Statistical Inference on Bandit Data

Kelly Zhang, Lucas Janson, Susan Murphy

April 29, 2021

Harvard John A. Paulson School of Engineering and Applied Sciences

Real World Sequential Decision Making Problems

- Mobile Health
    - Promote behavior change using personalized nudge messages
    - Learn when to send messages based on user’s context
- Online Education
    - Improve online learning experience by optimizing which teaching strategies are used

Objectives in Sequential Decision Making Problems

Regret Minimization

- Maximizing welfare of experimental population
- Personalize to provide best user experience
- Bandit algorithms designed to optimize this objective

Causal Inference Objective

- Use data collected by sequential decision making algorithm to gain generalizable knowledge
- For example, construct confidence intervals for a treatment effect

Objectives in Sequential Decision Making Problems

Regret Minimization

- Maximizing welfare of experimental population
- Personalize to provide best user experience
- <u>Bandit algorithms designed to optimize this objective</u>

Causal Inference Objective

- Use data collected by sequential decision making algorithm to gain knowledge
- For example, construct confidence intervals for a treatment effect

Contextual Bandit Problem

Variables:

- X_t are **contexts**
- A_t are **actions**
- Y_t are **outcomes**
- $R_t = f(Y_t)$ are **rewards**

Bandit Objective: Select actions to maximize total expected reward

$$E_\pi \left[\sum_{t=1}^{T} R_t(A_t) \right]$$

Contextual Bandit Examples

Online Education

- **Actions** A_t : teaching strategies
- **Context** X_t : student background, recent progress, ect.
- **Outcome** Y_t : student performance on quizzes or homework

Contextual Bandit Examples

Online Education

- **Actions** A_t : teaching strategies
- **Context** X_t : student background, recent progress, ect.
- **Outcome** Y_t : a student performance on a quizzes, homework

Online Advertising

- **Actions** A_t : different types of ads
- **Context** X_t : type of website, recent user behavior, ect.
- **Outcome** Y_t : click-through rate or amount of money made through the ad

Contextual Bandit Environment

- **Potential Outcomes:**
  $\left\{X_t,\ Y_t(a) : a \in \mathscr{A}\right\}_{t=1}^{T}$ i.i.d. over t

- **History:** $H_{t-1} = \left\{X_s, A_s, Y_s\right\}_{s=1}^{t-1}$

- Bandit algorithm determines **action selection probabilities:**
  $\pi_{t,a} = P\left(A_t = a \,\middle|\, H_{t-1}, X_t\right)$

Binary Treatment Case

Potential Outcomes	t=1	t=2	t=3	...	t=T
Contexts	X_1	X_2	X_3	...	X_T
Potential Outcomes Under Treatment 0	$Y_1(0)$	$Y_2(0)$	$Y_3(0)$	...	$Y_T(0)$
Potential Outcomes Under Treatment 1	$Y_1(1)$	$Y_2(1)$	$Y_3(1)$	...	$Y_T(1)$

Contextual Bandit Environment

- **Potential Outcomes:**
  $\left\{X_t,\ Y_t(a) : a \in \mathscr{A}\right\}_{t=1}^{T}$ i.i.d. over t

- **History:** $H_{t-1} = \left\{X_s, A_s, Y_s\right\}_{s=1}^{t-1}$

- Bandit algorithm determines **action selection probabilities:**
  $\pi_{t,a} = P\left(A_t = a \,\middle|\, H_{t-1}, X_t\right)$

- At the end, we have dataset
  $\left\{X_s, A_s, Y_s\right\}_{s=1}^{T}$

Binary Treatment Case

Potential Outcomes	**t=1**	**t=2**	**t=3**	**...**	**t=T**
Contexts	X_1	X_2	X_3	...	X_T
Potential Outcomes Under Treatment 0	$Y_1(0)$	$Y_2(0)$	$Y_3(0)$	...	$Y_T(0)$
Potential Outcomes Under Treatment 1	$Y_1(1)$	$Y_2(1)$	$Y_3(1)$	...	$Y_T(1)$
Actions Selected by Bandit Algorithm	$A_1 = 0$	$A_2 = 1$	$A_3 = 1$	**...**	$A_T = 0$

