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

James
Leiner

Carnegie Mellon University

Based on joint work with

Robin
Dunn

Novartis
Pharmaceutical
Corporation

Aaditya
Ramdas

Carnegie
Mellon

December 17, 2025

Consider the contextual bandit problem

- Suppose features (X_t), actions (A_t), and outcomes (Y_t) are observed sequentially
- At each time step t, the analyst chooses $\mathbb{P}\left(A_t = a \,|\, X_t, \mathscr{H}_{t-1}\right)$ for each a in a finite action space $\mathscr{A}$ given a pooled history $\mathscr{H}_{t-1} = \{(X_i, A_i, Y_i\}_{i=1}^{t-1}$.
- Denoting $Y_t(a)$ as the *potential outcome* had action a been chosen at time step t, we assume that

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

- We summarize the joint distribution at time step t z

$$(X_t, A_t, Y_t) \sim p\left(y \,|\, x, a\right) \mathbb{P}\left(A_t = a \,|\, X_t, \mathscr{H}_{t-1}\right) p(x)$$

Even in the well-specified linear case, adaptive data collection can make inference difficult

Suppose that

$$Y_t = \theta^T Z_t + \epsilon_t$$

where

- $Z_t = \phi(X_t, A_t)$ for some feature map $\phi : \mathbb{R}^p \times \mathscr{A} \to \mathbb{R}$
- ϵ_t is a random variable such that $\mathbb{E}[\epsilon_t \,|\, \mathscr{H}_{t-1}] = 0$, $\mathbb{E}[\epsilon_t^2 \,|\, \mathscr{H}_{t-1}] = \sigma^2$

Fact (Lai and Wei, 1982)

A sufficient condition for the OLS estimate $\hat{\theta}$ to be asymptotically normal is for there to exist a sequence of positive definite matrices $\{B_T\}_{T=1}^{\infty}$ such that

$$B_T^{-1} \sum_{t=1}^{T} Z_t Z_t^T \xrightarrow{p} I_d$$

This often will not be satisfied in bandit settings

This condition is often not obtained in bandit problems where the difference in expected rewards across arms is zero

Example

Let $A_t \in \{0,1\}$ with $\mathbb{E}[Y_t | A_t = 1] = \mathbb{E}[Y_t | A_t = 0]$

In this setup, Zhang et al. (2020) demonstrate that the OLS estimator will be **non-Gaussian** when data is collected using standard bandit algorithms such as epsilon-greedy, Thompson sampling, and UCB.

This is a well studied problem in the literature. Some solutions include:

- Estimating $\hat{\theta}$ across batches with the batch size tending to ∞ (Zhang et al., 2020).
- Adding bias correction term when estimating $\hat{\theta}$ (Deshpande et al., 2018; Khamaru et al., 2023)

But all approaches still assume the linear model is true

What can be said for generic M (and Z)- estimators

This is a less studied problem in the literature.

- Zhang et al. (2021) attempt to cover a target parameter that exists only when the conditional mean under the working model is **correctly specified.** That is,

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

- In a parallel work to our own, Guo and Xu (2025) cover separate $\{\theta_a^\star\}_{a\in\mathscr{A}}$ that solve for the roots

$$0 = \mathbb{E}\left[m_{\theta_a^\star}(X_t, Y_t(a))\right] \text{ for all } t \in [T]$$

Their work does allow for **model misspecification**, but:

- Assumes that the policy $\mathbb{P}(A_t = a \,|\, X_t, \mathscr{H}_{t-1})$ converges to a deterministic function independent of history.
- Does not allow for a model with a lower-dimensional $\theta^\star$ that is defined *across* actions.

What target makes sense under misspecification?

Our Choice: $\theta^\star := \text{argmax}_{\theta \in \Theta} \mathbb{E}_{\mathscr{P}, A \sim \pi_e} \left[m_\theta(X, A, Y) \right]$

The evaluation policy $\pi_e(a|\text{x})$ is a choice of density that is independent of history.

If the policy converges, letting
$\pi_e(A_t = a \,|\, X_t) := \lim_{t \to \infty} \mathbb{P}(A_t = a \,|\, X_t, \mathscr{H}_{t-1})$ is perfectly sensible.

