# Adaptive Off-Policy Inference for M-Estimators Under Model Misspecification

**Anonymous authors**
**Paper under double-blind review**

## Abstract

When data are collected adaptively, such as in bandit algorithms, classical statistical approaches such as ordinary least squares and $M$-estimation will often fail to achieve asymptotic normality. Although recent lines of work have modified the classical approaches to ensure valid inference on adaptively collected data, most of these works assume that the model is correctly specified. The misspecified setting poses unique challenges because the parameter of interest itself may not be well-defined over a non-stationary distribution of rewards. We therefore tackle the problem of *off-policy* inference in adaptive settings, where we uniquely define a projected solution over a stationary evaluation policy. Our method provides valid inference for $M$-estimators that use adaptively collected bandit data with a possibly misspecified working model. A key ingredient in our approach is the use of flexible approaches to stabilize the variance induced by adaptive data collection. A major novelty is that the procedure enables the construction of valid confidence sets even in settings where treatment policies are unstable and non-converging, such as when there is no unique optimal arm and standard bandit algorithms are used. Empirical results on semi-synthetic datasets constructed from the Osteoarthritis Initiative demonstrate that the method maintains type I error control, while existing methods for inference in adaptive settings do not cover in the misspecified case.

## 1 Introduction

Adaptive data collection has become a common practice in modern data analysis due to the rise of reinforcement learning and sequential data collections in contexts such as personalized healthcare (Yom-Tov et al., 2017), web recommendations (Afsar et al., 2022), and clinical trials (Zhao et al., 2009; Liu et al., 2020). In contrast to classical settings, where i.i.d. data are observed, adaptively collected data allows an analyst to interact with the decision-making algorithm in various ways, leading to sequentially dependent data. A canonical example is the (contextual) bandit problem where an analyst is allowed to make a choice of treatment (or "arm") at each time step and then observes a reward. Bandit algorithms attempt to maximize the total expected reward over time. These algorithms tend to result in data collection policies that explore treatment options in the early rounds and tend to shift to a greedier strategy that chooses the treatment with the highest estimated reward in the later rounds. When inference is conducted on data collected by bandit algorithms, the greedy search process can result in unstable variances (Deshpande et al., 2018; Zhang et al., 2020) and biased estimates of the average reward in each treatment arm (Shin et al., 2019) that make naive approaches to inference invalid.

There is a rich literature for adapting to the difficulties that arise for inference on adaptively collected data. Typical approaches focus on limiting the variance in each data collection using techniques such as propensity score truncation (Cook et al., 2024) or post-hoc reweighting of the data when constructing an estimator (Bibaut et al., 2021; Syrgkanis & Zhan, 2024). These approaches allow for assumption-lean inference in semiparametric settings, but they limit the study to relatively simple functionals such as the average reward for a fixed treatment or contrasts between average rewards for fixed arms.

An alternative set of approaches assumes a linear model linking treatment ($X_t \in \mathbb{R}^d$) and rewards ($Y_t \in \mathbb{R}$),

$$Y_t = \theta^T X_t + \epsilon_t,$$

where $\epsilon_t$ is i.i.d. random noise. It is striking that even in this simple setting, estimators such as ordinary least squares fail to achieve asymptotic normality when $X_t$ depends on previous rounds $\{X_i, Y_i\}_{i=1}^{t-1}$. A classical result (Lai & Wei, 1982) demonstrates that a necessary condition for the OLS estimate of $\theta$ to be asymptotically normal is when there exists a deterministic sequence of positive definite matrices $\{B_T\}_{T=1}^{\infty}$ such that $B_T^{-1} \sum_{t=1}^{T} X_t X_t^T \xrightarrow{p} I_d$. More recent literature (Zhang et al., 2020; Deshpande et al., 2018; Khamaru et al., 2025) makes the point that this assumption typically does not apply for common bandit algorithms when the margin (that is, the difference in mean rewards between arms) is zero. Zhang et al. (2020) proposes a solution through batching, by fixing the sampling rule for each batch and letting the number of observations within each batch go to infinity. This means that even if the sampling rule does not concentrate *across time*, it is fixed within each batch, allowing the empirical covariance matrix within each batch to concentrate. An alternative set of approaches (Deshpande et al., 2018; Khamaru et al., 2025) uses an online debiasing approach, where the finite sample bias is controlled through regularization. Other approaches generalize this this to the generalized partial linear model Lin et al. (2025) and for estimating equations (Ying et al., 2023).

We note that across all of these approaches, *correct specification* of the working model is crucial to ensuring valid inference. Correct model specification is an unlikely assumption to hold in practical settings. In the non-adaptive setting, it is common to instead conduct inference on a projected solution,

$$\theta^\star := \operatorname*{argmin}_{\theta} \mathbb{E}\left[\left(Y - \theta^T X\right)^2\right] \tag{1}$$

It is well known that under mild regularity conditions, the ordinary least squares estimate with sandwich estimators of the variance will cover the projected solution (White, 1982), but no analogous result exists in the adaptive setting. Moving beyond the linear case, Zhang et al. (2021) considers statistical inference on $M$-estimators in the contextual bandit problem, allowing for more complex models to describe the reward structure, such as generalized linear models. However, this work again requires that the conditional mean of the model is correctly specified in the working model.

Beyond its technical convenience, the assumption of a correctly specified working model has an important conceptual implication: it induces a target estimand that is invariant to the distribution of actions under which the data are collected or evaluted. Under model misspecification, however, the target parameter is generally defined through a projection, and projections depend on how different regions of the covariate and action space are weighted. Consequently, defining a meaningful projection target becomes a nontrivial aspect of the problem when the distribution of actions evolves over time and may not converge. Our approach is to define the estimand with respect to a fixed evaluation policy, which yields an off-policy projection parameter as an inferential target that remains well defined under persistent adaptivity; formal details are deferred to Section 2. For example, in an online platform or information-technology setting, a company may already have a deployed policy and wish to run adaptive experiments to determine whether improved policies exist relative to this status quo. In this case, the evaluation policy can be taken to be the existing deployed policy, so that the resulting projection parameter is interpreted relative to the operational regime currently used in practice.

An alternative approach (Zhang et al., 2023) performs inference for $Z$-estimators and allows for a misspecified model, but only under a finite amount of adaptivity. In this framework, $n$ individuals are tracked over $T$ time periods, with $T$ fixed and $n$ growing to $\infty$. The advantage of this approach is that it allows for non-stationary reward and context distributions over time, but it relies on the number of individuals enrolled in the trial to tend to $\infty$ and does not allow adaptive decisions to be made separately for each individual in the trial.

### 1.1 Our Contributions

In this work, we propose a methodology for conductive inference for finite-dimensional parameters in $M$-estimation problems under a potentially misspecified working model and non-vanishing levels of adaptivity

(i.e. without assuming the adaptivity eventually vanishes). The main result, Theorem 1, provides a Central Limit Theorem that enables the construction of confidence intervals when the variance (which varies over time and may not converge) of the score function can be estimated consistently.

The problem of estimating the variance of the score function is nontrivial because using naive empirical estimates when applying Theorem 1 is only valid when the variance is a fixed deterministic quantity. In many adaptive settings, such as bandit problems where expected rewards between arms are comparable, action selection probabilities will be non-convergent random quantities, leading to unstable variances (Zhang et al., 2020). To solve this problem, we propose methods to construct valid time-varying plug-in estimates for the variances at each time step using machine learning techniques.

In a work that is concurrent to ours, Guo & Xu (2025) also provides a procedure for inference on $Z$-estimators trained on adaptively collected data via inverse propensity weighting, but only under the assumption that treatment policies converge as $T \to \infty$. Since many common bandit algorithms fail to converge under model misspecification, the authors provide a sufficient set of conditions to guarantee policy convergence (e.g., continuous policies that are sufficiently smooth). We take a different approach — rather than restricting ourselves to only particular classes of policy that are guaranteed to converge, we allow for potentially non-convergent policies but stabilize the estimator post hoc by estimating the conditional variance *separately* at each time step. In cases where the policy converges, these machine-learning-based approaches can be replaced with simple empirical estimates of the variance, and our procedure can be simplified considerably.

We note that before now, the problem of estimating projection parameters under non-finite amounts of adaptivity was not solved even for simple linear models. Of course, this is merely a special case of $M$-estimation so our proposed methodology can be applied straightforwardly. We discuss this particular application as a running example throughout the remainder of the paper.

Our primary results are written under the assumption that the number of time steps tends to infinity, with only a single observation at each time step. However, this can be easily extended to the batched setting, where multiple observations are observed at each time step.

### 1.2 Paper Outline

In Section 2, we define the problem, most crucially introducing definitions of target parameters inspired by the off-policy evaluation literature that remain tractable in the misspecified setting. In Section 3, we present a CLT that enables inference for $M$-estimators in the adaptive setting given accurate (time-varying) estimates of the variance of the score function. Section 4 discusses practical strategies for estimating variance using flexible machine learning approaches. Section 5 presents empirical results on semi-synthetic datasets constructed from the Osteoarthritis Initiative, a publicly available longitudinal dataset provided by the NIH which tracks health outcomes of patients with osteoarthritis. The results verify the validity of the procedure and the failure of existing approaches to provide confidence sets with valid coverage. We provide concluding thoughts in Section 6, including potential limitations of this approach and possible directions for future work. All proofs for stated theorems are included in the Appendix.

### 1.3 Notation

We introduce some shorthand notation used throughout the paper. We define $[n] := \{1, ..., n\}$ for a positive integer $n$. We denote $e_j$ as the $j$-th standard basis vector in $\mathbb{R}^d$. For a function $f_\theta : \mathbb{R}^d \to \mathbb{R}$, we refer to $\dot{f}_\theta$ as the first derivative of $f$ with respect to $\theta$, $\ddot{f}_\theta$ as the Hessian matrix with respect to $\theta$, $\dddot{f}_\theta$ as the third derivative with respect to $\theta$, and so on. For $M \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, we write $\|M\|_1 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sum_{k=1}^{d_3} |a_{i,j,k}|$. For two matrices $A, B$, $A \succeq B$ means that $A - B$ is positive semidefinite.

When taking expectations, $\mathbb{E}_{\mathcal{P},\pi}$ denotes the expectation under the assumption that $(X_t, A_t, Y_t) \sim p(y|x,a)\,\pi(a|x)p(x)$. When no subscript is denoted, the expectation is over the entire joint distribution of $\{X_t, A_t, Y_t\}_{t=1}^T$.

## 2 Problem Setup

We assume a stochastic bandit environment in which features $X_t \in \mathbb{R}$ are observed in each round $t \in [T]$. After observing a feature, an analyst chooses an action $A_t$ from a finite set of $K$ actions $\mathcal{A} := \{1, ..., K\}$. After selecting an action, an outcome $Y_t$ is observed. We denote $Y_t(a)$, where $a \in \mathcal{A}$, to be the counterfactual result had the analyst chosen $a$ in round $t$, regardless of the actual action taken. We assume that the joint distribution of features and potential outcomes is independent and identically distributed across time.

**Assumption 1.** $\{(X_t, Y_t(1), \ldots, Y_t(K))\}_{t=1}^{T} \overset{\text{iid}}{\sim} \mathcal{P}$ .

Although potential outcomes are distributed i.i.d., we allow the analyst to choose the treatment adaptively based on a pooled history $\mathcal{H}_{t-1} := \{(X_i, A_i, Y_i)\}_{i=1}^{t-1}$. Formally, we say $\pi_t(a|X_t) := \mathbb{P}(A_t = a|X_t, \mathcal{H}_{t-1}) \in \sigma(\mathcal{H}_{t-1})$ for all $a \in \mathcal{A}$. We can summarize the conditional density (given $\mathcal{H}_{t-1}$) over actions, features, and outcomes at time step $t$ as

$$(X_t, A_t, Y_t) \sim p\,(y|x, a)\,\pi_t(a|x)p(x), \tag{2}$$

where densities $p(x)$ and $p(y|x, a)$ are invariant over time and defined by Assumption 1, but $\pi_t(a|x)$ is adaptive over time, because the analyst has the option to change the assignment probabilities in reaction to $\mathcal{H}_{t-1}$.

Our goal is to construct confidence regions for a parameter $\theta^\star$, which we define as the expected maximizer of some function $m_\theta$ in some parameter space $\Theta$. For example, $m_\theta(x, a, y) := -(y - \theta^T(x, a))^2$ corresponds to an ordinary least squares regression, and $m_\theta(x, a, y) := -y(x, a)^T\theta + \psi((x, a)^T\theta)$, where $\psi$ denotes the convex log-partition function, corresponds more generally to GLMs (Agresti, 2015). Some nuance is in order to determine the distribution over which the expected loss is minimized. Because the distribution of actions evolves over time, the distribution at each time step may be substantially different so the solution to $\text{argmax}_{\theta\in\Theta}\,\mathbb{E}\,[m_\theta(X_t, A_t, Y_t)]$ will also vary substantially across $t \in [T]$. In order to ensure there is a unique and stable inferential target, we tackle the problem of *off-policy learning* and seek inference under a hypothetical policy $\pi_e(a|x)$ that is distinct from the policy used during treatment and invariant over time.

$$\theta^\star := \underset{\theta\in\Theta}{\text{argmax}}\,\mathbb{E}_{\mathcal{P},A\sim\pi_e}\,[m_\theta(X, A, Y)]\,. \tag{3}$$

**Choice of evaluation policy** The evaluation policy $\pi_e(a \mid x)$ defines the distribution under which the target parameter $\theta^\star$ is projected. Under model misspecification, different evaluation policies may induce substantially different projection targets, since the model is approximating the underlying response surface over different regions of the action space. This phenomenon is not specific to our framework, but rather reflects a generic feature of misspecified projection parameters.

The evaluation policy may either be fixed and known a priori, or estimated after data collection. For example, one might choose a uniform policy satisfying $\pi_e(A_t = a \mid X_t) = \frac{1}{|\mathcal{A}|}$ for all $a \in \mathcal{A}$ and $t \in [T]$, corresponding to a hypothetical regime in which treatments are assigned uniformly across the population. Alternatively, $\pi_e$ may correspond to a deployment policy of practical interest, such as a target policy in off-policy evaluation, a hypothetical optimal policy, or a limiting treatment assignment rule such as $\lim_{t\to\infty}\mathbb{P}(A_t = a \mid X_t, \mathcal{H}_{t-1})$, which we do not assume exists in general.

When the target policy is estimated after data collection, an appeal to Slutsky's theorem allows estimated versions of $\pi_e$ to be substituted into the estimator, provided the resulting estimates are asymptotically independent of the adaptive history. In practice, the choice of $\pi_e$ should reflect the scientific or operational regime under which the model will ultimately be interpreted or deployed. In the absence of a natural deployment policy, a uniform policy over the action space is often a reasonable default, since it avoids concentrating the projection on only a restricted subset of treatments. The following example illustrates how the target parameter can vary across evaluation policies under model misspecification.

**Example 1** (Effect of $\pi_e$ on $\theta^\star$)**.** *Consider data generated as* $Y_t \sim N(6A_t^2, 1)$ *under two treatment regimes, where* $\mathcal{A} = \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. *The first policy is uniform over* $A_t \in$

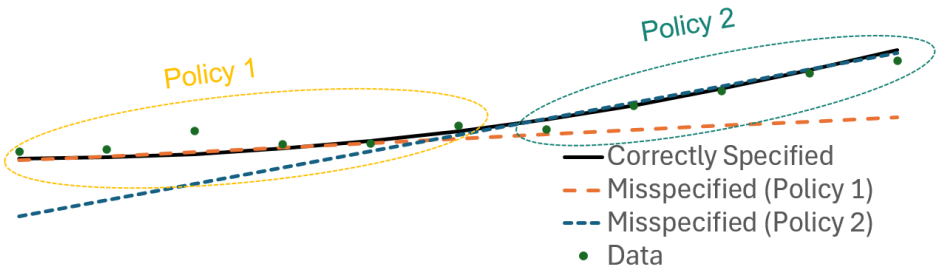



Figure 1: Illustration for Example 1. Under the misspecified linear model $m_\theta = -(Y_t - \theta_0 - \theta_1 A_t)^2$, $(\theta_0^\star, \theta_1^\star) = (-0.2, 4)$ for the first policy and $(\theta_0^\star, \theta_1^\star) = (-5.3, 15)$ for the second policy. Intuitively, the misspecified model will try to estimate a secant line at different points of the quadratic function, resulting in very different interpretations for the target parameter. As such, the choice of evaluation policy is critical for properly interpreting the target parameter.

$\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. *The second policy is uniform over* $A_t \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$. *Under the correctly specified model* $m_\theta(A_t, Y_t) := -(Y_t - \theta_0 - \theta_1 A_t^2)^2$, $(\theta_0^\star, \theta_1^\star) = (0, 6)$ *for both policies. On the other hand, under an incorrectly specified linear model,* $\theta_0^\star$ *and* $\theta_1^\star$ *can vary markedly as seen in Figure 1.*

In some special cases, however, the target parameter becomes invariant to the choice of evaluation policy. Consider the average treatment effect as a prototypical example.

**Remark 1** (Average Treatment Effect)**.** *We can use this framework to target the average treatment effect by using a vector of indicator variables as the chosen model with a square loss function. We can permit any evaluation policy that satisfies the standard causal assumption of* $Y_t(a) \perp\!\!\!\perp A_t$ *under* $\pi_e$*, and such that* $\pi_e(A_t = a) > 0$.

*Formally, we let* $\theta = (\theta_1, ..., \theta_{|\mathcal{A}|})$ *and consider setting* $m_\theta = -(Y_t - \sum_{a=1}^K 1_{A_t=a}\theta_a)^2$. *First order conditions imply that for all* $t$*,*

$$0 = \mathbb{E}_{\mathcal{P},\pi_e}\left[Y_t - \sum_{a=1}^{K} 1_{A_t=a}\theta_a^\star\right].$$

*Noting that* $Y_t = Y_t(a)1_{A_t=a}$ *and rearranging terms yields,*

$$\theta_a^\star = \frac{\mathbb{E}_{\mathcal{P},\pi_e}\left[Y_t(a)1_{A_t=a}\right]}{\mathbb{E}_{\mathcal{P},\pi_e}\left[1_{A_t=a}\right]} = \frac{\mathbb{E}_{\mathcal{P}}\left[Y_t(a)\right]\mathbb{E}_{\mathcal{P},\pi_e}\left[1_{A_t=a}\right]}{\mathbb{E}_{\mathcal{P},\pi_e}\left[1_{A_t=a}\right]} = \mathbb{E}_{\mathcal{P}}\left[Y_t(a)\right] = \mathbb{E}_{\mathcal{P}}\left[Y(a)\right].$$

*To infer the average effect of treatment* $j$ *versus treatment* $k$*, we can then target* $\eta^T\theta^\star$ *for some contrast vector* $\eta := e_j - e_k$ *(recall* $e_j$ *refers to the* $j$*-th standard basis vector in* $\mathbb{R}^K$*). For example* $\eta = (-1, 1)$ *in the setting of* $K = 2$ *would recover the average treatment effect* $\mathbb{E}\left[Y(1) - Y(0)\right]$.

Although the average treatment effect is well studied in contextual bandit problems (Hadad et al., 2021a; Bibaut et al., 2021; Cook et al., 2024), similar invariance properties can be obtained in richer parametric models by including sufficiently flexible interaction terms with the action variable so that the relevant conditional response surface is correctly represented. In such settings, the choice of evaluation policy no longer changes the interpretation of the target parameter, but instead primarily affects the statistical efficiency of the resulting estimator.

**Choice of estimator** Similar to Zhang et al. (2021), one option is to consider estimators of the form

$$\hat{\theta}_0 := \operatorname*{argmax}_\theta \sum_{t=1}^{T} w_t m_\theta(X_t, A_t, Y_t),$$

where $w_t \in \sigma(\mathcal{H}_{t-1}, X_t)$. Note that when the model is correctly specified, we have at every $t$ that

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}\left[m_\theta(X_t, A_t, Y_t)|A_t, X_t\right] \text{ for all } A_t \in \mathcal{A}, X_t \in \mathbb{R}. \tag{4}$$

When this is true conditionally, it will also be true marginally for *any choice* of policy. Therefore, under no model misspecification, we have at every time step $t$ that

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{P},\pi_e}\left[m_\theta(X_t, A_t, Y_t)\right] = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{P},\pi_t}\left[m_\theta(X_t, A_t, Y_t)|\mathcal{H}_{t-1}\right],$$

removing the need to consider the choice of *policy* when defining the target parameter. Assuming that the maximum occurs at a critical point of $m_\theta$, this implies that

$$0 = \mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)|X_t, A_t\right] = \mathbb{E}\left[w_t \dot{m}_{\theta^\star}(X_t, A_t, Y_t)|X_t, A_t, \mathcal{H}_{t-1}\right],$$

for all $X_t, A_t$ and all $w_t \in \sigma(X_t, A_t, \mathcal{H}_{t-1})$. Thus, $w_t$ can be arbitrarily chosen based on $X_t$, $A_t$, and $\mathcal{H}_{t-1}$ without violating the first-order conditions. For this reason, previous work often uses $w_t$ as a free parameter to control the variance without having to worry that this type of re-weighting will change the target away from $\theta^\star$. As an example, Zhang et al. (2021) chooses $w_t = \sqrt{\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})}}$, which allows the variance of the score function to converge to a value independent of history. Other examples include online debiasing approaches (Deshpande et al., 2018; Khamaru et al., 2025) which use regularization to choose $w_t$ so that the variance of $\hat{\theta}_T$ is controlled.

