Contents

A. Simulations

A.1. Simulation Details

Simulation Environment

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

Algorithm

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

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

A.2. Details on Constructing of Confidence Regions

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

A.2.1. LEAST SQUARES ESTIMATORS

- $\hat{\theta}_T = \left(\sum_{t=1}^T W_t Z_t Z_t^\top\right)^{-1} \sum_{t=1}^T W_t Z_t R_t$
  - For unweighted least squares, $W_t = 1$ and we call the estimator $\hat{\theta}_T^{\text{OLS}}$.
  - For adaptively weighted least squares, $W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$; this is equivalent to using square-root importance weights with a uniform stabilizing policy. We call the estimator $\hat{\theta}_T^{\text{AW-LS}}$.
- We assume homoskedastic errors and estimate the noise variance σ^2 as follows:

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

- We use a Hotelling t-squared test statistic to construct confidence regions for $\theta^*(\mathcal{P})$:

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

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

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

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

A.2.2. MLE Estimators

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

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

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

  We use Newton Raphson optimization to solve for $\hat{\theta}_T$.

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

- Second derivative of score function: $-\sum_{t=1}^{T} b''(\hat{\theta}_T^\top Z_t) Z_t Z_t^\top$.

- We use a Hotelling t-squared test statistic to construct confidence regions for $\theta^*(\mathcal{P})$:

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

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

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

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

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

- For the AW-MLE estimator, we also construct projected confidence regions for $\theta_1^*(\mathcal{P})$ using the confidence region defined in equation (2). See Section A.2.5 below for more details on constructing projected confidence regions.

A.2.3. W-Decorrelated

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

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

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

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

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

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

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

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

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

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

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

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

A.2.4. Self-Normalized Martingale Bound

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

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

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

A.2.5. Construction of Projected Confidence Regions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

  Thus the equation for the final projected ellipsoid is

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

A.3. Additional Simulation Results

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

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

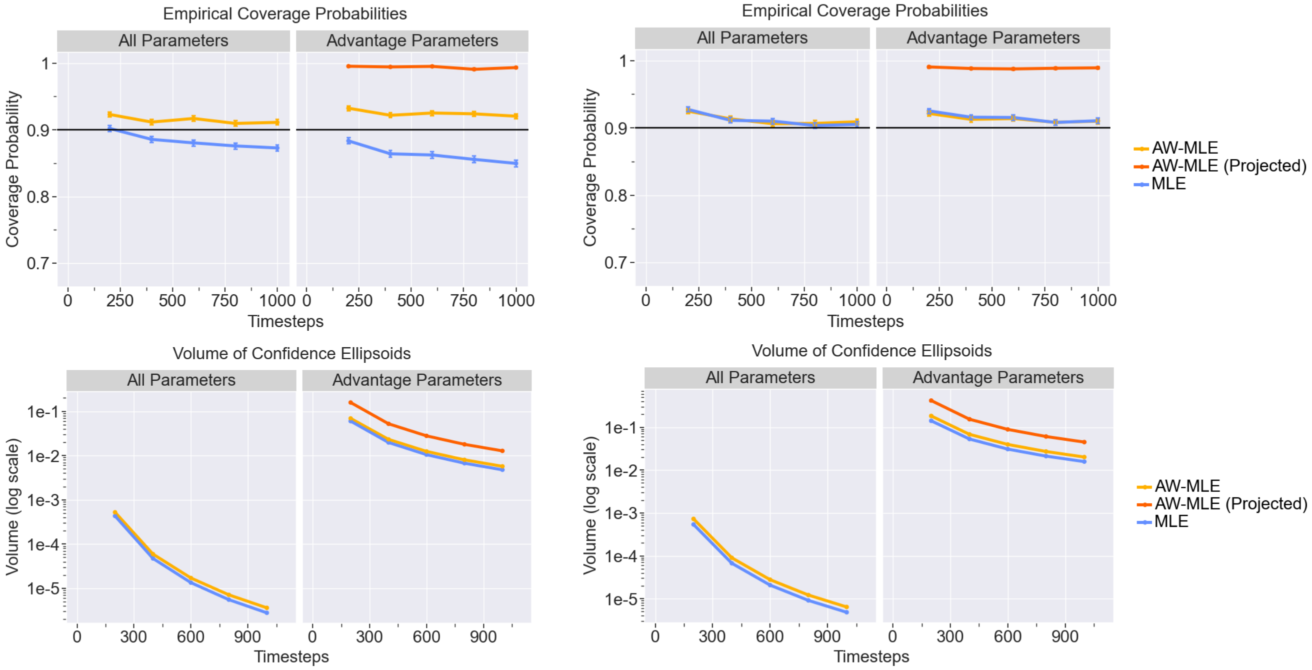

Figure 1. **Poisson Rewards:** Empirical coverage probabilities for 90% confidence ellipsoids for parameters $\theta^*(\mathcal{P})$ and parameters $\theta_1^*(\mathcal{P})$ (top row). We also plot the volumes of these 90% confidence ellipsoids for $\theta^*(\mathcal{P})$ and parameters $\theta_1^*(\mathcal{P})$ (bottom row). We set the true parameters to $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and to $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