If the policy does not converge (e.g. multi-armed bandits when expected rewards across arms is comparable), it an **interpretative choice**. Some examples:

- Uniform over the action space;
- Weighting certain actions based on prior assumptions about efficacy'
- Using some known deployment policy.

Choice of evaluation policy is crucial for model interpretation when it is misspecified

Assume $Y_t \sim N(6A_t^2, 1)$ but we erroneously assume a linear model

- Policy 1: $p_e(a \mid x)$ is uniform over $\{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
- Policy 2: $p_e(a \mid x)$ is uniform over $\{0.6, 0.7, 0.8, 0.9, 1.0\}$

Correct Specification

Let $m_\theta = (Y_t - \theta_0 - \theta_1 A_t^2)^2$

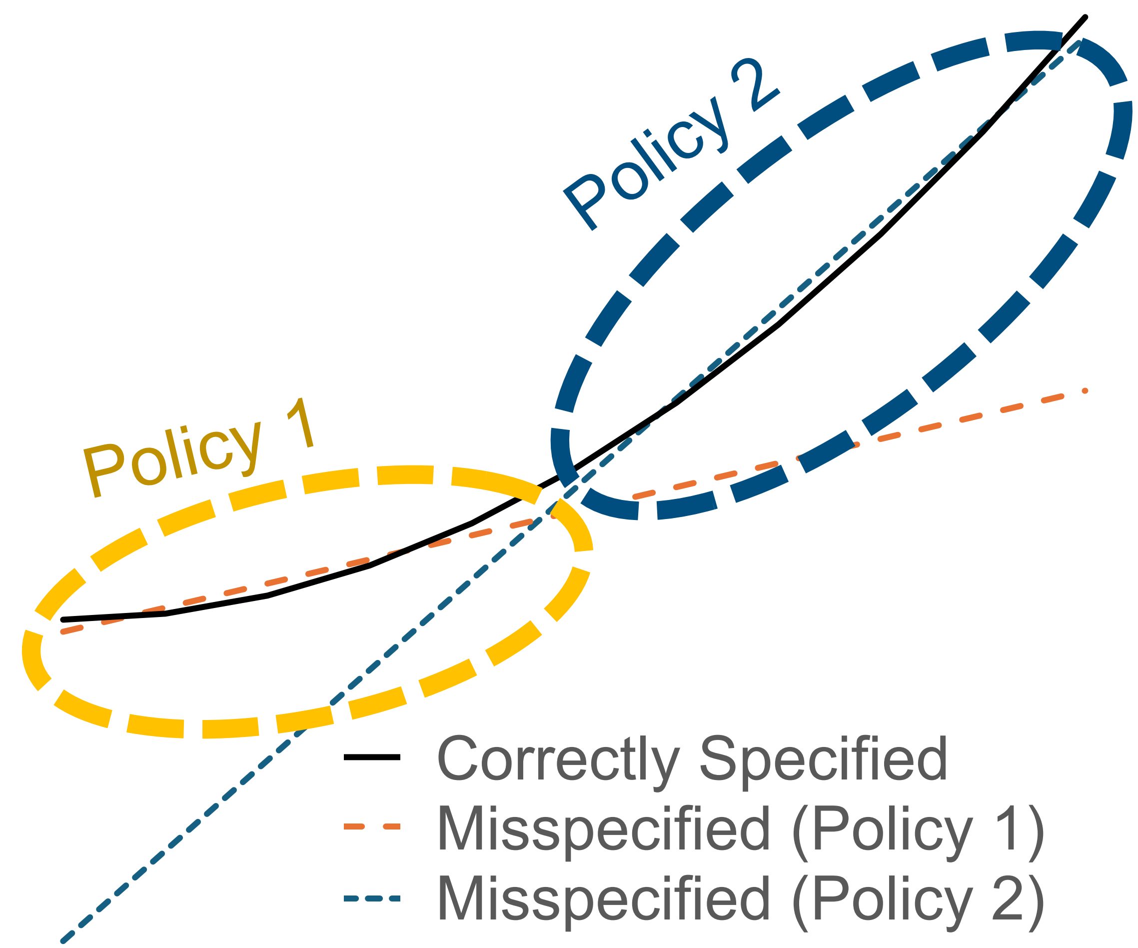
$\theta_0^\star = 0$ and $\theta_1^\star = 6$ under both policies

