Inference after Pooling Practical

Tuesday Afternoon Session

Overview

1. Boltzmann Sampling Algorithm
2. Are Adaptive and Standard Sandwich Variances ever equivalent?
3. Does standard Thompson Sampling and ϵ-greedy algorithms converge to limiting policies?

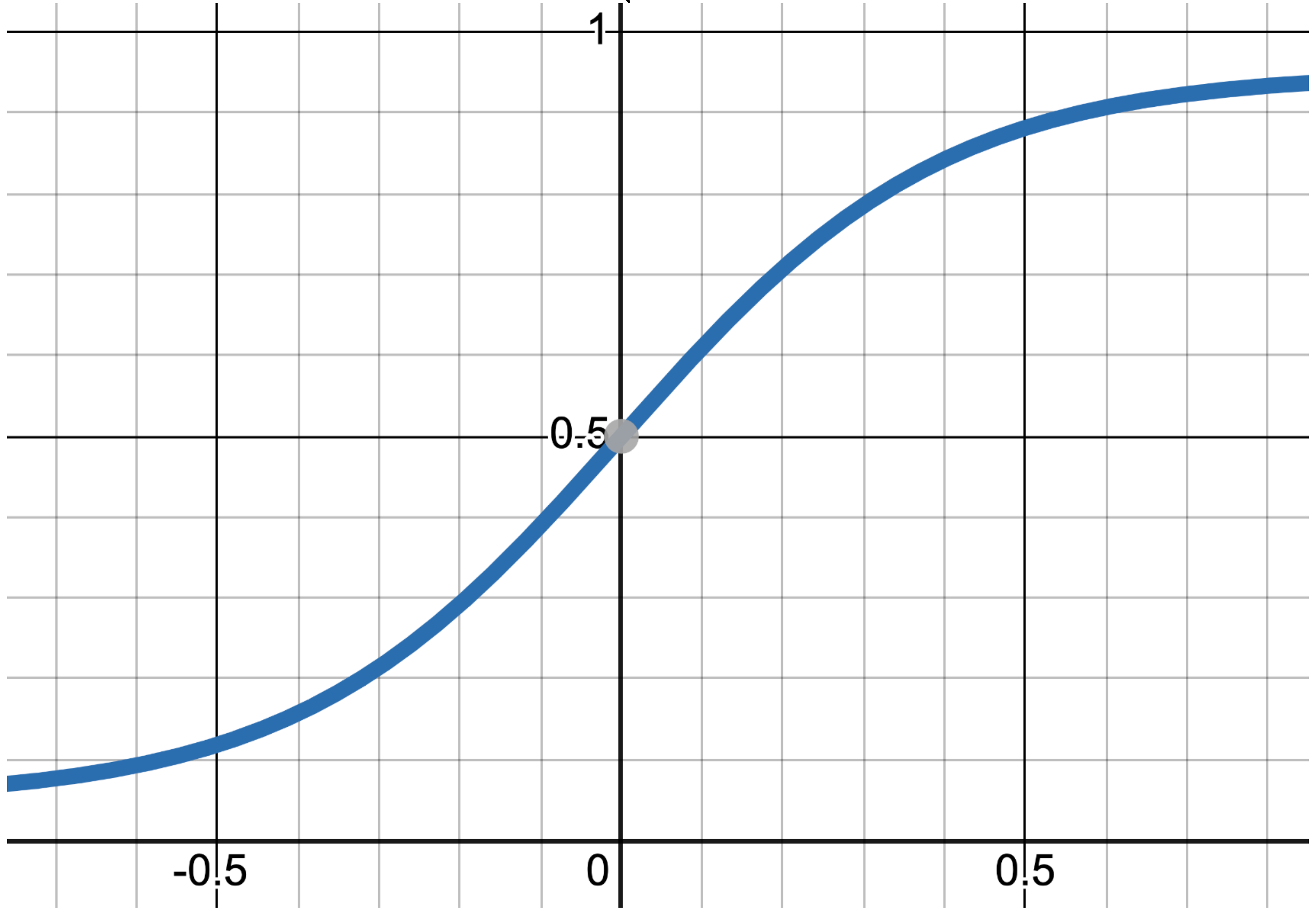
Allocation Function: What probability should the limiting policy send a message?