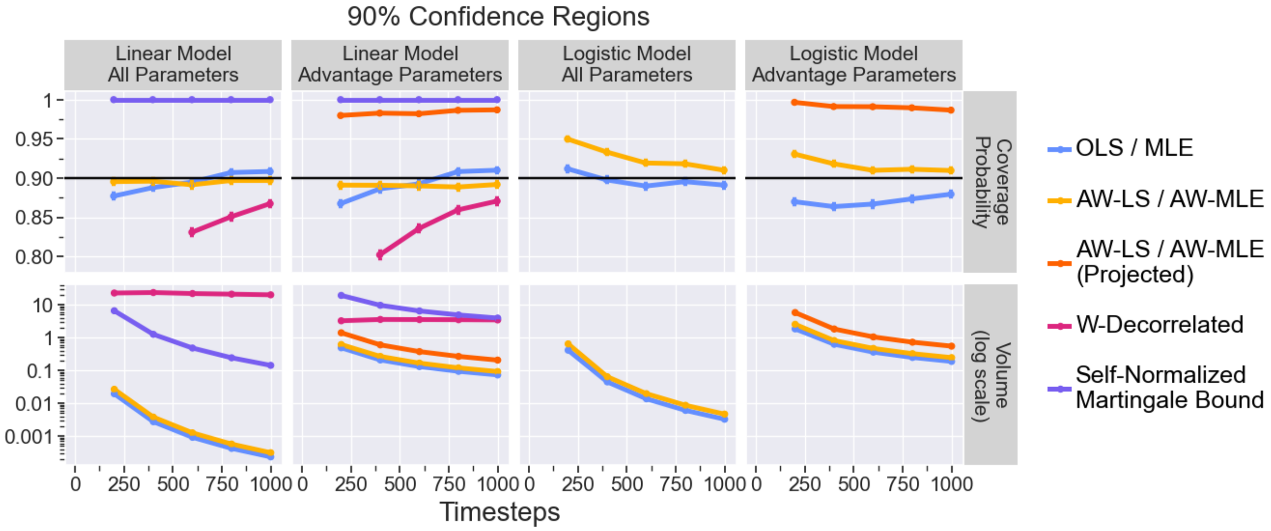

Figure 2. Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. In these simulations, $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$. The left two columns are for the linear reward model setting (t-distributed rewards) and the right two columns are for the logistic regression model setting (Bernoulli rewards). We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings.

In Figure 3, we plot the mean squared errors of all estimators for all three simulation settings (same simulation hyperparameters as described previously for the respective simulation settings).

