Efficient Adaptive Experimental Design for Average Treatment Effect Estimation

Masahiro Kato
CyberAgent Inc.
The University of Tokyo

Takuya Ishihara
The University of Tokyo

Junya Honda
The University of Tokyo
RIKEN

Yusuke Narita
Yale University

Abstract

The goal of many scientific experiments, including A/B testing, is to estimate the *average treatment effect* (ATE), which is defined as the difference between the expected outcomes of two or more treatments. In this paper, we consider a situation where an experimenter can assign a treatment to research subjects sequentially. In *adaptive experimental design*, the experimenter is allowed to change the probability of assigning a treatment using past observations for estimating the ATE efficiently. However, with this approach, it is not easy to apply a standard statistical method to construct an estimator because the observations are not independent and identically distributed. We thus propose an algorithm for efficient experiments with estimators constructed from dependent samples. We also introduce a *sequential testing* framework using the proposed estimator. To justify our proposed approach, we provide finite and infinite sample analyses. Finally, we experimentally show that the proposed algorithm exhibits preferable performance.

1 Introduction

Discovering causality from observations is a fundamental task in statistics and machine learning. In this paper, we follow Rubin (1974) to define a causal effect as the difference between the average outcomes resulting from two different actions, i.e., the *average treatment effect* (ATE). One of these actions corresponds to the *treatment* and the other corresponds to the *control* (Imbens & Rubin, 2015). One naive method for estimating the ATE using scientific experiments is the *randomized control trial* (RCT). In an RCT, we randomly assign one of the two actions to each research subject (Kendall, 2003) to obtain an unbiased estimator of the ATE (Imbens & Rubin, 2015).

However, while an RCT is a reliable method for scientific experiments, it often requires a large sample size for estimating the ATE precisely enough. To mitigate this problem, *adaptive experimental designs* have garnered increasing attention in various fields such as medicine and social science (Chow SC, 2005; van der Laan, 2008; Komiyama et al., 2009; Hahn et al., 2011; Chow & Chang, 2011; Villar et al., 2015; FDA, 2019). Compared to usual non-adaptive designs, adaptive designs often allow experimenters to detect the true causal effect while exposing fewer subjects to potentially harmful treatment. This motivates the US Food and Drug Administration (FDA) to recommend adaptive designs (FDA, 2019).

This paper proposes an adaptive experimental design that sequentially estimates a treatment assignment probability that minimizes the asymptotic variance of an estimator of the ATE and assigns a treatment according to the estimated probability. The proposed method is inspired by van der Laan

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

(2008) and Hahn et al. (2011), which conduct hypothesis testing based on the asymptotic distribution for samples gathered to minimize the asymptotic variance. In this paper, we show two directions of hypothesis testing based on the asymptotic distribution and the concentration inequality. The asymptotic variance-based hypothesis testing is an extension of the methods proposed by van der Laan (2008) and Hahn et al. (2011). The concentration inequality is a recently developed framework for hypothesis testing (Balsubramani & Ramdas, 2016), which is potentially useful in real-world applications, such as ad-optimization.

One of the theoretical difficulties comes from the dependency among samples. Because we update the assignment probability using past observations, samples are not *independent and identically distributed* (i.i.d.). Therefore, instead of using existing results under the i.i.d. assumption for deriving the theoretical properties, we use the theoretical results of *martingales*. The main contributions of this paper are as follows: (i) We establish a framework of causal inference from samples obtained from a time-dependent algorithm with theoretical properties. (ii) We propose an algorithm for scientific experiments that achieves the lower bound of the asymptotic variance with several statistical hypothesis testing methods. This paper thus contributes to the literature and practice of RCTs and A/B testing by proposing an efficient experimental design with theoretical guarantees.

Related Work: Among various methods for the adaptive experimental design, we share the motivation with Hahn et al. (2011) and van der Laan (2008). To the best of our knowledge, other existing studies have not pursued this direction of experimental design. Hahn et al. (2011) proposed the two-stage adaptive experimental design. Using the samples in the first stage, they estimated the conditional variance of outcomes to construct the optimal policy that minimizes the asymptotic variance of an estimator of the ATE (Proposition 1). In the second stage, they assigned the treatments to samples following the policy constructed in the first stage. There are three essential differences between them. First, because our method enables us to simultaneously construct the optimal policy and assign a treatment, we do not have to decide the sample size of the first stage in advance. Second, because of this property, our method and sequential testing introduced in Section 4 are compatible. Third, by applying martingale theory as van der Laan & Lendle (2014), we do not require Donsker's condition for the nuisance estimator. Our proposed method is also closely related to the method of van der Laan (2008). Compared with their method, our contributions are a generalization of the estimator using the martingale-based construction of the nuisance estimator and a proposition of nonparametric sequential testing based on concentration inequality. In addition, following Klaassen (1987), van der Laan & Lendle (2014), and Chernozhukov et al. (2018), we point out that Donkser's condition is unnecessary to the nuisance estimator.

Several existing studies offer statistical inference from samples with dependency (van der Laan, 2008; van der Laan & Lendle, 2014; Luedtke & van der Laan, 2016; Hadad et al., 2019). The asymptotic normality of our proposed estimator can be regarded as a special and simplified version of the estimator proposed by van der Laan (2008) and van der Laan & Lendle (2014), but the concentration inequality of the estimator is a new result. Although Hadad et al. (2019) recommended using adaptive weights for stabilization, we consider that we can control the behavior of the assignment probability by ourselves in our setting.

While the common goal of the MAB problem is to maximize the profit obtained from treatments, another framework called the *best arm identification* (BAI) aims to find actions with better rewards, whose motivation is similar to ours. For example, Yang et al. (2017) and Jamieson & Jain (2018) proposed a method to conduct a statistical test to find better actions using as a small sample size as possible. However, the approach is different. In the BAI without covariates, we typically compare the sample average of the rewards of each arm and tries to find an arm whose expected reward is the best among those arms with a high probability. On the contrary, in the best arm identification with covariates, we aim to find an arm whose expected reward conditional on the covariates is the best among the arms with a high probability. The problem setting in this paper shares the same goal as the best arm identification without covariates; however, we can also use the covariate information. In conclusion, the problem setting in this paper can be regarded as a new approach to the BAI. This setting can be called *semiparametric best arm identification*. Similarly, Deshpande et al. (2018) and Yao et al. (2020) consider controlling the power of the test in the MAB problem but have different motivations with different problem setting.

