Inference for Batched Bandits

Kelly W. Zhang [1] **Lucas Janson** [2] **Susan A. Murphy** [1 2]

Abstract

As bandit algorithms are increasingly utilized in scientific studies, there is an associated increasing need for reliable inference methods based on the resulting adaptively-collected data. In this work, we develop methods for inference regarding the treatment effect on data collected in batches using a bandit algorithm. We focus on the setting in which the total number of batches is fixed and develop approximate inference methods based on the asymptotic distribution as the size of the batches goes to infinity. We first prove that the ordinary least squares estimator (OLS), which is asymptotically normal on independently sampled data, is *not* asymptotically normal on data collected using standard bandit algorithms when the treatment effect is zero. This asymptotic non-normality result implies that the naive assumption that the OLS estimator is approximately normal can lead to Type-1 error inflation and confidence intervals with below-nominal coverage probabilities. Second, we introduce the Batched OLS estimator (BOLS) that we prove is asymptotically normal—even in the zero treatment effect case—on data collected from both multi-arm and contextual bandits. Moreover, BOLS is robust to changes in the baseline reward and can be used for obtaining simultaneous confidence intervals for the treatment effect from all batches in non-stationary bandits. We demonstrate in simulations that BOLS can be used reliably for hypothesis testing and obtaining a confidence interval for the treatment effect, even in small sample settings.

1. Introduction

Most statistical inference methods for data from randomized experiments assume that all treatments are assigned independently (Imbens & Rubin, 2015). In many settings though, in order to minimize regret, we would like to adjust the treatment assignment probabilities to assign treatments that appear to be performing better with higher probability. As a result, bandits have been increasingly utilized in scientific studies—for example, in mobile health (Yom-Tov et al., 2017) and online education (Rafferty et al., 2019). Such adaptively collected data, in which prior treatment outcomes are used to inform future treatment decisions, makes inference challenging due to the induced dependence. For example, estimators like the sample mean are often biased on adaptively-collected data (Nie et al., 2018; Shin et al., 2019).

Additionally, many settings in which bandit algorithms are used have significant non-stationarity. For example, in online advertising, the effectiveness of an ad may change over time due to exposure to competing ads and general societal changes that could affect perceptions of an ad. Thus, approximate inference methods based on asymptotics that rely on the number of (approximately stationary) time periods of an experiment going to infinity are often less applicable. For this reason, in this work we focus on the setting in which data is collected with bandit algorithms in *batches* and develop inference techniques that allow for non-stationarity between batches. For our asymptotic analysis we fix the total number of batches, T, and allow the samples in each batch, n, to go to infinity. Many real world settings naturally collect data in batches since data is often received from multiple users simultaneously and it can be costly to update the algorithm's parameters too frequently (Jun et al., 2016).

The first contribution of this work is showing that on adaptively collected data, rather surprisingly, whether standard estimators are asymptotically normal can depend on whether the treatment effect is zero. We prove that when data is collected using common bandit algorithms, the sampling probabilities can only concentrate if there is a unique optimal arm. Thus, for two-arm bandits, the sampling probabilities do not concentrate when the difference in the expected rewards between the arms (the treatment effect) is zero. We show that this leads the ordinary least squares (OLS) estimator to be asymptotically normal when the treatment effect is non-zero, and asymptotically *not* normal when the treatment effect is zero. Due to the asymptotic non-normality of the OLS estimator under zero treatment effect, we demonstrate that using a normal distribution to ap-

[1] School of Engineering and Applied Sciences, Harvard University [2] Department of Statistics, Harvard University. Correspondence to: Kelly Zhang <kellywzhang@seas.harvard.edu>.

proximate the OLS estimator's finite sample distribution for hypothesis testing can lead to inflated Type-1 error. More crucially, *the discontinuity in the OLS estimator's asymptotic distribution means that standard inference methods (normal approximations, bootstrap*[1]*) may lead to unreliable confidence intervals.* In particular, even when the treatment effect is non-zero, when the ratio of the magnitude of treatment effect to the standard deviation of the noise is small, *assuming the OLS estimator has an approximately normal distribution can lead to confidence intervals that have below-nominal coverage probability* (see Figure 2).

The second contribution of this work is introducing the Batched OLS (BOLS) estimator, which can be used for reliable inference—even in non-stationary and small-sample settings—on data collected with batched bandits. Regardless of whether the true treatment effect is zero or not, the BOLS estimator for the treatment effect for both multi-arm and contextual bandits is asymptotically normal. BOLS can be used for both hypothesis testing and obtaining confidence intervals for the treatment effect. Moreover, BOLS is also automatically robust to non-stationarity in the rewards and can be used for constructing valid confidence intervals even if there is non-stationarity in the baseline reward, i.e., if the rewards of the arms change from batch to batch, but the treatment effect remains constant. If the treatment effect itself is also non-stationary, BOLS can also be used for constructing simultaneous confidence intervals for the treatment effects for each batch. Additionally, we find in simulations that BOLS has very reliable Type-1 error control, even in small-sample settings.

2. Related Work

Batched Bandits Most work on batched bandits either focuses on minimizing regret (Perchet et al., 2016; Gao et al., 2019) or identifying the best arm with high probability (Agarwal et al., 2017; Jun et al., 2016). Note that best arm identification is distinct from obtaining a confidence interval for the treatment effect, as the former identifies the best arm with high probability (assuming there is a best arm), while the latter can be used to test whether one arm is better than the other and provides guarantees regarding the magnitude of the difference in expected rewards between two arms. Note that in contrast to other batched bandit literature that fix the total number of batches and allow batch sizes to be adjusted adaptively (Perchet et al., 2016), we assume that the size of the batches are not chosen adaptively.

Adaptive Inference Much of the recent work on inference for adaptively-collected data focuses on characterizing and reducing the bias of the OLS estimator in finite samples (Deshpande et al., 2018; Nie et al., 2018; Shin et al., 2019). Villar et al. (2015) thoroughly examine a variety of adaptive sampling algorithms for multi-arm bandits and empirically find that the OLS estimator has inflated finite-sample Type-1 error rates when the data is collected using these algorithms.

Deshpande et al. (2018) also consider inference for adaptively sampled arms. They develop the W-decorrelated estimator which is an adjusted version of OLS. Their estimator requires choosing a tuning parameter λ, which allows practitioners to trade off bias for variance. They prove a CLT for their estimator when λ is chosen in a particular way that depends on the number of times each arm is sampled. To gain further insight into the W-decorrelated estimator, we examine it in the two-armed bandit setting; see Appendix F. Most notably, we find that the W-decorrelated estimator down-weights samples that occur later in the study and up-weights samples from earlier in the study. Note that the W-decorrelated estimator does not have guarantees in non-stationary settings.

The Adaptively-Weighted Augmented-Inverse-Probability-Weighted Estimator (AW-AIPW) for multi-arm bandits reweights the samples of a regular AIPW estimator with adaptive weights that are non-anticipating (Hadad et al., 2019). The AW-AIPW estimator for the treatment effect can easily be adapted to the batched (triangular array) setting. In the stationary multi-arm bandit case, we make similar assumptions to those that Hadad et al. (2019) use to prove asymptotic normality of the AW-AIPW estimator; however, the AW-AIPW estimator does not have guarantees in non-stationary settings.

Lai & Wei (1982) prove that the OLS estimator is asymptotically normal on adaptively collected data under certain conditions. In Section 4.1, we examine these conditions and determine settings in which the necessary conditions are satisfied. Later in Section 4.2, we characterize natural settings in which the necessary conditions are violated and explain how this leads to the asymptotic non-normality of the OLS estimator on adaptively collected data.

3. Problem Formulation

3.1. Setup and Notation

Though our results generalize to K-arm, contextual bandits (see Section 5.2), we first focus on the two-arm bandit for expositional simplicity. Suppose there are T timesteps or batches in a study. In each timestep $t \in [1\colon T]$, we select n treatment arms $\{A_{t,i}^{(n)}\}_{i=1}^{n} \in \{0,1\}^n$. We then observe independent rewards $\{R_{t,i}^{(n)}\}_{i=1}^{n}$, one for each treatment arm sampled. Note that the distribution of these random

[1] Note that since the validity of bootstrap methods rely on uniform convergence (Romano et al., 2012), the non-uniformity in the asymptotic distribution of standard estimators also means that bootstrap methods can lead to confidence intervals with below-nominal coverage.

variables changes with the batch size, n. For example, the distribution of the actions one chooses for the 2nd batch, $\{A_{2,i}^{(n)}\}_{i=1}^{n}$, may change if one has observed $n = 10$ vs. $n = 100$ samples $(A_{1,i}^{(n)}, R_{1,i}^{(n)})_{i=1}^{n}$ in the first batch. We include an (n) superscript on these random variables as a reminder that their distribution changes with n; however, we often drop the (n) superscript for readability.

We define multi-arm bandit algorithms as functions $\{\mathcal{A}_t\}_{t=1}^{T}$ such that $\mathcal{A}_t(H_{t-1}^{(n)}) =: \pi_t^{(n)} \in [0,1]$, where $H_{t-1}^{(n)} := \{A_{t',i}^{(n)}, R_{t',i}^{(n)} : i \in [1\colon n], t' \in [1\colon t-1]\}$ is the history prior to batch t. The bandit selects treatment arms such that for each $t \in [1\colon T]$, $\{A_{t,i}^{(n)}\}_{i=1}^{n} \overset{i.i.d.}{\sim}$ Bernoulli$(\pi_t^{(n)})$ conditionally on $H_{t-1}^{(n)}$. We assume the following conditional mean for the rewards:

$$\mathbb{E}[R_{t,i}^{(n)} | H_{t-1}^{(n)}, A_{t,i}^{(n)}] = (1 - A_{t,i}^{(n)})\beta_{t,0} + A_{t,i}^{(n)}\beta_{t,1}. \quad (1)$$

Let $\mathbf{X}_{t,i}^{(n)} := [1 - A_{t,i}^{(n)}, A_{t,i}^{(n)}]^\top \in \mathbb{R}^2$ ($\mathbf{X}_{t,i}^{(n)}$ will be higher dimensional when we add more arms and/or contextual variables). Also define $N_{t,1}^{(n)} := \sum_{i=1}^{n} A_{t,i}^{(n)}$ and $N_{t,0}^{(n)} := \sum_{i=1}^{n}(1 - A_{t,i}^{(n)})$, the number of times each arm is sampled in the t^{th} batch. We define the errors as $\epsilon_{t,i}^{(n)} := R_{t,i}^{(n)} - (\mathbf{X}_{t,i}^{(n)})^\top \boldsymbol{\beta}_t$. Equation (1) implies that $\{\epsilon_{t,i}^{(n)} : i \in [1\colon n], t \in [1\colon T]\}$ are a martingale difference array with respect to the filtration $\{\mathcal{G}_t^{(n)}\}_{t=1}^{T}$, where $\mathcal{G}_t^{(n)} := \sigma\big(H_{t-1}^{(n)} \cup \{A_{t,i}^{(n)}\}_{i=1}^{n}\big)$. So $\mathbb{E}[\epsilon_{t,i}^{(n)} | \mathcal{G}_{t-1}^{(n)}] = 0$ for all t, i, n.

3.2. Parameters of Interest

The parameters $\boldsymbol{\beta}_t = (\beta_{t,0}, \beta_{t,1})$ can change across batches $t \in [1\colon T]$, which allows for non-stationarity between batches. Assuming that $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t'}$ for all $t, t' \in [1\colon T]$ simplifies to the stationary bandit case. Our goal is to estimate and obtain confidence intervals for $\boldsymbol{\beta}_t$ for $t \in [1\colon T]$ when the data is collected using common bandit algorithms like Thompson Sampling and ϵ-greedy. Specifically we want an estimator that has an asymptotic distribution that approximates its finite-sample distribution well, so we can both perform hypothesis testing for the zero treatment effect and construct a confidence interval for the treatment effect.

3.3. Action Probability Constraint (Clipping)

When running an experiment with adaptive sampling, it is desirable to minimize regret as much as possible. However, in order to perform inference on the resulting data it is also necessary to guarantee that the bandit algorithm explores sufficiently. Greater exploration in the multi-arm bandit case means sampling the treatments with closer to equal probability, rather than sampling one treatment arm almost exclusively. For example, the central limit theorems (CLTs) for both the W-decorrelated (Deshpande et al., 2018) and the AW-AIPW estimators (Hadad et al., 2019) have conditions that implicitly require that the bandit algorithms cannot sample any given treatment arm with probabilities that go to zero or one arbitrarily fast. Greater exploration also increases the power of statistical tests regarding the treatment effect, i.e., it makes it more probable that a true discovery will be made from the collected data. Moreover, if there is non-stationarity in the treatment effect between batches, it is desirable for the bandit algorithm to continue exploring, so it can adjust to these changes and not almost exclusively sample one arm that is no longer the best.

We explicitly guarantee exploration by constraining the probability that any given treatment arm can be sampled, as per Definition 1 below. Note that we allow the sampling probabilities $\pi_t^{(n)}$ to converge to 0 and/or 1 at some rate.

Definition 1. *A clipping constraint with rate $f(n)$ means that $\pi_t^{(n)}$ satisfies the following:*

$$\lim_{n\to\infty} \mathbb{P}\left(\pi_t^{(n)} \in [f(n), 1 - f(n)]\right) = 1 \quad (2)$$

4. Asymptotic Distribution of the Ordinary Least Squares Estimator

Suppose we are in the stationary case, and we would like to estimate $\boldsymbol{\beta}$. Consider the OLS estimator:

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{R}$$

where $\underline{\mathbf{X}} := [\mathbf{X}_{1,1}, .., \mathbf{X}_{1,n}, .., \mathbf{X}_{T,1}, .., \mathbf{X}_{T,n}]^\top \in \mathbb{R}^{nT\times 2}$ and $\mathbf{R} := [R_{1,1}, .., R_{1,n}, .., R_{T,1}, .., R_{T,n}]^\top \in \mathbb{R}^{nT}$. Note that $\underline{\mathbf{X}}^\top \underline{\mathbf{X}} = \sum_{t=1}^{T}\sum_{i=1}^{n} \mathbf{X}_{t,i}\mathbf{X}_{t,i}^\top$.

4.1. Conditions for Asymptotic Normality of the OLS estimator

If $(\mathbf{X}_{t,i}, \epsilon_{t,i})$ are i.i.d. (i.e., no adaptive sampling), $\mathbb{E}[\epsilon_{t,i}] = 0$, $\mathbb{E}[\epsilon_{t,i}^2] = \sigma^2$, and the first two moments of $\mathbf{X}_{t,i}$ exist, a classical result from statistics (Amemiya, 1985) is that the OLS estimator is asymptotically normal, i.e., as $n \to \infty$,

$$(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{1/2}(\hat{\boldsymbol{\beta}}^{\text{OLS}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \underline{\boldsymbol{I}}_K). \quad (3)$$

Lai & Wei (1982) generalize this result by proving that the OLS estimator is still asymptotically normal in the adaptive sampling case when $\underline{\mathbf{X}}^\top \underline{\mathbf{X}}$ satisfies a certain stability condition. To show that a similar result holds for the batched setting, we generalize the asymptotic normality result of Lai & Wei (1982) to triangular arrays, as stated in Theorem 5. Note, we must consider triangular array asymptotics since the distribution of our random variables vary as the batch size, n, changes.

Condition 1 (Moments)**.** *For all* t, n, i, $\mathbb{E}[(\epsilon_{t,i}^{(n)})^2 | \mathcal{G}_{t-1}^{(n)}] = \sigma^2$ *and* $\mathbb{E}\big[(\epsilon_{t,i}^{(n)})^4 | \mathcal{G}_{t-1}^{(n)}\big] < M < \infty$.

Condition 2 (Stability)**.** *For some non-random sequence of scalars* $\{a_i\}_{i=1}^{\infty}$, *as* $n \to \infty$,

$$a_n \cdot \frac{1}{nT} \sum_{t=1}^{T} N_{t,1}^{(n)} \xrightarrow{P} 1.$$

Theorem 1 (Triangular array version of Lai & Wei (1982), Theorem 3)**.** *Assuming Conditions 1 and 2, as* $n \to \infty$,

$$\left(\underline{\boldsymbol{X}}^{(n),\top}\underline{\boldsymbol{X}}^{(n)}\right)^{1/2}(\hat{\boldsymbol{\beta}}^{\text{OLS}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_p)$$

A strong assumption of Theorem 5 is Condition 2 above. Intuitively, if Condition 2 holds, then the rate at which each arm is sampled eventually will not depend on the values of the previous rewards, so the algorithm will be essentially sampling non-adaptively and the samples can be treated as if they are i.i.d. We now state simple sufficient conditions for Condition 2, and thus Theorem 5 in the bandit setting.

Condition 3 (Conditionally i.i.d. actions)**.** *For each* $t \in [1\colon T]$, $A_{t,i}^{(n)} \overset{i.i.d.}{\sim} \text{Bernoulli}(\pi_t^{(n)})$ *i.i.d. over* $i \in [1\colon n]$ *conditionally on* $H_{t-1}^{(n)}$.

Corollary 1 (Sufficient conditions for Theorem 5)**.** *If Conditions 1 and 3 hold, and* ***the treatment effect is non-zero****, data collected in batches using* ϵ*-greedy or Thompson Sampling with clipping constraint* $f(n) = c$ *for some* $0 < c \leq \frac{1}{2}$ *(see Definition 1) satisfy Theorem 5 conditions.*

In the next section, we show that when the treatment effect is zero, the OLS estimator is asymptotically non-normal; we also discuss how this is due to a violation of stability Condition 2. Note that when the treatment effect is zero, the OLS estimator of the treatment effect is an *unbiased* estimator. This is because in the zero treatment effect setting for a two-armed bandit, the two arms have the same expected reward and are exchangeable. Thus, the expected value of the OLS estimates for the expected reward for each arm are equal. So, when the treatment effect is zero, the OLS estimator of the treatment effect—which is exactly the difference in the OLS estimates for each arm—has expectation zero and is unbiased. From Nie et al. (2018) and Shin et al. (2019) we know that estimators based on adaptively collected data are often biased. As will be seen below, the inferential difficulties arising from the use of adaptively sampled data is not limited to just bias. Another challenge arises because whether standard estimators for the treatment effect converge uniformly (meaning the normalized errors of the estimator has the same asymptotic distribution no matter the size of the true treatment effect) can depend on whether the data is sampled independently or adaptively.

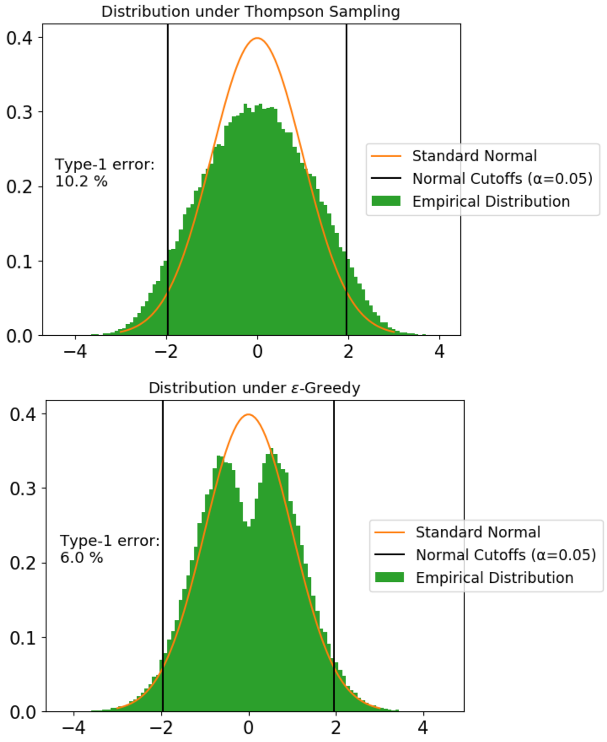

Figure 1. Empirical distribution of the normalized errors of the OLS estimator for the treatment effect with known noise variance, σ^2. All simulations are with no treatment effect ($\beta_1 = \beta_0 = 0$), $\mathcal{N}(0, 1)$ rewards, $T = 25$, and $n = 100$. For ϵ-greedy, $\epsilon = 0.1$.

4.2. Asymptotic Non-Normality under No Treatment Effect (Null)

We prove that when the treatment effect is zero, the OLS estimator is asymptotically non-normal under Thompson Sampling (Theorem 2) and ϵ-greedy (Appendix C). As seen in Figure 1, the normalized errors of the OLS estimator of the treatment effect can have fat tails, which can lead to poor control of Type-1 error in hypothesis testing.

It is sufficient to prove asymptotic non-normality when $T = 2$ and the treatment effect is the same for each batch, under Thompson Sampling with fixed clipping. Here the OLS estimator of $\Delta = \beta_1 - \beta_0$ is simply the difference in the sample means for each arm, i.e., $\hat{\Delta}^{\text{OLS}} = \hat{\beta}_1^{\text{OLS}} - \hat{\beta}_0^{\text{OLS}}$. The normalized errors of $\hat{\Delta}^{\text{OLS}}$, which are asymptotically normal under non-adaptive sampling, are as follows:

$$\sqrt{\frac{(N_{1,1} + N_{1,2})(N_{0,1} + N_{0,2})}{2\sigma^2 n}} (\hat{\Delta}^{\text{OLS}} - \Delta). \qquad (4)$$

Theorem 2 (Asymptotic non-normality of OLS estimator for Thompson Sampling under zero treatment effect)**.** *Let* $T = 2$ *and* $\pi_1^{(n)} = \frac{1}{2}$ *for all* n. *We put independent standard normal priors on the means of each arm,* $\tilde{\beta}_0, \tilde{\beta}_1 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$. *The algorithm assumes noise variance*

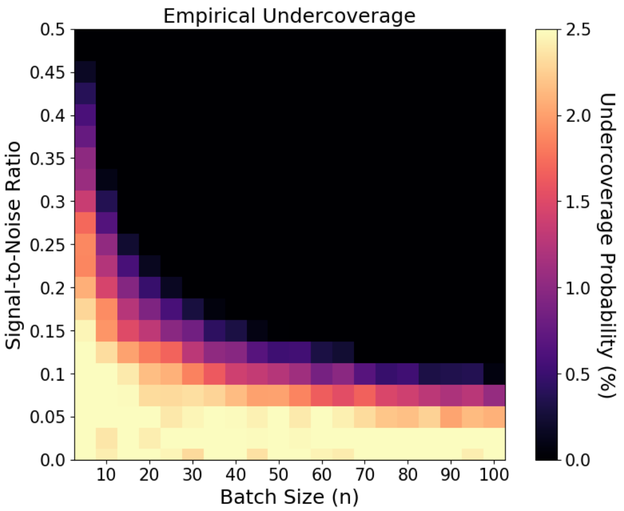

Figure 2. We display undercoverage probabilities (coverage probability below the nominal level, 95%) of confidence intervals based on the OLS estimator and constructed using the quantiles of a standard Normal. We use Thompson Sampling with $\mathcal{N}(0,1)$ priors, a fixed clipping constraint of $0.05 \leq \pi_t^{(n)} \leq 0.95$, $\mathcal{N}(0,1)$ rewards, set $T = 25$, and assume known noise variance σ^2. All standard errors for the estimates used in the plot above are < 0.001.