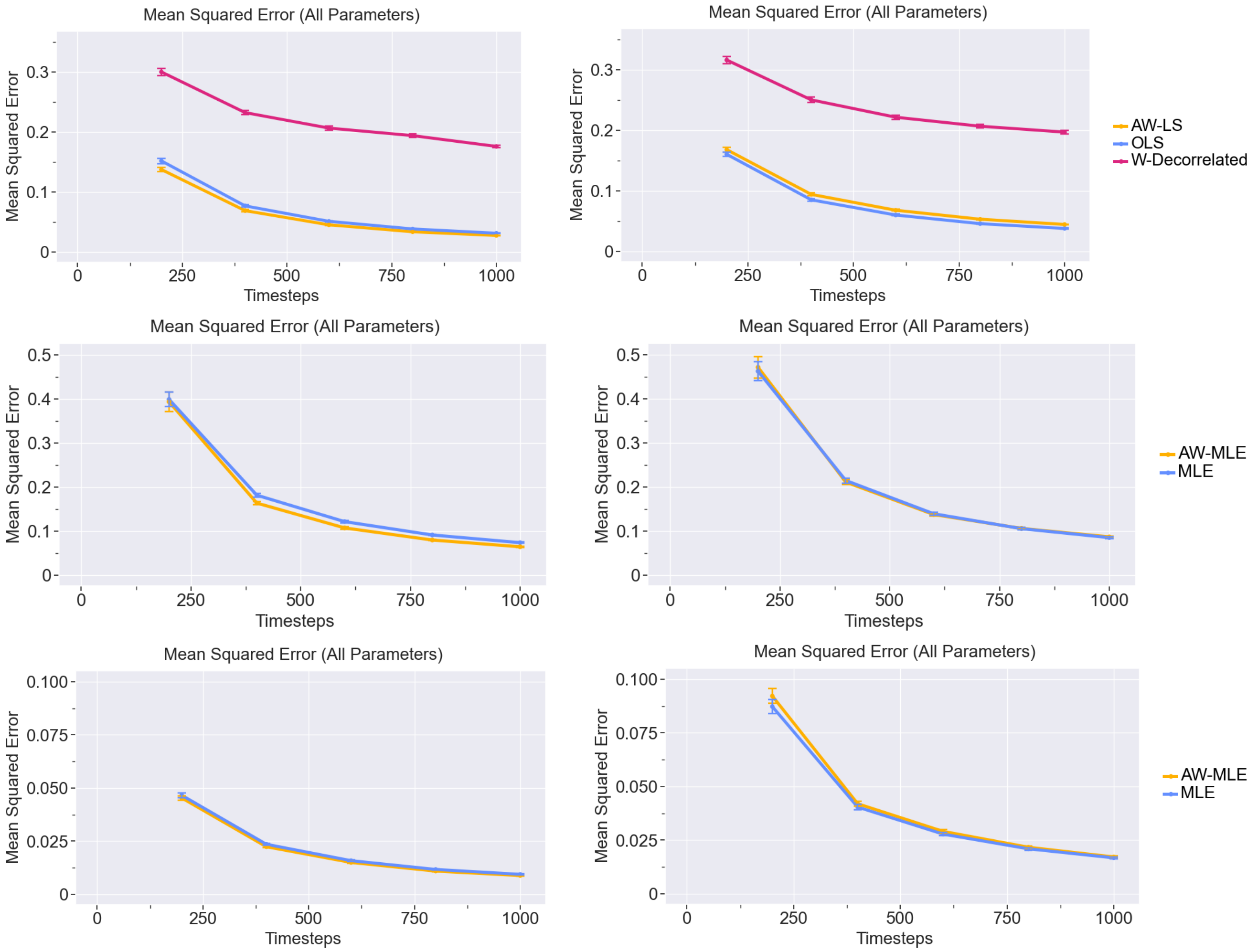

Figure 3. Mean squared error estimators of $\theta^*(\mathcal{P})$ for linear model (top), logistic regression model (middle), and generalized linear model for Poisson rewards (bottom). We consider simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0, 0, 0]$ (left) and simulations with $\theta^*(\mathcal{P}) = [0.1, 0.1, 0.1, 0.2, 0.1, 0]$ (right).

B. Asymptotic Results

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

B.1. Conditions

We now discuss conditions under which the adaptively weighted M-estimators are asymptotically normal in the following sense

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{3}$$

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

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

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

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

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

Condition 4 (Lipschitz)**.** *There exists some function* g *such that* $\sup_{\mathcal{P}\in\boldsymbol{P},t\geq 1} \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2]$ *is bounded and satisfies the following for all* $\theta, \theta' \in \Theta$,

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

Conditions 3 and 4 together restrict the complexity of the function m in order to ensure a martingale law of large numbers result holds uniformly over functions $\{m_\theta : \theta \in \Theta\}$; this is used to prove the consistency of $\hat{\theta}_T$. Similar conditions are commonly used to prove consistency of M-estimators based on i.i.d. data, although the boundedness of the parameter space can be dropped when m_θ is concave (as it is in many canonical examples such as least squares) (Van der Vaart, 2000; Engle, 1994; Bura et al., 2018); we expect that a similar result would hold for adaptively weighted M-estimators.

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

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

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

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

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

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

For matrices A, B, we define $A \succeq B$ to mean that $A - B$ is positive semi-definite, as used above. Condition 7 (i) ensures that $\theta^*(\mathcal{P})$ is a conditionally maximizing solution for all contexts X_t and actions A_t; this ensures that $\{\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)\}_{t=1}^T$

is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=1}^T$. Note it does not require $\theta^*(\mathcal{P})$ to always be a conditionally *unique* optimal solution. Condition 7 (ii) is related to the local curvature at the maximizing solution and the analogous condition in the i.i.d. setting is trivially satisfied; we specifically use this condition to ensure we can replace $\ddot{M}(\theta^*(\mathcal{P}))$ with $\ddot{M}(\hat{\theta}_T)$ in our asymptotic normality result, i.e., that $\ddot{M}(\theta^*(\mathcal{P}))^{-1}\ddot{M}(\hat{\theta}_T) \xrightarrow{P} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$.

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

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

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

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

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

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

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

The asymptotic normality result of equation (4) guarantees that for d-dimensional $\theta^*(\mathcal{P})$,

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

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

B.2. Definitions

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

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

$$\sup_{\mathcal{P} \in \boldsymbol{P}} \mathbb{P}_{\mathcal{P}, \eta} \left(\|Z_T(\mathcal{P}) - c\| > \epsilon \right) \to 0. \tag{5}$$

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

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

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

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

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

$$\sup_{\mathcal{P}\in\boldsymbol{P}} \sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P},\eta}\left[f\left(Z_T(\mathcal{P})\right)\right] - \mathbb{E}_{\mathcal{P},\eta}\left[f\left(Z(\mathcal{P})\right)\right] \right| \to 0, \tag{6}$$

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

As discussed in Kasy (2019), Equation (5) holds if and only if for any $\epsilon > 0$ and any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\mathbb{P}_{\mathcal{P}_T,\eta}\left(\|Z_T(\mathcal{P}_T) - c\| > \epsilon\right) \to 0$.