Balsubramani & Ramdas (2016) proposed nonparametric sequential testing based on the law of the iterated logarithm (LIL). We apply their results to adaptive ATE estimation with the AIPW estimator.

Some of the adversarial bandit algorithms also use IPW to obtain an unbiased estimator (Auer et al., 2003), but we have a different motivation. Further discussion of related work is in Appendix F.

Organization of this Paper: In the following sections, we introduce the proposed algorithm with its theoretical analysis and experimental results. First, in Section 2, we define the problem setting. In Section 3, we present a new estimator constructed from samples with dependency. In Section 4, we introduce sequential hypothesis testing, which can reduce the sample size compared with conventional hypothesis testing. We propose an algorithm for constructing an efficient estimator of the treatment effect in Section 5. Finally, in Section 6, we elucidate the empirical performance of the proposed algorithm using synthetic and semi-synthetic datasets.

2 Problem Setting

In the problem setting, a research subject arrives in a certain period, and an experimenter assigns a treatment to the research subject. For simplicity, we assume the immediate observation of the outcome of a treatment. After several trials, we decide whether the treatment has an effect.

2.1 Data Generating Process

We define the data generating process (DGP) as follows. In period $t \in \mathbb{N}$, a research subject visits an experimenter, and the experimenter assigns an action $A_t \in \mathcal{A} = \{0, 1\}$ based on the *covariate* $X_t \in \mathcal{X}$, where $\mathcal{X}$ denotes the space of the covariate. After assigning the action, the experimenter observes a reward $Y_t \in \mathbb{R}$ immediately, which has a potential outcome denoted by a random variable, $Y_t : \mathcal{A} \to \mathbb{R}$. We have access to a set $\mathcal{S}_T = \{(X_t, A_t, Y_t)\}_{t=1}^T$ with the following DGP:

$$\big\{(X_t, A_t, Y_t)\big\}_{t=1}^{T} \sim p(x)p_t(a \mid x, \Omega_{t-1})p(y \mid a, x), \tag{1}$$

where $Y_t = \mathbb{1}[A_t = 0]Y_t(0) + \mathbb{1}[A_t = 1]Y_t(1)$ for an indicator function $\mathbb{1}[\cdot]$, $p(x)$ denotes the density of the covariate X_t, $p_t(a \mid x, \Omega_{t-1})$ denotes the probability of assigning an action A_t conditioned on a covariate X_t, $p(y \mid a, x)$ denotes the density of an outcome Y_t conditioned on A_t and X_t, and $\Omega_{t-1} \in \mathcal{M}_{t-1}$ denotes the history defined as $\Omega_{t-1} = \{X_{t-1}, A_{t-1}, Y_{t-1}, \dots, X_1, A_1, Y_1\}$ with the space $\mathcal{M}_{t-1}$. We assume that $p(x)$ and $p(y \mid a, x)$ are invariant over time, but $p_t(a \mid x)$ can take different values. Further, we allow the decision maker to change $p_t(a \mid x, \Omega_{t-1})$ based on past observations. In this case, the samples $\big\{(X_t, A_t, Y_t)\big\}_{t=1}^{T}$ are correlated over time (i.e., the samples are not i.i.d.). The probability $p_t(a \mid x, \Omega_{t-1})$ is determined by a *policy* $\pi_t : \mathcal{A} \times \mathcal{X} \times \mathcal{M}_{t-1} \to (0, 1)$, which is a function of a covariate X_t, an action A_t, and a history Ω_{t-1}. For the policy $\pi_t(a \mid x, \Omega_{t-1})$, we consider the following process. First, we draw a random variable ξ_t following the uniform distribution on $[0, 1]$ in period t. Then, in each period t, we select an action A_t such that $A_t = \mathbb{1}[\xi_t \leq \pi_t(X_t, \Omega_{t-1})]$. Under this process, we regard the policy as the probability (i.e., $p_t(a \mid x, \Omega_{t-1}) = \pi_t(a \mid x, \Omega_{t-1})$).

Remark 1 (Observation of a Reward)**.** We assume that an outcome can be observed immediately after assigning an action. This setting is also referred to as *bandit feedback*. The case in which we observe a reward after some time can be regarded as a special case of bandit feedback.

2.2 Average Treatment Effect Estimation

Our goal is to estimate the treatment effect, which is a *counterfactual* value because we can only observe an outcome of an action when assigning the action. Therefore, following the causality formulated by Rubin (1974), we consider estimating the ATE between $d = 1$ and $d = 0$ as $\theta_0 = \mathbb{E}[Y_t(1) - Y_t(0)]$ (Imbens & Rubin, 2015). For identification of θ_0, we put the following assumption.

Assumption 1 (Boundedness)**.** There exist C_1 and C_2 such that $\frac{1}{p_t(a|x)} \leq C_1$ and $|Y_t| \leq C_2$.

Remark 2 (Stable Unit Treatment Value Assumption)**.** In the DGP, we assume that the *Stable Unit Treatment Value Assumption*, namely, $p(y \mid a, x)$, is invariant no matter what mechanism is used to assign an action (Rubin, 1986).

Remark 3 (Unconfoundedness)**.** Existing methods often make an assumption called unconfoundedness: the outcomes $(Y_t(1), Y_t(0))$ and the action A_t are conditionally independent on X_t. In the DGP, this assumption is satisfied because we choose an action based on the observed outcome.

Notations: Let k be an action in $\mathcal{A}$. Let us denote $\mathbb{E}[Y_t(k) \mid x]$, $\mathbb{E}[Y_t^2(k) \mid x]$, $\mathrm{Var}(Y_t(k) \mid x)$, and $\mathbb{E}[Y_t(1) - Y_t(0) \mid x]$ as $f^*(k,x)$, $e^*(k,x)$, $v^*(k,x)$, and $\theta_0(x)$, respectively. Let $\hat{f}_t(k,x)$ and $\hat{e}_t(k,x)$ be estimators of $f^*(k,x)$ and $e^*(k,x)$ constructed from Ω_t, respectively. Let $\mathcal{N}(\mu, \mathrm{var})$ be the normal distribution with the mean μ and the variance var.