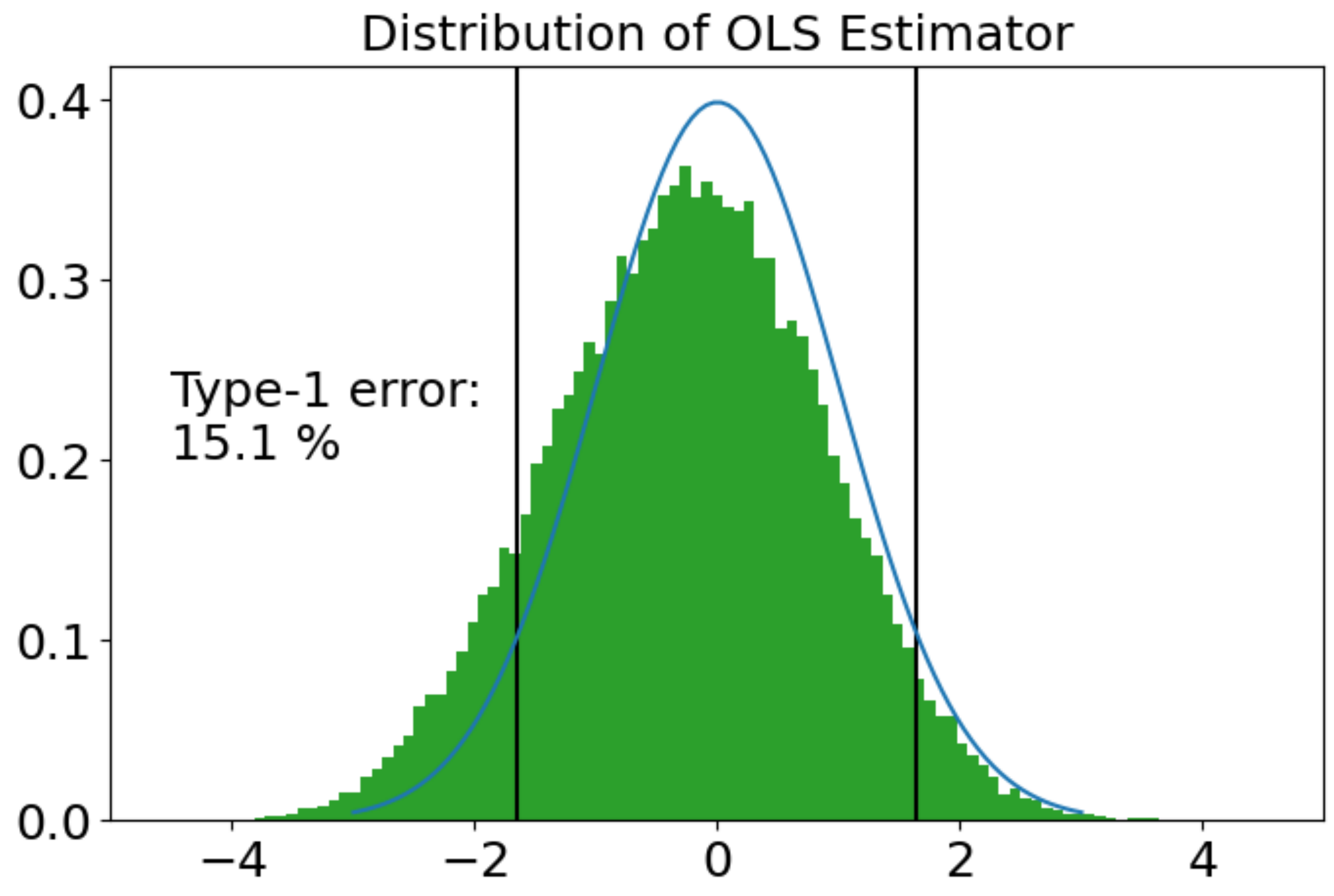
Blue indicates observed data

Bandit observations are not independent

Observations $\{X_t, A_t, Y_t\}$ **are not independent over** $t \in [1:T]$

- We use past observations H_{t-1} to inform what action A_t to select in new context X_t
- Bandit data considered "adaptively collected"

Standard Statistical Estimators Asymptotically Non-Normal on Bandit Data



Data generating process: Two-arm bandit with arm means $\theta^* = [\theta_1^*, \theta_2^*]^\top = [0,0]^\top$. Thompson Sampling. $T = 1000$.

$$\sqrt{\sum_{t=1}^{T} 1_{A_t=1}}\,(\hat{\theta}_{1,T}^{\text{OLS}} - \theta_1^*)$$

Related Work

W-Decorrelated Estimator (Deshpande et. al.)

- Construct confidence intervals for parameters for linear models of expected reward

Adaptively Weighted Augmented-Inverse-Probability Weighted Estimator (Hadad et al.)

- Construct confidence intervals for the expected outcomes in multi-armed bandit setting (no contextual)

Both methods utilize “adaptive weighting”. We show that adaptive weighting can be used to construct confidence regions for more general statistical models (e.g. non-linear models for expected reward).

Statistical Analysis Objectives

- Given dataset collected by a known bandit algorithm $\{X_t, Y_t, A_t\}_{t=1}^{T}$
- Examples of Outcome Models
  - **Linear Model:** $E[Y_t | X_t, A_t] = X_t^\top \theta_0 + A_t X_t^\top \theta_1$
  - **Logistic Regression Model:** $E[Y_t | X_t, A_t] = \left[1 + \exp\left(-X_t^\top \theta_0 - A_t X_t^\top \theta_1\right)\right]$
  - **Generalized Linear Model**
- **M-estimators** encompass many estimators including **least squares** and **maximum likelihood estimators.**

$$\hat{\theta}_T := \text{argmax}_{\theta \in \Theta} \left\{ \sum_{t=1}^{T} m_\theta(Y_t, X_t, A_t) \right\}$$

Adaptive Square-Root Inverse Propensity Weights

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

We choose weights as follows:

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}} = \frac{1}{\sqrt{P(A_t \mid H_{t-1}, X_t)}} \in \sigma\left(H_{t-1}, X_t, A_t\right)$$

W_t are **adaptive** because they depend on history H_{t-1}.

Weighted Least Squares (Example M-Estimator)

Linear Model for Expected Outcome:

$$E\left[Y_t(1)\middle|X_t\right] = X_t^\top \theta$$

Adaptively Weighted Least Squares Estimator

$$\hat{\theta}^{\mathrm{AW-LS}} = \mathrm{argmax}_\theta \left\{ -\sum_{t=1}^{T} W_t A_t \left(Y_t - X_t^\top \theta\right)^2 \right\}$$

- On independent data weights are used to minimize the variance of the estimator under heteroskedadicity.
- In contrast, adaptive weights are used to "stabilize" the variance of the estimator.

Weighted Least Squares (Example M-Estimator)

By standard Taylor Series arguments

$$\left(\frac{1}{T} \sum_{t=1}^{T} W_t A_t X_t X_t^\top \right) \sqrt{T} \left(\hat{\theta}^{\mathrm{AW-LS}} - \theta^* \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t A_t X_t \left(Y_t - X_t^\top \theta^* \right)$$

- **Approach:** Show right hand side is asymptotically normal by applying a martingale central limit theorem.
- Key condition we need to show is that the “variance stabilizes”.