Balance Maximizing Rewards and Inferring Treatment Effects

- Between trial learning / Continual learning

$$\pi^{\star}(s) = \text{Softmax}\Big(\text{Treatment Effect}(s)\Big)$$

Probability of Sending a Message



Treatment Effect in State s

Boltzmann Sampling: Learning Algorithm

RL Algorithm Reward Model

$$\mathbb{E}\left[R_{i,t+1} \,|\, H_{i,t}, S_{i,t+1}, A_{i,t}\right] = \phi_0(H_{i,t-1}, S_{i,t})^\top \beta_0 + A_{i,t}\phi_1(S_{i,t})\beta_1$$

Fit Reward Model

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta} g\left(D_{i,1:t};\beta\right)\Big|_{\hat{\beta}_t}$$

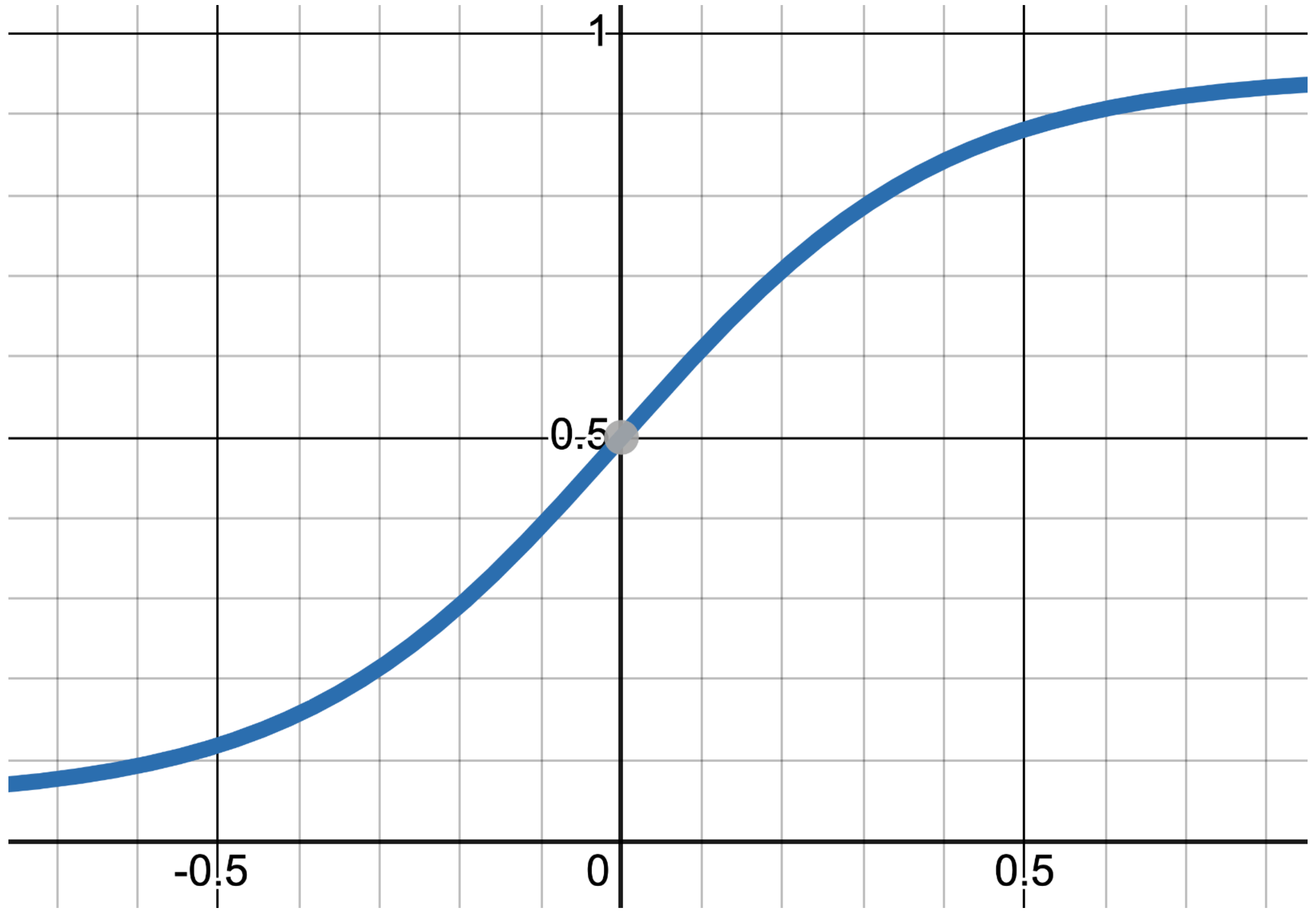
$$g\left(D_{i,1:t};\beta\right) = \sum_{t'=1}^{t} \left\{R_{i,t+1} - \phi_0(H_{i,t-1}, S_{i,t})^\top \beta_0 - A_{i,t}\phi_1(S_{i,t})\beta_1\right\}^2$$

Boltzmann Sampling: Optimization Algorithm

Use "Softmax" to form action selection probability

$$\mathbb{P}\left(A_{i,t+1} \,|\, H_{1:n,t}, S_{i,t+1}\right) = \frac{1}{1 + \exp\left(-\phi_1(S_{i,t+1})^\top \hat{\beta}_t\right)}$$