Similarly, Equation (6) holds if and only if for any sequence $\{\mathcal{P}_T\}_{T\geq 1}$ such that $\mathcal{P}_T \in \mathbf{P}$ for all $T \geq 1$, $\sup_{f\in BL_1} \left| \mathbb{E}_{\mathcal{P}_T,\eta}\left[f\left(Z_T(\mathcal{P}_T)\right)\right] - \mathbb{E}_{\mathcal{P}_T,\eta}\left[f\left(Z(\mathcal{P}_T)\right)\right] \right| \to 0$.

B.3. Consistency

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

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

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

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

Thus, the above two results imply the following inequality:

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

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

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

By triangle inequality,

$$\le \sup_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\Bigg(\underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} (W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}])\right\}}_{(a)} + \underbrace{\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\mathcal{P},\pi}\left[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})\middle|\mathcal{H}_{t-1}\right]\right\}}_{(b)} - \underbrace{\frac{1}{T}\sum_{t=1}^{T}\left\{W_t m_{\theta^*(\mathcal{P}),t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta^*(\mathcal{P}),t}|\mathcal{H}_{t-1}]\right\}}_{(c)} \ge 0\Bigg) \to 0. \quad (7)$$

We now show that the limit in Equation (7) above holds.

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

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

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|>\epsilon}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi}[W_t(m_{\theta,t} - m_{\theta^*(\mathcal{P}),t})|\mathcal{H}_{t-1}]\right\} \le -\delta' \text{ w.p. } 1. \quad (8)$$

By law of iterated expectations,

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

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

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

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

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

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

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

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

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

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

By law of iterated expectations,

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

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

B.4. Asymptotic Normality

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

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{9}$$

B.4.1. Main Argument

The three results we show to ensure Equation (9) holds are as follows:

$$\Sigma_T(\mathcal{P})^{-1/2} \sqrt{T} \dot{M}_T(\theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{10}$$

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

$$\sup_{\theta\in\Theta:\|\theta-\theta^*(\mathcal{P})\|\leq\epsilon_{\dddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1 = O_{\mathcal{P}\in\mathbf{P}}(1). \tag{11}$$

For matrix H positive definite,

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1). \tag{12}$$

For a reminder on the notation of $o_{\mathcal{P}\in\mathbf{P}}(1)$ and $O_{\mathcal{P}\in\mathbf{P}}(1)$ see definitions 5 and 2. For now, we assume that Equations (10), (11), and (12) hold; we will show they hold in Sections B.4.2, B.4.3, and B.4.4 respectively. Our argument is based on Van der Vaart (2000, Theorem of 5.41).

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

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

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

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

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

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

Note that by the above equation and Equation (10), we have that

$$\Sigma_T(\mathcal{P})^{-1/2} \left[\ddot{M}_T(\theta^*(\mathcal{P})) + \frac{1}{2}(\hat{\theta}_T - \theta^*(\mathcal{P}))^\top \dddot{M}_T(\tilde{\theta}_T) \right] \sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \\ \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{13}$$

By Equation (12), the probability that $\ddot{M}_T(\theta^*(\mathcal{P}))$ is invertible goes to 1 uniformly over $\mathcal{P} \in \mathbf{P}$. Thus by Equation (13), we have that

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

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

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \overset{D}{\to} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \quad (14)$$

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

- By Condition 5, the minimum eigenvalue of $\Sigma_T(\mathcal{P})$ is bounded uniformly above some constant greater than zero, so $\sup_{\mathcal{P} \in \mathbf{P}} \|\Sigma_T(\mathcal{P})^{-1/2}\| = O(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| = o_{\mathcal{P} \in \mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\mathbb{1}_{\|\tilde{\theta}_T - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} = o_{\mathcal{P} \in \mathbf{P}}(1)$. Thus by Equation (11), $\dddot{M}_T(\tilde{\theta}_T) = O_{\mathcal{P} \in \mathbf{P}}(1)$.
- By Equation (12), the minimum eigenvalue of $-\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ is bounded above that of positive definite matrix H. Thus $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\| = O_{\mathcal{P} \in \mathbf{P}}(1)$.
- By Condition 5, $\sup_{\mathcal{P} \in \mathbf{P}} \|\Sigma_T(\mathcal{P})^{1/2}\| = O(1)$.

Thus, by Slutsky's Theorem and Equation (14), we have that

$$\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\theta^*(\mathcal{P}))\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \overset{D}{\to} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \quad (15)$$