Under model misspecification, however, the choice of weights has an impact on the target parameter that is being covered because at $\theta^\star$ we will require that at each time step $t$,

$$0 = \mathbb{E}_{\mathcal{P},\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\right] = \mathbb{E}_{\mathcal{P},\pi_t}\left[w_t \mathbb{E}_{\mathcal{P}}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)|X_t, A_t, \mathcal{H}_{t-1}\right]|\mathcal{H}_{t-1}\right],$$

which is *only* true for the specific choice of $w_t = \frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})}$. Thus, we are constrained to define an estimator such as

$$\hat{\theta}_0 = \underset{\theta}{\operatorname{argmax}} \sum_{t=1}^{T} \frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})} m_\theta(X_t, A_t, Y_t). \tag{5}$$

Using inverse propensity weights ensures that $\hat{\theta}_0$ will be unbiased for $\theta^\star$, but we must now consider alternative methods to stabilize the variance of the estimator, which we will discuss in the next section.

**MAIPWM-Estimator** As an alternative to Equation (5), we can potentially improve power through the use of a predictive model that we update over time, $f_t : \mathbb{R} \times \mathcal{A} \to \mathbb{R}$. $f_t$ is trained on $\mathcal{H}_{t-1}$ (along with potentially other information known to the experimenter a priori before the experiment). $f_t$ then takes in $(X_t, A_t)$ as an input and outputs an estimate for $\mathbb{E}\left[Y_t|X_t, A_t\right]$. An alternative definition to Equation (5) is what we will term the **misspecified augmented inverse propensity weighted $M$-estimator (MAIPWM-Estimator)**

$$\tilde{\theta}_T = \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{a=1}^{K} \pi_e(A_t = a|X_t) \left( m_\theta(a, X_t, f_t(a, X_t)) + \mathbb{1}_{A_t=a} \frac{m_\theta(A_t, X_t, Y_t) - m_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)} \right). \tag{6}$$

The estimator improves power when the model $f_t$ is a good estimator of $\mathbb{E}[Y_t|X_t, A_t]$ by augmenting observations with predictions from a model. This approach is routinely used when estimating simple functionals like the average treatment effect in causal inference (Chernozhukov et al., 2018) and off-policy evaluation (Hadad et al., 2021b) and is referred to as the *augmented inverse probability weighted* estimator in these settings. More recently, Zrnic & Candès (2024) introduces a similar modification of the score function of $M$-estimators with predictions derived from black box machine learning models, but limits their findings to the non-causal, non-adaptive observational setting.

A common approach to finding the maximum is to find the solution defined by the root of the *score equation*,

$$s_{t,\theta} = \sum_{a=1}^{K} \pi_e(A_t = a|X_t) \left( \dot{m}_\theta(X_t, a, f_t(X_t, a)) + \mathbb{1}_{A_t=a} \frac{(\dot{m}_\theta(X_t, a, Y_t) - \dot{m}_\theta(X_t, a, f_t(X_t, a))}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} \right). \tag{7}$$

In order to limit the variance of this quantity, we will instead consider a second estimator $\hat{\theta}_T$ defined as the (approximate) root of $\sum_{t=1}^{T} \Sigma_t^{-1} s_{t,\theta}$ for specially constructed matrices $\Sigma_t \in \mathbb{R}^d$ designed to normalize the variance of $s_{t,\theta}$. However, the construction of $\Sigma_t$ will require consistent estimates of $\theta^\star$ so our procedure will involve a two-step process where $\tilde{\theta}_T$ is used to construct the stabilizing matrices $\Sigma_t$ and then a central limit theorem is proved for $\hat{\theta}_T$.

This approach will be most useful when an individual has access to powerful machine learning methods for prediction but wants to perform inference on a simpler parametric model that is more interpretable. If an analyst prefers not to manage a predictive model during the course of the experiments, substituting any constant for $f_t$ will recover Equation (5). Therefore, we will prove our main results with respect to $\tilde{\theta}_T$ and $\hat{\theta}_T$, but similar guarantees will exist for $\hat{\theta}_0$ when no predictive model is being updated as a special case.

## 3 Main Results

Our approach will be to construct two sets of weights based on different filtrations. As before, $w_t \in \sigma(X_t, \mathcal{H}_{t-1})$ is restricted to being the inverse propensity weights $\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1})}$ to ensure that $\hat{\theta}_T$ is unbiased for $\theta^\star$. Similarly, we consider $\Sigma_t \in \sigma(\mathcal{H}_{t-1})$ that will allow us to adjust the second moment of $s_{t,\theta}$, while preserving the first moment of the score function in $\theta^\star$ to be 0. By assumption, $\mathbb{E}\left[s_{t,\theta^\star}|\mathcal{H}_{t-1}\right] = 0$. Define a martingale difference sequence (MDS) $Z_t := \Sigma_t^{-1/2} s_{t,\theta^\star}$. Then,

$$\mathbb{E}[Z_t|\mathcal{H}_{t-1}] = \mathbb{E}[\Sigma_t^{-1/2} s_{t,\theta^\star}|\mathcal{H}_{t-1}] = \Sigma_t^{-1/2}\mathbb{E}[s_{t,\theta^\star}|\mathcal{H}_{t-1}] = 0,$$

by the fact that $\Sigma_t^{-1/2} \in \sigma(\mathcal{H}_{t-1})$. A natural choice is to choose $\Sigma_t$ as

$$V_{t,\theta^\star} = \mathbb{E}_{\mathcal{P},\pi_t}\left[s_{t,\theta^\star} s_{t,\theta^\star}^T|\mathcal{H}_{t-1}\right].$$

By re-weighting the score function to stabilize the variance but *requiring* that the weights be determined from a coarser filtration than what was used for constructing $w_t$, we are able to preserve the first-order conditions that will be stated in Assumption 3. Of course, $V_{t,\theta^\star}$ is unknown in practice and will need to be replaced with an estimate $\hat{V}_t$. We proceed with the remainder of this section assuming the existence of such an estimate but note that it is a non-trivial task as we only have a single observation corresponding to each $t$. The key idea for constructing an estimator is to leverage the potential outcomes framework of Assumption 1, which we discuss in Section 4.

Before stating the main result, we introduce several assumptions that will be required to ensure asymptotic normality.

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

The existence of the first two derivatives allows us to identify the quantity of interest as the expected maximizer of $m_\theta$ and estimate it by evaluating the critical points of this function, which we discuss in more detail in Assumption 3. The existence of a third derivative will allow us to use the Taylor expansion around the score function to form a confidence set covering $\theta^\star$.

**Assumption 3.** *Let $\theta^\star$ be defined as in Equation (3) and $\Theta \subset \mathbb{R}^d$ be a bounded parameter space such that Assumption 2 holds. We further assume that at every $t \in [T]$,*

1. $\mathbb{E}_{\mathcal{P},\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\right] = 0$,
2. $-\mathbb{E}_{\mathcal{P},\pi_e}\left[\ddot{m}_{\theta^\star}(X_t, A_t, Y_t)\right] \succeq H$ *for some positive definite matrix* $H$,
3. *For any $\epsilon > 0$, there exists constants $\delta_1, \delta_2 > 0$ such that*
   - $\inf_{\|\theta-\theta^\star\|\geq\epsilon} \left\{\mathbb{E}_{\pi_e}\left[m_{\theta^\star}(X_t, A_t, Y_t) - m_\theta(X_t, A_t, Y_t)\right]\right\} \geq \delta_1$,
   - $\inf_{\|\theta-\theta^\star\|\geq\epsilon} \left\|\mathbb{E}_{\pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)\right]\right\|_1 \geq \delta_2$.

4. *The first four moments of* $m_\theta(X_t, A_t, Y_t)$, $\dot{m}_\theta(X_t, A_t, Y_t)$, *and* $\ddot{m}_\theta(X_t, A_t, Y_t)$ *are bounded with respect to* $\mathcal{P}, \pi_e$.

The first two conditions of this assumption ensure that the quantity of interest $\theta^\star$ is the solution to a maximization problem and can be found by evaluating the critical points of $m_\theta$. The third condition mirrors classical assumptions to demonstrate the consistency of $Z$ estimators and $M$ estimators (van der Vaart, 2000). Having a well-separated solution ensures the uniqueness of $\theta^\star$ and allows estimators constructed from finite samples to converge appropriately. Note that we assume that $\theta^\star$ is well separated *both* as a maximizer of $m_\theta$ and as a root of the score equation because we consider estimators defined in both of these ways within our procedure. A special case of the third condition is when $m_\theta$ is a continuously differentiable concave function with the maximum obtained at $\theta^\star$. The fourth condition is used to ensure that Lindeberg conditions are achieved when invoking the martingale central limit theorem and is analogous to conditions in classical proofs of the normality of $M$-estimators (van der Vaart, 2000).

**Assumption 4** (Bounded importance ratios)**.** *There exist a constant* $C_1 > 0$ *such that* $\frac{1}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})} \leq C_1$ *for all* $a \in \mathcal{A}$.

Bounding the weights is important to ensure that the variance of the score function is bounded, so that the martingale law of large numbers and central limit theorems can be applied. It is possible to weaken this so that the $\mathbb{P}(A_t = a|X_t, \mathcal{H}_{t-1})$ are allowed to converge to either 0 or 1 at appropriately slow rates, but we leave this generalization to future work.

**Assumption 5** (Finite Bracketing Numbers)**.** *Define the set of functions* $\mathcal{M}_\Theta = \{m_\theta(X_t, A_t, Y_t) : \theta \in \Theta\}$ *and* $\dot{\mathcal{M}}_\Theta = \{c^T \dot{m}_\theta(X_t, A_t, Y_t) : \theta \in \Theta, \|c\| \leq 1\}$. *We assume that for all* $\epsilon > 0$, *the bracketing numbers* $N_{[]}(\epsilon, \mathcal{M}_\Theta, L_2(\mathcal{P}, \pi_e)) < \infty$ *and* $N_{[]}(\epsilon, \dot{\mathcal{M}}_\Theta, L_2(\mathcal{P}, \pi_e)) < \infty$.

Assumption 5 limits the complexity of the function class $\{m_\theta : \theta \in \Theta\}$ so that a martingale law of large numbers can be applied which is required when constructing an argument for the consistency of $\hat{\theta}_T$ and $\tilde{\theta}_T$. We note that Zhang et al. (2021) provides an intuitive sufficient condition to check to ensure that Assumption 5 is true. They show that whenever $m_\theta$ is Lipschitz in the sense that

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

for some function $h$ such that $\mathbb{E}_{\mathcal{P},\pi_e}[h(X_t, A_t, Y_t)^2] < m$ for some constant $m$, then $\mathcal{M}_\theta$ will have finite bracketing numbers for all $\epsilon > 0$. We place a similar assumption on the complexity of the model classes.

**Assumption 6** (Model Class Complexity)**.** *The predictive model* $f_t \in \sigma(\mathcal{H}_{t-1})$, *and* $f_t$ *belongs to a function class* $\mathcal{F}$. *For this function class, there exists a single function* $u_\mathcal{F}$ *and a constant* $m_u \in \mathbb{R}$ *such that* $\mathbb{E}[u_\mathcal{F}(X_t, a, f_t(a, X_t))^4] < m_u$ *for all* $a \in \mathcal{A}$ *and,*

$$\sup_{\theta\in\Theta} m_\theta(x, a, f_t(a, x)) \leq u_\mathcal{F}(x, a, f_t(a, x)),$$

$$\sup_{\theta\in\Theta} \|\dot{m}_\theta(x, a, f_t(a, x))\|_1 \leq u_\mathcal{F}(x, a, f_t(a, x)),$$

$$\sup_{\theta\in\Theta} \|\ddot{m}_\theta(x, a, f_t(a, x))\|_1 \leq u_\mathcal{F}(x, a, f_t(a, x)),$$

$$\sup_{\theta\in\Theta} \|\dddot{m}_\theta(x, a, f_t(a, x))\|_1 \leq u_\mathcal{F}(x, a, f_t(a, x)),$$

*for all* $a \in \mathcal{A}$, $x$ *in the support of* $\mathcal{P}$ *and* $f_t \in \mathcal{F}$. *Furthermore, we assume that there exists a constant* $C_2$ *not dependent on* $t$ *such thate*

$$\mathbb{E}\left[\|m_\theta(X_t, a, f_t(X_t, a)) - m_{\theta'}(X_t, a, f_t(X_t, a))\|_2^2 \,|\mathcal{H}_{t-1}\right] \leq C_2 \|\theta - \theta'\|_2^2$$

*almost surely.*

Note that a primary difference between Assumption 5 and Assumption 6 is that Assumption 6 is written in terms of expectations relative to the actual policy $\pi_t$ rather than $\pi_e$. This is necessary because Assumption 6 bounds the variance of a term related to $f_t$, which is dependent on history whereas Assumption 5 is bounding the complexity of a quantity that is *independent* of history. These assumptions allow us to prove a key lemma needed to ensure the consistency of $\hat{\theta}_T$ and $\tilde{\theta}_T$.

**Lemma 1.** *Assume $\mathcal{G}_\Theta = \{g_\theta(X_t, A_t, Y_t) : \theta \in \Theta\}$ is a class of functions such that for any $\epsilon > 0$, the bracketing number $N_{[]}(\epsilon, \mathcal{G}_\Theta, L_2(\mathcal{P}, \pi_e)) < \infty$. Define*

$$R_t(\theta) = \sum_{a=1}^{K} \pi_e(A_t = a|X_t) \left( g_\theta(a, X_t, f_t(a, X_t)) + \mathbb{1}_{A_t=a} \frac{g_\theta(X_t, A_t, Y_t) - g_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)} \right). \tag{8}$$

*Assume that there exists a constant $L$ not dependent on $t$ such that*

$$\mathbb{E}\left[ \|m_\theta(X_t, a, f_t(X_t, a)) - m_{\theta'}(X_t, a, f_t(X_t, a))\|_2^2 \,|\mathcal{H}_{t-1} \right] \leq L \|\theta - \theta'\|_2^2$$

*almost surely. Then under Assumptions 1-4,*

$$\sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \left( R_t(\theta) - \mathbb{E}[R_t(\theta)|\mathcal{H}_{t-1}] \right) \xrightarrow{p} 0.$$

We can think of Lemma 1 as a uniform version of the martingale weak law of large numbers, specifically adjusted for our setting where machine learning methods are used to augment data collection.

**Assumption 7.** *There exists a fixed, integrable function $u_m(X_t, A_t, Y_t)$ such that for some $\delta > 0$,*

$$\sup_{\theta \in \Theta: \|\theta - \theta^\star\| < \delta} \|\ddot{m}_\theta(X_t, A_t, Y_t)\|_1 \leq u_m(X_t, A_t, Y_t),$$

*and $\mathbb{E}_{\mathcal{P}, \pi_e}\left[u_m(X_t, A_t, Y_t)^2\right]$ is bounded.*

This assumption mirrors that of classical approaches to proving the normality of $M$-estimators (van der Vaart, 2000) and is used to make sure that certain quantities related to the third derivative of the objective function remain bounded when using Taylor expansion to construct the confidence set.

We are now ready to state the theorem.

**Theorem 1.** *Let $\theta^\star$ be defined as Equation (3) and $s_{t,\theta}$ defined as in Equation (7). Assume $V_{t,\theta^\star} := \mathbb{E}_{\mathcal{P}, \pi_t}\left[ s_{t,\theta^\star} s_{t,\theta^\star}^T | \mathcal{H}_{t-1} \right]$ is almost surely invertible and that there exists a sequence of random matrices $\{\hat{V}_t\}_{t=1}^T$ adapted to the filtration $\sigma(\mathcal{H}_{t-1})$ such that $\left\| \hat{V}_t^{-1/2} - V_{t,\theta^\star}^{-1/2} \right\|_{op} \xrightarrow{p} 0$. Assume that the eigenvalues of both $\hat{V}_t$ and $V_{t,\theta^\star}$ are bounded above and below by constants $\delta_{min}, \delta_{max}$. Let $\hat{\theta}_T$ be any estimator such that $\frac{1}{T}\sum_{t=1}^T \hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} = o_p(1/\sqrt{T})$. Then under Assumptions 1-6,*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \left( \hat{\theta}_T - \theta^\star \right) \xrightarrow{d} N(0, I_d).$$

Note that letting $\frac{1}{T}\sum_{t=1}^T \hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} = o_p(1/\sqrt{T})$ instead of exactly zero ensures that the results are still valid when using an approximate root finding algorithm instead of an exact root. As an illustrative example, we first demonstrate the application of Theorem 1 in the case of a linear model.

#### An illustration: correctly specified models

Let us first consider the case of correct specification, where $Y_t = Z_t^T \theta^* + \epsilon_t$ and $\epsilon_t$ are independent Gaussian noise with $\epsilon_t \sim N(0, \sigma_t^2)$. We assume $Z_t \in \sigma(\mathcal{H}_{t-1})$ is dependent on user choices based on a shared history

$\mathcal{H}_{t-1} := \{Z_1, Y_1, ..., Z_{t-1}, Y_{t-1}\}$. In our framework, we can think of $Z_t = \phi(A_t)$ as some deterministic function of the observed actions (e.g., one-hot encoding).

Letting $\Sigma_T := \sum_{t=1}^T Z_t Z_t^T$, $\lambda_{\min}(\Sigma_T)$ represent the minimum eigenvalue of $\Sigma_T$, and $\lambda_{\max}(\Sigma_T)$ represent the maximum eigenvalue of $\Sigma_T$, Lai & Wei (1982) show that a sufficient condition for the OLS estimator $\tilde{\theta}_T := \Sigma_T^{-1} \sum_{t=1}^T Z_t Y_t$ to converge to $\theta$ almost surely is that $\lambda_{\min}(\Sigma_T) \xrightarrow{\text{a.s.}} \infty$ and $\frac{\log \lambda_{\max}(\Sigma_T)}{\lambda_{\min}(\Sigma_T)} \xrightarrow{\text{a.s.}} 0$. However, the requirement for asymptotic normality is much stronger. They require that a deterministic series of matrices $\{B_T\}_{T=1}^\infty$ exist such that $B_T^{-1}(\sum_{t=1}^T Z_t Z_t^T)^{1/2} \xrightarrow{p} I_p$.

Suppose that we apply Theorem 1 in the case where $\tilde{\theta}_T$ is consistent but the asymptotic normality condition does not apply. We can still form a confidence interval for $\theta^*$ as follows. For simplicity, we assume that $f_t = c$ for some constant $c$ so the score function corresponds to the ordinary least squares loss without augmented components. In this case, we have $s_{t,\theta^\star} = -2w_t(Y_t - Z_t^T\theta^*)Z_t$ with $w_t = \frac{\pi_e(A_t)}{\mathbb{P}(A_t|\mathcal{H}_{t-1})}$. Moreover,

$$V_{t,\theta^\star} = \mathbb{E}\left[s_{t,\theta^\star} s_{t,\theta^\star}^T | \mathcal{H}_{t-1}\right] = \text{Var}\left(Z_t^T(Y_t - \theta^* Z_t^T)|\mathcal{H}_{t-1}\right) = \sigma_t^2 \mathbb{E}\left[w_t^2 Z_t Z_t^T | \mathcal{H}_{t-1}\right],$$

noting that $\text{Var}\left(Y_t|Z_t\right) = \text{Var}\left(\epsilon_t|\mathcal{H}_{t-1}\right) = \sigma_t^2$ and $\dot{s}_{t,\theta} = w_t Z_t Z_t^T$.

