Design-Based Inference for Multi-arm Bandits

Dae Woong Ham
Iavor Bojinov
Michael Lindon
Martin Tingley

Harvard Business School

Design-Based Inference for Multi-arm Bandits

Dae Woong Ham
Harvard University

Iavor Bojinov
Harvard Business School

Michael Lindon
Netflix

Martin Tingley
Netflix

Design-Based Inference for Multi-arm Bandits

Dae Woong Ham[*1], Iavor Bojinov[2], Michael Lindon[3], and Martin Tingley[3]

[1]*Department of Statistics, Harvard University*, [2]*Harvard School of Business*,
[3]*Netflix, U.S.A*

March 1, 2023

Abstract

Multi-arm bandits are gaining popularity as they enable real-world sequential decision-making across application areas, including clinical trials, recommender systems, and online decision-making. Consequently, there is an increased desire to use the available adaptively collected datasets to distinguish whether one arm was more effective than the other, e.g., which product or treatment was more effective. Unfortunately, existing tools fail to provide valid inference when data is collected adaptively or require many untestable and technical assumptions, e.g., stationarity, *iid* rewards, bounded random variables, etc. Our paper introduces the design-based approach to inference for multi-arm bandits, where we condition the full set of potential outcomes and perform inference on the obtained sample. Our paper constructs valid confidence intervals for both the reward mean of any arm and the mean reward difference between any arms in an assumption-light manner, allowing the rewards to be arbitrarily distributed, non-*iid*, and from non-stationary distributions. In addition to confidence intervals, we also provide valid design-based confidence *sequences*, sequences of confidence intervals that have uniform type-1 error guarantees over time. Confidence sequences allow the agent to perform a hypothesis test as the data arrives sequentially and stop the experiment as soon as the agent is satisfied with the inference, e.g., the mean reward of an arm is statistically significantly higher than a desired threshold.

Keywords: Adaptive inference, Sequential inference, Confidence sequences, Finite-population inference, Non-stationary bandits

---

[*]Email: daewoongham@g.harvard.edu

1 Introduction

Multi-armed bandit (MAB) algorithms are a popular and well-established framework for sequential decision-making Robbins (1952); Berry and Fristedt (1985); Bubeck and Nicolò (2012); Sutton and Barto (2018); Slivkins (2019). Because of their regret-minimizing properties, MABs are used across various applications such as optimal treatment allocation in clinical trials Durand et al. (2018), recommender systems improvements Mary, Gaudel and Preux (2015), anomaly detection for networks Ding, Li and Liu (2019), and aiding online decision-making Dimakopoulou, Ren and Zhou (2021). For example, Durand et al. uses an adaptive bandit algorithm to sequentially assign treatments against skin cancer in a mice experiment. Here, the authors care both about assigning the best possible treatment and estimating the relative effectiveness of the different treatments.

As the clinical trial example illustrates, real-world applications of MABs increasingly require inference (i.e., determining if one of the arms is significantly better than the alternatives). Unfortunately, inference is challenging because MAB algorithms collect data adaptively, breaking the standard independent and identically distributed (*iid*) assumption often evoked in the statistics literature. To make progress, researchers have focused on specific settings, such as linear bandits where the rewards follow a parametric linear model (Deshpande et al., 2017; Dimakopoulou et al., 2019; Banerjee et al., 2022), or have made strong untestable assumptions on both the rewards and the action space, such as stationarity and *iid* rewards (Zhang, Janson and Murphy, 2020; Hadad et al., 2021; Waudby-Smith et al., 2022a).

Broadly, there are two approaches for inference in MAB problems. The first is to perform inference at the *end* of a pre-specified time by constructing a single confidence *interval* (CI). The second is to perform inference *during* the study as new data arrives by constructing a confidence *sequence* — a sequence of confidence intervals that are uniformly valid over time. As we perform inference on every arm at every time, the second approach allows us to stop collecting the data once we detect that one of the arms is statistically significantly outperforming the others.

In our paper, we propose a generic and flexible framework for performing inference both during and at the end of the study. Our work leverages the design-based approach to causal inference, which has a long history in the statistics community dating back to Fisher and Neyman but has seen a resurgence in popularity as it permits for inference on the obtained sample while handling complicated settings such as interference in an assumption-light manner Fisher (1935); Neyman, Iwaszkiewicz and Kolodziejczyk (1935); Rubin (1974); Holland (1986); Imbens and Rubin (2015); Ding, Feller and Miratrix (2016); Basse and Airoldi (2017); Ham, Imai and Janson (2022). The design-based framework has several advantages in MAB settings. First, it lets us relax the stationary and independent assumption for the rewards because the design-based framework conditions on the (potential) outcomes, allowing our inferential results to hold for any general reward distribution.[1] Second, a central principle in any MAB algorithm is balancing exploration versus exploitation; however, this trade-off is inherently about analyzing how the agent has performed for the *current* sample, e.g., how much the agent "lost" by picking arm A over arm B for the finite-sample data. Unlike the more common super-population approach that performs inference on the general population eligible to interact with the MAB algorithm Hadad et al. (2021); Waudby-Smith et al. (2022a), the design-based framework directly performs inference for relevant finite-sample estimands. Lastly, while relaxing the stationary and independence assumptions, the design-based framework only requires light and mostly testable assumptions. In particular, the main assumption we require is that the probability of picking an arm at any time is bounded away from zero or one. Since the agent/experimenter is in control of the data collection process, this can not only be verified but also controlled by using the appropriate adaptive scheme.

The main contribution of this paper is that we formally introduce the design-based approach for MAB, allowing us to build valid confidence intervals (Section 3) and confidence sequences (Section 4) for the mean

---

[1]For example, the mean reward of each arm may be time-varying and also dependent on past data.

reward of each arm and the difference between two arms' mean rewards. We define our setting and design-based estimand and estimators in Section 2. Then we extend our results to contextual MAB in Section 5 and end with simulations in Section 6.

2 Setup: Estimands and Estimation

We now introduce relevant notation, the design-based inference framework, the estimands of interest in the MAB framework, and our proposed estimators.

2.1 Design-Based Causal Inference

Suppose we observe T samples of $\{W_t, Y_t\}_{t=1}^T$, where $W_t \in \{0, \ldots, K-1\}$ is the K possible actions (or treatments) the agent takes at time t and Y_t is the observed reward (or outcome) at time $t = 1, \ldots, T$. For readers more familiar with the MAB literature, the notation (A_t, R_t) is often used for the action and reward, respectively. We purposefully use the alternative notation to emphasize the connection with causal inference. In Section 5, we further assume that at each point in time, we also observe some covariates or contexts X_t to generalize our analysis to contextual MAB Li et al. (2010).

Borrowing from the standard causal inference literature, we assume that for each t there are K potential outcomes $Y_t(0), \ldots, Y_t(K-1)$, corresponding to the K possible actions Fisher (1935); Neyman, Iwaszkiewicz and Kolodziejczyk (1935); Imbens and Rubin (2015); Abadie et al. (2020). Note that we implicitly assume the no-interference assumption of potential outcomes for simplicity and brevity of notation. To connect the potential outcomes to the observed outcome, we assume that $y_t = Y_t(w_t)$, where lower case (w_t, y_t) denotes our observed treatment samples.

The design-based approach conditions on the full set of potential rewards $Y_t(0), \ldots, Y_t(K-1)$, allowing the rewards to be arbitrarily dependent on the past rewards and non-stationary. This setting generalizes the common assumptions invoked in MAB, where researchers assume $Y_t(W_t)$ is a function of only its current action and is generated from an *iid* distribution. The relaxation of the assumptions is possible because the design-based approach shifts the modeling burden away from the unknown outcome to the know action distribution.

Specifically, we assume for each t the action W_t to be adapted based on the historical data,

$$p_{t|t-1}(w) := \Pr(W_t = w_t \mid \mathcal{F}_{t-1}), \tag{1}$$

where the filtration $\mathcal{F}_{t-1}$ contains the information set of the past $(W_{1:(t-1)}, Y_{1:(t-1)})$. In Section 5, we extend Equation (1) to allow the action to further depend on the context variables. Lastly, we impose that $p_{t|t-1}(w)$ is bounded away from zero or one, often known as probabilistic treatment assignment assumption in causal inference Imbens and Rubin (2015).

Assumption 1 (Probabilistic Treatment Assignment)**.** For all times $t = 1, \ldots, T$ and possible actions $w \in \{0, \ldots, K-1\}$, we have that

$$0 < p_{t|t-1}(w) < 1.$$

Assumption 1 is fairly weak and holds as long as the proposed MAB algorithm does not converge to a zero/one probability event for any of the arms in finite time. Furthermore, we remark that many existing works require this regularity condition for inference Hadad et al. (2021); Zhang, Janson and Murphy (2021); Ham, Imai and Janson (2022).

2.2 Multi-arm Bandit Estimands

The primary estimand of interest in MAB algorithms is the reward mean distribution for arm $w \in \{0, \ldots, K-1\}$. In the typical MAB literature, the reward mean distribution of arm w is distributed *iid*, thus can be characterized through $\mu_t(w) = \mu(w)$ since it does not vary over time. As we do not impose any restriction on the reward mean, we choose to define our estimand as a finite-sample cumulative mean:

$$Q_t(w) := \frac{1}{t} \sum_{j=1}^{t} Y_j(w), \tag{2}$$

When constructing confidence intervals at the end of a fixed time T, we are interested in $Q_T(w)$. However, when constructing confidence sequences, we perform inference at every time t; hence our estimand is the time-varying $Q_t(w)$ at every time t.