Probability of Sending a Message



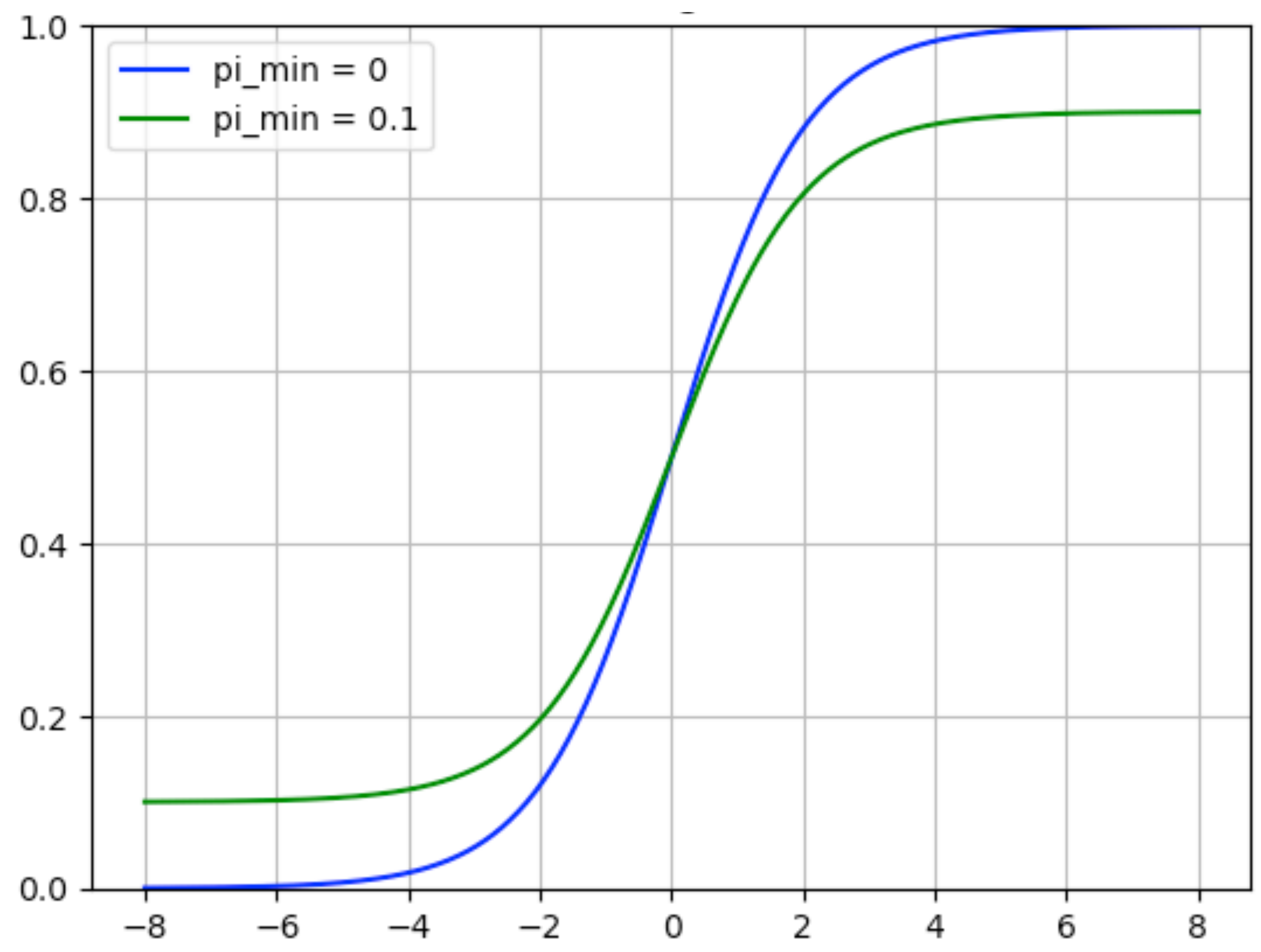
Treatment Effect in State s

Boltzmann Sampling: Optimization Algorithm

Use "Softmax" to form action selection probability

$$\mathbb{P}\left(A_{i,t+1} \mid H_{1:n,t}, S_{i,t+1}\right) = \pi_{\min} + \frac{1 - 2\pi_{\min}}{1 + \exp\left(-\phi_1(S_{i,t+1})^\top \hat{\beta}_t\right)}$$

Probability of Sending a Message



Treatment Effect in State s

Overview

1. Boltzmann Sampling Algorithm
2. Are Adaptive and Standard Sandwich Variances ever equivalent?
3. Does standard Thompson Sampling and ϵ-greedy algorithms converge to limiting policies?

