Statistical Inference with M-Estimators on Adaptively Collected Data

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

Objectives in Sequential Decision Making

1. Personalize treatment actions to provide best user experience

- Regret minimization / Choose best actions compared to an oracle policy
- <u>Bandit / RL algorithms are designed to optimize this objective</u>

2. Assess Causal Effects

- Use data collected to gain generalizable knowledge
- Example: construct confidence intervals for a treatment effect

Contextual Bandit Environment

- **Contextual Bandit Variables:**
  - A_t are **actions** (different types of ads)
  - X_t are **contexts** (type of website, recent user behavior)
  - Y_t are **outcomes** (click-through rate, money spent)
  - $R_t = f(Y_t)$ are **rewards**
- **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:** $\mathbb{P}\left(A_t = a \,|\, H_{t-1}, X_t\right)$

Bandit Algorithms Induce Dependence

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

- Use past observations H_{t-1} to inform what action A_t to select next
- Bandit data is "adaptively collected"

Consequences for Statistical Inference

- Violates independence assumptions of standard statistical inference methods —> Bias, Asymptotically non-normal

Binary Action 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	$A_1 = 0$	$A_2 = 1$	$A_3 = 1$	...	$A_T = 0$

Statistical Analysis Objective

We are interested in constructing confidence regions for the true value of θ, which parameterizes an outcome model, e.g.,

- **Linear Model:** $\mathbb{E}[Y_t \,|\, X_t, A_t] = X_t^\top \theta_0 + A_t X_t^\top \theta_1$
- **Logistic Regression Model:**

$$\mathbb{E}[Y_t \,|\, X_t, A_t] = \left[1 + \exp\left(-X_t^\top \theta_0 - A_t X_t^\top \theta_1\right)\right]$$

- **Generalized Linear Models**

Many standard estimators are **M-estimators:** least squares, logistic regression, maximum likelihood

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

Adaptive Square-Root Inverse Propensity Weights

Rather than consider standard M-estimators, we consider we use an **adaptively weighted M-estimator:**

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

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

$$W_t = \frac{1}{\sqrt{\mathbb{P}(A_t \,|\, H_{t-1}, X_t)}}$$

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

Asymptotic Normality Result

Estimand:

$$\theta^* = \mathrm{argmax}_{\theta \in \Theta} \left\{ \mathbb{E}\left[m_\theta(Y_t, X_t, A_t) \,\middle|\, X_t, A_t \right] \right\} \text{ for all } X_t, A_t$$

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

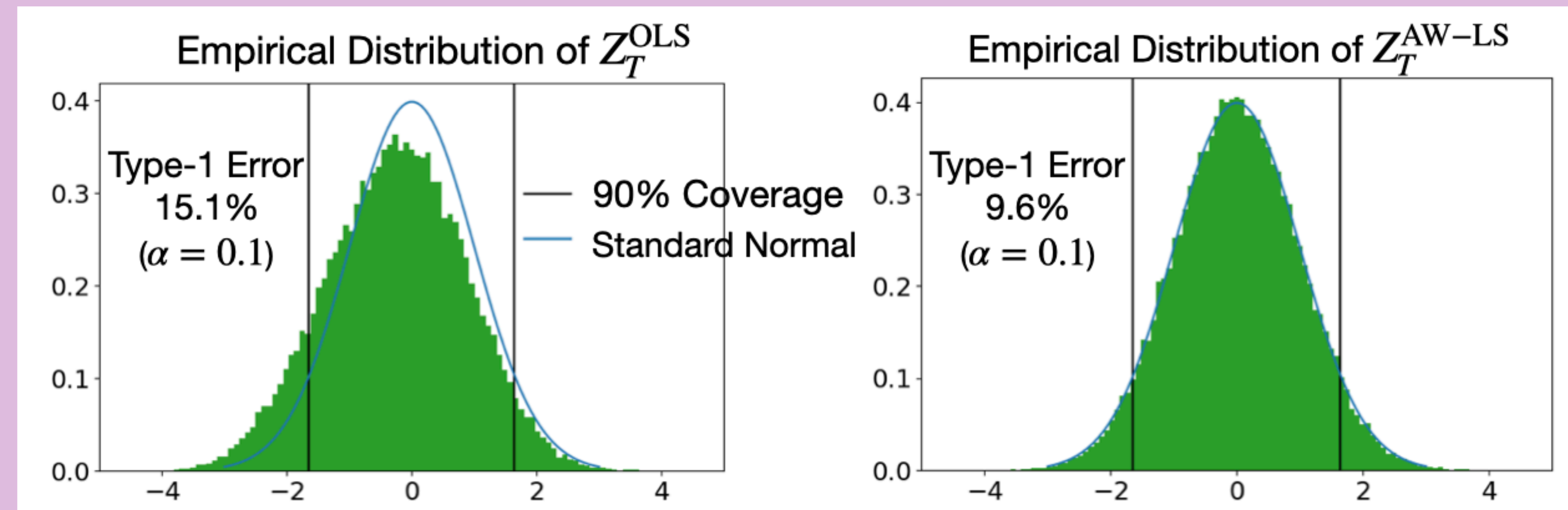
Asymptotic Normality:

$$\left[\frac{1}{T} \sum_{t=1}^{T} W_t \left(\frac{\partial^2}{\partial\theta\partial\theta^\top} m_{\hat{\theta}_T}(Y_t, X_t, A_t) \right) \right] \sqrt{T}(\hat{\theta}_T - \theta^*)$$

$$\xrightarrow{D} N\left(0, \sum_{a \in \mathscr{A}} \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} m_{\theta^*}(Y_t, X_t, a) \right)^{\otimes 2} \right] \right)$$