Conditional Variance With Adaptive Weights

$$E_{\theta^*}\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$$

Conditional Variance With Adaptive Weights

$$E_{\theta^*}\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$$

$$= E_{\theta^*}\left[E_{\theta^*}\left[\frac{1}{\pi_{t,A_t}} A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t\right] \middle| H_{t-1}\right]$$

Law of iterated expectations

Conditional Variance With Adaptive Weights

$$E_{\theta^*}\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$$

$$= E_{\theta^*}\left[E_{\theta^*}\left[\frac{1}{\pi_{t,A_t}} A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t\right] \middle| H_{t-1}\right]$$

Law of iterated expectations

$$= E_{\theta^*}\left[E_{\theta^*}\left[X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t, A_t = 1\right] \middle| H_{t-1}\right]$$

Conditioning on $A_t = 1$

Conditional Variance With Adaptive Weights

$$E_{\theta^*}\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$$

$$= E_{\theta^*}\left[E_{\theta^*}\left[\frac{1}{\pi_{t,A_t}} A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t\right] \middle| H_{t-1}\right]$$

Law of iterated expectations

$$= E_{\theta^*}\left[E_{\theta^*}\left[X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t, A_t = 1\right] \middle| H_{t-1}\right]$$

Conditioning on $A_t = 1$

$$= E_{\theta^*}\left[X_t X_t^\top \left(Y_t(1) - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

Conditional Variance With Adaptive Weights

$$E_{\theta^*}\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right]$$

$$W_t = \frac{1}{\sqrt{\pi_{t,A_t}}}$$

$$= E_{\theta^*}\left[E_{\theta^*}\left[\frac{1}{\pi_{t,A_t}} A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t\right] \middle| H_{t-1}\right]$$

Law of iterated expectations

$$= E_{\theta^*}\left[E_{\theta^*}\left[X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t, A_t = 1\right] \middle| H_{t-1}\right]$$