Will the adaptive sandwich and standard sandwich variances ever be equivalent?

$$\sqrt{n}\left(\hat{\theta}-\theta^\star\right) \xrightarrow{D} \mathcal{N}\left(0,\, \ddot{L}^{-1}\Sigma^{\text{adapt}}(\ddot{L}^{-1})^\top\right)$$

$$\Sigma^{\text{adapt}} = \mathbb{E}_{\pi^\star}\left[\left\{\dot{\ell}_i(\theta^\star) - V_1\ddot{G}_1^{-1}\dot{g}_{i,1}(\beta_1^\star)\right\}\left\{\dot{\ell}_i(\theta^\star) - V_1\ddot{G}_1^{-1}\dot{g}_{i,1}(\beta_1^\star))\right\}^\top\right]$$

Yes! In particular, in special cases V_1 may be zero!

$$V_1 = \frac{\partial}{\partial\beta_1}\mathbb{E}_{\pi_2(\beta_1)}\left[\dot{\ell}_i(\theta^\star)\right]\Big|_{\beta_1=\beta_1^\star}$$

Intuition: How does the solution $\theta^\star$ changing with small changes in $\pi_2(\beta_1)$?

V_1 under "Correct" Model Specification

$$V_1 = \frac{\partial}{\partial \beta_1} \mathbb{E}_{\pi_2(\beta_1)} \left[\dot{\ell}_i(\theta^\star) \right] \Big|_{\beta_1 = \beta_1^\star}$$

- For example,

$$\ell\left(H_{i,T}; \theta\right) = \sum_{t=1}^{T} \left(Y_{i,t+1} - X_{i,t}^\top \theta_0 - A_{i,t}\theta_1\right)^2$$

$$\dot{\ell}\left(H_{i,T}; \theta\right) = \sum_{t=1}^{T} \left(Y_{i,t+1} - X_{i,t}^\top \theta_0 - A_{i,t}\theta_1\right) \begin{bmatrix} X_{i,t} \\ A_{i,t} \end{bmatrix}$$

Intuition: Under "correct" model specification the solution $\theta^\star$ will not depend on $\pi(\beta_1)$

- $V_1 = 0$ under "correct" model specification

$$\mathbb{E}\left[Y_{i,t+1} \,\middle|\, H_{i,t-1}, X_{i,t}, A_{i,t} \right] = X_{i,t}^\top \theta_0^\star + A_{i,t}\theta_1^\star$$

Overview

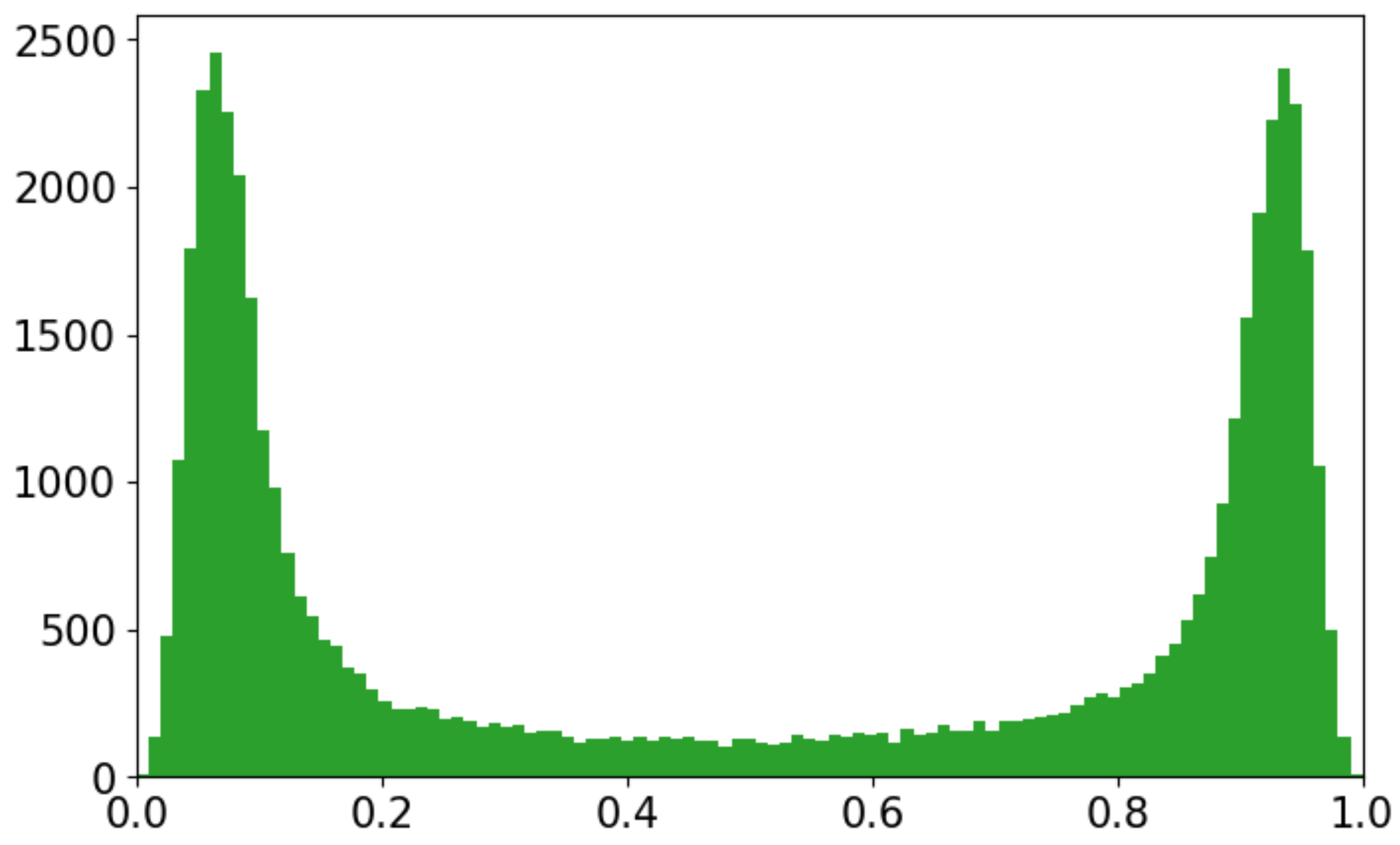
1. Boltzmann Sampling Algorithm
2. Are Adaptive and Standard Sandwich Variances ever equivalent?
3. Does standard Thompson Sampling and ϵ-greedy algorithms converge to limiting policies?

