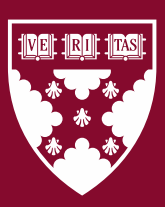
Working Paper 24-057

An Experimental Design for Anytime-Valid Causal Inference on Multi-Armed Bandits

Biyonka Liang
Iavor Bojinov



Harvard
Business
School

An Experimental Design for Anytime-Valid Causal Inference on Multi-Armed Bandits

Biyonka Liang
Harvard University

Iavor Bojinov
Harvard Business School

Working Paper 24-057

Funding for this research was provided in part by Harvard Business School.

An Experimental Design for Anytime-Valid Causal Inference on Multi-Armed Bandits

Biyonka Liang[1] and Iav Bojinov[2]

[1]Department of Statistics, Harvard University
[2]Harvard Business School

November 9th, 2023

Abstract

Typically, multi-armed bandit (MAB) experiments are analyzed at the *end* of the study and thus require the analyst to specify a fixed sample size in advance. However, in many online learning applications, it is advantageous to continuously produce inference on the average treatment effect (ATE) between arms as new data arrive and determine a data-driven stopping time for the experiment. Existing work on continuous inference for adaptive experiments assumes that the treatment assignment probabilities are bounded away from zero and one, thus excluding nearly all standard bandit algorithms. In this work, we develop the Mixture Adaptive Design (MAD), a new experimental design for multi-armed bandits that enables continuous inference on the ATE with guarantees on statistical validity and power *for nearly any bandit algorithm*. On a high level, the MAD "mixes" a bandit algorithm of the user's choice with a Bernoulli design through a tuning parameter δ_t, where δ_t is a deterministic sequence that controls the priority placed on the Bernoulli design as the sample size grows. We show that for $\delta_t = o\left(1/t^{1/4}\right)$, the MAD produces a confidence sequence that is asymptotically valid and guaranteed to shrink around the true ATE. We empirically show that the MAD improves the coverage and power of ATE inference in MAB experiments without significant losses in finite-sample reward.

Keywords: Adaptive Experimental Design, Multi-armed Bandit, Online Learning, Sequential Analysis, Always-valid inference, Asymptotic Confidence Sequence, A/B Test

1 Introduction

1.1 Motivation

Multi-armed bandits (MABs) are one of the most widely used frameworks for sequential decision making, with applications in clinical trial design (Villar et al. 2015; Durand et al. 2018), network anomaly detection (Ding et al. 2019), recommendation systems (Mary et al. 2015), and various other online learning applications (Dimakopoulou et al. 2018; Kohavi

et al. 2020; Zhang et al. 2020; Bojinov and Gupta 2022). In many modern online decision-making settings, MABs are often favored over traditional randomized control trials (or A/B tests) as they reduce the risk of experimentation by dynamically updating the assignment probabilities, thereby decreasing the proportion of units exposed to sub-par treatments. However, it is often desirable, even vital, for decision-makers to conduct statistical inference on the average treatment effect (ATE) between arms, not just to minimize risk. For example, additive features in software often come with a monetary cost, e.g., maintenance from engineers and higher CPU demands. To drive product innovation and augment managerial decision-making, it is often necessary to understand how much an additive feature moved some user metric, not just whether it did or not (Moe et al. 2012). Another example is in clinical trials for drug testing. As drugs often come with side effects or have interactions with other medications, having statistical inference on the treatment effect between the various drugs and/or doses being compared is important to guide policies around which drug and/or dose should be prescribed given different patient histories (Böttiger et al. 2009).

However, simultaneously conducting inference on the ATE while minimizing regret is particularly challenging because MABs determine treatment assignments adaptively; most statistical guarantees of ATE estimation methods for independent and identically distributed (i.i.d.) data no longer hold in the MAB setting. For instance, the difference in sample averages between a treatment arm and a control arm is no longer unbiased for the true ATE of these arms (Xu et al. 2013) and may no longer be asymptotically normal Dimakopoulou et al. (2021); Zhang et al. (2020). While Inverse Propensity Weighting (IPW) estimators are often unbiased in the MAB setting (Horvitz and Thompson 1952; Hadad et al. 2021), their variance can explode, thus making any subsequent hypothesis testing on the ATE essentially powerless. The variance explodes because the probability that a bandit algorithm pulls a sub-optimal arm rapidly approaches zero (Russo et al. 2020). For instance, the well-known Thompson Sampling (TS) algorithm (Agrawal and Goyal 2012) and variations of the Upper Confidence Bound (UCB) algorithm (Garivier and Cappé 2011; Ménard and Garivier 2017; Kaufmann 2018) achieve a regret bound of $O(\log T)$ and are said to be asymptotically optimal as the regret of any algorithm must have at least a $\log(T)$ rate (Lai and Robbins 1985). Hence, we expect the number of draws from sub-optimal arms to be on the order of $\log(T)/T$ in the long run (Kasy and Sautmann 2021).

Therefore, the design objective of this work is to develop an experimental design for MABs which provides statistical guarantees on ATE estimation while simultaneously allowing for regret minimization via adaptive assignment.

1.2 Background and Existing Work

As MABs can be viewed as an adaptive experimental design, various works on the design of MABs have proposed adaptations of existing bandit algorithms to be more amenable to ATE estimation. For instance, Kasy and Sautmann (2021) proposes a variation of Thompson sampling which re-weights the estimated mean reward of each arm to enforce additional exploration of sub-optimal arms. Simchi-Levi and Wang (2023) first formulates the trade-off between mean squared estimation error and regret via a minimax multi-objective optimization problem, then proposes a variation of the EXP3 algorithm (Auer et al. 2002)

which takes draws from the estimated sub-optimal arms via a specified rate and shows that this adaptation satisfies this characterization of optimality. Intuitively, it is sensible that some degree of exploration must be imposed to produce inferential guarantees, as the precision of ATE inference essentially depends on the variance of the treatment arm which has the fewest number of observations. However, these works focus on adapting *specific* bandit algorithms, thus restricting the analyst to a specific algorithm. Additionally, these works propose designs that do not allow the user to tune the degree of exploration in the algorithm (Kasy and Sautmann 2021) or have limitations on when during the experiment this tuning can take place (Simchi-Levi and Wang 2023). Hahn et al. (2011) proposes a two-stage adaptive design and proves an asymptotic normality result for ATE estimation in this setting, but requires the adaptive assignment algorithm to have assignment probabilities bounded away from zero and one, which is not satisfied by most common bandit algorithms such as UCB (a deterministic algorithm) and Thompson sampling. Works such as Zhang et al. (2020, 2021) develop CLT results for a general class of estimators for adaptively collected data without significantly altering the design of the bandit algorithm, but requires certain assumptions on the outcome distributions, such as moment bounds or independently and identically distributed outcomes. Most importantly, all of the above approaches are designed for the setting where the analyst is conducting inference at the *end* of an adaptive experiment with a pre-specified time horizon.

However, since the fundamental idea behind MABs and adaptive experiments in general is to use existing data to inform future decisions, a more natural inferential framework would be to allow the analyst to continuously produce inference on the ATE between arms as new data arrive. Such a framework would allow analysts to "peek" at the data in their experiment without invalidating the subsequent statistical inference. Therefore, experiments would not require the analyst to specify a sample size in advance, but determine a data-dependent stopping time in response to running inferential results (*e.g.*, stop the experiment once statistically significant results are achieved), further minimizing harm to experimental units and reducing the opportunity cost of running long experiments.

Such a continuous monitoring setting requires statistical tests that uniformly control Type I error at every time point, i.e., are *anytime-valid*. As classical tests do not satisfy this condition, prior works have proposed using *confidence sequences* to enable valid inference (Waudby-Smith et al. 2023). A confidence sequence (CS) is a set of confidence intervals $\{C_t\}_{t=1}^{\infty}$ at level $\alpha \in (0,1)$ such that, for a true (non-zero) treatment effect τ_t,

$$\mathbb{P}\left(\forall t, \tau_t \in C_t\right) \geq 1-\alpha.$$

However, non-asymptotic confidence sequences often require specific assumptions on the data such as known bounds on the random variables and/or tail behavior assumptions such as sub-Gaussianity (Waudby-Smith and Ramdas 2020; Howard et al. 2021; Howard and Ramdas 2022). In particular, such assumptions prevent non-asymptotic CSs from being applicable or even possible to construct in many real-world settings. To address this restriction, (Waudby-Smith et al. 2023) first introduced the notion of an *asymptotic confidence sequence* and derived a universal asymptotic CS that requires only CLT-like assumptions.

Definition 1 (Asymptotic Confidence Sequence)**.** $(\hat{\mu}_t \pm \hat{V}_t)$ *is an asymptotic* $1-\alpha$ *confidence*

sequence for a target parameter μ_t *if there exists some (unknown) exact* $1-\alpha$ *confidence sequence* $(\hat{\mu}_t \pm V_t)$ *for* μ_t *such that*

$$\frac{\hat{V}_t}{V_t} \xrightarrow{a.s.} 1.$$

The assumptions necessary for generating asymptotic CSs are comparatively much weaker, and hence, asymptotic CSs expand the utility of CSs to a wider array of settings. For instance, Ham et al. (2022) proposes a framework for continuous design-based causal inference for time series and panel experiments utilizing asymptotic CSs. However, existing work on continuous inference for *adaptive* or MAB experiments is rather limited. While Ham et al. (2022) and Howard et al. (2021) discuss extensions to adaptive experiments, their proposed CSs require *probabilistic treatment assignment*, i.e., they assume there exists $0 < p_{min} \leq 1/2$ s.t. $p_{t|t-1}(w) \in [p_{min}, 1-p_{min}]$ almost surely. This restriction excludes many commonly used bandit algorithms such as UCB, which has deterministic treatment assignments, and Thompson Sampling. Without this treatment assignment assumption, the asymptotic confidence sequences of Ham et al. (2022) are *not guaranteed to decrease with the sample size or even be asymptotically valid* (Waudby-Smith et al. 2023). The non-asymptotic confidence sequences of Howard et al. (2021) may be inapplicable since their CS requires the analyst to input p_{min}. Thus, naively applying such approaches in MAB settings without any adjustment to the experimental design provides no guarantees on validity and hence, cannot be used to reliably generate inference on the ATE.

1.3 Our Contributions

In this work, we develop the Mixture Adaptive Design (MAD), a new approach to the experimental design of multi-armed bandit experiments that enables anytime-valid inference on the ATE in MAB experiments with guarantees on validity and power. Intuitively, the MAD "mixes" any bandit algorithm with a Bernoulli design through a tuning parameter δ_t, where $\delta_t \in (0, 1]$ is a deterministic, i.e., non-random, sequence that controls the priority placed on the Bernoulli design as the sample size grows. For $\delta_t = o\left(\frac{1}{t^{1/4}}\right)$, we provide a confidence sequence that is asymptotically valid and guaranteed to shrink around the true ATE, *for nearly any choice of bandit algorithm*. This confidence sequence is, to our knowledge, the first confidence sequence for ATE estimation in bandit settings that does not require the treatment assignment probabilities to be bounded away from zero and one. Thus, the MAD expands the utility of confidence sequences to nearly any bandit algorithm while providing guarantees on the validity and power of the ATE estimation. From a practical perspective, the Mixture Adaptive Design guarantees that an analyst with the goal of stopping her experiment once the confidence sequence suggests a non-zero treatment effect is guaranteed to see statistically significant results in finite time. The condition that $\delta_t = o(\frac{1}{t^{1/4}})$ provides the user great flexibility to tune the experimental design based on the problem specifics. For instance, setting $\delta_t = 1$ would recover a Bernoulli design. We provide recommendations for setting δ_t based on the user's priorities; see Section 3.2. Finally, we show empirically that the MAD generates confidence sequences that shrink quickly around the true ATE and achieves the correct coverage in finite samples without major losses in reward compared to a standard bandit design.

2 Problem Statement

In this section, we formalize the multi-armed bandit problem setting and define the causal estimands of interest. Throughout, we adopt a design-based causal inference framework (Neyman and Iwaszkiewicz 1935; Fisher 1936; Ding et al. 2016); thus, instead of treating the outcomes as random variables, we condition on the set of potential outcomes and treat only the assignment as random. Our estimand of interest is the sample average treatment effect, rather than the population average treatment effect for some hypothetical super-population. This design-based framework allows us to make relatively few assumptions on the outcome distribution of our bandit algorithm, i.e., we do not assume any parametric form on the outcome distribution or that our outcomes are independent and identically distributed (i.i.d.) from some distribution (Dimakopoulou et al. 2018; Zhang et al. 2020; Hadad et al. 2021; Banerjee et al. 2023). In many real-world settings for bandit experiments, we expect that the outcomes will be dependent or non-stationary (Besbes et al. 2014; Allesiardo et al. 2017; Wu et al. 2018). In such settings, incorrectly assuming i.i.d. outcomes can result in effective sample sizes that are much smaller than assumed and cause variance estimation to be highly conservative (Meng 2018).

Assume that we observe a sequence of t units $\{W_i, Y_i\}_{i=1}^t$ where $W_i \in \{0, ..., K-1\}$ and Y_i are the treatment assignment and outcome respectively for unit i. In the MAB literature, $\{W_i, Y_i\}_{i=1}^t$ are analogous to the sequence of actions and rewards commonly notated $\{A_i, R_i\}_{i=1}^t$. Although we will ultimately show that our method applies more broadly, we first assume that we observe a single unit at each time and that we have binary treatment assignments, i.e., $W_i \in \{0, 1\}$. In Sections 3.3 and 3.4, we generalize our results to any $K \geq 2$ number of treatments and the batched bandit setting, respectively.