2.3 Existing Estimators

We review three types of standard estimators of the ATE in the case in which we know the probability of assigning an action and the samples are i.i.d., that is, the probability of assigning an action is invariant as $p(a \mid x) = p_1(a \mid x, \Omega_0) = p_2(a \mid x, \Omega_1) = \cdots$. The first estimator is an *inverse probability weighting* (IPW) estimator given by $\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]Y_t}{p(1|X_t)} - \frac{\mathbb{1}[A_t=0]Y_t}{p(0|X_t)}\right)$ (Horvitz & Thompson, 1952; Rubin, 1987; Hirano et al., 2003; Swaminathan & Joachims, 2015). Although this estimator is unbiased when the behavior policy is known, it suffers from high variance. The second estimator is a direct method (DM) estimator $\frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}_t(1,X_t) - \hat{f}_t(0,X_t)\right)$ (Hahn, 1998). This estimator is known to be weak against model misspecification for $\mathbb{E}[Y_t(k) \mid X_t]$. The third estimator is an augmented IPW (AIPW) estimator (Robins et al., 1994; Chernozhukov et al., 2018) defined as $\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]\left(Y_t - \hat{f}_T(1,X_t)\right)}{p(1|X_t)} + \hat{f}_T(1,X_t) - \frac{\mathbb{1}[A_t=0]\left(Y_t - \hat{f}_T(0,X_t)\right)}{p(0|X_t)} - \hat{f}_T(0,X_t)\right)$. For the unbiasedness of the IPW and AIPW estimators, we can calculate the variance explicitly. The variance of the IPW estimator is $\left(\mathbb{E}\left[\frac{e^*(1,X_t)}{p(1|X_t)}\right] + \mathbb{E}\left[\frac{e^*(0,X_t)}{p(0|X_t)}\right] - \theta_0^2\right)/T$. The variance of the AIPW estimator is $\left(\mathbb{E}\left[\frac{v^*(1,X_t)}{p(1|X_t)}\right] + \mathbb{E}\left[\frac{v^*(0,X_t)}{p(0|X_t)}\right] + \mathbb{E}\left[(f^*(1,X_t) - f^*(0,X_t) - \theta_0)^2\right]\right)/T$, when $\hat{f}_T = f^*$. The asymptotic variances of the IPW and AIPW estimators are the same as their respective variances. Further, the variance and asymptotic variance are equal to the mean squared error (MSE) and asymptotic MSE (AMSE), respectively. As an important property, the (asymptotic) variance of the AIPW estimator achieves the lower bound of the asymptotic variance among regular $\sqrt{T}$-consistent estimators (van der Vaart, 1998, Theorem 25.20).

2.4 Semiparametric Efficiency

The lower bound of the variance is defined for an estimator under some posited models of the DGP. If this posited model is parametric, it is equal to the Cramér–Rao lower bound. When this posited model is a non- or semiparametric, we can still define the corresponding lower bound Bickel et al. (1998). As Narita (2018) showed, the semiparametric lower bound of (1) under $p_1(a \mid x) = p_2(a \mid x) = \cdots = p_T(a \mid x) = p(a \mid x)$ is given as $\mathbb{E}\left[\left\{\sum_{k=0}^{1}\frac{v\left(k,X_t\right)}{p(k|X_t)} + \left(\theta_0(X_t) - \theta_0\right)^2\right\}\right]$.

2.5 Efficient Policy

We consider minimizing the variance by appropriately optimizing the policy. Following Hahn et al. (2011), the efficient policies for IPW and AIPW estimators are given in the following proposition.

Proposition 1 (Efficient Probability of Assigning an Action)**.** For an IPW estimator, a probability minimizing the variance is given as $\pi^{\mathrm{IPW}}(1 \mid X_t) = \frac{\sqrt{e^*(1,X_t)}}{\sqrt{e^*(1,X_t)} + \sqrt{e^*(0,X_t)}}$. For an AIPW estimator, a probability minimizing the variance is given as $\pi^{\mathrm{AIPW}}(1 \mid X_t) = \frac{\sqrt{v^*(1,X_t)}}{\sqrt{v^*(1,X_t)} + \sqrt{v^*(0,X_t)}}$.

The derivation of an AIPW estimator is shown in Hahn et al. (2011). For an IPW estimator, we show the proof in Appendix B. In the following sections, we show that using the probability in Proposition 1, which minimizes the variance, we can also minimize the asymptotic variance and upper bound of the concentration inequality of appropriately defined estimators. Because the variance is equivalent to the MSE, a policy minimizing the variance also minimizes the MSE.

3 Adaptive Policy for Efficient ATE Estimation

As shown in the previous section, by setting the policy as $\pi_t(1 \mid x, \Omega_{t-1}) = \pi^{\mathrm{AIPW}}(1 \mid x) = \frac{\sqrt{v^*(1,x)}}{\sqrt{v^*(1,x)}+\sqrt{v^*(0,x)}}$, we can minimize the variance of the estimators. However, how to conduct statistical inference from the policy is unclear. There are two problems. First, we do not know $v^*(k,x) = \sigma^2(k,x)$. The second problem is how to conduct statistical inference from samples with dependency, which comes from the construction of $\pi_t(1 \mid x, \Omega_{t-1})$ (i.e., the estimation of $v^*(k,x)$). We solve the first problem by estimating $v^*(k,x)$ sequentially. For example, we can estimate $v^*(k,x) = e^*(k,x) - (f^*(k,x))^2$ by estimating $f^*(k,x)$ and $e^*(k,x)$. In this section, for solving the second problem, we propose estimators for samples with dependency and analyze the behavior of the estimators for infinite and finite samples.

