Statistical Inference with M-Estimators on Adaptively Collected Data

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

Abstract

Bandit algorithms are increasingly used in real-world sequential decision-making problems. Associated with this is an increased desire to be able to use the resulting datasets to answer scientific questions like: Did one type of ad lead to more purchases? In which contexts is a mobile health intervention effective? However, classical statistical approaches fail to provide valid confidence intervals when used with data collected with bandit algorithms. Alternative methods have recently been developed for simple models (e.g., comparison of means). Yet there is a lack of general methods for conducting statistical inference using more complex models on data collected with (contextual) bandit algorithms; for example, current methods cannot be used for valid inference on parameters in a logistic regression model for a binary reward. In this work, we develop theory justifying the use of M-estimators—which includes estimators based on empirical risk minimization as well as maximum likelihood—on data collected with adaptive algorithms, including (contextual) bandit algorithms. Specifically, we show that M-estimators, modified with particular adaptive weights, can be used to construct asymptotically valid confidence regions for a variety of inferential targets.

1. Introduction

Due to the need for interventions that are personalized to users, (contextual) bandit algorithms are increasingly used to address sequential decision making problems in healthcare (Yom-Tov et al., 2017; Liao et al., 2020), online education (Liu et al., 2014; Shaikh et al., 2019), and public policy (Kasy and Sautmann, 2021). Contextual bandits personalize, that is, minimize regret, by learning to choose the best intervention in each context, i.e., the action that leads to the greatest expected reward. Besides the goal of regret minimization, another critical goal in these real-world problems is to be able to use the resulting data collected by bandit algorithms to advance scientific knowledge (Liu et al., 2014; Erraqabi et al., 2017). By scientific knowledge, we mean information gained by using the data to conduct a variety of statistical analyses, including confidence interval construction and hypothesis testing. **While regret minimization is a *within*-experiment learning objective, gaining scientific knowledge from the resulting adaptively collected data is a *between*-experiment learning objective**, which ultimately helps with regret minimization between deployments of bandit algorithms. Note that the data collected by bandit algorithms are *adaptively collected* because previously observed contexts, actions, and rewards are used to inform what actions to select in future timesteps.

There are a variety of between-experiment learning questions encountered in real-life applications of bandit algorithms. For example, in real-life sequential decision-making problems there are often a number of additional scientifically interesting outcomes besides the reward that are collected during the experiment. In the online advertising setting, the reward might be whether an ad is clicked on, but one may be interested in the outcome of amount of money spent or the subsequent time spent on the advertiser's website. If it was found that an ad had high click-through rate, but low amounts of money was spent after clicking on the ad, one may redesign the reward used in the next bandit experiment. One type of statistical analysis would be to construct confidence intervals for the relative effect of the actions on multiple outcomes (in addition to the reward) conditional on the context. Furthermore, due to engineering and practical limitations, some of the variables that might be useful as context are often not accessible to the bandit algorithm online. If after-study analyses find some such contextual variables to have sufficiently strong influence on the relative usefulness of an action, this might lead investigators to ensure these variables are accessible to the bandit algorithm in the next experiment.

As discussed above, we can gain scientific knowledge from data collected with (contextual) bandit algorithms by constructing confidence intervals and performing hypothesis tests for unknown quantities such as the expected outcome

[1]Department of Computer Science, Harvard University [2]Department of Statistics, Harvard University. Correspondence to: Kelly W. Zhang <kellywzhang@seas.harvard.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

for different actions in various contexts. Unfortunately, standard statistical methods developed for i.i.d. data fail to provide valid inference when applied to data collected with common bandit algorithms. For example, assuming the sample mean of rewards for an arm is approximately normal can lead to unreliable confidence intervals and inflated type-1 error; see Figure 2 for an illustration. Recently statistical inference methods have been developed for data collected using bandit algorithms (Hadad et al., 2019; Deshpande et al., 2018; Zhang et al., 2020); however, these methods are limited to inference for parameters of simple models. There is a lack of general statistical inference methods for data collected with (contextual) bandit algorithms in more complex data-analytic settings, including parameters in nonlinear models for outcomes; for example, there are currently no methods for constructing valid confidence intervals for the parameters of a logistic regression model for binary outcomes or for constructing confidence intervals based on robust estimators like minimizers of the Huber loss function.