In our binary treatment setting, each unit i has a pair of potential outcomes $\{Y_i(1), Y_i(0)\}$. In the design-based framework, we implicitly condition on these observed potential outcomes, i.e., we treat $\{Y_i(1), Y_i(0)\}$ as fixed. Since Y_i, our observed response for unit i, is a function of its treatment assignment W_i, we can write

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0).$$

Since we implicitly condition on the potential outcomes $Y_i(1), Y_i(0)$, Y_i is random only via the treatment assignments W_i. Let $\tau_i = Y_i(1) - Y_i(0)$. Then, the Average Treatment Effect (ATE) at time t is:

Definition 2 (Average Treatment Effect (ATE) at time t)**.**

$$\bar{\tau}_t := \frac{1}{t} \sum_{i=1}^{t} \tau_i.$$

Our objective is to generate a confidence sequence for $\bar{\tau}_t$ in MAB experiments with guarantees on statistical power.

3 The Mixture Adaptive Design

3.1 The Mixture Adaptive Design for Binary Treatments

In this section, we formalize our proposed experimental design and state our main result.

As in the MAB literature, we assume that the treatment assignment probabilities at a given time t can depend on previously observed data up to time $t-1$. Formally, let $H_t \coloneqq \{W_i, Y_i\}_{i=1}^{t}$. For any arbitrary adaptive assignment algorithm, even one that does not satisfy probabilistic treatment assignments, define $p_{i|i-1}^{\text{adapt}}(w) \coloneqq \mathbb{P}\left(W_i = w \mid H_{i-1}\right)$. For instance, if the user wanted to use Thompson sampling for their experiment, $p_{i|i-1}^{\text{adapt}}(w)$ would be the assignment probability that Thompson sampling would assign treatment w to unit i, given the history of the $i-1$ previous units. We can now define the Mixture Adaptive Design for the binary treatment setting.

Definition 3 (Mixture Adaptive Design (MAD) for Binary Treatments)**.** *For a real-valued sequence $\delta_i \in (0,1]$, $w \in \{0,1\}$,*

$$p_{i|i-1}^{MAD}(w) \coloneqq \mathbb{P}\left(W_i = 1 \mid \mathcal{F}_{i-1}\right) = \begin{cases} 1/2 & w/p\ \delta_i \\ p_{i|i-1}^{adapt}(w) & w/p\ 1-\delta_i. \end{cases}$$

On a high level, the MAD "mixes" a bandit algorithm with a Bernoulli design. The intuitive idea is that if we can balance these two designs via δ_t, we can gain the ATE precision of a Bernoulli design while maintaining some of the regret minimization of the bandit algorithm.

We now define our ATE estimators. Let $\mathcal{F}_{t,n}$ be the sigma-algebra that contains all pairs of potential outcomes $\{Y_i(1), Y_i(0)\}_{i=1}^{t}$ and all observed data $\{W_i, Y_i\}_{i=1}^{n}$ where $n \leq t$.

Based on the estimator for τ_i proposed in Bojinov and Shephard (2019); Bojinov et al. (2021) for adaptive experiment settings, we set

$$\hat{\tau}_i \coloneqq \frac{\mathbb{1}\{W_i = 1\} Y_i}{p_{i|i-1}^{\text{MAD}}(1)} - \frac{\mathbb{1}\{W_i = 0\} Y_i}{p_{i|i-1}^{\text{MAD}}(0)}, \tag{1}$$

which is an unbiased estimator for τ_i:

$$\mathbb{E}\left[\hat{\tau}_i \mid \mathcal{F}_{i,i-1}\right] = \tau_i.$$

Hence,

$$\frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left[\hat{\tau}_i \mid \mathcal{F}_{i,i-1}\right] = \bar{\tau}_t,$$

and we define our estimator of the Average Treatment effect at time t as:

$$\hat{\bar{\tau}}_t \coloneqq \frac{1}{t}\sum_{i=1}^{t} \hat{\tau}_i,$$

which is an unbiased estimator for $\bar{\tau}_t$. Additionally, we have that

$$Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1}) \leq \sigma_i^2, \text{ where } \sigma_i^2 \coloneqq \frac{Y_i(1)^2}{p^{\text{MAD}}_{i|i-1}(1)} + \frac{Y_i(0)^2}{p^{\text{MAD}}_{i|i-1}(0)}, \tag{2}$$

and thus, a natural unbiased estimator for σ_i^2 is

$$\hat{\sigma}_i^2 \coloneqq \frac{Y_i(1)^2 \mathbb{1}\{W_i = 1\}}{(p^{\text{MAD}}_{t|t-1}(1))^2} + \frac{Y_i(0)^2 \mathbb{1}\{W_i = 0\}}{(p^{\text{MAD}}_{t|t-1}(0))^2}. \tag{3}$$

Equations (2) and (3) follow from Bojinov and Shephard (2019); Bojinov et al. (2021), which propose the estimator of Equation (3) because the closed form does not admit a natural unbiased estimator. Let $S_t \coloneqq \sum_{i=1}^t \sigma_i^2$ and $\hat{S}_t \coloneqq \sum_{i=1}^t \hat{\sigma}_i^2$.

We now state the assumptions necessary for our main results.

Assumption 1 (Bounded (Realized) Potential Outcomes)**.** *There exists $M \in \mathbb{R}$ such that*

$$\limsup_{t \to \infty} |Y_t(w)| \leq M < \infty$$

for all $w \in \mathcal{W}$.

This assumption is used in existing work on CSs for the ATE (Howard et al. 2021; Ham et al. 2022) and commonly assumed in design-based causal inference settings (Bojinov and Shephard 2019; Bojinov et al. 2021; Lei and Ding 2021). Note, this is an assumption on the *realized* potential outcomes. While this assumption may seem limiting, the realized outcomes in any real-world experiment are almost always bounded, as the limitations of computing precision guarantee that any realized outcome collected using existing computing resources will be bounded by, e.g., the highest floating point number via IEEE-754 standards. Hence, even if $Y_t(w)$ were drawn from a Gaussian distribution using existing computing resources, the *realized* potential outcomes will never exceed this upper limit on floating point precision, and therefore we can argue that this assumption is satisfied.

We also require that the average conditional variance of our estimator $\frac{1}{t}\sum_{i=1}^t Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$ is not vanishing. Specifically, we need that the cumulative conditional variances $\sum_{i=1}^t Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$ are growing at at least a linear rate in t. Define $\Omega(t)$ such that if $f(t) = \Omega(t)$, there exists $k > 0$, t_0 such that for all $t \geq t_0$, $f(t) \geq kt$.

Assumption 2 (At Least Linear Rate of Cumulative Conditional Variances)**.**

$$\sum_{i=1}^t Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1}) = \Omega(t).$$

Assumption 2 states that the sums of the conditional variances go to infinity at least as fast as a constant rate. Existing works such as Ham et al. (2022) and Waudby-Smith et al. (2023) require that $\sum_{i=1}^t Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1}) \to \infty$, but do not assume a specific rate. For instance, they allow for the average conditional variance to vanish superlinearly, which would violate Assumption 2. Although Assumption 2 is slightly stronger, this assumption

should be easily satisfied in most realistic bandit settings. For instance, we expect the assumption would be satisfied with a Bernoulli design since all $Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$ would be equal and constant. Intuitively, adaptive assignment should only make $Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$ larger, so as long as we also don't have an adversarial sequence $(Y_i(1), Y_i(0))_{i=1}^{\infty}$ which "cancels out" the rate of these variance terms, this assumption should be satisfied. For instance, if there existed some time t such that beyond t, $Y_i(1), Y_i(0)$ are constantly 0, this assumption may not hold. However, such scenarios would be unusual in practice and/or may indicate practical issues with the experiment. Importantly, this assumption is the only condition we impose on the user-chosen bandit algorithm (the $p^{\text{adapt}}_{i|i-1}(w)$ in $p^{\text{MAD}}_{i|i-1}(w)$), and therefore, we expect our result to be valid for *nearly any bandit algorithm*.

Finally, for $a \geq 0$, define $\delta_t = o(1/t^a)$ to mean that $1/\delta_t = o(t^a)$.

Theorem 1. *Let $\{\hat{\tau}_i\}_{i=1}^{\infty}$ be the sequence of random variables where $W_i = w$ with probability $p^{MAD}_{i|i-1}(w)$ as in Definition 3 with $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 2 hold. Then $(\hat{\bar{\tau}}_t \pm \hat{V}_t)$ where*

$$\hat{V}_t = \sqrt{\frac{2(\hat{S}_t\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{\hat{S}_t\eta^2 + 1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$ asymptotic CS for $\bar{\tau}_t$ and $\hat{V}_t \xrightarrow{a.s.} 0$.

The proof is provided in Appendix A. Intuitively, Theorem 1 holds because a Bernoulli Design provides this guarantee (simply set $\delta_i = 1$ above), so a design that stochastically injects a Bernoulli design into a MAB experiment will maintain the same guarantee as long as the experiment does not deviate from the Bernoulli design too quickly, i.e., the rate at which δ_i approaches 0 is controlled.

At a high level, proving the validity of Theorem 1 proceeds by showing that $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$ along with Assumptions 1 and 2 ensure that the ATE estimator of Equation (1) satisfies a Lindeberg-type uniform integrability (see Lemma A.1 in Appendix A), thus allowing us to apply the universal asymptotic CS of Waudby-Smith et al. (2023) in this setting. The shrinking variance result proceeds by showing that $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$ and Assumption 1 guarantees that $\hat{S}_t \log \hat{S}_t = o(t^2)$ almost surely, and hence, $\hat{V}_t$ is shrinking towards 0 almost surely. Note, this CS is not a function of M in Assumption 1; M is only used to ensure the Lindeberg-type uniform integrability result and to control the asymptotic rate of $\hat{S}_t$. Therefore, M can be set arbitrarily large to satisfy Assumption 1 without affecting the width of our CS.

Practically, Theorem 1 enables analysts to perform bandit experiments using the MAD and, as long as Assumptions 1 and 2 hold, be guaranteed both valid and powerful inference. While Ham et al. (2022) and Howard et al. (2021) propose CSs for the ATE which allow for adaptive assignment, their CS requires probabilistic assignment assumption for their validity guarantees. Both Ham et al. (2022) and Howard et al. (2021) also require Assumption 1, so Theorem 1 removes the probabilistic assignment assumption entirely while only requiring a slightly stronger assumption on the cumulative conditional variances (which should be easily satisfied in most practical settings, as discussed above). Additionally, we prove the novel result that the width of our asymptotic CS is shrinking by deriving

the rate of $\hat{S}_t$ for the MAD. While Ham et al. (2022) discusses how their asymptotic CS scales with the variance of the estimator and state that they expect it to shrink in many adaptive settings, they do not prove that $\hat{S}_t$ has the correct rate to ensure a shrinking CS, even with probabilistic assignments.

As discussed in Ham et al. (2022), analysts at companies such as Netflix often use the first time zero is outside of a CS as a stopping rule for their online experiments. We show that as long as $\bar{\tau}_t$ is truly non-zero in the long term, such a stopping rule will occur in finite time.

Let $T_{MAD} \coloneqq \inf_t \left\{ t : 0 \notin (\hat{\bar{\tau}}_t \pm \hat{V}_t) \right\}$, i.e., this is the first time 0 is not within the confidence sequence specified in Theorem 1.

Theorem 2. *Let $\{\hat{\tau}_i\}_{i=1}^{\infty}$ be the sequence of random variables where $W_i = w$ with probability $p^{MAD}_{i|i-1}(w) = \frac{1}{2}\delta_i + (1-\delta_i)p^{adapt}_{i|i-1}(w)$, $w \in \{0,1\}$, and $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 2 hold. Assume $\bar{\tau}_t \to c$ for some $|c| > 0$ as $i \to \infty$. Then,*

$$\mathbb{P}\left(T_{MAD} < \infty\right) = 1.$$

The proof is provided in Appendix B. As part of the proof for Theorem 2, we show that $\bar{\tau}_t \to c$ implies that $\hat{\bar{\tau}}_t \xrightarrow{a.s.} \bar{\tau}_t$ (and hence, to c) and use the fact that $\hat{V}_t$ is converging almost surely to zero. So intuitively, the CS of Theorem 1 is shrinking around the true ATE, so if c is non-zero, the CS will at some point exclude zero. Note, existing work such as Ham et al. (2022) have no coverage guarantees in general bandit settings. As we exhibit empirically in Section 4, the CS of Ham et al. (2022) can severely undercover if naively applied to a bandit setting, i.e., the percentage of times that the true ATE is included in their CS is far below $1-\alpha$.

The assumption that $\bar{\tau}_t \to c$ is satisfied immediately in stationary outcome settings, e.g., if we assume $\tau_i = c$ for all i or if $\tau_i \overset{i.i.d.}{\sim} \mathcal{D}$ for some distribution $\mathcal{D}$ with finite expectation. In non-stationary settings, this assumption intuitively says that the non-stationarity should be limited (e.g., at the beginning of the experiment) such that the long-term behavior of $\bar{\tau}_t$ stabilizes towards some value.

3.2 Recommendations for Setting δ_i

Theorems 1 and 2 hold for any $\delta_i = o(\frac{1}{i^{1/4}})$, therefore, the MAD provides flexibility to select $\delta_i \in (0,1]$ based on the specifics of the problem. In this section, we provide a few recommendations of δ_i based on the analyst's priorities.

Example 1: $\delta_i = 1/i^{0.24}$; such a δ_i would allow the analyst to aggressively prioritize the bandit design (and hence regret minimization) while still maintaining validity and shrinking width guarantees.

Example 2: $\delta_i = c$ for some $c \in (0,1]$; such a δ_i would maintain a constant rate of priority in the Bernoulli design. Such a δ_i can be valuable in settings where the analyst may want to stop the experiment as soon as possible (e.g., due to an external deadline, limited or expensive resources, etc.) via some stopping rule (e.g. when 0 is outside the interval), and so, would like to generate precise inferential results as quickly as possible while still

allowing for some regret minimization. Note that choosing $c = 1$ reduces to a Bernoulli design.