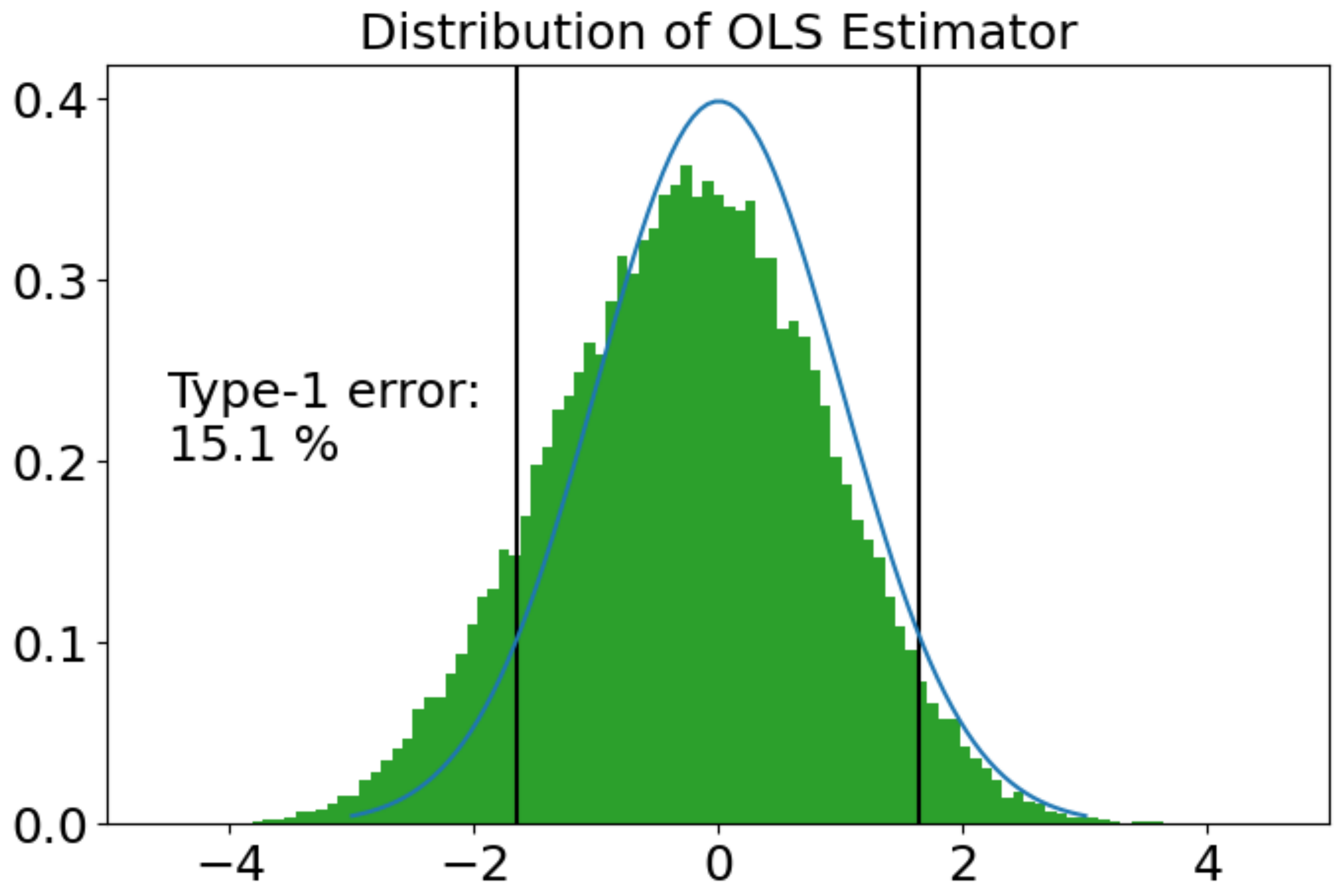
Conditioning on $A_t = 1$

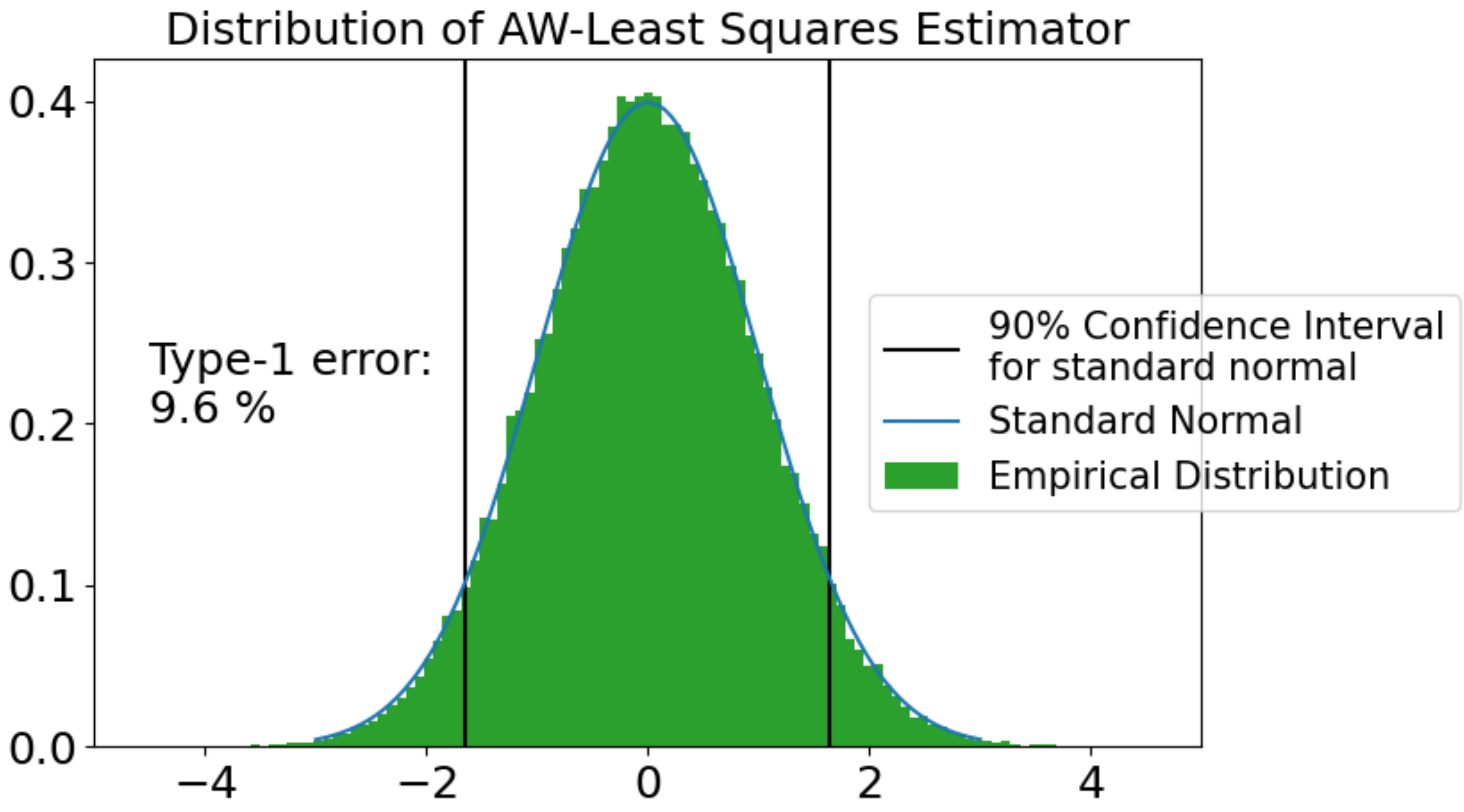
$$= E_{\theta^*}\left[X_t X_t^\top \left(Y_t(1) - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right] = E_{\theta^*}\left[X_t X_t^\top \left(Y_t(1) - X_t^\top \theta_1^*\right)^2\right]$$

i.i.d. Potential Outcomes

Least Squares With and Without Adaptive Weights

Data generating process: Two-arm bandit with arm means $\theta^* = [\theta_1^*, \theta_2^*]^\top = [0,0]^\top$. Thompson Sampling with $N(0,1)$ priors, $N(0,1)$ noise on rewards, and $T = 1000$.





$$\sqrt{\sum_{t=1}^{T} 1_{A_t=1}} (\hat{\theta}_{1,T}^{\text{OLS}} - \theta_1^*)$$

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} W_t 1_{A_t=1} (\hat{\theta}_{1,T}^{\text{AW-LS}} - \theta_1^*)$$

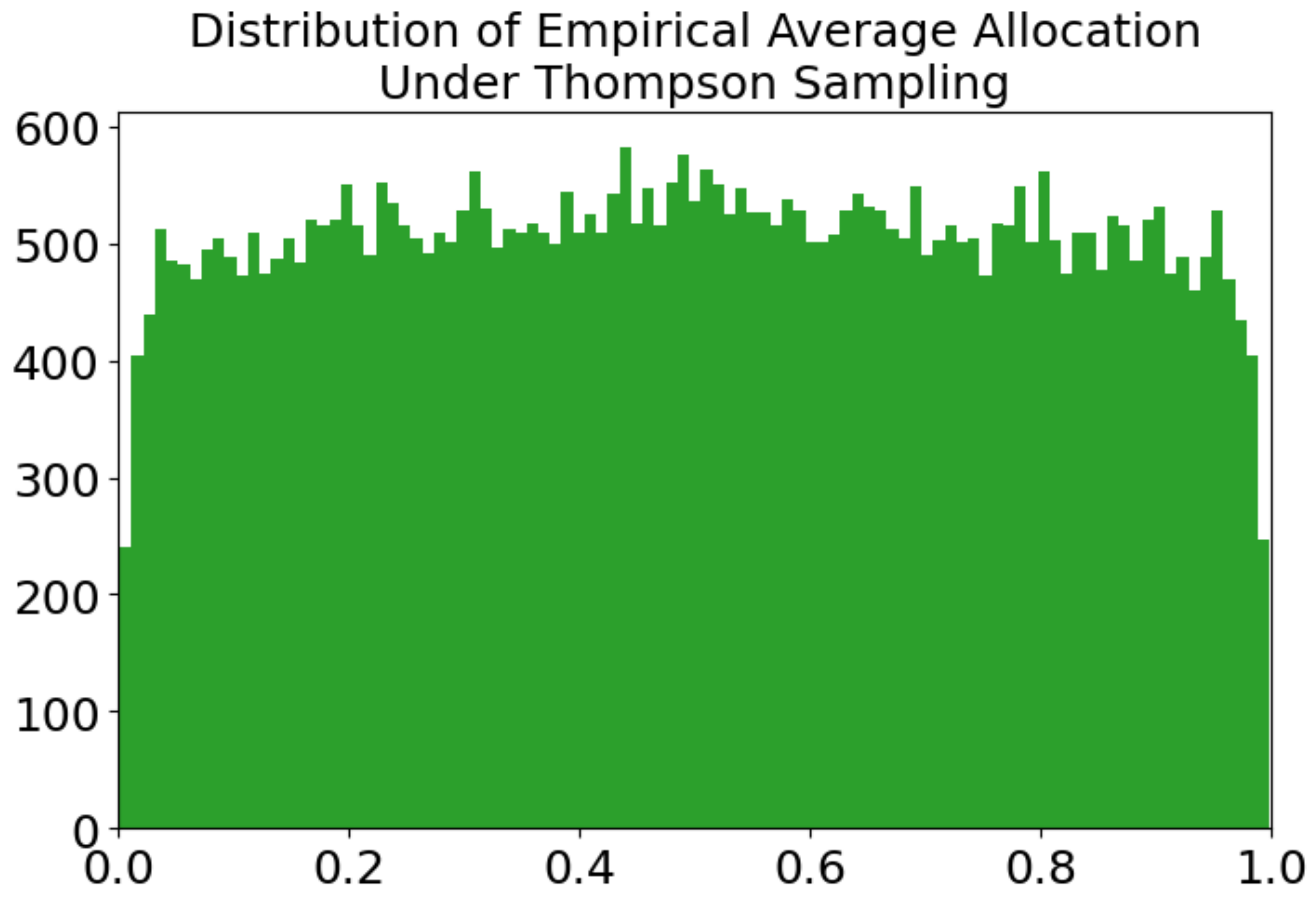
What Goes Wrong without Adaptive Weights?