$\sigma^2 = 1$. *If* $\epsilon_{t,i}^{(n)} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *and* $\pi_2^{(n)} = \pi_{\min} \vee \big[(1 - \pi_{\max}) \wedge \mathbb{P}\big(\tilde{\beta}_1 > \tilde{\beta}_0 \big| H_1^{(n)}\big)\big]$ *for constants* $\pi_{\min}, \pi_{\max}$ *with* $0 < \pi_{\min} \leq \pi_{\max} < 1$*, then the normalized errors of the OLS estimator are asymptotically* ***not*** *normal when the treatment effect is zero, i.e.,*

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \neq 0,$$

where Φ *is the standard Normal distribution CDF and* F_n *is the CDF of the normalized errors of the OLS estimator,* (4).

By Corollary 1, we know the OLS estimator is asymptotically normal when $\Delta \neq 0$. And since Theorem 2 shows that the OLS estimator is asymptotically *not* normal when $\Delta = 0$, we know that the OLS estimator does not converge *uniformly* on adaptively-collected data (see Kasy (2019) for a review of the problems with asymptotic approximations under non-uniform convergence). In real world applications, there is rarely exactly zero treatment effect. However, the discontinuity in the asymptotic distribution of the OLS estimator at zero treatment effect is still practically important. For hypothesis testing and constructing confidence intervals, we use the asymptotic distribution to approximate the finite-sample distribution. The asymptotic distribution of the OLS estimator when the treatment effect is zero is indicative of the finite-sample distribution when the treatment effect is statistically difficult to differentiate from zero, i.e., when the signal-to-noise ratio, $\frac{|\Delta|}{\sigma}$, is low. Figure 2 shows that *even when the treatment effect is non-zero, when the signal-to-noise ratio is low, the confidence interval constructed using cutoffs based on the normal distribution have coverage probabilities below the nominal level. Moreover, for any number of samples and noise variance* σ^2*, there exists a non-zero treatment effect size such that the finite-sample distribution will be poorly approximated by a normal distribution.*

The asymptotic non-normality of OLS occurs specifically when the treatment effect is zero because when there is no unique optimal arm, the sampling probabilities, $\pi_t^{(n)}$, do not concentrate, i.e., $\pi_t^{(n)}$ does not converge to a constant and can fluctuate indefinitely. Under Thompson Sampling, the posterior probability that one arm is better than another is approximately the p-value for the test of the null $H_0 : \Delta = 0$ using the z-statistic for the OLS estimator of the treatment effect. Thus, under the null of zero treatment effect, the posterior probability that arm 1 is better than arm 0 converges to a uniform distribution, as seen in Proposition 1; see Appendix C for details. Next, we introduce an inference method that is asymptotically normal even when the sampling probabilities do not concentrate.

5. Batched OLS Estimator

5.1. Batched OLS Estimator for Multi-Arm Bandits

We now introduce the Batched OLS (BOLS) estimator that is asymptotically normal, even when the treatment effect is zero. Instead of computing the OLS estimator on the data from all batches together, we compute the OLS estimator from each batch and normalize it by the variance estimated from that batch. For each $t \in [1 \colon T]$, the BOLS estimator of the treatment effect $\Delta_t := \beta_{t,1} - \beta_{t,0}$ is:

$$\hat{\Delta}_t^{\text{BOLS}} = \frac{\sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{N_{0,t}} - \frac{\sum_{i=1}^n A_{t,i} R_{t,i}}{N_{1,t}}.$$

Theorem 3 (Asymptotic normality of Batched OLS estimator for multi-arm bandits)**.** *Assuming Conditions 1 (moments) and 3 (conditionally i.i.d. actions), and a clipping rate* $f(n) = \omega(\frac{1}{n})$ *(see Definition 1),*[2]

$$\begin{bmatrix} \sqrt{\frac{N_{1,0} N_{1,1}}{n}} (\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{N_{2,0} N_{2,1}}{n}} (\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{N_{T,0} N_{T,1}}{n}} (\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_T)$$

By Theorem 3, for the stationary treatment effect case, we can test $H_0 : \Delta = c$ vs. $H_1 : \Delta \neq c$ with the following

[2] It is straightforward to show that these results hold in the case that batches are different sizes (for non-adaptively chosen batch sizes) as the size of the smallest batch goes to infinity.

statistic, which is asymptotically normal under the null:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\sqrt{\frac{N_{t,0}N_{t,1}}{n\sigma^2}}(\hat{\Delta}_t^{\text{BOLS}}-c). \tag{5}$$

The key to proving asymptotic normality for BOLS is that the following ratio converges in probability to one:

$$\frac{N_{1,t}^{(n)}N_{0,t}^{(n)}}{n}\frac{1}{n\pi_t^{(n)}(1-\pi_t^{(n)})}\xrightarrow{P}1.$$

Since $\pi_t^{(n)}\in\mathcal{G}_{t-1}^{(n)}$, $\frac{1}{n\pi_t^{(n)}(1-\pi_t^{(n)})}$ is a constant given $\mathcal{G}_{t-1}^{(n)}$. Thus, even if $\pi_t^{(n)}$ does not concentrate, we are still able to apply the martingale CLT (Dvoretzky, 1972) to prove asymptotic normality. See Appendix B for more details.

5.2. Batched OLS Estimator for Contextual Bandits

For the contextual, K-arm bandit case, the parameters of interest are $\boldsymbol{\beta}_t=(\boldsymbol{\beta}_{t,0},\boldsymbol{\beta}_{t,1},...,\boldsymbol{\beta}_{t,K-1})^\top\in\mathbb{R}^{Kd}$. For any two arms $x,y\in[0\colon K-1]$, we can estimate the treatment effect between them $\boldsymbol{\Delta}_{t,x-y}:=\boldsymbol{\beta}_{t,x}-\boldsymbol{\beta}_{t,y}\in\mathbb{R}^d$ for all $t\in[1\colon T]$. In each batch, we observe context vectors $\{\mathbf{C}_{t,i}^{(n)}\}_{i=1}^n$ where $\mathbf{C}_{t,i}^{(n)}\in\mathbb{R}^d$. We can deterministically set the first element of $\mathbf{C}_{t,i}^{(n)}$ to 1 to allow for a non-zero intercept. For our new definition of the history, $H_{t-1}^{(n)}:=\{\mathbf{C}_{t,i}^{(n)},A_{t,i}^{(n)},R_{t,i}^{(n)}\colon i\in[1\colon n],t\in[1\colon T]\}$, we define a new filtration $\mathcal{F}_t^{(n)}:=\sigma\big(H_{t-1}^{(n)}\cup\{A_{t,i}^{(n)},\mathbf{C}_{t,i}^{(n)}\}_{i=1}^n\big)$. We define contextual bandit algorithms to be functions $\{\mathcal{A}_t\}_{t=1}^T$ such that $\mathcal{A}_t(H_{t-1}^{(n)},\mathbf{C}_{t,i})=:\boldsymbol{\pi}_{t,i}^{(n)}\in[0,1]^K$. Note, $\boldsymbol{\pi}_{t,i}^{(n)}$ is now indexed by i because it depends on the context $\mathbf{C}_{t,i}^{(n)}$. Our policy $\boldsymbol{\pi}_{t,i}^{(n)}\in\mathbb{R}^K$ is now a vector representing the probability that any action is chosen, i.e., an action $A_{t,i}^{(n)}\in[0\colon K-1]$ and $A_{t,i}^{(n)}\sim\text{Categorial}(\boldsymbol{\pi}_{t,i}^{(n)})-1$. We assume the following conditional mean model of the reward:

$$\mathbb{E}\big[R_{t,i}^{(n)}\big|\mathcal{F}_{t-1}^{(n)}\big]=\sum_{k=0}^{K-1}\mathbb{I}_{(A_{t,i}^{(n)}=k)}\mathbf{C}_{t,i}^\top\boldsymbol{\beta}_{t,k}$$

and let $\epsilon_{t,i}^{(n)}:=R_{t,i}^{(n)}-\sum_{k=0}^{K-1}\mathbb{I}_{(A_{t,i}^{(n)}=k)}\mathbf{C}_{t,i}^\top\boldsymbol{\beta}_{t,k}$.

Condition 4 (Conditionally i.i.d. contexts)**.** *For each t, $\boldsymbol{C}_{t,1},\boldsymbol{C}_{t,2},...,\boldsymbol{C}_{t,n}$ are i.i.d. and its first two moments, $\boldsymbol{\mu}_t,\underline{\boldsymbol{\Sigma}}_t$, are non-random given $H_{t-1}^{(n)}$, i.e., $\boldsymbol{\mu}_t,\underline{\boldsymbol{\Sigma}}_t\in\sigma(H_{t-1}^{(n)})$.*

Condition 5 (Bounded context)**.** *$||\boldsymbol{C}_{t,i}||_{\max}\le u$ for all i,t,n for some constant u. Also, the minimum eigenvalue of $\underline{\boldsymbol{\Sigma}}_t^{(n)}$ is lower bounded, i.e., $\lambda_{\min}(\underline{\boldsymbol{\Sigma}}_t^{(n)})>l>0$.*

Definition 2. *A conditional clipping constraint with rate $f(n)$ means that the sampling probabilities $\boldsymbol{\pi}_{t,i}^{(n)}:=\mathcal{A}_t(H_{t-1}^{(n)},\boldsymbol{C}_{t,i})\in[0,1]^K$ satisfy the following:*

$$\mathbb{P}\bigg(\forall\,\boldsymbol{c}\in\mathbb{R}^d,\mathcal{A}_t(H_{t-1}^{(n)},\boldsymbol{c})\in\big[f(n),1-f(n)\big]^K\bigg)\to1$$

For each $t\in[1\colon T]$, we have the OLS estimator for $\boldsymbol{\Delta}_{t,x-y}$:

$$\hat{\boldsymbol{\Delta}}_t^{\text{OLS}}:=\big[\underline{\mathbf{C}}_{t,x}^{-1}+\underline{\mathbf{C}}_{t,y}^{-1}\big]^{-1}(\hat{\boldsymbol{\beta}}_{t,x}^{\text{OLS}}-\hat{\boldsymbol{\beta}}_{t,y}^{\text{OLS}})$$

where $\underline{\mathbf{C}}_{t,k}:=\sum_{i=1}^n\mathbb{I}_{A_{t,i}^{(n)}=k}\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top\in\mathbb{R}^{d\times d}$, $\hat{\boldsymbol{\beta}}_{t,k}^{\text{OLS}}=\underline{\mathbf{C}}_{t,k}^{-1}\sum_{i=1}^n\mathbb{I}_{A_{t,i}^{(n)}=k}\mathbf{C}_{t,i}R_{t,i}$.

Theorem 4 (Asymptotic normality of Batched OLS estimator for contextual bandits)**.** *Assuming Conditions 1 (moments)*[3]*, 3 (conditionally i.i.d. actions), 4, and 5, and a conditional clipping rate $f(n)=c$ for some $0\le c<\frac{1}{2}$ (see Definition 2),*

$$\begin{bmatrix}\big[\underline{\boldsymbol{C}}_{1,0}^{-1}+\underline{\boldsymbol{C}}_{1,1}^{-1}\big]^{1/2}(\hat{\boldsymbol{\Delta}}_1^{\text{OLS}}-\boldsymbol{\Delta}_1)\\ \big[\underline{\boldsymbol{C}}_{2,0}^{-1}+\underline{\boldsymbol{C}}_{2,1}^{-1}\big]^{1/2}(\hat{\boldsymbol{\Delta}}_2^{\text{OLS}}-\boldsymbol{\Delta}_2)\\ \vdots\\ \big[\underline{\boldsymbol{C}}_{T,0}^{-1}+\underline{\boldsymbol{C}}_{T,1}^{-1}\big]^{1/2}(\hat{\boldsymbol{\Delta}}_T^{\text{OLS}}-\boldsymbol{\Delta}_T)\end{bmatrix}\xrightarrow{D}\mathcal{N}(0,\sigma^2\underline{\boldsymbol{I}}_{Td}).$$

5.3. Batched OLS Statistic for Non-Stationary Bandits

Many real world problems we would like to use bandit algorithms for have non-stationarity over time. For example, in mobile health, bandit algorithms are used to decide when to send interventions to encourage users to engage in healthy behaviors; however, due to peoples' changing routines and because users can become desensitized to the interventions, there is significant non-stationarity in the effectiveness of these interventions over time. We now describe how BOLS can be used in the non-stationary multi-arm bandit setting.

Non-stationary in the baseline reward We may believe that the expected reward for a given treatment arm may vary over time, but that the treatment effect is constant from batch to batch. In this case, we can simply use the BOLS test statistic described earlier in equation (5) to test $H_0\colon\Delta=0$ vs. $H_1\colon\Delta\neq0$. Note that the BOLS test statistic for the treatment effect is robust to non-stationarity in the baseline reward without any adjustment. Moreover, in our simulation settings we estimate the variance σ^2 separately for each batch, which allows for non-stationarity in the variance between batches as well; see Appendix A for variance estimation details and see Section 6 for simulation results.

Non-stationary in the treatment effect Alternatively we may believe that the treatment effect itself varies from batch

[3] Assume an analogous moment condition for the contextual bandit case, where $\mathcal{G}_t^{(n)}$ is replaced by $\mathcal{F}_t^{(n)}$.

to batch. In this case, we are able to construct a confidence region that contains the true treatment effect Δ_t for each batch simultaneously with probability $1 - \alpha$.

Corollary 2 (Confidence band for treatment effect for non-stationary bandits)**.** *Assume the same conditions as Theorem 3. We let $\Phi^{-1}(\alpha)$ be the α^{th} quantile of the standard Normal distribution. For each $t \in [1\colon T]$, we define the interval*

$$\boldsymbol{L}_t = \hat{\Delta}_t^{\text{OLS}} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2T}\right)\sqrt{\frac{\sigma^2 n}{N_{t,0}N_{t,1}}}.$$

$\lim_{n\to\infty} \mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \boldsymbol{L}_t\big) \geq 1 - \alpha.$

We can also test the null hypothesis of no treatment effect against the alternative that at least one batch has non-zero treatment effect, i.e., $H_0\colon \forall t \in [1\colon T], \Delta_t = 0$ vs. $H_1\colon \exists t \in [1\colon T]\ s.t.\ \Delta_t \neq 0$. Note that the global null stated above is of great interest in the mobile health literature (Klasnja et al., 2015; Liao et al., 2016). Specifically we use the following test statistic:

$$\sum_{t=1}^{T} \frac{N_{t,0}N_{t,1}}{\sigma^2 n}(\hat{\Delta}_t^{\text{OLS}} - 0)^2,$$

which by Theorem 3 converges in distribution to a chi-squared distribution with T degrees of freedom under the null $\Delta_t = 0$ for all t.

6. Simulation Experiments

6.1. Procedure

We focus on the two-arm bandit setting and test whether the treatment effect is zero, specifically $H_0\colon \Delta = 0$ vs. $H_1\colon \Delta \neq 0$. We perform experiments for when the variance of the error $\epsilon_{t,i}$ is estimated. We assume homoscedastic errors throughout. See Appendix A.3 for more details about how we estimate the noise variance. In Figures 6 and 7, we display results for stationary bandits and in Figure 5 we show results for bandits with non-stationary baseline rewards. See Appendix A.4 for results for bandits with non-stationary treatment effects.

For the AW-AIPW estimator, we use the *variance stabilizing* weights, which provably satisfy the CLT conditions of Hadad et al. (2019) and performed well in their simulation results, in terms of MSE and low standard error; for the model of the expected rewards for each arm we use the respective arm's sample mean. For the W-decorrelated estimator, we choose λ based on the procedure used in Deshpande et al. (2018); see Appendix A.1 for details.

We found that several of the estimators, primarily OLS and AW-AIPW, have inflated finite-sample Type-1 error. Since Type-1 error control is a hard constraint, solutions with inflated Type-1 error are *infeasible* solutions. For the sake of comparison, in the power plots, we adjust the cutoffs of the estimators to ensure proper Type-1 error control under the null. Note that it is unfeasible to make this cutoff adjustment for real experiments (unless one found the worst case setting), as there are many nuisance parameters—like the expected rewards for each arm and the variance of the noise—which can affect these cutoff values. For BOLS, we use cutoffs based on the Student-t distribution rather than the normal distribution, as it is relatively straightforward to determine the number of degrees of freedom needed in the correction; see Appendix A.3 for details. We do not make a similar correction for the other estimators we compare to because it is unclear how to determine the number of degrees of freedom that should be used.

6.2. Results

Figure 6 shows that for small sample sizes ($nT \lesssim 300$), BOLS has more reliable Type-1 error control than AW-AIPW with variance stabilizing weights. After $nT \geq 500$ samples, AW-AIPW has proper Type-1 error, and by Figure 7 it always has slightly greater power than the BOLS estimator in the stationary setting. The W-decorrelated approach consistently has reliable Type-1 error control, but very low power compared to AW-AIPW and BOLS.

In Figure 5, we display simulation results for the non-stationary baseline reward setting. *Whereas other estimators have no Type-1 error guarantees, BOLS still has proper Type-1 error control in the non-stationary baseline reward setting. Moreover, BOLS can have much greater power than other estimators when there is non-stationarity in the baseline reward.* Overall, it makes sense to choose BOLS over other estimators (e.g. AW-AIPW) in small-sample settings or whenever the experimenter wants to be robust to non-stationarity in the baseline reward—at the cost of losing a little power if the environment is stationary.

7. Discussion

We found that the OLS estimator is asymptotically non-normal when the treatment effect is zero due to the non-concentration of the sampling probabilities. Since the OLS estimator is a canonical example of a method-of-moments estimator (Hazelton, 2011), our results suggest that the inferential guarantees of standard method-of-moments estimators may fail to hold on adaptively collected data when there is no unique optimal, regret-minimizing policy. We develop the the Batched OLS estimator, which is asymptotically normal even when the sampling probabilities do not concentrate. An open question is whether batched versions of general method-of-moments estimators could similarly be used for adaptive inference.

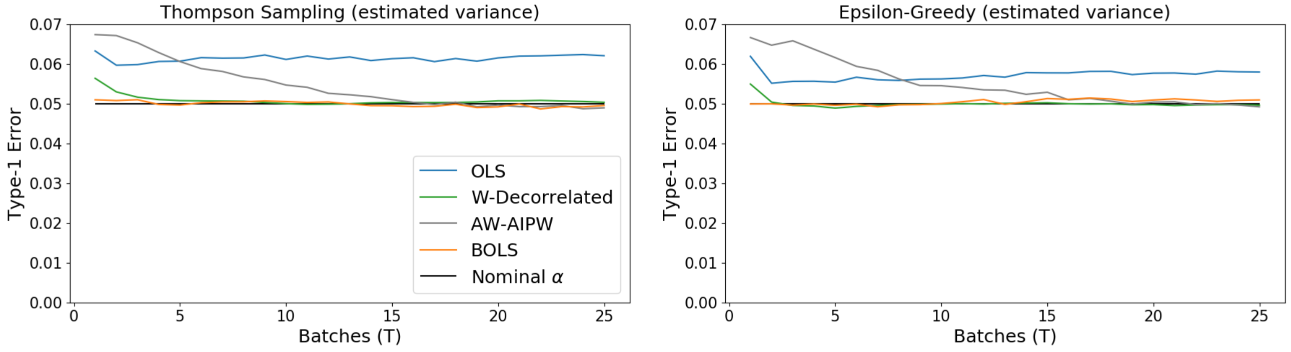

Figure 3. **Stationary Setting:** Type-1 error for estimators of treatment effect for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = \beta_0 = 0$ and use 25 samples per batch. We use a fixed clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. Standard errors for the above simulations are < 0.001.

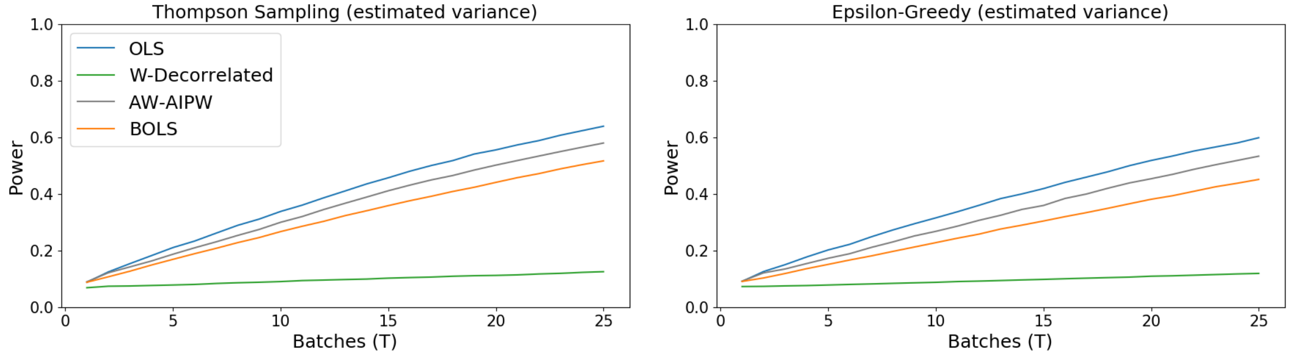

Figure 4. **Stationary Setting:** We plot the power for estimators of treatment effect for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = 0$, $\beta_0 = 0.25$, and use 25 samples per batch. We use a fixed clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. All standard errors for the above simulations are < 0.002.