Example 3: $\delta_i = \max\{\tilde{\delta}_i, c\}$ for some $c \in (0, 1]$, $\tilde{\delta}_i$ a sequence in $(0, 1]$; such a δ_i would prioritize the bandit design up until a point, then maintain a constant rate of priority on the Bernoulli design in the long run, thus never fully prioritizing the bandit over the Bernoulli design. For instance, since analysts often have an upper bound N in mind for the maximum time they would ideally like to run the experiment (e.g., due to an external deadline, limited or expensive resources, etc.), they could set $\tilde{\delta}_i = i^{0.24}$ and $c = 1/N^{0.24}$. Such a δ_i can be valuable in settings where the analyst primarily prioritizes regret minimization, and hence, initially chooses an aggressive $\tilde{\delta}_i$ to see if they can achieve statistically significant results before the N samples/time steps. If such results are not achieved by N samples, then the experimenter will maintain a constant rate of Bernoulli assignment from that point onward to sharpen the inference further.

Thus, analysts can flexibly determine a δ_i that balances the trade-off between regret minimization and statistical power throughout the experiment. Recall from Section 1.2 that most existing works in bandit experimental design prescribe specific algorithms that do not allow for user tuning at all (Kasy and Sautmann 2021) or limitations on when during the experiment the user can tune the exploration (Simchi-Levi and Wang 2023). Although potentially useful as out-of-the-box algorithms, such approaches provide the user less flexibility to tune the algorithm based on their specific priorities or experimental constraints.

3.3 Extension to $K \geq 2$ Treatments

We formalize the Mixture Adaptive Design for $K \geq 2$ treatments. Assume $W_t \in \mathcal{W} = \{0, ..., K-1\}$ for all t, where $K \geq 2$.

Definition 4 (Generalized Mixture Adaptive Design (MAD))**.** *For a real-valued sequence* $\delta_t \in (0, 1]$*, for* $w \in \{0, ..., K-1\}$*,*

$$p^{MAD}_{t|t-1}(w \mid \mathcal{F}_{t-1}) = \begin{cases} 1/K & w/p\ \delta_t \\ p^{adapt}_{t|t-1}(w) & w/p\ 1-\delta_t. \end{cases}$$

For any pair of treatments w, w' where a non-zero treatment effect exists, the analogous result of Theorem 1 holds for the Generalized Mixture Adaptive Design and follows almost exactly from the proof of Theorem 1 for the binary treatment setting; see Appendix C for proof and a full formalization of the problem setting for $K \geq 2$.

In particular, the $K > 2$ case showcases the advantages of our anytime valid guarantees, as the analyst can iteratively exclude "unpromising" treatments from consideration after observing sufficient inferential evidence, thus reducing harm to the experimental units. For example, assume the analyst has $K - 1$ total treatments, $w_1, ..., w_{K-1}$ under consideration and one control, w_0. At each time step t, the analyst can compute the CS of Theorem 1 for each of the $K - 1$ treatment-control pairs (w_j, w_0), $j = 1, .., .K - 1$. Then, if the analyst observes sufficient evidence of strong treatment effects between certain treatments and the control (e.g., the confidence sequence excludes 0 and/or seems to be centered around a non-zero value) while others seem to have weak or possibly no treatment effects, the analyst can

decide to remove these "weak" treatments from consideration and continue the experiment with only the most promising treatments.

3.4 Extension to Batched Bandits

In many applications, it is more practical to update the treatment assignment probabilities of an adaptive/bandit algorithm after observing a batch of experimental units. In this section, we show that an analogous result of Theorem 1 also holds in a batched bandit setting.

First, we formalize the problem setting and Mixture Adaptive Design for batched bandits. Assume we observe a sequence of batches, where for each batch j, we have a (non-random and finite) batch size of B. So, for each batch j, we observe $H_j^{batch} := \left\{W_i^{(j)}, Y_i^{(j)}\right\}_{i=1}^{B}$, where $W_i^{(j)} \in \mathcal{W} = \{0, ..., K-1\}$ and $Y_i^{(j)}$ are the treatment assignment and outcome for unit i in batch j, respectively. We assume the treatment assignment probabilities are fixed within a batch, i.e., the treatment assignment probabilities are only (adaptively) updated after observing a batch of B units.

Let $p_{j|j-1}^{\text{batch-adapt}}(w) := \mathbb{P}(W_i^{(j)} = w \mid H_{j-1}^{batch})$ be the assignment probability to treatment w for unit i in the jth batch for any user-provided batched bandit algorithm. We define the Mixture Adaptive Design for Batched Bandits as:

Definition 5 (Mixture Adaptive Design for Batched Bandits)**.** *For a real-valued sequence $\delta_j \in (0, 1]$, for $w \in \mathcal{W}$, the assignment probability to treatment w in batch j is:*

$$p_{j|j-1}^{B-MAD}(w) = \begin{cases} 1/K & w/p\ \delta_j \\ p_{j|j-1}^{batch\text{-}adapt}(w) & w/p\ 1-\delta_j. \end{cases}$$

The above looks very similar to the standard Generalized MAD. The primary difference is that the MAD assignment probabilities now update after each batch of B units rather than after every individual observation. Hence, in this setting, treatments within a batch are assigned independently, i.e., for each unit i in batch j, we assign treatment by rolling a K-sided dice where each treatment $w \in \mathcal{W}$ has probability $p_{j|j-1}^{B-MAD}(w)$ of appearing, and $p_{j|j-1}^{B-MAD}(w)$ is fixed within the batch j. Let $\{(Y_i^{(j)}(w))_{w\in\mathcal{W}}\}$ be the set of all potential outcome for unit i in batch j. Given a pair of treatments $w, w' \in \mathcal{W}$, we can define the Average Treatment Effect *within* a batch j as:

$$\tau_j^{\text{batch}}(w, w') = \frac{1}{B}\sum_{i=1}^{B} Y_i^{(j)}(w) - Y_i^{(j)}(w'),$$

and the corresponding unbiased estimator as:

$$\hat{\tau}_j^{\text{batch}}(w, w') = \frac{1}{B}\sum_{i=1}^{B} \hat{\tau}_i^{(j)}(w, w').$$

where

$$\hat{\tau}_i^{(j)}(w, w') = \frac{Y_i^{(j)}(w)\mathbb{1}\{W_i^{(j)} = w\}}{p_{j|j-1}^{MAD}(w)} - \frac{Y_i^{(j)}(w')\mathbb{1}\{W_i^{(j)} = w'\}}{p_{j|j-1}^{MAD}(w')}.$$

Thus, our target estimand, the *Batched-Average Treatment Effect* up to batch b, is:

$$\bar{\tau}_b^{\text{batch}}(w, w') = \frac{1}{b}\sum_{j=1}^{b} \tau_j^{\text{batch}}(w, w').$$

and the corresponding unbiased estimator is:

$$\hat{\bar{\tau}}_b^{\text{batch}}(w, w') = \frac{1}{b}\sum_{j=1}^{b} \hat{\tau}_j^{\text{batch}}(w, w').$$

Hence, we define $\mathcal{F}_{b,b-1}^{\text{batch}}$ as the filtration generated by all potential outcomes for all units up to and including the bth batch $\{\{(Y_i^{(j)}(w))_{w\in\mathcal{W}}\}_{i=1}^{B_h}\}_{j=1}^{b}$ and all observed history up to batch $b-1$, $\{H_j^{\text{batch}}\}_{j=1}^{b-1}$.

Then, for each $i = 1, ..., B$, $Var\left(\hat{\tau}_i^{(j)}(w, w') \mid \mathcal{F}_{j,j-1}^{\text{batch}}\right) \leq \sigma_i^{(j)2}(w, w')$ where

$$\sigma_i^{(j)2}(w, w') = \frac{Y_i(w)^2}{p_{j|j-1}^{MAD}(w)} + \frac{Y_i(w')^2}{p_{j|j-1}^{MAD}(w')}.$$

So, $\hat{\sigma}_i^{(j)2}(w, w') = \frac{Y_i(w)^2 \mathbb{1}\{W_i=w\}}{(p_{j|j-1}^{MAD}(w))^2} + \frac{Y_i(w')^2 \mathbb{1}\{W_i=w'\}}{(p_{j|j-1}^{MAD}(w'))^2}$.

Hence,

$$Var\left(\hat{\tau}_j^{\text{batch}}(w, w') \mid \mathcal{F}_{j,j-1}^{\text{batch}}\right) = \frac{1}{B^2}\sum_{i=1}^{B} Var\left(\hat{\tau}_i^{(j)}(w, w') \mid \mathcal{F}_{j,j-1}^{\text{batch}}\right) \leq \frac{1}{B^2}\sum_{i=1}^{B} \sigma_i^{(j)2}(w, w').$$

Hence, we define

$S_b^{\text{batch}}(w, w') := \sum_{j=1}^{b} \frac{1}{B^2} \sum_{i=1}^{B} \sigma_i^{(j)2}(w, w')$ and $\hat{S}_b^{\text{batch}}(w, w') := \sum_{j=1}^{b} \frac{1}{B^2} \sum_{i=1}^{H_j} \hat{\sigma}_i^{(j)2}(w, w')$.

Finally, we state the analogous assumptions as Assumptions 1 and 2 for the batched bandit setting.

Assumption 3 (Bounded Potential Outcomes for Batched Bandits)**.** *There exists $M \in \mathbb{R}$ such that*

$$|Y_i^{(j)}(w)| \leq M < \infty$$

for all $j, i \in \mathbb{N}^+, w \in \mathcal{W}$.

Assumption 4 (At Least Linear Rate of Cumulative Conditional Variances for Batched Bandits)**.** *For all $w, w' \in \mathcal{W}$,*

$$\sum_{j=1}^{b} Var\left(\hat{\tau}_j^{batch}(w, w') \mid \mathcal{F}_{j,j-1}^{batch}\right) = \Omega(b).$$

Theorem 3. *For $w, w' \in \mathcal{W}$, let $\{\hat{\tau}_j^{batch}(w, w')\}_{j=1}^{\infty}$ be the sequence of random variables where $W_i^{(j)} = w$ with probability $p_{j|j-1}^{B-MAD}(w) = \frac{1}{K}\delta_j + (1-\delta_j)p_{j|j-1}^{adapt}(w)$, $w \in \mathcal{W}$, and*

$\delta_j \in (0,1]$ *such that* $\delta_j = o\left(\frac{1}{j^{1/4}}\right)$. *Assume Assumptions 3 and 4 hold. Then* $(\hat{\bar{\tau}}_b^{batch}(w,w') \pm \hat{V}_b^{batch}(w,w'))$ *where*

$$\hat{V}_b^{batch}(w,w') := \sqrt{\frac{2(\hat{S}_b^{batch}\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{\hat{S}_b^{batch}\eta^2+1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$ *asymptotic CS for* $\bar{\tau}_b^{batch}$ *and* $\hat{V}_b^{batch}(w,w') \xrightarrow{a.s.} 0$.

For $w, w' \in \mathcal{W}$, , define $T_{\text{B-MAD}}(w,w') := \inf_t \left\{b : 0 \notin (\hat{\bar{\tau}}_b^{\text{batch}}(w,w') \pm V_b^{\text{batch}}(w,w'))\right\}$.

Theorem 4. *For* $w, w' \in \mathcal{W}$, *let* $\{\hat{\tau}_j^{batch}(w,w')\}_{j=1}^{\infty}$ *be the sequence of random variables where* $W_i^{(j)} = w$ *with probability* $p_{j|j-1}^{B-MAD}(w) = \frac{1}{K}\delta_j + (1-\delta_j)p_{j|j-1}^{adapt}(w)$, $w \in \mathcal{W}$, *and* $\delta_j \in (0,1]$ *such that* $\delta_j = o\left(\frac{1}{j^{1/4}}\right)$. *Assume Assumptions 3 and 4 hold. Then, if* $\hat{\bar{\tau}}_b^{batch}(w,w') \to c$ *as* $b \to \infty$ *for some* $|c| > 0$,

$$\mathbb{P}\left(T_{B-MAD}(w,w') < \infty\right) = 1.$$

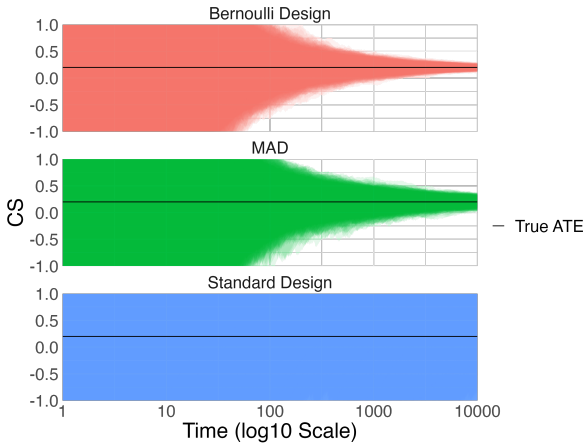
Hence, Theorems 3 and 4 show that the MAD provides the same validity, shrinking width, and stopping time guarantees in the batched bandit setting.