Note that we can obtain a closed form solution by solving for $\hat{\theta}_T$ in the equation

$$0 = \frac{1}{T}\sum_{t=1}^T V_{t,\theta^\star}^{-1/2} s_{t,\hat{\theta}} = \frac{1}{T}\sum_{t=1}^T - \left(w_t^2\sigma_t^2 \mathbb{E}\left[Z_t Z_t^T|\mathcal{H}_{t-1}\right]\right)^{-1/2} w_t(Y_t - Z_t^T\theta^*)Z_t.$$

Rearranging terms yields

$$\hat{\theta}_T = \left(\sum_{t=1}^T \sigma_t^{-1}\mathbb{E}\left[w_t^2 Z_t Z_t^T|\mathcal{H}_{t-1}\right]^{-1/2} Z_t Z_t^T\right)^{-1} \left(\sum_{t=1}^T \sigma_t^{-1} w_t^{-1}\mathbb{E}\left[w_t^2 Z_t Z_t^T|\mathcal{H}_{t-1}\right]^{-1/2} Z_t Y_t\right).$$

The formation of confidence intervals using Theorem 1 yields,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T w_t \sigma_t^{-1}\mathbb{E}\left[w_t^2 Z_t Z_t^T|\mathcal{H}_{t-1}\right]^{-1/2} Z_t Z_t^T \left(\hat{\theta}_T - \theta^\star\right) \xrightarrow{d} N(0, I_p).$$

In the case where the model is specified correctly, $\hat{\theta}_T$ is a consistent estimate of $\theta^\star$ for any choices of $w_t$ that follow Assumption 4. Therefore, the choice of $w_t$ becomes strictly a question of asymptotic efficiency. We can illustrate this with a few special cases

- Suppose $\mathbb{E}\left[Z_t Z_t^T|\mathcal{H}_{t-1}\right]$ is independent of $\mathcal{H}_{t-1}$, then $\frac{1}{T}\sum_{t=1}^T Z_t Z_t^T$ becomes a consistent estimate for $\mathbb{E}\left[Z_t Z_t^T\right]$.
  - When $\sigma_t$ is constant across $t$, it is optimal to choose $w_t = 1$. This reduces to the ordinary least squares solution (i.e., $\tilde{\theta}_T = \hat{\theta}_T$) and the confidence intervals become standard.
  - When $\sigma_t$ is not constant, it is optimal to choose $w_t \propto \frac{1}{\sigma_t^2}$. This reduces to the case of weighted least squares, and the confidence intervals again become standard.
- When each component of $Z_t$ corresponds to an indicator variable (i.e. $Z_{t,j} = 1_{A_t=j}$), then $\mathbb{E}\left[w_t^2 Z_t Z_t^T|\mathcal{H}_{t-1}\right] = \text{diag}\left(w_t^2 \mathbb{P}\left(Z_{t,j} = 1|\mathcal{H}_{t-1}\right)\right)$. In this case, the CLT simplifies to:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T \sigma_t^{-1/2}\text{diag}\left(\{\mathbb{P}(A_t = a|\mathcal{H}_{t=1})^{-1/2}\}_{a\in\mathcal{A}}\right)\left(\hat{\theta}_T - \theta^\star\right) \xrightarrow{d} N(0, I_p).$$

Of course, this example is unrealistic because it assumes oracle knowledge of $V_{t,\theta^\star} = \mathbb{E}[w_t^2 Z_t Z_t^T]$. In practice, the key challenge in applying Theorem 1 is to find a reliable method to estimate this quantity. In the next section, we tackle the problem of finding practical methods to construct a sequence of estimators $\{\hat{V}_t\}_{t=1}^T$ that can be used as plug-ins.

## 4 Practical Strategies for Covariance Estimation

Our strategy for building an estimator of the variance is to decompose $\text{Var}(s_{t,\theta^\star} \mid \mathcal{H}_{t-1})$ into terms that are either independent of the history or known to the experimenter and then find empirical estimators of these quantities from external data, similar in spirit to Kato (2020).

**Proposition 1.** *Recalling definitions for $\mathcal{H}_{t-1}$, $V_{t,\theta^\star}$, and $\theta^\star$ defined previously,*

$$\begin{aligned}\text{Var}(s_{t,\theta^\star} \mid \mathcal{H}_{t-1}) = \text{Var}\Big(\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\Big) + \\ \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi_e(a \mid X_t)^2}{\pi_t(a|X_t)} \mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a))\dot{m}_{\theta^\star}(X_t, a, Y_t(a))^T \mid X_t] \mid \mathcal{H}_{t-1}\right] - \\ \mathbb{E}\left[\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right] \mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T\right].\end{aligned}$$

This decomposition allows the dependence on $\mathcal{H}_{t-1}$ to be explicitly decomposed into pieces that are known conditional on $\mathcal{H}_{t-1}$ (action-selection probabilities) and terms that are independent of history. It is still challenging to use this decomposition without knowing the potential outcome distribution $(X_t, Y_t(a))$. To bridge this gap, we assume that we can sample freely from the marginal distribution of $X_t$ and then estimate the first and second moments of the conditional distribution $Y_t(a)|X_t$ using flexible machine learning approaches. In many applications, we may have access to an independent data set of features that can be used for this purpose. However, if this is unavailable, we can also use sample splitting to use the features from the experiment itself. We briefly discuss both of these strategies below.

### 4.1 Using external data

We first assume access to an external dataset composed of observations of $X$ independent of the history but with the same marginal distribution as $X_t$.

**Assumption 8.** *There exists an external data set $\tilde{X} := \{\tilde{X}_i\}_{i=1}^n$, independent of $\mathcal{H}_t$ for all $t \in T$, where $\tilde{X}_i \overset{\text{iid}}{\sim} p(x)$ and $n = [rT]$ for some fixed $r \in (0, \infty)$.*

We also assume access to models targeting the conditional means and variance that can be used to create a plug-in estimate of $\text{Var}\left(s_{t,\theta^\star}|\mathcal{H}_{t-1}\right)$.

**Assumption 9.** *We can construct functions $g_t : \mathbb{R}^p \times \mathcal{A} \to \mathbb{R}^d$ and $h_t : \mathbb{R}^p \times \mathcal{A} \to \mathbb{R}^{d\times d}$ adapted to $\sigma(\mathcal{H}_{t-1})$ such that $g_t(X_t, a) - \mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)|X_t, A_t = a\right] \xrightarrow{p} 0$ for all $a \in \mathcal{A}$ and*

$$h_t(X_t, a) - \mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\dot{m}_{\theta^\star}(X_t, A_t, Y_t)^T|X_t, A_t = a\right] \xrightarrow{p} 0$$

*for all $a \in \mathcal{A}$.*

We postpone discussion of how to construct the functions $g_t$ and $h_t$ to Proposition 3. In this section, we assume the existence of these quantities and use them to construct an estimate for $\text{Var}\left(s_{t,\theta^\star}|\mathcal{H}_{t-1}\right)$.

**Proposition 2.** *Define the functions*

$$\hat{\nu}_t(X_i) = \sum_{a=1}^{K} \pi_e(A_t = a \mid X_t) g_t(a, X_i), \tag{9}$$

*and $\bar{\nu} = \frac{1}{n}\sum_{X_i\in\tilde{X}} \hat{\nu}_t(X_i)$. Define:*

$$\begin{aligned}\hat{V}_t = \frac{1}{n-1}\left(\sum_{X_i\in\tilde{\mathcal{X}}} \hat{\nu}_t(X_i) - \bar{\nu}\right)\left(\sum_{X_i\in\tilde{\mathcal{X}}} \hat{\nu}_t(X_i) - \bar{\nu}\right)^T \\ + \frac{1}{n}\sum_{X_i\in\tilde{\mathcal{X}}}\sum_{a=1}^{K} \frac{\pi_e(A_t = a \mid X_i)^2\, h_t(X_i, a)}{\mathbb{P}(A_t = a \mid X_i, \mathcal{H}_{t-1})} - \frac{1}{n}\sum_{X_i\in\tilde{\mathcal{X}}} \hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T.\end{aligned}$$

*Then, under Assumptions 1-9,* $\left\|\hat{V}_t - V_{t,\theta^\star}\right\|_{op} \xrightarrow{p} 0$.

Of course, it is far from clear how to construct $g_t$ and $h_t$ so that Assumption 9 is satisfied. In the remainder of this section, we provide a construction in the case of certain classes of objective functions $m_\theta$ and demonstrate that generalized linear models are a special case of this.

**Proposition 3.** *Assume you have access to an estimator* $\bar{\theta}_T \in \sigma(\mathcal{H}_T)$ *such that* $\bar{\theta}_T \xrightarrow{p} \theta^\star$. *Assume that you can construct a sequence of continuous functions* $f_t : \mathbb{R} \times \mathcal{A} \to \mathbb{R}$ *and* $j_t : \mathbb{R} \times \mathcal{A} \to \mathbb{R}$ *adapted to* $\sigma(\mathcal{H}_{t-1})$ *such that* $f_t(X_t, a) - \mathbb{E}\left[Y_t | X_t, A_t = a\right] \xrightarrow{p} 0$ *and* $j_t(X_t, a) - Var[Y_t | X_t, A_t = a] \xrightarrow{p} 0$. *Furthermore, assume that* $\dot{m}_\theta$ *is linear in* $y$ *such that it can be decomposed as* $\dot{m}_\theta(x, a, y) = z_\theta(x, a)y + \nu_\theta(x, a)$ *for some continuous functions* $z, \nu$. *Then, defining*

$$g_t(x, a) = z_{\bar{\theta}_T}(x, a) f_t(x, a) + \nu_{\bar{\theta}_T}(x, a)$$
$$h_t(x, a) = z_{\bar{\theta}_T}(x, a) z_{\bar{\theta}_T}(x, a)^T j_t(x, a),$$

*will satisfy the conditions of Assumption 9.*

Although this may seem restrictive, GLMs are a notable class of models that have a score function that satisfies these conditions. In our settings, GLMs will satisfy a first order condition of the form

$$\sum_{t=1}^{T} (Y_t - \psi(\theta^T z(X_t, A_t))) z(X_t, A_t) = 0,$$

where $z : \mathbb{R}^d \times |\mathcal{A}| \to \mathbb{R}^p$ denotes some deterministic transformations of $(X_t, A_t)$ (e.g., one-hot encoding of $A_t$, interaction terms between $A_t$ and $X_t$) and $\psi(\cdot)$ is the inverse-link function that maps $\theta^T z(X_t, A_t)$ to the mean under the (possibly misspecified) working model. In this case, we have:

$$g_t(X_t, a) = \left(f_t(X_t, a) - \psi(\theta^T z(X_t, A_t))\right) z(X_t, A_t) = f_t(X_t, a) z(X_t, A_t) - \psi(\theta^T z(X_t, A_t)) z(X_t, A_t),$$
$$h_t(X_t, a) = z(X_t, a) z(X_t, a)^T j_t(X_t, a).$$

Proposition 3 simplifies the variance estimation task, as it converts the problem of estimation of potentially very high-dimensional covariance matrices into a simpler question of estimating conditional means and variances of $Y_t$. Note that $f_t$ can be the same model that was used in the definition of the MAIPWM estimator (Equation (6)) so this procedure only requires the management of an additional second model targeting $\mathbb{E}\left[Y_t^2 | X_t, A_t = a\right]$. In the case where $m_\theta$ is *not* linear in $y$, then Jensen's inequality implies that a straightforward plug-in estimate is not available since $\mathbb{E}\left[\dot{m}_\theta(X_t, A_t, Y_t)\right] \neq \dot{m}_\theta(X_t, A_t, \mathbb{E}[Y_t])$. In these cases, more complicated methods of density estimation will be required.

Some care must also be taken when choosing $\bar{\theta}_T$ in Proposition 3. Although $\hat{\theta}_T$ is consistent, it is defined as an estimator that satisfies $\frac{1}{T}\sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\hat{\theta}} = o_p(1)$ and therefore assumes that $\hat{V}_t$ has already been constructed in its definition, so it is not available when we need to construct $\hat{V}_t$. Fortunately, $\tilde{\theta}_T$ can be shown to be consistent.

**Proposition 4.** *Let* $\tilde{\theta}_T$ *be defined as in Equation* (6). *Under the conditions of Theorem 1,* $\left\|\tilde{\theta}_T - \theta^\star\right\|_1 = o_p(1)$.

This results in a two-step procedure in which $\tilde{\theta}_T$ is first estimated and used to construct $\{\hat{V}_t\}_{t=1}^T$. We then use these matrices to construct a corrected estimator $\hat{\theta}_T$ that allows us to form confidence sets. We summarize the end-to-end procedure for producing confidence sets under adaptive sampling for GLMs in Algorithm 1.

**Remark 2** (Target versus nuisance misspecification)**.** *The misspecification considered throughout this paper refers to the working model* $m_\theta$ *used to define the target parameter. This should not be confused with misspecification of the nuisance models used in Proposition 3 to estimate the covariance matrix. While our framework permits* $m_\theta$ *to be misspecified relative to the underlying data-generating process, valid inference still requires sufficiently accurate estimation of the conditional moments used to construct* $\widehat{V}_t$. *Thus, the burden of modeling is not eliminated entirely, but shifted from requiring a correctly specified parametric target model to requiring estimable nuisance quantities (conditional first and second moments) for covariance estimation.*

**Algorithm 1** Construction of Confidence Intervals for GLMs Under Adaptive Sampling

**Require:** Data $\{(X_t, A_t, Y_t)\}_{t=1}^T$; evaluation dataset $\tilde{X} := \{\tilde{X}_i\}_{i=1}^n$; target policy $\pi_e(a \mid X_t)$; model class $\{m_\theta : \theta \in \Theta\}$ for the GLM with transformation function $z(x, a)$ and link function $\psi(\cdot)$.

1: **for** each time step $t$ **do**

2: Construct predictive models $f_t, j_t$ using $\mathcal{H}_{t-1}$ targeting the conditional mean and variances:

$$f_t(X_i, a) \approx E[Y_i \mid X_i, A_i = a], \quad j_t(X_i, a) \approx \mathrm{Var}[Y_i \mid X_i, A_i = a].$$

3: **end for**

4: $\tilde{\theta}_T \leftarrow \mathrm{argmax}_{\theta\in\Theta} \sum_{t=1}^T \sum_{a=1}^K \pi_e(a \mid X_t) \left( m_\theta(a, X_t, f_t(a, X_t)) + \mathbb{1}_{\{A_t=a\}} \frac{m_\theta(A_t,X_t,Y_t) - m_\theta(X_t,A_t,f_t(X_t,A_t))}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})} \right)$.

5: Compute

$$g_t(x, a) \leftarrow f_t(x, a)z(x, a) - \psi(\tilde{\theta}_T^T z(x, a))z(x, a),$$
$$h_t(x, a) \leftarrow z(x, a)z(x, a)^T j_t(x, a).$$

6: Use $\tilde{X}$ and Proposition 2 to construct $\hat{V}_t$ for every $t \in [T]$.

7: Use $\{\hat{V}_t\}_{t=1}^T$ and apply Theorem 1 to construct $\hat{\theta}_T$ and corresponding confidence intervals.

8: **Output:** Estimated parameter $\hat{\theta}_T$ and asymptotic $(1 - \alpha)$ confidence region

$$\mathcal{C}_{\theta^\star} = \left\{ \theta \in \Theta : \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T}(\hat{\theta}_T - \theta) \right)^\top \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T}(\hat{\theta}_T - \theta) \right) \leq \chi^2_{d,1-\alpha} \right\}.$$

### 4.2 Using sequential sample splitting

Alternatively, when an external dataset is not available, we can use sample splitting to create an independent data set to estimate $\mathrm{Var}\left(s_{t,\theta^\star} | \mathcal{H}_{t-1}\right)$. At each $t$, we draw an independent random variable $\zeta_t \sim \mathrm{Ber}(r)$ and then assign $X_t$ into one of two parallel histories $\mathcal{H}_t^{(1)}$ and $\mathcal{H}_t^{(2)}$:

$$X_t \in \begin{cases} \mathcal{H}_t^{(1)} & \text{when } \zeta_t = 0 \\ \mathcal{H}_t^{(2)} & \text{when } \zeta_t = 1 \end{cases} \text{ and } \pi_t(a|X_t) \in \begin{cases} \sigma(\mathcal{H}_{t-1}^{(1)}, X_t) & \text{when } \zeta_t = 0 \\ \sigma(\mathcal{H}_{t-1}^{(2)}, X_t) & \text{when } \zeta_t = 1 \end{cases}.$$

Proceeding in this fashion constructs two separate histories with the same distribution that are independent of each other. Note that for this particular application, we are only trying to sample from the marginal distribution of $X_t$ so the assignment of $A_t$ is only required for one of the two histories. An alternative experimental setup might only track the triples $(X_t, A_t, Y_t)$ for $\mathcal{H}_t^{(1)}$ and then refrain from treating or tracking outcomes for individuals in $\mathcal{H}_t^{(2)}$. In contexts where there is a cost to administering the treatment, the latter design will be preferable, but in other situations it may not be possible or desirable to refrain from treatment in which case the former is required.

Regardless of the particular approach, the only requirement is that $\zeta_t$ is independent of both histories and the predictive models are trained on only one of the tracked histories.

**Assumption 10.** *$\zeta_t \sim Ber(p)$ for some fixed $p \in (0, 1)$ and is independent of $\mathcal{H}_{t-1}^{(1)}$ and $\mathcal{H}_{t-1}^{(2)}$ for all $t \in [T]$.*

Assumption 10 can be used in place of Assumption 8 to construct an independent dataset of features when one is not available.

**Assumption 11.** *There exist functions $g_t : \mathbb{R}^p \times \mathcal{A} \to \mathbb{R}$ and $h_t : \mathbb{R}^p \times \mathcal{A} \to \mathbb{R}^{d\times d}$ adapted to $\sigma(\mathcal{H}_{t-1}^{(1)})$ such that $g_t(X_t, a) - \mathbb{E}[Y_t|X_t, A_t = a] \xrightarrow{p} 0$ for all $a \in \mathcal{A}$ and*

$$h_t(X_t, a) - \mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\dot{m}_{\theta^\star}(X_t, A_t, Y_t)^T | X_t, A_t = a\right] \xrightarrow{p} 0$$

*for all $a \in \mathcal{A}$.*

Assumption 11 is almost the same as Assumption 9 but we require that the models are only trained on $\mathcal{H}_{t-1}^{(1)}$ to ensure that $\mathcal{H}_{t-1}^{(2)}$ can be preserved as an independent data set for calculating $\hat{V}_t^{-1/2}$. At this point, the procedure can proceed the same as in Section 4.1. For completeness, we present a slightly modified version of Proposition 2 adapted to the case of sample splitting.

**Proposition 5.** *Under Assumptions 1- 6, 10, and 11, define* $\hat{\nu}_t(X_i)$ *as in Equation* (9) *and* $n_p := \sum_{t=1}^{T} \zeta_t$. *Define:*

$$\hat{V}_t = \frac{1}{n_p - 1}\left(\sum_{X_i \in \mathcal{H}_t^{(2)}} \hat{\nu}_t(X_i) - \frac{1}{n_p}\sum_{X_i \in \mathcal{H}_t^{(2)}} \hat{\nu}_t(X_i)\right)\left(\sum_{X_i \in \mathcal{H}_t^{(2)}} \hat{\nu}_t(X_i) - \frac{1}{n_p}\sum_{X_i \in \mathcal{H}_t^{(2)}} \hat{\nu}_t(X_i)\right)^T + \frac{1}{n_p}\sum_{X_i \in \mathcal{H}_t^{(2)}}\sum_{a=1}^{K}\left[\frac{\pi_e(A_t = a \mid X_i)^2\, h_t(X_i, a)}{\mathbb{P}(A_t = a \mid X_i, \mathcal{H}_{t-1}^{(2)})}\right] - \frac{1}{n_p}\sum_{X_i \in \mathcal{H}_t^{(2)}} \hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T.$$

*Then* $\left\|\hat{V}_t - V_{t,\theta^\star}\right\|_{\text{op}} \xrightarrow{p} 0.$

Algorithm 1 remains the same when used in this setting, with the only modification being that Proposition 5 should be used at the sixth step instead of Proposition 2.

## 5 Empirical Results

We deploy these methods in the contextual bandit problem, where the goal is to choose a sequence of actions $\{A_t\}_{t=1}^T$ that minimize *regret*, $\sum_{t=1}^{T}(\mu_t^\star - Y_t)$, where $\mu_t^\star := \max\{\mathbb{E}[Y_t(a)] : a \in \mathcal{A}\}$ is the arm with the highest expected reward. This is a well-studied problem with a variety of proposed solutions; we focus on strategies for choosing $P(A_t = a|X_t, \mathcal{H}_{t-1})$ that simultaneously leverage flexible predictive models $f_t$ and $j_t$ described in Assumption 9 targeting $\mathbb{E}[Y_t|X_t, A_t]$ and $\mathbb{E}[Y_t^2|X_t, A_t]$, respectively.

**Strategies for selecting $A_t$:**

- A **uniform** strategy where $\mathbb{P}(A_t = a|X_t, \mathcal{H}_{t-1}) = \frac{1}{|\mathcal{A}|}$ for all $t$. Classical statistical guarantees will apply in this setting since the decision-making is stationary and non-adaptive.
- An **epsilon greedy** strategy that lets

$$A_t = \begin{cases} \text{argmax}_{a\in\mathcal{A}}\, f_t(X_t, a) & \text{with probability } 1-\epsilon \\ \text{any other } a \in \mathcal{A} & \text{with probability } \frac{\epsilon}{|\mathcal{A}|-1}. \end{cases}$$

- An **upper confidence bound (UCB)** strategy that constructs confidence intervals using $f_t$ to define the centerpoint and $j_t$ to control the width. Here,

$$A_t = \underset{a\in\mathcal{A}}{\text{argmax}}\left[f_t(X_t, a) - q_{\alpha/2}\sqrt{j_t(X_t, a)}\right],$$

  where $q_{\alpha/2}$ denotes the $\alpha/2$ quantile of a standard normal distribution.
- An approach inspired by **Thompson Sampling** which draws from a distribution of $Y_t|X_t, a$ for each $a$ to form an estimate $\hat{Y}_t(a)$ and then lets $A_t := \text{argmax}\,\hat{Y}_t(a)$. We apply this algorithm in the current setting using the working assumption that $Y_t(a) \sim N(f_t(X_t, a), j_t(X_t, a) - f_t^2(X_t, a))$. We also consider versions of this method with a clipping constraint so all probabilities are constrained to lie in $[0.05, 0.95]$.