To connect this to the commonly known MAB reward mean distribution, consider the following example where the agent is interested in the mean reward of arm $w = 1$.

Example 1. Suppose $t = 3$, then Equation (2) reduces to

$$Q_3(1) = \frac{1}{3}(Y_1(1) + Y_2(1) + Y_3(1)).$$

Notice that $Q_3(1)$ in Example 1 is the sample finite-population analogue of the more typical reward mean distribution function $E(Y_t(1))^2$. Because the design-based framework conditions on the full set of potential outcome, all corresponding estimands have a finite-sample interpretation for quantities of interest.

In practice, we are more interested in comparing the mean reward between two or more arms,

$$\tau_t(w, w') := Q_t(w) - Q_t(w'). \tag{3}$$

For simplicity, we focus on pairwise differences between two arms, but our paper can easily be extended to estimate any linear combinations of the K mean rewards. Our main result builds confidence intervals and sequences for both $Q_t(w)$ and $\tau_t(w, w')$.

2.3 Estimation

We now propose an estimators for $Q_t(w)$ and consequently $\tau_t(w, w')$ that serve as the main building blocks for theoretical results throughout the paper. We leverage the inverse propensity score estimator and its corresponding (conservative variance) estimator in Bojinov and Shephard (2019); Ham et al. (2022).

First, we introduce our unbiased estimator $\hat{Q}_t(w)$ along with an estimate of its variance

$$\hat{Q}_t(w) := \frac{1}{t} \sum_{j=1}^{t} \hat{\tau}_j(w) := \frac{1}{t} \sum_{j=1}^{t} \frac{\mathbf{1}\{W_j = w\}Y_j}{p_{j|j-1}(w)}$$

$$S_t(w) := \sum_{j=1}^{t} \hat{\sigma}_j^2(w) \tag{4}$$

$$:= \sum_{j=1}^{t} \frac{\mathbf{1}\{W_j = w\}Y_j^2(1 - p_{j|j-1}(w))}{p_{j|j-1}(w)^2}.$$

---

[2]Here the expectation is taken with respect to the random reward.

Similarly, we propose the following estimator of $\tau_t(w, w')$ and an estimate of the upper bound of its variance

$$\hat{\tau}_t(w, w') := \hat{Q}_t(w) - \hat{Q}_t(w') = \frac{1}{t}\sum_{j=1}^{t}\hat{\tau}_j(w) - \hat{\tau}_j(w')$$

$$S_t(w, w') := \sum_{j=1}^{t}\hat{\sigma}_j^2(w, w') \qquad (5)$$

$$:= \sum_{j=1}^{t} \frac{\mathbf{1}\{W_j = w\}Y_j^2}{p_{j|j-1}(w)^2} + \frac{\mathbf{1}\{W_j = w'\}Y_j^2}{p_{j|j-1}(w')^2}.$$

The above estimators are conditionally unbiased for the respective estimands, which we formally state in the following lemma that is proven in Appendix A. Here, and throughout the paper, the expectation is taken with respect to the random action W_t.

Lemma 2.1 (Unbiased properties of estimators). Under Assumption 1, we have that

$$E(\hat{Q}_t(w) \mid \mathcal{F}_{t-1}) = Q_t(w)$$

$$E(\hat{\tau}_t(w, w') \mid \mathcal{F}_{t-1}) = \tau_t(w, w')$$

holds for every $t = 1, \ldots, T$. Furthermore, we have that

$$\text{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1}) = E(\hat{\sigma}_t^2(w) \mid \mathcal{F}_{t-1})$$

$$\text{Var}(\hat{\tau}_t(w) - \hat{\tau}_t(w') \mid \mathcal{F}_{t-1}) \leq E(\hat{\sigma}_t^2(w, w') \mid \mathcal{F}_{t-1})$$

holds for every $t = 1, \ldots, T$.

We conclude this section with a few remarks about our framework. First, besides Assumption 1, which can be verified and controlled by the agent, we do not place restrictions on either the adaptive assignment process or the reward generation process. This allows us to assume that the rewards distribution is both non-*iid* and non-stationary. Hence, all our estimands are denoted with subscript t to show that these may change over time. Despite this general framework, the design-based approach allows us to perform inference since we leverage the randomness in the actions. Second, we can only obtain an upper bound of the variance for $\hat{\tau}_t(w) - \hat{\tau}_t(w')$ because the actual variance contains a product of $Y_t(w)Y_t(w')$, which is never observed. Finally, Lemma 2.1 holds for any t including $t = T$. Consequently, the estimates proposed in this section are relevant for constructing both confidence intervals and confidence sequences by leveraging the martingale convergence theory.

3 Design-based Confidence Intervals for Multi-arm Bandits

We now demonstrate how to construct asymptotically valid confidence intervals for $Q_t(w)$ and $\tau_t(w, w')$. Before providing our result, we require an additional assumption that restricts any realized reward to be bounded by an arbitrarily large constant.

Assumption 2 (Bounded realized rewards).

$$|Y_t(W_t)| \leq M$$

for all t and $W_t \in \{0, \ldots, K - 1\}$, where $M \in \mathbb{R}$.

5

Note that M can be extreme to make this assumption hold. Such assumptions are commonly used to satisfy the necessary regularity conditions used in design-based inference Bojinov and Shephard (2019); Lei and Ding (2020). For example, if T potential outcomes were generated from a $N(0,1)$ distribution, an unbounded distribution, each of the T realized rewards are still bounded.

3.1 Design-based Confidence Intervals

Given these assumptions, we state the first result that allows an agent to build confidence *intervals* at the end of the study at time $t = T$ for both the reward mean function $Q_T(w)$ and the reward mean difference between two arms $\tau_T(w, w')$ for general arbitrary distributions for the rewards.

Theorem 3.1 (Design-based CI for MAB). Suppose data $\{w_t, y_t\}_{t=1}^T$ are observed for a fixed pre-specified T, where Assumption 1- 2 are satisfied and W_t adapts based on the past as shown in Equation (1). Then, as $T \to \infty$,

$$\hat{Q}_T(w) \pm z_{1-\alpha/2} \frac{\sqrt{\sum_{j=1}^T \hat{\sigma}_j^2(w)}}{T}$$

forms an asymptotically valid $(1-\alpha)$ confidence interval for $Q_T(w)$, where z_a is the ath-quantile of a standard normal distribution. Furthermore,

$$\hat{\tau}_T(w, w') \pm z_{1-\alpha/2} \frac{\sqrt{\sum_{j=1}^T \hat{\sigma}_j^2(w, w')}}{T}$$

forms an asymptotically valid $(1 - \alpha)$ confidence interval for $\tau_T(w, w')$.

The proof is provided in Appendix B, which uses a central limit theorem for martingale sequence differences. The widths of the confidence intervals in Theorem 3.1 decrease with rate approximately $1/\sqrt{T}$, similar to that of a t-test. For Theorem 3.1 to hold, we only required bounded realized rewards and probabilistic treatment assignments. The first is a mild condition, while the second is within the agent's control. Otherwise, we assume nothing about the data-generating process of the reward distribution and allow that the action W_t is adapted based on the past.

4 Design-based Confidence Sequences for Multi-arm Bandits

We focus on the sequential nature of MAB and present a strategy to perform inference any time new data arrives by constructing valid confidence sequences. Confidence sequences are sequences of confidence intervals that are uniformly valid over time (often referred to as anytime-valid inference). Formally, a sequence set of confidence intervals $\{V_t\}_{t=1}^T$ is a valid confidence sequence with type-1 error α for the target parameter μ_t if

$$\Pr(\forall t, \mu_t \in V_t) \geq 1 - \alpha \tag{6}$$

holds for arbitrary data-dependent stopping rule T, where T is the final time of the experiment and can be determined in any data-dependent way Howard et al. (2020). In words, Equation (6) states that our confidence sequence V_t covers the desired potentially time-varying estimand estimand, e.g., mean reward of arm, uniformly at *any time* with probability $1 - \alpha$. This formally allows the analyst to stop the MAB as soon as the analyst is satisfied with the inference, e.g., mean reward of an arm is statistically higher than a threshold.

In Section 4.1, we provide non-asymptotic confidence sequences that have some practical issues. Consequently, we provide an improved asymptotic confidence sequence in Section 4.2.

4.1 Design-Based Exact Confidence Sequence

We begin by stating the exact closed-form confidence sequence in the following theorem that is proved in Appendix C.1.

Theorem 4.1 (Design-based Exact CS for MAB)**.** Suppose data $\{w_t, y_t\}_{t=1}^{T}$ are observed for any arbitrary data dependent stopping time T^3, where Assumption 1- 2 holds. Let $m := M/p_{min}$, where $p_{min} = \min_{t,w} p_{t|t-1}(w)$. Then, $\hat{Q}_t(w) \pm C_t(S_t(w))$ forms a valid $(1-\alpha)$ confidence sequence for $Q_t(w)$, where $C_t(S_t) :=$

$$\left[\frac{m(m+1)}{t} \log\left(\frac{2}{\alpha}\right) + \frac{S_t}{t} \left(\frac{m+1}{m} \log\left(1 + \frac{1}{m}\right) - \frac{1}{m} \right) \right]$$

Furthermore, $\hat{\tau}_t(w, w') \pm C_t(S_t(w, w'))$ forms a valid $(1 - \alpha)$ confidence sequence for $\tau_t(w, w')$, where $S_t(w), S_t(w, w')$ are defined in Equation (4) and (5), respectively.