4 Simulation Study

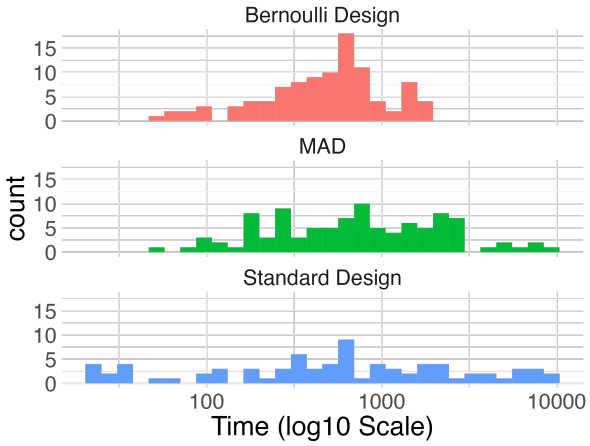
Empirically, we find that MAD improves both the precision and efficiency of ATE inference in bandit experiments. Consider a two-arm bandit with $Y_i(1) \overset{i.i.d.}{\sim} Bern(0.8)$ and $Y_i(0) \overset{i.i.d.}{\sim} Bern(0.6)$, so $\bar{\tau}_t = 0.2$ for all t. We implemented a Thompson sampler with uniform priors on both arms and computed the CS of Theorem 1 for the *Mixture Adaptive Design with* $\delta_i = \frac{1}{i^{0.24}}$. As baselines, we implement the same CS under a *Bernoulli design* and under a *Standard design*, which naively applies the CS of Ham et al. (2022) to the bandit setting. Note, since we make no adjustment to the assignment probabilities, there are no validity and coverage guarantees for the Standard design. We set $\alpha = 0.05$. At each time step, the Thompson sampler pulls a single arm and observes the outcome. Confidence sequences are computed over $T = 10,000$ time steps for 100 random seeds. We set $\eta = \sqrt{\frac{-2\log(\alpha)+\log(-2\log(\alpha)+1)}{t^*}} \approx 0.028$ for $t^* = 10,000$; we choose this value following the recommendations in Appendix B.2 of Waudby-Smith et al. (2023) on setting η to optimize the width of the CS for a specific time t^*, where we take t^* to be the end of the time horizon we set for this experiment.

As shown in Figure 1a, both Bernoulli and MAD confidence sequences are contained within $[-1,1]$ by ≈ 100 samples while 59% of the Standard design confidence sequences still included $[-1,1]$ at $10,000$ samples. In Figure 1b, all Bernoulli and MAD confidence sequences excluded zero by $10,000$ samples, while 22% of Standard design confidence sequences still included zero by $10,000$ samples. Figure 2a shows that the MAD has the correct coverage in finite samples, while the Standard design can severely undercover. Figure 2b shows that the time-averaged reward under the MAD approaches that of the Standard design and is

generally only slightly smaller at any given point in time. Thus, the MAD empirically makes notable gains with respect to coverage, statistical power, and early stopping without a significant loss of finite-sample reward. We repeat this experiment varying the true ATE and find similar results; see Appendix E for details.

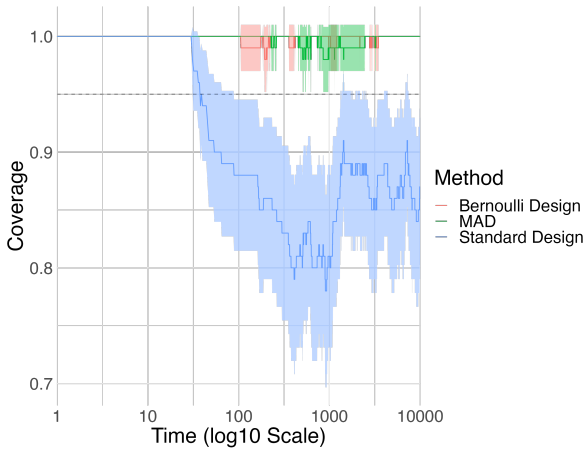


(a) Confidence sequences plotted over 10^5 samples/timesteps across 100 random replicates. 59% of the Standard design confidence sequences still included $[-1, 1]$ at $10,000$ samples.

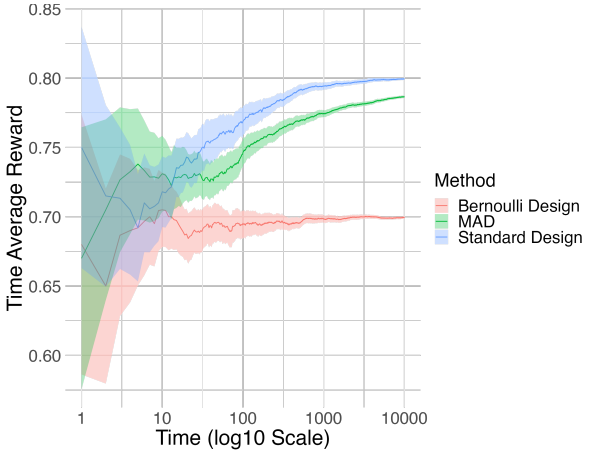


(b) Histogram of the first time zero was excluded from the CS across 100 random replicates. 22% of Standard design confidence sequences still included zero by $10,000$ samples, and hence, we did not observe a stopping time.

Figure 1



(a) Coverage of confidence sequences across 100 random replicates. Error bars represent 2SEs. The dashed line represents $1 - \alpha = 0.95$.



(b) Time averaged reward of the confidence sequences across 100 random replicates. Error bars represent 2SEs.

Figure 2

In the setting where all $Y_i(w) \in [0, 1]$, $w = 0, 1$ and there exists some p_{min} such that all treatment assignment probabilities are bounded within $[p_{min}, 1 - p_{min}]$, Corollary 2 of Howard et al. (2021) provides a nonasymptotic CS for the ATE with width shrinking towards 0. In the same Thompson sampler setting as above, we explore the performance of the MAD

when using this CS, setting $\tilde{\delta}_i = \max\{\frac{1}{i^{0.24}}, \frac{1}{100}\}$ to ensure the p_{min} condition is satisfied. Practically, we can interpret $\tilde{\delta}_i$ as imposing our adaptive algorithm to assign treatment according to a Bernoulli design every 1 out of 100 samples in the long run. By Corollary 2 of Howard et al. (2021), the confidence sequences produced are exactly valid and guaranteed to be shrinking over time. Hence, using the MAD with this CS provides an approach for performing nonasymptotic continuous inference on the ATE. As baselines, we also implement a Bernoulli design with this CS, and the analogous "Standard" design, which naively calculates this CS for the bandit experiment without any adjustment to the assignment probabilities. Since this CS requires p_{min} as input to the CS, we use $p_{min} = 1/10,000$ as a conservative lower bound. [1] However, since no true p_{min} exists in this setting, the validity of the CS is not guaranteed for the standard bandit setting. We find that the MAD improves the width of this CS while achieving comparable reward to the standard bandit design; see Appendix E for details. The MAD and Bernoulli designs produce confidence sequences that begin to shrink around 1000 samples while none of the confidence sequences generated using the Standard bandit design shrunk below $[-1, 1]$. Hence, even as a heuristic, using this CS on a standard bandit experiment is rather impractical. Though this CS is non-asymptotic, it tends to be much wider than the asymptotic CS of Theorem 1. For instance, even the confidence sequences produced using the Bernoulli design only excluded 0 once out of the 100 random replicates; see Appendix E for details. In practice, we recommend using the asymptotic CS of Theorem 1 as it empirically achieves the correct coverage and produces thinner confidence sequences. The script used to reproduce all simulation results is provided at https://github.com/biyonka/mixture_adaptive_design.

5 Conclusion

The Mixture Adaptive Design (MAD) provides a framework for anytime-valid design-based causal inference on the ATE in MAB experiments with guarantees on statistical validity and power. By controlling the rate of a bandit algorithm's adaptive assignment probabilities via "mixing" with a Bernoulli design, the MAD provides a confidence sequence for the ATE that is asymptotically valid and guaranteed to shrink around the true ATE under mild conditions and thus holds for nearly any choice of bandit algorithm. As existing approaches do not guarantee validity in general bandit settings, the MAD provides a practical tool for online and anytime valid causal inference in both standard and batched bandit settings.

6 Acknowledgements

We thank Nathan Cheng for his helpful discussions and feedback.

References

Agrawal, S. and Goyal, N. (2012), Analysis of thompson sampling for the multi-armed bandit problem, *in* "Conference on learning theory", JMLR Workshop and Conference

[1] Since we expect Thompson sampling to achieve a regret of $O(\log(T))$ asymptotically (Agrawal and Goyal 2012), the share of units that get assigned to the suboptimal treatment is on the order of $\log(T)/T$, so we use $1/T$ as a conservative, heuristic lower bound.

Proceedings, pp. 39–1.

Allesiardo, R., Féraud, R. and Maillard, O.-A. (2017), "The non-stationary stochastic multi-armed bandit problem", *International Journal of Data Science and Analytics* **3**, 267–283.

Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002), "Finite-time analysis of the multiarmed bandit problem", *Machine learning* **47**, 235–256.

Banerjee, D., Ghosh, A., Chowdhury, S. R. and Gopalan, A. (2023), "Exploration in linear bandits with rich action sets and its implications for inference".

Besbes, O., Gur, Y. and Zeevi, A. (2014), Stochastic multi-armed-bandit problem with non-stationary rewards, *in* Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Weinberger, eds, "Advances in Neural Information Processing Systems", Vol. 27, Curran Associates, Inc.

Bojinov, I. and Gupta, S. (2022), "Online Experimentation: Benefits, Operational and Methodological Challenges, and Scaling Guide", *Harvard Data Science Review* **4**(3). https://hdsr.mitpress.mit.edu/pub/aj31wj81.

Bojinov, I., Rambachan, A. and Shephard, N. (2021), "Panel experiments and dynamic causal effects: A finite population perspective".

Bojinov, I. and Shephard, N. (2019), "Time series experiments and causal estimands: Exact randomization tests and trading", *Journal of the American Statistical Association* **114**(528), 1665–1682.

Böttiger, Y., Laine, K., Andersson, M. L., Korhonen, T., Molin, B., Ovesjö, M.-L., Tirkkonen, T., Rane, A., Gustafsson, L. L. and Eiermann, B. (2009), "Sfinx—a drug-drug interaction database designed for clinical decision support systems", *European journal of clinical pharmacology* **65**, 627–633.

Csörgő, M. (1968), "On the strong law of large numbers and the central limit theorem for martingales", *Transactions of the American Mathematical Society* **131**(1), 259–275.

Dimakopoulou, M., Ren, Z. and Zhou, Z. (2021), "Online multi-armed bandits with adaptive inference".

Dimakopoulou, M., Zhou, Z., Athey, S. and Imbens, G. (2018), "Balanced linear contextual bandits".

Ding, K., Li, J. and Liu, H. (2019), Interactive anomaly detection on attributed networks, *in* "Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining", WSDM '19, Association for Computing Machinery, New York, NY, USA, p. 357–365.

Ding, P., Feller, A. and Miratrix, L. (2016), "Randomization inference for treatment effect variation", *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78**(3), 655–671.

Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D. and Pineau, J. (2018), Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis, *in* F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace and J. Wiens, eds, "Proceedings of the 3rd Machine Learning for Healthcare Conference", Vol. 85 of *Proceedings of Machine Learning Research*, PMLR, pp. 67–82.

Fisher, R. A. (1936), "Design of experiments", *British Medical Journal* **1**(3923), 554.

Garivier, A. and Cappé, O. (2011), The kl-ucb algorithm for bounded stochastic bandits and beyond, *in* "Proceedings of the 24th annual conference on learning theory", JMLR Workshop and Conference Proceedings, pp. 359–376.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S. and Athey, S. (2021), "Confidence intervals for policy evaluation in adaptive experiments", *Proceedings of the national academy of sciences* **118**(15), e2014602118.

Hahn, J., Hirano, K. and Karlan, D. (2011), "Adaptive experimental design using the propensity score", *Journal of Business & Economic Statistics* **29**(1), 96–108.

Ham, D. W., Bojinov, I., Lindon, M. and Tingley, M. (2022), "Design-based confidence sequences for anytime-valid causal inference".

Horvitz, D. G. and Thompson, D. J. (1952), "A generalization of sampling without replacement from a finite universe", *Journal of the American statistical Association* **47**(260), 663–685.

Howard, S. R. and Ramdas, A. (2022), "Sequential estimation of quantiles with applications to a/b testing and best-arm identification", *Bernoulli* **28**(3).

Howard, S. R., Ramdas, A., McAuliffe, J. and Sekhon, J. (2021), "Time-uniform, nonparametric, nonasymptotic confidence sequences", *The Annals of Statistics* **49**(2).

Kasy, M. and Sautmann, A. (2021), "Adaptive treatment assignment in experiments for policy choice", *Econometrica* **89**(1), 113–132.

Kaufmann, E. (2018), "On bayesian index policies for sequential resource allocation", *The Annals of Statistics* **46**(2), 842–865.

Kohavi, R., Tang, D. and Xu, Y. (2020), *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*, Cambridge University Press.

Lai, T. L. and Robbins, H. (1985), "Asymptotically efficient adaptive allocation rules", *Advances in applied mathematics* **6**(1), 4–22.

Lei, L. and Ding, P. (2021), "Regression adjustment in completely randomized experiments with a diverging number of covariates", *Biometrika* **108**(4), 815–828.

Mary, J., Gaudel, R. and Preux, P. (2015), Bandits and recommender systems, *in* "Machine Learning, Optimization, and Big Data: First International Workshop, MOD 2015, Taormina, Sicily, Italy, July 21-23, 2015, Revised Selected Papers 1", Springer, pp. 325–336.

Ménard, P. and Garivier, A. (2017), A minimax and asymptotically optimal algorithm for stochastic bandits, *in* "International Conference on Algorithmic Learning Theory", PMLR, pp. 223–237.

Meng, X.-L. (2018), "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election", *The Annals of Applied Statistics* **12**(2), 685–726.

Moe, N. B., Aurum, A. and Dybå, T. (2012), "Challenges of shared decision-making: A multiple case study of agile software development", *Information and Software Technology* **54**(8), 853–865. Special Issue: Voice of the Editorial Board.

