Statistical Inference with M-Estimators on Adaptively Collected Data

Kelly W. Zhang, Lucas Janson, Susan A. Murphy

Objectives in Sequential Decision Making

Bandit algorithms are increasingly used in real world problems due to their regret minimization properties. For example in online advertising, mobile health, and online education.

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

Problem Setup

- **Variables in Contextual Bandit Problem**
  - X_t are **contexts**
  - A_t are **actions**
  - Y_t are **outcomes**
  - $R_t = f(Y_t)$ are **rewards**
- **Potential Outcomes:** $\{X_t,\ Y_t(a) : a \in \mathscr{A}\}_{t=1}^{T}$ i.i.d. over t
- **History:** $H_{t-1} = \{X_s, A_s, Y_s\}_{s=1}^{t-1}$
- Bandit algorithm determines action selection probabilities: $\pi_t(A_t, X_t, H_{t-1}) = P(A_t \mid H_{t-1}, X_t)$

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

Note that while the potential outcomes are i.i.d. the observed data $\{X_t, A_t, Y_t\}_{t=1}^{T}$ are not independent over $t \in [1:T]$.

Adaptively Weighted M-Estimators

- We are interested in constructing confidence regions for the true value of θ, which parameterizes an outcome model, e.g.,
  - **Linear Model:** $E[Y_t \mid X_t, A_t] = X_t^\top \theta_0 + A_t X_t^\top \theta_1$
  - **Logistic Regression Model:** $E[Y_t \mid X_t, A_t] = \left[1 + \exp\left(-X_t^\top \theta_0 - A_t X_t^\top \theta_1\right)\right]$
  - **Generalized Linear Model**
- M-estimators encompass many estimators including least squares and maximum likelihood estimators.

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

- Rather we use an **adaptively weighted M-estimator**

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

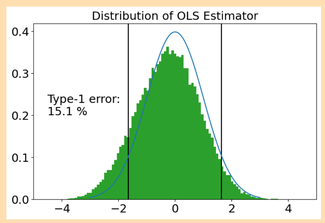
We choose **square-root propensity** weights as follows:

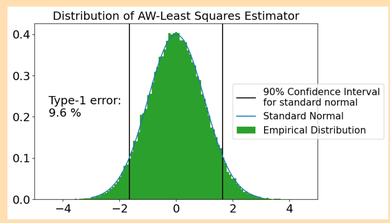
$$W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, H_{t-1})}} = P(A_t \mid X_t, H_{t-1}) \in \sigma\left(H_{t-1}, X_t, A_t\right)$$

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

Least Squares (with and without 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 $T = 1000$.





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

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

Asymptotic Normality Result (abridged)

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

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

Asymptotic Normality:

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

Why square-root propensity weights? Least Squares Example

Suppose we are interested in the following adaptively-weighted least squares estimator:

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

By standard Taylor Series arguments:

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

- **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". Sufficient for the following to equal a constant:

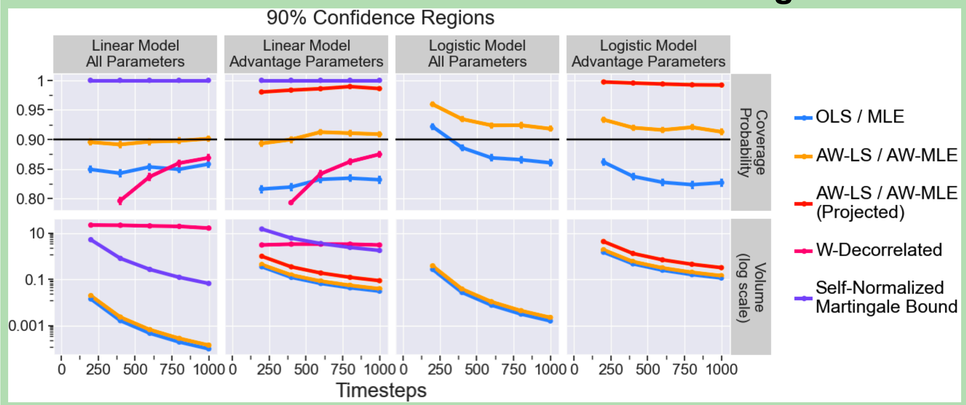
$$E\left[W_t^2 A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right] \qquad W_t = \frac{1}{\sqrt{\pi_t(A_t, X_t, H_{t-1})}}$$

$$= E\left[E\left[\frac{1}{\pi_t(A_t, X_t, H_{t-1})} A_t X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t\right] \middle| H_{t-1}\right] \quad \text{Law of iterated expectations}$$

$$= E\left[E\left[X_t X_t^\top \left(Y_t - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}, X_t, A_t = 1\right] \middle| H_{t-1}\right] \quad \text{Conditioning on } A_t = 1$$

$$= E\left[X_t X_t^\top \left(Y_t(1) - X_t^\top \theta_1^*\right)^2 \middle| H_{t-1}\right] = E\left[X_t X_t^\top \left(Y_t(1) - X_t^\top \theta_1^*\right)^2\right] \quad \text{i.i.d. Potential Outcomes}$$

Simulations in Contextual Bandit Setting



Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence regions. The left two columns are for the continuous reward setting and the right two columns are for the binary reward setting.

- $\tilde{X}_t = [1, X_t]$ and $\theta^* = [\theta_0^*, \theta_1^*] = [0.1, 0.1, 0.1, 0, 0, 0]$ **(θ_1^* are advantage parameters)**
- Thompson Sampling contextual bandit algorithm
- **Weighted Least Squares (continuous reward):** $E_{\theta^*}[R_t \mid A_t, X_t] = \tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*$
- **Weighted Logistic Regression (binary reward):** $E_{\theta^*}[R_t \mid A_t, X_t] = \text{Logistic}(\tilde{X}_t^\top \theta_0^* + A_t \tilde{X}_t^\top \theta_1^*)$

Acknowledgements: This work is supported by NIAAA (award number R01AA23187), NIDA (award number P50DA039838), NCI (award number U01CA229437), and by NIH/NIBIB and OD (number P41EB028242). This work is also supported by the NSF Graduate Research Fellowship Program (Grant No. DGE1745303).