$ for all $t \in [1\colon T]$, which will make the evaluation policy $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a, \mathcal{H}_{t-1})]$ equivalent to the uniform policy.

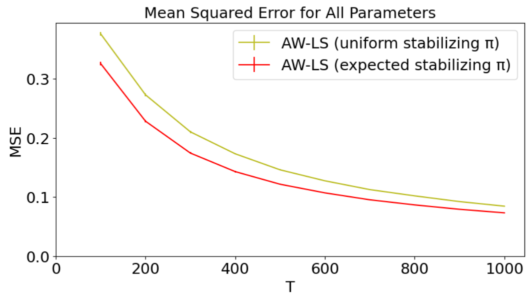

Figure 6: Above we plot the mean squared errors for the adaptively-weighted least squares estimator with evaluation policies: (1) uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) expected $\pi_t(a, \mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}\left[\pi_t(a)\right]$ (oracle quantity). In a two arm bandit setting we perform Thompson Sampling with standard normal priors, 0.01 clipping, $\theta^*(\mathcal{P}) = [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0, 1]$, standard normal errors, and $T = 1000$. Error bars denote standard errors computed over 5,000 Monte Carlo simulations.

D Need for Uniformly Valid Inference on Data Collected with Bandit Algorithms

Here we consider the two-armed bandit setting where $\mathbb{E}_{\mathcal{P}}[R_t(a)] = \theta_{0,a}(\mathcal{P})$, $\text{Var}_{\mathcal{P}}(R_t(a)) = \sigma^2$, and $\mathbb{E}_{\mathcal{P}}[R_t(a)^4] < c < \infty$ for $a \in \{0, 1\}$. The unweighted least squares estimator is asymptotically normal on adaptively collected data under the following condition of Lai and Wei [1982],

$$b_T \cdot \sum_{t=1}^{T} A_t \xrightarrow{P} 1. \tag{35}$$

Specifically, by Theorem 3 of Lai and Wei [1982], under (35),

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

However, as discussed in Deshpande et al. [2018] and Zhang et al. [2020], (35) can fail to to hold for common bandit algorithms when there is no unique optimal policy, i.e., when $\theta_0^*(\mathcal{P}) - \theta_1^*(\mathcal{P}) = 0$. For example, in Figure 7 we plot $\frac{1}{T}\sum_{t=1}^{T} A_t$ for Thompson Sampling and ϵ-greedy for a bandit with two identical arms.

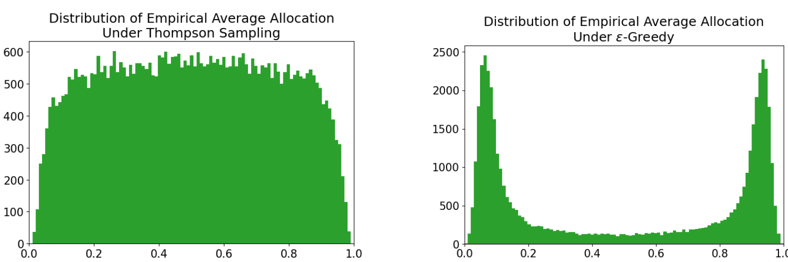

Figure 7: Above we plot empirical allocations, $\frac{1}{T}\sum_{t=1}^{T} A_t$, under both Thompson Sampling (standard normal priors, 0.01 clipping) and ϵ-greedy ($\epsilon = 0.1$) under zero margin $\theta_0^*(\mathcal{P}) = \theta_1^*(\mathcal{P}) = 0$. For our simulations $T = 100$, errors are standard normal, and we use $50k$ Monte Carlo repetitions.

In order to construct reliable confidence intervals using asymptotic approximations, it is crucial that that estimators 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 $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) = 0$. 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:

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

Under standard Thompson Sampling arm one is chosen according to the posterior probability that is optimal, so $\pi_t(1, \mathcal{H}_{t-1}) = \mathbb{P}(\tilde{\theta}_1 > \tilde{\theta}_0 | \mathcal{H}_{t-1})$. Above, $\mu_{a,t}$ denotes the posterior mean for the mean reward for arm 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

$$\frac{1}{T}\sum_{t=1}^{T} A_t \xrightarrow{P} \begin{cases} 1 - \pi_{\min} & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) > 0 \\ \pi_{\min} & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) < 0 \\ 0.5 & \text{if } \theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P}) = 0 \end{cases} \tag{36}$$

By equation (36), we satisfy (35) pointwise for every fixed $\mathcal{P}$ and we have that the OLS estimator is asymptotically normal *pointwise* [Lai and Wei, 1982]. However, equation (36) fails to hold uniformly over $\mathcal{P} \in \mathbf{P}$. Specifically, it fails to hold for any sequence of $\{\mathcal{P}_t\}_{t=1}^{\infty}$ such that $\theta_1^*(\mathcal{P}_t) - \theta_0^*(\mathcal{P}_t) = t^{-4}$. In Figure 8, 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 $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$.