Consider multi-armed bandit (no context). Let $E\left[Y_t(1)\right] = \theta*$ and $\mathrm{Var}(Y_t(1)) = \sigma^2$.

$$\frac{1}{T}\sum_{t=1}^{T} E_{\theta*}\left[A_t\left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right] = \sigma^2 \frac{1}{T}\sum_{t=1}^{T} \pi_{t,1}$$

Under common bandit algorithms, $\frac{1}{T}\sum_{t=1}^{T}\pi_{t,1}$ is not stable in the limit when there is no unique optimal action.

Data generating process: Two-arm bandit with arm means $\theta* = [\theta_1^*, \theta_2^*]^\top = [0,0]^\top$. Thompson Sampling with $N(0,1)$ priors, $N(0,1)$ noise on rewards, and $T = 1000$.



Empirical Distribution of $\frac{1}{T}\sum_{t=1}^{T}\pi_{t,1}$

Asymptotic Normality Result (abridged)

Estimand: $\theta^* := \operatorname{argmax}_{\theta \in \Theta} \left\{ E_{\theta^*} \left[m_\theta(Y_t, X_t, A_t) \middle| X_t, A_t \right] \right\}$

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

Asymptotic Normality:

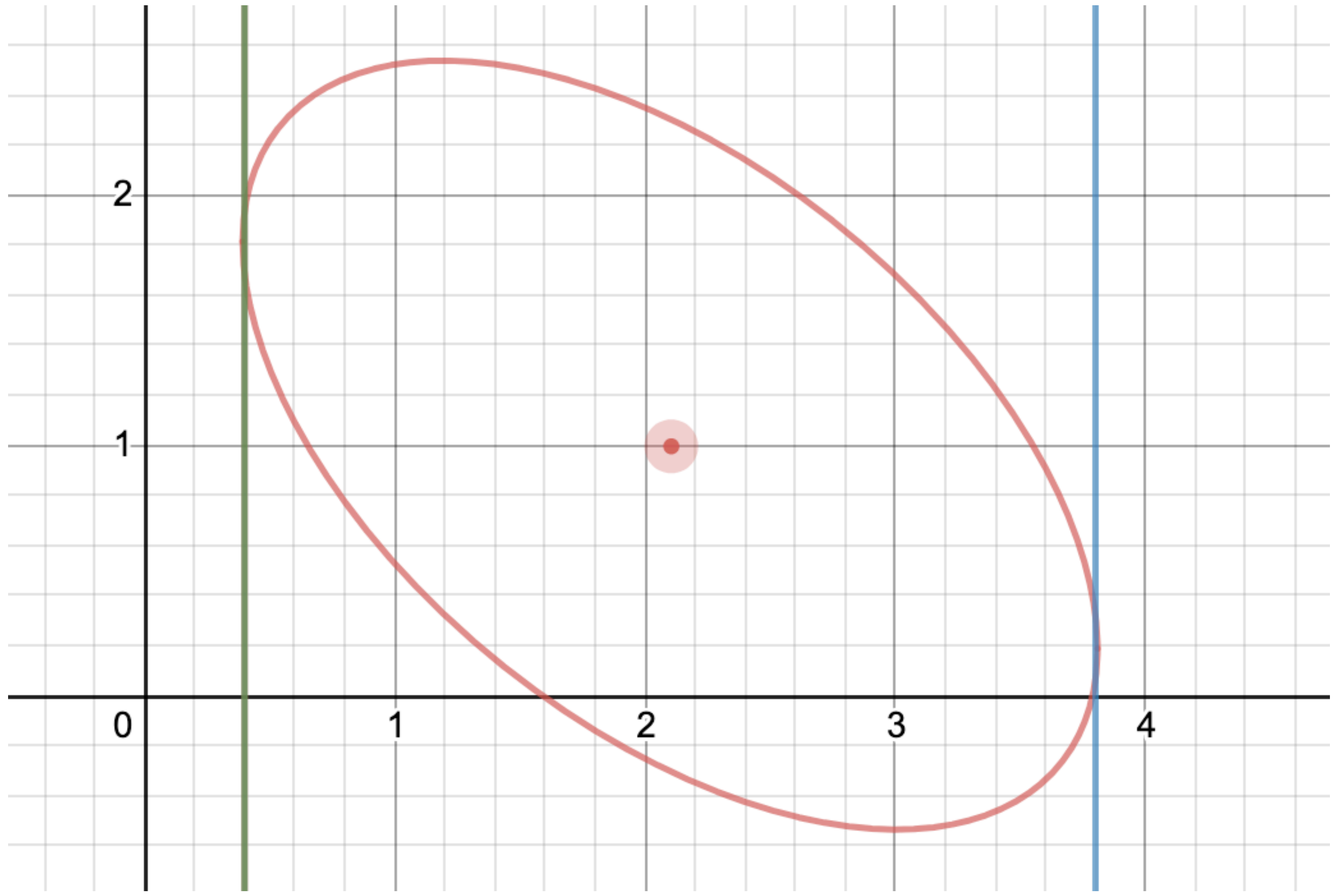
$$\left[\frac{1}{T} \sum_{t=1}^{T} W_t \ddot{m}_{\hat{\theta}_T}(Y_t, X_t, A_t) \right] \sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N} \left(0, E_{\theta^*, \pi^{\text{eval}}} \left[\dot{m}_{\theta^*}(Y_t, X_t, A_t)^{\otimes 2} \right] \right)$$