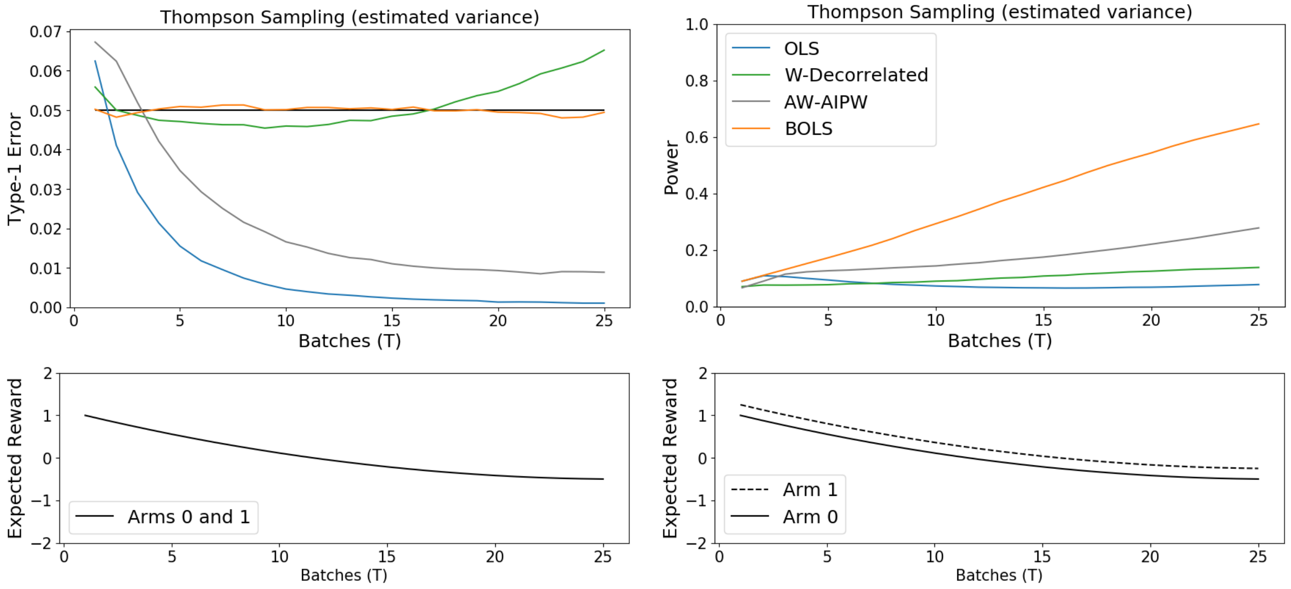

Figure 5. **Non-stationary baseline reward setting:** Type-1 error (upper left) and power (upper right) for estimators of treatment effect for a two-sided test of $H_0 : \Delta = 0$ vs. $H_1 : \Delta \neq 0$ ($\alpha = 0.05$). The expected rewards for each arm over time (batches) are plotted in the lower two plots; note that the treatment effect is constant across all batches. We use 25 samples per batch and a fixed clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. All standard errors for the above simulations are < 0.002.

Acknowledgements

KZ is supported by the NSF Graduate Research Fellowship under Grant No. DGE 1745303. Further funding was provided the National Institutes of Health, National Institute on Alcohol Abuse and Alcoholism (R01AA23187).

References

Agarwal, A., Agarwal, S., Assadi, S., and Khanna, S. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Conference on Learning Theory*, pp. 39–75, 2017.

Amemiya, T. *Advanced Econometrics*. Harvard University Press, 1985.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1194–1203, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Dvoretzky, A. Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. Batched multi-armed bandits problem. *Conference on Neural Information Processing Systems*, 2019.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019.

Hazelton, M. L. *Methods of Moments Estimation*, pp. 816–817. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_364. URL https://doi.org/10.1007/978-3-642-04898-2_364.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jun, K.-S., Jamieson, K. G., Nowak, R. D., and Zhu, X. Top arm identification in multi-armed bandits with batch arm pulls. In *AISTATS*, pp. 139–148, 2016.

Kasy, M. Uniformity and the delta method. *Journal of Econometric Methods*, 8(1), 2019.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

Lai, T. L. and Wei, C. Z. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.

Liao, P., Klasnja, P., Tewari, A., and Murphy, S. A. Sample size calculations for micro-randomized trials in mhealth. *Statistics in medicine*, 35(12):1944–1971, 2016.

Nie, X., Tian, X., Taylor, J., and Zou, J. Why adaptively collected data have negative bias and how to correct for it. *International Conference on Artificial Intelligence and Statistics*, 2018.

Perchet, V., Rigollet, P., Chassang, S., Snowberg, E., et al. Batched bandit problems. *The Annals of Statistics*, 44(2): 660–681, 2016.

Rafferty, A., Ying, H., and Williams, J. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *JEDM— Journal of Educational Data Mining*, 11(1):47–79, 2019.

Romano, J. P., Shaikh, A. M., et al. On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, 40(6):2798–2822, 2012.

Shin, J., Ramdas, A., and Rinaldo, A. Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems*, pp. 7100–7109, 2019.

Villar, S. S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Yom-Tov, E., Feraru, G., Kozdoba, M., Mannor, S., Tennenholtz, M., and Hochberg, I. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

A. Simulation Details

A.1. W-Decorrelated Estimator

For the W-decorrelated estimator (Deshpande et al., 2018), for a batch size of n and for T batches, we set λ to be the $\frac{1}{nT}$ quantile of $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})/\log(nT)$, where $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})$ denotes the minimum eigenvalue of $\underline{\mathbf{X}}^\top \underline{\mathbf{X}}$. This procedure of choosing λ is derived from Theorem 4 in Deshpande et al. (2018) and is based on what Deshpande et al. (2018) do in their simulation experiments. We had to adjust the original procedure for choosing λ used by Deshpande et al. (2018) (who set λ to the 0.15 quantile of $\lambda_{\min}(\underline{\mathbf{X}}^\top \underline{\mathbf{X}})$), because they only evaluated the W-decorrelated method for when the total number of samples was $nT = 1000$ and valid values of λ changes with the sample size.

A.2. AW-AIPW Estimator

Since the AW-AIPW test statistic for the treatment effect is not explicitly written in the original paper (Hadad et al., 2019), we now write the formulas for the AW-AIPW estimator of the treatment effect: $\hat{\Delta}^{\text{AW-AIPW}} := \hat{\beta}_1^{\text{AW-AIPW}} - \hat{\beta}_0^{\text{AW-AIPW}}$. We use the variance stabilizing weights, equal to the square root of the sampling probabilities, $\sqrt{\pi_t^{(n)}}$ and $\sqrt{1 - \pi_t^{(n)}}$.

$$Y_{t,1} := \frac{A_{t,i}^{(n)}}{\pi_t^{(n)}} R_{t,i}^{(n)} + \left(1 - \frac{A_{t,i}^{(n)}}{\pi_t^{(n)}}\right) \frac{\sum_{t'=1}^{t-1} \sum_{i=1}^n A_{t,i}^{(n)} R_{t,i}^{(n)}}{\sum_{t'=1}^{t-1} N_{t',1}}$$

$$Y_{t,0} := \frac{1 - A_{t,i}^{(n)}}{1 - \pi_t^{(n)}} R_{t,i}^{(n)} + \left(1 - \frac{1 - A_{t,i}^{(n)}}{1 - \pi_t^{(n)}}\right) \frac{\sum_{t'=1}^{t-1} \sum_{i=1}^n (1 - A_{t,i}^{(n)}) R_{t,i}^{(n)}}{\sum_{t'=1}^{t-1} N_{t',0}}$$

$$\hat{\beta}_1^{\text{AW-AIPW}} := \frac{\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}} Y_{t,1}}{\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}}} \quad \text{and} \quad \hat{\beta}_0^{\text{AW-AIPW}} := \frac{\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}} Y_{t,0}}{\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}}}$$

The variance estimator for $\hat{\Delta}^{\text{AW-AIPW}}$ is $\hat{V}_0 + \hat{V}_1 + 2\hat{C}_{0,1}$ where

$$\hat{V}_1 := \frac{\sum_{t=1}^T \sum_{i=1}^n \pi_t^{(n)} (Y_{t,1} - \hat{\beta}_1^{\text{AW-AIPW}})^2}{\left(\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}}\right)^2} \quad \text{and} \quad \hat{V}_0 := \frac{\sum_{t=1}^T \sum_{i=1}^n (1 - \pi_t^{(n)}) (Y_{t,0} - \hat{\beta}_0^{\text{AW-AIPW}})^2}{\left(\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}}\right)^2}$$

$$\hat{C}_{0,1} := -\frac{\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}(1 - \pi_t^{(n)})} (Y_{t,1} - \hat{\beta}_1^{\text{AW-AIPW}})(Y_{t,0} - \hat{\beta}_0^{\text{AW-AIPW}})}{\left(\sum_{t=1}^T \sum_{i=1}^n \sqrt{\pi_t^{(n)}}\right)\left(\sum_{t=1}^T \sum_{i=1}^n \sqrt{1 - \pi_t^{(n)}}\right)}$$

A.3. Estimating Noise Variance

OLS Estimator Given the OLS estimators for the means of each arm, $\hat{\beta}_1^{\text{OLS}}, \hat{\beta}_0^{\text{OLS}}$, we estimate the noise variance σ^2 as follows:

$$\hat{\sigma}^2 := \frac{1}{nT - 2} \sum_{t=1}^T \sum_{i=1}^n \left(R_{t,i} - A_{t,i} \hat{\beta}_1^{\text{OLS}} - (1 - A_{t,i}) \hat{\beta}_0^{\text{OLS}}\right)^2.$$

We use a degrees of freedom bias correction by normalizing by $nT - 2$ rather than nT. Since the W-decorrelated estimator is a modified version of the OLS estimator, we also use this same noise variance estimator for the W-decorrelated estimator; we found that this worked well in practice, in terms of Type-1 error control.

Batched OLS Given the Batched OLS estimators for the means of each arm for each batch, $\hat{\beta}_{t,1}^{\text{BOLS}}, \hat{\beta}_{t,0}^{\text{BOLS}}$, we estimate the noise variance for each batch σ_t^2 as follows:

$$\hat{\sigma}_t^2 := \frac{1}{n - 2} \sum_{i=1}^n \left(R_{t,i} - A_{t,i} \hat{\beta}_{t,1}^{\text{BOLS}} - (1 - A_{t,i}) \hat{\beta}_{t,0}^{\text{BOLS}}\right)^2.$$

Again, we use a degrees of freedom bias correction by normalizing by $n-2$ rather than n. Using BOLS to test $H_0 : \Delta = a$ vs. $H_1 : \Delta \neq a$, we use the following test statistic:

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sqrt{\frac{N_{t,0}N_{t,1}}{n\hat{\sigma}_t^2}} (\hat{\Delta}_t^{\text{BOLS}} - a).$$

For this test statistic, we use cutoffs based on the Student-t distribution, i.e., for $Y_t \overset{i.i.d.}{\sim} t_{n-2}$ we use a cutoff $c_{\alpha/2}$ such that

$$\mathbb{P}\left(\left| \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_t \right| > c_{\alpha/2} \right) = \alpha.$$

We found $c_{\alpha/2}$ by simulating draws from the Student-t distribution.

A.4. Non-Stationary Treatment Effect

In the non-stationary treatment effect simulations, we test the null that $H_0 \colon \forall t \in [1\colon T], \beta_{t,1} - \beta_{t,0} = 0$ vs. $H_1 \colon \exists t \in [1\colon T], \beta_{t,1} - \beta_{t,0} \neq 0$. To test this we use the following test statistic:

$$\frac{1}{T} \sum_{t=1}^{T} \frac{N_{t,0}N_{t,1}}{n\hat{\sigma}_t^2} (\hat{\Delta}_t^{\text{BOLS}} - 0)^2. \tag{6}$$

For this test statistic, we use cutoffs based on the Student-t distribution, i.e., for $Y_t \overset{i.i.d.}{\sim} t_{n-2}$ we use a cutoff $c_{\alpha/2}$ such that

$$\mathbb{P}\left(\frac{1}{T} \sum_{t=1}^{T} Y_t^2 > c_\alpha \right) = \alpha.$$

We found c_α by simulating draws from the Student-t distribution.

In the plots below we call the test statistic in (6) "BOLS Non-Stationary Treatment Effect" (BOLS NSTE). BOLS NSTE performs poorly in terms of power compared to other test statistics in the stationary setting; however, *in the non-stationary setting, BOLS NSTE significantly outperforms all other test statistics, which tend to have low power when the average treatment effect is close to zero.* Note that the W-decorrelated estimator performs well in the left plot of Figure 8; this is because as we show in Appendix F, the W-decorrelated estimator upweights samples from the earlier batches in the study. So when the treatment effect is large in the beginning of the study, the W-decorrelated estimator has high power and when the treatment effect is small or zero in the beginning of the study, the W-decorrelated estimator has low power.

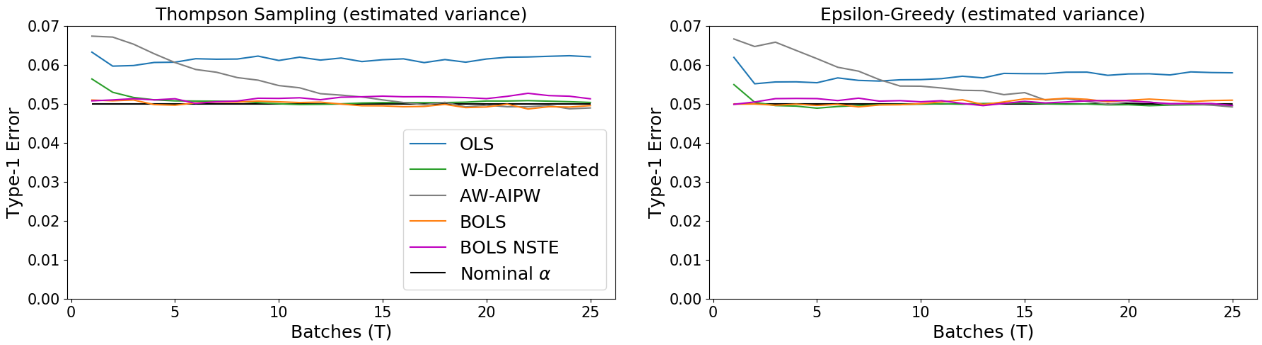

Figure 6. **Stationary Setting:** Type-1 error for estimators of treatment effect for a two-sided test of $H_0 \colon \Delta = 0$ vs. $H_1 \colon \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = \beta_0 = 0$ and use 25 samples per batch. We use a fixed clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. Standard errors for the above simulations are < 0.001.

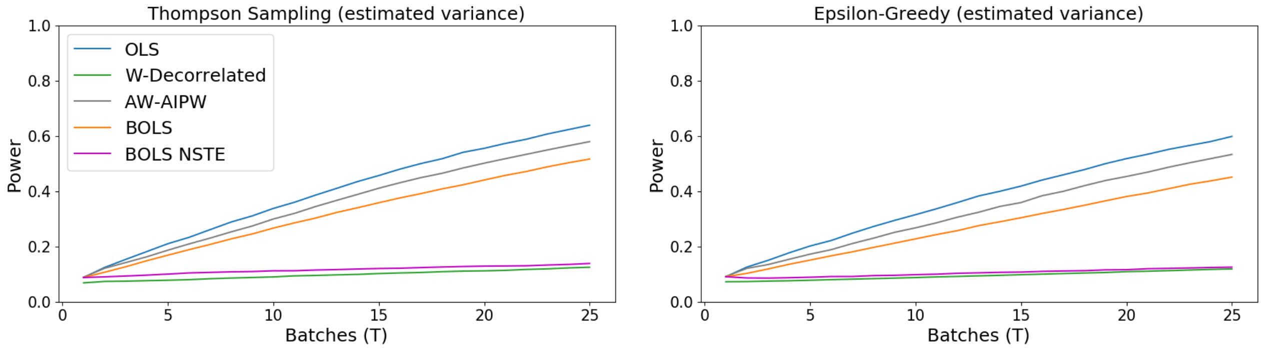

Figure 7. **Stationary Setting:** We plot the power for estimators of treatment effect for a two-sided test of $H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0$ ($\alpha = 0.05$). We set $\beta_1 = 0$, $\beta_0 = 0.25$, and use 25 samples per batch. We use a fixed clipping constraint of $0.1 \leq \pi_t^{(n)} \leq 0.9$. All standard errors for the above simulations are < 0.002.

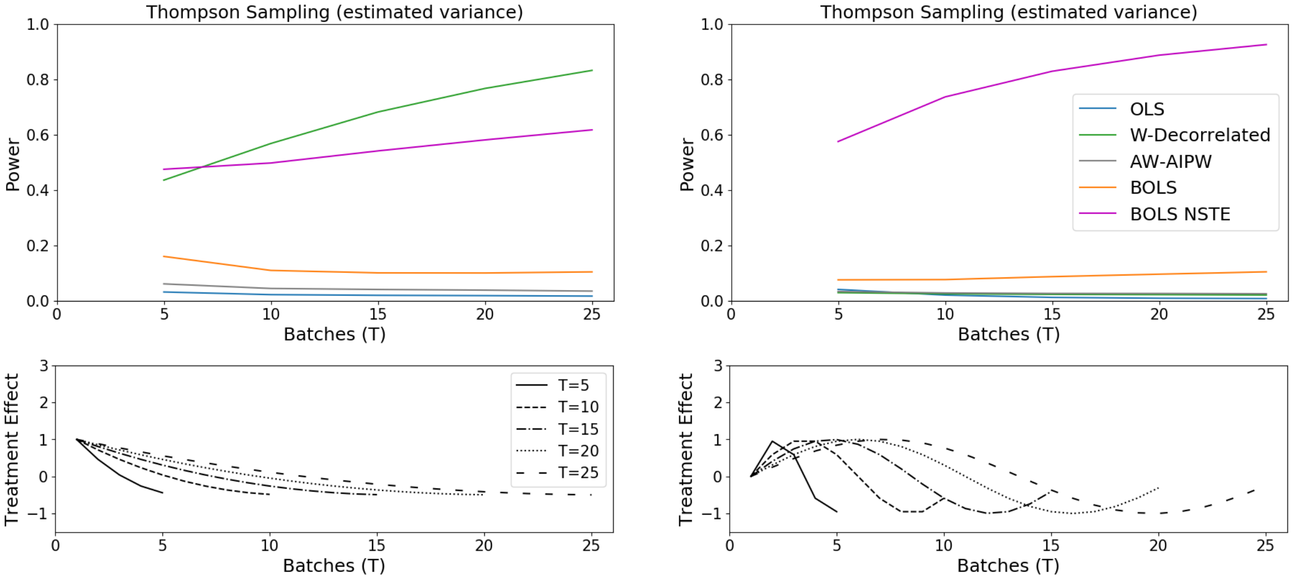

Figure 8. **Nonstationary setting:** The two upper plots display the power of estimators for a two-sided test of $H_0: \forall t \in [1: T],\ \beta_{t,1} - \beta_{t,0} = 0$ vs. $H_1: \exists t \in [1: T], \beta_{t,1} - \beta_{t,0} \neq 0$ ($\alpha = 0.05$). The two lower plots display two treatment effect trends; the left plot considers a decreasing trend (quadratic function) and the right plot considers a oscillating trend (sin function). We use 25 samples per batch and use a clipping constraint is $0.1 \leq \pi_t \leq 0.9$. All standard errors for the above simulations are < 0.002.

B. Asymptotic Normality of the OLS Estimator

Condition 6 (Weak moments). $\forall t, n, i$, $\mathbb{E}[(\epsilon_{t,i}^{(n)})^2 \big| \mathcal{G}_{t-1}^{(n)}] = \sigma^2$ *and for all* $\forall t, n, i$, $\mathbb{E}[\varphi([\epsilon_{t,i}^{(n)}]^2)|\mathcal{G}_{t-1}^{(n)}] < M < \infty$ *a.s. for some function* φ *where* $\lim_{x\to\infty} \frac{\varphi(x)}{x} \to \infty$.

Condition 7 (Stability). *There exists a sequence of nonrandom positive-definite symmetric matrices,* $\underline{\boldsymbol{V}}_n$*, such that*

(a) $\underline{\boldsymbol{V}}_n^{-1}\big(\sum_{t=1}^T \sum_{i=1}^n \boldsymbol{X}_{t,i}\boldsymbol{X}_{t,i}^\top\big)^{\frac{1}{2}} = \underline{\boldsymbol{V}}_n^{-1}\big(\underline{\boldsymbol{X}}^\top \underline{\boldsymbol{X}}\big)^{\frac{1}{2}} \xrightarrow{P} \underline{\boldsymbol{I}}_p$

(b) $\max_{i\in[1:\,n], t\in[1:\,T]} \|\underline{\boldsymbol{V}}_n^{-1}\boldsymbol{X}_{t,i}\|_2 \xrightarrow{P} 0$

Theorem 5 (Triangular array version of Lai & Wei (1982), Theorem 3). *Let* $\boldsymbol{X}_{t,i}^{(n)} \in \mathbb{R}^p$ *be non-anticipating with respect to filtration* $\{\mathcal{G}_t^{(n)}\}_{t=1}^T$*, so* $\boldsymbol{X}_{t,i}^{(n)}$ *is* $\mathcal{G}_{t-1}^{(n)}$ *measurable. We assume the following conditional mean model for rewards:*

$$\mathbb{E}\big[R_{t,i}^{(n)}\big|\mathcal{G}_{t-1}^{(n)}\big] = (\boldsymbol{X}_{t,i}^{(n)})^\top \boldsymbol{\beta}.$$

We define $\epsilon_{t,i}^{(n)} := R_{t,i}^{(n)} - (\boldsymbol{X}_{t,i}^{(n)})^\top\boldsymbol{\beta}$*. Note that* $\{\epsilon_{t,i}^{(n)}\}$ *is a martingale difference array with respect to filtration* $\{\mathcal{G}_t^{(n)}\}_{t=1}^T$*. Assuming Conditions 6 and 7, as* $n \to \infty$*,*

$$\big(\underline{\boldsymbol{X}}^{(n),\top}\underline{\boldsymbol{X}}^{(n)}\big)^{1/2}(\hat{\boldsymbol{\beta}}^{\text{OLS}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_p)$$

Note, in the body of the paper we state that this theorem holds in the two-arm bandit case assuming Conditions 2 and 1. Note that Condition 1 is sufficient for Condition 6 and Condition 2 is sufficient for Condition 7 in the two-arm bandit case.