Neyman, J. and Iwaszkiewicz, K. (1935), "Statistical problems in agricultural experimentation", *Supplement to the Journal of the Royal Statistical Society* **2**(2), 107–180.

Russo, D., Roy, B. V., Kazerouni, A., Osband, I. and Wen, Z. (2020), "A tutorial on thompson sampling".

Simchi-Levi, D. and Wang, C. (2023), Multi-armed bandit experimental design: Online decision-making and adaptive inference, *in* "International Conference on Artificial Intelligence and Statistics", PMLR, pp. 3086–3097.

Villar, S. S., Bowden, J. and Wason, J. (2015), "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.

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H. and Ramdas, A. (2023), "Time-uniform central limit theory and asymptotic confidence sequences".

Waudby-Smith, I. and Ramdas, A. (2020), "Estimating means of bounded random variables by betting", *arXiv preprint arXiv:2010.09686* .

Wu, Q., Iyer, N. and Wang, H. (2018), Learning contextual bandits in a non-stationary environment, *in* "The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval", SIGIR '18, Association for Computing Machinery, New York, NY, USA, p. 495–504.

Xu, M., Qin, T. and Liu, T.-Y. (2013), Estimation bias in multi-armed bandit algorithms for search advertising, *in* C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds, "Advances in Neural Information Processing Systems", Vol. 26, Curran Associates, Inc.

Zhang, K., Janson, L. and Murphy, S. (2020), Inference for batched bandits, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin, eds, "Advances in Neural Information Processing Systems", Vol. 33, Curran Associates, Inc., pp. 9818–9829.

Zhang, K., Janson, L. and Murphy, S. (2021), Statistical inference with m-estimators on adaptively collected data, *in* M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan, eds, "Advances in Neural Information Processing Systems", Vol. 34, Curran Associates, Inc., pp. 7460–7471.

SUPPLEMENTARY MATERIAL

A Proof of Theorem 1

To prove Theorem 1, we must first show that $(\hat{\tau}_i)_{i=1}^{\infty}$ satisfies a Lindeberg-type uniform integrability condition (Condition L-2 of Waudby-Smith et al. (2023)).

Recall, $\hat{\tau}_i := \frac{\mathbb{1}\{W_i=1\}Y_i}{p_{i|i-1}(1)} - \frac{\mathbb{1}\{W_i=0\}Y_i}{p_{i|i-1}(0)}$, $p^{\text{adapt}}_{t|t-1}(w) := \mathbb{P}\left(W_t = w \mid \mathcal{F}_{t-1}\right)$, and $p^{\text{MAD}}_{i|i-1}(w) = \delta_i \frac{1}{2} + (1-\delta_i)p^{\text{adapt}}_{i|i-1}(w)$ for $w = 0, 1$. Recall, we write $\delta_t = o(1/t^a)$ to mean that $1/\delta_t = o(i^a)$.

Lemma A.1. *Let $\{\hat{\tau}_i\}_{i=1}^{\infty}$ be a sequence of random variables where $W_i = w$ with probability $p^{MAD}_{i|i-1}(w) = \frac{1}{2}\delta_i + (1-\delta_i)p^{adapt}_{i|i-1}(w)$ for $w \in \{0,1\}$ and $\delta_i \in (0,1]$ such that $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 2 hold. Then, $\{\hat{\tau}_i\}_{i=1}^{\infty}$ satisfies a Lindeberg-type uniform integrability condition, i.e., then there exists $\kappa \in (0,1)$ such that*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\left[(\hat{\tau}_t - \tau_t)^2 \mathbb{1}\{(\hat{\tau}_t - \tau_t)^2 > (B_t)^{\kappa}\}\right]}{(B_t)^{\kappa}} < \infty \ a.s.$$

where $B_t = \sum_{i=1}^{t} Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$.

Proof. By Assumption 1,

$$|\hat{\tau}_t| = \left| \frac{Y_t(1)\mathbb{1}\{W_t = 1\}}{p^{\text{MAD}}_{t|t-1}(1)} - \frac{Y_t(0)\mathbb{1}\{W_t = 0\}}{p^{\text{MAD}}_{t|t-1}(0)} \right| \leq \frac{2M}{\min(p^{\text{MAD}}_{t|t-1}(0), p^{\text{MAD}}_{t|t-1}(1))}$$

and $|\tau_t| \leq 2M$ for all t.

Hence,

$$\begin{aligned}(\hat{\tau}_{t|t-1} - \tau_t)^2 &\leq \left(\frac{2M}{\min(p^{\text{MAD}}_{t|t-1}(0), p^{\text{MAD}}_{t|t-1}(1))}\right)^2 + (2M)^2 + 2\left(\frac{2M}{\min(p^{\text{MAD}}_{t|t-1}(0), p^{\text{MAD}}_{t|t-1}(1))}\right)2M \\ &\leq \frac{24M^2}{(\min(p^{\text{MAD}}_{t|t-1}(0), p^{\text{MAD}}_{t|t-1}(1)))^2}\end{aligned}$$

First, note that for all w, $p^{\text{MAD}}_{t|t-1}(w) \geq \delta_t(1/2)$ and so $\frac{1}{p^{\text{MAD}}_{t|t-1}(w)} = o(t^{1/4})$. So, $(\hat{\tau}_{t|t-1} - \tau_t)^2 = o(t^{2a})$ almost surely.

By Assumption 2, $B_t = \Omega(t)$. Therefore, $B_t^{\kappa} = \Omega(t^{\kappa})$. Set $\kappa > 2a$. Then, there exists some $\tilde{t}$ such that for all $t \geq \tilde{t}$, $B_t^{\kappa} > (\hat{\tau}_{t|t-1} - \tau_t)^2$ a.s.. Hence, for all $t \geq \tilde{t}$, $\mathbb{1}\{(\hat{\tau}_{t|t-1} - \tau_t)^2 > (B_t)^{\kappa}\} = 0$, and so,

$$\sum_{i=1}^{t} \frac{\mathbb{E}\left[(\hat{\tau}_{t|t-1} - \tau_t)^2 \mathbb{1}\{(\hat{\tau}_{t|t-1} - \tau_t)^2 > (B_t)^{\kappa}\}\right]}{(B_t)^{\kappa}} < \infty \text{ a.s..}$$

□

Theorem 1. *Let $\{\hat{\tau}_i\}_{i=1}^{\infty}$ be the sequence of random variables where $W_i = w$ with probability $p^{MAD}_{i|i-1}(w)$ as in Definition 3 with $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 2 hold. Then $(\hat{\bar{\tau}}_t \pm \hat{V}_t)$ where*

$$\hat{V}_t = \sqrt{\frac{2(\hat{S}_t\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{\hat{S}_t\eta^2 + 1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$ asymptotic CS for $\bar{\tau}_t$ and $\hat{V}_t \xrightarrow{a.s.} 0$.

Proof. By Assumptions 1 and 2 imply that Lemma A.1 holds. Lemma A.1 and Assumption 2 satisfy Conditions L-1 and L-2 of Theorem 2.5 in Waudby-Smith et al. (2023), so, by Steps 1 and 2 of the proof of Theorem 2.5 in Waudby-Smith et al. (2023),

$(\hat{\bar{\tau}}_t \pm V^*_t)$ where

$$V^*_t = \sqrt{\frac{2(B_t\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{B_t\eta^2 + 1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$asymptotic CS and $B_t = \sum_{i=1}^{t} Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1})$.

As noted in Equation (2),

$$Var(\hat{\tau}_i \mid \mathcal{F}_{i,i-1}) \leq \sigma_i^2, \text{ where } \sigma_i^2 := \frac{Y_i(1)^2}{p_{i|i-1}(1)} + \frac{Y_i(0)^2}{p_{i|i-1}(0)}. \tag{4}$$

Hence, $(\hat{\bar{\tau}}_t \pm \tilde{V}_t)$ where

$$\tilde{V}_t = \sqrt{\frac{2(S_t\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{S_t\eta^2 + 1}}{\alpha}\right)}$$

is still a valid $(1-\alpha)$asymptotic CS where recall, $S_t = \sum_{i=1}^{t} \sigma_i^2$. This holds because replacing B_t with S_t only makes the CS wider since $S_t \geq B_t$.

As noted in Equation (3), an unbiased estimator for σ_i^2 is

$$\hat{\sigma}_i^2 := \frac{Y_i(1)^2 \mathbb{1}\{W_i = 1\}}{p^2_{i|i-1}(1)} + \frac{Y_i(0)^2 \mathbb{1}\{W_i = 0\}}{p^2_{i|i-1}(0)}. \tag{5}$$

and we define $\hat{S}_t = \sum_{i=1}^{t} \hat{\sigma}_i^2$.

To establish the validity of the CS in Theorem 1, we must show that $\frac{1}{t}\hat{S}_t - \frac{1}{t}S_t \xrightarrow{a.s.} 0$. Then, Condition L-3 of Theorem 2.5 of Waudby-Smith et al. (2023) is satisfied and we can conclude that $(\hat{\bar{\tau}}_t \pm \hat{V}_t)$ where

$$\hat{V}_t = \sqrt{\frac{2(\hat{S}_t\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{\hat{S}_t\eta^2 + 1}}{\alpha}\right)}$$

is still a valid $(1-\alpha)$ asymptotic CS.

First, note that

$$\begin{aligned}
(\hat{\sigma}_i^2)^2 &\leq \left(\frac{M^2}{p^{MAD}_{i|i-1}(1)^2} + \frac{M^2}{p^{MAD}_{i|i-1}(0)^2} \right)^2 \\
&= \frac{M^4}{p^{MAD}_{i|i-1}(1)^4} + \frac{M^4}{p^{MAD}_{i|i-1}(0)^4} + 2\frac{M^2}{p^{MAD}_{i|i-1}(1)^2 p^{MAD}_{i|i-1}(0)^2} \\
&\leq \frac{M^4}{(\delta_i/2)^4} + \frac{M^4}{(\delta_i/2)^4} + 2\frac{M^2}{(\delta_i/2)^4} \\
&= M^4 \left(2^4 \frac{1}{\delta_i^4} + 2^4 \frac{1}{\delta_i^4} + 2^5 \frac{1}{\delta_i^4} \right) \\
&= o(i)
\end{aligned}$$

where the last line follows because $\frac{1}{\delta_i} = o\left(i^{1/4}\right)$. Define $X_i = \hat{\sigma}_i^2 - \sigma_i^2$ and note that X_i is a martingale difference sequence. Hence,

$$\begin{aligned}
\mathbb{E}[X_i] &= E[(\hat{\sigma}_i^2)^2] - (\sigma_i^2)^2 \\
&\leq E[(\hat{\sigma}_i^2)^2] \\
&= o(i)
\end{aligned}$$

So, $\frac{\mathbb{E}[X_i^2]}{i^2} = \frac{o(i)}{i^2}$ and hence, $\sum_{i=1}^{\infty} \frac{\mathbb{E}[X_i^2]}{i^2} < \infty$. For instance, if $\delta_i = \frac{1}{i^a}$ for some $0 \leq a < 1/4$, then $\frac{\mathbb{E}[X_i^2]}{i^2} \leq C i^{4a-2}$ for some constant $C < \infty$, and since $4a - 2 < -1$, $\sum_{i=1}^{\infty} \frac{\mathbb{E}[X_i^2]}{i^2} \leq C \sum_{i=1}^{\infty} i^{4a-2} < \infty$ by the p-series test.

Then, by the SLLN for martingale difference sequences (Theorem 1 of (Csörgő 1968)), we can conclude that

$$\frac{1}{t} \sum_{i=1}^{t} X_i \overset{a.s.}{\to} 0.$$

Hence, Condition L-3 of Waudby-Smith et al. (2023) is satisfied and by Step 3 of the proof for Theorem 2.5 of Waudby-Smith et al. (2023), we can conclude that $(\hat{\tau}_t \pm \hat{V}_t)$ where

$$\hat{V}_t = \sqrt{\frac{2(\hat{S}_t \eta^2 + 1)}{t^2 \eta^2} \log\left(\frac{\sqrt{\hat{S}_t \eta^2 + 1}}{\alpha} \right)}$$

is still a valid $(1-\alpha)$ asymptotic CS.

To show $\hat{V}_t \overset{a.s.}{\to} 0$, note that

$$\begin{aligned}
\hat{\sigma}_i^2 &\leq \frac{M^2}{p_{i|i-1}^{MAD}(1)^2} + \frac{M^2}{p_{i|i-1}^{MAD}(0)^2} \\
&\leq \frac{M^2}{(\delta_i/2)^2} + \frac{M^2}{(\delta_i/2)^2} \\
&= \frac{4M^2}{\delta_i^2} + \frac{4M^2}{\delta_i^2} \\
&= 8M^2 \frac{1}{\delta_i^2} \\
&= 8M^2 o(i^{1/2})
\end{aligned}$$

So, $\hat{S}_t \leq 8M^2 \sum_{i=1}^t o(i^{1/2})$ a.s..

So, there exists positive real numbers N and x_0 such that for all $t \geq x_0$, $\hat{S}_t \leq N \sum_{i=1}^t i^{1/2} < N \sum_{i=1}^t t^{1/2} = N t^{1+1/2}$, and hence, we can conclude that $\hat{S}_t = O(t^{1+\frac{1}{2}})$.

We can show that $\log\left(\hat{S}_t\right) = o(t^{1/2})$ a.s.. For $t \geq x_0$,

$$\frac{\log\left(\hat{S}_t\right)}{t^{1/2}} \leq \frac{\log\left(N t^{3/2}\right)}{t^{1/2}} = \frac{\log(N) + (3/2)\log(t))}{t^{1/2}}$$

and $\frac{\log(N)+(3/2)\log(t))}{t^{1/2}} \to 0$ as $t \to \infty$. Therefore, $\log\left(\hat{S}_t\right) = o(t^{1/2})$ a.s. and $\hat{S}_t \log\left(\hat{S}_t\right) = o(t^2)$ a.s.. Hence, $\hat{V}_t = \frac{2(\hat{S}_t\eta^2+1)}{t^2\eta^2} \log\left(\frac{\sqrt{\hat{S}_t\eta^2+1}}{\alpha}\right) = o(1)$ a.s..