In this work we show that a wide variety of estimators which are frequently used both in science and industry on i.i.d. data, namely, M-estimators (Van der Vaart, 2000), can be used to conduct valid inference on data collected with (contextual) bandit algorithms when adjusted with particular adaptive weights, i.e., weights that are a function of previously collected data. Different forms of adaptive weights are used by existing methods for simple models. Our work is a step towards developing a general framework for statistical inference on data collected with adaptive algorithms, including (contextual) bandit algorithms.

2. Problem Formulation

We assume that the data we have after running a contextual bandit algorithm is comprised of contexts $\{X_t\}_{t=1}^T$, actions $\{A_t\}_{t=1}^T$, and primary outcomes $\{Y_t\}_{t=1}^T$. T is deterministic and known. We assume that rewards are a deterministic function of the primary outcomes, i.e., $R_t = f(Y_t)$ for some known function f. We are interested in constructing confidence regions for the parameters of the conditional distribution of Y_t given (X_t, A_t). Below we consider $T \to \infty$ in order to derive the asymptotic distributions of estimators and construct asymptotically valid confidence intervals. We use potential outcome notation (Imbens and Rubin, 2015) and let $\{Y_t(a) : a \in \mathcal{A}\}$ denote the potential outcomes of the primary outcome and let $Y_t := Y_t(A_t)$ be the observed outcome. We assume a stochastic contextual bandit environment in which $\{X_t, Y_t(a) : a \in \mathcal{A}\} \overset{i.i.d.}{\sim} \mathcal{P} \in \mathbf{P}$ for $t \in [1\colon T]$; the contextual bandit environment distribution $\mathcal{P}$ is in a space of possible environment distributions $\mathbf{P}$. We define the history $\mathcal{H}_t := \{X_{t'}, A_{t'}, Y_{t'}\}_{t'=1}^t$ for $t \geq 1$ and $\mathcal{H}_0 := \emptyset$. Actions $A_t \in \mathcal{A}$ are selected according to policies $\pi := \{\pi_t\}_{t\geq 1}$, which define action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) := \mathbb{P}(A_t|\mathcal{H}_{t-1}, X_t)$. Even though the potential outcomes are i.i.d., the *observed* data $\{X_t, A_t, Y_t\}_{t=1}^T$ are *not* because the actions are selected using policies π_t which are a function of past data, $\mathcal{H}_{t-1}$.

We are interested in constructing confidence regions for some unknown $\theta^*(\mathcal{P}) \in \Theta \subset \mathbb{R}^d$, which is a parameter of the conditional distribution of Y_t given (X_t, A_t). Specifically, we assume that $\theta^*(\mathcal{P})$ is a conditionally maximizing value of criterion m_θ, i.e., for all $\mathcal{P} \in \mathbf{P}$,

$$\theta^*(\mathcal{P}) \in \operatorname*{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P}}[m_\theta(Y_t, X_t, A_t)|X_t, A_t] \text{ w.p. 1.} \quad (1)$$

Note that it is an implicit modelling assumption that such a $\theta^*(\mathcal{P})$ exists for a given m_θ. To estimate $\theta^*(\mathcal{P})$, we build on M-estimation (Huber, 1992), which classically selects the estimator $\hat{\theta}$ to be the $\theta \in \Theta$ that maximizes the empirical analogue of Equation (1):

$$\hat{\theta}_T := \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^T m_\theta(Y_t, X_t, A_t). \quad (2)$$