Proof:

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = \left((\underline{\mathbf{X}}^{(n)})^\top \underline{\mathbf{X}}^{(n)}\right)^{-1} \underline{\mathbf{X}}^{(n),\top}\mathbf{R}^{(n)} = (\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top(\underline{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon})$$

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} - \boldsymbol{\beta} = (\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top\boldsymbol{\epsilon} = \left(\sum_{t=1}^T\sum_{i=1}^n \mathbf{X}_{t,i}\mathbf{X}_{t,i}^\top\right)^{-1} \sum_{t=1}^T\sum_{i=1}^n \mathbf{X}_{t,i}\epsilon_{t,i}$$

It is sufficient to show that as $n \to \infty$:

$$(\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1/2}\sum_{t=1}^T\sum_{i=1}^n \mathbf{X}_{t,i}\epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2\underline{\mathbf{I}}_p)$$

By Slutsky's Theorem and Condition 7 (a), it is also sufficient to show that as $n \to \infty$,

$$\underline{\mathbf{V}}_n^{-1}\sum_{t=1}^T\sum_{i=1}^n \mathbf{X}_{t,i}\epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2\underline{\mathbf{I}}_p)$$

By Cramer-Wold device, it is sufficient to show multivariate normality if for any fixed $\mathbf{c} \in \mathbb{R}^p$ s.t. $\|\mathbf{c}\|_2 = 1$, as $n \to \infty$,

$$\mathbf{c}^\top\underline{\mathbf{V}}_n^{-1}\sum_{t=1}^T\sum_{i=1}^n \mathbf{X}_{t,i}\epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

We will prove this central limit theorem by using a triangular array martingale central limit theorem, specifically Theorem 2.2 of Dvoretzky (1972). We will do this by letting $Y_{t,i}^{(n)} = \mathbf{c}^T\underline{\mathbf{V}}_n^{-1}\mathbf{X}_{t,i}\epsilon_{t,i}$. The theorem states that as $n \to \infty$, $\sum_{t=1}^T\sum_{i=1}^n Y_{t,i}^{(n)} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$ if the following conditions hold as $n \to \infty$:

(a) $\sum_{t=1}^T\sum_{i=1}^n E[Y_{t,i}^{(n)}|\mathcal{G}_{t-1}^{(n)}] \xrightarrow{P} 0$

(b) $\sum_{t=1}^T\sum_{i=1}^n E[Y_{t,i}^{(n),2}|\mathcal{G}_{t-1}^{(n)}] \xrightarrow{P} \sigma^2$

(c) $\forall\delta > 0, \sum_{t=1}^T\sum_{i=1}^n E\big[Y_{t,i}^{(n),2}\mathbb{I}_{(|Y_{t,i}^{(n)}|>\delta)}\big|\mathcal{G}_{t-1}^{(n)}\big] \xrightarrow{P} 0$

Useful Properties Note that by Cauchy-Schwartz and Condition 7 (b), as $n \to \infty$,

$$\max_{i \in [1:\, n], t \in [1:\, T]} \left| \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right| \leq \max_{i \in [1:\, n], t \in [1:\, T]} \|\mathbf{c}\|_2 \|\underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i}\|_2 \xrightarrow{P} 0$$

By continuous mapping theorem and since the square function on non-negative inputs is order preserving, as $n \to \infty$,

$$\left(\max_{i \in [1:\, n], t \in [1:\, T]} \left| \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right| \right)^2 = \max_{i \in [1:\, n], t \in [1:\, T]} \left(\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \right)^2 \xrightarrow{P} 0 \tag{7}$$

By Condition 7 (a) and continuous mapping theorem, $\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} \xrightarrow{P} \mathbf{c}^\top$, so

$$\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} (\mathbf{X}_{t,i}^\top \mathbf{X}_{t,i})^{1/2} \underline{\mathbf{V}}_n^{-1} \mathbf{c} \xrightarrow{P} \mathbf{c}^\top \mathbf{c} = 1$$

Thus,

$$\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \bigg(\sum_{t=1}^T \sum_{i=1}^n \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top \bigg) \underline{\mathbf{V}}_n^{-1} \mathbf{c} = \sum_{t=1}^T \sum_{i=1}^n \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \mathbf{X}_{t,i}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{c} \xrightarrow{P} 1$$

Since $\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i}$ is a scalar, as $n \to \infty$,

$$\sum_{t=1}^T \sum_{i=1}^n (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \xrightarrow{P} 1 \tag{8}$$

Condition (a): Martingale

$$\sum_{t=1}^T \sum_{i=1}^n \mathbb{E}[\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \epsilon_{t,i} | \mathcal{G}_{t-1}^{(n)}] = \sum_{t=1}^T \sum_{i=1}^n \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i} \mathbb{E}[\epsilon_{t,i} | \mathcal{G}_{t-1}^{(n)}] = 0$$

Condition (b): Conditional Variance

$$\sum_{t=1}^T \sum_{i=1}^n \mathbb{E}[(\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \epsilon_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] = \sum_{t=1}^T \sum_{i=1}^n (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \mathbb{E}[\epsilon_{t,i}^2 | \mathcal{G}_{t-1}^{(n)}] = \sigma^2 \sum_{t=1}^T \sum_{i=1}^n (\mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i})^2 \xrightarrow{P} \sigma^2$$

where the last equality holds by Condition 6 and the limit holds by (8) as $n \to \infty$.

Condition (c): Lindeberg Condition Let $\delta > 0$. We want to show that as $n \to \infty$,

$$\sum_{t=1}^T \sum_{i=1}^n Z_{t,i}^2 \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2 \epsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)} \right] \xrightarrow{P} 0$$

where above, we define $Z_{t,i}^{(n)} := \mathbf{c}^\top \underline{\mathbf{V}}_n^{-1} \mathbf{X}_{t,i}$. By Condition 6, we have that for all $n \geq 1$,

$$\max_{t \in [1:\, T], i \in [1:\, n]} \mathbb{E}[\varphi(\epsilon_{t,i}^2) | \mathcal{G}_{t-1}^{(n)}] < M$$

Since we assume that $\lim_{x \to \infty} \frac{\varphi(x)}{x} = \infty$, for all $m \geq 1$, there exists a b_m s.t. $\varphi(x) \geq mMx$ for all $x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) | \mathcal{G}_{t-1}^{(n)}] \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | \mathcal{G}_{t-1}^{(n)}] \geq mM \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | \mathcal{G}_{t-1}^{(n)}]$$

Thus,

$$\max_{t \in [1:\, T], i \in [1:\, n]} \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | \mathcal{G}_{t-1}^{(n)}] \leq \frac{1}{m}$$

So we have that

$$\sum_{t=1}^T \sum_{i=1}^n Z_{t,i}^2 \mathbb{E}\big[\epsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2 \epsilon_{t,i}^2 > \delta^2)} | \mathcal{G}_{t-1}^{(n)} \big]$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n} Z_{t,i}^2 \left(\mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2 \epsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)} \right] \mathbb{I}_{(Z_{t,i}^2 \leq \delta^2 / b_m)} + \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(Z_{t,i}^2 \epsilon_{t,i}^2 > \delta^2)} \middle| \mathcal{G}_{t-1}^{(n)} \right] \mathbb{I}_{(Z_{t,i}^2 > \delta^2 / b_m)} \right)$$

$$\leq \sum_{t=1}^{T} \sum_{i=1}^{n} Z_{t,i}^2 \left(\mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 > b_m)} \middle| \mathcal{G}_{t-1}^{(n)} \right] + \sigma^2 \mathbb{I}_{(Z_{t,i}^2 > \delta^2 / b_m)} \right)$$

$$\leq \left(\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} \right) \sum_{t=1}^{T} \sum_{i=1}^{n} Z_{t,i}^2$$

By Slutsky's Theorem and (8), it is sufficient to show that as $n \to \infty$,

$$\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} \xrightarrow{P} 0$$

For any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{m} + \sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \epsilon \right) \leq \mathbb{I}_{(\frac{1}{m} > \frac{\epsilon}{2})} + \mathbb{P}\left(\sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \frac{\epsilon}{2} \right)$$

We can choose m such that $\frac{1}{m} \leq \frac{\epsilon}{2}$, so $\mathbb{P}\big(\frac{1}{m} > \frac{\epsilon}{2}\big) = 0$. For the second term (note that m is now fixed),

$$\mathbb{P}\left(\sigma^2 \mathbb{I}_{(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m)} > \frac{\epsilon}{2} \right) \leq \mathbb{P}\left(\max_{t' \in [1:\, T], j \in [1:\, n]} Z_{t',j}^2 > \delta^2 / b_m \right) \to 0$$

where the last limit holds by (7) as $n \to \infty$. □

B.1. Corollary 1 (Sufficient conditions for Theorem 5)

Under Conditions 1 and 3, when ***the treatment effect is non-zero*** *data collected in batches using ϵ-greedy or Thompson Sampling with a fixed clipping constraint (see Definition 1) will satisfy Theorem 5 conditions.*

Proof: The only condition of Theorem 5 that needs to verified is Condition 2. To satisfy Condition 2, it is sufficient to show that for any given Δ, for some constant $c \in (0, T)$,

$$\frac{1}{n} \sum_{t=1}^{T} N_{t,1}^{(n)} \xrightarrow{P} c.$$

ϵ-greedy We assume without loss of generality that $\Delta > 0$ and $\pi_1^{(n)} = \frac{1}{2}$. Recall that for ϵ-greedy, for $a \in [2\colon T]$,

$$\pi_a^{(n)} = \begin{cases} 1 - \frac{\epsilon}{2} & \text{if } \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} A_{t,i} R_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} > \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} (1 - A_{t,i}) R_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} \\ \frac{\epsilon}{2} & \text{otherwise} \end{cases}$$

Thus to show that $\pi_a^{(n)} \xrightarrow{P} 1 - \frac{\epsilon}{2}$ for all $a \in [2\colon T]$, it is sufficient to show that

$$\mathbb{P}\left(\frac{\sum_{t=1}^{a} \sum_{i=1}^{n} A_{t,i} R_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} > \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} (1 - A_{t,i}) R_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} \right) \to 1 \tag{9}$$

To show (9), it is equivalent to show that

$$\mathbb{P}\left(\Delta > \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} (1 - A_{t,i}) \epsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} - \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} A_{t,i} \epsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} \right) \to 1 \tag{10}$$

To show (10), it is sufficient to show that

$$\frac{\sum_{t=1}^{a} \sum_{i=1}^{n} (1 - A_{t,i}) \epsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',0}} - \frac{\sum_{t=1}^{a} \sum_{i=1}^{n} A_{t,i} \epsilon_{t,i}}{\sum_{t'=1}^{a} N_{t',1}} \xrightarrow{P} 0. \tag{11}$$

To show (11), it is equivalent to show that

$$\sum_{t=1}^{a} \frac{\sqrt{N_{t,0}}}{\sum_{t'=1}^{a} N_{t',0}} \frac{\sum_{i=1}^{n}(1-A_{t,i})\epsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} \frac{\sqrt{N_{t,1}}}{\sum_{t'=1}^{a} N_{t',1}} \frac{\sum_{i=1}^{n} A_{t,i}\epsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0. \tag{12}$$

By Lemma 1, for all $t \in [1\colon T]$,

$$\frac{N_{t,1}}{\pi_t^{(n)} n} \xrightarrow{P} 1$$

Thus by Slutsky's Theorem, to show (12), it is sufficient to show that

$$\sum_{t=1}^{a} \frac{\sqrt{n(1-\pi_t^{(n)})}}{n\sum_{t'=1}^{a}(1-\pi_{t'}^{(n)})} \frac{\sum_{i=1}^{n}(1-A_{t,i})\epsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} \frac{\sqrt{n\pi_t^{(n)}}}{n\sum_{t'=1}^{a}\pi_{t'}^{(n)}} \frac{\sum_{i=1}^{n} A_{t,i}\epsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0. \tag{13}$$

Since $\pi_t^{(n)} \in [\frac{\epsilon}{2}, 1-\frac{\epsilon}{2}]$ for all t, n, the left hand side of (13) equals the following:

$$\sum_{t=1}^{a} o_p(1) \frac{\sum_{i=1}^{n}(1-A_{t,i})\epsilon_{t,i}}{\sqrt{N_{t,0}}} - \sum_{t=1}^{a} o_p(1) \frac{\sum_{i=1}^{n} A_{t,i}\epsilon_{t,i}}{\sqrt{N_{t,1}}} \xrightarrow{P} 0.$$

The above limit holds because by Thereom 3, we have that

$$\left(\frac{\sum_{i=1}^{n} A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^{n}(1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}}, ..., \frac{\sum_{i=1}^{n} A_{T,i}\epsilon_{T,i}}{\sqrt{N_{T,1}}}, \frac{\sum_{i=1}^{n}(1-A_{T,i})\epsilon_{T,i}}{\sqrt{N_{T,0}}} \right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \underline{\boldsymbol{I}}_{2T}). \tag{14}$$

Thus, by Slutsky's Theorem and Lemma 1, we have that

$$\frac{1}{n}\sum_{t=1}^{T} N_{t,1} \xrightarrow{P} \frac{1}{2} + (T-1)(1-\frac{\epsilon}{2}) \qquad \text{and} \qquad \frac{1}{n}\sum_{t=1}^{T} N_{t,0} \xrightarrow{P} \frac{1}{2} + (T-1)\frac{\epsilon}{2}$$

Thompson Sampling We assume without loss of generality that $\Delta > 0$ and $\pi_1^{(n)} = \frac{1}{2}$. Recall that for Thompson Sampling with independent standard normal priors ($\tilde{\beta}_1, \tilde{\beta}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$) for $a \in [2\colon T]$,

$$\pi_a^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \mathbb{P}(\tilde{\beta}_1 > \tilde{\beta}_0 \mid H_{a-1}^{(n)})\right]$$

Given the independent standard normal priors on $\tilde{\beta}_1, \tilde{\beta}_0$, we have the following posterior distribution:

$$\tilde{\beta}_1 - \tilde{\beta}_0 \mid H_{a-1}^{(n)} \sim \mathcal{N}\left(\frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n} A_{t,i}R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1}\sum_{i=1}^{n}(1-A_{t,i})R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}}, \frac{\sigma^2(\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}) + \sigma^2(\sigma^2 + \sum_{t=1}^{a-1} N_{t,0})}{(\sigma^2 + \sum_{t=1}^{a-1} N_{t,0})(\sigma^2 + \sum_{t=1}^{a-1} N_{t,1})} \right)$$

$$=: \mathcal{N}\left(\mu_{a-1}^{(n)}, (\sigma_{a-1}^{(n)})^2\right)$$

Thus to show that $\pi_a^{(n)} \xrightarrow{P} \pi_{\max}$ for all $a \in [2\colon T]$, it is sufficient to show that $\mu_{a-1}^{(n)} \xrightarrow{P} \Delta$ and $(\sigma_{a-1}^{(n)})^2 \xrightarrow{P} 0$ for all $a \in [2\colon T]$. By Lemma 1, for all $t \in [1\colon T]$,

$$\frac{N_{t,1}}{\pi_t^{(n)} n} \xrightarrow{P} 1$$

Thus, to show $(\sigma_{a-1}^{(n)})^2 \xrightarrow{P} 0$, it is sufficient to show that

$$\frac{\sigma^2(\sigma^2 + n\sum_{t=1}^{a-1}\pi_t^{(n)}) + \sigma^2(\sigma^2 + n\sum_{t=1}^{a-1}(1-\pi_t^{(n)}))}{(\sigma^2 + n\sum_{t=1}^{a-1}(1-\pi_t^{(n)}))(\sigma^2 + n\sum_{t=1}^{a-1}\pi_t^{(n)})} \xrightarrow{P} 0$$

The above limit holds because $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ for $0 < \pi_{\min} \leq \pi_{\max} < 1$ by the clipping condition.

We now show that $\mu_{a-1}^{(n)} \xrightarrow{P} \Delta$, which is equivalent to showing that the following converges in probability to Δ

$$\frac{\sum_{t=1}^{a-1} \sum_{i=1}^n A_{t,i} R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}}$$

$$= \frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n A_{t,i} R_{t,i}}{\sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{\sum_{t=1}^{a-1} N_{t,0}}$$

$$= \frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \left(\beta_1 + \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sum_{t=1}^{a-1} N_{t,1}} \right) - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \left(\beta_0 + \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sum_{t=1}^{a-1} N_{t,0}} \right) \tag{15}$$

Note that

$$\frac{\sum_{t=1}^{a-1} N_{t,1}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} \beta_1 - \frac{\sum_{t=1}^{a-1} N_{t,0}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \beta_0 \xrightarrow{P} \Delta \tag{16}$$

Equation (16) above holds by Lemma 1, because

$$\frac{n \sum_{t=1}^{a-1} \pi_t^{(n)}}{\sigma^2 + n \sum_{t=1}^{a-1} \pi_t^{(n)}} \xrightarrow{P} 1 \qquad\qquad \frac{n \sum_{t=1}^{a-1} (1 - \pi_t^{(n)})}{\sigma^2 + n \sum_{t=1}^{a-1} (1 - \pi_t^{(n)})} \xrightarrow{P} 1 \tag{17}$$

which hold because $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ due to our clipping condition.

By Slutsky's Theorem and (16), to show (15), it is sufficient to show that

$$\frac{\sum_{t=1}^{a-1} \sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,1}} - \frac{\sum_{t=1}^{a-1} \sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sigma^2 + \sum_{t=1}^{a-1} N_{t,0}} \xrightarrow{P} 0. \tag{18}$$

Equation (18) is equivalent to the following:

$$\sum_{t=1}^{a-1} \frac{\sqrt{N_{t,1}}}{\sigma^2 + \sum_{t'=1}^{a-1} N_{t',1}} \frac{\sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} \frac{\sqrt{N_{t,0}}}{\sigma^2 + \sum_{t'=1}^{a-1} N_{t',0}} \frac{\sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0 \tag{19}$$

By Lemma 1, to show (19) it is sufficient to show that

$$\sum_{t=1}^{a-1} \frac{\sqrt{n \pi_t^{(n)}}}{\sigma^2 + n \sum_{t'=1}^{a-1} \pi_{t'}^{(n)}} \frac{\sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} \frac{\sqrt{n (1 - \pi_t^{(n)})}}{\sigma^2 + n \sum_{t'=1}^{a-1} (1 - \pi_{t'}^{(n)})} \frac{\sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0 \tag{20}$$

Since $\pi_t^{(n)} \in [\pi_{\min}, \pi_{\max}]$ due to our clipping condition, the left hand side of (20) equals the following

$$\sum_{t=1}^{a-1} o_p(1) \frac{\sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{t,1}}} - \sum_{t=1}^{a-1} o_p(1) \frac{\sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sqrt{N_{t,0}}} \xrightarrow{P} 0$$

The above limit holds by (14).

Thus, by Slutsky's Theorem and Lemma 1, we have that

$$\frac{1}{n} \sum_{t=1}^T N_{t,1} \xrightarrow{P} \frac{1}{2} + (T-1)\pi_{\max} \qquad \text{and} \qquad \frac{1}{n} \sum_{t=1}^T N_{t,0} \xrightarrow{P} \frac{1}{2} + (T-1)\pi_{\min} \qquad \square$$

C. Non-uniform convergence of the OLS Estimator

Definition 3 (Non-concentration of a sequence of random variables). *For a sequence of random variables $\{Y_i\}_{i=1}^n$ on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we say Y_n does not concentrate if for each $a \in \mathbb{R}$ there exists an $\epsilon_a > 0$ with*

$$P\big(\{\omega \in \Omega : \big|Y_n(\omega) - a\big| > \epsilon_a\}\big) \not\to 0.$$

Proposition 1 (Non-concentration of sampling probabilities under Thompson Sampling). *Under the assumptions of Theorem 2, the posterior distribution that arm 1 is better than arm 0 converges as follows:*

$$\mathbb{P}\big(\tilde{\beta}_1 > \tilde{\beta}_0 \mid H_1^{(n)}\big) \xrightarrow{D} \begin{cases} 1 & \text{if } \Delta > 0 \\ 0 & \text{if } \Delta < 0 \\ \text{Uniform}[0,1] & \text{if } \Delta = 0 \end{cases}$$

Thus, the sampling probabilities $\pi_t^{(n)}$ do not concentrate when $\Delta = 0$.