Misspecification

Let $m_\theta = (Y_t - \theta_0 - \theta_1 A_t)^2$

Policy 1: $\theta_0^\star = -0.2$ ***and*** $\theta_1^\star = 4$

Policy 2: $\theta_0^\star = -5.3$ ***and*** $\theta_1^\star = 15$



Let us first consider a naive estimator

Let us consider $\hat{\theta}_0 := \text{argmax}_\theta \sum_{t=1}^{T} w_t m_\theta(X_t, A_t, Y_t)$.

Assume that $\hat{\theta}_0$ corresponds to the solution to the estimating equation

$$0 = \sum_{t=1}^{T} w_t \dot{m}_\theta m_\theta(X_t, A_t, Y_t)$$

Assume $\theta^\star$ corresponds to the root of $0 = \mathbb{E}_{A\sim\pi_e}\left[\dot{m}_\theta(X, A, Y)\right]$

Generic Strategy

1. Show $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \dot{m}_{t,\theta^\star}(X_t, A_t, Y_t) \xrightarrow{d} N(0, I_d)$ using martingale CLT
2. Taylor expand around this quantity to form a confidence ellipsoid center at $\theta^\star$
3. Prove $\hat{\theta}_0 \xrightarrow{p} \theta^\star$ to allow for plug-ins

Proving a martingale CLT is non-trivial in the misspecified, adaptive setting

Desiderata: $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} w_t \dot{m}_{t,\theta^\star}(X_t, A_t, Y_t) \xrightarrow{d} N(0, I_d)$

We need to check that for any fixed $c \in \mathbb{R}^d$ and $\epsilon > 0$

1. **Martingale Difference Sequence**

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

2. **Conditional Variance Converges**

   $$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[c^T w_t^2 \dot{m}_{\theta^\star}(X, A, Y) \dot{m}_{\theta^\star}(X_t, A_t, Y_t)^T c \mid \mathscr{H}_{t-1}\right] \xrightarrow{p} \sigma_c^2$$

   for some fixed σ_c^2

3. **Asymptotic Negligibility**

   $$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[c^T w_t^2 \dot{m}_{\theta^\star}(X, A, Y) \dot{m}_{\theta^\star}(X, A, Y)^T c 1_{|w_t \dot{m}_{\theta^\star}(X,A,Y) > \epsilon|} \mid \mathscr{H}_{t-1}\right] \xrightarrow{p} 0$$

Both conditions tend to fail under misspecification + adaptivity

Controlling the first moment requires inverse propensity weighting

Assume that $w_t \in \sigma(\mathscr{H}_{t-1}, X_t)$

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

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

(but $\mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t, A_t\right] = 0$ under correct specification)

So $\mathbb{E}\left[w_t \dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid \mathscr{H}_{t-1}\right] = 0$ for all $w_t \in \sigma(\mathscr{H}_{t-1}, X_t)$ when model is **correctly specified**

If the **model is misspecified**, pick $w_t = \dfrac{p_e(A_t = a \mid X_t)}{\mathbb{P}\left(A_t = a \mid \mathscr{H}_{t-1}, X_t\right)}$. Then,

$$\mathbb{E}\left[\mathbb{E}\left[\frac{p_e(A_t = a \mid X_t)}{\mathbb{P}\left(A_t = a \mid \mathscr{H}_{t-1}, X_t\right)} \dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t, A_t, \mathscr{H}_{t-1}\right] \mid \mathscr{H}_{t-1}\right] = \mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\right]$$

$$= 0$$

Controlling the second moment requires square root IPW-weighting

On the other hand, $\mathbb{E}\left[w_t^2 \dot{m}_{\theta^\star}(X,A,Y)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T \,|\, \mathscr{H}_{t-1}\right]$ may also be quite unstable in bandit settings under a **non-converging** policy.

If w_t has **not already been used** to control the first moment, we can simply let $w_t = \left(\frac{p_e(A_t = a \,|\, X_t)}{\mathbb{P}\left(A_t = a \,|\, \mathscr{H}_{t-1}, X_t\right)}\right)^{1/2}$ (Zhang et al., 2021).