For example, in a classical linear regression setting with $|\mathcal{A}| < \infty$, a natural choice for m_θ is the negative of the squared loss function, $m_\theta(Y_t, X_t, A_t) = -(Y_t - X_t^\top \theta_{A_t})^2$. When Y_t is binary, a natural choice is instead the negative log-likelihood function for a logistic regression model. More generally, m_θ is commonly chosen to be a log-likelihood function or the negative of a robust loss function such as the Huber loss. If the data, $\{X_t, A_t, Y_t\}_{t=1}^T$, were independent across time, classical approaches could be used to prove the consistency and asymptotic normality of M-estimators. However, on data collected with bandit algorithms, standard M-estimators like the ordinary least-squares estimator fail to provide valid confidence intervals.

3. Adaptively Weighted M-Estimators

We consider a weighted M-estimating criteria with adaptive weights $W_t \in \sigma(\mathcal{H}_{t-1}, X_t, A_t)$ given by $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$. Here $\{\pi_t^{\text{sta}}\}_{t\geq 1}$ are pre-specified *stabilizing policies* that do not depend on data $\{Y_t, X_t, A_t\}_{t\geq 1}$. A default choice for the stabilizing policy when the action space is of size $|\mathcal{A}| < \infty$ is just $\pi_t^{\text{sta}}(a, x) = 1/|\mathcal{A}|$ for all x, a, and t; we discuss considerations for the choice of $\{\pi_t^{\text{sta}}\}_{t=1}^T$ in Appendix C. We call these weights *square-root importance weights* because they are the square-root of the standard importance weights (Hammersley, 2013). We use estimators for $\theta^*(\mathcal{P})$, $\hat{\theta}_T$, is the maximizer of a *weighted* version of the M-estimation criterion of Equation (2):

$$\hat{\theta}_T := \operatorname*{argmax}_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^T W_t m_\theta(Y_t, X_t, A_t)$$

Let $M_T(\theta) := \frac{1}{T}\sum_{t=1}^T W_t m_\theta(Y_t, X_t, A_t)$. We provide asymptotically valid confidence regions for $\theta^*(\mathcal{P})$ by de-

riving the asymptotic distribution of $\hat{\theta}_T$ as $T \to \infty$ and by proving that the convergence in distribution is *uniform* over $\mathcal{P} \in \mathbf{P}$. Such convergence allows us to construct a uniformly asymptotically valid $1-\alpha$ level confidence region, $C_T(\alpha)$, for $\theta^*(\mathcal{P})$, which is a confidence region that satisfies

$$\liminf_{T\to\infty} \inf_{\mathcal{P}\in\mathbf{P}} \mathbb{P}_{\mathcal{P},\pi}\left(\theta^*(\mathcal{P}) \in C_T(\alpha)\right) \geq 1-\alpha. \quad (3)$$

Confidence regions that are asymptotically valid, but not *uniformly* asymptotically valid, fail to be reliable in practice (Leeb and Pötscher, 2005; Romano et al., 2012). To construct uniformly valid confidence regions for $\theta^*(\mathcal{P})$ we prove that $\hat{\theta}_T$ is uniformly asymptotically normal in that

$$\Sigma_T(\mathcal{P})^{-1/2}\ddot{M}_T(\hat{\theta}_T)\sqrt{T}(\hat{\theta}_T - \theta^*(\mathcal{P})) \xrightarrow{D} \mathcal{N}(0, I_d) \text{ uniformly over } \mathcal{P} \in \mathbf{P}, \quad (4)$$

where $\dot{m}_\theta := \frac{\partial}{\partial\theta} m_\theta$, $\ddot{M}_T(\theta) := \frac{\partial^2}{\partial^2\theta} M_T(\theta)$, and $\Sigma_T(\mathcal{P}) := \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{P},\pi_t^{\text{sta}}}\left[\dot{m}_{\theta^*(\mathcal{P})}(Y_t, X_t, A_t)^{\otimes 2}\right]$. For any vector z we define $z^{\otimes 2} := zz^\top$. In Appendix B, we state the full conditions we use to prove for general M-estimators that $\hat{\theta}_T$ is uniformly consistent and uniformly asymptotically normal in the sense of Equation (4).