Proof: Posterior means:

$$\tilde{\beta}_0 | H_1^{(n)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n (1 - A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}}, \frac{\sigma^2}{\sigma_a^2 + N_{0,1}} \right)$$

$$\tilde{\beta}_1 | H_1^{(n)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}}, \frac{\sigma_a^2}{\sigma_a^2 + N_{1,1}} \right)$$

$$\tilde{\beta}_1 - \tilde{\beta}_0 \mid H_1^{(n)} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\sum_{i=1}^n (1 - A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}}, \frac{\sigma_a^2(\sigma_a^2 + N_{1,1}) + \sigma_a^2(\sigma_a^2 + N_{1,0})}{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})} \right) =: \mathcal{N}(\mu_n, \sigma_n^2)$$

$$P(\tilde{\beta}_1 > \tilde{\beta}_0 \mid H_1^{(n)}) = P(\tilde{\beta}_1 - \tilde{\beta}_0 > 0 \mid H_1^{(n)}) = P\left(\frac{\tilde{\beta}_1 - \tilde{\beta}_0 - \mu_n}{\sigma_n} > -\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right)$$

For $Z \sim \mathcal{N}(0,1)$ independent of μ_n, σ_n.

$$= P\left(Z > -\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right) = P\left(Z < \frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right) = \Phi\left(\frac{\mu_n}{\sigma_n} \,\middle|\, H_1^{(n)} \right)$$

$$\frac{\mu_n}{\sigma_n} = \left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\sum_{i=1}^n (1 - A_{1,i}) R_{1,i}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}}$$

$$= \left(\frac{\beta_1 N_{1,1} + \sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sigma_a^2 + N_{1,1}} - \frac{\beta_0 N_{1,0} + \sum_{i=1}^n (1 - A_{1,i}) \epsilon_{1,i}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}}$$

$$= \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} \sqrt{\frac{N_{1,1}(\sigma_a^2 + N_{1,0})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,1})}} - \frac{\sum_{i=1}^n (1 - A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \sqrt{\frac{N_{1,0}(\sigma_a^2 + N_{1,1})}{(2\sigma_a^4 + \sigma_a^2 n)(\sigma_a^2 + N_{1,0})}}$$

$$+ \left(\beta_1 \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} \right) \sqrt{\frac{(\sigma_a^2 + N_{1,0})(\sigma_a^2 + N_{1,1})}{2\sigma_a^4 + \sigma_a^2 n}} =: B_n + C_n$$

Let's first examine C_n. Note that $\beta_1 = \beta_0 + \Delta$, so $\beta_1 \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}}$ equals

$$= (\beta_0 + \Delta) \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \beta_0 \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} = \Delta \frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} + \beta_0 \left(\frac{N_{1,1}}{\sigma_a^2 + N_{1,1}} - \frac{N_{1,0}}{\sigma_a^2 + N_{1,0}} \right)$$

$$= \Delta \frac{N_{1,1}/n}{(\sigma_a^2 + N_{1,1})/n} + \beta_0 \left(\frac{N_{1,1}(\sigma_a^2 + N_{1,0}) - N_{1,0}(\sigma_a^2 + N_{1,1})}{(\sigma_a^2 + N_{1,1})(\sigma_a^2 + N_{1,1})} \right)$$

$$= \Delta \frac{\frac{1}{2} + o(1)}{\frac{1}{2} + o(1)} + \beta_0 \sigma_a^2 \left(\frac{N_{1,1} - N_{1,0}}{(\sigma_a^2 + N_{1,1})(\sigma_a^2 + N_{1,1})} \right) = \Delta[1 + o(1)] + o\left(\frac{1}{n} \right)$$

where the last equality holds by the Strong Law of Large Numbers because

$$\frac{\frac{1}{n^2}(N_{1,1}-N_{1,0})}{\frac{1}{n^2}(\sigma_a^2+N_{1,1})(\sigma_a^2+N_{1,1})}=\frac{\frac{1}{n}[\frac{1}{2}-\frac{1}{2}+o(1)]}{[\frac{1}{2}+o(1)][\frac{1}{2}+o(1)]}=\frac{\frac{1}{n}o(1)}{\frac{1}{4}+o(1)}=o\bigg(\frac{1}{n}\bigg)$$

Thus,

$$C_n=\bigg[\Delta[1+o(1)]+o\bigg(\frac{1}{n}\bigg)\bigg]\sqrt{\frac{(\sigma_a^2+N_{1,0})(\sigma_a^2+N_{1,1})}{2\sigma_a^4+\sigma_a^2 n}}$$

$$=\bigg[\Delta[1+o(1)]+o\bigg(\frac{1}{n}\bigg)\bigg]\sqrt{\frac{n[\frac{1}{2}+o(1)][\frac{1}{2}+o(1)]}{o(1)+\sigma_a^2}}=\sqrt{n}\Delta\big[1/(2\sigma_a)+o(1)\big]+o\bigg(\frac{1}{\sqrt{n}}\bigg)$$

Let's now examine B_n.

$$\sqrt{\frac{N_{1,1}(\sigma_a^2+N_{1,0})}{(2\sigma_a^4+\sigma_a^2 n)(\sigma_a^2+N_{1,1})}}=\sqrt{\frac{[\frac{1}{2}+o(1)][\frac{1}{2}+o(1)]}{[\sigma_a^2+o(1)][\frac{1}{2}+o(1)]}}=\sqrt{\frac{1}{2\sigma_a^2}}+o(1)$$

$$\sqrt{\frac{N_{1,0}(\sigma_a^2+N_{1,1})}{(2\sigma_a^4+\sigma_a^2 n)(\sigma_a^2+N_{1,0})}}=\sqrt{\frac{[\frac{1}{2}+o(1)][\frac{1}{2}+o(1)]}{[\sigma_a^2+o(1)][\frac{1}{2}+o(1)]}}=\sqrt{\frac{1}{2\sigma_a^2}}+o(1)$$

Note that by Theorem 3, $\Big[\frac{1}{\sqrt{N_{1,1}}}\sum_{i=1}^n \epsilon_{1,i}A_{1,i}, \frac{1}{\sqrt{N_{1,0}}}\sum_{i=1}^n \epsilon_{1,i}(1-A_{1,i})\Big]\overset{D}{\to}\mathcal{N}(\mathbf{0},\underline{\mathbf{I}}_2)$. Thus by Slutky's Theorem,

$$\begin{bmatrix}\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}\sqrt{\frac{N_{1,1}(\sigma_a^2+N_{1,0})}{(2\sigma_a^4+\sigma_a^2 n)(\sigma_a^2+N_{1,1})}}\\ \frac{\sum_{i=1}^n(1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}}\sqrt{\frac{N_{1,0}(\sigma_a^2+N_{1,1})}{(2\sigma_a^4+\sigma_a^2 n)(\sigma_a^2+N_{1,0})}}\end{bmatrix}=\begin{bmatrix}\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}\big[\sqrt{\frac{1}{2\sigma_a^2}}+o(1)\big]\\ \frac{\sum_{i=1}^n(1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}}\big[\sqrt{\frac{1}{2\sigma_a^2}}+o(1)\big]\end{bmatrix}\overset{D}{\to}\mathcal{N}\bigg(\mathbf{0},\frac{1}{2\sigma_a^2}\underline{\mathbf{I}}_2\bigg)$$

Thus, we have that, $B_n\overset{D}{\to}\mathcal{N}\big(0,\frac{1}{\sigma_a^2}\big)$. Since we assume that the algorithm's variance is correctly specified, so $\sigma_a^2=1$,

$$B_n+C_n\overset{D}{\to}\begin{cases}\infty & \text{if }\Delta>0\\ -\infty & \text{if }\Delta<0\\ \mathcal{N}(0,1) & \text{if }\Delta=0\end{cases}$$

Thus, by continuous mapping theorem,

$$\mathbb{P}\big(\tilde{\beta}_1>\tilde{\beta}_0\big|H_1^{(n)}\big)=\Phi\bigg(\frac{\mu_n}{\sigma_n}\bigg)=\Phi(B_n+C_n)\overset{D}{\to}\begin{cases}1 & \text{if }\Delta>0\\ 0 & \text{if }\Delta<0\\ \text{Uniform}[0,1] & \text{if }\Delta=0\end{cases}$$

Proof of Theorem 2 (Non-uniform convergence of the OLS estimator of the treatment effect for Thompson Sampling): The normalized errors of the OLS estimator for Δ, which are asymptotically normal under i.i.d. sampling are as follows:

$$\sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\Big(\hat{\beta}_1^{\text{OLS}}-\hat{\beta}_0^{\text{OLS}}-\Delta\Big)$$

$$=\sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\bigg(\frac{\sum_{t=1}^2\sum_{i=1}^n A_{t,i}R_{t,i}}{N_{1,1}+N_{2,1}}-\frac{\sum_{t=1}^2\sum_{i=1}^n(1-A_{t,i})R_{t,i}}{N_{1,0}+N_{2,0}}-\Delta\bigg)$$

$$=\sqrt{\frac{(N_{1,1}+N_{2,1})(N_{1,0}+N_{2,0})}{2n}}\bigg((\beta_1-\beta_0)-\Delta+\frac{\sum_{t=1}^2\sum_{i=1}^n A_{t,i}\epsilon_{t,i}}{N_{1,1}+N_{2,1}}-\frac{\sum_{t=1}^2\sum_{i=1}^n(1-A_{t,i})\epsilon_{t,i}}{N_{1,0}+N_{2,0}}\bigg)$$

$$=\sqrt{\frac{N_{1,0}+N_{2,0}}{2n}}\frac{\sum_{t=1}^2\sum_{i=1}^n A_{t,i}\epsilon_{t,i}}{\sqrt{N_{1,1}+N_{2,0}}}-\sqrt{\frac{N_{1,1}+N_{2,1}}{2n}}\frac{\sum_{t=1}^2\sum_{i=1}^n(1-A_{t,i})\epsilon_{t,i}}{\sqrt{N_{1,0}+N_{2,0}}}$$

$$= [1, -1, 1, -1] \begin{bmatrix} \sqrt{\frac{N_{1,0}+N_{2,0}}{2n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}+N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2n}} \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}+N_{2,0}}} \\ \sqrt{\frac{N_{1,0}+N_{2,0}}{2n}} \frac{\sum_{i=1}^n A_{2,i}\epsilon_{2,i}}{\sqrt{N_{1,1}+N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2n}} \frac{\sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i}}{\sqrt{N_{1,0}+N_{2,0}}} \end{bmatrix} = [1, -1, 1, -1] \begin{bmatrix} \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{1,1}}{n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2(N_{1,0}+N_{2,0})}} \sqrt{\frac{N_{1,0}}{n}} \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{2,1}}{n}} \frac{\sum_{i=1}^n A_{2,i}\epsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{N_{1,1}+N_{2,1}}{2(N_{1,0}+N_{2,0})}} \sqrt{\frac{N_{2,0}}{n}} \frac{\sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} \tag{21}$$

By Theorem 3, $\left(\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}}, \frac{\sum_{i=1}^n A_{2,i}\epsilon_{2,i}}{\sqrt{N_{2,1}}}, \frac{\sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i}}{\sqrt{N_{2,0}}} \right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \underline{\mathbf{I}}_4)$. By Lemma 1 and Slutsky's Theorem, $\sqrt{\frac{2n(N_{1,1}+N_{2,1})}{N_{1,1}(N_{1,0}+N_{2,0})}} \sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}} = 1 + o_p(1)$, thus,

$$\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{1,1}}{n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}$$

$$= \left(\sqrt{\frac{2n(N_{1,1}+N_{2,1})}{N_{1,1}(N_{1,0}+N_{2,0})}} \sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}} + o_p(1) \right) \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{1,1}}{n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}$$

$$= \sqrt{\frac{\frac{1}{2}(\frac{1}{2}+[1-\pi_2])}{2(\frac{1}{2}+\pi_2)}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{1,1}}{n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}}$$

Note that $\sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}}$ is stochastically bounded because for any $K > 2$,

$$\mathbb{P}\left(\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})} > K \right) \leq \mathbb{P}\left(\frac{n}{N_{1,1}} > K \right) = \mathbb{P}\left(\frac{1}{K} > \frac{N_{1,1}}{n} \right) \to 0$$

where the limit holds by the law of large numbers since $N_{1,1}^{(n)} \sim \text{Binomial}(n, \frac{1}{2})$. Thus, since $\frac{N_{1,1}}{n} \leq 1$ and $\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} \xrightarrow{D} \mathcal{N}(0,1)$,

$$o_p(1) \sqrt{\frac{N_{1,0}+N_{2,0}}{2(N_{1,1}+N_{2,1})}} \sqrt{\frac{N_{1,1}}{n}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} = o_p(1)$$

We can perform the above procedure on the other three terms. Thus, equation (21) is equal to the following:

$$[1, -1, 1, -1] \begin{bmatrix} \sqrt{\frac{1/2+1-\pi_2}{4(1/2+\pi_2)}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{1/2+\pi_2}{4(1/2+1-\pi_2)}} \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{(1/2+1-\pi_2)\pi_2}{2(1/2+\pi_2)}} \frac{\sum_{i=1}^n A_{2,i}\epsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{(1/2+\pi_2)(1-\pi_2)}{2(1/2+1-\pi_2)}} \frac{\sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} + o_p(1)$$

Recall that we showed earlier in Proposition 1 that

$$\pi_2^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{\mu_n}{\sigma_n} \right) \right] = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(B_n + C_n \right) \right]$$

$$= \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{2N_{1,1}}} - \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{2N_{1,0}}} + \sqrt{n}\Delta\left[\frac{1}{2} + o(1) \right] + o(1) \right) \right]$$

When $\Delta > 0$, $\pi_2^{(n)} \xrightarrow{P} \pi_{\max}$ and when $\Delta < 0$, $\pi_2^{(n)} \xrightarrow{P} \pi_{\min}$. We now consider the $\Delta = 0$ case.

$$\pi_2^{(n)} = \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{1}{\sqrt{2}} \left[\frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} - \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}} \right] + o(1) \right) \right]$$

$$= \pi_{\min} \vee \left[\pi_{\max} \wedge \Phi\left(\frac{1}{\sqrt{2}} \left[\frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} - \frac{\sum_{i=1}^n (1 - A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \right] \right) \right] + o(1)$$

By Slutsky's Theorem, for $Z_1, Z_2, Z_3, Z_4 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

$$[1, -1, 1, -1] \begin{bmatrix} \sqrt{\frac{1/2+1-\pi_2}{4(1/2+\pi_2)}} \frac{\sum_{i=1}^n A_{1,i}\epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{1/2+\pi_2}{4(1/2+1-\pi_2)}} \frac{\sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i}}{\sqrt{N_{1,0}}} \\ \sqrt{\frac{(1/2+1-\pi_2)\pi_2}{2(1/2+\pi_2)}} \frac{\sum_{i=1}^n A_{2,i}\epsilon_{2,i}}{\sqrt{N_{2,1}}} \\ \sqrt{\frac{(1/2+\pi_2)(1-\pi_2)}{2(1/2+1-\pi_2)}} \frac{\sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i}}{\sqrt{N_{2,0}}} \end{bmatrix} + o_p(1) \xrightarrow{D} [1, -1, 1, -1] \begin{bmatrix} \sqrt{\frac{1/2+1-\pi_*}{4(1/2+\pi_*)}} Z_1 \\ \sqrt{\frac{1/2+\pi_*}{4(1/2+1-\pi_*)}} Z_2 \\ \sqrt{\frac{(1/2+1-\pi_*)\pi_*}{2(1/2+\pi_*)}} Z_3 \\ \sqrt{\frac{(1/2+\pi_*)(1-\pi_*)}{2(1/2+1-\pi_*)}} Z_4 \end{bmatrix}$$

$$= \sqrt{\frac{1/2 + 1 - \pi_*}{2(1/2 + \pi_*)}} \left(\sqrt{1/2} Z_1 + \sqrt{\pi_*} Z_3 \right) - \sqrt{\frac{1/2 + \pi_*}{2(1/2 + 1 - \pi_*)}} \left(\sqrt{1/2} Z_2 + \sqrt{1 - \pi_*} Z_4 \right)$$

where $\pi_* = \begin{cases} \pi_{\max} & \text{if } \Delta > 0 \\ \pi_{\min} & \text{if } \Delta < 0 \\ \pi_{\min} \vee (\pi_{\max} \wedge \Phi[\sqrt{1/2}(Z_1 - Z_2)]) & \text{if } \Delta = 0 \quad \square \end{cases}$

Proposition 2 (Non-concentration of the sampling probabilities under zero treatment effect for ϵ-greedy)**.** *Let $T = 2$ and $\pi_1^{(n)} = \frac{1}{2}$ for all n. We assume that $\{\epsilon_{t,i}^{(n)}\}_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and*

$$\pi_2^{(n)} = \begin{cases} 1 - \frac{\epsilon}{2} & \text{if } \frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,0}} \\ \frac{\epsilon}{2} & \text{otherwise} \end{cases}$$

Thus, the sampling probability $\pi_2^{(n)}$ does not concentrate when $\beta_1 = \beta_0$.

Proof: We define $M_n := \mathbb{I}_{\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,0}} \right)} = \mathbb{I}_{\left((\beta_1 - \beta_0) + \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{N_{1,0}} \right)}$. Note that when $M_n = 1$, $\pi_2^{(n)} = 1 - \frac{\epsilon}{2}$ and when $M_n = 0$, $\pi_2^{(n)} = \frac{\epsilon}{2}$.

When the margin is zero, M_n does not concentrate because for all $N_{1,1}, N_{1,0}$, since $\epsilon_{1,i} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$,

$$\mathbb{P}\left(\frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{N_{1,0}} \right) = \mathbb{P}\left(\frac{1}{\sqrt{N_{1,1}}} Z_1 - \frac{1}{\sqrt{N_{1,0}}} Z_2 > 0 \right) = \frac{1}{2}$$

for $Z_1, Z_2 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Thus, we have shown that $\pi_2^{(n)}$ does not concentrate when $\beta_1 - \beta_0 = 0$. □

Theorem 6 (Non-uniform convergence of the OLS estimator of the treatment effect for ϵ-greedy)**.** *Assuming the setup and conditions of Proposition 2, and that $\beta_1 = b$, we show that the normalized errors of the OLS estimator converges in distribution as follows:*

$$\sqrt{N_{1,1} + N_{2,1}} \big(\hat{\beta}_1^{\text{OLS}} - b \big) \xrightarrow{D} Y$$

$$Y = \begin{cases} Z_1 & \text{if } \beta_1 - \beta_0 \neq 0 \\ \sqrt{\frac{1}{3-\epsilon}} (Z_1 - \sqrt{2-\epsilon} Z_3) \mathbb{I}_{(Z_1 > Z_2)} + \sqrt{\frac{1}{1+\epsilon}} (Z_1 - \sqrt{\epsilon} Z_3) \mathbb{I}_{(Z_1 < Z_2)} & \text{if } \beta_1 - \beta_0 = 0 \end{cases}$$

for $Z_1, Z_2, Z_3 \overset{i.i.d.}{\sim} N(0,1)$. Note the $\beta_1 - \beta_0 = 0$ case, Y is non-normal.

Proof: The normalized errors of the OLS estimator for β_1 are

$$\sqrt{N_{1,1} + N_{2,1}} \left(\frac{\sum_{t=1}^2 \sum_{i=1}^n A_{t,i} R_{t,i}}{N_{1,1} + N_{2,1}} - b \right) = \frac{\sum_{t=1}^2 \sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{1,1} + N_{2,1}}}$$

$$= [1,1] \begin{bmatrix} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1} + N_{2,1}}} \\ \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{1,1} + N_{2,1}}} \end{bmatrix} = [1,1] \begin{bmatrix} \sqrt{\frac{N_{1,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{N_{2,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} \end{bmatrix}$$

By Slutsky's Theorem and Lemma 1, $\left(\sqrt{\frac{1/2}{1/2 + \pi_2^{(n)}}} \sqrt{\frac{N_{1,1} + N_{2,1}}{N_{1,1}}}, \sqrt{\frac{\pi_2^{(n)}}{1/2 + \pi_2^{(n)}}} \sqrt{\frac{N_{1,1} + N_{2,1}}{N_{2,1}}} \right) \xrightarrow{P} (1,1)$, so

$$= [1,1] \begin{bmatrix} \left(\sqrt{\frac{1/2}{1/2 + \pi_2^{(n)}}} \sqrt{\frac{N_{1,1} + N_{2,1}}{N_{1,1}}} + o_p(1) \right) \sqrt{\frac{N_{1,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \left(\sqrt{\frac{\pi_2^{(n)}}{1/2 + \pi_2^{(n)}}} \sqrt{\frac{N_{1,1} + N_{2,1}}{N_{2,1}}} + o_p(1) \right) \sqrt{\frac{N_{2,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} \end{bmatrix}$$

$$= [1,1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2 + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \sqrt{\frac{N_{1,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} \\ \sqrt{\frac{\pi_2^{(n)}}{1/2 + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \sqrt{\frac{N_{2,1}}{N_{1,1} + N_{2,1}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} \end{bmatrix} = [1,1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2 + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\pi_2^{(n)}}{1/2 + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix}$$

The last equality holds because by Theorem 3, $\left(\frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} \right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \underline{\mathbf{I}}_2)$.

Let's define $M_n := \mathbb{I}_{\left(\frac{\sum_{i=1}^n A_{1,i} R_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) R_{1,i}}{N_{1,0}} \right)} = \mathbb{I}_{\left((\beta_1 - \beta_0) + \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{N_{1,0}} \right)}$. Note that when $M_n = 1$, $\pi_2^{(n)} = 1 - \frac{\epsilon}{2}$ and when $M_n = 0$, $\pi_2^{(n)} = \frac{\epsilon}{2}$.

$$M_n = \mathbb{I}_{\left((\beta_1 - \beta_0) + \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{N_{1,1}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{N_{1,0}} \right)} = \mathbb{I}_{\left(\sqrt{N_{1,0}}(\beta_1 - \beta_0) + \sqrt{\frac{N_{1,0}}{N_{1,1}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \right)}$$

$$= \mathbb{I}_{\left(\sqrt{N_{1,0}}(\beta_1 - \beta_0) + [1 + o_p(1)] \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \right)}$$

where the last equality holds because $\sqrt{\frac{N_{1,0}}{N_{1,1}}} \xrightarrow{P} 1$ by Lemma 1, Slutsky's Theorem, and continuous mapping theorem. Thus, by Proposition 2,

$$M^{(n)} \xrightarrow{P} \begin{cases} 1 & \text{if } \beta_1 - \beta_0 > 0 \\ 0 & \text{if } \beta_1 - \beta_0 < 0 \\ \text{does not concentrate} & \text{if } \beta_1 - \beta_0 = 0 \end{cases}$$

Note that

$$\begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\pi_2^{(n)}}{\frac{1}{2} + \pi_2^{(n)}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + 1 - \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{1 - \epsilon/2}{\frac{1}{2} + 1 - \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} M_n + \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\frac{\epsilon}{2}}{\frac{1}{2} + \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} (1 - M_n)$$

Also note that by Theorem 3, $\left(\frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}}, \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}}, \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}}, \frac{\sum_{i=1}^n (1-A_{2,i}) \epsilon_{2,i}}{\sqrt{N_{2,1}}} \right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \underline{\mathbf{I}}_4)$.

When $\beta_1 > \beta_0$, $M_n \xrightarrow{P} 1$ and when $\beta_1 < \beta_0$, $M_n \xrightarrow{P} 0$; in both these cases the normalized errors are asymptotically normal. We now focus on the case that $\beta_1 = \beta_0$. By continuous mapping theorem and Slutsky's theorem for $Z_1, Z_2, Z_3, Z_4 \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$,

$$= [1, 1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + 1 - \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{1 - \epsilon/2}{\frac{1}{2} + 1 - \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} \mathbb{I}_{\left([1 + o(1)] \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \right)}$$

$$+ [1, 1] \begin{bmatrix} \sqrt{\frac{\frac{1}{2}}{\frac{1}{2} + \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} + o_p(1) \\ \sqrt{\frac{\frac{\epsilon}{2}}{\frac{1}{2} + \frac{\epsilon}{2}}} \frac{\sum_{i=1}^n A_{2,i} \epsilon_{2,i}}{\sqrt{N_{2,1}}} + o_p(1) \end{bmatrix} \left(1 - \mathbb{I}_{\left([1 + o(1)] \frac{\sum_{i=1}^n A_{1,i} \epsilon_{1,i}}{\sqrt{N_{1,1}}} > \frac{\sum_{i=1}^n (1-A_{1,i}) \epsilon_{1,i}}{\sqrt{N_{1,0}}} \right)} \right)$$

$$\xrightarrow{D} [1, 1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2 + 1 - \epsilon/2}} Z_1 \\ \sqrt{\frac{1 - \epsilon/2}{1/2 + 1 - \epsilon/2}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1 > Z_2)} + [1, 1] \begin{bmatrix} \sqrt{\frac{1/2}{1/2 + \epsilon/2}} Z_1 \\ \sqrt{\frac{\epsilon/2}{1/2 + \epsilon/2}} Z_3 \end{bmatrix} \mathbb{I}_{(Z_1 < Z_2)}$$

$$= \left(\sqrt{\frac{1}{3 - \epsilon}} Z_1 + \sqrt{\frac{2 - \epsilon}{3 - \epsilon}} Z_3 \right) \mathbb{I}_{(Z_1 > Z_2)} + \left(\sqrt{\frac{1}{1 + \epsilon}} Z_1 + \sqrt{\frac{\epsilon}{1 + \epsilon}} Z_3 \right) \mathbb{I}_{(Z_1 < Z_2)}$$

Thus,

$$\frac{\sum_{t=1}^2 \sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{1,1} + N_{2,1}}} \xrightarrow{D} Y$$

$$Y := \begin{cases} \sqrt{\frac{1}{3-\epsilon}} \big(Z_1 - \sqrt{2 - \epsilon} Z_3 \big) & \text{if } \beta_1 - \beta_0 > 0 \\ \sqrt{\frac{1}{1+\epsilon}} \big(Z_1 - \sqrt{\epsilon} Z_3 \big) & \text{if } \beta_1 - \beta_0 < 0 \\ \sqrt{\frac{1}{3-\epsilon}} \big(Z_1 - \sqrt{2 - \epsilon} Z_3 \big) \mathbb{I}_{(Z_1 > Z_2)} + \sqrt{\frac{1}{1+\epsilon}} \big(Z_1 - \sqrt{\epsilon} Z_3 \big) \mathbb{I}_{(Z_1 < Z_2)} & \text{if } \beta_1 - \beta_0 = 0 \ \square \end{cases}$$

D. Asymptotic Normality of the Batched OLS Estimator: Multi-Arm Bandits

Theorem 3 (Asymptotic normality of Batched OLS estimator for multi-arm bandits) *Assuming Conditions 6 (weak moments) and 3 (conditionally i.i.d. actions), and a clipping rate of $f(n) = \omega(\frac{1}{n})$ (Definition 1),*

$$\begin{bmatrix} \begin{bmatrix} N_{1,0} & 0 \\ 0 & N_{1,1} \end{bmatrix}^{1/2} (\hat{\boldsymbol{\beta}}_1^{\text{BOLS}} - \boldsymbol{\beta}_1) \\ \begin{bmatrix} N_{2,0} & 0 \\ 0 & N_{2,1} \end{bmatrix}^{1/2} (\hat{\boldsymbol{\beta}}_2^{\text{BOLS}} - \boldsymbol{\beta}_2) \\ \vdots \\ \begin{bmatrix} N_{T,0} & 0 \\ 0 & N_{T,1} \end{bmatrix}^{1/2} (\hat{\boldsymbol{\beta}}_T^{\text{BOLS}} - \boldsymbol{\beta}_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{2T})$$

where $\boldsymbol{\beta}_t = (\beta_{t,0}, \beta_{t,1})$. Note in the body of this paper, we state Theorem 3 with conditions that are are sufficient for the weaker conditions we use here.