There are two practical limitations to Theorem 4.1. First, the confidence sequence scales with M/p_{min}. M may be unknown unless under special cases such as binary rewards. Furthermore, in the spirit of a sequential test, the agent may desire to change $p_{t|t-1}(w)$ as more units enter. However, Theorem 4.1 requires p_{min} to be determined before the start of the algorithm, thus restricting the action probabilities to a pre-specified range.

Second, the confidence width is of order approximately $O(S_t/t)$, which likely only shrinks asymptotically to zero if the estimated variances grow sub-linearly or there are stronger assumptions on the potential outcomes. We can fix the second issue by leveraging a mixture distribution with the truncated gamma distribution to build another confidence sequence shown in Appendix C.2 with order approximately $O(\sqrt{S_t \log(S_t)}/t)$. The price we pay for the additional performance gain is that the CS no longer has a closed-form solution and requires root-solving algorithms.

4.2 Design-Based Asymptotic Confidence Sequence

We now improve the confidence sequence presented in Theorem 4.1 by providing asymptotic confidence sequences. Informally, asymptotic confidence sequences are valid confidence sequences after a "sufficiently large" time. We further show the robustness of our confidence sequence through simulations in Section 6.

For completeness, we formally define an asymptotic confidence sequence first introduced in Waudby-Smith et al. (2021).

Definition 4.1 (Asymptotic Confidence Sequences)**.** We say that $(\hat{\mu}_i \pm V_i)$ is a two-sided $(1 - \alpha)$ asymptotic confidence sequence for a target parameter μ_i if there exists a non-asymptotic confidence sequence $(\hat{\mu}_i \pm \tilde{V}_i)$ for μ_i such that

$$\frac{\tilde{V}_t}{V_t} \xrightarrow{a.s.} 1. \tag{7}$$

To give some intuition of the above definition, "couplings" has been used in the literature of strong approximations Einmahl (1987); Komlos, Major and Tusnády (1976) to formally define asymptotic confidence *intervals*. In this literature, asymptotic confidence interval is defined by a "coupled" finite-sample valid confidence interval centered at the same statistic such that the difference between the two non-asymptotic and

---

[3]With a slight abuse of notation we also use T (the final data collection time) as a stopping time. Furthermore, because T is data-dependent, we require that it is formally a well-defined stopping time, i.e., a measurable function dependent on the current and previous data (not on the future).

asymptotic confidence intervals is negligible in the limit. Equation (7) captures the same notion except with the almost-sure convergence to satisfy the time *uniform* guarantee required.[4]

Before stating the theorem, we require an additional assumption that restricts the variance from vanishing in the limit.

Assumption 3 (None Vanishing Variance). Let $\sigma_t^2(w) := \mathrm{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1})$ and $\sigma_t^2(w, w')$ is defined similarly. Then we assume that both

$$\sum_{j=1}^{t} \sigma_j^2(w) \to \infty, \quad \sum_{j=1}^{t} \sigma_j^2(w, w') \to \infty$$

almost surely.

Assumption 3 holds if $1/t \sum_{j=1}^{t} \sigma_j^2(w) \xrightarrow{a.s.} \sigma_*^2$ or if $\sigma_1^2(w) = \sigma_2^2(w) = \cdots = \sigma_T^2(w)$. Informally, Assumption 3 is satisfied as long as the potential outcomes do not vanish to zero as time grows.

Theorem 4.2 (Design-based Asymptotic CS for MAB). Suppose data $\{w_t, y_t\}_{t=1}^{T}$ are observed for any arbitrary data dependent stopping time T, where Assumption 1- 3 are satisfied and W_t adapts based on the past as shown in Equation (1). Then, $\hat{Q}_t(w) \pm D_t(S_t(w))$ forms a valid $(1 - \alpha)$ asymptotic confidence sequence for $Q_t(w)$, where

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

forms a valid $(1 - \alpha)$ asymptotic confidence sequence for $Q_t(w)$. Furthermore, $\hat{\tau}_t(w, w') \pm D_t(S_t(w, w'))$ forms a valid $(1 - \alpha)$ asymptotic confidence sequence for $\tau_t(w, w')$, where $\eta > 0$ is any pre-specified constant

The proof is also provided in Appendix B. The confidence sequences provided in Theorem 4.2 can also serve as valid confidence *intervals* at any time t. The width of the confidence sequences scale similar to that in Theorem 4.2 except there is an extra log term to control the type-1 error at every time uniformly. Lastly, η is typically chosen by the analyst to minimize the confidence width at a certain fixed time. Following the advice in Ham et al. (2022), for all examples and simulations, we choose η so that the width is minimized at time 10. We do this because the confidence sequence width is largest at early times, and the choice of η becomes insignificant as more data arrives. We additionally show a closed-form expression to calculate η in Appendix D. For practice, we recommend using the same η we do throughout the paper, i.e., $\eta \approx 0.77$.

To illustrate Theorems 3.1 and 4.2, consider the following basic example.

Example 2 (Stationary and independent four-arm bandit). Suppose a four-arm bandit problem, where an agent pulls from arms $0, 1, 2, 3$. Suppose the rewards for arm w are stationary and independently generated from a $N(\mu_w, 1)$, where $\mu_0 = \mu_4 = 1, \mu_1 = 2, \mu_2 = 1.5$, so that arm 1 has the highest mean reward.

We use the following adaptive probabilities for the tth observation

$$p_{t|t-1}(w) = \frac{\bar{Y}_{w,t-1}}{\sum_{w=1}^{K} \bar{Y}_{w,t-1}}, \quad t > 0.1 * T, \tag{8}$$

where $\bar{Y}_{w,t-1}$ is the sample mean of arm w using the obtained samples for up to time $t - 1$. In other words, the agent up-weights the arm that produces higher mean rewards. Lastly, the agent does a fair coin flip for all the arms for the first 10% of the sample (exploration period), where $T = 1000$.

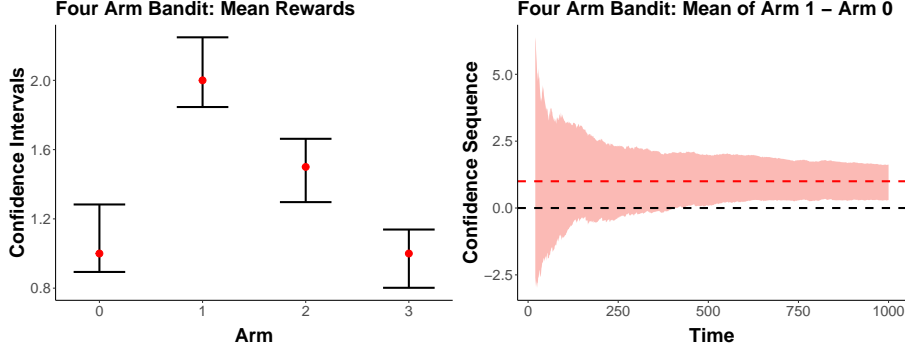[4]This is formally proven in Appendix C.4 of Waudby-Smith et al. (2021).

Figure 1: Four Arm Bandit (Example 2). The left panel uses Theorem 3.1 to construct confidence intervals for the four-arm means, where the red dot represents the truth. On the right panel, the red contours show the lower and upper confidence sequence using Theorem 4.2. The horizontal red dotted line represents the true mean difference of the rewards, and the black horizontal dotted line represents the zero (null) line.

The left panel of Figure 1 shows our proposed confidence intervals cover the true mean reward for all arms even under adaptively sampled data. Furthermore, the right panel shows our confidence sequence tightens to the desired truth and covers the true mean difference at all times. For this case, if the agent desired to find any arm that is better than the default arm 0, then this confidence sequence shows the agent can terminate the experiment as early as $t = 400$ (approximately the first time the red contours are statistically significantly above zero), saving further budget and sample.

5 Contextual multi-arm bandits

We now extend our results to contextual MAB, where an agent has access to some "context" or covariates X_t before deciding the tth action. Consequently, at the end of the tth roud the full data are $\{X_t, W_t, Y_t\}_{t=1}^T$, where X_t may be multi-dimensional. Although we can ignore X_t and only use (W_t, Y_t) to construct valid confidence intervals and confidence sequences, we can leverage X_t as a variance reduction technique.

To formalize this, we first denote

$$\hat{Y}_{t|t-1} := \hat{f}_t(X_{1:t}, Y_{1:(t-1)}, W_{1:(t-1)}),$$

where $\hat{f}_t$ denotes the prediction for the next observation Y_t using all the past data and the current X_t. With a slight abuse of notation, we make the filtration $\mathcal{F}_{t-1}$ to further contain the information set for $X_{1:t}$. Since our results leverage martingale theory, $\hat{Y}_{t|t-1}$ is equivalent to a constant because we always condition on $\mathcal{F}_t$.

More formally, our estimand $\tau_t(w, w')$ in Equation (3) can be re-written as $\tau_t(w, w') =$

$$Q_t(w) - Q_t(w')$$
$$= \frac{1}{t} \sum_{j=1}^t \{Y_j(w) - \hat{Y}_{j|j-1}\} - \{Y_j(w') - \hat{Y}_{j|j-1}\}. \tag{9}$$

Using the above formulation, we can define our new "residualized" outcome $\tilde{Y}_t := Y_t - \hat{Y}_{t|t-1}$ and use the same estimators in Section 2.3 and results in Theorems 3.1 and 4.2 except replacing Y_t with $\tilde{Y}_t$.