Lastly, to show our desired result, that $\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \overset{D}{\to} \mathcal{N}(0, I_d)$ uniformly over $\mathcal{P} \in \mathbf{P}$, by Equation (15) and Slutsky's Theorem it is sufficient to show that $\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \ddot{M}_T(\theta^*(\mathcal{P}))^{-1} \Sigma_T(\mathcal{P})^{1/2} \overset{P}{\to} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$. Note if we can show that $\ddot{M}_T(\hat{\theta}_T) \ddot{M}_T(\theta^*(\mathcal{P}))^{-1} \overset{P}{\to} I_d$ uniformly over $\mathcal{P} \in \mathbf{P}$, then $\Sigma_T(\mathcal{P})^{-1/2} \ddot{M}_T(\hat{\theta}_T) \ddot{M}_T(\theta^*(\mathcal{P}))^{-1} \Sigma_T(\mathcal{P})^{1/2} = \Sigma_T(\mathcal{P})^{-1/2} \left[I_d + o_{\mathcal{P} \in \mathbf{P}}(1) \right] \Sigma_T(\mathcal{P})^{1/2} = I_d + \Sigma_T(\mathcal{P})^{-1/2} o_{\mathcal{P} \in \mathbf{P}}(1) \Sigma_T(\mathcal{P})^{1/2} = I_d + o_{\mathcal{P} \in \mathbf{P}}(1)$. The last limit holds since $\|\Sigma_T(\mathcal{P})^{-1/2}\| = O_{\mathcal{P} \in \mathbf{P}}(1)$ and $\|\Sigma_T(\mathcal{P})^{1/2}\| = O_{\mathcal{P} \in \mathbf{P}}(1)$ by Condition 5 (use the same argument as that used in the bullet points below Equation (14)).

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

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

Recall that the probability the inverse of $\ddot{M}_T(\theta^*(\mathcal{P}))$ exists goes to 1 by Equation (12) (use the same argument as that used in the bullet points below Equation (14)). Thus we have that $\ddot{M}_T(\hat{\theta}_T) \ddot{M}_T(\theta^*(\mathcal{P}))^{-1}$ equals the following:

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

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

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

- By uniform consistency of $\hat{\theta}_T$, $\mathbb{1}_{\|\tilde{\theta}_T - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} = o_{\mathcal{P} \in \mathbf{P}}(1)$. Thus by Equation (11), $\dddot{M}_T(\tilde{\theta}_T) = O_{\mathcal{P} \in \mathbf{P}}(1)$.
- By uniform consistency of $\hat{\theta}_T$, $\|\hat{\theta}_T - \theta^*(\mathcal{P})\| = o_{\mathcal{P} \in \mathbf{P}}(1)$.
- By Equation (12), $\|\ddot{M}_T(\theta^*(\mathcal{P}))^{-1}\| = O_{\mathcal{P} \in \mathbf{P}}(1)$.

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

We will show that Equation (10) holds by applying a martingale central limit theorem. For notational convenience, we let $\dot{m}_{\theta,t} := \dot{m}_\theta(Y_t, X_t, A_t)$. Note that by definition $\Sigma_T(\mathcal{P})^{-1/2}\sqrt{T}\dot{M}_T(\theta^*(\mathcal{P})) = \Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t \dot{m}_{\theta^*(\mathcal{P}),t}$. We first show that $\left\{\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}W_t\dot{m}_{\theta^*(\mathcal{P}),t}\right\}_{t=1}^T$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t=0}^T$. For any $t \in [1\colon T]$,

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

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

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

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

By Cramer-Wold device, to show that Equation (10) holds, it is sufficient to show that for any fixed $\mathbf{c} \in \mathbb{R}^d$ with $\|\mathbf{c}\|_2 = 1$, that $\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t\dot{m}_{\theta^*(\mathcal{P}),t} \xrightarrow{D} \mathcal{N}\left(0, \mathbf{c}^\top I_d\mathbf{c}\right)$ uniformly over $\mathcal{P} \in \mathbf{P}$. We now apply Theorem 2, a uniform version of the martingale central limit theorem of Dvoretzky (1972); while the original theorem holds for any fixed $\mathcal{P}$, we can show uniform convergence in distribution by ensuring that the conditions of the theorem hold uniformly over $\mathcal{P} \in \mathbf{P}$ (see Definition 3). By Theorem 2, it is sufficient to show that the following two conditions hold:

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

1. Conditional Variance

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

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

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

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

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

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

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

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

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

2. Conditional Lindeberg

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

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

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

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

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

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

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

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

- Above, inequality (a) holds because $\mathbb{1}_{\left|W_t\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}_{\theta^*(\mathcal{P}),t}\right|>\sqrt{T}\delta} = 1$ if and only if $W_t^2\frac{1}{T\delta^2}\mathbf{c}^\top\Sigma_T(\mathcal{P})^{-1/2}\dot{m}^{\otimes 2}_{\theta^*(\mathcal{P}),t}\Sigma_T(\mathcal{P})^{-1/2}\mathbf{c} > 1$.
- Inequality (b) holds because by Condition 9, $W_t^2 \leq \rho_{\max}$ with probability 1.
- Equality (c) holds by the law of iterated expectations.