Lemma 1. *Assuming the conditions of Theorem 3, for any batch $t \in [1\colon T]$,*

$$\frac{N_{t,1}^{(n)}}{n\pi_t^{(n)}} = \frac{\sum_{i=1}^n A_{t,i}^{(n)}}{n\pi_t^{(n)}} \xrightarrow{P} 1 \quad \text{and} \quad \frac{N_{t,0}^{(n)}}{n(1-\pi_t^{(n)})} = \frac{\sum_{i=1}^n (1 - A_{t,i}^{(n)})}{n(1-\pi_t^{(n)})} \xrightarrow{P} 1$$

Proof of Lemma 1: To prove that $\frac{N_{t,1}}{n\pi_t^{(n)}} \xrightarrow{P} 1$, it is equivalent to show that $\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \xrightarrow{P} 0$. Let $\epsilon > 0$.

$$\mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right| > \epsilon \right) = \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right| \left[\mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} + \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])} \right] > \epsilon \right)$$

$$\leq \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} > \frac{\epsilon}{2} \right) + \mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \notin [f(n), 1-f(n)])} > \frac{\epsilon}{2} \right)$$

Since by our clipping assumption, $\mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \xrightarrow{P} 1$, the second probability in the summation above converges to 0 as $n \to \infty$. We will now show that the first probability in the summation above also goes to zero. Note that $\mathbb{E}\big[\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)})\big] = \mathbb{E}\big[\frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (\mathbb{E}[A_{t,i} | H_{t-1}^{(n)}] - \pi_t^{(n)})\big] = 0$. So by Chebychev inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n\pi_t^{(n)}} \sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right| \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} > \epsilon \right) \leq \frac{1}{\epsilon^2 n^2} \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \left(\sum_{i=1}^n (A_{t,i} - \pi_t^{(n)}) \right)^2 \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$\leq \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} (A_{t,i} - \pi_t^{(n)})(A_{t,j} - \pi_t^{(n)}) \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \mathbb{E}\big[A_{t,i} A_{t,j} - \pi_t^{(n)} (A_{t,i} + A_{t,j}) + (\pi_t^{(n)})^2 \big| H_{t-1}^{(n)} \big] \right]$$

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \left(\mathbb{E}\big[A_{t,i} A_{t,j} \big| H_{t-1}^{(n)} \big] - (\pi_t^{(n)})^2 \right) \right] \tag{22}$$

Note that if $i \neq j$, since $A_{t,i} \overset{i.i.d.}{\sim}$ Bernoulli$(\pi_t^{(n)})$, $\mathbb{E}[A_{t,i} A_{t,j} | H_{t-1}^{(n)}] = \mathbb{E}[A_{t,i} | H_{t-1}^{(n)}] E[A_{t,j} | H_{t-1}^{(n)}] = (\pi_t^{(n)})^2$, so (22) above equals the following

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \mathbb{E}\left[\frac{1}{(\pi_t^{(n)})^2} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \left(\mathbb{E}\big[A_{t,i} \big| H_{t-1}^{(n)} \big] - (\pi_t^{(n)})^2 \right) \right] = \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \mathbb{E}\left[\frac{1 - \pi_t^{(n)}}{\pi_t^{(n)}} \mathbb{I}_{(\pi_t^{(n)} \in [f(n), 1-f(n)])} \right]$$

$$= \frac{1}{\epsilon^2 n}\mathbb{E}\left[\frac{1-\pi_t^{(n)}}{\pi_t^{(n)}}\mathbb{I}_{(\pi_t^{(n)}\in[f(n),1-f(n)])}\right] \leq \frac{1}{\epsilon^2 n}\frac{1}{f(n)} \to 0$$

where the limit holds because we assume $f(n) = \omega(\frac{1}{n})$ so $f(n)n \to \infty$. We can make a very similar argument for $\frac{N_{t,0}}{n(1-\pi_t^{(n)})} \xrightarrow{P} 1$. $\square$

Proof for Theorem 3 (Asymptotic normality of Batched OLS estimator for multi-arm bandits): For readability, for this proof we drop the (n) superscript on $\pi_t^{(n)}$. Note that

$$\begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{1/2} (\hat{\boldsymbol{\beta}}_t^{\text{BOLS}} - \boldsymbol{\beta}_t) = \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{1/2} \begin{bmatrix} \frac{\sum_{i=1}^n (1-A_{t,i})\epsilon_{t,i}}{N_{t,0}} \\ \\ \frac{\sum_{i=1}^n A_{t,i}\epsilon_{t,i}}{N_{t,1}} \end{bmatrix} = \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i}.$$

We want to show that

$$\begin{bmatrix} \begin{bmatrix} N_{0,1} & 0 \\ 0 & N_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \epsilon_{1,i} \\ \begin{bmatrix} N_{0,2} & 0 \\ 0 & N_{1,2} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \epsilon_{2,i} \\ \vdots \\ \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \epsilon_{T,i} \end{bmatrix} = \begin{bmatrix} N_{0,1}^{-1/2} \sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i} \\ N_{1,1}^{-1/2} \sum_{i=1}^n A_{1,i}\epsilon_{1,i} \\ N_{0,2}^{-1/2} \sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i} \\ N_{1,2}^{-1/2} \sum_{i=1}^n A_{2,i}\epsilon_{2,i} \\ \vdots \\ N_{t,0}^{-1/2} \sum_{i=1}^n (1-A_{T,i})\epsilon_{T,i} \\ N_{t,1}^{-1/2} \sum_{i=1}^n A_{T,i}\epsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_{2T}).$$

By Lemma 1 and Slutsky's Theorem it is sufficient to show that as $n \to \infty$,

$$\begin{bmatrix} \frac{1}{\sqrt{n(1-\pi_1)}} \sum_{i=1}^n (1-A_{1,i})\epsilon_{1,i} \\ \frac{1}{\sqrt{n\pi_1}} \sum_{i=1}^n A_{1,i}\epsilon_{1,i} \\ \frac{1}{\sqrt{n(1-\pi_2)}} \sum_{i=1}^n (1-A_{2,i})\epsilon_{2,i} \\ \frac{1}{\sqrt{n\pi_2}} \sum_{i=1}^n A_{2,i}\epsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n(1-\pi_T)}} \sum_{i=1}^n (1-A_{T,i})\epsilon_{T,i} \\ \frac{1}{\sqrt{n\pi_T}} \sum_{i=1}^n A_{T,i}\epsilon_{T,i} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_{1,1} & 0 \\ 0 & \pi_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \epsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_2 & 0 \\ 0 & \pi_2 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \epsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \epsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_{2T})$$

By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{c} \in \mathbb{R}^{2T}$ s.t. $\|\mathbf{c}\|_2 = 1$ that as $n \to \infty$,

$$\mathbf{c}^\top \begin{bmatrix} n^{-1/2} \begin{bmatrix} 1-\pi_{1,1} & 0 \\ 0 & \pi_{1,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \epsilon_{1,i} \\ n^{-1/2} \begin{bmatrix} 1-\pi_2 & 0 \\ 0 & \pi_2 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \epsilon_{2,i} \\ \vdots \\ n^{-1/2} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \epsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let us break up $\mathbf{c}$ so that $\mathbf{c} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_T]^\top \in \mathbb{R}^{2T}$ with $\mathbf{c}_t \in \mathbb{R}^2$ for $t \in [1\colon T]$. The above is equivalent to

$$\sum_{t=1}^T n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let us define $Y_{t,i}^{(n)} := n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_{t,i} & 0 \\ 0 & \pi_{t,i} \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i}$.

The sequence $\{Y_{1,1}^{(n)}, Y_{1,2}^{(n)}, ..., Y_{1,n}^{(n)}, ..., Y_{T,1}^{(n)}, Y_{T,2}^{(n)}, ..., Y_{T,n}^{(n)}\}$ is a martingale with respect to sequence of histories

$\{H_t^{(n)}\}_{t=1}^T$, since

$$\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}] = n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \middle| H_{t-1}^{(n)} \right]$$

$$= n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} (1-\pi_t) E[\epsilon_{t,i} | H_{t-1}^{(n)}, A_{t,i}=0] \\ \pi_{t,i} E[\epsilon_{t,i} | H_{t-1}^{(n)}, A_{t,i}=1] \end{bmatrix} \middle| H_{t-1}^{(n)} \right] = 0$$

for all $i \in [1\colon n]$ and all $t \in [1\colon T]$. We then apply Dvoretzky (1972) martingale central limit theorem to $Y_{t,i}^{(n)}$ to show the desired result (see the proof of Theorem 5 in Appendix B for the statement of the martingale CLT conditions).

Condition(a): Martingale Condition The first condition holds because $\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}] = 0$ for all $i \in [1\colon n]$ and all $t \in [1\colon T]$.

Condition(b): Conditional Variance

$$\sum_{t=1}^T \sum_{i=1}^n E[Y_{n,t,i}^2 | H_{t-1}^{(n)}] = \sum_{t=1}^T \sum_{i=1}^n \mathbb{E}\left[\left(\frac{1}{\sqrt{n}} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \right)^2 \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^T \sum_{i=1}^n \mathbb{E}\left[\frac{1}{n} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} & 0 \\ 0 & A_{t,i} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbf{c}_t \epsilon_{t,i}^2 \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^T \sum_{i=1}^n \frac{1}{n} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} \mathbb{E}[(1-A_{t,i})\epsilon_{t,i}^2 | H_{t-1}^{(n)}] & 0 \\ 0 & \mathbb{E}[A_{t,i}\epsilon_{t,i}^2 | H_{t-1}^{(n)}] \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbf{c}_t$$

Since $\mathbb{E}[A_{t,i}\epsilon_{t,i}^2 | H_{t-1}^{(n)}] = \pi_t \mathbb{E}[\epsilon_{t,i}^2 | H_{t-1}^{(n)}, A_{t,i}=1] = \sigma^2 \pi_t$ and $\mathbb{E}[(1-A_{t,i})\epsilon_{t,i}^2 | H_{t-1}^{(n)}] = (1-\pi_t)\mathbb{E}[\epsilon_{t,i}^2 | H_{t-1}^{(n)}, A_{t,i}=0] = \sigma^2(1-\pi_t)$,

$$= \sum_{t=1}^T \sum_{i=1}^n n^{-1} \mathbf{c}_t^\top \mathbf{c}_t \sigma^2 = \sum_{t=1}^T \mathbf{c}_t^\top \mathbf{c}_t \sigma^2 = \sigma^2$$

Condition(c): Lindeberg Condition Let $\delta > 0$.

$$\sum_{t=1}^T \sum_{i=1}^n E\big[Y_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \big| H_{t-1}^{(n)}\big] = \sum_{t=1}^T \sum_{i=1}^n \mathbb{E}\left[\left(n^{-1/2} \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \right)^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} & 0 \\ 0 & A_{t,i} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbf{c}_t \epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)} \right]$$

$$= \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n \mathbf{c}_t^\top \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-\frac{1}{2}} \begin{bmatrix} \mathbb{E}[(1-A_{t,i})\epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} | H_{t-1}^{(n)}] & 0 \\ 0 & \mathbb{E}[A_{t,i}\epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} | H_{t-1}^{(n)}] \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-\frac{1}{2}} \mathbf{c}_t$$

Note that for $\mathbf{c}_t = [c_{t,0}, c_{t,1}]^\top$, $\mathbb{E}\big[(1-A_{t,i})\epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \big| H_{t-1}^{(n)}\big] = \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\frac{c_{t,0}^2}{1-\pi_t} \epsilon_{t,i}^2 > n\delta^2 \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=0 \right](1-\pi_t)$ and $\mathbb{E}\big[A_{t,i}\epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \big| H_{t-1}^{(n)}\big] = \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\frac{c_{t,1}^2}{\pi_t} \epsilon_{t,i}^2 > n\delta^2 \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=1 \right]\pi_t$. Thus, we have that

$$= \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n c_{t,0}^2 \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2(1-\pi_t)}{c_{t,0}^2} \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=0 \right] + c_{t,1}^2 \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 \pi_t}{c_{t,1}^2} \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=1 \right]$$

$$\leq \sum_{t=1}^T \max_{i \in [1\colon n]} \left\{ c_{t,0}^2 \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2(1-\pi_t)}{c_{t,0}^2} \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=0 \right] + c_{t,1}^2 \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 \pi_t}{c_{t,1}^2} \right)} \middle| H_{t-1}^{(n)}, A_{t,i}=1 \right] \right\}$$

Note that for any $t \in [1\colon T]$ and $i \in [1\colon n]$,

$$\mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 \pi_t}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1\right] = \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 \pi_t}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1\right] \left(\mathbb{I}_{(\pi_t \in [f(n), 1-f(n)])} + \mathbb{I}_{(\pi_t \notin [f(n), 1-f(n)])}\right)$$

$$\leq \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 f(n)}{c_{t,1}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1\right] + \sigma^2 \mathbb{I}_{(\pi_t \notin [f(n), 1-f(n)])}$$

The second term converges in probability to zero as $n \to \infty$ by our clipping assumption. We now show how the first term goes to zero in probability. Since we assume $f(n) = \omega(\frac{1}{n})$, $nf(n) \to \infty$. So, it is sufficient to show that for all t, n,

$$\lim_{m \to \infty} \max_{i \in [1\colon n]} \left\{ \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 > m)} \middle| H_{t-1}^{(n)}, A_{t,i} = 1\right] \right\} = 0$$

By Condition 6, we have that for all $n \geq 1$,

$$\max_{t \in [1\colon T], i \in [1\colon n]} \mathbb{E}[\varphi(\epsilon_{t,i}^2) | H_{t-1}^{(n)}, A_{t,i} = 1] < M$$

Since we assume that $\lim_{x \to \infty} \frac{\varphi(x)}{x} = \infty$, for all m, there exists a b_m s.t. $\varphi(x) \geq mMx$ for all $x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) | H_{t-1}^{(n)}, A_{t,i} = 1] \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = 1] \geq mM \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = 1]$$

Thus,

$$\max_{t \in [1\colon T], i \in [1\colon n]} \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = 1] \leq \frac{1}{m}$$

We can make a very similar argument that for all $t \in [1\colon T]$, as $n \to \infty$,

$$\max_{i \in [1\colon n]} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{\left(\epsilon_{t,i}^2 > \frac{n\delta^2 (1-\pi_t)}{c_{t,0}^2}\right)} \middle| H_{t-1}^{(n)}, A_{t,i} = 0\right] \xrightarrow{P} 0 \qquad \square$$

Corollary 3 (Asymptotic Normality of the Batched OLS Estimator of Margin; two-arm bandit setting)**.** *Assume the same conditions as Theorem 3. For each* $t \in [1\colon T]$, *we have the BOLS estimator of the margin* $\beta_1 - \beta_0$*:*

$$\hat{\Delta}_t^{\text{BOLS}} = \frac{\sum_{i=1}^n (1 - A_{t,i}) R_{t,i}}{N_{t,0}} - \frac{\sum_{i=1}^n A_{t,i} R_{t,i}}{N_{t,1}}$$

We show that as $n \to \infty$*,*

$$\begin{bmatrix} \sqrt{\frac{N_{1,0}N_{1,1}}{n}} (\hat{\Delta}_1^{\text{BOLS}} - \Delta_1) \\ \sqrt{\frac{N_{2,0}N_{2,1}}{n}} (\hat{\Delta}_2^{\text{BOLS}} - \Delta_2) \\ \vdots \\ \sqrt{\frac{N_{T,0}N_{T,1}}{n}} (\hat{\Delta}_T^{\text{BOLS}} - \Delta_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_T)$$

Proof:

$$\sqrt{\frac{N_{t,0}N_{t,1}}{n}} (\hat{\Delta}_t^{\text{BOLS}} - \Delta_t) = \sqrt{\frac{N_{t,0}N_{t,1}}{n}} \left(\frac{\sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{N_{t,0}} - \frac{\sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{N_{t,1}} \right)$$

$$= \sqrt{\frac{N_{t,1}}{n}} \frac{\sum_{i=1}^n (1 - A_{t,i}) \epsilon_{t,i}}{\sqrt{N_{t,0}}} - \sqrt{\frac{N_{t,0}}{n}} \frac{\sum_{i=1}^n A_{t,i} \epsilon_{t,i}}{\sqrt{N_{t,1}}} = \begin{bmatrix} \sqrt{\frac{N_{t,1}}{n}} & -\sqrt{\frac{N_{t,0}}{n}} \end{bmatrix} \begin{bmatrix} N_{t,0} & 0 \\ 0 & N_{t,1} \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1 - A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i}$$

By Slutsky's Theorem and Lemma 1, it is sufficient to show that as $n \to \infty$,

$$\begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_1} & -\sqrt{1 - \pi_1} \end{bmatrix} \begin{bmatrix} 1 - \pi_1 & 0 \\ 0 & \pi_1 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1 - A_{1,i} \\ A_{1,i} \end{bmatrix} \epsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_2} & -\sqrt{1 - \pi_2} \end{bmatrix} \begin{bmatrix} 1 - \pi_2 & 0 \\ 0 & \pi_2 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1 - A_{2,i} \\ A_{2,i} \end{bmatrix} \epsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1 - \pi_t} \end{bmatrix} \begin{bmatrix} 1 - \pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1 - A_{T,i} \\ A_{T,i} \end{bmatrix} \epsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\mathbf{I}}_T)$$

By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{d} \in \mathbb{R}^T$ s.t. $\|\mathbf{d}\|_2 = 1$ that

$$\mathbf{d}^\top \begin{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_1} & -\sqrt{1-\pi_1} \end{bmatrix} \begin{bmatrix} 1-\pi_1 & 0 \\ 0 & \pi_1 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{1,i} \\ A_{1,i} \end{bmatrix} \epsilon_{1,i} \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_2} & -\sqrt{1-\pi_2} \end{bmatrix} \begin{bmatrix} 1-\pi_2 & 0 \\ 0 & \pi_2 \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{2,i} \\ A_{2,i} \end{bmatrix} \epsilon_{2,i} \\ \vdots \\ \frac{1}{\sqrt{n}} \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1-\pi_t} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{T,i} \\ A_{T,i} \end{bmatrix} \epsilon_{T,i} \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Let $[d_1, d_2, ..., d_T]^\top := \mathbf{d} \in \mathbb{R}^T$. The above is equivalent to

$$\sum_{t=1}^T \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1-\pi_t} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \sum_{i=1}^n \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Define $Y_{t,i}^{(n)} := \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1-\pi_t} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i}$.
$\{Y_{1,1}^{(n)}, Y_{1,2}^{(n)}, ..., Y_{1,n}^{(n)}, ..., Y_{T,1}^{(n)}, Y_{T,2}^{(n)}, ..., Y_{T,n}^{(n)}\}$ is a martingale difference array with respect to the sequence of histories $\{H_t^{(n)}\}_{t=1}^T$ because for all $i \in [1\colon n]$ and $t \in [1\colon T]$,

$$\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}] = \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1-\pi_t} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} 1-A_{t,i} \\ A_{t,i} \end{bmatrix} \epsilon_{t,i} \middle| H_{t-1}^{(n)} \right]$$

$$= \frac{1}{\sqrt{n}} d_t \begin{bmatrix} \sqrt{\pi_t} & -\sqrt{1-\pi_t} \end{bmatrix} \begin{bmatrix} 1-\pi_t & 0 \\ 0 & \pi_t \end{bmatrix}^{-1/2} \mathbb{E}\left[\begin{bmatrix} (1-\pi_t) E[\epsilon_{t,i} | H_{t-1}^{(n)}, A_{t,i} = 0] \\ \pi_{t,i} E[\epsilon_{t,i} | H_{t-1}^{(n)}, A_{t,i} = 1] \end{bmatrix} \middle| H_{t-1}^{(n)} \right] = 0$$