We use a real data example to test the confidence intervals constructed using Theorem 1 under the above sampling strategies. The data set is sourced from the Osteoarthritis Initiative (OAI) (Nevitt et al., 2006),

a publicly available data set provided by the NIH. The OAI is a ten-year longitudinal observational study of men and women affected by osteoarthritis. The study collects several baseline measurements about each participant's knee health, such as self-reported measurements of pain, disability status, and flexion contracture, in addition to demographic measurements such as age, BMI, and sex. After enrollment in the study, measurements of knee health, such as Kellgren and Lawrence (KL) grade, are taken at yearly intervals.

We consider the outcome ($Y_t$) to be the four year change in KL grade for the affected knee of an osteoarthritic patient. Although KL grade is a discrete numeric variable, we treat KL grade as a continuous variable for the purposes of this example. The chosen features ($X_t$) include the aforementioned baseline measurements of knee health and demographic variables as well as additional risk factors identified in Dunn et al. (2020) such as self-reported quality of life scores and use of pain medication. Since OAI was an observational study to characterize disease progression rather than a trial to compare treatments, to evaluate methodology, we construct a semi-synthetic data set that applies hypothetical treatments ($A_t$) and then enforces a synthetic relationship between $A_t$ and $Y_t$ while preserving the conditional distribution of $Y_t \mid X_t$. This is done using machine learning methods to first learn the distribution of $Y_t \mid X_t$ and then to create synthetic outcomes that combine the output of the machine learning model with hypothetical treatment effects. For a detailed description of the semi-synthetic dataset creation process, see Appendix C. The end result is a dataset with similar complexity to the original longitudinal study but with the option to sample patients sequentially with ground-truth knowledge of the treatment effect and target parameters. Note that to ensure that this setting would align with the hardest case where treatment policies would not be likely to converge, we picked treatments so that there would be the same expected reward across multiple arms and there would be no unique optimal policy.

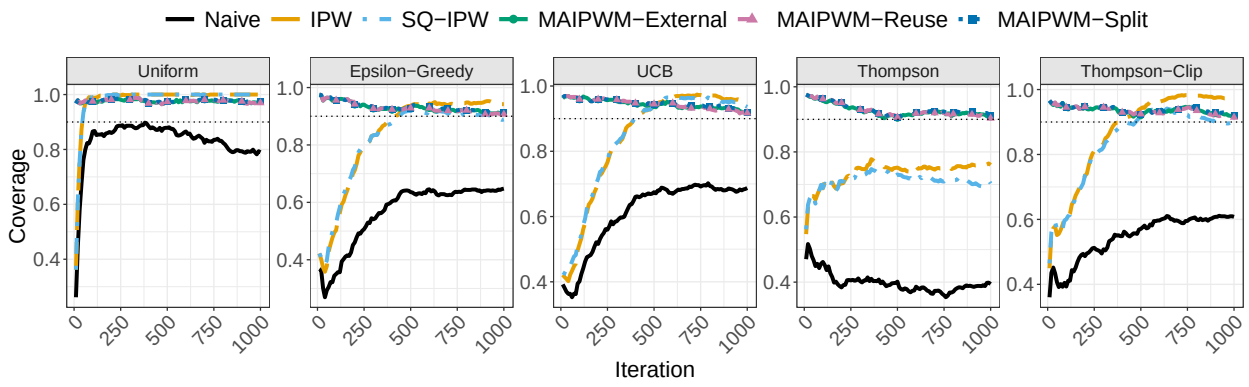
In simulations, patients come online one at a time and $\hat{\theta}_t$ is recomputed for each $t$ with Theorem 1 used to form confidence intervals. As discussed in Section 4, online model-based estimates of the conditional mean and variance are trained sequentially (using random forests) as inputs into the MAIPWM estimator. To increase computational efficiency, we choose to re-train the predictive models after every 100 data points, though in principle they could be re-trained after every sample. For continuous outcomes, we consider a linear model with $m_\theta := -(Y_t - \sum_{a \in \mathcal{A}} \theta_a \mathbb{1}_{A_t=a})^2$.

We track the length of the confidence intervals and their nominal coverage rate (given target coverage of 0.9) in Figure 2. The methods we consider are:

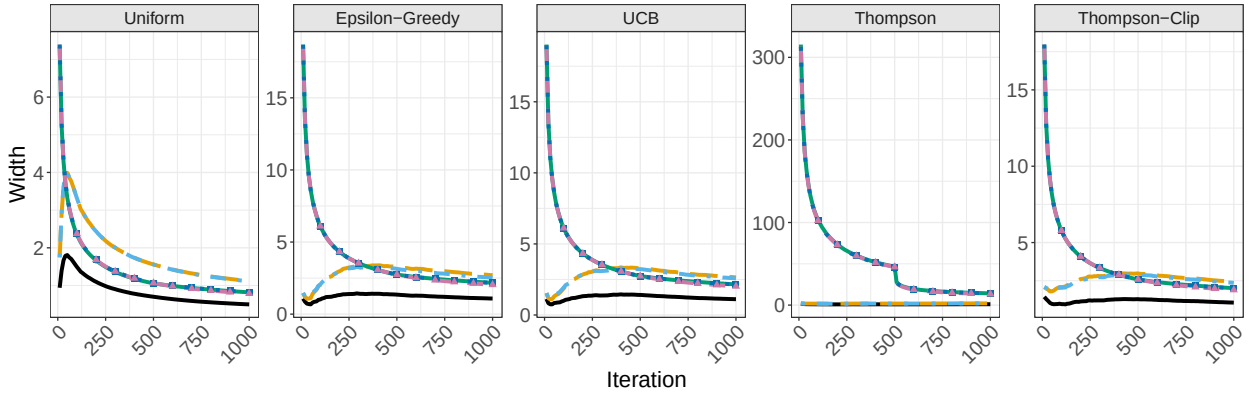
**Methods for confidence interval construction:**

- **Naive** Use off-the shelf maximum likelihood estimation with equal weighting of outcomes and sandwich estimates of the variance.
- **IPW** Uses the methodology of Guo & Xu (2025) to construct confidence intervals, which relies on inverse propensity weighting of the score function and sandwich estimates of the variance.
- **SQ-IPW** Use the methodology of Zhang et al. (2021) to construct confidence intervals, which weight observations by the *square root* of the propensity scores, setting $w_t = \sqrt{\frac{\pi_e(A_t|X_t)}{P(A_t|X_t,\mathcal{H}_{t-1})}}$ and using sandwich estimates of the variances.
- **MAIPWM–External** Using the method described in Theorem 1 with plug-in estimates $V_{t,\theta}$ constructed using external data as described in Section 4.1.
- **MAIPWM–Sample-Splitting** Using the method described in Theorem 1 with plug-in estimates $V_{t,\theta}$ constructed using sequential sample splitting as described in Section 4.2.
- **MAIPWM–Reuse** Using the method described in Theorem 1 with plug-in estimates $V_{t,\theta^\star}$ constructed by naively reusing all $X_i \in \mathcal{H}_{t-1}$ at time step $t$.

We tested the MAIPWM estimator in all combinations of strategies used to estimate $V_{t,\theta^\star}$ and select actions. The empirical results shown in Figure 2 align with the theory — naive maximum likelihood estimates undercover in all situations other than the uniform (nonadaptive case). IPW and SQ-IPW methods require




(a) Coverage for target $1 - \alpha = 0.9$




(b) Confidence interval width for confidence intervals shown above

Figure 2: Confidence intervals constructed using Theorem 1 using ML-based estimates of the variance cover in all scenarios. GLMs using naive inverse propensity weighting often undercover, especially in situations where the assignment probabilities vary substantially over time. We note that both sample splitting and external data reuse for covariance estimation are valid, but sample splitting has significantly wider confidence intervals. Reusing the same data for variance estimation and parameter estimation performs similarly to using external data, though we lack theoretical guarantees for this method.

a significantly larger number of samples before they achieve the nominal coverage rate and do not cover at all in the case of Thompson sampling. The MAIPW estimator using sample splitting and external data both cover correctly, with the sample splitting method paying a price in terms of notably wider confidence intervals. Although we have no theoretical results supporting MAIPWM-Reuse, which reuses the historical data to estimate the covariance, this method has nearly identical results to the two MAIPWM methods that use external samples. Future work can potentially justify this formally.

## 6 Conclusion

We present a method for inference in $M$ estimation problems under a potentially misspecified working model. The method requires the estimation of time-varying covariance matrices, for which we provide an algorithm for generalized linear models using flexible machine learning approaches. Empirical results demonstrate that the method has correct nominal coverage, while existing approaches often undercover in situations where action-selection probabilities vary considerably and the variance of the score function does not converge. Our work suggests several potential avenues for follow-up work, which we detail below:

**Estimation of action-selection probabilities** Throughout, we assume that the probabilities of each action being selected in the experiment are determined by the experimenter and known, which sometimes cannot be guaranteed in practical settings. Weakening this assumption first requires modifying the asymptotic arguments for Theorem 1 and ensuring the rate of estimation of these probabilities to converge sufficiently fast. It also complicates the estimation of the time-varying variance as explained in Section 4 since Proposition 1 crucially relies on knowledge of the action-selection probabilities to decompose the variance into a tractable form for plug-in estimation.

**Data reuse** In estimating the variance of the score function, we either assume access to an external dataset of features with the same marginal distribution as in the sequential experiment or are forced to split the data (and consequently sacrifice power). Empirical results suggest that simply reusing the collected sequential data for both variance estimation and parameter estimation provides nearly identical results compared to when an external dataset is available, but we have not provided a rigorous guarantee.

**Efficiency of estimators** Although we have demonstrated that the proposed estimator results in valid confidence intervals, we have not shown it is optimal in any sense or demonstrated results about semiparametric efficiency, which is a logical next step for future work.

### Acknowledgments

Data and/or research tools used in the preparation of this manuscript were obtained and analyzed from the controlled access data sets distributed by the Osteoarthritis Initiative (OAI), a data repository located within the NIMH Data Archive (NDA). OAI is a collaborative informatics system created by the National Institute of Mental Health and the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMS) to provide a worldwide resource to quicken the pace of biomarker identification, scientific investigation and OA drug development. Dataset identifier(s): [outcomes99, kxrsqbu06, kxrsqbu00, enrollees, allclinical00, allclinical06].

## References

M Mehdi Afsar, Trafford Crump, and Behrouz Far. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38, 2022.

Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley, 2015. ISBN 9781118730034. URL `https://books.google.com/books?id=jlIqBgAAQBAJ`.

Aurelien Bibaut, Antoine Chambaz, Maria Dimakopoulou, Nathan Kallus, and Mark Laan. Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, 34:28548–28559, 12 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 1033–1064, 2024. URL https://proceedings.mlr.press/v236/cook24a.html.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1194–1203, 2018. URL https://proceedings.mlr.press/v80/deshpande18a.html.

R. Dunn, J. Greenhouse, D. James, D. Ohlssen, and P. Mesenbrink. Risk scoring for time to end-stage knee osteoarthritis: data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 28(8):1020–1029, 2020. ISSN 1063-4584. doi: https://doi.org/10.1016/j.joca.2019.12.013. URL https://www.sciencedirect.com/science/article/pii/S1063458420309985.

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*, volume 6, pp. 513–536. University of California Press, 1972.

Yongyi Guo and Ziping Xu. Statistical inference for misspecified contextual bandits, 2025. URL https://arxiv.org/abs/2509.06287.

Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15): e2014602118, 2021a. doi: 10.1073/pnas.2014602118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2014602118.

Vitor Hadad, David A. Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15): e2014602118, 2021b. doi: 10.1073/pnas.2014602118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2014602118.

Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

Masahiro Kato. Confidence interval for off-policy evaluation from dependent samples via bandit algorithm: Approach from standardized martingales. *arXiv:2006.06982*, 2020.

Koulik Khamaru, Yash Deshpande, Tor Lattimore, Lester Mackey, and Martin J. Wainwright. Near-optimal inference in adaptive linear regression. *The Annals of Statistics*, 53(6):2329 – 2355, 2025. doi: 10.1214/24-AOS2450. URL https://doi.org/10.1214/24-AOS2450.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, 10(1):154–166, 1982. ISSN 00905364. URL http://www.jstor.org/stable/2240506.

Licong Lin, Koulik Khamaru, and Martin J. Wainwright. Semiparametric inference based on adaptively collected data. *The Annals of Statistics*, 53(3):989 – 1014, 2025. doi: 10.1214/24-AOS2485. URL https://doi.org/10.1214/24-AOS2485.

Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. Reinforcement learning for clinical decision support in critical care: Comprehensive review. *J Med Internet Res*, 22 (7):e18477, Jul 2020. ISSN 1438-8871. doi: 10.2196/18477. URL https://www.jmir.org/2020/7/e18477.

Michael Nevitt, David Felson, and Gayle Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1:2, 2006.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*, 2019.

Vasilis Syrgkanis and Ruohan Zhan. Post reinforcement learning inference. *axiv:2302.08854*, 2024. URL https://arxiv.org/abs/2302.08854.

Aad van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1), 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

Mufang Ying, Koulik Khamaru, and Cun-Hui Zhang. Adaptive linear estimating equations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52051–52072. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a399456a191ca36c7c78dff367887f0a-Paper-Conference.pdf.

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. *J Med Internet Res*, 19(10):e338, Oct 2017. ISSN 1438-8871. doi: 10.2196/jmir.7994. URL http://www.jmir.org/2017/10/e338/.

Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf.

Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with M-estimators on adaptively collected data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7460–7471, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3d7d9461075eb7c37fbbfcad1d7042c1-Paper.pdf.

Kelly W. Zhang, Lucas Janson, and Susan A. Murphy. Statistical inference after adaptive sampling for longitudinal data. *arXiv:2202.07098*, 2023.

Yufan Zhao, Michael R. Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28(26):3294–3315, 2009. doi: https://doi.org/10.1002/sim.3720. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3720.

Tijana Zrnic and Emmanuel J. Candès. Active statistical inference. *Proceedings of the 41st International Conference on Machine Learning*, 2024.

## A Preliminaries

We introduce several key facts about martingales that are used throughout. For a more thorough discussion, consult Hall & Heyde (2014).

**Fact 1** (Martingale Weak Law of Large Numbers, Hall & Heyde (2014))**.** *Let $(S_n, \mathcal{H}_n)_{n\geq 1}$ be a martingale where $S_n = \sum_{i=1}^n X_i$, and let $\{b_n\}$ be a sequence of positive constants with $b_n \to \infty$. Writing $X_{ni} = X_i \mathbb{1}[|X_i| \leq b_n]$ for $1 \leq i \leq n$, we have $b_n^{-1} S_n \xrightarrow{p} 0$ as $n \to \infty$ if*

1. $\sum_{i=1}^n P\left(|X_i| > b_n\right) \to 0$
2. $b_n^{-1} \sum_{i=1}^n \mathbb{E}\left[X_{ni} \mid \mathcal{H}_{n-1}\right] \xrightarrow{\mathrm{p}} 0$
3. $b_n^{-2} \sum_{i=1}^n \left\{\mathbb{E}\left[X_{ni}^2\right] - \mathbb{E}\left[\mathbb{E}\left[X_{ni} \mid \mathcal{H}_{n-1}\right]\right]^2\right\} \to 0$

**Remark 3.** *A sufficient condition for Fact 1 to hold is when $X_i$ is bounded by a constant.*

**Remark 4.** *A sufficient condition for Fact 1 to hold is when $X_i$ is square integrable and $\frac{1}{T^2}\sum_{t=1}^T \mathbb{E}[X_i^2] \to 0$.*

# B Deferred Proofs

## B.1 Proof of Theorem 1

The overall structure of the proof is similar to Zhang et al. (2021), which in turn is based on classical proofs of the asymptotic normality of $M$-estimators such as the ones described in van der Vaart (2000). The major differences in the approach are that we now need to deal with the variance stabilizing matrices $\hat{V}_t$ and bound the complexity of the predictive model $f_t$.

The proof relies on two Lemmas:

1. Lemma 3 uses the first order conditions to show that an appropriately rescaled version of the score function $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\theta^\star}$ is asymptotically normal using a Martingale Central Limit Theorem (Hall & Heyde, 2014).

2. Lemma 4 shows that $\left\| \hat{\theta}_T - \theta^\star \right\|_1 = o_p(1)$.

Given these two lemma, we can now use Taylor expansion around the score function to prove Theorem 1.

First, by Taylor's theorem we can write for each $t$ and some $\tilde{\theta}_t$ on the line segment between $\hat{\theta}_T$ and $\theta^\star$ that

$$s_{t,\theta^\star} = s_{t,\hat{\theta}_T} + \dot{s}_{t,\hat{\theta}_T}(\theta^\star - \hat{\theta}_T) + \frac{1}{2} \left( \theta^\star - \hat{\theta}_T \right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \theta^\star - \hat{\theta}_T \right).$$

For each $t$, we multiply by $\frac{1}{\sqrt{T}} \hat{V}_t^{-1/2}$ and sum over $t \in [T]$ to arrive at

$$\begin{aligned} -\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\theta^\star} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \Big( & s_{t,\hat{\theta}_T} - \dot{s}_{t,\hat{\theta}_T}(\hat{\theta}_T - \theta^\star) \\ & - \frac{1}{2} \left( \hat{\theta}_T - \theta^\star \right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \hat{\theta}_T - \theta^\star \right) \Big). \end{aligned}$$

Recall that by assumption $\frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} = o_p(1/\sqrt{T})$, this simplifies the expression to

$$\begin{aligned} -\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\theta^\star} &= o_p(1) + \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \left( \dot{s}_{t,\hat{\theta}_T}(\hat{\theta}_T - \theta^\star) + \frac{1}{2} \left( \hat{\theta}_T - \theta^\star \right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \hat{\theta}_T - \theta^\star \right) \right) \\ &= o_p(1) + \frac{1}{\sqrt{T}} \Bigg( I_d + \frac{1}{2} \sum_{t=1}^{T} \hat{V}_t^{-1/2} (\hat{\theta}_T - \theta^\star)^T \ddot{s}_{t,\tilde{\theta}_t} \\ &\qquad \times \left( \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \Bigg) \\ &\qquad \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} (\hat{\theta}_T - \theta^\star). \end{aligned}$$

Recall that $-\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\theta^\star} \xrightarrow{d} N(0, I_d)$. By Slutsky's Theorem, it is therefore sufficient to to show that

$$\left\| \sum_{t=1}^{T} \hat{V}_t^{-1/2} \left( \hat{\theta}_T - \theta^\star \right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 = o_p(1),$$

in order to conclude that $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \left(\hat{\theta}_T - \theta^\star\right) \xrightarrow{d} N(0, I_d)$. We write this as follows.

$$
\begin{aligned}
&\left\| \sum_{t=1}^{T} \hat{V}_t^{-1/2} \left(\hat{\theta}_T - \theta^\star\right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 \\
&= \left\| \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \left(\hat{\theta}_T - \theta^\star\right)^T \ddot{s}_{t,\tilde{\theta}_t} \left( \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 \\
&\leq \left\| \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \left(\hat{\theta}_T - \theta^\star\right)^T \ddot{s}_{t,\tilde{\theta}_t} \right\|_2 t \left\| \left( \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 \\
&\leq \frac{1}{T} \left\| \hat{\theta}_T - \theta^\star \right\| \sum_{t=1}^{T} \left\| \hat{V}_t^{-1/2} \ddot{s}_{t,\tilde{\theta}_t} \right\|_2 \left\| \left( \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 \\
&\leq \left\| \hat{\theta}_T - \theta^\star \right\|_2 \frac{1}{\delta_{\min}} \frac{1}{T} \sum_{t=1}^{T} \left\| \ddot{s}_{t,\tilde{\theta}_t} \right\|_2 \left\| \left( \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 .
\end{aligned}
$$

By Lemma 4, $\left\| \hat{\theta}_T - \theta^\star \right\|_2 = o_p(1)$ and by Lemma 2, $\left\| \left( \frac{1}{T} \sum_{t=1}^{T} \hat{V}_t^{-1/2} \dot{s}_{t,\hat{\theta}_T} \right)^{-1} \right\|_2 = O_p(1)$ . Therefore, it is sufficient to show that the remaining terms are $O_p(1)$ to conclude that the entire term is $o_p(1)$. We know that $1/\delta_{\min}$ is bounded because $\delta_{\min} > 0$ by assumption. We tackle each of these terms individually (note that by equivalence of matrix norms, it is sufficient to show convergence in 1-norm).