3.1 Adaptive Estimators from Samples with Dependency

Here, we define the estimators constructed from samples with dependency. First, we define the adaptive IPW (AdaIPW) estimator as $\hat{\theta}^{\mathrm{AdaIPW}}_T = \frac{1}{T}\sum^T_{t=1}\left(\frac{\mathbb{1}[A_t=1]Y_t}{\pi_t(1|X_t,\Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]Y_t}{\pi_t(0|X_t,\Omega_{t-1})}\right)$. Second, we define the adaptive AIPW (A2IPW) estimator as $\hat{\theta}^{\mathrm{A2IPW}}_T = \frac{1}{T}\sum^T_{t=1} h_t$, where $h_t = \left(\frac{\mathbb{1}[A_t=1]\left(Y_t - \hat{f}_{t-1}(1,X_t)\right)}{\pi_t(1|X_t,\Omega_{t-1})} - \frac{\mathbb{1}[A_t=0]\left(Y_t - \hat{f}_{t-1}(0,X_t)\right)}{\pi_t(0|X_t,\Omega_{t-1})} + \hat{f}_{t-1}(1,X_t) - \hat{f}_{t-1}(0,X_t)\right)$. For $z_t = h_t - \theta_0$, $\{z_t\}^T_{t=1}$ is a martingale difference sequence (MDS), that is, $\mathbb{E}\big[z_t \mid \Omega_{t-1}\big] = \theta_0$. Using this property, we derive the theoretical results of $\hat{\theta}^{\mathrm{A2IPW}}_T$ in the following section. We omit the discussion for $\hat{\theta}^{\mathrm{AdaIPW}}_T$, but can derive the theoretical properties as well as $\hat{\theta}^{\mathrm{A2IPW}}_T$.

3.2 Asymptotic Distribution of A2IPW

For the A2IPW estimator $\hat{\theta}^{\mathrm{A2IPW}}_T$, we derive the asymptotic distribution.

Theorem 1 (Asymptotic Distribution of A2IPW)**.** *Suppose that (i) point convergence in probability of $\hat{f}_{t-1}$ and π_t, i.e., for all $x \in \mathcal{X}$ and $k \in \{0,1\}$, $\hat{f}_{t-1}(k,x) - f^*(k,x) \xrightarrow{\mathrm{p}} 0$ and $\pi_t(k \mid x, \Omega_{t-1}) - \tilde{\pi}(k \mid x) \xrightarrow{\mathrm{p}} 0$, where $\tilde{\pi} : \mathcal{A} \times \mathcal{X} \to (0,1)$ and (ii) there exits a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$. Then, under Assumption 1, for the A2IPW estimator, we have $\sqrt{T}\left(\hat{\theta}^{\mathrm{A2IPW}}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma^2 = \mathbb{E}\left[\sum^1_{k=0}\frac{\nu^*\left(k, X_t\right)}{\tilde{\pi}(k|X_t)} + \left(f^*(1,X_t) - f^*(0,X_t) - \theta_0\right)^2\right]$.*

The proof is shown in Appendix C. This result is a special case of van der Laan & Lendle (2014). Unlike van der Laan & Lendle (2014), we do not impose the convergence rate of $\hat{f}_{t-1}$ owing to the unbiasedness. As well as cross-fitting (Klaassen, 1987; Zheng & van der Laan, 2011; Chernozhukov et al., 2018), we do not have to impose Donsker condition as pointed by van der Laan & Lendle (2014). The asymptotic variance is semiparametric efficient under the policy $\tilde{\pi}$. It can also be regarded as the AMSE defined between θ_0 and $\hat{\theta}^{\mathrm{A2IPW}}_T$. In Appendix E, we show the corresponding estimator and its asymptotic variance for the off-policy evaluation, which is a generalization of the ATE estimation. Finally, we also show the consistency by using the weak law of large numbers for an MDS (Proposition 4 in Appendix A). We omit the proof because we can easily show it from the boundedness of z_t.

Theorem 2 (Consistency of A2IPW)**.** *Suppose that there exits a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$. Then, under Assumption 1, $\hat{\theta}^{\mathrm{A2IPW}}_T \xrightarrow{\mathrm{p}} \theta_0$.*

3.3 Regret Bound of A2IPW

For the finite sample analysis, instead of asymptotic theory, we introduce the *regret analysis* framework often used in the literature on the MAB problem. In this paper, we define regret based on the MSE. We define the optimal policy Π^{OPT} as a policy that chooses a treatment with the probability π^{AIPW} defined in (1), and an estimator $\hat{\theta}^{\mathrm{OPT}}_T$ with oracle f^* as $\hat{\theta}^{\mathrm{OPT}}_T =$

$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mathbb{1}[A_t=1]\left(Y_t-f^*(1,X_t)\right)}{\pi^{\mathrm{AIPW}}(1|X_t)}-\frac{\mathbb{1}[A_t=0]\left(Y_t-f^*(0,X_t)\right)}{1-\pi^{\mathrm{AIPW}}(1|X_t)}+f^*(1,X_t)-f^*(0,X_t)\right)$. Then, for a policy Π adapted by the experimenter, we define the regret of between Π and Π^{OPT} as $\texttt{regret} = \mathbb{E}_{\Pi}\left[\left(\theta_0-\hat{\theta}_T^{\mathrm{A2IPW}}\right)^2\right]-\mathbb{E}_{\Pi^{\mathrm{OPT}}}\left[\left(\theta_0-\hat{\theta}_T^{\mathrm{OPT}}\right)^2\right]$, where the expectations are taken over each policy. The upper bound is in the following theorem.

Theorem 3 (Regret Bound of A2IPW)**.** *Suppose that there exits a constant C_3 such that $|\hat{f}_{t-1}| \leq C_3$. Then, under Assumption 1, the* $\texttt{regret}$ *is bounded by* $\frac{1}{T^2}\sum_{t=1}^{T}\sum_{k=0}^{1}\left\{\mathrm{O}\left(\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k\mid X_t)}-\sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}\right|\right]\right)+\mathrm{O}\left(\mathbb{E}\left[\left|f^*(k,X_t)-\hat{f}_{t-1}(k,X_t)\right|\right]\right)\right\}$, *where the expectation is taken over the random variables including Ω_{t-1}.*

The proof is shown in Appendix D. Then, by substituting the finite sample bounds of $\mathbb{E}\left[\left|\sqrt{\pi^{\mathrm{AIPW}}(k\mid X_t)}-\sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}\right|\right]$ and $\mathbb{E}\left[\left|f^*(k,X_t)-\hat{f}_{t-1}(k,X_t)\right|\right]$, the regret bound for finite samples can be obtained. We can bound $\hat{f}_{t-1}(k,X_t)$ and $\sqrt{\pi_t(k\mid X_t,\Omega_{t-1})}$ by the same argument as existing work on the MAB problem such as Yang & Zhu (2002).