Discussion: Consider the following plots

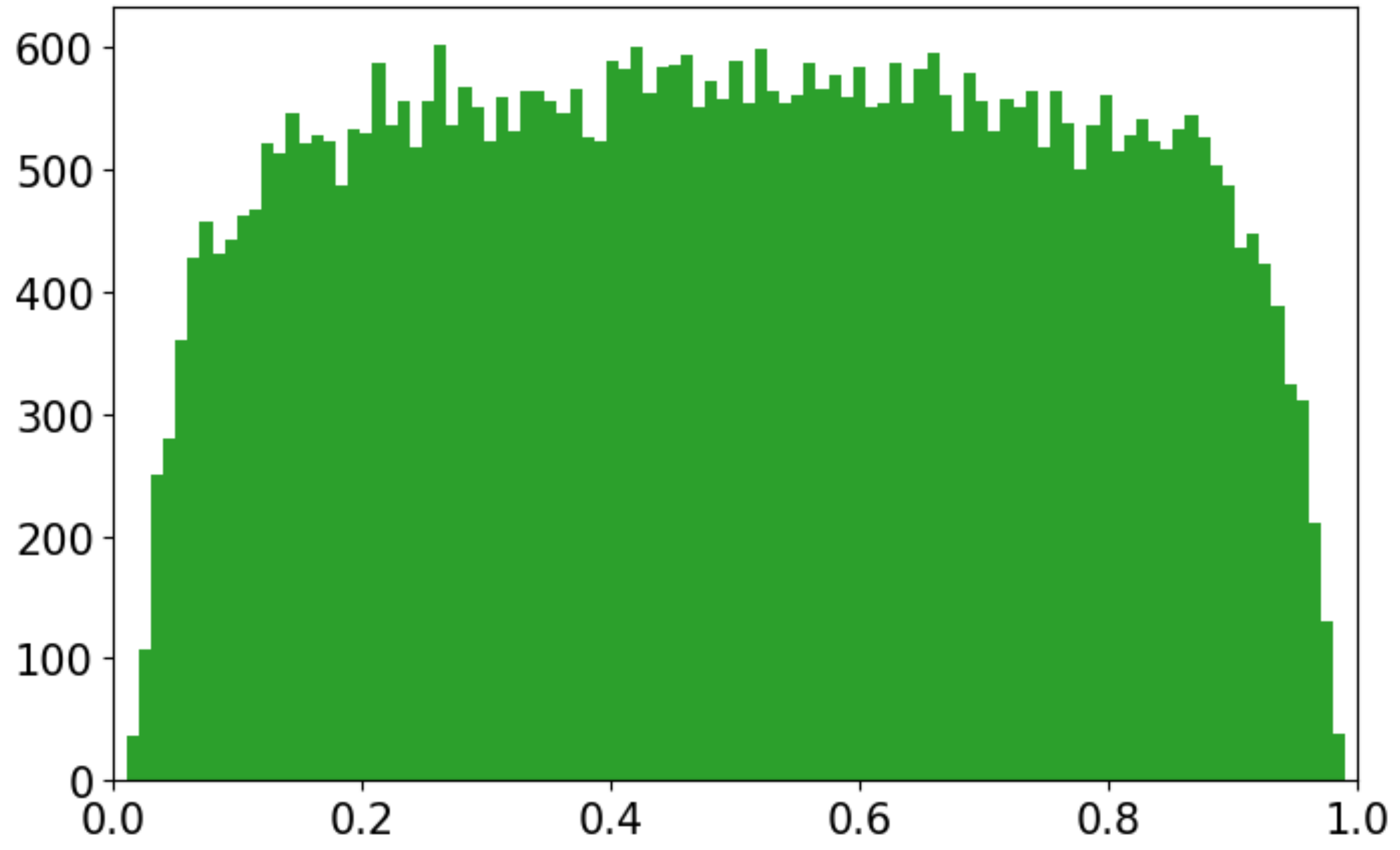
Collected in a multi-arm bandit environment where