**Showing $\frac{1}{T} \sum_{t=1}^{T} \left\| \ddot{s}_{t,\tilde{\theta}_t} \right\|_1 = O_p(1)$.** We for some

$$
\begin{aligned}
\left\| \ddot{s}_{t,\tilde{\theta}_t} \right\|_1 &= \Bigg\| \sum_{a=1}^{K} \pi_e(A_t = a | X_t) \Big( \dddot{m}_{\tilde{\theta}_t}(a, X_t, f_t(a, X_t)) \\
&\qquad + \mathbb{1}_{A_t = a} \frac{\dddot{m}_{\tilde{\theta}_t}(X_t, A_t, Y_t) - \dddot{m}_{\tilde{\theta}_t}(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a | X_t, \mathcal{H}_{t-1}\right)} \Big) \Bigg\|_1 \\
&\leq \frac{\pi_e(A_t | X_t)}{\mathbb{P}\left(A_t | X_t, \mathcal{H}_{t-1}\right)} \left\| \dddot{m}_{\tilde{\theta}_t}(X_t, A_t, Y_t) \right\|_1 \\
&\qquad + \pi_e(A_t = a | X_t) \sum_{a=1}^{K} \left( 1 - \frac{\mathbb{1}_{A_t = a}}{\mathbb{P}\left(A_t = a | X_t, \mathcal{H}_{t-1}\right)} \right) \left\| \dddot{m}_{\tilde{\theta}_t}(X_t, a, f_t(X_t, a)) \right\|_1 \\
&\leq \frac{\pi_e(A_t | X_t)}{\mathbb{P}\left(A_t | X_t, \mathcal{H}_{t-1}\right)} u_m(X_t, A_t, Y_t) \\
&\qquad + \pi_e(A_t = a | X_t) \sum_{a=1}^{K} \left( 1 - \frac{\mathbb{1}_{A_t = a}}{\mathbb{P}\left(A_t = a | X_t, \mathcal{H}_{t-1}\right)} \right) u_{\mathcal{F}}(X_t, a, f_t(X_t, a)),
\end{aligned}
$$

where the last inequalities follow from Assumption 7 and Assumption 6 after noting that $\left\| \tilde{\theta}_t - \theta^\star \right\| < \delta$ for any fixed $\delta$ since $\tilde{\theta}_t$ must lie on the line segment between $\hat{\theta}_T$ and $\theta^\star$. Therefore,

$$
\begin{aligned}
\sum_{t=1}^{T} \left\| \ddot{s}_{t,\tilde{\theta}_t} \right\| &\leq \sum_{t=1}^{T} \frac{\pi_e(A_t \mid X_t)}{\mathbb{P}(A_t \mid X_t, \mathcal{H}_{t-1})} u_m(X_t, A_t, Y_t) \\
&\qquad + \sum_{t=1}^{T} \pi_e(A_t = a \mid X_t) \sum_{a=1}^{K} \left( 1 - \frac{\mathbb{1}_{A_t = a}}{\mathbb{P}(A_t = a \mid X_t, \mathcal{H}_{t-1})} \right) u_{\mathcal{F}}(X_t, a, f_t(X_t, a)) .
\end{aligned}
$$

We will now show that both the first and second terms converge to their expected values by Fact 1 to conclude the proof. For the first term, $\mathbb{E}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1})}u_m(X_t,A_t,Y_t)|\mathcal{H}_{t-1}\right] = \mathbb{E}_{\mathcal{P},\pi_e}\left[u_m(X_t,A_t,Y_t)\right]$. On the other hand, letting $w_t = \frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1})}$ implies that $w_t \leq C_1$ by Assumption 4. Therefore,

$$\begin{aligned}
\mathbb{E}[w_t^2 u_m(X_t,A_t,Y_t)^2] &\leq \mathbb{E}[C_1 w_t u_m(X_t,A_t,Y_t)^2] \\
&= C_1\mathbb{E}[w_t u_m(X_t,A_t,Y_t)^2] \\
&= C_1\mathbb{E}[\mathbb{E}[w_t u_m(X_t,A_t,Y_t)^2|\mathcal{H}_{t-1}]] \\
&= C_1\mathbb{E}[\mathbb{E}_{\mathcal{P},\pi_e}[u_m(X_t,A_t,Y_t)^2]] \\
&= C_1\mathbb{E}_{\mathcal{P},\pi_e}[u_m(X_t,A_t,Y_t)^2],
\end{aligned}$$

where the last term is bounded by a constant due to Assumption 7.

Similarly, we have

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\pi_e(A_t=a|X_t)\sum_{a=1}^{K}\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)u_{\mathcal{F}}(X_t,a,f_t(X_t,a)|\mathcal{H}_{t-1}\right] \\
&\leq \sum_{a=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)u_{\mathcal{F}}(X_t,a,f_t(X_t,a)|X_t,\mathcal{H}_{t-1}\right]|,\mathcal{H}_{t-1}\right] \\
&= \sum_{a=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)|X_t,\mathcal{H}_{t-1}\right]u_{\mathcal{F}}(X_t,a,f_t(X_t,a)|\mathcal{H}_{t-1}\right] \\
&= 0,
\end{aligned}$$

where the last line follows from the fact that $\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right)|X_t,\mathcal{H}_{t-1}\right] = 0$.

On the other hand, the variance can also be bounded. By Assumption 4 and the fact that $\pi_e(A_t=a|X_t) < 1$, we have that $\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right)^2 < (1-C_2)^2$. Therefore,

$$\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)^2 u_{\mathcal{F}}(X_t,a,f_t(X_t,a))^2\right] \leq (1-C_2)^2\mathbb{E}\left[u_{\mathcal{F}}(X_t,a,f_t(X_t,a))^2\right],$$

which is finite by Assumption 6.

**Lemma 2.** *Under the conditions of Theorem 1,* $\left\|\left(\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\dot{s}_{t,\hat{\theta}_T}\right)^{-1}\right\|_2 = O_p(1)$.

*Proof.* Let us decompose

$$\begin{aligned}
\dot{s}_{t,\hat{\theta}_T} = &\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1}\right)}\ddot{m}_{\hat{\theta}_T}(X_t,A_t,Y_t) \\
&+ \pi_e(A_t=a|X_t)\sum_{a=1}^{K}\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\ddot{m}_{\hat{\theta}_T}(X_t,a,f_t(X_t,a)).
\end{aligned}$$

It is sufficient to show the following statements:

1. $\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1},)}\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)$ is invertible and has bounded eigenvalues with probability tending to 1.
2. $\left\|\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1},)}\left(\ddot{m}_{\hat{\theta}_T}(X_t,A_t,Y_t)-\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right)\right\| = o_p(1)$
3. $\left\|\frac{1}{T}\sum_{t=1}^{T}\pi_e(A_t=a|X_t)\sum_{a=1}^{K}\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right)\ddot{m}(X_t,a,f_t(X_t,a))\right\| = o_p(1)$.

**Demonstrating (1)** First, note that

$$\begin{aligned}
&\mathbb{E}\left[\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)|\mathcal{H}_{t-1}\right] \\
&\quad=\hat{V}_t^{-1/2}\mathbb{E}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)|\mathcal{H}_{t-1}\right] \\
&\quad=\hat{V}_t^{-1/2}\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right]
\end{aligned}$$

By the weak law of large numbers, we can therefore conclude that

$$\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)=o_p(1)+\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right].$$

Rearranging terms, we can write this as

$$o_p(1)+\frac{1}{T}\sum_{t=1}^{T}(\hat{V}_t^{-1/2}-V_{t,\theta^\star}^{1/2})\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right]+\frac{1}{T}\sum_{t=1}^{T}V_{t,\theta^\star}^{-1/2}\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right].$$

For the second term, we note that:

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}(\hat{V}_t^{-1/2}-V_{t,\theta^\star}^{1/2})\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right]\right\|_{\text{op}} \\
&\quad\leq\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{V}_t^{-1/2}-V_{t,\theta^\star}^{1/2}\right\|_{\text{op}}\left\|\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right]\right\|_{\text{op}} \\
&\quad=o_p(1).
\end{aligned}$$

The last line follows because $\left\|\hat{V}_t^{-1/2}-V_{t,\theta^\star}^{1/2}\right\|_{\text{op}}=o_p(1)$ by assumption, while $\left\|\mathbb{E}_{\mathcal{P},p_e}[\ddot{m}_{\theta^\star}]\right\|_{\text{op}}$ is bounded by Assumption 3.

For the final term, we have that $-\mathbb{E}_{\mathcal{P},p_e}\left[\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right]\succeq H$ for some positive definite matrix $H$ and therefore it has bounded eigenvalues. Similarly, $\left\|\frac{1}{T}\sum_{t=1}^{T}V_{t,\theta^\star}^{-1/2}\right\|_{\text{op}}\leq\delta_{\min}^{-1/2}$ and $\left\|\left(\frac{1}{T}\sum_{t=1}^{T}V_{t,\theta^\star}^{-1/2}\right)^{-1}\right\|_{\text{op}}\geq\delta_{\max}^{-1/2}$. Therefore, the product of both terms has bounded eigenvalues and therefore the third term is invertible with bounded eigenvalues.

**Demonstrating (2)** We have by Taylor's theorem that for some $\tilde{\theta}$ on the line segment connecting $\hat{\theta}_T$ and $\tilde{\theta}_T$ that,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\left(\ddot{m}_{\hat{\theta}_T}(X_t,A_t,Y_t)-\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right) \\
&=\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\dddot{m}_{\tilde{\theta}}(X_t,A_t,Y_t).
\end{aligned}$$

Taking expectations, we have:

$$\begin{aligned}
&\left\|\mathbb{E}\left[\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\ddot{m}_{\tilde{\theta}}(X_t,A_t,Y_t)|\mathcal{H}_{t-1}\right]\right\|_1 \\
&\quad=\left\|\hat{V}_t^{-1/2}\mathbb{E}_{\mathcal{P},\pi_e}\left[\ddot{m}_{\tilde{\theta}}(X_t,A_t,Y_t)\right]\right\|_1 \\
&\quad\leq\left(\left\|\hat{V}_t^{-1/2}-V_{t,\theta^\star}\right\|_1+\|V_{t,\theta_\star}\|_1\right)\left\|\mathbb{E}_{\mathcal{P},\pi_e}\left[u_m(X_t,A_t,Y_t)\right]\right\|_1.
\end{aligned}$$

Therefore, by Fact 1 we have that

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\hat{V}_t^{-1/2}\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},\right)}\left(\ddot{m}_{\hat{\theta}_T}(X_t,A_t,Y_t)-\ddot{m}_{\theta^\star}(X_t,A_t,Y_t)\right)\right\| \\
&\leq o_p(1)+\frac{1}{T}\sum_{t=1}^{T}\left(\left\|\hat{V}_t^{-1/2}-V_{t,\theta^\star}^{-1/2}\right\|_1+\|V_{t,\theta_\star}\|_1\right) \\
&\qquad\cdot\left\|\mathbb{E}_{\mathcal{P},\pi_e}\left[u_m(X_t,A_t,Y_t)\right]\right\|_1 \\
&=o_p(1)+\frac{1}{T}\sum_{t=1}^{T}\|V_{t,\theta_\star}\|_1\left\|\mathbb{E}_{\mathcal{P},\pi_e}\left[u_m(X_t,A_t,Y_t)\right]\right\|_1.
\end{aligned}$$

However, the last term is also $o_p(1)$ because $V_{t,\theta^\star}$ has bounded eigenvalues and $\mathbb{E}_{\mathcal{P},\pi_e}[u_m]$ is bounded by Assumption 7.

**Demonstrating (3)** Note that for all $a\in\mathcal{A}$, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a\mid X_t,\mathcal{H}_{t-1})}\right)\ddot{m}_{\hat{\theta}_T}(X_t,a,f_t(X_t,a))\;\middle|\;\mathcal{H}_{t-1},X_t\right] \\
&\quad=\mathbb{E}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a\mid X_t,\mathcal{H}_{t-1})}\right)\;\middle|\;\mathcal{H}_{t-1},X_t\right]\mathbb{E}\left[\ddot{m}_{\hat{\theta}_T}(X_t,a,f_t(X_t,a))\;\middle|\;\mathcal{H}_{t-1},X_t\right]=0.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\pi_e(A_t=a|X_t)\sum_{a=1}^{K}\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\ddot{m}_{\hat{\theta}_T}(X_t,a,f_t(X_t,a))|\mathcal{H}_{t-1}\right]=0$$

By the martingale WLLN, we then have that

$$\frac{1}{T}\sum_{t=1}^{T}\pi_e(A_t=a|X_t)\sum_{a=1}^{K}\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\ddot{m}(X_t,a,f_t(X_t,a))=o_p(1).$$

□

**Lemma 3.** *Under the conditions of Theorem 1,* $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{V}_t^{-1/2}s_{t,\theta^\star}\xrightarrow{d}N(0,I_d)$.

*Proof.* Consider the term $Z_t:=c^T\hat{V}_t^{-1/2}s_{t,\theta^\star}$, for any $c\in\mathbb{R}^d$. By the Cramer-Wold device, it suffices to show that $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Z_t\xrightarrow{d}N(0,c^2)$. To do this, we demonstrate that $Z_t$ is a martingale difference sequence and then apply the martingale CLT of Dvoretzky (1972). In order to apply this fact, we need the following to be true:

- **Conditional Expectation** $\mathbb{E}[Z_t|\mathcal{H}_{t-1}]=0$ for all $t\in[T]$;
- **Conditional Variance** $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Z_t^2|\mathcal{H}_{t-1}]=\|c\|^2$;
- **Lindeberg Condition** $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[Z_t^2\mathbb{1}_{Z_t>\epsilon}|\mathcal{H}_{t-1}]\xrightarrow{p}0$.

Checking each condition separately:

**Conditional Expectation** By construction, we have that

$$\mathbb{E}[Z_t|\mathcal{H}_{t-1}] = \mathbb{E}[c^T\hat{V}_t^{-1/2}s_{t,\theta^\star}|\mathcal{H}_{t-1}] = c^T\hat{V}_t^{-1/2}\mathbb{E}[s_{t,\theta^\star}|\mathcal{H}_{t-1}],$$

based on the fact that $\hat{V}_t^{-1/2} \in \sigma(\mathcal{H}_{t-1})$. So it suffices to demonstrate that $\mathbb{E}[s_{t,\theta^\star}|\mathcal{H}_{t-1}] = 0$.

First, we will rewrite

$$\begin{aligned}\mathbb{E}[s_{t,\theta^\star}|\mathcal{H}_{t-1}] =& \mathbb{E}_{\mathcal{P},\pi_t}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1},X_t\right)}\dot{m}_{\theta^\star}(A_t,X_t,Y_t)|\mathcal{H}_{t-1}\right] \\ &+\sum_{a=1}^{K}\mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1}\right)}\right)\dot{m}_{\theta^\star}(a,X_t,f_t(a,X_t))|\mathcal{H}_{t-1}\right].\end{aligned}$$

For the first term, note that

$$\mathbb{E}_{\mathcal{P},\pi_t}\left[\frac{\pi_e(A_t=a|X_t)}{\mathbb{P}\left(A_t|X_t,\mathcal{H}_{t-1}\right)}\dot{m}_{\theta^\star}(A_t,X_t,Y_t)|\mathcal{H}_{t-1}\right] = \mathbb{E}_{\mathcal{P},\pi_e}\left[\dot{m}_{\theta^\star}(A_t,X_t,Y_t)|\mathcal{H}_{t-1}\right] = 0,$$

where the first equality follows because of a change of measure and the last line follows because $\theta^\star$ is the minimizer of a score equation.

For the second term, we note that

$$\begin{aligned}&\sum_{a=1}^{K}\mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\dot{m}_{\theta^\star}(a,X_t,f_t(a,X_t))|\mathcal{H}_{t-1}\right] \\ &=\sum_{a=1}^{K}\mathbb{E}_{\mathcal{P},\pi_t}\left[\mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\dot{m}_{\theta^\star}(a,X_t,f_t(a,X_t))|\mathcal{H}_{t-1},X_t\right]|\mathcal{H}_{t-1}\right] \\ &=\sum_{a=1}^{K}\mathbb{E}_{\mathcal{P},\pi_t}\left[\dot{m}_{\theta^\star}(a,X_t,f_t(a,X_t))\mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)|\mathcal{H}_{t-1},X_t\right]|\mathcal{H}_{t-1}\right].\end{aligned}$$

However, $\mathbb{E}_{\mathcal{P},\pi_t}\left[1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t|X_t,\mathcal{H}_{t-1})}|\mathcal{H}_{t-1},X_t\right] = 0$. Therefore, both the first and second terms are 0.

**Conditional Variance** Let us rewrite $\hat{V}_t^{-1/2} = \hat{V}_t^{-1/2} + V_{t,\theta^\star}^{-1/2} - V_{t,\theta^\star}^{-1/2}$. We then have that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[c^T\hat{V}_t^{-1/2}s_{t,\theta^\star}s_{t,\theta^\star}^T\hat{V}_t^{-1/2}c|\mathcal{H}_{t-1}\right] \\ &=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[c^T(\hat{V}_t^{-1/2}+V_{t,\theta^\star}^{-1/2}-V_{t,\theta^\star}^{-1/2})s_{t,\theta^\star}s_{t,\theta^\star}^T(\hat{V}_t^{-1/2}+V_{t,\theta^\star}^{-1/2}-V_{t,\theta^\star}^{-1/2})c|\mathcal{H}_{t-1}\right] \\ &=\frac{1}{T}\sum_{t=1}^{T}c^T(\hat{V}_t^{-1/2}+V_{t,\theta^\star}^{-1/2}-V_{t,\theta^\star}^{-1/2})\mathbb{E}\left[s_{t,\theta^\star}s_{t,\theta^\star}|\mathcal{H}_{t-1}^T\right](\hat{V}_t^{-1/2}+V_{t,\theta^\star}^{-1/2}-V_{t,\theta^\star}^{-1/2})c \\ &=c^T\frac{1}{T}\sum_{t=1}^{T}\Big[V_{t,\theta^\star}^{-1/2}V_{t,\theta^\star}V_{t,\theta^\star}^{-1/2}-\left(V_{t,\theta^\star}^{-1/2}-\hat{V}_t^{-1/2}\right)V_{t,\theta^\star}^{1/2}-V_{t,\theta^\star}^{1/2}\left(V_{t,\theta^\star}^{-1/2}-\hat{V}_t^{-1/2}\right) \\ &\qquad+\left(V_{t,\theta^\star}^{-1/2}-\hat{V}_t^{-1/2}\right)V_{t,\theta^\star}\left(V_{t,\theta^\star}^{-1/2}-\hat{V}_t^{-1/2}\right)\Big]c \\ &=\|c\|^2+c^T\left(\frac{1}{T}\sum_{t=1}^{T}A_t\right)c,\end{aligned}$$

where

$$A_t = \left(V_{t,\theta^\star}^{-1/2} - \hat{V}_t^{-1/2}\right) V_{t,\theta^\star}^{1/2} + V_{t,\theta^\star}^{1/2}\left(V_{t,\theta^\star}^{-1/2} - \hat{V}_t^{-1/2}\right) + \left(V_{t,\theta^\star}^{-1/2} - \hat{V}_t^{-1/2}\right) V_{t,\theta^\star}\left(V_{t,\theta^\star}^{-1/2} - \hat{V}_t^{-1/2}\right).$$

It suffices to show that $\|A_t\|_{\text{op}} \xrightarrow{p} 0$. However, this is true because $\left\|V_{t,\theta^\star}^{-1/2} - \hat{V}_t^{-1/2}\right\|_{\text{op}} \xrightarrow{p} 0$ by assumption, and $V_{t,\theta^\star}^{1/2}$ has bounded eigenvalues.