The critical role of the square-root importance weights $W_t = \sqrt{\frac{\pi_t^{\text{sta}}(A_t, X_t)}{\pi_t(A_t, X_t, \mathcal{H}_{t-1})}}$ is to adjust for instability in the *variance* of M-estimators due to the bandit algorithm. These weights act akin to standard importance weights when squared and adjust a key term in the variance of M-estimators from depending on adaptive policies $\{\pi_t\}_{t=1}^T$, which can be ill-behaved, to depending on the pre-specified stabilizing policies $\{\pi_t^{\text{sta}}\}_{t=1}^T$. See Zhang et al. (2020) and Deshpande et al. (2018) for more discussion of the ill-behavior of common bandit algorithms, which occurs particularly when there is no unique optimal policy.

To better understand the role of the weights, we consider the least-squares estimators in a finite-arm linear contextual bandit setting as an illustrating example. Assume that $\mathbb{E}_{\mathcal{P}}[Y_t | X_t, A_t = a] = X_t^\top \theta_a^*(\mathcal{P})$ w.p. 1. We focus on estimating $\theta_a^*(\mathcal{P})$ for some $a \in \mathcal{A}$. The least-squares estimator corresponds to an M-estimator with $m_{\theta_a}(Y_t, X_t, A_t) = -\mathbb{1}_{A_t=a}(Y_t - X_t^\top\theta_a)^2$. The adaptively weighted least-squares (AW-LS) estimator is $\hat{\theta}_{T,a}^{\text{AW-LS}} := \text{argmax}_{\theta_a}\{-\sum_{t=1}^T W_t \mathbb{1}_{A_t=a}(Y_t - X_t^\top\theta_a)^2\}$. For simplicity, suppose that the stabilizing policy does not change with t and drop the index t to get π^{sta}. Taking the derivative of this criterion, we get $0 = \sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t\left(Y_t - X_t^\top \hat{\theta}_{T,a}^{\text{AW-LS}}\right)$, and rearranging terms gives

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t X_t^\top \left(\hat{\theta}_{T,a}^{\text{AW-LS}} - \theta_a^*(\mathcal{P})\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^T W_t \mathbb{1}_{A_t=a} X_t \left(Y_t - X_t^\top \theta_a^*(\mathcal{P})\right). \quad (5)$$

Note the right hand side of Equation (5) is a martingale difference sequence with respect to history $\{\mathcal{H}_t\}_{t=0}^T$ because $\mathbb{E}_{\mathcal{P},\pi}[W_t \mathbb{1}_{A_t=a}(Y_t - X_t^\top\theta_a^*(\mathcal{P}))|\mathcal{H}_{t-1}] = 0$ for all t, by our i.i.d. potential outcomes assumption and since $\mathbb{E}_{\mathcal{P}}[Y_t|X_t, A_t = a] = X_t^\top\theta_a^*(\mathcal{P})$. We prove Equation (5) is uniformly asymptotically normal by applying a martingale central limit theorem; the key condition we need to ensure is that the conditional variance converges uniformly, for which it is sufficient to show that $\mathbb{E}_{\mathcal{P},\pi}[W_t^2 \mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2|\mathcal{H}_{t-1}] = \Sigma(\mathcal{P})$ for every t. By law of iterated expectations, $\mathbb{E}_{\mathcal{P},\pi}[W_t^2 \mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2|\mathcal{H}_{t-1}]$ equals