9

More formally, we rewrite Equation (5) as

$$\hat{\tau}_t(w, w')^X := \frac{1}{t} \sum_{j=1}^{t} \hat{\tau}_j(w)^X - \hat{\tau}_j(w')^X$$

$$S_t(w, w')^X := \sum_{j=1}^{t} \hat{\gamma}_j^2(w, w'),$$

(10)

where $\hat{\tau}_t(w)^X, \hat{\gamma}_t^2(w, w')$ are defined similar to $\hat{\tau}_t(w), \hat{\sigma}_t^2(w, w')$ except replacing Y_t with $\tilde{Y}_t$. The result of Lemma 2.1 directly extends to these new estimators because we condition on the filtration; thus $\hat{Y}_{t|t-1}$ is equivalent to a constant (hence it is crucial that the predictions are constructed based on the past data without using W_t). This allows the analyst to formally use $\hat{Y}_{t|t-1}$ to incorporate any machine learning algorithm or prior knowledge to reduce the variance. This reduction is proportional to how small $\{Y_t - \hat{Y}_{t|t-1}\}^2$ is, i.e., how well the analyst can use the prior data to predict the next response.

Furthermore, we can also allow $p_{t|t-1}(w)$, defined in Equation (1), to further adapt based on the context variables. Note that, we can only use $\hat{Y}_{t|t-1}$ for tackling $\tau_t(w, w')$ since $\hat{Y}_{t|t-1}$ does not cancel in Equation (9) when just estimating $Q_t(w)$ alone.

5.1 Variance Reduction Technique

We now extend Theorems 3.1 and 4.2 for building valid confidence intervals and sequences for $\tau_t(w, w')$ using $\hat{Y}_{t|t-1}$ in the contextual MAB setting.

Theorem 5.1 (Design-based CI and CS for contextual MAB). Suppose data $\{x_t, w_t, y_t\}_{t=1}^{T}$ are observed, where Assumption 1- 2 are satisfied and W_t adapts based on the past as shown in Equation (1). Further suppose $\hat{Y}_{t|t-1}$ is bounded for all t. Then,

$$\hat{\tau}_T(w, w')^X \pm z_{1-\alpha/2} \frac{\sqrt{\sum_{j=1}^{T} \hat{\gamma}_j^2(w, w')}}{T}$$

forms an asymptotically valid $(1 - \alpha)$ confidence interval for $\tau_T(w, w')$ for a pre-specified T.

Denote $\tilde{\sigma}_j^2(w, w') := \text{Var}(\hat{\tau}_t(w, w)^X \mid \mathcal{F}_{t-1})$. Then if we further assume none vanishing variance for the new residualized outcome, i.e., $\sum_{j=1}^{t} \tilde{\sigma}_j^2(w, w') \to \infty$ almost surely, then

$$\hat{\tau}_t(w, w')^X \pm D_t(S_t(w, w')^X)$$

forms a valid $(1 - \alpha)$ asymptotic confidence sequence for $\tau_t(w, w')$ for any arbitrary data-dependent stopping rule T, where $\eta > 0$ is any pre-specified constant and $D_t(S_t)$ is defined in Theorem 4.2.

The proof is omitted because under the stated assumptions, the setting is identical to that of Theorems 3.1 and 4.2 except we replace the response Y_t with the new residualized response $\tilde{Y}_t$. We show through simulations in Section 6.2 that residualizing the outcome can lead to a substantial reduction in the variance.

6 Simulations and Related Work

We now provide a simulation with two goals. First, we show the empirical coverage of the proposed methods and compare the advantages and disadvantages of using confidence sequences over confidence intervals in

Method	Coverage	Width	Stopping Time	Power
Asymp-CI	95%	0.14	NA	92%
Asymp-CS	95%	0.23	580	NA
Exact-CS	99%	0.25	640	NA
CS with X	98%	1.92	115	NA

Table 1: Binary *iid* rewards simulation. The first row shows the performance of the asymptotic CI in Theorem 3.1 while the second and third row shows the performance of the asymptotic and exact CS using Theorem 4.2 and Theorem 4.1, respectively under $\alpha = 0.05$.

a simple *iid* binary reward setting. Next, we shift to a more complex setting with non-stationary continuous rewards with context variable X_t. We demonstrate both the validity and gain we gain from incorporating X_t into our confidence interval and sequence. Finally, we end with a discussion of our results in context with existing works.

6.1 Independent and Identically Distributed Binary Rewards

We begin our simulations in a simplistic setting where we have two arms $w = 0, 1$ with *iid* binary rewards from $\text{Bern}(\mu_w)$, where $\mu_1 = 0.27, \mu_0 = 0.15$, i.e., a 12% expected increase from choosing arm 1. Although this is a simplified setting, it represents typical adaptive A/B tests or MAB with a treatment and control group, e.g., learning whether the new product is better than the standard offering through bandits.

For simulations in Section 6.1- 6.2, we build confidence interval and confidence sequence for $\tau_t(1, 0)$, i.e., the mean reward difference from choosing arm 1 over arm 0 when $\alpha = 0.05$. For simplicity, we use the adaptive sampling procedure outlined in Equation (8), where we use the first 10% of samples for exploration and adapt based on the sample means of each arm. Additionally, we run 1000 Monte-Carlo experiments and report four statistics. We first report the coverage, i.e., the proportion of times the confidence interval covers $\tau_T(1, 0)$ for a fixed T. Then, for confidence sequences, we report the proportion of times our confidence sequence covers $\tau_t(1, 0)$ for all times $t > 10$. Following the advice in Ham et al. (2022), we check only after an initial 10 samples because our method is asymptotic, and it is practically unlikely for the analyst to terminate the experiment only after 10 samples. Second, we report the average width at a pre-specified time $t = T$ for all methods. Third, we report the average stopping time for only the confidence sequences, where our stopping time is defined as the first time the confidence sequence does not cover zero. Therefore, for all confidence sequences, we run the simulation for sufficiently large $t > T$. Lastly, we report the statistical power for the confidence interval.

Table **??** shows the simulation results under the simple *iid* setting for $T = 700$ samples. We find that all our proposed methods have the desired coverage. As expected, the width for the confidence sequences is wider than that of the confidence interval by less than two times. However, on average, the asymptotic confidence sequence can detect an effect as early as $t = 580$ (approximately 80% of $T = 700$ the total sample). Although the analyst would reject close to 92% of the times with the proposed confidence interval by $t = 700$, the confidence sequences are attractive alternatives if the agent wishes to terminate the experiment as soon as the agent detects a statistically significant effect. Lastly, we find that the asymptotic confidence sequence has improved width and stopping time compared to the non-asymptotic confidence sequence while maintaining proper coverage at even early times, likely due to our conservative variance estimator.

Method	Coverage	Width	Stopping Time	Power
CI no X	95%	1.84	NA	88%
CS no X	99%	3.68	340	NA
CI with X	95%	1.02	NA	98%
CS with X	98%	1.92	115	NA

Table 2: Continuous non-stationary rewards simulation. The first two row shows the performance of the asymptotic CI and CS in Theorem 3.1 and Theorem 4.2, respectively. The last two rows show the corresponding CI and CS incorporating the context variable X_t through Theorem 5.1.

6.2 Non-stationary Continuous Rewards

We now change our reward distribution to a non-stationary continuous distribution with four arms ($K = 4$). We further add one binary context variable X_t to illustrate Theorem 5.1. More specifically, our data generating process is the following AR(1) linear model

$$
\begin{aligned}
Y_t(0) &= \rho Y_{t-1}(0) + \beta X_t + \epsilon_t, |\rho| \le 1, \epsilon_t \overset{iid}{\sim} N(\mu_{t,0}, 1) \\
Y_t(w) &= Y_t(0) + N(\mu_{t,w}, 1) \\
Y_0(0) &= 0, \quad X_t \sim \text{Bern}(0.5),
\end{aligned}
\tag{11}
$$

where ρ represents how the next potential outcome is lag-1 dependent on its immediate history. Furthermore, $\mu_{t,1} - \mu_{t,0}$ is our target parameter, i.e., the mean causal effect of being in arm 1 over arm 0. Although we can make $\mu_{t,w}$ time-varying, we fix $\mu_{t,0} = 1, \mu_{t,1} = 1.5, \mu_{t,2} = 1.25, \mu_{t,3} = 1$ for all t for simplicity. Our reward distribution is non-stationary and no longer *iid* due to ρ, which we set at $\rho = 0.1$. Finally, we let $\beta = 1$ and use a linear regression of $Y_{1:(t-1)}$ on $X_{1:(t-1)}$ to predict Y_t to demonstrate Theorem 5.1.

Table **??** shows the simulation results under the non-stationary setting described in Equation (11) for $T = 300$ samples. As expected, our methods have an over-conservative coverage due to the conservative variance estimator. Nevertheless, incorporating X using Theorem 5.1 successfully reduces the stopping time by a third, approximately halving the width of the confidence sequence and interval, and increasing the power substantially. Furthermore, we similarly find that the confidence sequence width is larger than that of the confidence intervals but has the potential to end the MAB substantially earlier than $T = 300$.

6.3 Comparison with Related Works

We now extend the previous simulation to compare with existing work. Our confidence interval result is most closely related to Hadad et al. (2021), where the aforementioned paper takes a super-population approach to inference, i.e., the potential outcomes $Y_t(W_t)$ are generated *iid* from a distribution satisfying technical conditions, e.g., bounded moments. Additionally, to the best of our knowledge, Waudby-Smith et al. (2022a); Karampatziakis, Mineiro and Ramdas (2021) are the only existing works that build confidence sequence for MAB. However, the main results presented in the aforementioned papers assume bounded $[0, 1]$ stationary rewards. Furthermore, there also exist many technical and often untestable assumptions on the rewards, which our work bypasses through the design-based approach.