Least Squares With and Without Adaptive Weights

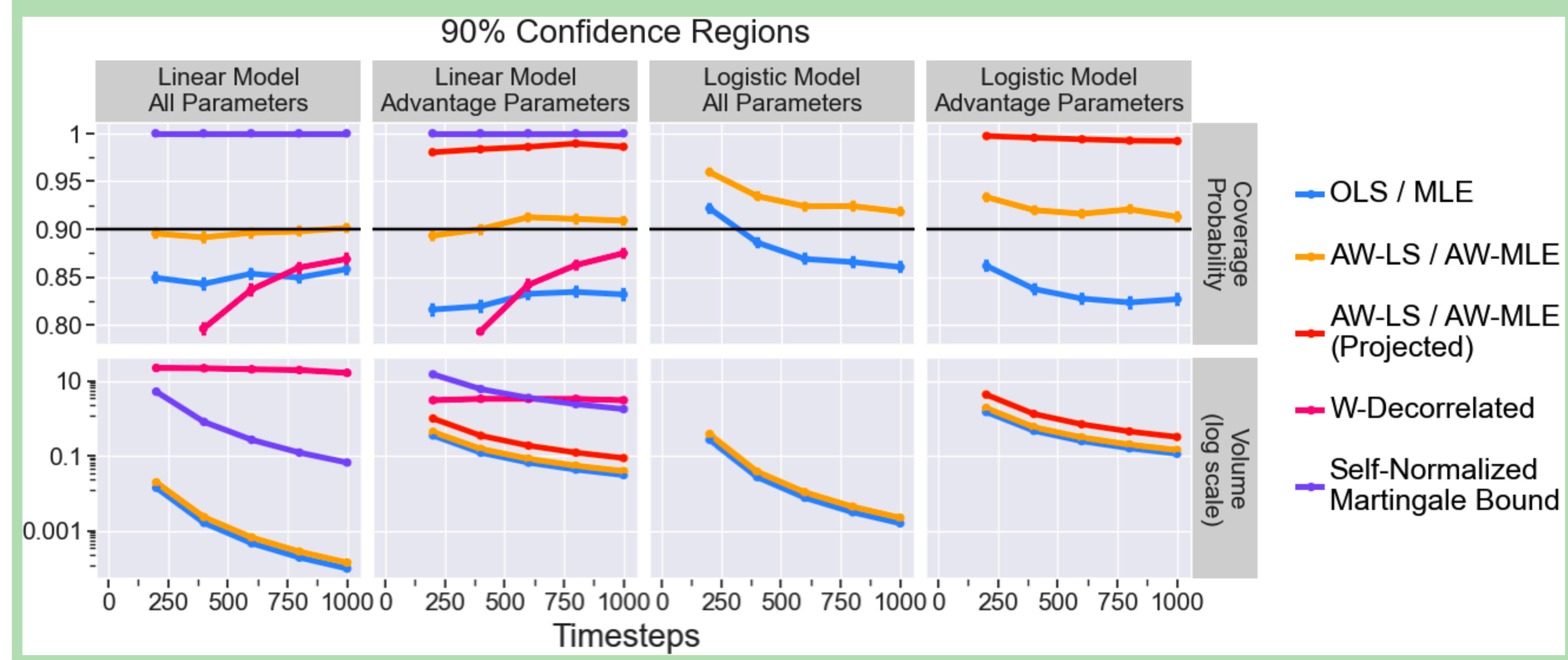
Data generating process: Two-arm bandit with arm means $\theta^* = [\theta_1^*, \theta_2^*]^\top = [0,0]^\top$. Thompson Sampling with $N(0,1)$ priors, $N(0,1)$ noise on rewards, and $T = 1000$.



$$Z_T^{\mathrm{OLS}} := \sqrt{\sum_{t=1}^{T} A_t} \left(\hat{\theta}_{1,T}^{\mathrm{OLS}} - \mathbb{E}[Y_t(1)] \right) \qquad Z_T^{\mathrm{AW-LS}} := \left(\frac{1}{T} \sum_{t=1}^{T} W_t A_t \right) \sqrt{T} \left(\hat{\theta}_{1,T}^{\mathrm{AW-LS}} - \mathbb{E}[Y_t(1)] \right)$$

Simulations in Contextual Bandit Setting

- Context X_t is 3-dimensional (including intercept)
- Binary actions $A_t \in \{0,1\}$
- Reward Types
  - **Continuous:** $R_t = X_t^\top \theta_0^* + A_t X_t^\top \theta_1^* + \epsilon_t$ for ϵ_t t-distributed
  - **Binary:** $R_t \,|\, X_t, A_t \sim \mathrm{Bernoulli}\left(X_t^\top \theta_0^* + A_t X_t^\top \theta_1^* \right)$
- $\theta^* = [\theta_0^*, \theta_1^*]$ where $\theta_1^* = [0,0,0]$ (advantage parameters) and $\theta_0^* = [0.1, 0.1, 0.1]$
- Posterior Sampling contextual bandit algorithm used to select actions A_t
- Estimators
  - **Continuous rewards:** Least Squares (OLS) and Adaptively Weighted-Least Squares
  - **Binary Rewards:** Logistic Regression / MLE and Adaptively Weighted-MLE



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).