$$\begin{aligned}
&\mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P}}[W_t^2 \mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2|\mathcal{H}_{t-1}, X_t]\middle|\mathcal{H}_{t-1}\right] \\
&\underset{(a)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}[\mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2|\mathcal{H}_{t-1}, X_t]\middle|\mathcal{H}_{t-1}\right] \\
&\underset{(b)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[\mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2\middle|X_t\right]\middle|\mathcal{H}_{t-1}\right] \\
&\underset{(c)}{=} \mathbb{E}_{\mathcal{P}}\left[\mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}\left[\mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2\middle|X_t\right]\right] \\
&\underset{(d)}{=} \mathbb{E}_{\mathcal{P},\pi^{\text{sta}}}[\mathbb{1}_{A_t=a} X_t X_t^\top (Y_t - X_t^\top\theta_a^*(\mathcal{P}))^2] =: \Sigma(\mathcal{P}).
\end{aligned}$$

Above, (a) holds because the importance weights change the sampling measure from the adaptive policy π_t to the pre-specified stabilizing policy π^{sta}. (b) holds by our i.i.d. potential outcomes assumption and because π^{sta} is a pre-specified policy. (c) holds because X_t does not depend on $\mathcal{H}_{t-1}$ by our i.i.d. potential outcomes assumption. (d) holds by the law of iterated expectations. Note that $\Sigma(\mathcal{P})$ does not depend on t because π^{sta} is not time-varying. In contrast, without the adaptive weighting, i.e., when $W_t = 1$, the conditional covariance of $\mathbb{1}_{A_t=a}\left(Y_t - X_t^\top\theta_a^*(\mathcal{P})\right)$ on $\mathcal{H}_{t-1}$ is a random variable, due to the adaptive policy π_t.

In Figure 2 we plot the empirical distributions of the z-statistic for the least-squares estimator both with and without adaptive weighting. Note the unweighted version gives the ordinary least-squares (OLS) estimator. It is clear that the least-squares estimator with adaptive weighting has a z-statistic that is much closer to a normal distribution.

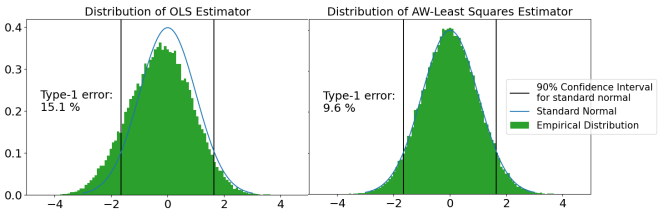

Figure 2. The empirical distributions of the weighted and unweighted least-squares estimators for $\theta_1^*(\mathcal{P}) := \mathbb{E}_{\mathcal{P}}[Y_t(1)]$ in a two arm bandit setting where $\mathbb{E}_{\mathcal{P}}[Y_t(1)] = \mathbb{E}_{\mathcal{P}}[Y_t(0)] = 0$. We perform Thompson Sampling with $\mathcal{N}(0,1)$ priors, $\mathcal{N}(0,1)$ errors, and $T = 1000$. We plot $\sqrt{\sum_{t=1}^T A_t}(\hat{\theta}_{T,1}^{\text{OLS}} - \theta_1^*(\mathcal{P}))$ on the left and $\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \sqrt{\frac{0.5}{\pi_t(1)}} A_t\right)(\hat{\theta}_{T,1}^{\text{AW-LS}} - \theta_1^*(\mathcal{P}))$ on the right.