**Lindeberg Condition** Fix any $\epsilon > 0$. Note that in general $\mathbb{1}_{|Z_t|\geq\epsilon} \leq \frac{Z_t^2}{\epsilon^2}$. We therefore have that,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[c^T\hat{V}_t^{-1/2}s_{t,\theta^\star}s_{t,\theta^\star}^T\hat{V}_t^{-1/2}c\mathbb{1}_{|c^T\frac{1}{\sqrt{T}}\hat{V}_t^{-1/2}s_{t,\theta^\star}|\geq\epsilon}|\mathcal{H}_{t-1}\right] \\
&\quad\leq \frac{1}{\epsilon^2T^2}\sum_{t=1}^{T}\mathbb{E}\left[\left(c^T\hat{V}_t^{-1/2}s_{t,\theta^\star}s_{t,\theta^\star}^T\hat{V}_t^{-1/2}c\right)^2|\mathcal{H}_{t-1}\right] \\
&\quad= \frac{1}{\epsilon^2T^2}\sum_{t=1}^{T}c^T\hat{V}_t^{-1/2}\mathbb{E}\left[(c^Ts_{t,\theta^\star})^2(c^T\hat{V}_t^{-1/2}s_{t,\theta^\star})^2|\mathcal{H}_{t-1}\right].
\end{aligned}$$

It is sufficient to show that $\mathbb{E}\left[(c^Ts_{t,\theta^\star})^2(c^T\hat{V}_t^{-1/2}s_{t,\theta^\star})^2|\mathcal{H}_{t-1}\right]$ is bounded. By Cauchy-Schwarz, we have:

$$\begin{aligned}
\mathbb{E}\left[(c^Ts_{t,\theta^\star})^2(c^T\hat{V}_t^{-1/2}s_{t,\theta^\star})^2|\mathcal{H}_{t-1}\right] &\leq \mathbb{E}\left[(c^Ts_{t,\theta^\star})^4|\mathcal{H}_{t-1}\right]^{1/2}\mathbb{E}\left[(c^T\hat{V}_t^{-1/2}s_{t,\theta^\star})^4|\mathcal{H}_{t-1}\right]^{1/2} \\
&\leq 2c^4\mathbb{E}\left[\|s_{t,\theta^\star}\|^4|\mathcal{H}_{t-1}\right]\left\|\hat{V}_t^{-1/2}\right\|_{\text{op}}^2.
\end{aligned}$$

By assumption, we know that $\|\widehat{V}_t^{-1/2}\|_{\text{op}}^2$ is bounded. Therefore, it is sufficient to demonstrate that $\mathbb{E}\big[\|s_{t,\theta^\star}\|^4 \mid \mathcal{H}_{t-1}\big] = O_p(1)$ to conclude the proof. Recall that

$$s_{t,\theta^\star} = \sum_{a=1}^{K}\pi_e(a \mid X_t)\left\{\dot{m}_{\theta^\star}(X_t,a,f_t(X_t,a)) + \mathbb{1}_{\{A_t=a\}}\frac{\dot{m}_{\theta^\star}(X_t,a,Y_t) - \dot{m}_{\theta^\star}(X_t,a,f_t(X_t,a))}{\mathbb{P}(A_t=a \mid X_t,\mathcal{H}_{t-1})}\right\}.$$

Define

$$U_{t,a} := \dot{m}_{\theta^\star}(X_t,a,f_t(X_t,a)), \qquad V_{t,a} := \mathbb{1}_{\{A_t=a\}}\frac{\dot{m}_{\theta^\star}(X_t,a,Y_t) - \dot{m}_{\theta^\star}(X_t,a,f_t(X_t,a))}{\mathbb{P}(A_t=a \mid X_t,\mathcal{H}_{t-1})}.$$

Since $\pi_e(a \mid X_t) \in [0,1]$ and $K < \infty$, the inequality

$$\left\|\sum_{a=1}^{K}w_az_a\right\|^4 \leq K^3\sum_{a=1}^{K}\|z_a\|^4 \quad \text{for } |w_a| \leq 1,$$

applied with $w_a = \pi_e(a \mid X_t)$ and $z_a = U_{t,a} + V_{t,a}$ yields

$$\begin{aligned}
\|s_{t,\theta^\star}\|^4 &\leq K^3\sum_{a=1}^{K}\|U_{t,a} + V_{t,a}\|^4 \\
&\leq 8K^3\sum_{a=1}^{K}\left(\|U_{t,a}\|^4 + \|V_{t,a}\|^4\right),
\end{aligned}$$

where the second inequality follows from $(x+y)^4 \leq 8(x^4+y^4)$.

We now bound the two terms separately. For $U_{t,a}$, Assumption 6 implies that

$$\|U_{t,a}\|^4 \leq u_F(X_t, a, f_t(X_t, a))^4,$$

where $u_F$ is an envelope satisfying $\mathbb{E}[u_F(X_t, a, f_t(X_t, a))^4] < \infty$. Consequently,

$$\mathbb{E}\big[\|U_{t,a}\|^4 \mid \mathcal{H}_{t-1}\big] \leq \mathbb{E}\big[u_F(X_t, a, f_t(X_t, a))^4 \mid \mathcal{H}_{t-1}\big] = O_p(1).$$

For $V_{t,a}$, bounded importance ratios (Assumption 4) imply that $\mathbb{P}(A_t = a \mid X_t, \mathcal{H}_{t-1})^{-1} \leq C_1$, and therefore

$$\begin{aligned}\|V_{t,a}\|^4 &\leq C_1^4 \big\|\dot{m}_{\theta^\star}(X_t, a, Y_t) - \dot{m}_{\theta^\star}(X_t, a, f_t(X_t, a))\big\|^4 \\ &\leq 8C_1^4\Big(\|\dot{m}_{\theta^\star}(X_t, a, Y_t)\|^4 + \|U_{t,a}\|^4\Big),\end{aligned}$$

where the second line again uses $(x+y)^4 \leq 8(x^4+y^4)$. Taking conditional expectations and using the fourth-moment bound on $\dot{m}_{\theta^\star}(X_t, a, Y_t)$ from Assumption 3, together with the bound on $\mathbb{E}[\|U_{t,a}\|^4 \mid \mathcal{H}_{t-1}]$, yields

$$\mathbb{E}\big[\|V_{t,a}\|^4 \mid \mathcal{H}_{t-1}\big] = O_p(1).$$

Combining these bounds and summing over the fixed number of actions $K$ gives

$$\mathbb{E}\big[\|s_{t,\theta^\star}\|^4 \mid \mathcal{H}_{t-1}\big] \leq 8K^3 \sum_{a=1}^{K} \big(\mathbb{E}[\|U_{t,a}\|^4 \mid \mathcal{H}_{t-1}] + \mathbb{E}[\|V_{t,a}\|^4 \mid \mathcal{H}_{t-1}]\big) = O_p(1),$$

as required.

□

**Lemma 4.** *Under the conditions of Theorem 1,* $\left\|\hat{\theta}_T - \theta^\star\right\|_1 = o_p(1)$.

*Proof.* We note that by Assumption 3, there exists some $\delta_2 > 0$ such that $\left\|\mathbb{E}[\dot{m}_{\hat{\theta}_T}]\right\|_1 > \delta_2$ will imply that $\left\|\hat{\theta}_T - \theta^\star\right\| \geq \epsilon$. Therefore,

$$\mathbb{P}\left(\left\|\hat{\theta}_T - \theta^\star\right\| \geq \epsilon\right) \leq \mathbb{P}\left(\left\|\mathbb{E}\left[\dot{m}_{\hat{\theta}_T}(X_t, A_t, Y_t)\right]\right\|_1 > \delta_2\right).$$

However, we know that for all $\theta \in \Theta$,

$$\begin{aligned}\mathbb{E}[s_{t,\theta} \mid \mathcal{H}_{t-1}] &= \sum_{a=1}^{K} \mathbb{E}\Bigg[\pi_e(a \mid X_t)\Bigg(m_\theta(a, X_t, f_t(a, X_t)) \\ &\qquad + \mathbb{1}_{\{A_t=a\}} \frac{m_\theta(X_t, A_t, Y_t) - m_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}(A_t = a \mid X_t, \mathcal{H}_{t-1})}\Bigg)\Bigg| \mathcal{H}_{t-1}\Bigg] \\ &= \mathbb{E}_{\mathcal{P},\pi_t}\left[\frac{\pi_e(A_t \mid X_t)}{\mathbb{P}(A_t \mid X_t, \mathcal{H}_{t-1})}\, m_\theta(X_t, A_t, Y_t)\right] \\ &\quad + \sum_{a=1}^{K} \mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1 - \frac{\mathbb{1}_{\{A_t=a\}}}{\mathbb{P}(A_t = a \mid X_t, \mathcal{H}_{t-1})}\right) m_\theta(a, X_t, f_t(a, X_t))\right] \\ &= \mathbb{E}_{\pi_e}[m_\theta(X_t, A_t, Y_t)]\,.\end{aligned}$$

Note that,

$$\begin{aligned}\left\|\mathbb{E}\left[\hat{V}_t^{-1/2} s_{t,\theta} | \mathcal{H}_{t-1}\right]\right\|_1 &= \left\|\hat{V}_t^{-1/2} \mathbb{E}\left[s_{t,\theta} | \mathcal{H}_{t-1}\right]\right\|_1 \\ &= \left\|\hat{V}_t^{-1/2} \mathbb{E}\left[\dot{m}_\theta(X_t, A_t, Y_t)\right]\right\|_1 \\ &\leq \delta_{\max} \left\|\mathbb{E}\left[\dot{m}_{\hat{\theta}_T}(X_t, A_t, Y_t)\right]\right\|_1.\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\left\|\mathbb{E}\left[\dot{m}_{\hat{\theta}_T}(X_t, A_t, Y_t)\right]\right\|_1 > \delta_2\right) = \mathbb{P}\left(\delta_{\max}\left\|\mathbb{E}\left[\dot{m}_{\hat{\theta}_T}(X_t, A_t, Y_t)\right]\right\|_1 > \delta_{\max}\delta_2\right)$$
$$\leq \mathbb{P}\left(\left\|\mathbb{E}\left[\hat{V}_t^{-1/2} s_{t,\theta} | \mathcal{H}_{t-1}\right]\right\|_1 > \delta_{\max}\delta_2\right).$$

By assumption, $\frac{1}{T}\sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} = o_p(1/\sqrt{T})$. Therefore,

$$\mathbb{P}\left(\left\|\mathbb{E}\left[\dot{m}_{\hat{\theta}_T}(X_t, A_t, Y_t)\right]\right\|_1 > \delta_2\right)$$
$$\leq \mathbb{P}\left(\left\|\frac{1}{T}\sum_{t=1}^{T}\left(\hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} - \mathbb{E}\left[\hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} | \mathcal{H}_{t-1}\right]\right)\right\|_1 > \delta_{\max}\delta_2 + o_p(1)\right).$$

Therefore, it is sufficient to show that

$$\sup_{\theta\in\Theta}\left\|\frac{1}{T}\left(\sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\theta} - \mathbb{E}\left[\hat{V}_t^{-1/2} s_{t,\theta} | \mathcal{H}_{t-1}\right]\right)\right\|_1 \xrightarrow{p} 0.$$

We show this is true for each component individually. Define $g_\theta(x, a, y) := e_j^T \hat{V}_t^{-1/2} \dot{m}_\theta(x, a, y)$. By Assumption 5, for any $\epsilon > 0$, $N_{[]}(\epsilon, \dot{\mathcal{M}}_\Theta, L_2(\mathcal{P}, \pi_e)) < \infty$. Since $\hat{V}_t^{-1/2}$ has bounded eigenvalues and $e_j$ is a unit vector, $N_{[]}(\epsilon, \{g_\theta\}, L_2(\mathcal{P}, \pi_e)) < \infty$. Now note that

$$e_j^T \hat{V}_t^{-1/2} s_{t,\theta} = \sum_{a=1}^{K} \pi_e(A_t = a|X_t)\Big(g_\theta(a, X_t, f_t(a, X_t))$$
$$+ \mathbb{1}_{A_t=a}\frac{g_\theta(X_t, A_t, Y_t) - g_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\Big),$$

and we can immediately apply Lemma 1 to conclude the proof. □

### B.2 Proof of Lemma 1

*Proof.* We will decompose $R_t(\theta)$ into two terms and consider them separately.

$$R_t(\theta) = \sum_{a=1}^{K} \pi_e(A_t = a|X_t)\Big(g_\theta(a, X_t, f_t(a, X_t))$$
$$+ \mathbb{1}_{A_t=a}\frac{g_\theta(X_t, A_t, Y_t) - g_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\Big)$$
$$= \underbrace{\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} g_\theta(X_t, A_t, Y_t)}_{R_t^{(1)}(\theta)}$$
$$+ \underbrace{\sum_{a=1}^{K} \pi_e(A_t = a|X_t)\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\right) g_\theta(X_t, a, f_t(X_t, a))}_{R_t^{(2)}(\theta)}.$$

**Showing first term converges** We have by assume that for any $\delta > 0$, there exists a set of brackets $B_\delta$ of finite size. This means that for every $\theta$, there exists a pair of functions $(l, u) \in B_\delta$ such that:

1. $l(x, a, y) \leq g_\theta(x, a, y) \leq u(x, a, y)$ for every $a \in \mathcal{A}$, $(x, y)$ in the support of $\mathcal{P}$;

2. $\mathbb{E}_{\pi_e}\left[u(X_t, A_t, Y_t) - l(X_t, A_t, Y_t)\right] \leq \delta$;
3. $\mathbb{E}\left[u(X_t, A_t, Y_t)^2\right] < \infty$ and $\mathbb{E}\left[l(X_t, A_t, Y_t)^2\right] < \infty$.

Fixing $\delta > 0$, we then have the following:

$$
\begin{aligned}
&\sup_{\theta\in\Theta} R_t^{(1)}(\theta) - \mathbb{E}[R_t^{(1)}(\theta)|\mathcal{H}_{t-1}] \\
&\quad= \sup_{\theta\in\Theta} \frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} g_\theta(X_t, A_t, Y_t) \\
&\qquad - \mathbb{E}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} g_\theta(X_t, A_t, Y_t) \mid \mathcal{H}_{t-1}\right] \\
&\quad\leq \max_{(l,u)\in B_\delta} \left\{ \frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} u(X_t, A_t, Y_t) \right. \\
&\qquad \left. - \mathbb{E}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} l(X_t, A_t, Y_t) \mid \mathcal{H}_{t-1}\right] \right\} \\
&\quad= \max_{(l,u)\in B_\delta} \left\{ \frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} u(X_t, A_t, Y_t) \right. \\
&\qquad - \mathbb{E}\left[\frac{\pi_e(A_t|X_t)}{\mathbb{P}\left(A_t|X_t, \mathcal{H}_{t-1}\right)} u(X_t, A_t, Y_t) \mid \mathcal{H}_{t-1}\right] \\
&\qquad \left. + \mathbb{E}_{\mathcal{P},\pi_e}\left[u(X_t, A_t, Y_t) - l(X_t, A_t, Y_t)\right] \right\}.
\end{aligned}
$$

First, note that $\mathbb{E}_{\mathcal{P},\pi_e}\left[u(X_t, A_t, Y_t) - l(X_t, A_t, Y_t)\right] \leq \delta$ by assumption. For the other term, let $w_t = \frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})}$. Now, note that

$$
\begin{aligned}
&\max_{(l,u)\in B_\delta} w_t u(X_t, A_t, Y_t) - \mathbb{E}\left[w_t u(X_t, A_t, Y_t)|\mathcal{H}_{t-1}\right] \\
&\qquad \leq \sum_{(l,u)\in B_\delta} \left|w_t u(X_t, A_t, Y_t) - \mathbb{E}\left[w_t u(X_t, A_t, Y_t)|\mathcal{H}_{t-1}\right]\right|.
\end{aligned}
$$

Putting this all together, we have that

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} R_t^{(1)}(\theta) - \mathbb{E}[R_t^{(1)}(\theta)|\mathcal{H}_{t-1}] \\
&\qquad \leq \delta + \frac{1}{T}\sum_{t=1}^{T} \sum_{(l,u)\in B_\delta} \left|w_t u(X_t, A_t, Y_t) - \mathbb{E}\left[w_t u(X_t, A_t, Y_t)|\mathcal{H}_{t-1}\right]\right|.
\end{aligned}
$$

Since $|B_\delta|$ is finite, it is sufficient to show that $w_t u(X_t, A_t, Y_t) - \mathbb{E}\left[w_t u(X_t, A_t, Y_t)|\mathcal{H}_{t-1}\right] = o_p(1)$ for all $(l, u) \in B_\delta$. To do this, we invoke the martingale weak law of large numbers from Fact 1. Following Remark 4,it is sufficient to show that $E_{\mathcal{P},\pi_t}[w_t^2 u(X_t, A_t, Y_t)^2]$ is bounded by a constant for all $t$ to demonstrate the third criterion. For this, note that because of Assumption 4, there exists a constant $C_1 > 0$ such that $\frac{\pi_e(A_t|X_t)}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})} \leq \frac{1}{\mathbb{P}(A_t|X_t, \mathcal{H}_{t-1})} < C_1$. Then

$$
\begin{aligned}
\mathbb{E}[w_t^2 u(X_t, A_t, Y_t)^2] &\leq \mathbb{E}[C_1 w_t u(X_t, A_t, Y_t)^2] \\
&= C_1 \mathbb{E}[w_t u(X_t, A_t, Y_t)^2] \\
&= C_1 \mathbb{E}[\mathbb{E}[w_t u(X_t, A_t, Y_t)^2|\mathcal{H}_{t-1}]] \\
&= C_1 \mathbb{E}[\mathbb{E}_{\mathcal{P},\pi_e}[u(X_t, A_t, Y_t)^2]] \\
&= C_1 \mathbb{E}_{\mathcal{P},\pi_e}[u(X_t, A_t, Y_t)^2] < \infty.
\end{aligned}
$$

Therefore, Fact 1 applies, and we can conclude that $\frac{1}{T}\sum_{t=1}^{T} R_t^{(1)}(\theta) - \mathbb{E}[R_t^{(1)}(\theta)|\mathcal{H}_{t-1}] \leq \delta + o_p(1)$. Taking $\delta \to 0$ concludes the proof.

**Showing second term converges** We have that $\pi_e(A_t|X_t) < 1$ and $|\mathcal{A}| < \infty$ so it is sufficient to show that for all $a \in \mathcal{A}$

$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T}\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\right) g_\theta(X_t, a, f_t(X_t, a)) = o_p(1).$$

We will first show pointwise convergence and then use a covering argument to get the uniform result. Define $Z_{t,\theta} = \left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right) g_\theta(X_t, a, f_t(X_t, a))$. Note that for all $\theta$,

$$\mathbb{E}[Z_{t,\theta}|\mathcal{H}_{t-1}] = \mathbb{E}\left[\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\right) g_\theta(X_t, a, f_t(X_t, a))|\mathcal{H}_{t-1}\right] = 0.$$

By Assumption 6, $\sup_\theta \|g_\theta(X_t, a, f_t(X_t))\|_1 \leq u_{\mathcal{F}}(X_t, a, f_t(X_t, a))$ for all $X_t$ and $a$. Note that it is sufficient to show that $\mathbb{E}\left[\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right)^2 u_{\mathcal{F}}(X_t, a, f_t(X_t, a))^2\right]$ is bounded by a constant for all $t$ in order for Fact 1 to be applied. However, we note that by Assumption 4 and the fact that $\pi_e(A_t = a|X_t) < 1$ that $\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}(A_t=a|X_t,\mathcal{H}_{t-1})}\right)^2 < \max((1 - C_1)^2, 1)$. Therefore,

$$\begin{aligned}\mathbb{E}&\left[\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\right)^2 u_{\mathcal{F}}(X_t, a, f_t(X_t, a))^2\right] \\ &\leq \max((1 - C_1)^2, 1)\mathbb{E}\left[u_{\mathcal{F}}(X_t, a, f_t(X_t, a))^2\right],\end{aligned}$$

which is finite by assumption. This implies that for each $\theta$, $\frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta} = 0$.

Because $\Theta$ is a bounded parameter space, for any $\epsilon > 0$, we can cover $\Theta \subseteq \bigcup_{j=1}^{M} \Theta_j$. Here, each $\Theta_j$ is an $\epsilon$-ball with centers denoted $\theta_1, ..., \theta_k$. For any $\theta$, we can decompose $Z_{t,\theta} = Z_{t,\theta} + Z_{t,\theta_j}$ for some $\theta_j$ such that $|\theta - \theta_j| < \epsilon$. Now, let us write out

$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta} \leq \max_{1\leq j\leq M} |\frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta_j}| + |\frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta} - Z_{t,\theta_j}|.$$

For the first term, we know that $\max_{1\leq j\leq M}|\frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta}| = o_p(1)$ because $M$ is finite for each $\epsilon$. We therefore just need to show that $\sup_{\theta\in\Theta_j}|\frac{1}{T}\sum_{t=1}^{T} Z_{t,\theta}| = o_p(1)$ to finish the proof.

First, note that $Z_{t,\theta} - Z_{t,\theta_j}$ is a martingale difference sequence. Therefore, it is sufficient to show that $\mathbb{E}\left[\left\|Z_{t,\theta} - Z_{t,\theta_j}\right\|^2\right]$ is bounded to conclude the proof. This is true becaue

$$\begin{aligned}\mathbb{E}\left[\left\|Z_{t,\theta} - Z_{t,\theta_j}\right\|^2\right] &\leq \mathbb{E}\Bigg[\left(1 - \frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t = a|X_t, \mathcal{H}_{t-1}\right)}\right)^2 \\ &\qquad \cdot \left(g_\theta(X_t, a, f_t(X_t, a)) - g_{\theta_j}(X_t, a, f_t(X_t, a))\right)^2\Bigg] \\ &\leq \max((1 - C_1)^2, 1)\epsilon^2.\end{aligned}$$

□

### B.3 Proof of Proposition 1

*Proof.* For ease of notation, we denote the quantities $\pi_t(a \mid X_t) := \mathbb{P}(A_t \mid X_t, \mathcal{H}_{t-1})$, $\mu^\star(X_t, a) := \mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a)) \mid X_t]$, and $v^\star(X_t, a) := \mathrm{Var}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a)) \mid X_t]$.