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[w_t^2 \dot{m}_{\theta^\star}(X,A,Y)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T \,|\, \mathscr{H}_{t-1}\right] &= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{p_e(A_t = a \,|\, X_t)}{\mathbb{P}\left(A_t = a \,|\, \mathscr{H}_{t-1}, X_t\right)}\dot{m}_{\theta^\star}(X,A,Y)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T \,|\, \mathscr{H}_{t-1}\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X,A,Y)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T\right] \\
&= \mathbb{E}_{A_t\sim\pi_e}\left[\dot{m}_{\theta^\star}(X,A,Y)\dot{m}_{\theta^\star}(X_t,A_t,Y_t)^T\right]
\end{aligned}
$$

… but this is only a tractable strategy under correct specification.

Controlling the first two moments simultaneously requires additional free parameters

Consider nested filtrations

$$\sigma(\mathscr{H}_{t-1}) \subseteq \sigma(\mathscr{H}_{t-1}, X_t) \subseteq \sigma(\mathscr{H}_{t-1}, X_t, A_t) \subseteq \sigma(\mathscr{H}_{t-1})$$

Choose $\Sigma_t \in \sigma(\mathscr{H}_{t-1})$ to **stabilize the variance**

Choose $w_t \in \sigma(X_t, \mathscr{H}_{t-1})$ to **ensure the score function is a MDS** (i.e. control first moment)

This is a viable strategy if we can estimate the time-varying variance well

Let $w_t = \dfrac{p_e(A_t = a \mid X_t)}{\mathbb{P}\left(A_t = a \mid \mathscr{H}_{t-1}, X_t\right)}$

Let Σ_t be an estimate of $\mathbb{E}\left[s_{t,\theta^\star} s_{t,\theta^\star}^T \mid \mathscr{H}_{t-1}\right]$ for $s_{t,\theta} := w_t \dot{m}_{\theta^\star}(X_t, A_t, Y_t)$

Checking first condition…

$$\begin{aligned}
\mathbb{E}[\Sigma_t^{-1/2} s_{t,\theta^\star} \mid \mathscr{H}_{t-1}] &= \Sigma_t^{-1/2}\mathbb{E}[s_{t,\theta^\star} \mid \mathscr{H}_{t-1}] \\
&= \Sigma_t^{-1/2}\mathbb{E}\left[\mathbb{E}\left[w_t \dot{m}_{\theta^\star}(X_t, A_t, Y_t) \mid X_t, A_t, \mathscr{H}_{t-1}\right] \mathscr{H}_{t-1}\right] \\
&= \Sigma_t^{-1/2}\mathbb{E}_{A_t \sim \pi_e}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t)\right] \\
&= 0
\end{aligned}$$

Checking second condition…

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\Sigma_t^{-1/2} s_{t,\theta^\star} s_{t,\theta^\star}^T \Sigma_t^{-1/2} \mid \mathscr{H}_{t-1}] &= \frac{1}{T}\sum_{t=1}^{T}\Sigma_t^{-1/2}\mathbb{E}[s_{t,\theta^\star} s_{t,\theta^\star}^T \mid \mathscr{H}_{t-1}]\Sigma_t^{-1/2} \\
&\approx \frac{1}{T}\sum_{t=1}^{T}\Sigma_t^{-1/2}\Sigma_t\Sigma_t^{-1/2} \\
&\approx I_d
\end{aligned}$$

(assuming sufficiently good estimate…)

Our final estimator also allows for the use of flexible ML to reduce variance further

We define the MAIPWM (misspecified augmented inverse propensity weighted M-) estimator as

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

where $f_t : \mathbb{R} \times \mathcal{A}$ is trained only on $\mathcal{H}_{t-1}$ and targets conditional mean $\mathbb{E}\left[Y_t \mid X_t, A_t\right]$.

- If f_t is accurate, we empirically find a significant reduction in the variance of the estimator.
- If no predictive model is available, letting $f_t = c$ for any constant reduces the estimate to $\hat{\theta}_0$.
- Similar constructions are commonly used when targeting ATE (Hadad but M-estimators is a more recent application of the idea (Zrnic and Candes, 2024).