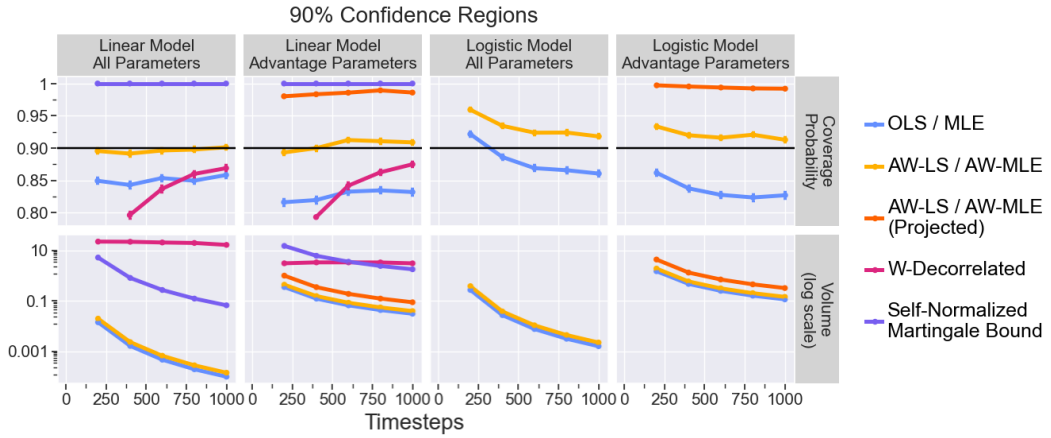

Figure 1. Empirical coverage probabilities (upper row) and volume (lower row) of 90% confidence ellipsoids. The left two columns are for the continuous reward setting and the right two columns are for the binary reward setting. We consider confidence ellipsoids for all parameters $\theta^*(\mathcal{P})$ and for advantage parameters $\theta_1^*(\mathcal{P})$ for both settings. Error bars are standard errors computed over 5k repetitions.

4. Related Work

Recent work has discussed the non-normality of OLS on data collected with bandit algorithms and proposed alternative methods for statistical inference. The common thread through these methods is the use of *adaptive weighting*. Deshpande et al. (2018) introduced the W-decorrelated estimator, which adjusts the OLS estimator with a sum of adaptively weighted residuals. Hadad et al. (2019) introduce adaptively weighted versions of both the standard augmented-inverse propensity weighted estimator and the sample mean. They introduce a class of adaptive "variance stabilizing" weights, for which the variance of a normalized version of their estimators converges in probability to a constant. In their discussion section they note open questions, two of which this work addresses: 1) "What additional estimators can be used for normal inference with adaptively collected data?" and 2) How do their results generalize to more complex sampling designs, like data collected with contextual bandit algorithms? We demonstrate that variance stabilizing adaptive weights can be used to modify a large class of M-estimators to guarantee valid inference. This generalization allows us to perform valid inference for a large class of important inferential targets: parameters of models for expected outcomes that are context dependent.

An alternative to using asymptotic approximations to construct confidence intervals is to use high-probability anytime confidence bounds. These bounds provide stronger guarantees than those based on asymptotic approximations, as they are guaranteed to hold for finite samples and hold simultaneously for $T \geq 1$. However, these bounds are typically much wider, which is why much of classical statistics uses asymptotic approximations. We empirically compare to the self-normalized martingale bound (Abbasi-Yadkori et al., 2011), a bound commonly used in the bandit literature.

5. Simulation Results

In this section, $R_t = Y_t$. We consider two settings: a continuous reward setting and a binary reward setting. In the continuous reward setting, the rewards are generated with mean $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = \tilde{X}_t^\top \theta_0^*(\mathcal{P}) + A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P})$ and noise drawn from a student's t distribution with five degrees of freedom; here $\tilde{X}_t = [1, X_t] \in \mathbb{R}^3$ (X_t with intercept term), actions $A_t \in \{0, 1\}$, and parameters $\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P}) \in \mathbb{R}^3$. In the binary reward setting, the reward R_t is generated as a Bernoulli with success probability $\mathbb{E}_{\mathcal{P}}[R_t|X_t, A_t] = [1 + \exp(-\tilde{X}_t^\top \theta_0^*(\mathcal{P}) - A_t \tilde{X}_t^\top \theta_1^*(\mathcal{P}))]^{-1}$. Furthermore, in both simulation settings we set $\theta_0^*(\mathcal{P}) = [0.1, 0.1, 0.1]$ and $\theta_1^*(\mathcal{P}) = [0, 0, 0]$, so there is no unique optimal arm; we call vector parameter $\theta_1^*(\mathcal{P})$ the *advantage* of selecting $A_t = 1$ over $A_t = 0$. Also in both settings, the contexts X_t are drawn i.i.d. from a uniform distribution.