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

B.4.3. SHOWING THAT $\sup_{\theta \in \Theta: \|\theta - \theta^*(\mathcal{P})\| \leq \epsilon_{\dddot{m}}} \left\| \dddot{M}_T(\theta) \right\|_1$ IS BOUNDED IN PROBABILITY

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

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

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

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

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

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

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

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

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

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

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

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

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

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

B.4.4. Lower bounding $-\ddot{M}_T(\theta^*(\mathcal{P}))$

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

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

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

By law of iterated expectations,

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

By Condition 1,

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

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

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

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

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

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

$$-\ddot{M}_T(\theta^*(\mathcal{P})) \succeq H + o_{\mathcal{P}\in\mathbf{P}}(1). \tag{16}$$

B.5. Lemmas and Other Helpful Results

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

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

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

Under the above conditions,

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

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

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

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

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

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

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

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

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ W_t f(Y_t, X_t, A_t) - \mathbb{E}_{\mathcal{P}, \pi}[W_t f(Y_t, X_t, A_t) | \mathcal{H}_{t-1}] \right\} = O_{\mathcal{P} \in \boldsymbol{P}}(1). \tag{17}$$

Note that the above equation implies that

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

Lemma 1 is a type of martingale weak law of large number result and the proof is similar to the weak law of large numbers proofs for i.i.d. random variables.

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

$$\begin{aligned}
&\sup_{\mathcal{P} \in \mathbf{P}} \mathbb{P}_{\mathcal{P}, \pi} \left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left\{ W_t f_t - \mathbb{E}_{\mathcal{P}, \pi}[W_t f_t | \mathcal{H}_{t-1}] \right\} \right| > \epsilon \right) \\
&\underset{(a)}{\leq} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \mathbb{E}_{\mathcal{P}, \pi} \left[\left(\sum_{t=1}^{T} \left\{ W_t f_t - \mathbb{E}_{\mathcal{P}, \pi}[W_t f_t | \mathcal{H}_{t-1}] \right\} \right)^2 \right] \\
&\underset{(b)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi} \left[\left\{ W_t f_t - \mathbb{E}_{\mathcal{P}, \pi}[W_t f_t | \mathcal{H}_{t-1}] \right\}^2 \right] \\
&\underset{(c)}{\leq} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi} \left[W_t^2 f_t^2 \right] \\
&\underset{(d)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} W_t^2 \pi_t(a, X_t, \mathcal{H}_{t-1}) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right] \\
&\underset{(e)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}} \left[\int_{a \in \mathcal{A}} \pi_t^{\text{sta}}(a, X_t) \mathbb{E}_{\mathcal{P}}[f_t^2 | \mathcal{H}_{t-1}, X_t, A_t = a] da \right] \\
&\underset{(f)}{=} \frac{1}{T\epsilon^2} \sup_{\mathcal{P} \in \mathbf{P}} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[f_t^2 \right] \underset{(g)}{\leq} \frac{4m}{\epsilon^2}
\end{aligned}$$

- Above (a) holds by Chebyshev's inequality.
- (b) holds because the above terms form a martingale difference sequence with respect to $\mathcal{H}_{t-1}$, i.e., $\mathbb{E}_{\mathcal{P},\pi}\big[W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]\big|\mathcal{H}_{t-1}\big] = 0$; this implies that cross terms disappear, i.e., for $t > s$,

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

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

Since $s > t$,

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

- (c) holds because $\mathbb{E}_{\mathcal{P},\pi}\left[\{W_t f_t - \mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]\}^2\right] = \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2\right] - \mathbb{E}_{\mathcal{P},\pi}\left[\mathbb{E}_{\mathcal{P},\pi}[W_t f_t|\mathcal{H}_{t-1}]^2\right] \leq \mathbb{E}_{\mathcal{P},\pi}\left[W_t^2 f_t^2\right]$.
- (d) holds by law of iterated expectations.
- (e) holds because $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t,X_t)}{\pi_t(A_t,X_t,\mathcal{H}_{t-1})}}$.
- (f) holds since by Condition 1, $\mathbb{E}_{\mathcal{P}}[f_t^2|\mathcal{H}_{t-1},X_t,A_t] = \mathbb{E}_{\mathcal{P}}[f_t^2|X_t,A_t]$ and by law of iterated expectations $\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[f_t^2\right] = \mathbb{E}_{\mathcal{P}}\left[\int_{a\in\mathcal{A}}\pi_t^{\text{sta}}(a,X_t)\mathbb{E}_{\mathcal{P}}[f_t^2|X_t,A_t=a]da\right]$.
- (g) holds since $\sup_{\mathcal{P}\in\mathbf{P},t\geq1}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[f_t^2\right] < m < \infty$.

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

$$\sup_{\theta\in\Theta}\left\{\frac{1}{T}\sum_{t=1}^{T}W_t m_{\theta,t} - \mathbb{E}_{\mathcal{P},\pi}[W_t m_{\theta,t}|\mathcal{H}_{t-1}]\right\} = O_{\mathcal{P}\in\boldsymbol{P}}(1). \tag{18}$$

Lemma 1 is a type of martingale functionally uniform law of large number result and the proof is similar to the functionally uniform law of large numbers proofs for i.i.d. random variables Van Der Vaart and Wellner (1996, Theorem 2.4.1).