Note: it is not required to train f_t for our results to hold, but we do find a significant reduction in variance empirically.

We prove a CLT assuming that the time-varying variance can be approximated well

Key Assumptions:

1. $V_{t,\theta^\star} := \mathbb{E}_{\mathcal{P},\pi_t}\left[s_{t,\theta^\star}s_{t,\theta^\star}^T \,|\, \mathcal{H}_{t-1}\right]$ is almost surely invertible for each $t \in [T]$.
2. There exists a sequence of estimators $\{\hat{V}_t\}_{t=1}^T$ adapted to the filtration $\sigma(\mathcal{H}_{t-1})$ such that $||\hat{V}_t^{-1/2} - V_{t,\theta^\star}^{-1/2}||_{\text{op}} \xrightarrow{p} 0$.
3. There exists a constant C such that $\frac{p_e(A_t = a \,|\, X_t)}{\mathbb{P}\left(A_t = a \,|\, \mathcal{H}_{t-1}, X_t\right)} < C$

Theorem (simplified)

Assume $\frac{1}{T}\sum_{t=1}^{T} \hat{V}_t^{-1/2} s_{t,\hat{\theta}_T} = o_p(1/\sqrt{T})$ and the eigenvalues of both $V_{t,\theta^\star}$ and $\hat{V}_t$ are bounded above and below. Then under appropriate regularity conditions (e.g. bracketing entropy, well-separated solutions):

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

How do we estimate $\hat{V}_t^{-1/2}$?

We can use the law of total variance to decompose $V_{t,\theta^\star}$ into pieces that are either **invariant to history** or **known by the experimenter.**

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

Assumption: We have access to an external dataset $\{\tilde{X}_i\}_{i=1}^n$, independent of $\mathscr{H}_{t-1}$ such that $\tilde{X}_i \overset{\text{iid}}{\sim} p(x)$.

(alternatively can use sequential sample splitting)

Strategy (simplified):

- Learn model $f_t : \mathbb{R}^p \times \mathscr{A} \to \mathbb{R}^d$ s.t. $f_t(a, X_t) - \mathbb{E}\left[\dot{m}_{\theta^\star}(X_t, A_t, Y_t) \,|\, X_t, A_t = a\right] \xrightarrow{p} 0$
- Learn model $g_t : \mathbb{R} \times \mathscr{A} \to \mathbb{R}^{d\times d}$ s.t. $e_t(a, X_t) - \mathbb{E}\left[\dot{m}_\theta^\star(a, X_t, Y_t(a))\dot{m}_\theta^\star(a, X_t, Y_t(a))^T \,|\, X_t\right] \xrightarrow{p} 0$
- Plug external data into these models to estimate $\mathrm{Var}\left(s_{t,\theta^\star} \,|\, \mathscr{H}_{t-1}\right)$

Results

We test the methodology with semi-synthetic datasets

- The Osteoarthritis Initiative (OAI) is a ten year longitudinal study tracking long-term outcomes of patients with osteoarthritis.
  - Y_t is the four year change in KL grade (measure of knee health)
  - X_t includes baseline measurements of knee health, demographic variables, baseline health risk factors
  - A_t is not available (observational study), so we create semi-synthetic dataset

Create semi-synthetic dataset

- We use random forests to train a model $f(X_t)$ estimating $\mathbb{E}\left[Y_t \mid X_t\right]$
- For user-chosen parameters $\{\beta_1^a\}_{a\in\mathscr{A}}$ and $\{\beta_2^a\}_{a\in\mathscr{A}}$, we generate semi-synthetic outcomes where

$$\mathbb{E}[Y_t \mid X_t, A_t] = \sum_{a=1}^{K} \beta_1^a 1_{A_t=a} + \sum_{a=1}^{K} \beta_2^a f(X_t) \times 1_{A_t=a}$$