Therefore, we only compare our method with Hadad et al. (2021). In particular, we use Theorem 2 in Hadad et al. (2021) with unity weights since our adaptive allocation probabilities do not diverge or oscillate asymptotically. Since Hadad et al. (2021) takes a super-population approach, the corresponding estimand is $E(Y(0))$, where the expectation is taken with respect to the stochastic potential outcomes.

We keep an identical simulation setting of that in Section 6.2 except we set $\beta = 0$ to simplify the setting and vary ρ in the x-axis to induce dependency and break the *iid* stationarity assumption. We also focus on estimating $Q_T(0)$, i.e., the mean reward for arm 0. We remark that $E(Y(0)) = \mu_{t,0}$ marginally, which is the target estimand for Hadad et al. (2021). Since ρ induces dependency among the potential outcomes, we expect poor type-1 error control in Hadad et al. (2021) while our proposed confidence interval in Theorem 3.1 remains valid regardless of any potential outcome distribution. Figure 2 shows that the coverage for the design-based confidence interval (DBCS) remains at the desired 95% level while the confidence interval in Hadad et al. (2021) quickly loses validity as ρ grows. Therefore, our work extends existing work for inference in multi-arm bandits to general non-stationary MAB in an assumption-light manner through the design-based approach.
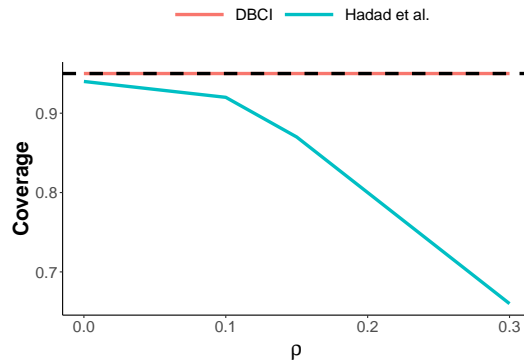

Figure 2: The simulation study settings remains identical to that in Section 6.2 except we vary ρ to break stationarity and set $\beta = 0$. The red line shows the coverage of $Q_T(0)$ using Theorem 3.1. The blue line shows the coverage of $E(Y(0))$ using the method proposed in Hadad et al. (2021), respectively. Finally, the black dotted line shows the desired type-1 error at $\alpha = 0.05$.

References

Abadie, Alberto, Susan Athey, Guido W Imbens and Jeffrey M Wooldridge. 2020. "Sampling-Based versus Design-Based Uncertainty in Regression Analysis." Econometrica 88:265–296.

Banerjee, Debangshu, Avishek Ghosh, Sayak Ray Chowdhury and Aditya Gopalan. 2022. "Exploration in Linear Bandits with Rich Action Sets and its Implications for Inference.".
URL: *https://arxiv.org/abs/2207.11597*

Basse, Guillaume and Edoardo Airoldi. 2017. "Limitations of Design-based Causal Inference and A/B Testing under Arbitrary and Network Interference." Sociological Methodology 48.

Berry, D. A. and B. Fristedt. 1985. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, London.

Bojinov, Iavor and Neil Shephard. 2019. "Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading." Journal of the American Statistical Association 114:1665–1682.
URL: *https://doi.org/10.1080/01621459.2018.1527225*

Bubeck, Sébastien and Cesa-Bianchi Nicolò. 2012. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems.

Chow, Y. S. 1971. "On the Lp-Convergence." The Annals of Mathematical Statistics 42:393 – 394.
URL: *https://doi.org/10.1214/aoms/1177693530*

Deshpande, Yash, Lester Mackey, Vasilis Syrgkanis and Matt Taddy. 2017. "Accurate Inference for Adaptive Linear Models.".

Dimakopoulou, Maria, Zhengyuan Zhou, Susan Athey and Guido Imbens. 2019. Balanced Linear Contextual Bandits. AAAI'19/IAAI'19/EAAI'19 AAAI Press.
URL: *https://doi.org/10.1609/aaai.v33i01.33013445*

Dimakopoulou, Maria, Zhimei Ren and Zhengyuan Zhou. 2021. "Online Multi-Armed Bandits with Adaptive Inference.".
URL: *https://arxiv.org/abs/2102.13202*

Ding, Kaize, Jundong Li and Huan Liu. 2019. Interactive Anomaly Detection on Attributed Networks. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. WSDM '19 New York, NY, USA: Association for Computing Machinery p. 357–365.
URL: *https://doi.org/10.1145/3289600.3290964*

Ding, Peng, Avi Feller and Luke Miratrix. 2016. "Randomization inference for treatment effect variation." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78:655–671.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12124*

Durand, Audrey, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios Mitsis and Joelle Pineau. 2018. "Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis." Journal of Machine Learning Research 85:1–15.

Einmahl, Uwe. 1987. "Strong Invariance Principles for Partial Sums of Independent Random Vectors." The Annals of Probability 15:1419 – 1440.
URL: *https://doi.org/10.1214/aop/1176991985*

Fan, Xiequan, Ion Grama and Quansheng Liu. 2015. "Exponential inequalities for martingales with applications." Electronic Journal of Probability 20:1 – 22.
URL: *https://doi.org/10.1214/EJP.v20-3496*

Fisher, Ronald A. 1935. The Design of Experiments. London: Oliver and Boyd Chapter II.

Hadad, Vitor, David A. Hirshberg, Ruohan Zhan, Stefan Wager and Susan Athey. 2021. "Confidence intervals for policy evaluation in adaptive experiments." Proceedings of the National Academy of Sciences 118:e2014602118.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.2014602118*

Ham, Dae Woong, Iavor Bojinov, Michael Lindon and Martin Tingley. 2022. "Design-Based Confidence Sequences for Anytime-valid Causal Inference.".
URL: *https://arxiv.org/abs/2210.08639*

Ham, Dae Woong, Kosuke Imai and Lucas Janson. 2022. "Using Machine Learning to Test Causal Hypotheses in Conjoint Analysis.".

Holland, Paul W. 1986. "Statistics and Causal Inference." Journal of the American Statistical Association 81:945–960.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478354*

Howard, Steven R., Aaditya Ramdas, Jon D. McAuliffe and Jasjeet S. Sekhon. 2020. "Time-uniform, non-parametric, nonasymptotic confidence sequences." The Annals of Statistics.

Imbens, Guido W. and Donald B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Karampatziakis, Nikos, Paul Mineiro and Aaditya Ramdas. 2021. Off-Policy Confidence Sequences. In Proceedings of the 38th International Conference on Machine Learning, ed. Marina Meila and Tong Zhang. Vol. 139 of Proceedings of Machine Learning Research PMLR pp. 5301–5310.

Komlos, Janos, Peter Major and G. Tusnády. 1976. "An approximation of partial sums of independent RV's, and the sample DF. II." Probability Theory and Related Fields 34:33–58.

Lei, Lihua and Peng Ding. 2020. "Regression adjustment in completely randomized experiments with a diverging number of covariates." Biometrika 108:815–828.
URL: *https://doi.org/10.1093/biomet/asaa103*

Li, Lihong, Wei Chu, John Langford and Robert E. Schapire. 2010. "A Contextual-Bandit Approach to Personalized News Article Recommendation." CoRR abs/1003.0146.
URL: *http://arxiv.org/abs/1003.0146*

Mary, Jeremie, Romaric Gaudel and Philippe Preux. 2015. Bandits and Recommender Systems. In Revised Selected Papers of the First International Workshop on Machine Learning, Optimization, and Big Data Volume 9432. Berlin, Heidelberg: Springer Verlag p. 325–336.

Neyman, J., K. Iwaszkiewicz and St. Kolodziejczyk. 1935. "Statistical Problems in Agricultural Experimentation (with discussion)." Supplement of Journal of the Royal Statistical Society 2:107–180.

Ramdas, Aaditya, Johannes Ruf, Martin Larsson and Wouter M. Koolen. 2020. "Admissible anytime-valid sequential inference must rely on nonnegative martingales." arXiv: Statistics Theory.

Robbins, Herbert. 1952. "Some aspects of the sequential design of experiments." Bulletin of the American Mathematical Society 58:527 – 535.
URL: *https://doi.org/*

Robbins, Herbert. 1970. "Statistical Methods Related to the Law of the Iterated Logarithm." The Annals of Mathematical Statistics 41:1397 – 1409.
URL: *https://doi.org/10.1214/aoms/1177696786*

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of Educational Psychology 66:688–701.

Slivkins, Aleksandrs. 2019. "Introduction to Multi-Armed Bandits." Foundations and Trends® in Machine Learning 12:1–286.
URL: *http://dx.doi.org/10.1561/2200000068*

Strassen, Volker. 1967. Almost sure behavior of sums of independent random variables and martingales.

Sutton, Richard S. and Andrew G. Barto. 2018. Reinforcement Learning: An Introduction. Cambridge, MA, USA: A Bradford Book.

Ville, Jean-Luc. 1939. Étude critique de la notion de collectif.

Waudby-Smith, Ian, David Arbour, Ritwik Sinha, Edward H. Kennedy and Aaditya Ramdas. 2021. "Time-uniform central limit theory with applications to anytime-valid causal inference.".