In both simulation settings we collect data using Thompson Sampling with a linear model for the expected reward and normal priors (Agrawal and Goyal, 2013) (so even when the reward is binary). We constrain the action selection probabilities with *clipping* at a rate of 0.05; this means that while typical Thompson Sampling produces action selection probabilities $\pi_t^{\text{TS}}(A_t, X_t, \mathcal{H}_{t-1})$, we instead use action selection probabilities $\pi_t(A_t, X_t, \mathcal{H}_{t-1}) = 0.05 \vee \left(0.95 \wedge \pi_t^{\text{TS}}(A_t, X_t, \mathcal{H}_{t-1})\right)$ to select actions. We constrain the action selection probabilities in order to ensure weights W_t are bounded when using a uniform stabilizing policy; see Appendix B.1 and Section 6 for more discussion on this boundedness assumption. Also note that increasing the amount the algorithm explores (clipping) decreases the expected width of confidence intervals constructed on the resulting data (see Section 6).

To analyze the data, in the continuous reward setting, we use least-squares estimators with a correctly specified

model for the expected reward, i.e., M-estimators with $m_\theta(R_t, X_t, A_t) = -(R_t - \tilde{X}_t^\top \theta_0 - A_t \tilde{X}_t^\top \theta_1)^2$. We consider both the unweighted and adaptively weighted versions. We also compare to the self-normalized martingale bound (Abbasi-Yadkori et al., 2011) and the W-decorrelated estimator (Deshpande et al., 2018), as they were both developed for the linear expected reward setting. For the self-normalized martingale bound, which requires explicit bounds on the parameter space, we set $\Theta = \{\theta \in \mathbb{R}^6 : \|\theta\|_2 \leq 6\}$. In the binary reward setting, we also assume a correctly specified model for the expected reward. We use both unweighted and adaptively weighted maximum likelihood estimators (MLEs), which correspond to an M-estimators with $m_\theta(R_t, X_t, A_t)$ set to the negative log-likelihood of R_t given X_t, A_t. We solve for these estimators using Newton–Raphson optimization and do not put explicit bounds on the parameter space Θ (note in this case m_θ is concave in θ (Agresti, 2015, Chapter 5.4.2)). See Appendix A for additional details and simulation results.

In Figure 1 we plot the empirical coverage probabilities and volumes of 90% confidence regions for $\theta^*(\mathcal{P}) := [\theta_0^*(\mathcal{P}), \theta_1^*(\mathcal{P})]$ and $\theta_1^*(\mathcal{P})$ in both the continuous and binary reward settings. While the confidence regions based on the unweighted least-squares estimator (OLS) and the unweighted MLE have significant undercoverage that does not improve as T increases, the confidence regions based on the adaptively weighted versions, AW-LS and AW-MLE, have very reliable coverage. For the confidence regions for $\theta_1^*(\mathcal{P})$ based on the AW-LS and AW-MLE, we include both projected confidence regions (for which we have theoretical guarantees) and non-projected confidence regions. The confidence regions based on projections (see Appendix B.1) are conservative but nevertheless have comparable volume to those based on OLS and MLE respectively. We do not prove theoretical guarantees for the non-projection confidence regions for AW-LS and AW-MLE, however they perform well across in our simulations. Both types of confidence regions based on AW-LS have significantly smaller volumes than those constructed using the self-normalized martingale bound and W-decorrelated estimator. Note that the W-decorrelated estimator and self-normalized martingale bounds are designed for linear contextual bandits and are thus not applicable for the logistic regression model setting. The confidence regions constructed using the self-normalized martingale bound have reliable coverage as well, but are very conservative. Empirically, we found that the coverage probabilities of the confidence regions based on the W-decorrelated estimator were very sensitive to the choice of tuning parameters. We use $5,000$ Monte-Carlo repetitions and the error bars plotted are standard errors.