Remark 4. This result tells us that regret is bounded by $\mathrm{o}(1/T)$ under the appropriate convergence rates of π_t and $\hat{f}_t$. By contrast, if we use a constant value for π_t, regret is $\mathrm{O}(1/T)$.

4 Sequential Hypothesis Testing with A2IPW Estimator

The goal of various applications including A/B testing is to conduct decision making between *null* ($\mathcal{H}_0$) and an *alternative* ($\mathcal{H}_1$) hypothesis while controlling both *false positives* (*Type I error*) and *false negatives* (*Type II error*). *Standard hypothesis testing* generates a confidence interval based on a fixed sample size T. In this case, we can use the asymptotic distribution derived in Theorem 1. On the contrary, for the case in which samples arrive in a stream, there is interest in conducting decision making without waiting for the sample size to reach T. Under this motivation, we discuss *sequential hypothesis testing*, which decides to accept or reject the null hypothesis at any time $t = 1, 2, \dots, T$. The preliminaries of the hypothesis testing are in Appendix G.

4.1 Sequential Testing and Control of Type I error

In sequential testing, we sequentially conduct decision making and stop whenever we want (Wald, 1945). However, if we sequentially conduct standard hypothesis testing based on the p-value defined for fixed sample size, the probability of the Type I error increases (Balsubramani & Ramdas, 2016). Therefore, the main issue of sequential testing is to control the Type I error, and various approaches have been proposed (Wald, 1945). One classical method is to correct the p-value based on multiple testing corrections, such as the Bonferroni (BF) and Benjamini–Hochberg procedures. For example, when we conduct standard hypothesis testing at $t = 100, 200, 300, 400, 500$ by constructing the corresponding p-values of p_{100}, p_{200}, p_{300}, p_{400}, and p_{500}, the BF procedure corrects the p-values to p_{100}, $p_{200}/2$, $p_{300}/3$, $p_{400}/4$, and $p_{500}/5$. Although this correction enables us to control the Type I error, it is also known to be exceedingly conservative and tends to produce suboptimal results (Balsubramani & Ramdas, 2016; Jamieson & Jain, 2018). Further, owing to this conservativeness, we cannot conduct decision making in each period. For example, in the case in which we conduct standard hypothesis testing in period $t = 1, 2, 3, \dots, t, \dots$, the corresponding p-values become too small $(p_1, p_2/2, p_3/3, p_4/4, \dots, p_t/t, \dots)$. Therefore, when conducting sequential testing based on multiple testing, we need to split the stream of samples into several batches (Balsubramani & Ramdas, 2016). To avoid the drawback of multiple testing, recent work has proposed using *adaptive concentration inequalities* for an adaptively chosen number of samples (i.e., the inequality holds at any randomly chosen $t = 1, 2, \dots$) (Balsubramani, 2014; Jamieson et al., 2014; Johari et al., 2015; Balsubramani & Ramdas, 2016; Zhao et al., 2016; Jamieson & Jain, 2018). This concentration inequality enables us to conduct sequential testing without separating samples into batches while controlling the Type I error under appropriate conditions.

There are two approaches for introducing such concentration inequalities into sequential testing: *confidence sequence* (Darling & Robbins, 1967; Lai, 1984; Zhao et al., 2016) and *always valid p-values* (Johari et al., 2015; Jamieson & Jain, 2018). These two approaches are equivalent, as shown by Ramdas (2018), and we adapt the former herein. For simplicity, let us define the null and alternative hypotheses as $\mathcal{H}_0 : \theta_0 = \mu$ and $\mathcal{H}_1 : \theta_0 \neq \mu$, respectively, where μ is a constant, and consider controlling the Type I error at α. Then, for the A2IPW estimator $\hat{\theta}^{\mathrm{A2IPW}}_t$ of θ_0, we define a sequence of positive values $\big\{q_t\big\}^T_{t=1}$, which satisfies $\mathbb{P}(\exists t \in \mathbb{N} : t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu > q_t) \leq \alpha$ when the null hypothesis is true. Using $\big\{q_t\big\}^T_{t=1}$, we consider the following process: if $t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu > q_t$, we reject the null hypothesis $\mathcal{H}_0$; otherwise, we temporally accept the null hypothesis $\mathcal{H}_0$. Because $\big\{q_t\big\}_{t\in\mathbb{N}}$ satisfies $\mathbb{P}\big(\text{reject } \mathcal{H}_0\big) = \mathbb{P}\left(\exists t \in \mathbb{N} : |t\hat{\theta}^{\mathrm{A2IPW}}_t - t\mu| > q_t\right) \leq \alpha$ when the null hypothesis is true, we can control the Type I error at α. This procedure of hypothesis testing has some desirable properties. First, it controls the Type I error with α in any period t. Second, the Type II error of the hypothesis testing with this procedure is less than or equal to that under standard hypothesis testing (Balsubramani & Ramdas, 2016). Third, it enables us to stop the experiment whenever we obtain sufficient samples for decision making.

4.2 Sequential Testing with Law of Iterated Logarithm (LIL)

Next, we consider constructing $\big\{q_t\big\}_{t\in\mathbb{N}}$ with the Type I error α using the proposed A2IPW estimator. Among the various candidates, concentration inequalities based on the LIL have garnered attention recently. The LIL was originally derived as an asymptotic property of independent random variables by Khintchine (1924) and Kolmogoroff (1929). Following their methods, several works have derived an asymptotic LIL for an MDS under some regularity conditions (Stout, 1970; Fisher, 1992), and Balsubramani & Ramdas (2016) derived a nonasymptotic LIL-based concentration inequality for hypothesis testing. Sequential testing with the LIL-based confidence sequence $\big\{q_t\big\}_{t\in\mathbb{N}}$ requires the smallest sample size needed to identify the parameter of interest (Jamieson et al., 2014; Balsubramani & Ramdas, 2016). For this tightness of the inequality, LIL-based concentration inequalities have been widely accepted in sequential testing (Balsubramani & Ramdas, 2016) and in the best arm identification in the MAB problem (Jamieson et al., 2014; Jamieson & Jain, 2018). Therefore, we also construct the confidence sequence $\big\{q_t\big\}_{t\in\mathbb{N}}$ based on the LIL-based concentration inequality for the A2IPW estimator derived in the following theorem.