Proof:

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

(a) For any $\theta\in\Theta$, we can find $(l,u)\in B_\delta$ such that
(i) $l(y,x,a) \leq m_\theta(y,x,a) \leq u(y,x,a)$ for all (x,y) in the joint support of $\{\mathcal{P}\in\mathbf{P}\}$ and all $a\in\mathcal{A}$.
(ii) $\sup_{\mathcal{P}\in\mathbf{P},t\geq1}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[|u(Y_t,X_t,A_t) - l(Y_t,X_t,A_t)|\right] \leq \delta$.

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

(c) For any $(l,u)\in B_\delta$, for some $m<\infty$ which does no depend on δ,
$\sup_{\mathcal{P}\in\mathbf{P},t\geq1}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[u(Y_t,X_t,A_t)^2\right] \leq m$ and $\sup_{\mathcal{P}\in\mathbf{P},t\geq1}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[l(Y_t,X_t,A_t)^2\right] \leq m$.

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

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

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

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}}[g(Y_t, X_t, A_t)^2] \leq m_g. \tag{19}$$

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

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

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

Note that

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

The inequalities above hold by Equation (19) and since $\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[g(Y_t, X_t, A_t)\right] \leq \sqrt{\mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[g(Y_t, X_t, A_t)^2\right]}$ by Jensen's inequality. (a) above holds for our choice of B_δ by letting meshwidth $\lambda = \delta/(2\sqrt{m_g})$.

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

$$\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\big\{m_\theta(Y_t, X_t, A_t) + g(Y_t, X_t, A_t)\big\}^2\right]$$
$$\leq 3 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[m_\theta(Y_t, X_t, A_t)^2\right] + 3 \sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[g(Y_t, X_t, A_t)^2\right]. \tag{20}$$

Note that the above upper bound, Equation (20), also holds for $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[\big\{m_\theta(Y_t, X_t, A_t) - g(Y_t, X_t, A_t)\big\}^2\right]$.

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

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

Note that $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[m_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^2\right]$ is bounded by our moment Condition 5 and that $\sup_{\mathcal{P} \in \mathbf{P}, t \geq 1} \mathbb{E}_{\mathcal{P}, \pi_t^{\text{sta}}} \left[g(Y_t, X_t, A_t)^2\right]$ is bounded by Equation (19).

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

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

The above is bounded by Equation (19) and by compactness of Θ, Condition 3. Thus (c) above holds for our choice of B_δ.

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

$$\sup_{\mathcal{P} \in \mathbf{P}} \mathbb{P}_{\mathcal{P}, \pi} \left(\sup_{\theta \in \Theta} \left\{ \frac{1}{T} \sum_{t=1}^{T} W_t m_{\theta, t} - \mathbb{E}_{\mathcal{P}, \pi}[W_t m_{\theta, t} | \mathcal{H}_{t-1}] \right\} > \epsilon \right) \to 0. \tag{21}$$

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

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

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

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

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

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

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

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

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

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

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

Equation (21) holds by choosing $\delta = \sqrt{\rho_{\max}}\epsilon/2$.

B.6. Least-Squares Estimator

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

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

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

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

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

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

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

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

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

$$\hat{\theta}_T^{\text{AW-LS}} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmax}} \left\{ - \sum_{t=1}^{T} W_t \left(Y_t - \phi(X_t, A_t)^\top \theta \right)^2 \right\}. \tag{22}$$

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

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

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

Proof: By taking the derivative of Equation (22) with respect to the parameters, we have that

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

By rearranging terms, we have that

$$- \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right) \\ = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P}) \right). \tag{23}$$

We first show that the following holds:

$$\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{24}$$

Equation (24) holds by a similar argument as that used in Section B.4.2, for $\dot{m}_\theta(Y_t, X_t, A_t) = \phi(X_t, A_t) \left(Y_t - \phi(X_t, A_t)^\top \theta^*(\mathcal{P}) \right)$ by showing that the conditions of Theorem 2 hold. It can be checked that all the arguments hold even when we allow $\rho_{\max,T}$ to grow at a rate such that $\frac{\rho_{\max,T}}{T} \to 0$.

By Equations (23) and (24),

$$\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2} \left(\hat{\theta}_T^{\text{AW-LS}} - \theta^*(\mathcal{P}) \right) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}. \tag{25}$$

By Equation (25), to ensure that $\hat{\theta}_T^{\text{AW-LS}} \xrightarrow{P} \theta^*(\mathcal{P})$ uniformly over $\mathcal{P} \in \mathbf{P}$, it is sufficient to show that the minimum eigenvalue of $\Sigma_T(\mathcal{P})^{-1/2} \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2}$ goes to infinity uniformly over $\mathcal{P} \in \mathbf{P}$ as $T \to \infty$.