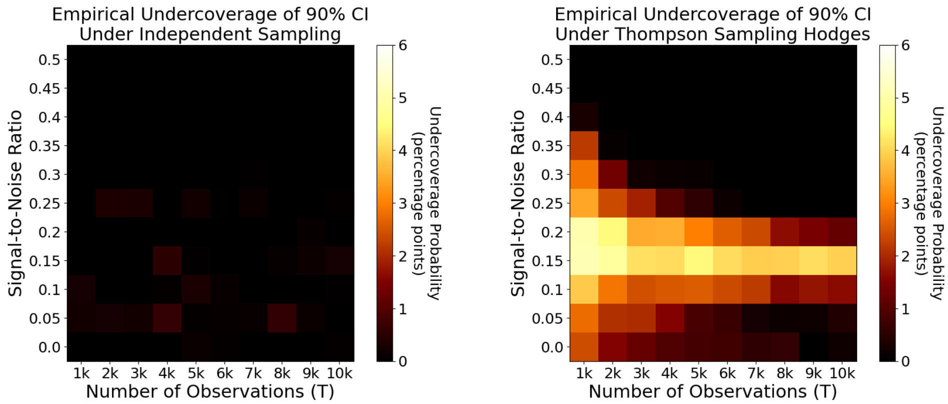

Figure 8: Above we construct confidence intervals for $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$ using a normal approximation for the OLS estimator. We compare independent sampling ($\pi_t = 0.5$) and TS Hodges, both with standard normal priors, 0.01 clipping, standard normal errors, and $T = 10,000$. We vary the value of $\theta_1^*(\mathcal{P}) - \theta_0^*(\mathcal{P})$ in the simulations to demonstrate the non-uniformity of the confidence intervals.

E Discussion of Chen et al. [2020]

Here we show formally that Theorem 3.1 in Chen et al. [2020], which proves that the OLS estimator is asymptotically normal on data collected with an ϵ-greedy algorithm, does not cover the case in which there is no unique optimal policy.

They assume that for rewards R_t, context vectors X_t, and binary actions $A_t \in \{0, 1\}$,

$$\mathbb{E}[R_t|X_t, A_t] = A_t \mathbf{X}_t^\top \boldsymbol{\beta}_1 + (1 - A_t)\mathbf{X}_t^\top \boldsymbol{\beta}_0.$$

They define $\boldsymbol{\beta} := \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$.

Specifically at part 1(b) of their proof on page 4 of the supplementary material, they claim that $g(\hat{\boldsymbol{\beta}}_t, \epsilon) \xrightarrow{P} g(\boldsymbol{\beta}, \epsilon)$, where $\hat{\boldsymbol{\beta}}_t$ is the OLS estimator for $\boldsymbol{\beta} := \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ and g is defined as follows:

$$g(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \epsilon) = \frac{\epsilon}{2} \int \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} d\mathcal{P}_x + (1 - \epsilon) \int \mathbb{1}_{\boldsymbol{\beta}^\top \mathbf{x} \geq 0} \mathbf{v}^\top \mathbf{x}\mathbf{x}^\top \mathbf{v} d\mathcal{P}_x$$

Above $v \in \mathbb{R}^d$ is arbitrary fixed vector and $x \in \mathbb{R}^d$ are the context vectors. $\mathcal{P}_x$ is the distribution of the context vectors $\boldsymbol{X}_t$.

Specifically, they claim that $g(\hat{\boldsymbol{\beta}}_t, \epsilon) \xrightarrow{P} g(\boldsymbol{\beta}, \epsilon)$ because $\hat{\boldsymbol{\beta}}_t \xrightarrow{P} \boldsymbol{\beta}$ (Corollary 3.1) and by continuous mapping theorem.

Recall the continuous mapping theorem for convergence in probability [Van der Vaart, 2000, Theorem 2.3]:

Theorem 4 (Continuous Mapping Theorem)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be continuous at every point of a set C such that $\mathbb{P}(X \in C) = 1$. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.*

Note that g is not continuous in $\boldsymbol{\beta}$ at the value $\boldsymbol{\beta} = \mathbf{0} \in \mathbb{R}^d$; this is due to the indicator term $\mathbb{1}_{\boldsymbol{\beta}^\top \mathbf{x} \geq 0}$. Thus, the standard continuous mapping theorem can not be applied in this setting. Note that the case that $0 = \boldsymbol{\beta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$, is exactly when there is no unique optimal policy. This means that Theorem 3.1 in Chen et al. [2020] does not cover the setting in which there is no unique optimal policy.