Waudby-Smith, Ian, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis and Paul Mineiro. 2022a. "Anytime-valid off-policy inference for contextual bandits.".
URL: *https://arxiv.org/abs/2210.10768*

Waudby-Smith, Ian, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis and Paul Mineiro. 2022b. "Anytime-valid off-policy inference for contextual bandits.".
URL: *https://arxiv.org/abs/2210.10768*

Zhang, Kelly, Lucas Janson and Susan Murphy. 2020. Inference for Batched Bandits. In Advances in Neural Information Processing Systems, ed. H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin. Vol. 33 Curran Associates, Inc. pp. 9818–9829.
URL: *https://proceedings.neurips.cc/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf*

Zhang, Kelly W., Lucas Janson and Susan A. Murphy. 2021. "Statistical Inference with M-Estimators on Adaptively Collected Data.".
URL: *https://arxiv.org/abs/2104.14074*

A Proof of Lemma 2.1

$$E(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1}) = E\left(\frac{\mathbf{1}\{W_t = w\}Y_t}{p_{t|t-1}(w)}\right)$$

$$= \frac{p_{t|t-1}(w)Y_t(w)}{p_{t|t-1}(w)}$$

$$= Y_t(w)$$

Hence, from above we immediately have that the first conditional moments are indeed unbiased.

Next, we first calculate the closed form expression of $\mathrm{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1})$.

$$\mathrm{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1}) = \mathrm{Var}\left(\frac{\mathbf{1}\{W_t = w\}Y_t}{p_{t|t-1}(w)} \mid \mathcal{F}_{t-1}\right)$$

$$= \frac{p_{t|t-1}(w)(1 - p_{t|t-1}(w))Y_t(w)^2}{p_{t|t-1}(w)^2}$$

$$= \frac{(1 - p_{t|t-1}(w))Y_t(w)^2}{p_{t|t-1}(w)}$$

Therefore, it follows that $\mathrm{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1}) = E(\hat{\sigma}_t^2(w) \mid \mathcal{F}_{t-1})$

Finally, we calculate the closed form expression of $\mathrm{Var}(\hat{\tau}_t(w) - \hat{\tau}_t(w') \mid \mathcal{F}_{t-1})$.

$$\mathrm{Var}(\hat{\tau}_t(w) - \hat{\tau}_t(w') \mid \mathcal{F}_{t-1}) = \mathrm{Var}(\hat{\tau}_t(w) \mid \mathcal{F}_{t-1}) + \mathrm{Var}(\hat{\tau}(w') \mid \mathcal{F}_{t-1})$$
$$- 2\mathrm{Cov}(\hat{\tau}_t(w), \hat{\tau}_t(w') \mid \mathcal{F}_{t-1}).$$

From the previous result we can calculate the two variance terms. First, we have that

$$\mathrm{Cov}(\mathbf{1}\{W_t = w\}, \mathbf{1}\{W_t = w'\} \mid \mathcal{F}_{t-1}) = -p_{t|t-1}(w)p_{t|t-1}(w').$$

Putting this together, we have that $\mathrm{Var}(\hat{\tau}_t(w) - \hat{\tau}_t(w') \mid \mathcal{F}_{t-1}) =$

$$= \frac{(1 - p_{t|t-1}(w))Y_t(w)^2}{p_{t|t-1}(w)} + \frac{(1 - p_{t|t-1}(w'))Y_t(w')^2}{p_{t|t-1}(w')}$$

$$+ 2Y_t(w)Y_t(w')$$

$$\leq \frac{\left(p_{t|t-1}(w)Y_t(w') + p_{t|t-1}(w)Y_t(w')\right)^2}{p_{t|t-1}(w)p_{t|t-1}(w')}$$

$$\leq \frac{Y_t(w)^2}{p_{t|t-1}(w)} + \frac{Y_t(w')^2}{p_{t|t-1}(w')}$$

where the third line follows because $p_{t|t-1}(w') \leq (1 - p_{t|t-1}(w))$ for any w, w' and the last line follows because $(a - b)^2 \geq 0 \to a^2 + b^2 \geq 2ab$ and we let $a = p_{t|t-1}(w')Y_t(w)$ and $b = p_{t|t-1}(w)Y_t(w')$. Finally we see that it is straight forward to show that

$$E(\hat{\sigma}_t^2(w, w') \mid \mathcal{F}_{t-1}) = \frac{Y_t(w)^2}{p_{t|t-1}(w)} + \frac{Y_t(w')^2}{p_{t|t-1}(w')},$$

completing the proof.

B Proof of Theorems 3.1 and 4.2

Throughout this proof, we omit subscript "a.s." from $o_{a.s.}(.)$ and $O_{a.s.}(.)$ to simplify notation. The proof proceeds in several steps, where the first step is enough to prove Theorem 3.1. The proof is identical for proving valid confidence interval and confidence sequence for either $Q_t(w)$ or $\hat{\tau}_t(w, w')$. Therefore, we show the proof for $\hat{\tau}_t(w, w')$.

Step 1: Strong Approximation via Martingale Sequence Differences The following is identical to Step 2 of the proof in Appendix D Ham et al. (2022), but we provide it here for completeness.

We first define

$$u_t = \hat{\tau}_t(w, w') - \tau_t(w, w').$$

By Lemma 2.1, $\{u_t\}$ is a martingale difference sequence with respect to $F_{T,t-1}$. Similar to the proof of Step 2 of Waudby-Smith et al. (2021), we also use the strong approximation theorem presented in Strassen (1967). In particular, we require Equation (159) in Theorem 4.4 of Strassen's paper (further details in Lemma A.3 of Waudby-Smith et al. (2021)) for our strong approximation theorem. However, our proof is different than that in Step 2 of Waudby-Smith et al. (2021) for the following reason.

The original Theorem 4.4 in Strassen (1967) is stated for martingales difference sequence of the form $E(X_n \mid \sigma(X_1, \dots, X_{n-1})) = 0$, where we use Strassen's notation and X_i are random variables with defined second moment. Although our martingale is of the form $E(f(X_n) \mid \sigma(X_1, \dots, X_{n-1})) = 0$, where $f(.)$ is the function that maps the data to u_t. More formally, to use the strong approximation theorem in Strassen (1967), we replace the beginning conditions of Theorem 4.4 in the following way.

"Let $X_1, X_2, \dots$ be random variables such that $0 \le E(f(X_n)^2 \mid X_1, \dots, X_{n-1}) \le C$ is bounded by some constant C (this directly holds under Assumption 2) and $E(f(X_n) \mid X_1, \dots, X_{n-1}) = 0$, a.s. for all n. Put $S_n = \sum_{i \le n} f(X_i)$ and $V_n = \sum_{i \le n} E(f(X_i)^2 \mid X_1, \dots, X_{i-1})$, where, in order to avoid trivial complications, we assume $V_1 = E(f(X_1)^2) > 0$."

The remaining conditions are identical and we omit the uniform integrability condition in Equation (138) of Strassen (1967) since it holds trivially under our bounded potential outcome for Assumption 2. We remark that the proof leading to Equation (159) remains identical and valid except replacing X_n with $f(X_n)$ in the appropriate steps. In particular, all random variables are still measurable with respect to $\sigma(X_1, \dots, X_{n-1})$. Lastly, although this theorem uses the actual variance (not an upper bound), using an upper bound only makes the confidence sequence width strictly wider and hence the validity still holds.

Let $Z_1, \dots, Z_t$ be *iid* random standard Gaussian. Utilizing Theorem 4.4 Equation (159) in Strassen (1967) we have that

$$\frac{1}{t}\sum_{j=1}^{t} u_j = \frac{1}{t}\sum_{j=1}^{t}\sigma_j(w, w')Z_j + o\left(\frac{\tilde{S}_t^{3/8}\log(\tilde{S}_t)}{t}\right) \quad a.s., \tag{12}$$

where $\tilde{S}_t := \sum_{j=1}^{t}\sigma_j(w, w')^2$. Simplifying Equation (12), we have that

$$\frac{(\hat{\tau}_t(w, w') - \tau_t(w, w'))}{\sqrt{1/t^2 \sum_{j=1}^{t}\sigma_j(w, w')^2}} \approx N(0, 1).$$

Our final result in Theorem 3.1 follows directly after replacing a conservative or exact consistent estimator for the variance $\sigma_j(w, w')$ with $\hat{\sigma}_j(w, w')$ (akin to the difference between z-test and t-test). We show that we indeed get a consistent estimator in the last step of this proof. The remainder of the proof aims to prove Theorem 4.2.

Step 2: Building martingale using Gaussian distribution Recently, (Ramdas et al., 2020) shows that all sequential tests must have an explicit or implicit construction of a non-negative martingale. Although one of the major advantages of an asymptotic confidence sequences is that it avoids explicitly constructing a martingale, the proof still relies on constructing a martingale with the asymptotic Gaussian distribution. Consequently, the first step of the proof builds a martingale from a sequence of iid standard Gaussian random variables.