convergence holds uniformly over $\theta^* \in \Theta$

Projected Confidence Regions

$$\left[\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\hat{\theta}_T}(Y_t, X_t, A_t)\right]\sqrt{T}(\hat{\theta}_T - \theta^*) \xrightarrow{D} \mathcal{N}\left(0, E_{\theta^*,\pi^{\text{eval}}}\left[\dot{m}_{\theta^*}(Y_t, X_t, A_t)^{\otimes 2}\right]\right)$$

- $\frac{1}{T}\sum_{t=1}^{T} W_t \ddot{m}_{\hat{\theta}_T}(Y_t, X_t, A_t)$ does not converge under common bandit algorithms.
- Constructing confidence regions for subsets of parameters of θ^* requires using projections, which are conservative.



Simulation Environment

Environment Details

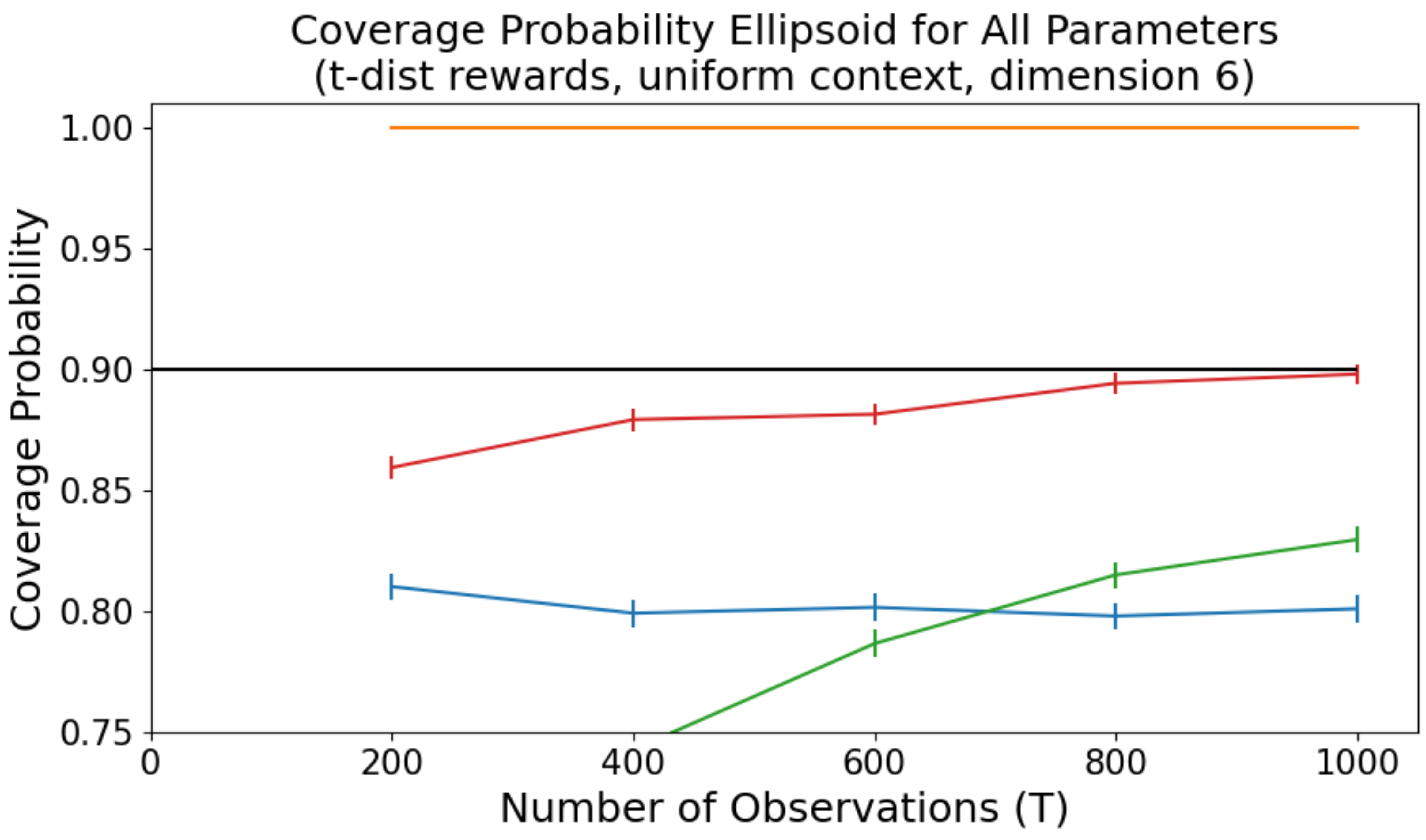
- $\tilde{X}_t = [1, X_t]$ and $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$
- Thompson Sampling contextual bandit algorithm