We now apply Dvoretzky (1972) martingale central limit theorem to $Y_{t,i}^{(n)}$ to show the desired result. Verifying the conditions for the martingale CLT is equivalent to what we did to verify the conditions in the conditions in the proof of Theorem 3—the only difference is that we replace $\mathbf{c}_t^\top$ in the Theorem 3 proof with $d_t \begin{bmatrix} \sqrt{1-\pi_t} & -\sqrt{\pi_t} \end{bmatrix}$ in this proof. Even though $\mathbf{c}_t$ is a constant vector and $d_t \begin{bmatrix} \sqrt{1-\pi_t} & -\sqrt{\pi_t} \end{bmatrix}$ is a random vector, the proof still goes through with this adjusted $\mathbf{c}_t$ vector, since (i) $d_t \begin{bmatrix} \sqrt{1-\pi_t} & -\sqrt{\pi_t} \end{bmatrix} \in H_{t-1}^{(n)}$, (ii) $\| \begin{bmatrix} \sqrt{1-\pi_t} & -\sqrt{\pi_t} \end{bmatrix} \|_2 = 1$, and (iii) $\frac{n\delta^2 \pi_t}{c_{t,1}^2} = \frac{n\delta^2 \pi_t}{d_t^2 \pi_t} \to \infty$ and $\frac{n\delta^2 (1-\pi_t)}{c_{t,0}^2} = \frac{n\delta^2 (1-\pi_t)}{d_t^2 (1-\pi_t)} \to \infty$. □

Proof of Corollary 2 (Confidence interval for treatment effect for non-stationary bandits)
Note that by Corollary 3,

$$\mathbb{P}\big(\text{exists some } t \in [1:T] \; s.t. \; \Delta_t \notin \mathbf{L}_t\big) \leq \sum_{t=1}^T \mathbb{P}\big(\Delta_t \notin \mathbf{L}_t\big) \to \sum_{t=1}^T \frac{\alpha}{T} = \alpha$$

where the limit is as $n \to \infty$. Since

$$\mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \mathbf{L}_t\big) = 1 - \mathbb{P}\big(\text{exists some } t \in [1:T] \; s.t. \; \Delta_t \notin \mathbf{L}_t\big)$$

Thus,

$$\lim_{n \to \infty} \mathbb{P}\big(\forall t \in [1:T], \Delta_t \in \mathbf{L}_t\big) \geq 1 - \alpha \qquad \square$$

E. Asymptotic Normality of the Batched OLS Estimator: Contextual Bandits

Theorem 4 (Asymptotic Normality of the Batched OLS Statistic) *For a K-armed contextual bandit, we for each $t \in [1\colon T]$, we have the BOLS estimator:*

$$\hat{\boldsymbol{\beta}}_t^{\text{BOLS}} = \begin{bmatrix} \underline{\boldsymbol{C}}_{t,0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \underline{\boldsymbol{C}}_{t,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underline{\boldsymbol{C}}_{t,2} & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \underline{\boldsymbol{C}}_{t,K-1} \end{bmatrix}^{-1} \sum_{i=1}^n \begin{bmatrix} \mathbb{I}_{A_{t,i}=0}\boldsymbol{C}_{t,i}^{(n)} \\ \mathbb{I}_{A_{t,i}=1}\boldsymbol{C}_{t,i}^{(n)} \\ \vdots \\ \mathbb{I}_{A_{t,i}=K-1}\boldsymbol{C}_{t,i}^{(n)} \end{bmatrix} R_{t,i} \in \mathbb{R}^{Kd}$$

where $\underline{\boldsymbol{C}}_{t,k} := \sum_{i=1}^n \mathbb{I}_{A_{t,i}^{(n)}=k} \boldsymbol{C}_{t,i}^{(n)} (\boldsymbol{C}_{t,i}^{(n)})^\top \in \mathbb{R}^{d \times d}$. Assuming Conditions 6 (weak moments), 3 (conditionally i.i.d. actions), 4 (conditionally i.i.d. contexts), and 5 (bounded contexts), and a conditional clipping rate $f(n) = c$ for some $0 \leq c < \frac{1}{2}$ (see Definition 2), we show that as $n \to \infty$,

$$\begin{bmatrix} \text{Diagonal}\big[\underline{\boldsymbol{C}}_{1,0}, \underline{\boldsymbol{C}}_{1,1}, ..., \underline{\boldsymbol{C}}_{1,K-1}\big]^{1/2} (\hat{\boldsymbol{\beta}}_1^{\text{BOLS}} - \boldsymbol{\beta}_1) \\ \text{Diagonal}\big[\underline{\boldsymbol{C}}_{2,0}, \underline{\boldsymbol{C}}_{2,1}, ..., \underline{\boldsymbol{C}}_{2,K-1}\big]^{1/2} (\hat{\boldsymbol{\beta}}_2^{\text{BOLS}} - \boldsymbol{\beta}_2) \\ \vdots \\ \text{Diagonal}\big[\underline{\boldsymbol{C}}_{T,0}, \underline{\boldsymbol{C}}_{T,1}, ..., \underline{\boldsymbol{C}}_{T,K-1}\big]^{1/2} (\hat{\boldsymbol{\beta}}_T^{\text{BOLS}} - \boldsymbol{\beta}_T) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{TKd})$$

Lemma 2. *Assuming the conditions of Theorem 4, for any batch $t \in [1\colon T]$ and any arm $k \in [0\colon K-1]$, as $n \to \infty$,*

$$\left[\sum_{i=1}^n \mathbb{I}_{A_{t,i}=k} \boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^\top\right] \left[n \underline{\boldsymbol{Z}}_{t,k} P_{t,k}\right]^{-1} \xrightarrow{P} \underline{\boldsymbol{I}}_d \tag{23}$$

$$\left[\sum_{i=1}^n \mathbb{I}_{A_{t,i}=k} \boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^\top\right]^{1/2} \left[n \underline{\boldsymbol{Z}}_{t,k} P_{t,k}\right]^{-1/2} \xrightarrow{P} \underline{\boldsymbol{I}}_d \tag{24}$$

where $P_{t,k} := \mathbb{P}(A_{t,i} = k | H_{t-1}^{(n)})$ and $\underline{\boldsymbol{Z}}_{t,k} := \mathbb{E}\big[\boldsymbol{C}_{t,i} \boldsymbol{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k\big]$.

Proof of Lemma 2: We first show that as $n \to \infty$, $\frac{1}{n} \sum_{i=1}^n \left(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top - \underline{\mathbf{Z}}_{t,k} P_{t,k}\right) \xrightarrow{P} \underline{\mathbf{0}}$. It is sufficient to show that convergence holds entry-wise so for any $r, s \in [0\colon d-1]$, as $n \to \infty$, $\frac{1}{n} \sum_{i=1}^n \left(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right) \xrightarrow{P} 0$. Note that

$$\mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right] = \mathbb{E}\left[\mathbb{E}\big[\mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) \big| H_{t-1}, A_{t,i} = k\big] P_{t,k} - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right] = 0$$

By Chebychev inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right| > \epsilon\right) \leq \frac{1}{\epsilon^2 n^2} \mathbb{E}\left[\left(\sum_{i=1}^n \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right)^2\right]$$

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left[\left(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right)\left(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,j} \mathbf{C}_{t,j}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right)\right] \tag{25}$$

By conditional independence and by law of iterated expectations (conditioning on $H_{t-1}^{(n)}$), for $i \neq j$, $\mathbb{E}\big[\big(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\big)\big(\mathbb{I}_{A_{t,j}=k} \mathbf{C}_{t,j} \mathbf{C}_{t,j}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\big)\big] = 0$. Thus, (25) above equals the following:

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \mathbb{E}\left[\left(\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top(r,s) - P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s)\right)^2\right]$$

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \mathbb{E}\Big[\mathbb{I}_{A_{t,i}=k} \big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top(r,s) \big]^2 - 2\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top(r,s) P_{t,k} \underline{\mathbf{Z}}_{t,k}(r,s) + P_{t,k}^2 \big[\underline{\mathbf{Z}}_{t,k}(r,s) \big]^2 \Big]$$

$$= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^n \mathbb{E}\Big[\mathbb{I}_{A_{t,i}=k} \big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top(r,s) \big]^2 - P_{t,k}^2 \big[\underline{\mathbf{Z}}_{t,k}(r,s) \big]^2 \Big]$$

$$= \frac{1}{\epsilon^2 n} \mathbb{E}\Big[\mathbb{I}_{A_{t,i}=k} \big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top(r,s) \big]^2 - P_{t,k}^2 \big[\underline{\mathbf{Z}}_{t,k}(r,s) \big]^2 \Big] \leq \frac{2d \max(u^2,1)}{\epsilon^2 n} \to 0$$

as $n \to \infty$. The last inequality above holds by Condition 5.

Proving Equation (23)**:**

It is sufficient to show that

$$\left\| \frac{2\max(du^2,1)}{\epsilon^2 n} \big[n \underline{\mathbf{Z}}_{t,k} P_{t,k} \big]^{-1} \right\|_{op} = \left\| \frac{2\max(du^2,1)}{\epsilon^2 n^2 P_{t,k}} \underline{\mathbf{Z}}_{t,k}^{-1} \right\|_{op} \xrightarrow{P} 0 \tag{26}$$

We define random variable $M_t^{(n)} = \mathbb{I}_{(\forall\ \mathbf{c} \in \mathbb{R}^d,\ \mathcal{A}_t(H_{t-1}^{(n)}, \mathbf{c}) \in [f(n), 1-f(n)]^K)}$, representing whether the conditional clipping condition is satisfied. Note that by our conditional clipping assumption, $M_t^{(n)} \xrightarrow{P} 1$ as $n \to \infty$. The left hand side of (26) is equal to the following

$$\left\| \frac{2\max(du^2,1)}{\epsilon^2 n^2 P_{t,k}} \underline{\mathbf{Z}}_{t,k}^{-1} (M_t^{(n)} + (1 - M_t^{(n)})) \right\|_{op} = \left\| \frac{2\max(du^2,1)}{\epsilon^2 n^2 P_{t,k}} \underline{\mathbf{Z}}_{t,k}^{-1} M_t^{(n)} \right\|_{op} + o_p(1) \tag{27}$$

By our conditional clipping condition and Bayes rule we have that for all $\mathbf{c} \in [-u, u]^d$,

$$\mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | A_{t,i} = k, H_{t-1}^{(n)}, M_t^{(n)} = 1) = \frac{\mathbb{P}(A_{t,i} = k | \mathbf{C}_{t,i} = \mathbf{c}, H_{t-1}^{(n)}, M_t^{(n)} = 1) \mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | H_{t-1}^{(n)}, M_t^{(n)} = 1)}{\mathbb{P}(A_{t,i} = k | H_{t-1}^{(n)}, M_t^{(n)} = 1)}$$

$$\geq \frac{f(n)\, \mathbb{P}(\mathbf{C}_{t,i} = \mathbf{c} | H_{t-1}^{(n)}, M_t^{(n)} = 1)}{1}.$$

Thus, we have that

$$\underline{\mathbf{Z}}_{t,k} M_t^{(n)} = \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k \big] M_t^{(n)} = \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k, M_t^{(n)} = 1 \big] M_t^{(n)}$$

$$\succcurlyeq f(n) \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, M_t^{(n)} = 1 \big] M_t^{(n)} = f(n) \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)} \big] M_t^{(n)} = f(n) \underline{\mathbf{\Sigma}}_t^{(n)} M_t^{(n)}.$$

By apply matrix inverses to both sides of the above inequality, we get that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1} M_t^{(n)}) \leq \frac{1}{f(n)} \lambda_{\max}\left(\left(\underline{\mathbf{\Sigma}}_t^{(n)} \right)^{-1} \right) M_t^{(n)} \leq \frac{1}{l\, f(n)} \tag{28}$$

where the last inequality above holds for constant l by Condition 5. Recall that $P_{t,k} = \mathbb{P}(A_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (M_t^{(n)} = 1) \geq f(n)$. Thus, equation (27) is bounded above by the following

$$\leq \frac{2\max(du^2,1)}{\epsilon^2 n^2 l f(n)^2} + o_p(1) \xrightarrow{P} 0$$

where the limit above holds because we assume that $f(n) = c$ for some $0 < c \leq \frac{1}{2}$ $\square$.

Proving Equation (24)**:** By Condition 5, $\|\frac{1}{n}\underline{\mathbf{C}}_{t,k}\|_{\max} \leq u$ and $\|\underline{\mathbf{Z}}_{t,k} P_{t,k}\|_{\max} \leq u$. Thus, any continuous function of $\frac{1}{n}\underline{\mathbf{C}}_{t,k}$ and $\underline{\mathbf{Z}}_{t,k} P_{t,k}$ will have compact support and thus be uniformly continuous. For any uniformly continuous function

$f : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$, for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any matrices $\underline{\mathbf{A}}, \underline{\mathbf{B}} \in \mathbb{R}^{d\times d}$, whenever $\|\underline{\mathbf{A}} - \underline{\mathbf{B}}\|_{\text{op}} < \delta$, then $\|f(\underline{\mathbf{A}}) - f(\underline{\mathbf{B}})\|_{\text{op}} < \epsilon$. Thus, for any $\epsilon > 0$, there exists some $\delta > 0$ such that

$$\mathbb{P}\left(\left\| \left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(A_{t,k}=k)} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \right) - \underline{\mathbf{Z}}_{t,k} P_{t,k} \right\|_{\text{op}} > \delta \right) \to 0$$

implies

$$\mathbb{P}\left(\left\| f\left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(A_{t,k}=k)} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \right) - f(\underline{\mathbf{Z}}_{t,k} P_{t,k}) \right\|_{\text{op}} > \epsilon \right) \to 0$$

Thus, by letting f be the matrix square-root function,

$$\left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(A_{t,k}=k)} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \right)^{1/2} - (\underline{\mathbf{Z}}_{t,k} P_{t,k})^{1/2} \xrightarrow{P} \underline{\mathbf{0}}.$$

We now want to show that for some constant $r > 0$, $\mathbb{P}\big(\|\underline{\mathbf{Z}}_{t,k}^{-1} \frac{1}{P_{t,k}}\|_{\text{op}} > r\big)$, because this would imply that

$$\left[\left(\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(A_{t,k}=k)} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \right)^{1/2} - (\underline{\mathbf{Z}}_{t,k} P_{t,k})^{1/2} \right] (\underline{\mathbf{Z}}_{t,k} P_{t,k})^{-1/2} \xrightarrow{P} \underline{\mathbf{0}}.$$

Recall that for $M_t^{(n)} = \mathbb{I}_{(\forall\, \mathbf{c}\in\mathbb{R}^d,\ \mathcal{A}_t(H_{t-1}^{(n)}, \mathbf{c}) \in [f(n), 1-f(n)]^K)}$, representing whether the conditional clipping condition is satisfied,

$$\underline{\mathbf{Z}}_{t,k}^{-1} = \underline{\mathbf{Z}}_{t,k}^{-1}(M_t^{(n)} + (1 - M_t^{(n)})) = \underline{\mathbf{Z}}_{t,k}^{-1} M_t^{(n)} + o_p(1).$$

Thus it is sufficient to show that $\mathbb{P}\big(\|\underline{\mathbf{Z}}_{t,k}^{-1} \frac{1}{P_{t,k}} M_t^{(n)}\|_{\text{op}} > r\big)$. Recall that by equation (28) we have that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1} M_t^{(n)}) \leq \frac{1}{f(n)} \lambda_{\max}\left(\big(\underline{\mathbf{\Sigma}}_t^{(n)}\big)^{-1} \right) M_t^{(n)} \leq \frac{1}{l\ f(n)}$$

Also note that $P_{t,k} = \mathbb{P}(A_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (M_t^{(n)} = 1) \geq f(n)$. Thus we have that

$$\mathbb{P}\left(\left\| \underline{\mathbf{Z}}_{t,k}^{-1} \frac{1}{P_{t,k}} M_t^{(n)} \right\|_{\text{op}} > r \right) \leq \mathbb{I}_{(\frac{1}{l\ f(n)^2} > r)} = 0$$

for $r > \frac{1}{l\ f(n)^2} = \frac{1}{lc^2}$, since we assume that $f(n) = c$ for some $0 < c \leq \frac{1}{2}$. □

Proof of Theorem 4: We define $P_{t,k} := \mathbb{P}(A_{t,i} = k | H_{t-1}^{(n)})$ and $\underline{\mathbf{Z}}_{t,k} := \mathbb{E}\big[\mathbf{C}_{t,i}\mathbf{C}_{t,i}^\top \big| H_{t-1}^{(n)}, A_{t,i} = k\big]$. We also define

$$\mathbf{D}_t^{(n)} := \text{Diagonal}\big[\underline{\mathbf{C}}_{t,0}, \underline{\mathbf{C}}_{t,1}, ..., \underline{\mathbf{C}}_{t,K-1}\big]^{1/2}(\hat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t) = \sum_{i=1}^n \begin{bmatrix} \underline{\mathbf{C}}_{t,0}^{-1/2}\ \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=0} \\ \underline{\mathbf{C}}_{t,1}^{-1/2}\ \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=1} \\ \vdots \\ \underline{\mathbf{C}}_{t,K-1}^{-1/2}\ \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=K-1} \end{bmatrix} \epsilon_{t,i}$$

We want to show that $[\mathbf{D}_1^{(n)}, \mathbf{D}_2^{(n)}, ..., \mathbf{D}_T^{(n)}]^\top \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \underline{\mathbf{I}}_{TKd})$. By Lemma 2 and Slutsky's Theorem, it sufficient to show that as $n \to \infty$, $[\mathbf{Q}_1^{(n)}, \mathbf{Q}_2^{(n)}, ..., \mathbf{Q}_T^{(n)}]^\top \xrightarrow{D} \mathcal{N}(\mathbf{0}, \sigma^2 \underline{\mathbf{I}}_{TKd})$ for

$$\mathbf{Q}_t^{(n)} := \sum_{i=1}^n \begin{bmatrix} \frac{1}{\sqrt{nP_{t,0}}} \underline{\mathbf{Z}}_{t,0}^{-1/2} \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=0} \\ \frac{1}{\sqrt{nP_{t,1}}} \underline{\mathbf{Z}}_{t,1}^{-1/2} \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=1} \\ \vdots \\ \frac{1}{\sqrt{nP_{t,K-1}}} \underline{\mathbf{Z}}_{t,K-1}^{-1/2} \mathbf{C}_{t,i}\mathbb{I}_{A_{t,i}=K-1} \end{bmatrix} \epsilon_{t,i}$$

By Cramer Wold device, it is sufficient to show that for any $\mathbf{b} \in \mathbb{R}^{TKd}$ with $\|\mathbf{b}\|_2 = 1$, where $\mathbf{b} = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_T]$ for $\mathbf{b}_t \in \mathbb{R}^{Kd}$, as $n \to \infty$.

$$\sum_{t=1}^{T} \mathbf{b}_t^\top \mathbf{Q}_t^{(n)} \xrightarrow{D} \mathcal{N}(0, \sigma^2) \tag{29}$$

We can further define for all $t \in [1\colon T]$, $\mathbf{b}_t = [\mathbf{b}_{t,0}, \mathbf{b}_{t,1}, ..., \mathbf{b}_{t,K-1}]$ with $\mathbf{b}_{t,k} \in \mathbb{R}^d$. Thus to show (29) it is equivalent to show that

$$\sum_{t=1}^{T} \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \underline{\mathbf{Z}}_{t,k}^{-1/2} \sum_{i=1}^{n} \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

We define $Y_{t,i}^{(n)} := \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \mathbb{I}_{A_{t,i}=k} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \epsilon_{t,i}$. The sequence $Y_{1,1}^{(n)}, Y_{1,2}^{(n)}, ..., Y_{1,n}^{(n)}, ... Y_{T,1}^{(n)}, Y_{T,2}^{(n)}, ..., Y_{T,n}^{(n)}$ is a martingale difference array with respect to the sequence of histories $\{H_{t-1}^{(n)}\}_{t=1}^{T}$ because $\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}] = \mathbb{E}\left[\mathbb{E}[Y_{t,i}^{(n)} | H_{t-1}^{(n)}, A_{t,i}, \mathbf{C}_{t,i}] \middle| H_{t-1}^{(n)}\right] = 0$ for all $i \in [1\colon n]$ and all $t \in [1\colon T]$. We then apply the martingale central limit theorem of Dvoretzky (1972) to $Y_{t,i}^{(n)}$ to show the desired result (see the proof of Theorem 5 in Appendix B for the statement of the martingale CLT conditions). Note that the first condition (a) of the martingale CLT is already satisfied, as we just showed that $Y_{t,i}^{(n)}$ form a martingale difference array with respect to $H_{t-1}^{(n)}$.

Condition(b): Conditional Variance

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}[Y_{t,i}^2 | H_{t-1}^{(n)}] = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \epsilon_{t,i} \right)^2 \middle| H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \epsilon_{t,i}^2 \middle| H_{t-1}^{(n)}\right] \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{b}_{t,k}$$

By law of iterated expectations (conditioning on $H_{t-1}^{(n)}, A_{t,i}, \mathbf{C}_{t,i}$) and Condition 6,

$$= \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \frac{1}{P_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \middle| H_{t-1}^{(n)}\right] \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{b}_{t,k} \sigma^2$$

$$= \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \frac{1}{P_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbb{E}\left[\mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \middle| H_{t-1}^{(n)}, A_{t,i} = k\right] P_{t,k} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{b}_{t,k} \sigma^2$$

$$= \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \underline{\mathbf{I}}_d \mathbf{b}_{t,k} \sigma^2 = \sigma^2 \sum_{t=1}^{T} \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k} = \sigma^2$$

Condition(c): Lindeberg Condition Let $\delta > 0$.

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}\left[Y_{t,i}^2 \mathbb{I}_{(|Y_{t,i}| > \delta)} \middle| H_{t-1}^{(n)}\right] = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbb{E}\left[\left(\sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \mathbf{Z}_{t,i}^{-1/2} \mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \epsilon_{t,i} \right)^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}\right]$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=0}^{K-1} \frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^\top \mathbf{Z}_{t,i}^{-1/2} \mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}\right] \mathbf{Z}_{t,i}^{-1/2} \mathbf{b}_{t,k}$$

It is sufficient to show that for any $t \in [1\colon T]$ and any $k \in [0 : K-1]$ the following converges in probability to zero:

$$\sum_{i=1}^{n} \frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^\top \mathbf{Z}_{t,i}^{-1/2} \mathbb{E}\left[\mathbb{I}_{A_{t,i}=k} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \epsilon_{t,i}^2 \mathbb{I}_{(Y_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}\right] \mathbf{Z}_{t,i}^{-1/2} \mathbf{b}_{t,k}$$

Recall that $Y_{t,i} = \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \frac{1}{\sqrt{nP_{t,k}}} \mathbb{I}_{A_{t,i}=k} \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \epsilon_{t,i}$.