Theorem 4 (Concentration Inequality of A2IPW Estimator)**.** *Suppose that there exists C such that $|z_t| \leq C$. Suppose that there exists C_4 such that $|(z_t - z_{t-1})^2 - \mathbb{E}[(z_t - z_{t-1})^2 \mid \Omega_{t-1}]| \leq C_4$. For any δ, with probability $\geq 1 - \delta$, for all $t \geq \tau_0$ simultaneously, $\left|\sum^t_{i=1} z_i\right| = \left|t\hat{\theta}^{\mathrm{A2IPW}}_t - t\theta_0\right| \leq \frac{2C}{e^2}\left(C_0(\delta) + \sqrt{2C_1\hat{V}^*_t\left(\log\log\hat{V}^*_t + \log\left(\frac{4}{\delta}\right)\right)}\right)$, where $\hat{V}^*_t = C_3\left(\frac{e^4}{4C^2}\sum^t_{i=1} z^2_i + \frac{2C_0(\delta)C_4}{e^2}\right)$, $C_0(\delta) = 3(e-2) + 2\sqrt{\frac{173}{2(e-2)}}\log\left(\frac{4}{\delta}\right)$, $C_1 = 6(e-2)$ and C_3 is an absolute constant.*

We can obtain this result by applying the result of Balsubramani (2014). The proof is in Appendix D.1. Then, we obtain confidence sequences, $\big\{q_t\big\}^T_{t=1}$, with the Type I error at α from the results of Theorem 4 and Balsubramani & Ramdas (2016) as $q_t \propto \log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum^t_{i=1} z^2_i\left(\log\frac{\log\sum^t_{i=1} z^2_i}{\alpha}\right)}$. Balsubramani & Ramdas (2016) proposed using the constant 1.1 to specify q_t, namely, $q_t = 1.1\left(\log\left(\frac{1}{\alpha}\right) + \sqrt{2\sum^t_{i=1} z^2_i\left(\log\frac{\log\sum^t_{i=1} z^2_i}{\alpha}\right)}\right)$. This choice is motivated by the asymptotic property of the LIL such that $\limsup_{t\to\infty}\frac{\left|t\hat{\theta}^{\mathrm{A2IPW}}_t - t\theta_0\right|}{\sqrt{2\tilde{V}^*_t\left(\log\log\tilde{V}^*_t\right)}} = 1$ with probability 1 for sufficiently large samples (Stout, 1970; Balsubramani & Ramdas, 2016), where $\tilde{V}^2_t = \sum^t_{i=1}\mathbb{E}[z^2_i \mid \Omega_{i-1}]$, and the empirical results of Balsubramani & Ramdas (2016).

5 Main Algorithm: AERATE

In this section, we define our main algorithm, referred to as *Adaptive ExpeRiments for efficient ATE estimation* (AERATE). The details are in Appendix H. First, we consider estimating $f^*(a,x) = \mathbb{E}\big[Y_t(a) \mid x\big]$ and $e^*(a,x) = \mathbb{E}\big[Y^2_t(a) \mid x\big]$. When estimating $f^*(a,x)$ and $e^*(a,x)$, we need to construct consistent estimators from dependent samples obtained from an adaptive policy.

Table 1: Experimental results using Datasets 1–2. The best performing method is in bold.

	Dataset 1: $\mathbb{E}[Y(1)] = 0.8$, $\mathbb{E}[Y(0)] = 0.3$, $\theta_0 \neq 0$								Dataset 2: $\mathbb{E}[Y(1)] = 0.5$, $\mathbb{E}[Y(0)] = 0.5$, $\theta_0 = 0$							
	$T = 150$			$T = 300$			ST		$T = 150$			$T = 300$			ST	
	MSE	STD	Testing	MSE	STD	Testing	LIL	BF	MSE	STD	Testing	MSE	STD	Testing	LIL	BF
RCT	0.145	0.178	25.0%	0.073	0.100	46.0%	455.4	370.4	0.084	0.129	4.7%	0.044	0.062	4.9%	497.2	481.8
A2IPW (K-nn)	0.085	0.116	38.4%	0.038	0.054	67.9%	389.5	302.8	0.050	0.071	5.6%	0.026	0.037	5.6%	497.2	477.3
A2IPW (NW)	0.064	0.092	51.4%	0.025	0.035	88.1%	303.8	239.8	**0.029**	0.045	**4.4%**	**0.012**	0.018	4.7%	496.2	480.6
MA2IPW (K-nn)	0.092	0.126	38.5%	0.044	0.058	66.2%	387.5	303.4	0.052	0.073	5.4%	0.025	0.034	4.7%	497.9	477.0
MA2IPW (NW)	**0.062**	0.085	52.7%	**0.023**	0.033	90.2%	303.3	236.6	0.032	0.047	6.3%	**0.012**	0.018	4.4%	496.6	475.3
AdaIPW (K-nn)	0.151	0.208	26.1%	0.075	0.103	43.6%	446.3	367.0	0.088	0.126	5.6%	0.043	0.062	5.2%	495.8	478.1
AdaIPW (NW)	0.161	0.232	23.4%	0.081	0.115	41.1%	446.6	375.0	0.094	0.140	5.8%	0.045	0.064	5.3%	495.6	471.6
DM (K-nn)	0.175	0.252	**88.7%**	0.086	0.126	**96.1%**	**59.9**	**164.6**	0.096	0.129	85.3%	0.046	0.063	89.5%	97.3	188.3
DM (NW)	0.111	0.167	82.1%	0.045	0.066	95.6%	119.6	176.2	0.054	0.075	53.7%	0.023	0.032	55.4%	312.8	305.3
Hahn 50 (K-nn)	0.109	0.149	35.2%	0.046	0.064	63.3%	398.5	316.0	0.060	0.089	5.4%	0.029	0.041	6.6%	493.8	473.4
Hahn 50 (NW)	0.085	0.128	45.7%	0.033	0.046	82.8%	313.1	257.0	0.040	0.057	5.6%	0.016	0.025	6.9%	493.7	477.7
Hahn 100 (K-nn)	0.141	0.200	29.6%	0.057	0.081	60.%	408.2	332.6	0.071	0.104	6.3%	0.029	0.044	5.2%	495.2	475.6
Hahn 100 (NW)	0.107	0.146	32.1%	0.036	0.050	75.2%	365.3	294.6	0.043	0.063	4.8%	0.014	0.019	**3.7%**	**498.2**	**483.5**
OPT	0.008	0.011	100.0%	0.004	0.005	100.0%	63.9	150.0	0.005	0.007	4.4%	0.002	0.003	4.4%	498.4	483.0

In a MAB problem, several nonparametric estimators are consistent, such as the K-nearest neighbor regression estimator and Nadaraya–Watson kernel regression estimator (Yang & Zhu, 2002; Qian & Yang, 2016).