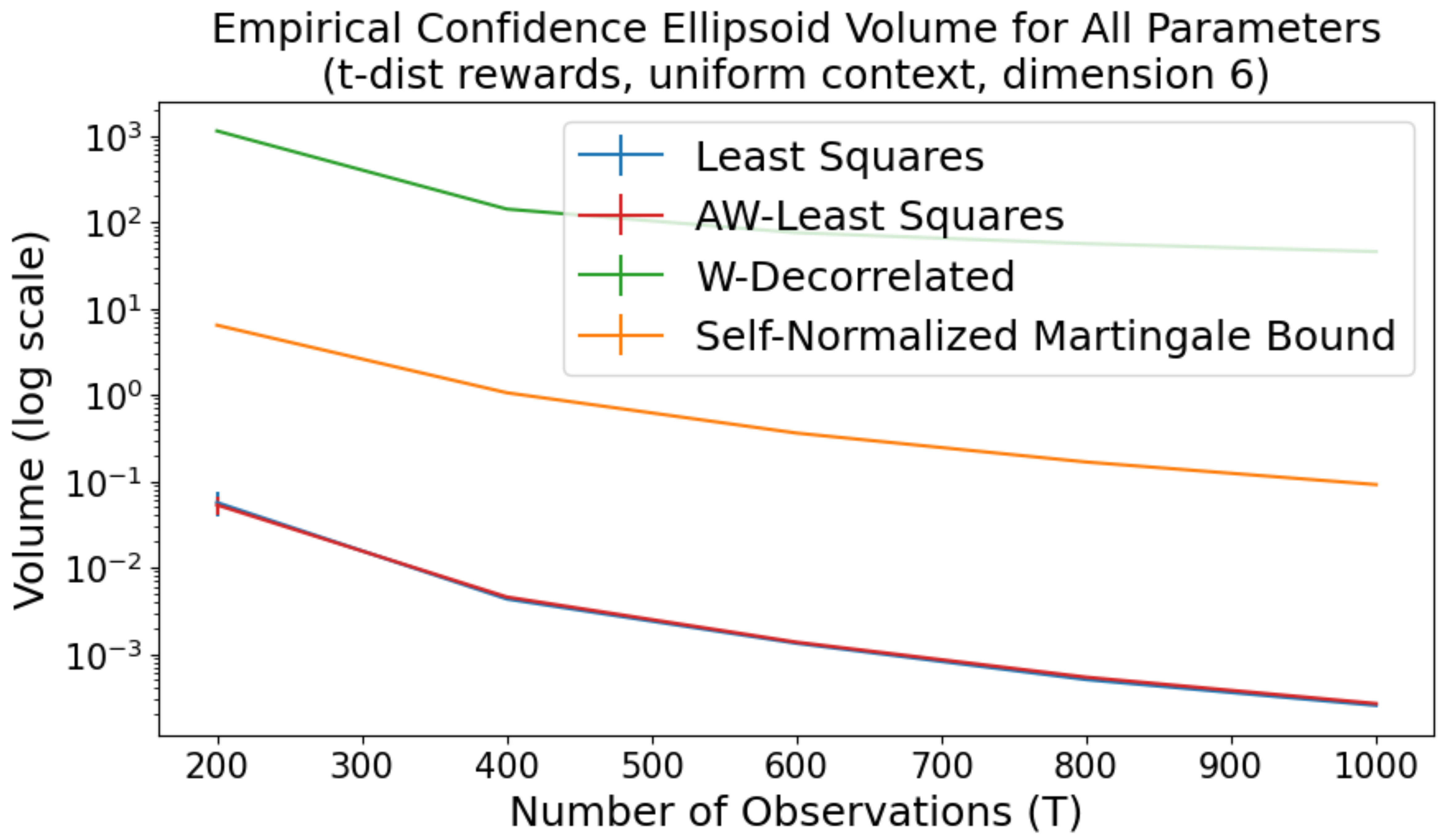
Weighted Least Squares

- $E_{\theta^*}[R_t \mid A_t, X_t] = \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$
- t-distributed rewards

Adaptively weighted M-estimator performs similarly for generalized linear models for Bernoulli and Poisson distributed rewards