Estimators and sampling strategies

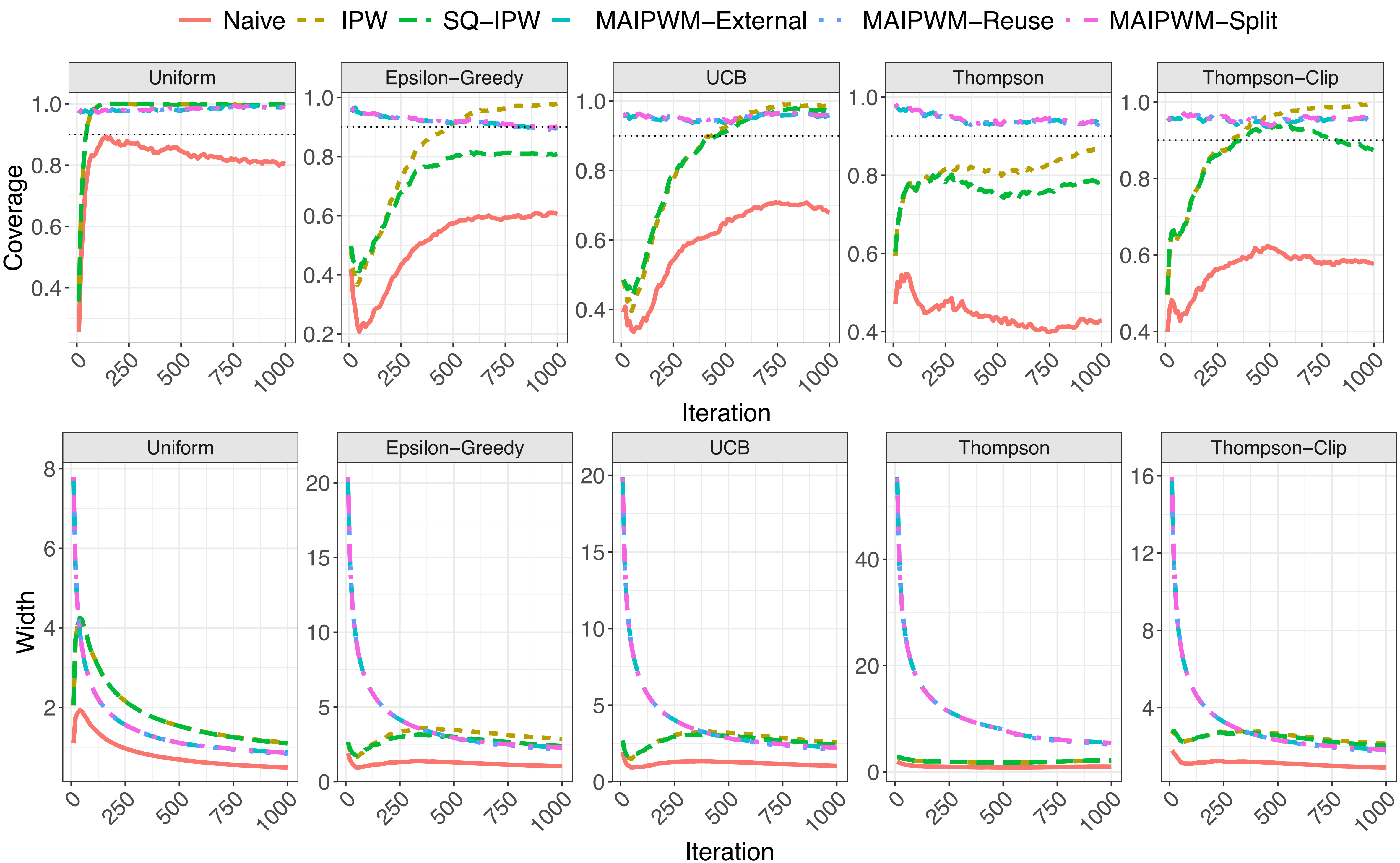
We test the methodology using uniform (i.e. non-adaptive), epsilon-greedy, UCB, and Thompsons sampling.