For simplicity, we only show the algorithm using A2IPW, and we can derive the procedure when using the AdaIPW estimator similarly. The proposed algorithm consists of three main steps: in period t, (i) estimate $\nu(k, x)$ using nonparametric estimators in the MAB problem (Yang & Zhu, 2002; Qian & Yang, 2016); (ii) assign an action with an estimator of the optimal policy, which is defined as $\pi^{\mathrm{A2IPW}}(1 \mid x) = \frac{\sqrt{\nu^*(1,x)}}{\sqrt{\nu^*(1,x)}+\sqrt{\nu^*(0,x)}}$; and (iii) conduct testing when sequential testing is chosen as the hypothesis testing method. Moreover, to stabilize the algorithm, we introduce the following three elements: (a) the estimator $\hat{\nu}_{t-1}(k, x)$ of $\nu^*(k, x)$ is constructed as $\max\left(\underline{\nu}, \hat{e}_{t-1}(k, x) - \hat{f}^2_{t-1}(k, x)\right)$, where $\underline{\nu}$ is the lower bound of ν^*, and $\hat{f}_{t-1}$ and $\hat{e}_{t-1}$ are the estimators of f^* and e^* only using Ω_{t-1}, respectively; (b) let a policy be $\pi_t(1 \mid x, \Omega_{t-1}) = \gamma\frac{1}{2} + (1-\gamma)\frac{\sqrt{\hat{\nu}_{t-1}(1,x)}}{\sqrt{\hat{\nu}_{t-1}(1,x)}+\sqrt{\hat{\nu}_{t-1}(0,x)}}$, where $\gamma = \mathrm{O}(1/\sqrt{T})$; and (c) as a candidate of the estimators, we also propose the mixed A2IPW (MA2IPW) estimator defined as $\hat{\theta}^{\mathrm{MA2IPW}}_t = \zeta\hat{\theta}^{\mathrm{AdaIPW}}_t + (1-\zeta)\hat{\theta}^{\mathrm{A2IPW}}_t$, where $\zeta = \mathrm{o}(1/\sqrt{t})$. The motivation of (a) is to prevent $\hat{\nu}_{t-1}$ from taking a negative value or zero technically, and we do not require accurate knowledge of the lower bound. The motivation of (b) is to stabilize the probability of assigning an action. The motivation of (c) is to control the behavior of an estimator by avoiding the situation in which $\hat{f}_{t-1}$ takes an unpredicted value in the early stage. Because the nonparametric convergence rate is lower bounded by $\mathrm{O}(1/\sqrt{t})$ in general, the convergence rate of π_t to π^{AIPW} is also upper bounded by $\mathrm{O}(1/\sqrt{t})$. Therefore, $\gamma = \mathrm{O}(1/\sqrt{t})$ does not affect the convergence rate of the policy. Similarly, the asymptotic distribution of $\hat{\theta}^{\mathrm{MA2IPW}}_T$ is the same as $\hat{\theta}^{\mathrm{A2IPW}}_T$. The pseudo code is in Appendix H.

6 Experiments

In this section, we show the effectiveness of the proposed algorithm experimentally. We compare the proposed AdaIPW, A2IPW, and MA2IPW estimators in AERATE with an RCT with $p(A_t = 1|X_t) = 0.5$, the method of Hahn et al. (2011), the estimator $\hat{\theta}^{\mathrm{OPT}}_T$ under the optimal policy, and the standard DM estimators. To best our knowledge, there is no recent method proposed in this problem setting. In AERATE, we set $\gamma = 1/\sqrt{t}$. For the MA2IPW estimator, we set $\zeta = t^{-1/1.5}$. When estimating f^* and e^*, we use K-nearest neighbor regression and Nadaraya–Watson regression. In the method of Hahn et al. (2011), we first use 50 and 100 samples to estimate the optimal policy. In this experiment, we use synthetic and semi-synthetic datasets. In each dataset, we conduct the following three patterns of hypothesis testing. For all the settings, the null and alternative hypotheses are $\mathcal{H}_0 : \theta_0 = 0$ and $\mathcal{H}_1 : \theta_0 \neq 0$, respectively. We conduct standard hypothesis testing with T-statistics when the sample sizes are 250 and 500, sequential testing based on multiple testing with the BF correction when the sample sizes are 150, 250, 350, and 450, and sequential testing with the LIL based on the concentration inequality shown in Theorem 4.

First, we conducted an experiment using the following synthetic datasets. We generated a covariate $X_t \in \mathbb{R}^5$ at each round as $X_t = (X_{t1}, X_{t2}, X_{t3}, X_{t4}, X_{t5})^\top$, where $X_{tk} \sim \mathcal{N}(0, 1)$ for $k =$

$1, 2, 3, 4, 5$. In this experiment, we used $Y_t(d) = \mu_d + \sum^{5}_{k=1} X_{tk} + e_{td}$ as a model of a potential outcome, where μ_d is a constant, e_{td} is the error term, and $\mathbb{E}[Y_t(d)] = \mu_d$ The error term e_{td} follows the normal distribution, and we denote the standard deviation as std_d. We made two datasets with different μ_d and std_d, Datasets 1–2, with 500 periods (samples). For Datasets 1, we set $\mu_1 = 0.8$ and $\mu_0 = 0.3$ with $\mathrm{std}_1 = 0.8$ and $\mathrm{std}_1 = 0.3$. For Datasets 1, we set $\mu_1 = \mu_0 = 0.5$ with $\mathrm{std}_1 = 0.8$ and $\mathrm{std}_1 = 0.3$. We ran 1000 independent trials for each setting. The results of experiment are shown in Table 1. We show the MSE between θ and $\hat{\theta}$, the standard deviation of MSE (STD), and percentages of rejections of hypothesis testing using T-statistics at the 150th (mid) round and the 300th (final) periods. Besides, we also showed the stopping time of the LIL based algorithm (LIL) and multiple testing with BF correction. When using BF correction, we conducted testing at $t = 150, 250, 350, 450$. In sequential testing, if we do not reject the hypothesis, we return the stopping time as 500. In many datasets, the proposed algorithm achieves the lower MSE than an the other methods. The DM estimators rejects the null hypothesis with small samples in Dataset 1, but also often reject the null hypothesis in Dataset II, i.e, the Type II error is large. The details of experiments is shown in Appendix I.