$R_t(0), R_t(1) \sim \mathcal{N}(0,1)$ and $T = 1000$

ϵ-Greedy



Thompson Sampling



Histograms of $\frac{1}{T}\sum_{t=1}^{T} A_t$

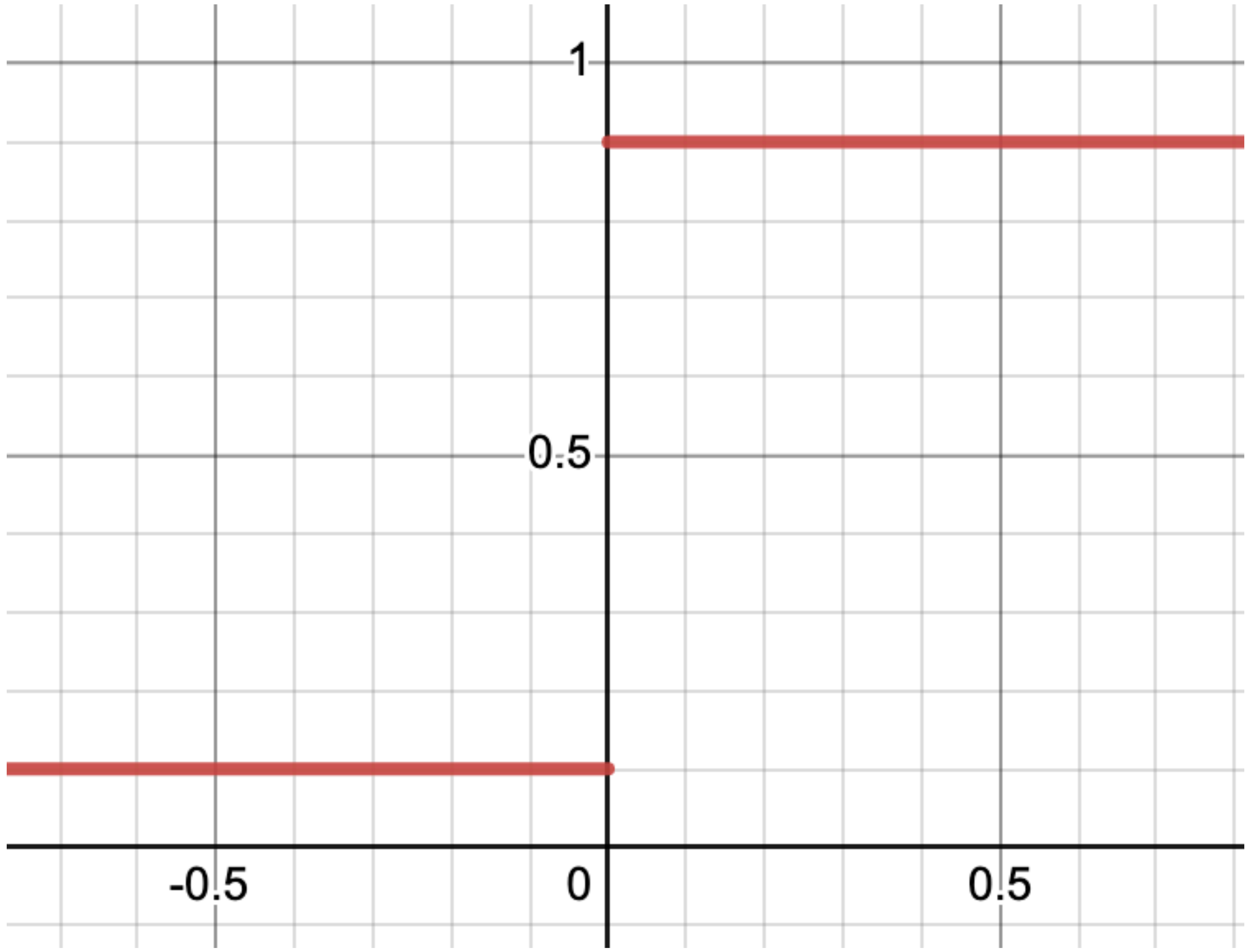
when experiment is repeated for many trials

- How can we explain the above plots?
- Are the above plots surprising? Why or why not?