- We also experiment with clipping the probabilities in the interval $[0.05, 0.95]$

Estimators tested:

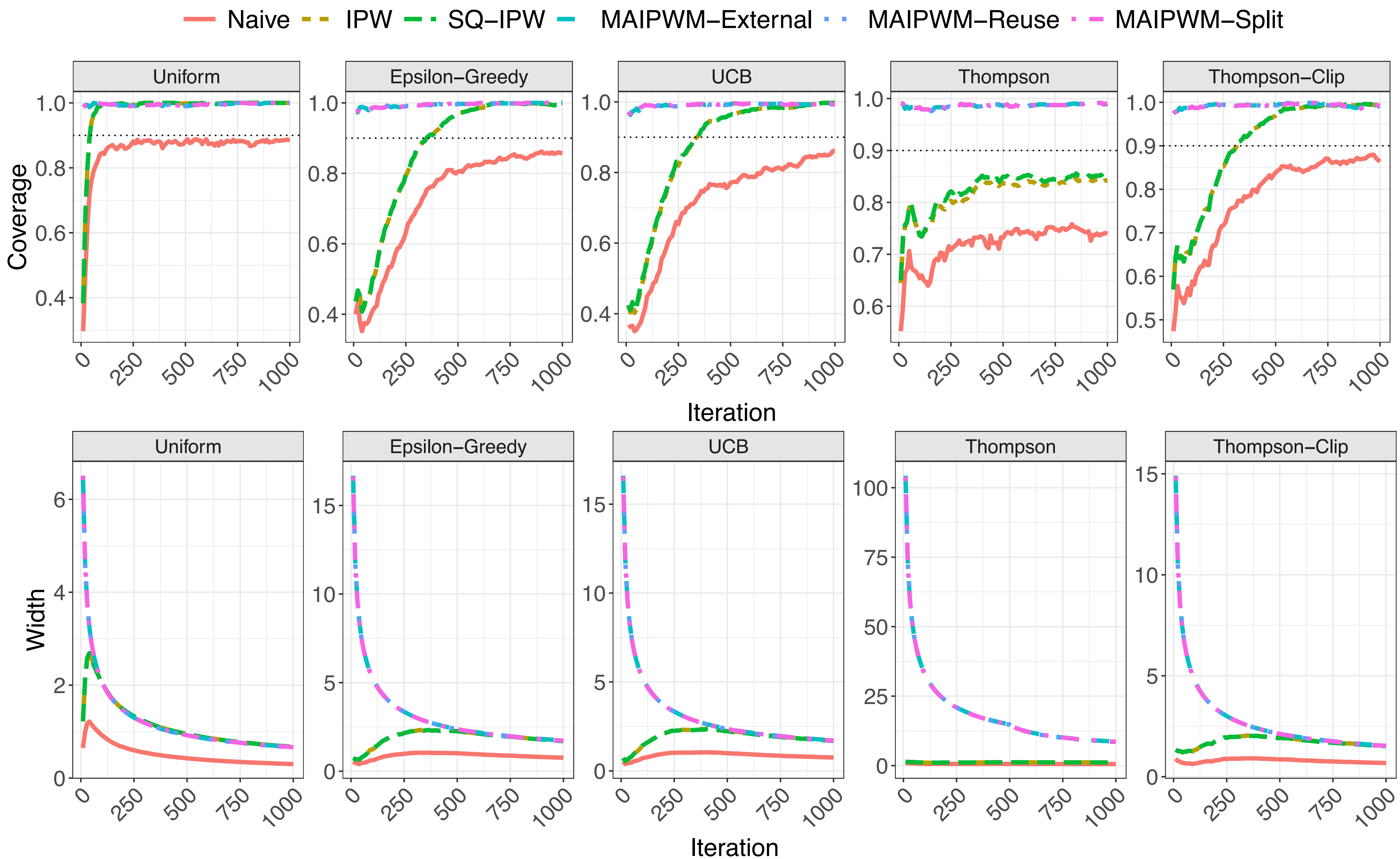
- Naive estimator with $w_t = 1$
- IPW estimator
- Square-root IPW estimator
- MAIPW estimators. We experiment with estimating $\hat{V}_t$ using:
    - External dataset independent of $\mathscr{H}_{t-1}$
    - Sequential sample splitting
    - Reusing $\mathscr{H}_{t-1}$ to estimate the variance *(no theoretical guarantees)*

Misspecified Case



$\beta_1 = (0,0,1,2,2,3,4,4)$ $\quad \beta_2 = (1,-1,1,0,-3,1,1,1)$

Correctly Specified Case



$\beta_1 = (0,1,2,3,4,5,6,7) \quad \beta_2 = (0,0,0,0,0,0,0,0)$

Thank you!