We note that

$$M_t(\lambda) := \exp\left(\sum_{j=1}^{t}(\lambda\sigma_j(w, w')Z_j - \lambda^2\sigma_j(w, w')^2/2)\right)$$

is a non-negative martingale starting at one for any $\lambda \in \mathbb{R}$ with respect to the canonical filtration (Robbins, 1970). For algebraic simplicity, we also define $L_t := \sum_{j=1}^{t}\sigma_j(w, w')Z_j$ and $\bar{\sigma}_t^2 = \frac{1}{t}\sum_{j=1}^{t}\sigma_j(w, w')^2$. Moreover, for any probability distribution $F(\lambda)$ on $\mathbb{R}$, we also have the mixture,

$$\int_{\lambda\in\mathbb{R}} M_t(\lambda)dF(\lambda)$$

is again a non-negative martingale with initial value one (Robbins, 1970). In particular, we consider the probability distribution function $f(\lambda; 0, \eta^2)$ for the Gaussian distribution with mean zero and variance η^2 as the mixing distribution. The resulting martingale is

$$\begin{aligned}
M_t &:= \int_{\lambda\in\mathbb{R}} M_t(\lambda)f(\lambda; 0, \eta^2)d\lambda \\
&= \frac{1}{\sqrt{2\pi\eta^2}}\int_{\lambda}\exp\left(\lambda L_t - \frac{t\lambda^2\bar{\sigma}_t^2}{2}\right)\exp\left(\frac{-\lambda^2}{2\eta^2}\right)d\lambda \\
&= \frac{1}{\sqrt{2\pi\eta^2}}\int_{\lambda}\exp\left(\lambda L_t - \frac{\lambda^2(1 + t\eta^2\bar{\sigma}_t^2)}{2\eta^2}\right)d\lambda \\
&= \frac{1}{\sqrt{2\pi\eta^2}}\int_{\lambda}\exp\left(\frac{-\lambda^2(1 + t\eta^2\bar{\sigma}_t^2) + 2\lambda\eta^2 L_t}{2\eta^2}\right)d\lambda \\
&= \frac{1}{\sqrt{2\pi\eta^2}}\int_{\lambda}\exp\left(\frac{-a(\lambda^2 + \frac{b}{a}2\lambda)}{2\eta^2}\right)d\lambda,
\end{aligned}$$

where $a = t\eta^2\bar{\sigma}_t^2 + 1$ and $b = \eta^2 L_t$. Completing the square, we have that the integrand is:

$$\exp\left(\frac{-a(\lambda^2 + \frac{b}{a}2\lambda)}{2\eta^2}\right) = \exp\left(\frac{-(\lambda - b/a)^2}{2\eta^2/a}\right)\exp\left(\frac{b^2}{2a\eta^2}\right).$$

Putting the expression back into M_t we have that,

$$\begin{aligned}
M_t &= \frac{1}{\sqrt{2\pi\eta^2/a}}\int_{\lambda}\exp\left(\frac{-(\lambda - b/a)^2}{2\eta^2/a}\right)d\lambda\frac{\exp\left(\frac{b^2}{2a\eta^2}\right)}{\sqrt{a}} \\
&= \exp\left(\frac{\eta^2(\sum_{j=1}^{t}\sigma_j(w, w')Z_j)^2}{2(t\bar{\sigma}_t^2\eta^2 + 1)}\right)(t\bar{\sigma}_t^2\eta^2 + 1)^{-1/2},
\end{aligned}$$

where the last line follows because the first part of the first line is one and we plug back in the definition of a and b.

Since M_t is a non-negative martingale with initial value one we can use Ville's maximal inequality Ville (1939) to claim that

$$\Pr(\forall t \geq 1, M_t < 1/\alpha) \geq 1 - \alpha,$$

which simplifies to the following after taking the logarithm and simple algebraic manipulation.

$$\Pr\left(\forall t \geq 1, \left|\frac{1}{t} L_t\right| < \sqrt{\frac{2(t\bar{\sigma}_t^2\eta^2 + 1)}{t^2\eta^2}\log\left(\frac{\sqrt{t\bar{\sigma}_t^2\eta^2 + 1}}{\alpha}\right)}\right) \geq 1 - \alpha \tag{13}$$

Combining Equation (13) and Equation (12) implies that with probability at least $(1 - \alpha)$,

$$\forall t \geq 1, \left|\frac{1}{t}\sum_{i=1}^{t} u_i\right| < \sqrt{\frac{2(t\bar{\sigma}_t^2\eta^2 + 1)}{t^2\eta^2}\log\left(\frac{\sqrt{t\bar{\sigma}_t^2\eta^2 + 1}}{\alpha}\right)} + o\left(R_t\right), \tag{14}$$

where $R_t = \tilde{S}_t^{3/8}\log(\tilde{S}_t)/t$. Using Assumption 3, we have that

$$\hat{\tau}_t(w, w') \pm \sqrt{\frac{2(t\bar{\sigma}_t^2\eta^2 + 1)}{t^2\eta^2}\log\left(\frac{\sqrt{t\bar{\sigma}_t^2\eta^2 + 1}}{\alpha}\right)} \tag{15}$$

forms an $(1 - \alpha)$-asymptotic confidence sequence, where we used Assumption 3 so that the $\tilde{V}_t/V_t \xrightarrow{a.s.} 1$ holds where $\tilde{V}_t$ is the non-asymptotic confidence width in Equation (14) (with the little o term) and V_t is defined in Equation (15) (without the little o term).

Step 3: Using empirical variance   Unfortunately, the confidence sequence in Equation (15) can not be directly used because $\bar{\sigma}_t$ is based off the true variance and hence not obtainable from the data. The last step is to replace Equation (15) with our estimated variance.

First, if $1/t \sum_{j=1}^{t} \hat{\sigma}_j(w, w')^2 \xrightarrow{a.s.} \bar{\sigma}_t^2$ (strongly consistent variance estimate), then we first show our confidence sequence in Theorem 4.2 forms a $(1 - \alpha)$-asymptotic confidence sequence for $\tau_t(w, w')$, giving us the desired result.

We begin by rewrite the assumption of $1/t \sum_{j=1}^{t} \hat{\sigma}_j(w, w')^2 \xrightarrow{a.s.} \bar{\sigma}_t^2$ as $\tilde{\sigma}_t^2 - \bar{\sigma}_t^2 = o(\bar{\sigma}_t^2)$, where $\tilde{\sigma}_t^2 = $

$1/t \sum_{j=1}^{t} \hat{\sigma}_j(w, w')^2$. Then, we have that our width from Equation (15)

$$\sqrt{\frac{2(t\bar{\sigma}_t^2\eta^2 + 1)}{t^2\eta^2}\log\left(\frac{\sqrt{t\bar{\sigma}_t^2\eta^2 + 1}}{\alpha}\right)}$$

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

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

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

$$= \sqrt{\left(\frac{t\tilde{\sigma}_t^2\eta^2 + 1}{t^2\eta^2} + o(\bar{\sigma}_t^2/t)\right)\log\left(\frac{t\tilde{\sigma}_t^2\eta^2 + o(t\bar{\sigma}_t^2) + 1}{\alpha^2}\right)}$$

Focusing on the second logarithmic term, we have

$$\log\left(\frac{t\tilde{\sigma}_t^2\eta^2 + o(t\bar{\sigma}_t^2) + 1}{\alpha^2}\right) = \log\left(\frac{t\tilde{\sigma}_t^2\eta^2 + 1}{\alpha^2} + o(t\bar{\sigma}_t^2)\right)$$

$$= \log\left(\frac{t\tilde{\sigma}_t^2\eta^2 + 1}{\alpha^2}\left[1 + o(1)\right]\right)$$

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

$$= \log\left(\frac{t\tilde{\sigma}_t^2\eta^2 + 1}{\alpha^2}\right) + o(1),$$

where the last line follows because $\log(1 + x) = x + o(1)$ for $|x| < 1$. Returning back to the main expression

we have

$$\sqrt{\frac{2(t\bar{\sigma}_t^2\eta^2+1)}{t^2\eta^2}\log\left(\frac{\sqrt{t\bar{\sigma}_t^2\eta^2+1}}{\alpha}\right)}$$

$$=\sqrt{\left(\frac{t\tilde{\sigma}_t^2\eta^2+1}{t^2\eta^2}+o(V_t/t^2)\right)\left[\log\left(\frac{t\tilde{\sigma}_t^2\eta^2+1}{\alpha^2}\right)+o(1)\right]}$$

$$=\sqrt{\frac{t\tilde{\sigma}_t^2\eta^2+1}{t^2\eta^2}\log\left(\frac{t\tilde{\sigma}_t^2\eta^2+1}{\alpha^2}\right)+o(V_t/t^2)+o(V_t\log V_t/t^2)+o(V_t/t^2)}$$

$$=\sqrt{\frac{t\tilde{\sigma}_t^2\eta^2+1}{t^2\eta^2}\log\left(\frac{t\tilde{\sigma}_t^2\eta^2+1}{\alpha^2}\right)+o(V_t\log V_t/t^2)}$$

$$=\sqrt{\frac{t\tilde{\sigma}_t^2\eta^2+1}{t^2\eta^2}\log\left(\frac{t\tilde{\sigma}_t^2\eta^2+1}{\alpha^2}\right)}+o(\sqrt{V_t\log V_t}/t),$$

where the last line follows because $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for $a,b\ge 0$. This formally shows how our confidence sequence in Theorem 4.2 is a valid $(1-\alpha)$−asymptotic confidence sequence for given that our variance estimator is strongly consistent.

However, Lemma 2.1 only tells us that $\tilde{\sigma}_t^2$ is conditionally unbiased for $\bar{\sigma}_t^2$. To establish the consistency result, we again use a version of strong law of large numbers for martingale sequence difference. We denote $U_t := \tilde{\sigma}_t^2 - \bar{\sigma}_t^2$. We remark that U_t is a martingale sequence difference with respect to the filtration $F_{T,t-1}$. Using classical results in (Chow, 1971), we have that $U_t \xrightarrow{a.s.} 0$ since Assumption 2 immediately satisfies the needed uniformly integrability condition. Since all the convergence statements above are almost-sure convergence, steps 1-3 give the desired claim.