6. Discussion

Immediate questions We assume that ratios $\pi_t^{\text{sta}}(A_t, X_t)/\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ are bounded for our theoretical results; this precludes $\pi_t(A_t, X_t, \mathcal{H}_{t-1})$ from going to zero for a fixed stabilizing policy. For simple models, e.g., the AW-LS estimator, we can let these ratios grow at a certain rate and still guarantee asymptotic normality (Appendix B.6); we conjecture similar results hold more generally.

Generality and robustness This work assumes $\theta^*(\mathcal{P}) \in \text{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P}}[m_\theta(Y_t, X_t, A_t)|X_t, A_t]$ w.p. 1. Our theorems use this assumption to ensure that $\{W_t \dot{m}_\theta(Y_t, X_t, A_t)\}_{t\geq 1}$ is a martingale difference sequence with respect to $\{\mathcal{H}_t\}_{t\geq 0}$. On i.i.d. data it is common to define $\theta^*(\mathcal{P})$ to be the best *projected* solution, i.e., $\theta_0(\mathcal{P}) \in \text{argmax}_{\theta\in\Theta} \mathbb{E}_{\mathcal{P},\pi}[m_\theta(Y_t, X_t, A_t)]$. Note that the best projected solution, $\theta^*(\mathcal{P})$, depends on the distribution of the action selection policy π. It would be ideal to also be able to perform inference for a projected solution on adaptively collected data.

Trading-off regret minimization and statistical inference objectives In sequential decision-making problems there is a fundamental trade-off between minimizing regret and minimizing estimation error for parameters of the environment using the resulting data (Bubeck et al., 2009; Dean et al., 2018). Given this trade-off there are many open problems regarding how to minimize regret while still guaranteeing a certain amount of power or expected confidence interval width, e.g., developing sample size calculators for use in justifying the number of users in a mobile health trial, and developing new adaptive algorithms (Liu et al., 2014; Erraqabi et al., 2017; Yao et al., 2020).

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. Regret bounds for robust adaptive control

of the linear quadratic regulator. In *Advances in Neural Information Processing Systems*, 2018.

Yash Deshpande, Lester Mackey, Vasilis Syrgkanis, and Matt Taddy. Accurate inference for adaptive linear models. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1194–1203, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Akram Erraqabi, Alessandro Lazaric, Michal Valko, Emma Brunskill, and Yun-En Liu. Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics*, pages 709–717. PMLR, 2017.

Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*, 2019.

John Hammersley. *Monte carlo methods*. Springer Science & Business Media, 2013.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

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

Maximilian Kasy and Anja Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, pages 21–59, 2005.

Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

Yun-En Liu, Travis Mandel, Emma Brunskill, and Zoran Popovic. Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM*, pages 161–168, 2014.

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

Hammad Shaikh, Arghavan Modiri, Joseph Jay Williams, and Anna N Rafferty. Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*, 2019.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Jiayu Yao, Emma Brunskill, Weiwei Pan, Susan Murphy, and Finale Doshi-Velez. Power-constrained bandits. *arXiv preprint arXiv:2004.06230*, 2020.

Elad Yom-Tov, Guy Feraru, Mark Kozdoba, Shie Mannor, Moshe Tennenholtz, and Irit Hochberg. Encouraging physical activity in patients with diabetes: intervention using a reinforcement learning system. *Journal of medical Internet research*, 19(10):e338, 2017.

Kelly W Zhang, Lucas Janson, and Susan A Murphy. Inference for batched bandits. In *Advances in Neural Information Processing Systems*, 2020.