Note, if $1/\delta_i$ was increasing at a rate faster than $i^{1/4}$ asymptotically, we are not guaranteed that $\hat{V}_t = o(1)$ a.s.. □

B Proof of Theorem 2

Theorem 2. *Let $\{\hat{\tau}_i\}_{i=1}^{\infty}$ be the sequence of random variables where $W_i = w$ with probability $p_{i|i-1}^{MAD}(w) = \frac{1}{2}\delta_i + (1-\delta_i)p_{i|i-1}^{adapt}(w)$, $w \in \{0,1\}$, and $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 2 hold. Assume $\bar{\tau}_t \to c$ for some $|c| > 0$ as $i \to \infty$. Then,*

$$\mathbb{P}\left(T_{MAD} < \infty\right) = 1.$$

Proof. We will first show that $\frac{1}{t}\sum_{i=1}^t \hat{\tau}_i \overset{a.s.}{\to} c$. Let $u_i \coloneqq \hat{\tau}_i - \tau_i$.

Note that

$$\begin{aligned}
\mathbb{E}[u_i^2] &= \mathbb{E}[\hat{\tau}_i^2] - \tau_i^2 \\
&\leq \mathbb{E}[\hat{\tau}_i^2]
\end{aligned}$$

where

$$\begin{aligned}\hat{\tau}_i^2 &= \left(\frac{Y_i(1)\mathbb{1}\{W_i=1\}}{p_{i|i-1}^{MAD}(1)} - \frac{Y_i(0)\mathbb{1}\{W_i=0\}}{p_{i|i-1}^{MAD}(0)}\right)^2 \\ &\leq \frac{Y_i(1)^2}{(\delta_i/2)^2} + \frac{Y_i(0)^2}{(\delta_i/2)^2} \\ &= 8M^2\frac{1}{\delta_i^2} \\ &= 8M^2 o(i^{1/2})\end{aligned}$$

Hence,

$$\sum_{i=1}^{t}\frac{\mathbb{E}[u_i^2]}{i^2} \leq \sum_{i=1}^{t}\frac{8M^2 o(i^{1/2})}{i^2} < \infty$$

by the p-series test.

Hence, by the SLLN for martingale difference sequences (Theorem 1 of (Csörgő 1968)), we can conclude that

$$\frac{1}{t}\sum_{i=1}^{t} u_i \overset{a.s.}{\to} 0.$$

By Theorem 1, we have that $\hat{V}_t \overset{a.s.}{\to} 0$.

Therefore, applying Slutsky's Theorem, we conclude that

$$\hat{\bar{\tau}}_t + \hat{V}_t \overset{a.s.}{\to} c,$$

and

$$\hat{\bar{\tau}}_t - \hat{V}_t \overset{a.s.}{\to} c.$$

So,

$$\begin{aligned}&\mathbb{P}\left(0 \notin \left(\hat{\bar{\tau}}_t \pm \hat{V}_t\right)\right) \\ &= \mathbb{P}\left(0 < \hat{\bar{\tau}}_t - \hat{V}_t\right) + \mathbb{P}\left(0 > \hat{\bar{\tau}}_t + \hat{V}_t\right) \\ &\to 1,\end{aligned}$$

where the last line follows because $\mathbb{P}\left(0 > \hat{\bar{\tau}}_t + \hat{V}_t\right) \to 0$ and $\mathbb{P}\left(0 < \hat{\bar{\tau}}_t - \hat{V}_t\right) \to 1$.

Let $p_t = \mathbb{P}\left(0 \in (\hat{\bar{\tau}}_t \pm \hat{V}_t)\right)$. Since $p_t \to 0$ as $t \to \infty$, there exists a subsequence t_k such that $p_{t_k} \leq \frac{1}{k^2}$, for all $k \geq 1$. Let A_{t_k} be the event $\{0 \in (\frac{1}{t_k}\sum_{i=1}^{t_k}\hat{\tau}_i \pm \hat{V}_{t_k})\}$. Then, $\sum_{k=1}^{\infty}\mathbb{P}(A_{t_k}) < \infty$, and by the Borel-Cantelli lemma,

$$\mathbb{P}\left(\limsup_{k\to\infty} A_{t_k}\right) = 0.$$

Hence,

$$\mathbb{P}\left(T_{MAD} = \infty\right) \leq \mathbb{P}\left(\limsup_{k\to\infty} A_{t_k}\right) = 0.$$

□

C ATE Inference for $K \geq 2$ Treatments

We first formalize the problem setting for $K \geq 2$ treatments and prove more general versions of Theorems 1 and Theorem 2 for $K \geq 2$ treatments. Assume $W_t \in \mathcal{W} \coloneqq \{0, ..., K-1\}$.

Let $\mathcal{F}_{t,n}$ be the sigma-algebra that contains all potential outcomes $\{(Y_i(w))_{w\in\mathcal{W}}\}_{i=1}^{t}$ and all observed data $\{W_i, Y_i\}_{i=1}^{n}$ where $n \leq t$.

As before, let the assignment probabilities for any user-chosen adaptive algorithm be denoted as $p^{\text{adapt}}_{i|i-1}(w) = \mathbb{P}(W_t = w \mid H_{t-1})$ where $H_{t-1} = \{W_i, Y_i\}_{i=1}^{t-1}$. Hence, the Generalized Mixture Adaptive Design has assignment probabilities $p^{\text{MAD}}_{i|i-1}(w) = \delta_i \frac{1}{K} + (1-\delta_i)p^{\text{adapt}}_{i|i-1}(w)$, for all $w \in \mathcal{W}$.

For any pair of treatments $w, w' \in \mathcal{W}$, let $\tau_i(w, w') = Y_i(w) - Y_i(w')$ and define the Average Treatment Effect (ATE) between w and w' up to t as $\bar{\tau}_t(w, w') \coloneqq \frac{1}{t}\sum_{i=1}^{t} \tau_i(w, w')$. So, using the Generalized Mixture Adaptive Design, our corresponding estimator for the ATE is:

$$\hat{\bar{\tau}}_t(w, w') = \frac{1}{t}\sum_{i=1}^{t} \hat{\tau}_i(w, w'),$$

where $\hat{\tau}_i(w, w') \coloneqq \frac{\mathbb{1}\{W_i=w\}Y_i(w)}{p^{\text{MAD}}_{t|t-1}(w)} - \frac{\mathbb{1}\{W_i=w'\}Y_i(w')}{p^{MAD}_{i|i-1}(w')}$. We also have the corresponding upper bound on the variance:

$$Var(\hat{\tau}_i(w, w') \mid \mathcal{F}_{i,i-1}) \leq \sigma_i^2(w, w'), \text{ where } \sigma_i^2(w, w') \coloneqq \frac{Y_i(w)^2}{p^{\text{MAD}}_{t|t-1}(w)} + \frac{Y_i(w')^2}{p^{MAD}_{i|i-1}(w')},$$

and the analogous unbiased estimator of $\sigma_i^2(w, w')$:

$$\hat{\sigma}_i^2(w, w') \coloneqq \frac{Y_i(w)^2\mathbb{1}\{W_i = w\}}{(p^{\text{MAD}}_{t|t-1}(w))^2} + \frac{Y_i(w')^2\mathbb{1}\{W_i = w'\}}{(p^{MAD}_{i|i-1}(w'))^2}.$$

Finally, let $S_t(w, w') \coloneqq \sum_{i=1}^{t} \sigma_i^2(w, w')$ and $\hat{S}_t(w, w') \coloneqq \sum_{i=1}^{t} \hat{\sigma}_i^2(w, w')$.

Define $T_{MAD}(w, w') \coloneqq \inf_t \left\{t : 0 \notin (\hat{\bar{\tau}}_t(w, w') \pm \hat{V}_t(w, w'))\right\}$ where the treatment assignments are generated via the Generalized Mixture Adaptive Design in Definition (4).

Though an analogous result of Theorem 1 for this setting follows almost directly from the fact that we still have $1/\delta_t = o(t^{1/4})$, we provide a full statement of the result and its proof here for completeness.

For precision, we need to state our cumulative conditional variance condition for the $K \geq 2$ setting.

Assumption 5 (At Least Linear Rate of Cumulative Conditional Variances for All Pairs of Treatments)**.** *For all* $w, w' \in \mathcal{W}$, $\sum_{i=1}^{t} Var(\hat{\tau}_i(w,w') \mid \mathcal{F}_{i,i-1}) = \Omega(t)$.

Lemma A.2. *Let* $w, w' \in \mathcal{W}$. *Let* $\{\hat{\tau}_i(w,w')\}_{i=1}^{\infty}$ *be a sequence of random variables where* $W_i = w$ *with probability* $p_{i|i-1}^{MAD}(w) = \frac{1}{K}\delta_i + (1-\delta_i)p_{i|i-1}^{adapt}(w)$ *for* $w \in \mathcal{W}$ *and* $\delta_i \in (0,1]$ *such that* $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. *Assume Assumptions 1 and 5 hold. Then,* $\{\hat{\tau}_i(w,w')\}_{i=1}^{\infty}$ *satisfies the Lindeberg-type uniform integrability condition of Waudby-Smith et al. (2023), i.e., there exists* $\kappa \in (0,1)$ *such that*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\left[(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 \mathbb{1}\{(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 > (B_t(w,w'))^{\kappa}\}\right]}{(B_t(w,w'))^{\kappa}} < \infty \ a.s.$$

where $B_t(w,w') = \sum_{i=1}^{t} Var(\hat{\tau}_i(w,w') \mid \mathcal{F}_{i,i-1})$.

Proof. By Assumption 1, for all t, w, w',

$$|\hat{\tau}_t(w,w')| = \left| \frac{Y_t(w)\mathbb{1}\{W_t = w\}}{p_{t|t-1}^{\mathrm{MAD}}(w)} - \frac{Y_t(w')\mathbb{1}\{W_t = w'\}}{p_{t|t-1}^{\mathrm{MAD}}(w')} \right| \leq \frac{2M}{\min(p_{t|t-1}^{\mathrm{MAD}}(w'), p_{t|t-1}^{\mathrm{MAD}}(w))}$$

and $|\tau_t(w,w')| \leq 2M$.

Hence,

$$\begin{aligned}
&(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 \\
&\leq \left(\frac{2M}{\min(p_{t|t-1}^{\mathrm{MAD}}(w), p_{t|t-1}^{\mathrm{MAD}}(w'))} \right)^2 + (2M)^2 + 2\left(\frac{2M}{\min(p_{t|t-1}^{\mathrm{MAD}}(w), p_{t|t-1}^{\mathrm{MAD}}(w'))} \right) 2M \\
&\leq \frac{24M^2}{(\min(p_{t|t-1}^{\mathrm{MAD}}(w), p_{t|t-1}^{\mathrm{MAD}}(w')))^2}.
\end{aligned}$$

First, note that for all $w \in \mathcal{W}$, $p_{t|t-1}^{\mathrm{MAD}}(w) \geq \delta_t(1/K)$ and so $\frac{1}{p_{t|t-1}^{\mathrm{MAD}}(w)} = o(t^{1/4})$. So, $(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 = o(t^{1/2})$ almost surely.

By Assumption 2, $B_t(w,w') = \Omega(t)$. Therefore, $B_t^{\kappa}(w,w') = \Omega(t^{\kappa})$. Set $\kappa > 1/2$. Then, there exists some $\tilde{t}$ such that for all $t \geq \tilde{t}$, $B_t^{\kappa}(w,w') > (\hat{\tau}_t(w,w') - \tau_t(w,w'))^2$ a.s.. Hence, for all $t \geq \tilde{t}$, $\mathbb{1}\{(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 > (B_t)^{\kappa}(w,w')\} = 0$, and so,

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\left[(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 \mathbb{1}\{(\hat{\tau}_t(w,w') - \tau_t(w,w'))^2 > (B_t(w,w'))^{\kappa}\}\right]}{(B_t(w,w'))^{\kappa}} < \infty \text{ a.s.}$$

□

Theorem 1*. *For* $w, w' \in \mathcal{W}$, *let* $\{\hat{\tau}_i(w,w')\}_{i=1}^{\infty}$ *be the sequence of random variables where* $W_i = w$ *with probability* $p_{i|i-1}^{MAD}(w) = \frac{1}{K}\delta_i + (1-\delta_i)p_{i|i-1}^{adapt}(w)$, *and* $\delta_i \in (0,1]$ *such that* $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. *Assume Assumptions 1 and 5 hold. Then* $(\hat{\bar{\tau}}_t(w,w') \pm \hat{V}_t(w,w'))$ *where*

$$\hat{V}_t(w,w') = \sqrt{\frac{2(\hat{S}_t(w,w')\eta^2 + 1)}{t^2\eta^2} \log\left(\frac{\sqrt{\hat{S}_t(w,w')\eta^2 + 1}}{\alpha} \right)}$$

is a valid $(1-\alpha)$ *asymptotic CS for* $\bar{\tau}_t(w,w')$ *and* $\hat{V}_t(w,w') \xrightarrow{a.s.} 0$.