C   Exact confidence sequences

C.1   Proof of Closed-form Exact Confidence Sequence

Proof. We build off the proof of Theorem 4 in (Howard et al., 2020). For brevity, we prove the theorem only for constructing valid confidence sequences for $\tau_t(w,w')$ and denote S_t for $S_t(w,w')$ for brevity. We first show that

$$\exp\left[\frac{\sum_{j=1}^t(\hat{\tau}_j(w,w')-\tau_j(w,w'))}{m(m+1)}+\frac{S_t}{m^2}\left(\log\left(\frac{m}{m+1}\right)+\frac{1}{m+1}\right)\right]$$

is a non-negative supermartingale with respect to the filtration $\mathcal{F}_{t-1}$. Fan, Grama and Liu (2015) shows that

$$\exp\left(\lambda\kappa+\kappa^2(\lambda+\log(1-\lambda))\right)\le 1+\lambda\kappa$$

for $\kappa\ge -1$ and $\lambda\in[0,1)$. We let

$$\kappa=\frac{\hat{\tau}_t(w,w')}{m},$$

where $\kappa \geq -1$ since $|\hat{\tau}_t(w, w')| \leq m$ for every t by Assumption 2. Let $h(\lambda) := (\lambda + \log(1 - \lambda))$ we have

$$\exp\left(\lambda \frac{\hat{\tau}_t(w, w')}{m} + \frac{\hat{\tau}_t(w, w')^2}{m^2} h(\lambda)\right) \leq 1 + \lambda \frac{\hat{\tau}_t(w, w')}{m}$$

$$E\left[\exp\left(\frac{\lambda \hat{\tau}_t(w, w')}{m} + \frac{\hat{\sigma}_t^2(w, w')}{m^2} h(\lambda)\right) \mid \mathcal{F}_{t-1}\right] \leq 1 + \lambda \frac{\tau_t(w, w')}{m}$$

$$E\left[\exp\left(\frac{\lambda(\hat{\tau}_t(w, w') - \tau_t(w, w'))}{m} + \frac{\hat{\sigma}_t^2(w, w')}{m^2} h(\lambda)\right) \mid \mathcal{F}_{t-1}\right] \leq 1$$

where the second line follows because $\hat{\tau}_t(w, w')^2 = \hat{\sigma}_t^2(w, w')$ and Lemma 2.1 and the last line follows because $1 - x \leq \exp(-x)$. We plug $\lambda = 1/(m + 1)$ and because the above is a non-negative quantity this directly implies that

$$\exp\left[\frac{\sum_{j=1}^t (\hat{\tau}_j(w, w') - \tau_j(w, w'))}{m(m + 1)} + \frac{S_t}{m^2}\left(\log\left(\frac{m}{m + 1}\right) + \frac{1}{m + 1}\right)\right]$$

is indeed a non-negative super martingale with respect to $\mathcal{F}_{t-1}$ as desired with initial value less than one. Therefore, we apply Ville's maximal inequality Ville (1939) to obtain

$$\Pr\left(\exists t : \exp\left[\frac{\sum_{j=1}^t u_j}{m(m + 1)} + \frac{S_n}{m^2}\left(\log\left(\frac{m}{m + 1}\right) + \frac{1}{m + 1}\right)\right] \geq \frac{1}{\tilde{\alpha}}\right)$$

$$= \Pr\left(\exists t : \left[\frac{\sum_{j=1}^t u_j}{m(m + 1)} + \frac{S_n}{m^2}\left(\log\left(\frac{m}{m + 1}\right) + \frac{1}{m + 1}\right)\right] \geq \log\left(\frac{1}{\tilde{\alpha}}\right)\right)$$

$$= \Pr\left(\exists t : \sum_{j=1}^t u_j \geq m(m + 1)\log\left(\frac{1}{\tilde{\alpha}}\right) - \frac{(m + 1)S_n}{m}\left(\log\left(\frac{m}{m + 1}\right) + \frac{1}{m + 1}\right)\right)$$

$$= \Pr\left(\exists t : \sum_{j=1}^t u_j \geq \left[m(m + 1)\log\left(\frac{1}{\tilde{\alpha}}\right) + S_n\left(\frac{m + 1}{m}\log\left(1 + \frac{1}{m}\right) - \frac{1}{m}\right)\right]\right)$$

is less than $\tilde{\alpha}$. This gives the one-sided confidence sequence and we can do the same trick and build the same statement instead for $\kappa = -\hat{\tau}_t(w, w')/m$. Taking $\alpha = \tilde{\alpha}/2$ and applying the union bound completes the proof. $\square$

C.2 Alternative exact confidence sequence

In this section, we correct the order of the confidence sequence presented above. We leverage the results presented in (Waudby-Smith et al., 2022b; Howard et al., 2020) by applying a mixture martingale over a truncated gamma distribution.

The above proof shows that

$$M_t := \exp\left(\lambda A_t + B_t(\lambda + \log(1 - \lambda))\right) \tag{16}$$

is a super-martingale with initial value 1, where

$$A_t := \frac{\sum_{j=1}^t u_j}{m}, \quad B_t := \frac{S_t}{m^2}.$$

For any distribution F on $(0, 1)$, we have by Fubini's theorem that

$$\tilde{M}_t := \int_{\lambda \in (0,1)} M_t dF(\lambda)$$

is again another super-martingale with initial value 1. Following the proof of Theorem 2 in (Waudby-Smith et al., 2022b), we choose the truncated gamma distribution given by

$$f(\lambda) = \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)}$$

for any $\rho \geq 0$. Therefore, we have that

$$\begin{aligned}
\tilde{M}_t &= \int_0^1 \exp\left\{\lambda A_t + B_t(\lambda + \log(1-\lambda))\right\} f(\lambda) d\lambda \\
&= \int_0^1 \exp\left\{\lambda A_t + B_t(\lambda + \log(1-\lambda))\right\} \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)} d\lambda \\
&= \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \int_0^1 \exp\{\lambda (A_t + B_t + \rho)\} (1-\lambda)^{B_t+\rho-1} d\lambda \\
&= \left(\frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)}\right) \left(\frac{1}{B_t + \rho}\right) {}_1F_1(1, B_t + \rho + 1, A_t + B_t + \rho),
\end{aligned}$$

where the last line follows from the definition of the Kummer's confluent hypergeometric function.

Therefore, we have Ville's maximal inequality Ville (1939) that

$$\Pr\left(\exists t : \left(\frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)}\right) \left(\frac{1}{B_t + \rho}\right) {}_1F_1(1, B_t + \rho + 1, A_t + B_t + \rho) \geq \frac{1}{\tilde{\alpha}}\right)$$

is less than α. Consequently, a one-sided lower confidence sequence can be obtained by a root-finding algorithm to find all

$$\{\tau_t(w, w') : V_t(w, w') \geq \frac{1}{\tilde{\alpha}}\},$$

where

$$V_t(w, w') := \left(\frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)}\right) \left(\frac{1}{B_t + \rho}\right) {}_1F_1(1, B_t + \rho + 1, A_t + B_t + \rho).$$

An upper confidence sequence can be obtained in a similar way. We remark that this confidence sequence does not solve the issue where it requires the analyst to know M and p_{min} before the experiment. Furthermore, this confidence sequence does not have a closed-form expression, thus it requires a root-solving algorithm to build the confidence sequence. However, (Waudby-Smith et al., 2022b) show that this provably has an asymptotic rate of $O(\sqrt{B_t \log(B_t)}/t)$, which does solve the issue related to the order of the confidence sequence width.

D   Optimizing and choosing hyper-parameter

In this section, we show in detail how an analyst can choose η to optimize the confidence sequence width for a desired specific time t^*. We remark that the derivations are nearly identical to those presented in Waudby-Smith et al. (2021), but we repeat them here for completeness.

Our proposed confidence sequence width presented in this paper all have the following structure

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

where we have omitted the variance terms and instead substituted each 1 since we want η to be data-independent.[5] We remark that

$$\operatorname*{argmin}_{\eta>0} B_t(\alpha) = \sqrt{\operatorname*{argmin}_{x>0} f(x)},$$

$$\text{where} \quad f(x) := \frac{tx + 1}{t^2x}\log\left(\frac{tx + 1}{\alpha^2}\right), \quad x := \eta^2.$$

Furthermore, $\lim_{x\to 0} f(x) = \lim_{x\to\infty} f(x)$ and thus if we can find the critical point by finding a solution for $\partial f/\partial x = 0$, then this must be the unique minimum.

Therefore, we have that

$$\frac{\partial f}{\partial x} = -\frac{1}{t^2x^2}\log\left(\frac{tx + 1}{\alpha^2}\right) + \frac{1}{tx}.$$

Setting the above to zero, we obtain

$$-\alpha^2\exp(1) = -(tx + 1)\exp(-(tx + 1)).$$

Therefore, we have that the solution is $-(tx + 1) = W_{-1}(-\alpha^2\exp(1))$, where W_{-1} is the lower branch of the Lambert W function. The solution only exists if

$$-\alpha^2\exp(1) \geq -\exp(1),$$

or equivalently if $\alpha^2 \leq 1$, which is always true for any $\alpha \in [0, 1]$. Therefore, we have that

$$\operatorname*{argmin}_{\eta>0} B_t(\alpha) = \sqrt{\frac{-W_{-1}(-\alpha^2\exp(1)) - 1}{t^*}}.$$

---

[5]Consequently, we are not formally optimizing η for the actual confidence width, but η can still be conceptually interpreted as minimizing the confidence sequence width at a desired time t^* (See Appendix C.3 in Waudby-Smith et al. (2021) for more details).