Appendix I shows the additional experimental results. In Appendix I, we investigate the performance of the proposed algorithm for other synthetic and semi-synthetic datasets constructed from the Infant Health and Development Program (IHDP). The IHDP dataset consists of simulated outcomes and covariate data from a real study following the simulation proposed by Hill (2011). In the IHDP data, we reduce the sample size by $1/5$ compared with the RCT.

7 Conclusion

In this paper, we proposed an algorithm of the MAB problem that yields an efficient estimator of the treatment effect. Using martingale theory, we derived the theoretical properties of the proposed algorithm for cases with both infinite and finite samples with the framework of sequential testing.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Math. J. (2)*, 19(3): 357–367, 1967.

Balsubramani, A. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Balsubramani, A. and Ramdas, A. Sequential nonparametric testing with the law of the iterated logarithm. In *UAI*, pp. 42–51. AUAI Press, 2016.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, 1998.

Casella, G. *Statistical inference*. Duxbury advanced series. Duxbury/Thomson Learning, Australia ; Pacific Grove, Calif., 2nd ed. edition, 2002.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Chow, S.-C. and Chang, M. *Adaptive Design Methods in Clinical Trials*. Chapman and Hall/CRC, 2 edition, 2011.

Chow SC, Chang M, P. A. Statistical consideration of adaptive methods in clinical development. *J Biopharm Stat*, 2005.

Darling, D. A. and Robbins, H. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967.

Deshpande, Y., Mackey, L., Syrgkanis, V., and Taddy, M. Accurate inference for adaptive linear models. In *ICML*, pp. 1194–1203, 2018.

Durrett, R. *Probability: Theory and Examples*. Cambridge University Press, USA, 4th edition, 2010.

FDA. Adaptive designs for clinical trials of drugs and biologics: Guidance for industry. Technical report, U.S. Department of Health and Human Services Food and Drug Administration (FDA), Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), 2019.

Fisher, E. On the law of the iterated logarithm for martingales. *The Annals of Probability*, 20(2): 675–680, 1992.

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

Hahn, J. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.

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

Hall, P. and Hayde, C. *Martingale Limit Theory and Its Application*. Probability and mathematical statistics. Academic Press, 1980.

Hall, P., Heyde, C., Birnbaum, Z., and Lukacs, E. *Martingale Limit Theory and Its Application*. Communication and Behavior. Elsevier Science, 2014.

Hamilton, J. *Time series analysis*. Princeton Univ. Press, Princeton, NJ, 1994.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hirano, K., Imbens, G. W., and Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

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

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

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *COLT*, volume 35, 2014.

Jamieson, K. G. and Jain, L. A bandit approach to sequential experimental design with false discovery control. In *NeurIPS*, pp. 3664–3674. Curran Associates, Inc., 2018.

Johari, R., Pekelis, L., and Walsh, D. J. Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*, 2015.

Kallus, N. and Uehara, M. Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*, 2019.

Kendall, J. M. Designing a research project: randomised controlled trials and their principles. *Emergency Medicine Journal*, 20(2):164–168, 2003.

Khintchine, A. Über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, 6(1): 9–20, 1924.

Klaassen, C. A. J. Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 1987.

Kolmogoroff, A. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101:126–135, 1929.

Komiyama, O., Koshimizu, T., Suganami, H., Sakai, H., Orhihashi, Y., and Tomiyama, H. Adaptive designs in clinical drug development: An executive summary of the phrma working group. *Rinsho yakuri/Japanese Journal of Clinical Pharmacology and Therapeutics*, 40(6):303–310, 2009.

Kosorok, M. R. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer New York, New York, NY, 2008.

Lai, T. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: a sequential approach. *Communications in Statistics - Theory and Methods*, 13(19):2355–2368, 1984.

Loeve, M. *Probability Theory*. Graduate Texts in Mathematics. Springer, 1977.

Luedtke, A. R. and van der Laan, M. J. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of statistics*, 2016.

Nardini, C. The ethics of clinical trials. *Ecancermedicalscience*, 8:387, 2014.

Narita, Y. Experiment-as-Market: Incorporating Welfare into Randomized Controlled Trials. Cowles Foundation Discussion Papers 2127, Cowles Foundation for Research in Economics, Yale University, 2018.

Qian, W. and Yang, Y. Kernel estimation and model combination in a bandit problem with covariates. *Journal of Machine Learning Research*, 17(149):1–37, 2016.

Ramdas, A. Sequential testing, always valid p-values. *Martingales 2 : Sequential Analysis*, 2018.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Rubin, D. B. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, 1987.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *NeurIPS*, pp. 828–836. Curran Associates, Inc., 2014.

Stout, W. F. A martingale analogue of kolmogorov's law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 15(4):279–290, 1970.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16:1731–1755, 2015.

Tomking, R. J. Some iterated logarithm results related to the central limit theorem. *Transactions of the American Mathematical Society*, 156, 1971.

van der Laan, M. J. The construction and analysis of adaptive group sequential designs. 2008.

van der Laan, M. J. and Lendle, S. D. Online targeted learning. 2014.

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

Villar, S., Bowden, J., and Wason, J. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30:199–215, 2015.

Wald, A. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 1945.

Yang, F., Ramdas, A., Jamieson, K. G., and Wainwright, M. J. A framework for multi-a(rmed)/b(andit) testing with online fdr control. In *NeurIPS*, pp. 5957–5966. Curran Associates, Inc., 2017.

Yang, Y. and Zhu, D. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.*, 30(1):100–121, 2002.

Yao, J., Brunskill, E., Pan, W., Murphy, S., and Doshi-Velez, F. Power-constrained bandits, 2020.

Zhao, S., Zhou, E., Sabharwal, A., and Ermon, S. Adaptive concentration inequalities for sequential decision problems. In *NeurIPS*, pp. 1343–1351. Curran Associates, Inc., 2016.

Zheng, W. and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics. 2011.