Instability of the Adaptive Policy

Limiting Action Selection Probabilities

Probability of Selecting $A_t = 1$



Treatment Effect:

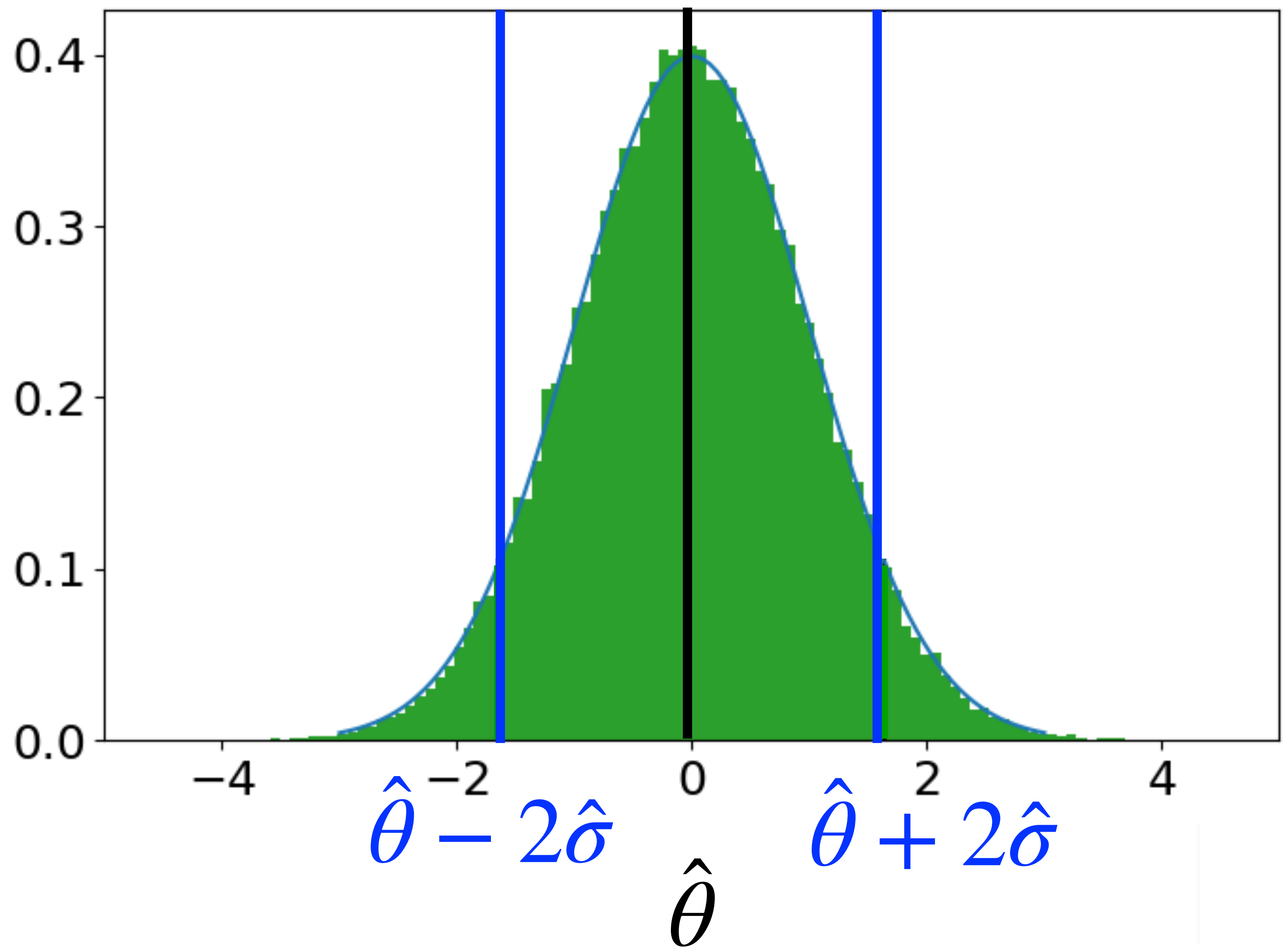
$$\Delta = \mathbb{E}\left[R_t(1)\right] - \mathbb{E}\left[R_t(0)\right]$$

Other examples non-smoothness problems:

- CI for test error of classifier
- Bootstrap
- Hodges estimator

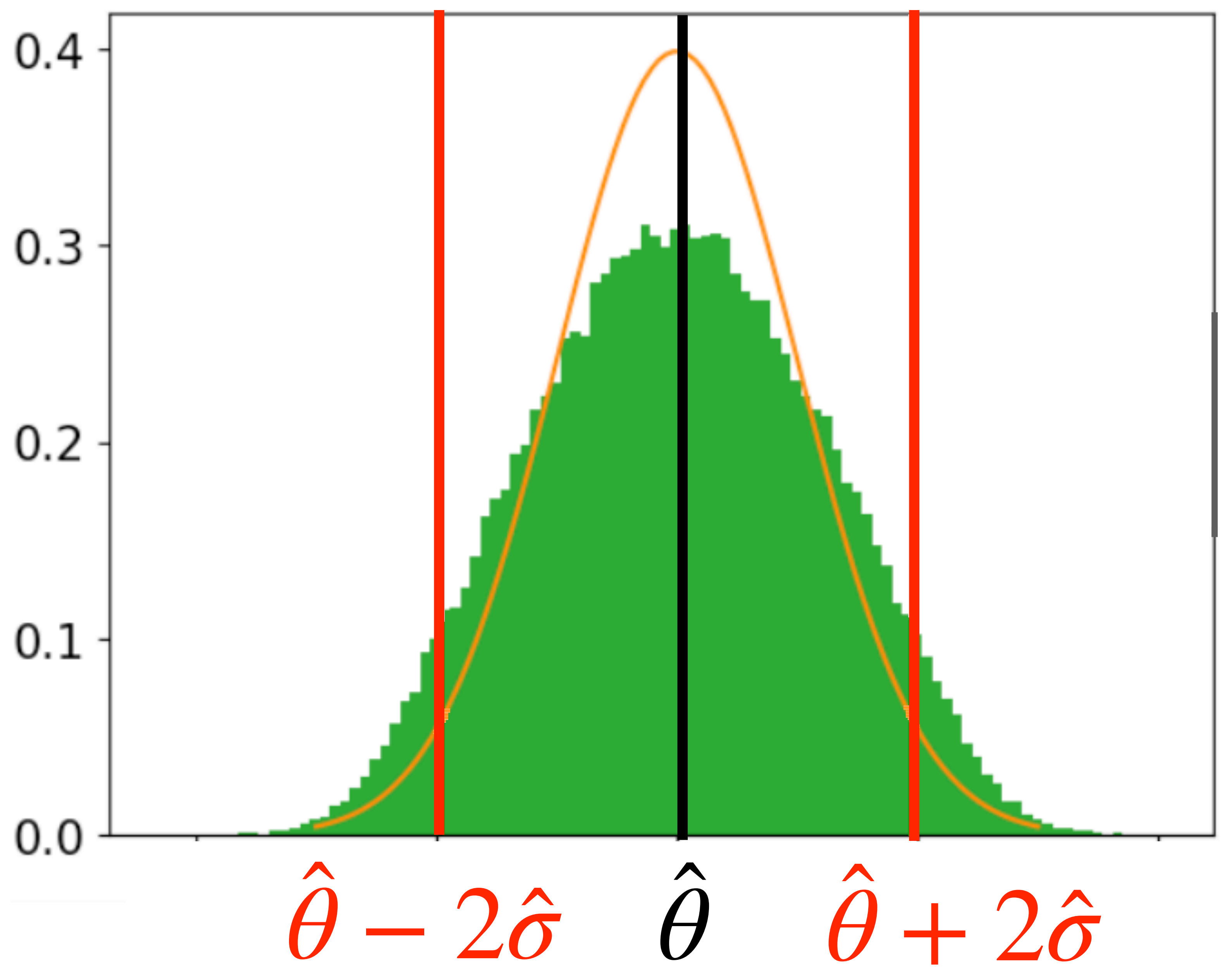
Consequences of Instability of Adaptive Policy

Difference in Sample Means
Independently Collected Data



95% Percent Confidence Interval

Difference in Sample Means
Under Thompson Sampling



Only 89.5% coverage (expect 95%)

For more details see see Statistical Inference with M-Estimators on Adaptively Collected Data by Zhang et al. 2021

Uniformity in Underlying Distribution

Reward Potential Outcomes: $\{R_t(0), R_t(1)\} \sim P$

Weak Convergence: $Z_{n,P} \xrightarrow{D} Z_P$ where $Z_{n,P} = \sqrt{n}\left(\frac{\sum_{t=1}^T A_t R_t}{\sum_{t=1}^T A_t} - \mathbb{E}_P[R_t(1)] \right)$

$$\sup_{f \in \mathrm{BL}_1} \mathbb{E}_P \left[f(Z_{n,P}) - f(Z_P) \right] \to 0$$

Weak Convergence Uniformly in Underlying Distribution:

$$\sup_{P \in \mathscr{P}} \sup_{f \in \mathrm{BL}_1} \mathbb{E}_P \left[f(Z_{n,P}) - f(Z_P) \right] \to 0$$

For more on uniformity see “On the uniform asymptotic validity of subsampling and the bootstrap” by Romano et al. 2012

Discussion Questions

(1) For standard Thompson sampling, what kinds of issues arise if the treatment effect is very small or zero?

(2) When the treatment effect is very small or zero, should the average randomization probability be around 0.5 when the decision is whether or not to treat?