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^\top \mathbf{Z}_{t,i}^{-1/2} \mathbb{E}\left[\mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \epsilon_{t,i}^2 \mathbb{I}_{(\frac{1}{nP_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{C}_{t,i} \mathbf{C}_{t,i}^\top \underline{\mathbf{Z}}_{t,k}^{-1/2} \mathbf{b}_{t,k} \epsilon_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] \mathbf{Z}_{t,i}^{-1/2} \mathbf{b}_{t,k}$$

Since $\mathbf{c} \in [-u, u]$, by the Gershgorin circle theorem, we can bound the maximum eigenvalue of $\mathbf{c}\mathbf{c}^\top$ by some constant $a > 0$.

$$\leq \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^\top \mathbf{Z}_{t,i}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \epsilon_{t,i}^2 > \delta^2)} \;\middle|\; H_{t-1}^{(n)}, A_{t,i} = k \right]$$

We define random variable $M_t^{(n)} = \mathbb{I}_{(\forall \ \mathbf{c} \in \mathbb{R}^d, \ \mathcal{A}_t(H_{t-1}^{(n)}, \mathbf{c}) \in [f(n), 1-f(n)]^K)}$, representing whether the conditional clipping condition is satisfied. Note that by our conditional clipping assumption, $M_t^{(n)} \overset{P}{\to} 1$ as $n \to \infty$.

$$= \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \epsilon_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] \left(M_t^{(n)} + (1 - M_t^{(n)}) \right)$$

$$= \frac{a}{n} \sum_{i=1}^{n} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\frac{a}{nP_{t,k}} \mathbf{b}_{t,k}^\top \underline{\mathbf{Z}}_{t,k}^{-1} \mathbf{b}_{t,k} \epsilon_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] M_t^{(n)} + o_p(1) \tag{30}$$

By equation (28), have that

$$\lambda_{\max}(\underline{\mathbf{Z}}_{t,k}^{-1}) \leq \frac{1}{f(n)} \lambda_{\max}\left(\left(\underline{\mathbf{\Sigma}}_t^{(n)} \right)^{-1} \right) \leq \frac{1}{l \ f(n)}$$

Recall that $P_{t,k} = \mathbb{P}(A_{t,i} = k \mid H_{t-1}^{(n)})$, so $P_{t,k} \mid (M_t^{(n)} = 1) \geq f(n)$. Thus we have that equation (30) is upper bounded by the following:

$$\leq \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{l \ f(n)} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\frac{a}{nf(n)} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{l \ f(n)} \epsilon_{t,i}^2 > \delta^2)} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] + o_p(1)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}}{l \ f(n)} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 > \delta^2 \frac{l \ nf(n)^2}{a \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}})} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] + o_p(1)$$

It is sufficient to show that

$$\lim_{n \to \infty} \max_{i \in [1 \colon n]} \frac{1}{f(n)} \mathbb{E}\left[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 > \delta^2 \frac{l \ nf(n)^2}{a \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}})} \middle| H_{t-1}^{(n)}, A_{t,i} = k \right] = 0. \tag{31}$$

By Condition 6, we have that for all $n \geq 1$, $\max_{t \in [1 \colon T], i \in [1 \colon n]} \mathbb{E}[\varphi(\epsilon_{t,i}^2) | H_{t-1}^{(n)}, A_{t,i} = k] < M$.

Since we assume that $\lim_{x \to \infty} \frac{\varphi(x)}{x} = \infty$, for all $m \geq 1$, there exists a b_m s.t. $\varphi(x) \geq mMx$ for all $x \geq b_m$. So, for all n, t, i,

$$M \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) | H_{t-1}^{(n)}, A_{t,i} = k] \geq \mathbb{E}[\varphi(\epsilon_{t,i}^2) \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k] \geq mM \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k]$$

Thus, $\max_{i \in [1 \colon n]} \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k] \leq \frac{1}{m}$; so $\lim_{m \to \infty} \max_{i \in [1 \colon n]} \mathbb{E}[\epsilon_{t,i}^2 \mathbb{I}_{(\epsilon_{t,i}^2 \geq b_m)} | H_{t-1}^{(n)}, A_{t,i} = k] = 0$. Since by our conditional clipping assumption, $f(n) = c$ for some $0 < c \leq \frac{1}{2}$ thus $nf(n)^2 \to \infty$. So equation (31) holds. □

Corollary 4 (Asymptotic Normality of the Batched OLS for Margin with Context Statistic)**.** *Assume the same conditions as Theorem 4. For any two arms $x, y \in [0\colon K-1]$ for all $t \in [1\colon T]$, we have the BOLS estimator for $\boldsymbol{\Delta_{t,x-y}} := \boldsymbol{\beta}_{t,x} - \boldsymbol{\beta}_{t,y}$. We show that as $n \to \infty$,*

$$\begin{bmatrix} \left[\underline{\boldsymbol{C}}_{1,x}^{-1} + \underline{\boldsymbol{C}}_{1,y}^{-1}\right]^{1/2} (\hat{\boldsymbol{\Delta}}_{1,x-y}^{\text{BOLS}} - \boldsymbol{\Delta}_{1,x-y}) \\ \left[\underline{\boldsymbol{C}}_{2,x}^{-1} + \underline{\boldsymbol{C}}_{2,y}^{-1}\right]^{1/2} (\hat{\boldsymbol{\Delta}}_{2,x-y}^{\text{BOLS}} - \boldsymbol{\Delta}_{2,x-y}) \\ \vdots \\ \left[\underline{\boldsymbol{C}}_{T,x}^{-1} + \underline{\boldsymbol{C}}_{T,y}^{-1}\right]^{1/2} (\hat{\boldsymbol{\Delta}}_{T,x-y}^{\text{BOLS}} - \boldsymbol{\Delta}_{T,x-y}) \end{bmatrix} \xrightarrow{D} \mathcal{N}(0, \sigma^2 \underline{\boldsymbol{I}}_{Td})$$

where

$$\hat{\boldsymbol{\Delta}}_{t,x-y}^{\text{BOLS}} = \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1} \left(\underline{\mathbf{C}}_{t,y}^{-1} \sum_{i=1}^n A_{t,i} \mathbf{C}_{t,i} R_{t,i} - \underline{\mathbf{C}}_{t,x}^{-1} \sum_{i=1}^n (1 - A_{t,i}) \mathbf{C}_{t,i} R_{t,i} \right).$$

Proof: By Cramer-Wold device, it is sufficient to show that for any fixed vector $\mathbf{d} \in \mathbb{R}^{Td}$ s.t. $\|\mathbf{d}\|_2 = 1$, where $\mathbf{d} = [\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_T]$ for $\mathbf{d}_t \in \mathbb{R}^d$, $\sum_{t=1}^T \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} (\hat{\boldsymbol{\Delta}}_{t,x-y}^{\text{BOLS}} - \boldsymbol{\Delta}_{t,x-y}) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$, as $n \to \infty$.

$$\sum_{t=1}^T \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} (\hat{\boldsymbol{\Delta}}_{t,x-y}^{\text{BOLS}} - \boldsymbol{\Delta}_{t,x-y}) = \sum_{t=1}^T \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1/2} \left(\underline{\mathbf{C}}_{t,y}^{-1} \sum_{i=1}^n A_{t,i} \mathbf{C}_{t,i} \epsilon_{t,i} - \underline{\mathbf{C}}_{t,x}^{-1} \sum_{i=1}^n (1 - A_{t,i}) \mathbf{C}_{t,i} \epsilon_{t,i} \right)$$

By Lemma 2, as $n \to \infty$, $\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} \underline{\mathbf{C}}_{t,x} \xrightarrow{P} \underline{\mathbf{I}}_d$ and $\frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \underline{\mathbf{C}}_{t,y} \xrightarrow{P} \underline{\mathbf{I}}_d$, so by Slutsky's Theorem it is sufficient to that as $n \to \infty$,

$$\sum_{t=1}^T \mathbf{d}_t^\top \left[\underline{\mathbf{C}}_{t,x}^{-1} + \underline{\mathbf{C}}_{t,y}^{-1}\right]^{-1/2} \left(\frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \sum_{i=1}^n A_{t,i} \mathbf{C}_{t,i} \epsilon_{t,i} - \frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} \sum_{i=1}^n (1 - A_{t,i}) \mathbf{C}_{t,i} \epsilon_{t,i} \right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

We know that $\left[\frac{1}{P_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \left[\frac{1}{P_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{1/2} \xrightarrow{P} \underline{\mathbf{I}}_d$.

By Lemma 2 and continuous mapping theorem, $nP_{t,x} \underline{\mathbf{Z}}_{t,x} \underline{\mathbf{C}}_{t,x}^{-1} \xrightarrow{P} \underline{\mathbf{I}}_d$ and $nP_{t,y} \underline{\mathbf{Z}}_{t,y} \underline{\mathbf{C}}_{t,y}^{-1} \xrightarrow{P} \underline{\mathbf{I}}_d$. So by Slutsky's Theorem,

$$\left[\frac{1}{P_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{P_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \left[n \underline{\mathbf{C}}_{t,x}^{-1} + n \underline{\mathbf{C}}_{t,y}^{-1}\right]^{1/2} \xrightarrow{P} \underline{\mathbf{I}}_d$$

So, returning to our CLT, by Slutsky's Theorem, it is sufficient to show that as $n \to \infty$,

$$\sum_{t=1}^T \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \sum_{i=1}^n A_{t,i} \mathbf{C}_{t,i} \epsilon_{t,i}$$

$$- \sum_{t=1}^T \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} \sum_{i=1}^n (1 - A_{t,i}) \mathbf{C}_{t,i} \epsilon_{t,i} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

The above sum equals the following:

$$= \sum_{t=1}^T \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} \left(\frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} \sum_{i=1}^n A_{t,i} \mathbf{C}_{t,i} \epsilon_{t,i} \right)$$

$$- \sum_{t=1}^T \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,y}^{-1}\right]^{-1/2} \frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} \left(\frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} \sum_{i=1}^n (1 - A_{t,i}) \mathbf{C}_{t,i} \epsilon_{t,i} \right)$$

Asymptotic normality holds by the same martingale CLT as we used in the proof of Theorem 4. The only difference is that we adjust our $\mathbf{b}_{t,k}$ vector from Theorem 4 to the following:

$$\mathbf{b}_{t,k} := \begin{cases} \mathbf{0} & \text{if } k \notin \{x,y\} \\ \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,x}}} \underline{\mathbf{Z}}_{t,x}^{-1/2} & \text{if } k = x \\ \mathbf{d}_t^\top \left[\frac{1}{nP_{t,x}} \underline{\mathbf{Z}}_{t,x}^{-1} + \frac{1}{nP_{t,y}} \underline{\mathbf{Z}}_{t,y}^{-1} \right]^{-1/2} \frac{1}{\sqrt{nP_{t,y}}} \underline{\mathbf{Z}}_{t,y}^{-1/2} & \text{if } k = y \end{cases}$$

The proof still goes through with this adjustment because for all $k \in [0\colon K-1]$, (i) $\mathbf{b}_{t,k} \in H_{t-1}^{(n)}$, (ii) $\sum_{t=1}^T \sum_{k=0}^{K-1} \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k} = \sum_{t=1}^T \mathbf{d}_t^\top \mathbf{d}_t = 1$. and (iii) $\frac{l\, n f(n)^2}{a \mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}} \to \infty$ still holds because $\mathbf{b}_{t,k}^\top \mathbf{b}_{t,k}$ is bounded above by one. □

F. W-Decorrelated Estimator (Deshpande et al., 2018)

To better understand why the W-decorrelated estimator has relatively low power, but is still able to guarantee asymptotic normality, we now investigate the form of the W-decorrelated estimator in the two-arm bandit setting.

F.1. Decorrelation Approach

We now assume we are in the unbatched setting (i.e., batch size of one), as the W-decorrelated estimator was developed for this setting; however, these results easily translate to the batched setting. We now let n index the number of samples total (previously this was nT) and examine asymptotics as $n \to \infty$. We assume the following model:

$$\mathbf{R}_n = \underline{\mathbf{X}}_n^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_n$$

where $\mathbf{R}_n, \boldsymbol{\epsilon}_n \in \mathbb{R}^n$ and $\underline{\mathbf{X}}_n \in \mathbb{R}^{n \times p}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$. The W-decorrelated OLS estimator is defined as follows:

$$\hat{\boldsymbol{\beta}}^d = \hat{\boldsymbol{\beta}}_{\text{OLS}} + \underline{\mathbf{W}}_n (\mathbf{R}_n - \underline{\mathbf{X}}_n \hat{\boldsymbol{\beta}}_{\text{OLS}})$$

With this definition we have that,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^d - \boldsymbol{\beta} &= \hat{\boldsymbol{\beta}}_{\text{OLS}} + \underline{\mathbf{W}}_n (\mathbf{R}_n - \underline{\mathbf{X}}_n \hat{\boldsymbol{\beta}}_{\text{OLS}}) - \boldsymbol{\beta} \\
&= \hat{\boldsymbol{\beta}}_{\text{OLS}} + \underline{\mathbf{W}}_n (\underline{\mathbf{X}}_n \boldsymbol{\beta} + \boldsymbol{\epsilon}_n) - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n \hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta} \\
&= (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n)(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}) + \underline{\mathbf{W}}_n \boldsymbol{\epsilon}_n
\end{aligned}$$

Note that if $\mathbb{E}[\underline{\mathbf{W}}_n \boldsymbol{\epsilon}_n] = \mathbb{E}\big[\sum_{i=1}^n \mathbf{W}_i \epsilon_i\big] = 0$ (where $\mathbf{W}_i$ is the i^{th} column of $\underline{\mathbf{W}}_n$), then $\mathbb{E}[(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n)(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})]$ would be the bias of the estimator. We assume $\{\epsilon_i\}$ is a martingale difference sequence w.r.t. filtration $\{\mathcal{G}_i\}_{i=1}^n$. Thus, if we constrain $\mathbf{W}_i$ to be $\mathcal{G}_{i-1}$ measurable,

$$\mathbb{E}[\underline{\mathbf{W}}_n \boldsymbol{\epsilon}_n] = \mathbb{E}\big[\sum_{i=1}^n \mathbf{W}_i \epsilon_i\big] = \sum_{i=1}^n \mathbb{E}\Big[\mathbb{E}[\mathbf{W}_i \epsilon_i | \mathcal{G}_{i-1}]\Big] = \sum_{i=1}^n \mathbb{E}\Big[\mathbf{W}_i \mathbb{E}[\epsilon_i | \mathcal{G}_{i-1}]\Big] = 0$$

Trading off Bias and Variance While decreasing $\mathbb{E}[(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n)(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})]$ will decrease the bias, making $\underline{\mathbf{W}}_n$ larger in norm will increase the variance. So the trade-off between bias and variance can be adjusted with different values of λ for the following optimization problem:

$$||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n||_F^2 + \lambda ||\underline{\mathbf{W}}_n||_F^2 = ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_n \underline{\mathbf{X}}_n||_F^2 + \lambda \text{Tr}(\underline{\mathbf{W}}_n \underline{\mathbf{W}}_n^\top)$$

Optimizing for $\underline{\mathbf{W}}_n$ The authors propose to optimize for $\underline{\mathbf{W}}_n$ in a recursive fashion, so that the i^{th} column, $\mathbf{W}_i$, only depends on $\{\mathbf{X}_j\}_{j \leq i} \cup \{\epsilon_j\}_{j \leq i-1}$ (so $\sum_{i=1}^n \mathbb{E}[\mathbf{W}_i \epsilon_i] = 0$). We let $\mathbf{W}_0 = 0$, $\mathbf{X}_0 = 0$, and recursively define $\underline{\mathbf{W}}_n := [\underline{\mathbf{W}}_{n-1} \mathbf{W}_n]$ where

$$\mathbf{W}_n = \text{argmin}_{\mathbf{W} \in \mathbb{R}^p} ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top||_F^2 + \lambda ||\mathbf{W}||_2^2$$

where $\underline{\mathbf{W}}_{n-1} = [\mathbf{W}_1; \mathbf{W}_2; ...; \mathbf{W}_{n-1}] \in \mathbb{R}^{p \times (n-1)}$ and $\underline{\mathbf{X}}_{n-1} = [\mathbf{X}_1; \mathbf{X}_2; ...; \mathbf{X}_{n-1}]^\top \in \mathbb{R}^{(n-1) \times p}$. Now, let us find the closed form solution for each step of this minimization:

$$\frac{d}{d\mathbf{W}} ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top||_F^2 + \lambda ||\mathbf{W}||_2^2 = 2(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top)(-\mathbf{X}_n) + 2\lambda \mathbf{W}$$

Note that since the Hessian is positive definite, so we can find the minimizing $\mathbf{W}$ by setting the first derivative to 0:

$$\frac{d^2}{d\mathbf{W} d\mathbf{W}^\top} ||\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top||_F^2 + \lambda ||\mathbf{W}||_2^2 = 2\mathbf{X}_n \mathbf{X}_n^\top + 2\lambda \underline{\mathbf{I}}_p \succcurlyeq 0$$

$$\begin{gathered}
0 = 2(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top)(-\mathbf{X}_n) + 2\lambda \mathbf{W} \\
(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1} - \mathbf{W} \mathbf{X}_n^\top) \mathbf{X}_n = \lambda \mathbf{W} \\
(\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1}) \mathbf{X}_n = \lambda \mathbf{W} + \mathbf{W} \mathbf{X}_n^\top \mathbf{X}_n = (\lambda + \|\mathbf{X}_n\|_2^2) \mathbf{W} \\
\mathbf{W}^* = (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{n-1} \underline{\mathbf{X}}_{n-1}) \frac{\mathbf{X}_n}{\lambda + \|\mathbf{X}_n\|_2^2}
\end{gathered}$$

Proposition 3 (W-decorrelated estimator and time discounting in the two-arm bandit setting)**.** *Suppose we have a 2-arm bandit. A_i is an indicator that equals 1 if arm 1 is chosen for the i^{th} sample, and 0 if arm 0 is chosen. We define $\boldsymbol{X}_i := [1 - A_i, A_i] \in \mathbb{R}^2$. We assume the following model of rewards:*

$$R_i = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \epsilon_i = A_i \beta_1 + (1 - A_i)\beta_0 + \epsilon_i$$

We further assume that $\{\epsilon_i\}_{i=1}^n$ are a martingale difference sequence with respect to filtration $\{\mathcal{G}_i\}_{i=1}^n$. We also assume that $\boldsymbol{X}_i$ are non-anticipating with respect to filtration $\{\mathcal{G}_i\}_{i=1}^n$. Note the W-decorrelated estimator:

$$\hat{\boldsymbol{\beta}}^d = \hat{\boldsymbol{\beta}}_{\text{OLS}} + \underline{\boldsymbol{W}}_n(\boldsymbol{R}_n - \underline{\boldsymbol{X}}_n \hat{\boldsymbol{\beta}}_{\text{OLS}})$$

We show that for $\underline{\boldsymbol{W}}_n = [\boldsymbol{W}_1; \boldsymbol{W}_2; ...; \boldsymbol{W}_n] \in \mathbb{R}^{p \times n}$ and choice of constant λ,

$$\boldsymbol{W}_i = \begin{bmatrix} (1 - \frac{1}{\lambda+1})^{\sum_{i=1}^n (1-A_i)} \frac{1}{\lambda+1} \\ (1 - \frac{1}{\lambda+1})^{\sum_{i=1}^n A_i} \frac{1}{\lambda+1} \end{bmatrix} \in \mathbb{R}^2$$

Moreover, we show that the W-decorrelated estimator for the mean of arm 1, β_1, is as follows:

$$\hat{\beta}_1^d = \left(1 - \sum_{i=1}^n A_t \frac{1}{\lambda+1}\left(1 - \frac{1}{\lambda+1}\right)^{N_{1,i}-1}\right)\hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^n A_t R_t \cdot \frac{1}{\lambda+1}\left(1 - \frac{1}{\lambda+1}\right)^{N_{1,i}-1}$$

where $\hat{\beta}_1^{\text{OLS}} = \frac{\sum_{i=1}^n A_i R_i}{N_{1,n}}$ for $N_{1,n} = \sum_{i=1}^n A_i$. Since Deshpande et al. (2018) require that $\lambda \geq 1$ for their CLT results to hold, thus, the W-decorrelated estimators is down-weighting samples drawn later on in the study and up-weighting earlier samples.

Proof: Recall the formula for $\mathbf{W}_i$,

$$\mathbf{W}_i = (\underline{\mathbf{I}}_p - \underline{\mathbf{W}}_{i-1}\underline{\mathbf{X}}_{i-1})\frac{\mathbf{X}_i}{\lambda + \|\mathbf{X}_i\|_2^2}$$

We let $\mathbf{W}_i = [W_{0,i}, W_{1,i}]^\top$. For notational simplicity, we let $r = \frac{1}{\lambda+1}$. We now solve for $W_{1,n}$:

$$W_{1,1} = (1 - 0) \cdot rA_1 = rA_1$$

$$W_{1,2} = (1 - W_{1,1} \cdot A_1) \cdot rA_2 = (1 - rA_1)rA_2$$

$$W_{1,3} = \left(1 - \sum_{i=1}^2 \mathbf{W}_{1,i} \cdot A_i\right) \cdot rA_3 = \left(1 - rA_1 - (1 - rA_1)rA_2\right) \cdot rA_3 = (1 - rA_1)(1 - rA_2) \cdot rA_3$$

$$W_{1,4} = \left(1 - \sum_{i=1}^3 \mathbf{W}_{1,i} \cdot A_i\right) \cdot rA_4 = \left(1 - rA_1 - (1 - rA_1)rA_2 - (1 - rA_1)(1 - rA_2) \cdot rA_3\right) \cdot rA_4$$
$$= (1 - rA_1)\big(1 - rA_2 - (1 - rA_2)rA_3)\big) \cdot rA_4 = (1 - rA_1)(1 - rA_2)(1 - rA_3) \cdot rA_4$$

We have that for arbitrary n,

$$W_{1,n} = \left(1 - \sum_{i=1}^{n-1} \mathbf{W}_{1,i} \cdot A_i\right) \cdot rA_n = rA_n \prod_{i=1}^{n-1}(1 - rA_i) = rA_n(1 - r)^{\sum_{i=1}^{n-1} A_i} = rA_n(1 - r)^{N_{1,n-1}}$$

By symmetry, we have that

$$W_{0,n} = \left(1 - \sum_{i=1}^{n-1} \mathbf{W}_{1,i} \cdot (1 - A_i)\right) \cdot r(1 - A_n) = r(1 - A_n)(1 - r)^{N_{0,n-1}}$$

Note the W-decorrelated estimator for β_1:

$$\hat{\beta}_1^d = \hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^n A_i \left(R_i - \hat{\beta}_1^{\text{OLS}}\right) r(1 - r)^{N_{1,i-1}}$$

$$= \left(1 - \sum_{i=1}^n A_i r(1 - r)^{N_{1,i-1}}\right)\hat{\beta}_1^{\text{OLS}} + \sum_{i=1}^n A_i R_i \cdot r(1 - r)^{N_{1,i-1}} \quad \square$$