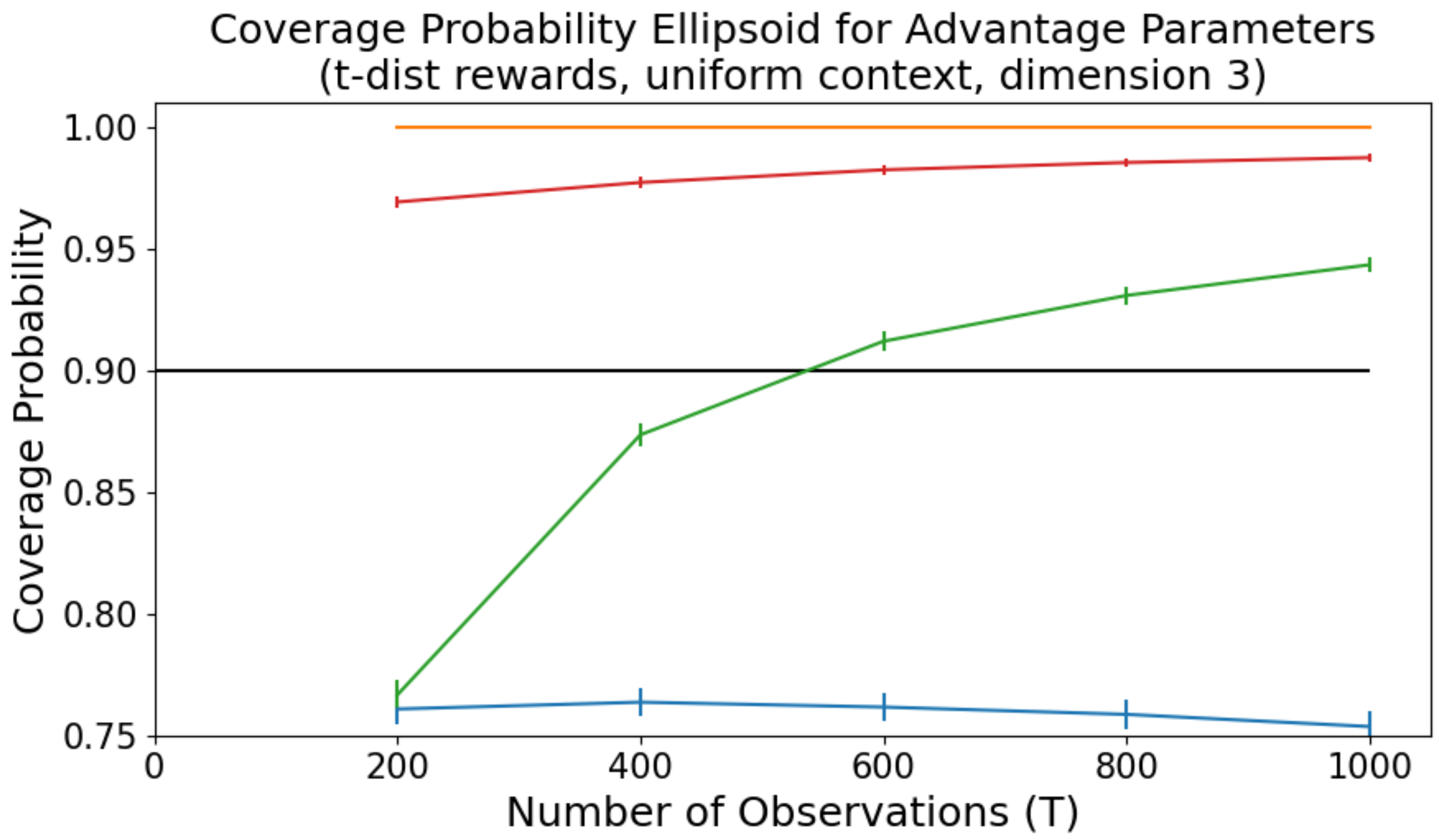
Simulations: Weighted Least Squares

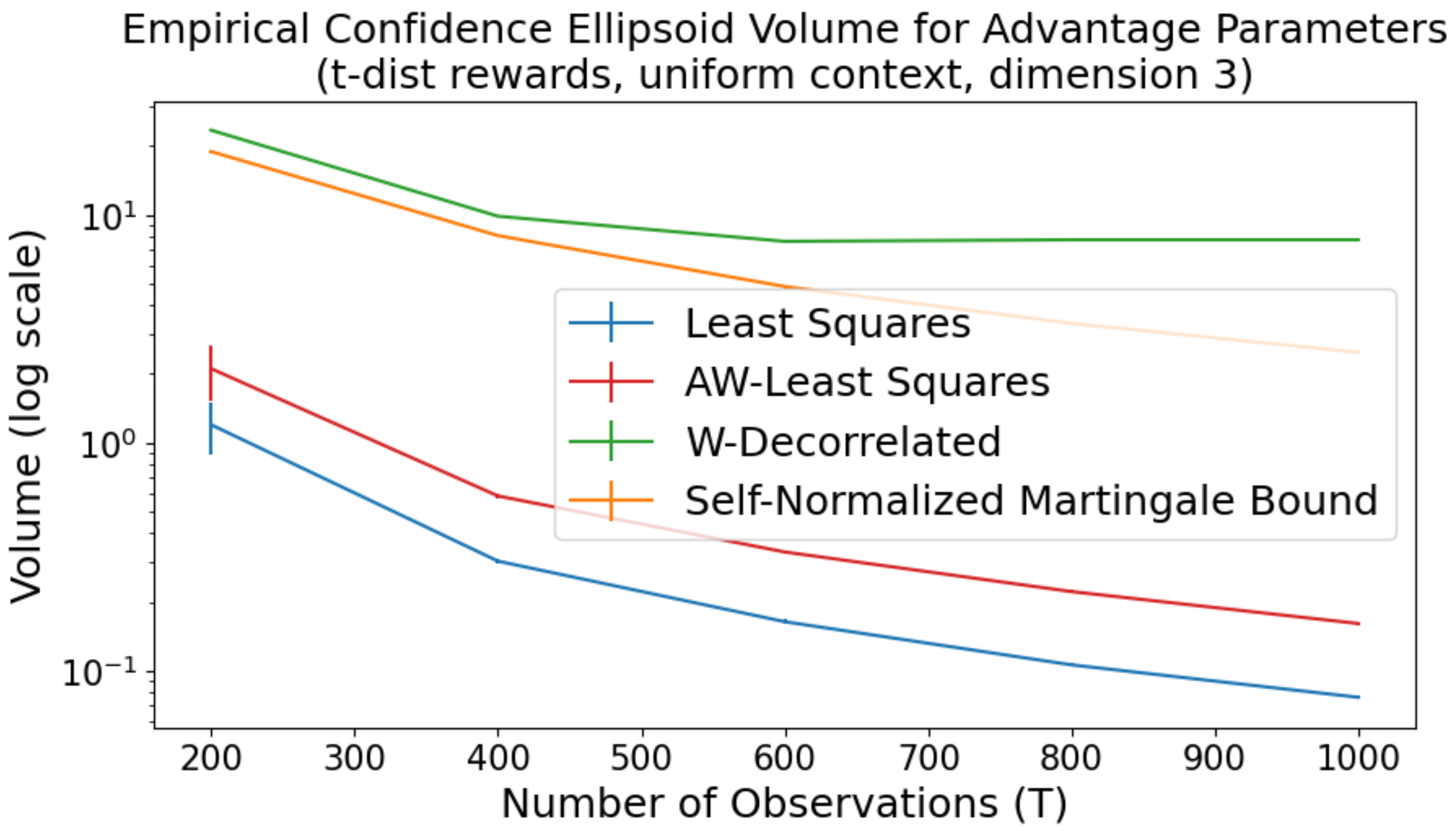




Confidence Regions for θ^* (all parameters)

Simulations: Weighted Least Squares





Confidence Regions for θ_1^* (Advantage)

Open Questions

Immediate Next Questions

- **Model misspecification:** Inference for projected parameters
- **More complex data analytic settings:** What if environment is Markov Decision Process?

Trade-off regret minimization and statistical inference objectives

- Algorithms that trade-off regret and width of confidence intervals
- Sample size calculators

Oralytics: Mobile Health for Oral Health Behavior

- Collaboration with dentists and behavioral scientists that I'm motivated by!!
- Promote users to brush teeth using personalized nudge messages
- Bandit algorithm learns when to send messages based on the user's context