Proof. By Assumptions 1 and 5 imply that Lemma A.2 holds. Lemma A.2 and Assumption 5 satisfy Conditions L-1 and L-2 of Theorem 2.5 in Waudby-Smith et al. (2023), so, by Steps 1 and 2 of the proof of Theorem 2.5 in Waudby-Smith et al. (2023),

$(\hat{\bar{\tau}}_t(w,w') \pm V_t^*(w,w'))$ where

$$V_t^*(w,w') = \sqrt{\frac{2(B_t(w,w')\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{B_t(w,w')\eta^2+1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$asymptotic CS and $B_t(w,w') = \sum_{i=1}^t Var(\hat{\tau}_i(w,w') \mid \mathcal{F}_{i,i-1})$.

As noted in Equation (2),

$$Var(\hat{\tau}_i(w,w') \mid \mathcal{F}_{i,i-1}) \leq \sigma_i^2, \text{ where } \sigma_i^2 \coloneqq \frac{Y_i(w)^2}{p^{\text{MAD}}_{t|t-1}(w)} + \frac{Y_i(w')^2}{p^{MAD}_{i|i-1}(w')}. \tag{6}$$

Hence, $(\hat{\bar{\tau}}_t(w,w') \pm \tilde{V}_t(w,w'))$ where

$$\tilde{V}_t(w,w') = \sqrt{\frac{2(S_t(w,w')\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{S_t(w,w')\eta^2+1}}{\alpha}\right)}$$

is still a valid $(1-\alpha)$asymptotic CS where recall, $S_t(w,w') = \sum_{i=1}^t \sigma_i^2(w,w')$. This holds because replacing $B_t(w,w')$ with $S_t(w,w')$ only makes the CS wider since $S_t(w,w') \geq B_t(w,w')$.

As noted in Equation (3), an unbiased estimator for $\sigma_i^2(w,w')$ is

$$\hat{\sigma}_i^2(w,w') \coloneqq \frac{Y_i(w)^2\mathbb{1}\{W_i=w\}}{(p^{\text{MAD}}_{t|t-1}(w))^2} + \frac{Y_i(w')^2\mathbb{1}\{W_i=w'\}}{(p^{MAD}_{i|i-1}(w'))^2}. \tag{7}$$

and we define $\hat{S}_t(w,w') = \sum_{i=1}^t \hat{\sigma}_i^2(w,w')$.

To establish the validity of the CS in Theorem 1, we must show that $\frac{1}{t}\hat{S}_t(w,w') - \frac{1}{t}S_t(w,w') \xrightarrow{a.s.} 0$. Then, Condition L-3 of Theorem 2.5 of Waudby-Smith et al. (2023) is satisfied and we can conclude that $(\hat{\bar{\tau}}_t(w,w') \pm \hat{V}_t(w,w'))$ where

$$\hat{V}_t(w,w') = \sqrt{\frac{2(\hat{S}_t(w,w')\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{\hat{S}_t(w,w')\eta^2+1}}{\alpha}\right)}$$

is still a valid $(1-\alpha)$ asymptotic CS.

First, note that

$$\begin{aligned}(\hat{\sigma}_i^2(w,w'))^2 &\leq \left(\frac{M^2}{p_{i|i-1}^{MAD}(w)^2}+\frac{M^2}{p_{i|i-1}^{MAD}(w')^2}\right)^2 \\ &= \frac{M^4}{p_{i|i-1}^{MAD}(w)^4}+\frac{M^4}{p_{i|i-1}^{MAD}(w')^4}+2\frac{M^2}{p_{i|i-1}^{MAD}(w)^2 p_{i|i-1}^{MAD}(w')^2} \\ &\leq \frac{M^4}{(\delta_i/K)^4}+\frac{M^4}{(\delta_i/K)^4}+2\frac{M^2}{(\delta_i/K)^4} \\ &= M^4\left(K^4\frac{1}{\delta_i^4}+K^4\frac{1}{\delta_i^4}+K^5\frac{1}{\delta_i^4}\right) \\ &= o(i)\end{aligned}$$

where the last line follows because $\frac{1}{\delta_i} = o\left(i^{1/4}\right)$ as shown in Lemma A.1. Define $X_i(w,w') = \hat{\sigma}_i^2(w,w') - \sigma_i^2(w,w')$ and note that $X_i(w,w')$ is a martingale difference sequence. Hence,

$$\begin{aligned}\mathbb{E}[X_i(w,w')] &= E[(\hat{\sigma}_i(w,w')^2] - (\sigma_i(w,w')^2 \\ &\leq E[(\hat{\sigma}_i(w,w')^2] \\ &= o(i)\end{aligned}$$

So, $\frac{\mathbb{E}[X_i^2(w,w')]}{i^2} = \frac{o(i)}{i^2}$ and hence, $\sum_{i=1}^{\infty}\frac{\mathbb{E}[X_i^2(w,w')]}{i^2} < \infty$. For instance, if $\delta_i = \frac{1}{i^a}$ for some $0 \leq a < 1/4$, then $\frac{\mathbb{E}[X_i^2(w,w')]}{i^2} \leq Ci^{4a-2}$ for some constant $C < \infty$, and since $4a-2 < -1$, $\sum_{i=1}^{\infty}\frac{\mathbb{E}[X_i^2(w,w')]}{i^2} \leq C\sum_{i=1}^{\infty} i^{4a-2} < \infty$ by the p-series test.

Then, by the SLLN for martingale difference sequences (Theorem 1 of (Csörgő 1968)), we can conclude that

$$\frac{1}{t}\sum_{i=1}^{t} X_i(w,w') \overset{a.s.}{\to} 0.$$

Hence, Condition L-3 of Waudby-Smith et al. (2023) is satisfied and by Step 3 of the proof for Theorem 2.5 of Waudby-Smith et al. (2023), we can conclude that $(\hat{\tilde{\tau}}_t(w,w') \pm \hat{V}_t(w,w'))$ where

$$\hat{V}_t(w,w') = \sqrt{\frac{2(\hat{S}_t(w,w')\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{\hat{S}_t(w,w')\eta^2+1}}{\alpha}\right)}$$

is still a valid $(1-\alpha)$ asymptotic CS.

To show $\hat{V}_t(w,w') \xrightarrow{a.s.} 0$, note that

$$\begin{aligned}
\hat{\sigma}_i^2 &\leq \frac{M^2}{p_{i|i-1}^{MAD}(w)^2} + \frac{M^2}{p_{i|i-1}^{MAD}(w')^2} \\
&\leq \frac{M^2}{(\delta_i/K)^2} + \frac{M^2}{(\delta_i/K)^2} \\
&= \frac{K^2M^2}{\delta_i^2} + \frac{K^2M^2}{\delta_i^2} \\
&= 2K^2M^2\frac{1}{\delta_i^2} \\
&= 2K^2M^2o(i^{1/2})
\end{aligned}$$

So, $\hat{S}_t(w,w') \leq 8M^2\sum_{i=1}^t o(i^{1/2})$ a.s..

So, there exists positive real numbers N and x_0 such that for all $t \geq x_0$, $\hat{S}_t(w,w') \leq N\sum_{i=1}^t i^{1/2} < N\sum_{i=1}^t t^{1/2} = Nt^{1+1/2}$, and hence, we can conclude that $\hat{S}_t(w,w') = O(t^{1+\frac{1}{2}})$.

We can show that $\log\left(\hat{S}_t(w,w')\right) = o(t^{1/2})$ a.s.. For $t \geq x_0$,

$$\frac{\log\left(\hat{S}_t(w,w')\right)}{t^{1/2}} \leq \frac{\log\left(Nt^{3/2}\right)}{t^{1/2}} = \frac{\log(N) + (3/2)\log(t))}{t^{1/2}}$$

and $\frac{\log(N)+(3/2)\log(t))}{t^{1/2}} \to 0$ as $t \to \infty$. Therefore, $\log\left(\hat{S}_t(w,w')\right) = o(t^{1/2})$ a.s. and $\hat{S}_t(w,w')\log\left(\hat{S}_t(w,w')\right) = o(t^2)$ a.s.. Hence, $\hat{V}_t(w,w') = \frac{2(\hat{S}_t(w,w')\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{\hat{S}_t(w,w')\eta^2+1}}{\alpha}\right) = o(1)$ a.s..

Note, if $1/\delta_i$ was increasing at a rate faster than $i^{1/4}$ asymptotically, e.g., $\delta_i = O(i^{1/2})$, we are not guaranteed that $\hat{V}_t(w,w') = o(1)$ a.s.. □

Theorem 2*. *For $w,w' \in \mathcal{W}$, let $\{\hat{\tau}_i(w,w')\}_{i=1}^\infty$ be the sequence of random variables where $W_i = w$ with probability $p_{i|i-1}^{MAD}(w) = \frac{1}{2}\delta_i + (1-\delta_i)p_{i|i-1}^{adapt}(w)$ for $w \in \mathcal{W}$ and $\delta_i = o\left(\frac{1}{i^{1/4}}\right)$. Assume Assumptions 1 and 5 hold. Assume $\bar{\tau}_i(w,w') \to c$ for some $|c| > 0$ as $i \to \infty$. Then,*

$$\mathbb{P}\left(T_{MAD}(w,w') < \infty\right) = 1.$$

Proof. We will first show that $\frac{1}{t}\sum_{i=1}^t \hat{\tau}_i(w,w') \xrightarrow{a.s.} c$. Let $u_i(w,w') \coloneqq \hat{\tau}_i(w,w') - \tau_i(w,w')$.

Note that

$$\begin{aligned}
\mathbb{E}[u_i(w,w')^2] &= \mathbb{E}[\hat{\tau}_i(w,w')^2] - \tau_i(w,w')^2 \\
&\leq \mathbb{E}[\hat{\tau}_i(w,w')^2]
\end{aligned}$$

where

$$\begin{aligned}\hat{\tau}_i(w,w')^2 &= \left(\frac{Y_i(w)\mathbb{1}\{W_i=w\}}{p^{MAD}_{i|i-1}(w)} - \frac{Y_i(w')\mathbb{1}\{W_i=w'\}}{p^{MAD}_{i|i-1}(w')}\right)^2 \\ &\leq \frac{Y_i(w)^2}{(\delta_i/K)^2} + \frac{Y_i(w')^2}{(\delta_i/K)^2} \\ &= 2K^2M^2\frac{1}{\delta_i^2} \\ &= 2K^2M^2o(i^{1/2})\end{aligned}$$

Hence,

$$\sum_{i=1}^{t}\frac{\mathbb{E}[u_i(w,w')^2]}{i^2} \leq \sum_{i=1}^{t}\frac{8M^2o(i^{1/2})}{i^2} < \infty$$

by the p-series test. Hence, by the SLLN for martingale difference sequences (Theorem 1 of (Csörgő 1968)), we can conclude that

$$\frac{1}{t}\sum_{i=1}^{t}u_i(w,w') \overset{a.s.}{\to} 0.$$

By Theorem 1*, we have that $\hat{V}_t(w,w') \overset{a.s.}{\to} 0$.

Therefore, applying Slutsky's Theorem, we conclude that

$$\hat{\bar{\tau}}_t(w,w') + \hat{V}_t(w,w') \overset{a.s.}{\to} c,$$

and

$$\hat{\bar{\tau}}_t(w,w') - \hat{V}_t(w,w') \overset{a.s.}{\to} c.$$

So,

$$\begin{aligned}&\mathbb{P}\left(0 \notin \left(\hat{\bar{\tau}}_t(w,w') \pm \hat{V}_t(w,w')\right)\right) \\ &= \mathbb{P}\left(0 < \hat{\bar{\tau}}_t(w,w') - \hat{V}_t(w,w')\right) + \mathbb{P}\left(0 > \hat{\bar{\tau}}_t(w,w') + \hat{V}_t(w,w')\right) \\ &\to 1,\end{aligned}$$

where the last line follows because $\mathbb{P}\left(0 > \hat{\bar{\tau}}_t(w,w') + \hat{V}_t(w,w')\right) \to 0$ and

$\mathbb{P}\left(0 < \hat{\bar{\tau}}_t(w,w') - \hat{V}_t(w,w')\right) \to 1.$

Let $p_t = \mathbb{P}\left(0 \in (\hat{\bar{\tau}}_t(w,w') \pm \hat{V}_t(w,w'))\right)$. Since $p_t \to 0$ as $t \to \infty$, there exists a subsequence t_k such that $p_{t_k} \leq \frac{1}{k^2}$, for all $k \geq 1$. Let A_{t_k} be the event $\{0 \in (\frac{1}{t_k}\sum_{i=1}^{t_k}\hat{\tau}_i(w,w') \pm \hat{V}_{t_k}(w,w'))\}$. Then, $\sum_{k=1}^{\infty}\mathbb{P}(A_{t_k}) < \infty$, and by the Borel-Cantelli lemma,

$$\mathbb{P}\left(\limsup_{k\to\infty} A_{t_k}\right) = 0.$$

Hence,

$$\mathbb{P}\left(\tilde{T}_{MAD}(w,w')=\infty\right) \leq \mathbb{P}\left(\limsup_{k\to\infty} A_{t_k}\right) = 0.$$

□

D ATE inference for Batched Bandits

We restate Theorems 3 and 4 here.

Theorem 3. *For $w, w' \in \mathcal{W}$, let $\{\hat{\tau}_j^{batch}(w,w')\}_{j=1}^{\infty}$ be the sequence of random variables where $W_i^{(j)} = w$ with probability $p_{j|j-1}^{B-MAD}(w) = \frac{1}{K}\delta_j + (1-\delta_j)p_{j|j-1}^{adapt}(w)$, $w \in \mathcal{W}$, and $\delta_j \in (0,1]$ such that $\delta_j = o\left(\frac{1}{j^{1/4}}\right)$. Assume Assumptions 3 and 4 hold. Then $(\hat{\bar{\tau}}_b^{batch}(w,w') \pm \hat{V}_b^{batch}(w,w'))$ where*

$$\hat{V}_b^{batch}(w,w') \coloneqq \sqrt{\frac{2(\hat{S}_b^{batch}\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{\hat{S}_b^{batch}\eta^2+1}}{\alpha}\right)}$$

is a valid $(1-\alpha)$ asymptotic CS for $\bar{\tau}_b^{batch}$ and $\hat{V}_b^{batch}(w,w') \xrightarrow{a.s.} 0$.