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

Note that by Lemma 1 and Condition 11,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t \phi(X_t, A_t)^{\otimes 2} - \mathbb{E}_{\mathcal{P},\pi} \left[W_t \phi(X_t, A_t)^{\otimes 2} \middle| \mathcal{H}_{t-1} \right] = O_{\mathcal{P} \in \mathbf{P}}(1). \tag{26}$$

Note that by law of iterated expectations,

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

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

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

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

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

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

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

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

By law of iterated expectations,

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

The above result and Equation (26) implies that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}W_t\phi(X_t,A_t)^{\otimes 2}\succeq O_{\mathcal{P}\in\mathbf{P}}(1)+\sqrt{\frac{T}{\rho_{\max,T}}}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\phi(X_t,A_t)^{\otimes 2}\right]. \tag{27}$$

By Condition 11, the minimum eigenvalue of $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\phi(X_t,A_t)^{\otimes 2}\right]$ is bounded above some constant greater than zero for all $\mathcal{P}\in\mathbf{P}$. By Condition 12, $\sqrt{\frac{T}{\rho_{\max,T}}}\to\infty$. Thus by Equation (25) and Equation (27), we have that $\hat{\theta}_T^{\text{AW-LS}}\overset{P}{\to}\theta^*(\mathcal{P})$ uniformly over $\mathcal{P}\in\mathbf{P}$.

C. Choice of Stabilizing Policy

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

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

$$\hat{\theta}^{\text{AW-LS}} := \underset{\theta\in\Theta}{\operatorname{argmax}} \left\{ \frac{1}{T} \sum_{t=1}^{T} W_t \left(Y_t - X_t^\top \theta_{A_t} \right)^2 \right\}. \tag{28}$$

Recall that our use of adaptive weights is to adjust for instability in the variance of M-estimators induced by the bandit algorithm in order to construct valid confidence regions; note that weighted estimators are not typically used for this reason. On i.i.d. data, the least-squares criterion is weighted like in Equation (28) in order to minimize the variance of estimators under noise heteroskedasticity; in this setting, the best linear unbiased estimator has weights $W_t = 1/\sigma^2(A_t, X_t)$ where $\sigma^2(A_t, X_t) := \mathbb{E}_{\mathcal{P}}[(Y_t - X_t^\top \theta^*_{A_t}(\mathcal{P}))^2 | X_t, A_t]$; this up-weights the importance of observations with low noise variance. Intuitively, if we do not need to variance stabilize, $\{W_t\}_{t\geq 1}$ should be determined by the relative importance of minimizing the errors for different observations, i.e., their noise variance.

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

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

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

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

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

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

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

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

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

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

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

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

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

C.2. Approximating the Optimal Stabilizing Policy

One way to approximately choose the optimal evaluation policy is to select $\pi_t^{\text{sta}}(a,x) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$. Note that $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ depends on the $\mathcal{P}$, which is unknown. Thus it is natural to choose $\pi_t^{\text{sta}}(a,x)$ to be $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,x,\mathcal{H}_{t-1})]$ weighted by a prior on $\mathcal{P}$. Note that as long as the evaluation policy ensures that weights W_t are bounded, the choice of evaluation policy does not affect the asymptotic validity of the estimator.

In Figure 4, we display the difference in mean squared error for the AW-LS estimator in a two-armed bandit setting for two different choices of evaluation policy: (1) the uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) the expected $\pi_t(a,\mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$. We can see in this setting that by setting $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$ we are able to decrease the mean squared error of the AW-LS estimator compared AW-LS with the uniform evaluation policy. Note though that in some cases setting $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$ is equivalent to choosing the uniform evaluation policy. For example, a two-armed bandit with identical arms so under common bandit algorithms $\mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})] = 0.5$ for all $t \in [1\colon T]$, which will make the evaluation policy $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a,\mathcal{H}_{t-1})]$ equivalent to the uniform policy.

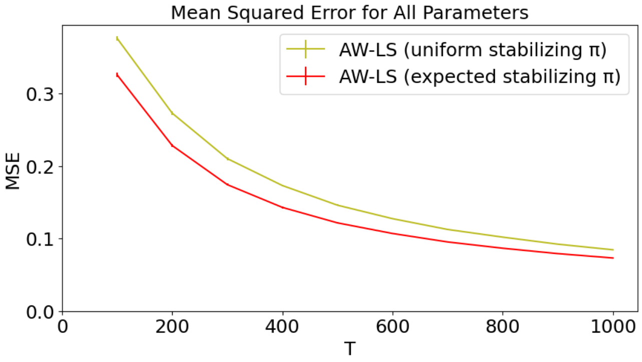

Figure 4. Above we plot the mean squared errors for the adaptively-weighted least squares estimator with evaluation policies: (1) uniform evaluation policy which selects actions uniformly from $\mathcal{A}$ and (2) expected $\pi_t(a,\mathcal{H}_{t-1})$ evaluation policy for which $\pi_t^{\text{sta}}(a) = \mathbb{E}_{\mathcal{P},\pi}[\pi_t(a)]$ (oracle quantity). In a two arm bandit setting we perform Thompson Sampling with standard normal priors, 0.01 clipping, $\theta^*(\mathcal{P}) = [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})] = [0,1]$, standard normal errors, and $T = 1000$. Error bars denote standard errors computed over 5,000 Monte Carlo simulations.

References

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

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

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

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

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

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

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

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

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

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

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

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