Let us rewrite using the law of total variance:

$$\mathrm{Var}\left(s_{t,\theta^\star} \mid \mathcal{H}_{t-1}\right) = \mathrm{Var}\Big(\mathbb{E}\left[s_{t,\theta^\star} \mid \mathcal{H}_{t-1}, X_t\right] \mid \mathcal{H}_{t-1}\Big) + \mathbb{E}\Big[\mathrm{Var}\left(s_{t,\theta^\star} \mid \mathcal{H}_{t-1}, X_t\right) \mid \mathcal{H}_{t-1}\Big].$$

Now, analyzing each of these terms separately:

$$\begin{aligned}
\mathbb{E}\left[s_{t,\theta^\star} \mid \mathcal{H}_{t-1}, X_t\right] &= \mathbb{E}\Bigg[\sum_{a=1}^{K} \pi_e(A_t = a \mid X_t)\Big(\dot{m}_{\theta^\star}(X_t, a, f_t(X_t, a)) \\
&\qquad + \mathbb{1}_{A_t=a}\frac{\dot{m}_{\theta^\star}(X_t, a, Y_t) - \dot{m}_{\theta^\star}(X_t, a, f_t(X_t, a)}{\pi_t(a \mid X_t)}\Big) \,\Big|\, \mathcal{H}_{t-1}, X_t\Bigg] \\
&= \sum_{a=1}^{K} \pi_e(A_t = a \mid X_t)\mathbb{E}\left[\frac{\mathbb{1}_{A_t=a}}{\pi_t(a \mid X_t)}\dot{m}_{\theta^\star}(X_t, a, Y_t(a)) \mid X_t, \mathcal{H}_{t-1}\right] \\
&\qquad + \sum_{a=1}^{K} \pi_e(A_t = a \mid X_t)\mathbb{E}\Bigg[\left(1 - \frac{\mathbb{1}_{A_t=a}}{\pi_t(a \mid X_t)}\right) \\
&\qquad\qquad \cdot \dot{m}_{\theta^\star}(X_t, a, f_t(X_t, a)) \,\Big|\, X_t, \mathcal{H}_{t-1}\Bigg] \\
&= \sum_{a=1}^{K} \pi_e(A_t = a \mid X_t)\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, a, Y_t(a)) \mid X_t\right] \\
&= \mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right].
\end{aligned}$$

On the other hand, note that $\dot{m}_\theta(a, X_t, f_t(a, X_t)) \in \sigma(X_t, \mathcal{H}_{t-1})$. Therefore, we can rewrite the conditional variance given $X_t$ as:

$$\begin{aligned}
\mathrm{Var}\left(s_{t,\theta^\star} \mid \mathcal{H}_{t-1}, X_t\right) &= \mathrm{Var}\left(\sum_{a=1}^{K} \pi_e(a \mid X_t)\frac{\mathbb{1}_{A_t=a}}{\pi_t(a \mid X_t)}\dot{m}_{\theta^\star}(X_t, a, Y_t) \,\Big|\, \mathcal{H}_{t-1}, X_t\right) \\
&= \sum_{a=1}^{K} \frac{\pi_e(a \mid X_t)^2}{\pi_t(a|X_t)}\mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a))\dot{m}_{\theta^\star}(X_t, a, Y_t(a))^T \mid X_t] \\
&\qquad - \Big(\sum_{a=1}^{K} \pi_e(a \mid X_t)\mu^\star(X_t, a)\Big)\Big(\sum_{a=1}^{K} \pi_e(a \mid X_t)\mu^\star(X_t, a)\Big)^T \\
&= \sum_{a=1}^{K} \frac{\pi_e(a \mid X_t)^2}{\pi_t(a|X_t)}\mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a))\dot{m}_{\theta^\star}(X_t, a, Y_t(a))^T \mid X_t] \\
&\qquad - \mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T.
\end{aligned}$$

Finally, combining the two terms, we obtain the full decomposition:

$$\begin{aligned}
\mathrm{Var}(s_{t,\theta^\star} \mid \mathcal{H}_{t-1}) &= \mathrm{Var}\Big(\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\Big) \\
&\quad + \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi_e(a \mid X_t)^2}{\pi_t(a|X_t)}\mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a))\dot{m}_{\theta^\star}(X_t, a, Y_t(a))^T \mid X_t] \mid \mathcal{H}_{t-1}\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T\right].
\end{aligned}$$

□

### B.4 Proof of Proposition 2

For ease of notation, let us write $\nu_t^\star(X_t) := \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)|X_t\right]$. We will tackle this proof in three parts:

1. Showing
$$\frac{1}{n-1}\left(\sum_{X_i \in \tilde{\mathcal{X}}} \hat{\nu}_t(X_i) - \bar{\nu}\right)\left(\sum_{X_i \in \tilde{\mathcal{X}}} \hat{\nu}_t(X_i) - \bar{\nu}\right)^T - \mathrm{Var}(\mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)|X_t\right]) = o_p(1);$$

2. Showing
$$\frac{1}{n}\sum_{X_i \in \tilde{\mathcal{X}}}\sum_{a=1}^{K} \frac{\pi_e(A_t = a \mid X_i)^2\, h_t(X_i, a)}{\mathbb{P}(A_t = a \mid X_i, \mathcal{H}_{t-1})} - \mathbb{E}\left[\sum_{a=1}^{K} \frac{\pi_e(a \mid X_t)^2}{\pi_t(a|X_t)} \mathbb{E}[\dot{m}_{\theta^\star}(X_t, a, Y_t(a))\dot{m}_{\theta^\star}(X_t, a, Y_t(a))^T \mid X_t] \mid \mathcal{H}_{t-1}\right] = o_p(1);$$

3. Showing
$$\frac{1}{n}\sum_{X_i \in \tilde{\mathcal{X}}} \hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T - \mathbb{E}\Big[\mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right] \cdot \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T \mid \mathcal{H}_{t-1}\Big] = o_p(1).$$

Adding term (1)-(3) together will demonstrate that $\left\|\hat{V}_t - V_{t,\theta^\star}\right\|_{\mathrm{op}} \xrightarrow{p} 0$.

**Showing (1)** First, we will show that $\hat{\nu}_t(X_i) - \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)|X_t\right] \xrightarrow{p} 0$. To see this, write

$$\begin{aligned}
\hat{\nu}_t(X_i) - \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)|X_t\right] &= \sum_{a=1}^{K} \pi_e(a|X_t) g_t(a, X_i) - \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_\theta(X_t, A_t, Y_t)\right] \\
&= \sum_{a=1}^{K} \pi_e(a|X_t) g_t(a, X_i) - \sum_{a=1}^{K} \pi_e(a|X_t)\mathbb{E}\left[\dot{m}_\theta(X_t, a, Y_t)|X_t\right] \\
&= \sum_{a=1}^{K} \pi_e(a|X_t)\left(g_t(a, X_i) - \mathbb{E}\left[\dot{m}_\theta(X_t, A_t, Y_t)|A_t = a, X_t\right]\right) \\
&\leq K\left(g_t(a, X_i) - \mathbb{E}\left[\dot{m}_\theta(X_t, A_t, Y_t)|A_t = a, X_t\right]\right) \\
&= o_p(1).
\end{aligned}$$

The final line holds by Assumption 9.

Let us consider the first term in the expression $\frac{1}{n}\left(\sum_{X_i \in \tilde{X}} \hat{\nu}_t(X_i) - \bar{\nu}\right)\left(\sum_{X_i \in \tilde{X}} \hat{\nu}_t(X_i)^2 - \bar{\nu}\right)^T$. Let us substitute $\hat{\nu}_t(X_i) = \hat{\nu}_t(X_i) - \nu_t^\star(X_t) + \nu_t^\star(X_t)$. We then obtain the result that the above expression is equal

to

$$\frac{1}{n}\sum_{X_i\in\tilde{X}}\nu_t^\star(X_t)\nu_t^\star(X_t)^T-\left(\frac{1}{n}\sum_{X_i\in\tilde{X}}\nu_t^\star(X_t)\right)\left(\frac{1}{n}\sum_{X_i\in\tilde{X}}\nu_t^\star(X_t)\right)^T+o_p(1),$$

which converges to $\mathrm{Var}\left(\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_\theta(X_t,A_t,Y_t)|X_t\right]\right)$ by the weak law of large numbers.

**Showing (2)** Now, let us consider the expression

$$\begin{aligned}&\frac{1}{n}\sum_{X_i\in\tilde{X}}\sum_{a=1}^{K}\frac{\pi_e(A_t=a\mid X_t)^2}{\mathbb{P}(A_t=a\mid X_t,\mathcal{H}_{t-1})}\,h_t(X_i,a)\\&\quad-\mathbb{E}\left[\sum_{a=1}^{K}\frac{\pi_e(A_t=a\mid X_t)^2}{\mathbb{P}(A_t=a\mid X_t,\mathcal{H}_{t-1})}\,\mathbb{E}\Big[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T\;\Big|\;X_t,A_t=a\Big]\;\Big|\;\mathcal{H}_{t-1}\right].\end{aligned}$$

This is simply equal to

$$\begin{aligned}&\frac{1}{n}\sum_{X_i\in\tilde{X}}\sum_{a=1}^{K}\pi_e(A_t=a|X_t)^2\frac{h_t(X_i,a)-\mathbb{E}\left[\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]\right]}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\\&\qquad\le\frac{C_1}{n}\left[\sum_{X_i\in\tilde{X}}\sum_{a=1}^{K}h_t(X_i,a)\right]\\&\qquad\qquad-C_1\mathbb{E}\left[\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]\right].\end{aligned}$$

To show that this expression is $o_p(1)$, it is sufficient to show that for all $a\in\mathcal{A}$,

$$\frac{1}{n}\sum_{X_i\in\tilde{X}}h_t(X_i,a)-\mathbb{E}\left[\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]\right]=o_p(1).$$

Note that by Assumption 9,

$$\begin{aligned}\frac{1}{n}\sum_{X_i\in\tilde{X}}h_t(X_i,a)&=\frac{1}{n}\sum_{X_i\in\tilde{X}}h_t(X_i,a)+\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]-\\&\qquad\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]\\&=o_p(1)+\frac{1}{n}\sum_{X_i\in\tilde{X}}\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right].\end{aligned}$$

Therefore, $\frac{1}{n}\sum_{X_i\in\tilde{X}}h_t(X_i,a)-\mathbb{E}\left[\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t,A_t,Y_t)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T|X_t,A_t=a\right]\right]=o_p(1)$ by the weak law of large numbers.

**Showing (3)** Finally, consider the expression

$$\hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T-\Big(\sum_{a=1}^{K}\pi_e(a\mid X_t)\mu^\star(X_t,a)\Big)\Big(\sum_{a=1}^{K}\pi_e(a\mid X_t)\mu^\star(X_t,a)\Big)^T.\tag{10}$$

Let $\Delta_t(x):=\widehat{\nu}_t(x)-\nu_t^\star(x)$ which we already showed was $o_p(1)$. Now, we can write:

$$\begin{aligned}\widehat{\nu}_t(X_i)\widehat{\nu}_t(X_i)^\top&=\big(\Delta_t(X_i)+\nu_t^\star(X_i)\big)\big(\Delta_t(X_i)+\nu_t^\star(X_i)\big)^\top\\&=\Delta_t(X_i)\Delta_t(X_i)^\top\;+\;\Delta_t(X_i)\nu_t^\star(X_i)^\top\;+\;\nu_t^\star(X_i)\Delta_t(X_i)^\top\;+\;\nu_t^\star(X_i)\nu_t^\star(X_i)^\top.\end{aligned}$$

Subtracting the last term on both sides yields the identity

$$\widehat{\nu}_t(X_i)\widehat{\nu}_t(X_i)^\top - \nu_t^\star(X_i)\nu_t^\star(X_i)^\top = \Delta_t(X_i)\Delta_t(X_i)^\top + \Delta_t(X_i)\nu_t^\star(X_i)^\top + \nu_t^\star(X_i)\Delta_t(X_i)^\top.$$

Each individual term is $o_p(1)$ because $\nu_t^\star(X_i)$ is integrable. Therefore, we have that Equation (10) is $o_p(1)$. Now, consider the expression.

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T \\
&\quad - \mathbb{E}\left[\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T \mid \mathcal{H}_{t-1}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\nu}_t(X_i)\hat{\nu}_t(X_i)^T - \nu_t^\star(X_i)\nu_t^\star(X_i)^\top\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\nu_t^\star(X_i)\nu_t^\star(X_i)^\top \\
&\quad - \mathbb{E}\left[\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t\right]^T \mid \mathcal{H}_{t-1}\right].
\end{aligned}$$

The first term is $o_p(1)$ by the arguments given above, the second term is $o_p(1)$ by the weak law of large numbers.

### B.5 Proof of Proposition 3

We have for all $\theta$ that

$$\begin{aligned}
\mathbb{E}\left[\dot{m}_\theta(X_i, A_i, Y_i)|X_i, A_i = a\right] &= \mathbb{E}\left[z_\theta(X_i, a)Y_i + v_\theta(X_i, a)|X_i, A_i = a\right] \\
&= z_\theta(X_i, a)\mathbb{E}[Y_i|X_i, A_i = a] + v_\theta(X_i, a).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&g_t(X_i, a) - \mathbb{E}\left[m_{\theta^\star}(X_i, A_i, Y_i)|X_i, A_i = a\right] \\
&\quad = z_{\bar{\theta}_T}(X_i, a)f_t(X_i, a) - z_{\theta^\star}(X_i, a)E[Y_i|X_i, A_i = a] \\
&\qquad + v_{\bar{\theta}_T}(X_i, a) - v_{\theta^\star}(X_i, a).
\end{aligned}$$

We have that $v_{\bar{\theta}_T}(X_i, A_i) - v_{\theta^\star}(X_i, A_i) = o_p(1)$ by the consistency of $\bar{\theta}_T$ combined with the continuous mapping theorem. Note for the remaining terms that

$$\begin{aligned}
&z_{\bar{\theta}_T}(X_i, a)f_t(X_i, a) - z_{\theta^\star}(X_i, a)E[Y_i|X_i, A_i = a] \\
&= z_{\bar{\theta}_T}(X_i, a)f_t(X_i, a) - z_{\theta^\star}(X_i, a)E[Y_i|X_i, A_i = a] \\
&\quad + z_{\bar{\theta}_T}(X_i, a)E[Y_i|X_i, A_i = a] - z_{\bar{\theta}_T}(X_i, a)E[Y_i|X_i, A_i = a] \\
&= (z_{\bar{\theta}_T}(X_i, a) - z_{\theta^\star}(X_i, a))E[Y_i|X_i, A_i = a] + z_{\bar{\theta}_T}(X_i, a)\left(f_t(X_i, a) - E[Y_i|X_i, A_i = a]\right).
\end{aligned}$$

Applying the assumptions that $\bar{\theta}_t \xrightarrow{p} \theta^\star$ and $f_t(X_i, a) - E[Y_i|X_i, A_i = a] \xrightarrow{p} 0$ along with the continuous mapping theorem immediately yields the result that $g_t(X_i, a)$ is consistent.

Next, we will prove the second part of the theorem. By the properties of variance and because $z_\theta(X_i, A_i)$ is fixed conditional on $X_i$ and $A_i$, we have that

$$\mathrm{Var}\left[\dot{m}_\theta(X_i, A_i, Y_i)|X_i, A_i = a\right] = z_\theta(X_i, a)z_\theta(X_i, a)^T\mathrm{Var}(Y_i|X_i, A_i = a).$$

Analyzing the term,

$$\begin{aligned}
h_t(X_i, a) &- \mathrm{Var}\big(\dot{m}_{\theta^\star}(X_i, A_i, Y_i) \mid X_i, A_i\big) \\
&= z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top j_t(X_i, a) - z_{\theta^\star}(X_i, a) z_{\theta^\star}(X_i, a)^\top \mathrm{Var}(Y_i | X_i, A_i = a) \\
&= z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top j_t(X_i, a) \\
&\quad - z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top \mathrm{Var}(Y_i | X_i, A_i = a) \\
&\quad + z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top \mathrm{Var}(Y_i | X_i, A_i = a) \\
&\quad - z_{\theta^\star}(X_i, a) z_{\theta^\star}(X_i, a)^\top \mathrm{Var}(Y_i | X_i, A_i = a) \\
&= z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top \left(j_t(X_i, a) - \mathrm{Var}(Y_i | X_i, A_i = a)\right) \\
&\quad + \left(z_{\bar{\theta}_T}(X_i, a) z_{\bar{\theta}_T}(X_i, a)^\top - z_{\theta^\star}(X_i, a) z_{\theta^\star}(X_i, a)^\top\right) \mathrm{Var}(Y_i | X_i, A_i = a).
\end{aligned}$$

Applying the assumptions that $\bar{\theta}_t \xrightarrow{p} \theta^\star$, $j_t(X_i, a) - \mathrm{Var}[Y_i | X_i, A_i = a] \xrightarrow{p} 0$ together with the continuous mapping theorem immediately yields the result that $h_t(X_i, a)$ is consistent.

### B.6 Proof of Proposition 4

*Proof.* Define

$$\begin{aligned}
R_t(\theta) = \sum_{a=1}^{K} \pi_e(A_t = a | X_t) \Big( & m_\theta(a, X_t, f_t(a, X_t)) \\
& + \mathbb{1}_{A_t = a} \frac{m_\theta(X_t, A_t, Y_t) - m_\theta(X_t, A_t, f_t(X_t, A_t))}{\mathbb{P}\left(A_t = a | X_t, \mathcal{H}_{t-1}\right)} \Big).
\end{aligned}$$

By definition of $\tilde{\theta}_T$, we have that

$$\sum_{t=1}^{T} R_t(\tilde{\theta}_T) = \sup_{\theta \in \Theta} \sum_{t=1}^{T} R_t(\tilde{\theta}_T) \geq \sum_{t=1}^{T} R_t(\theta^\star).$$

This implies that

$$\begin{aligned}
\mathbb{P}\left(\left\|\tilde{\theta}_T - \theta^\star\right\| \geq \epsilon\right) &\leq \mathbb{P}\left( \sup_{\|\theta - \theta^\star\| \geq \epsilon} \sum_{t=1}^{T} R_t(\theta) \geq \sum_{t=1}^{T} R_t(\theta^\star) \right) \\
&= \mathbb{P}\left( \sup_{\|\theta - \theta^\star\| \geq \epsilon} \left\{ \frac{1}{T} \sum_{t=1}^{T} R_t(\theta) \right\} - \frac{1}{T} \sum_{t=1}^{T} R_t(\theta^\star) \geq 0 \right) \\
&\leq \mathbb{P}\Bigg( \sup_{\|\theta - \theta^\star\| \geq \epsilon} \left\{ \frac{1}{T} \sum_{t=1}^{T} R_t(\theta) - \mathbb{E}\left[R_t(\theta) | \mathcal{H}_{t-1}\right] \right\} \\
&\qquad + \sup_{\|\theta - \theta^\star\| \geq \epsilon} \left\{ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[R_t(\theta) - R_t(\theta^\star) | \mathcal{H}_{t-1}\right] \right\} \\
&\qquad - \frac{1}{T} \sum_{t=1}^{T} \left\{ R_t(\theta^\star) - \mathbb{E}\left[R_t(\theta^\star) | \mathcal{H}_{t-1}\right] \right\} \geq 0 \Bigg).
\end{aligned}$$

We consider each term separately. First, note that in all cases we have that

$$
\begin{aligned}
\mathbb{E}\left[R_t(\theta)|\mathcal{H}_{t-1}\right] &= \mathbb{E}_{\mathcal{P},\pi_t}\Bigg[\sum_{a=1}^{K}\pi_e(A_t=a|X_t)\Big(m_\theta(a,X_t,f_t(a,X_t)) \\
&\qquad + \mathbb{1}_{A_t=a}\frac{m_\theta(X_t,A_t,Y_t)-m_\theta(X_t,A_t,f_t(X_t,A_t))}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\Big)\Big|\mathcal{H}_{t-1}\Bigg] \\
&= \mathbb{E}_{\mathcal{P},\pi_t}\left[\frac{\pi_e(A_t=a|X_t)}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}m_\theta(X_t,A_t,Y_t)\right] + \\
&\qquad \sum_{a=1}^{K}\pi_e(A_t=a|X_t)\mathbb{E}_{\mathcal{P},\pi_t}\left[\left(1-\frac{\mathbb{1}_{A_t=a}}{\mathbb{P}\left(A_t=a|X_t,\mathcal{H}_{t-1}\right)}\right)\right. \\
&\qquad\qquad \left.\cdot m_\theta(a,X_t,f_t(a,X_t))\right] \\
&= \mathbb{E}_{\pi_e}\left[m_\theta(X_t,A_t,Y_t)\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sup_{\|\theta-\theta^\star\|\geq\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[R_t(\theta)-R_t(\theta^\star)|\mathcal{H}_{t-1}\right]\right\} \\
&\quad = \sup_{\|\theta-\theta^\star\|\geq\epsilon}\left\{\mathbb{E}_{\pi_e}\left[m_\theta(X_t,A_t,Y_t)-m_{\theta^\star}(X_t,A_t,Y_t)\right]\right\} < -\delta_1,
\end{aligned}
$$

for some $\delta_1 > 0$ by Assumption 3. Note that Assumption 5 implies that the bracketing number $N_{[]}(\epsilon,\mathcal{M}_\Theta,L_2(\mathcal{P},\pi_e)) < \infty$ for $\mathcal{M}_\Theta = \{m_\theta(X_t,A_t,Y_t) : \theta\in\Theta\}$. This combined with Assumption 6 allows us to apply Lemma 1 conclude that the first and third terms also conerge to 0. Therefore, $\mathbb{P}\left(\left\|\tilde{\theta}_T-\theta^\star\right\|_1\geq\epsilon\right)\to 0$.