Proof. Note, the above result is equivalent to Theorem 1* except we replace $\hat{\tau}_i(w,w')$ with $\hat{\tau}_j^{\text{batch}}(w,w')$, $B_t(w,w')$ with $B_b^{\text{batch}}(w,w') = \sum_{j=1}^{b} Var\left(\hat{\tau}_j^{\text{batch}}(w,w') \mid \mathcal{F}_{j,j-1}^{\text{batch}}\right)$, and $\hat{S}_t(w,w')$ and $S_t(w,w')$ with $S_b^{\text{batch}}(w,w') \coloneqq \sum_{j=1}^{b}\frac{1}{B^2}\sum_{i=1}^{B}\sigma_i^{(j)2(w,w')}$ and $\hat{S}_b^{\text{batch}}(w,w') \coloneqq \sum_{j=1}^{b}\frac{1}{B^2}\sum_{i=1}^{H_j}\hat{\sigma}_i^{(j)2(w,w')}$ respectively. Assumption 4 ensures $\sum_{j=1}^{b} Var\left(\hat{\tau}_j^{\text{batch}}(w,w') \mid \mathcal{F}_{j,j-1}^{\text{batch}}\right) = \Omega(b)$ and, because $p_{j|j-1}^{B-MAD}(w) \geq \frac{1}{K}\delta_j$, we have that $(\hat{\tau}_j^{\text{batch}}(w,w') - \tau_j^{\text{batch}}(w,w'))^2 = o(j^{1/2})$ and following the steps of the proof of Lemma A.2, we establish that $\{\hat{\tau}_j^{\text{batch}}(w,w')\}_{j=1}^{\infty}$ satisfies the Lindeberg-type uniform integrability condition of Waudby-Smith et al. (2023), i.e., there exists $\kappa \in (0,1)$ such that

$$\sum_{b=1}^{\infty}\frac{\mathbb{E}\left[(\hat{\tau}_b^{\text{batch}}(w,w')-\tau_b^{\text{batch}}(w,w'))^2\mathbb{1}\{(\hat{\tau}_b^{\text{batch}}(w,w')-\tau_b^{\text{batch}}(w,w'))^2 > (B_b^{\text{batch}}(w,w'))^{\kappa}\}\right]}{(B_b^{\text{batch}}(w,w'))^{\kappa}} < \infty \text{ a.s.}$$

Hence, the remainder of the proof follows directly from Theorem 1*, replacing the corresponding terms for the batched bandit setting, because the fact that $p_{j|j-1}^{B-MAD}(w) \geq \frac{1}{K}\delta_j$ ensures that all analogous terms have the same rates as in Theorem 1*. □

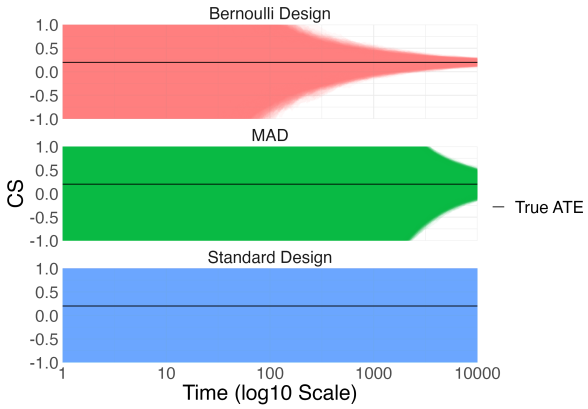
For $w, w' \in \mathcal{W}$, , define $T_{\text{B-MAD}}(w,w') \coloneqq \inf_t\left\{b : 0 \notin (\hat{\bar{\tau}}_b^{\text{batch}}(w,w') \pm V_b^{\text{batch}}(w,w'))\right\}$.

Theorem 4. *For $w, w' \in \mathcal{W}$, let $\{\hat{\tau}_j^{batch}(w,w')\}_{j=1}^{\infty}$ be the sequence of random variables where $W_i^{(j)} = w$ with probability $p_{j|j-1}^{B-MAD}(w) = \frac{1}{K}\delta_j + (1-\delta_j)p_{j|j-1}^{adapt}(w)$, $w \in \mathcal{W}$, and $\delta_j \in (0,1]$ such that $\delta_j = o\left(\frac{1}{j^{1/4}}\right)$. Assume Assumptions 3 and 4 hold. Then, if $\hat{\bar{\tau}}_b^{batch}(w,w') \to c$ as $b \to \infty$ for some $|c| > 0$,*

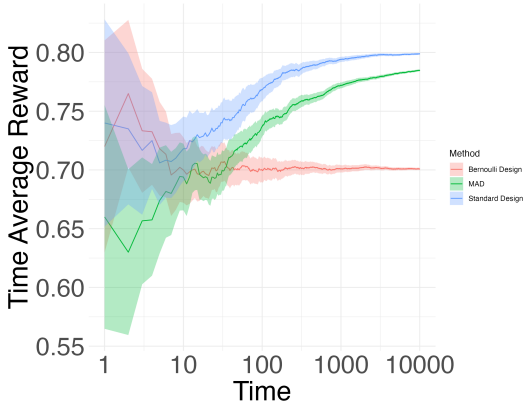
$$\mathbb{P}\left(T_{B-MAD}(w,w') < \infty\right) = 1.$$

Proof. Having established Theorem 3, the remainder of the proof follows from Theorem 2*, replacing the corresponding terms for the batched bandit setting, because the fact that $p^{B-MAD}_{j|j-1}(w) \geq \frac{1}{K}\delta_j$ ensures that all analogous terms have the same rates as in Theorem 2*. □

E Additional Simulation Results

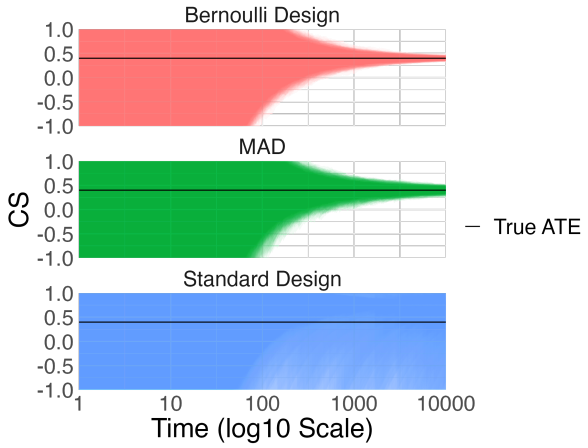


(a) Confidence sequences of Howard et al. (2021) plotted over 10^5 samples across 100 random replicates with $\alpha = 0.05$ and $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.6)$.



(b) Time averaged reward of the confidence sequences of Howard et al. (2021) plotted across 100 random replicates with $\alpha = 0.05$ and $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.6)$.

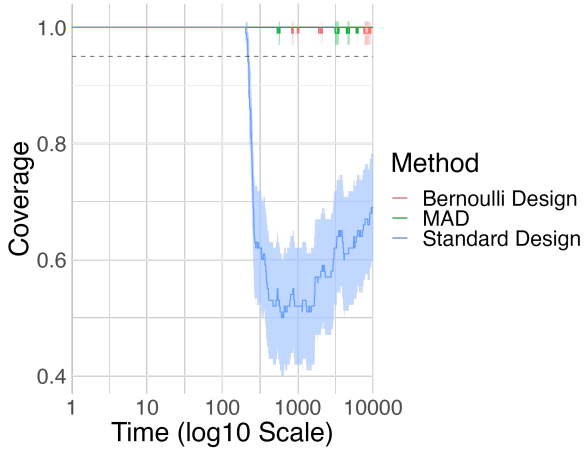
Figure 3



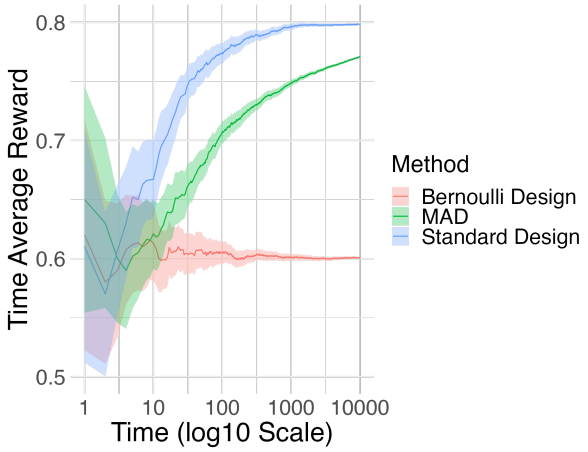
(a) Confidence sequences from Theorem 1 plotted over 10^5 samples across 100 random replicates with $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.4)$.

(b) Histogram of the first time zero was excluded from the CSs of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.4)$. In this setting, all experiments stopped for all designs.

Figure 4

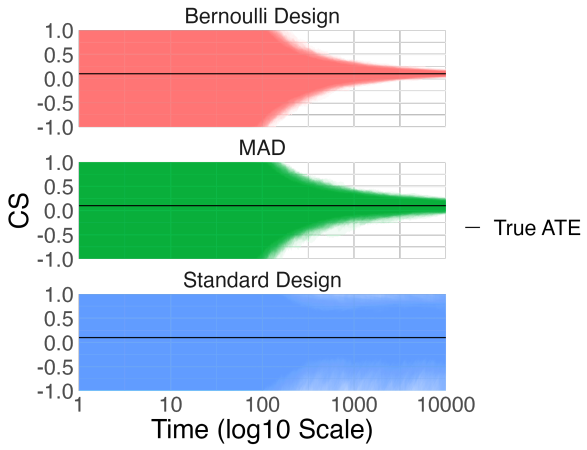


(a) Coverage of confidence sequences of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.4)$. Error bars represent 2SEs. The dashed line represents $1 - \alpha = 0.95$.

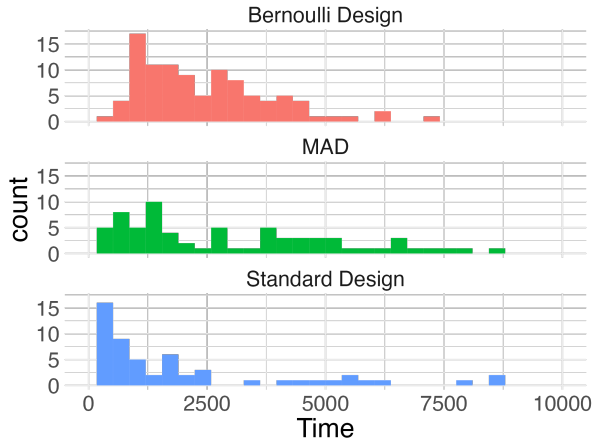


(b) Time averaged reward of the confidence sequences of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.4)$. Error bars represent 2SEs.

Figure 5

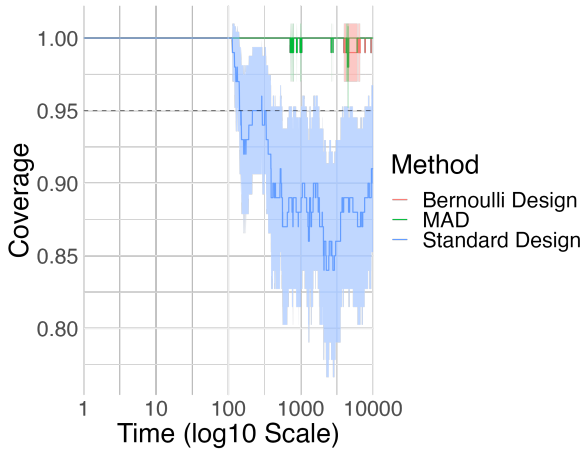


(a) Confidence sequences from Theorem 1 plotted over 10^5 samples across 100 random replicates with $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.7)$.

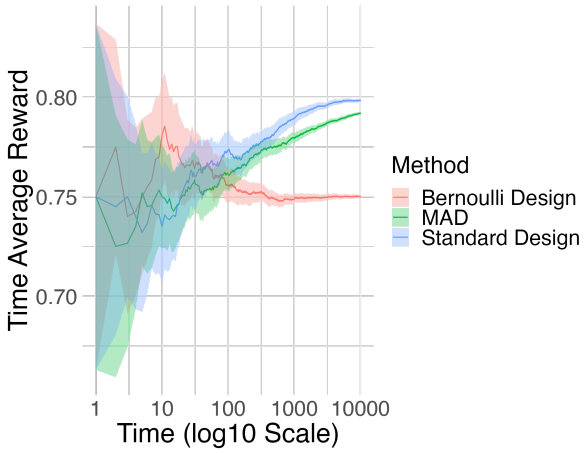


(b) Histogram of the first time zero was excluded from the CSs of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.7)$. 35% of Standard design CSs did not stop, 29% of MAD CSs did not stop, and all Bernoulli design CSs stopped.

Figure 6



(a) Coverage of confidence sequences of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.7)$. Error bars represent 2SEs. The dashed line represents $1-\alpha = 0.95$. Note, that compared to the other experiments where the ATE is larger, the coverage of the Standard approach is closer to the desired level. Since the bandit is not as able to easily distinguish between the two arms in this setting, intuitively, we would expect it to draw more often from the suboptimal arm and thus, have a more stable IPW estimator.



(b) Time averaged reward of the confidence sequences of Theorem 1 across 100 random replicates, $\alpha = 0.05$, $Y_i(1) \sim \text{Bern}(0.8)$, $Y_i(0) \sim \text{Bern}(0.7)$. Error bars represent 2SEs.

Figure 7

Supplementary References