□

### B.7 Proof of Proposition 5

By the law of large numbers, we have that $\frac{n_p}{T}\to pp$ and therefore $\frac{1}{n_p}=\frac{1}{pT}+o_p(T^{-1})$. Define the per-round variance summand

$$
\Psi_t := \Big(\widehat{\nu}_t(X_t)-\overline{\nu}_t\Big)\Big(\widehat{\nu}_t(X_t)-\overline{\nu}_t\Big)^\top + \sum_{a=1}^{K}\frac{\pi_e(a\mid X_t)^2}{\mathbb{P}(A_t=a\mid X_t,\mathcal{H}_{t-1})}\,h_t(X_t,a)\;-\;\widehat{\nu}_t(X_t)\widehat{\nu}_t(X_t)^\top.
$$

We can therefore write $\widehat{V}_t=\frac{1}{n_p}\sum_{t=1}^{T}\zeta_t\,\Psi_t$. Moreover, $\widehat{V}_t=\frac{1}{pT}\sum_{t=1}^{T}\zeta_t\,\Psi_t+o_p(1)$.

By assumption, $\{\zeta_t\}$ are independent of all histories, therefore $\mathbb{E}[\zeta_t\Psi_t\mid\mathcal{H}_{t-1}]=p\,\mathbb{E}[\Psi_t\mid\mathcal{H}_{t-1}]$. Consequently, by the law of large numbers and Slutsky's theorem, $\widehat{V}_t=\mathbb{E}[\Psi_t\mid\mathcal{H}_{t-1}]+o_p(1)$. The remainder of the argument is identical to Proposition 2 and, therefore,

$$
\left\|\widehat{V}_t-V_{t,\theta^\star}\right\|_{\mathrm{op}}\xrightarrow{p}0,
$$

as claimed.

## C Procedure for Semi-Synthetic Data Construction and Additional Results

All simulations were conducted over a horizon of ($T=2000$) sequential decisions with an initial burn-in period of 500 observations used to fit nuisance models. Results are averaged over 100 independent replications. The semi-synthetic experiments were constructed using the right-knee subsample of the Osteoarthritis Initiative (OAI), consisting of ($n=3486$) patients and 12 baseline covariates, including demographic characteristics,

Kellgren-Lawrence grade, WOMAC pain and disability scores, KOOS quality-of-life measures, prior surgery indicators, radiographic osteoarthritis indicators, body mass index, hip-knee-foot angle, and age. Outcome models were estimated using random forests trained on the burn-in sample and subsequently refit every 500 observations using all accumulated data. The conditional variance model was estimated using a second random forest trained on squared residuals. When required by the covariance estimators of Section 4.1, an additional external sample of 1000 synthetic patients drawn from the empirical OAI covariate distribution was used.

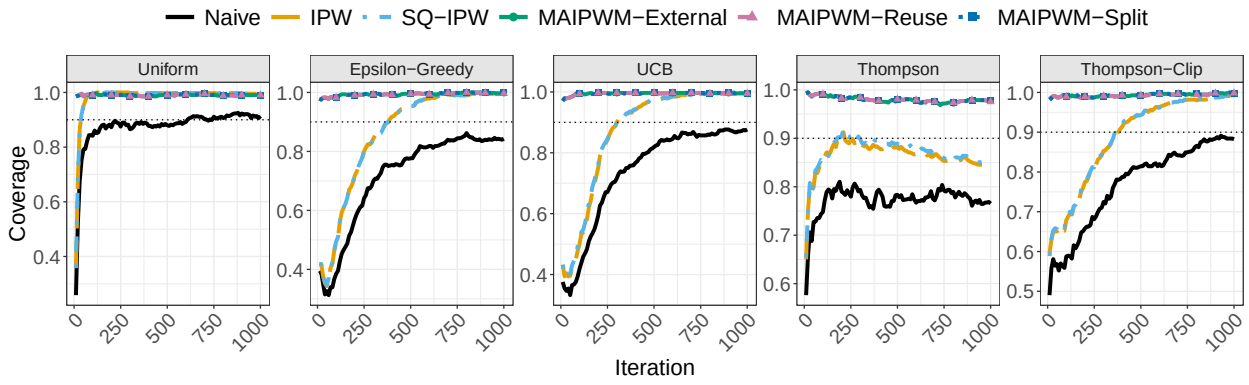
For the semi-synthetic dataset construction, we learn $E[Y_t|X_t]$ via a machine learning model, and then enforce a linear model to describe $E[Y_t|X_t, A_t]$. The procedure is

1. Train a machine learning model $f$ to predict $E[Y_t|X_t]$ using available dataset.
2. Generate synthetic causal outcomes, where it is assumed that the mean
$$E[Y_t|X_t, A_t] = \sum_{a=1}^{K} \beta_1^a \mathbb{1}_{A_t=a} + \sum_{a=1}^{K} \beta_2^a f(X_t) \times \mathbb{1}_{A_t=a},$$
for user chosen parameters $\beta_1^a$, $\beta_2^a$.
3. We observe each $X_t$ sequentially (sampled with replacement from the actual OAI data), and then in each round $A_t$ are chosen using each of the methods in Section 5. We then draw $Y_t \sim N\left(\sum_{a=1}^{k} \beta_1^a \mathbb{1}_{A_t=k} + \sum_{a=1}^{k} \beta_2^a f(X_t) \times \mathbb{1}_{A_t=k}, 1\right)$.
4. Alternatively, we can introduce heteroskedasticity into the dataset by learning a secondary model $v$ to predict $E\left[(Y_t - f(X_t))^2 |X_t\right]$. We then sample
$$\tilde{Y}_t \sim N\left(\sum_{a=1}^{k} \beta_1^a \mathbb{1}_{A_t=a} + \sum_{a=1}^{k} \beta_2^a f(X_t) \times \mathbb{1}_{A_t=a}, \sum_{a=1}^{K} v(X_t)\gamma_a \mathbb{1}_{A_t=a}\right),$$
where $\gamma_k$ is a user-chosen multiplier to the variance.

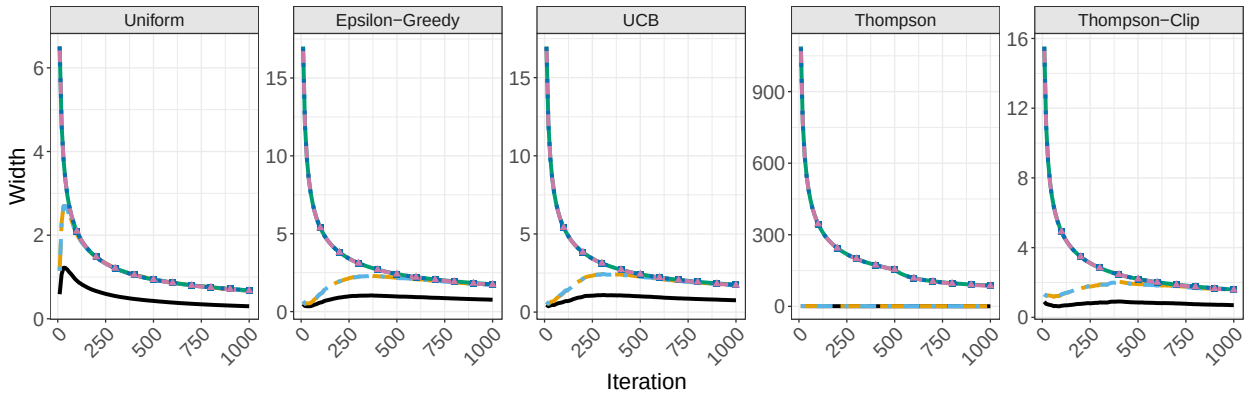
For the empirical simulations, we choose six different initializations

1. **Scenario 1: Correct Specification, Homoscedasticity, Unique Optimal Arm**
$$\beta_1 = (0, 1, 2, 3, 4, 5, 6, 7), \beta_2 = (0, 0, 0, 0, 0, 0, 0, 0), \text{and unit variance uniformly.}$$
In this scenario, all methodologies other than naive unweighted MLE should perform relatively well as bandit algorithms *will concentrate*. Results are shown in Figure 3.
2. **Scenario 2: Correct Specification, Homoscedasticity, No Unique Optimal Arm.**
$$\beta_1 = (0, 0, 1, 2, 2, 3, 5, 5), \beta_2 = (0, 0, 0, 0, 0, 0, 0, 0), \text{and unit variance uniformly.}$$
In this scenario, IPW-style estimators may have difficulties but confidence intervals constructed as in Zhang et al. (2021) should still cover correctly. Results are shown in Figure 4.
3. **Scenario 3: Misspecification, Homoscedasticity.**
$$\beta_1 = (0, 0, 1, 2, 2, 3, 4, 4), \beta_2 = (1, -1, 1, 0, 1, 1, 1, -3), \text{and unit variance uniformly.}$$
Results are shown in Figure 5.
4. **Scenario 4: Misspecification, Heteroskedasticity**
$$\beta_1 = (0, 0, 1, 2, 2, 3, 4, 5), \ \beta_2 = (1, -1, 1, 0, 1, 1, 1, -2),$$
$$\text{and } \gamma = (1, 2, 3, 4, 5, 5, 5, 5) \times 0.2.$$
Results are shown in Figure 6.

We showed the results of only Scenario 2 in the main body due to space constraints, but the other scenarios reveal broadly comparable results.

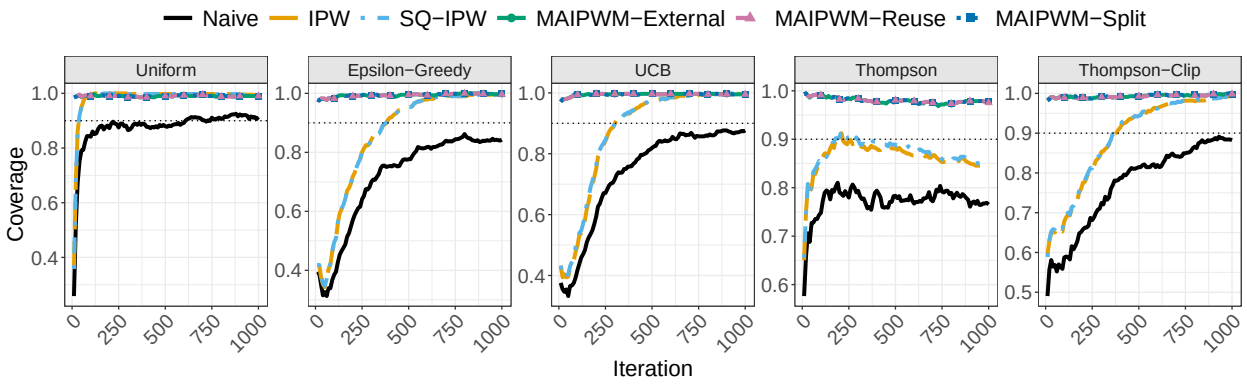



(a) Coverage for target $1 - \alpha = 0.9$

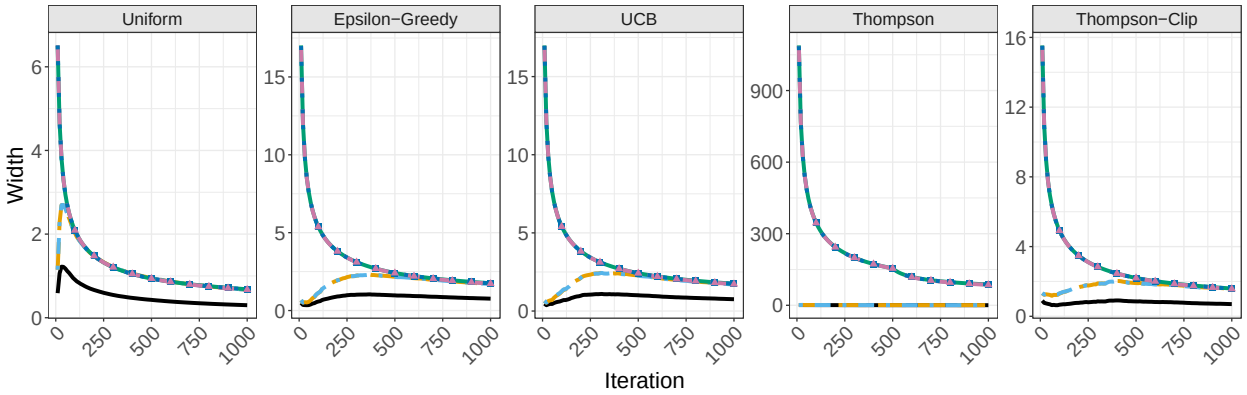



(b) Confidence interval width

Figure 3: Simulation results for scenario 1, where the model is correctly specified. In these settings, we see that all methods cover, but it often takes the naive approaches (IPW and SQ-IPW) estimates significantly more samples to reach the nominal coverage rates. Thompson sampling undercovers for several methods when propensity scores are not clipped, consistent with theory.

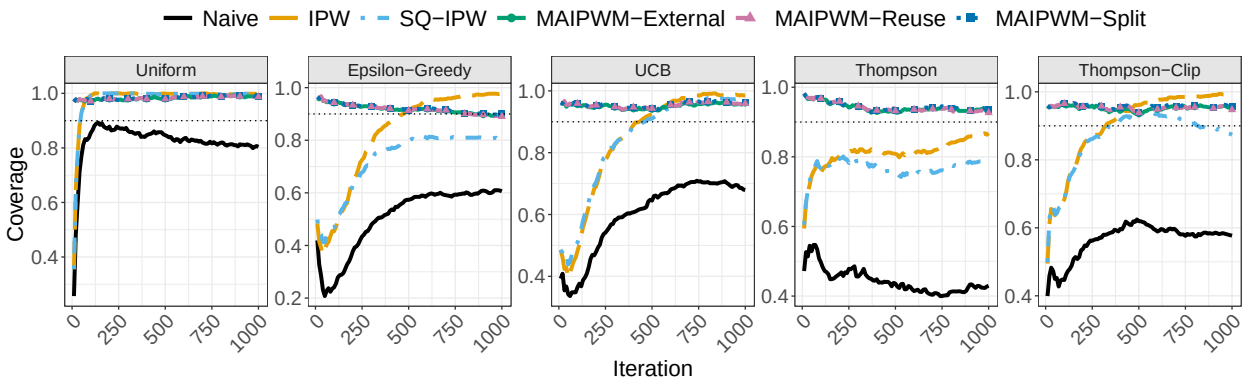



(a) Coverage for target $1 - \alpha = 0.9$

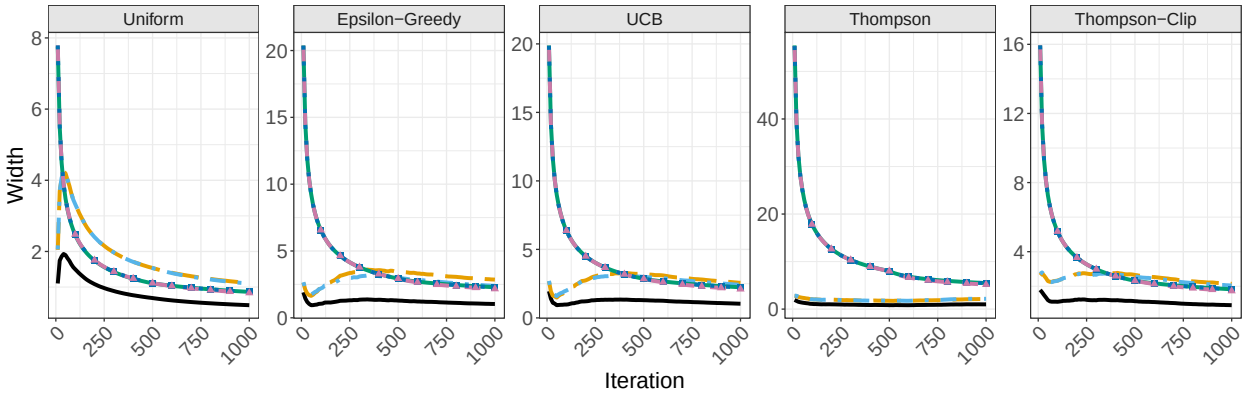



(b) Confidence interval width

Figure 4: Simulation results for scenario 2, where the model is correctly specified but there is not a unique optimal arm. The results are broadly similar as in Figure 4, though we note there is no theoretical guarantee that IPW estimates should cover in this setting (though they do empirically).

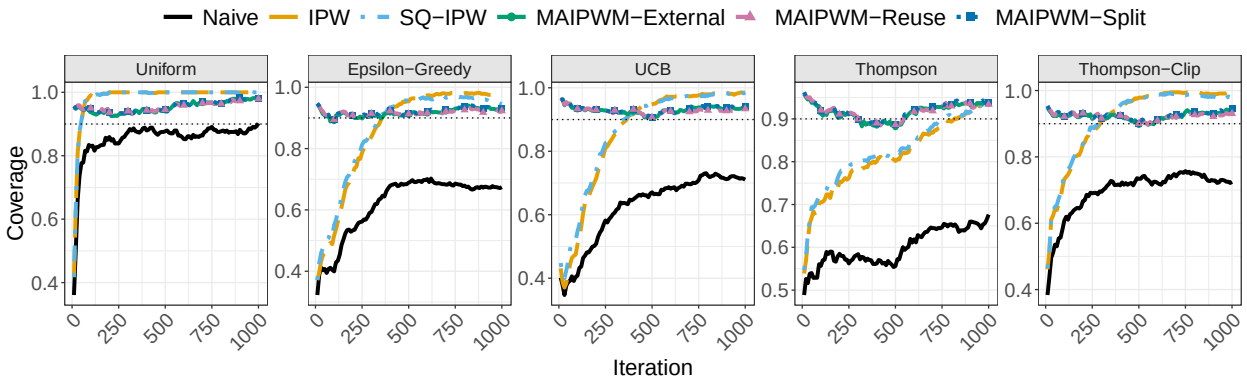



(a) Coverage for target $1 - \alpha = 0.9$

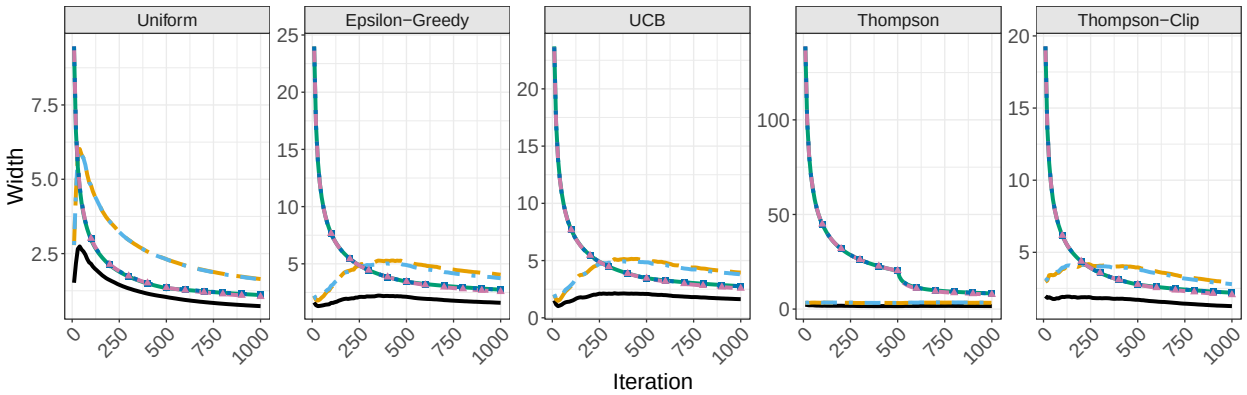



(b) Confidence interval width

Figure 5: Simulation results for scenario 3, when there is model misspecification but homoskedastic errors. IPW and SQ-IPW estimators now do not cover in every scenario, as their theoretical guarantees depend on *correct specification* of the working models.




(a) Coverage for target $1 - \alpha = 0.9$




(b) Confidence interval width

Figure 6: Simulation results for scenario 3, when there is model misspecification and heteroskedastic errors. The results are broadly similar to Figure 5.