Power-Constrained Bandits

Jiayu Yao
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
jiy328.harvard.edu

Emma Brunskill
Computer Science Department
Stanford University
Stanford, CA 94305 ebrun@cs.stanford.edu

Weiwei Pan
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
weiweipan@g.harvard.edu

Susan Murphy
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
samurphy11@gmail.com

Finale Doshi-Velez
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
finale@seas.harvard.edu

Abstract

Contextual bandits often provide simple and effective personalization in decision making problems, making them popular in many domains including digital health. However, when bandits are deployed in the context of a scientific study, the aim is not only to personalize for an individual, but also to determine, with sufficient statistical power, whether or not the system's intervention is effective. In this work, we develop general meta-algorithms to modify existing algorithms such that sufficient power is guaranteed, without significant decrease in average return.

1 Introduction

Contextual bandits provide an attractive middle ground between multi-arm bandits and full Markov Decision Processes. Their simplicity, robustness, and effectiveness had made them popular in domains ranging from online education to ads recommendation. In this work, we are motivated by situations in which we not only want to personalize actions to the user, but we also want to rigorously detect the effect of the treatment. Such situations are common when the contextual bandit is used as part of a study: imagine a mobile app that will help patients manage their mental illness by delivering reminders to self-monitor their mental state. In this case, not only may we want to personalize reminders, but we also want to measure the marginal effect of reminders on self-monitoring. Quantifying these effects is often essential for downstream science and development.

Currently, there exist algorithms that either have principled bounds on regret (e.g. Abbasi-Yadkori et al. [2011], Agrawal and Goyal [2012], Krishnamurthy et al. [2018]), largely coming from the RL community, or aim to rigorously determine an effect (e.g. micro-randomized trials Liao et al. [2016], Klasnja et al. [2015], Kramer et al. [2019]), which have been a focus in the experimental design community. Practical implementation of these algorithms often results in tensions: for regret minimization, one may make assumptions that are likely not true, but close enough to result in fast personalization. However, for treatment effect analysis, one must be able to make strong statistical

Preprint. Under review.

claims in the face of a potentially non-stationary user—e.g. one who is initially excited by the novelty of a new app, and then disengages—as well as highly stochastic, hidden aspects of the environment—e.g. if the user has a deadline looming, or started watching a new television series. It is not obvious whether an algorithm that does a decent job of personalization under one set of assumptions would guarantee desired power under more general assumptions.

In this work, we *both* rigorously guarantee that a trial will be sufficiently powered to provide inference about treatment effects (produce generalizable knowledge about a population of users) *and* minimize regret (improve each user's well-being). In minimizing regret, each user represents a different task; in assessing treatment effects, the sample of users is the task. We specifically focus on settings where trials are expensive, and thus not only must one be sufficiently powered, one must also leave open the option for post-hoc analyses via off-policy evaluation techniques; this requirement will imply that all action probabilities must be bounded away from 0 or 1. For this context, we adjust a wide variety of popular regret-minimization algorithms such that sufficient power is guaranteed *and* we get optimal regret for each user with respect to an oracle that selects from a class of power-preserving policies. We provide formal analyses and supporting experiments for specific algorithms as well as general approaches for adapting existing contextual bandit algorithms to meet these goals.

2 Related Work

A variety of works focus on ways to quantify the properties of various arms of a bandit. These include works for estimating the means of all arms (e.g. Carpentier et al. [2011]) and approaches focused on best-arm identification aim to find the best treatment with confidence [Audibert and Bubeck, 2010]. Best-arm identification has been applied to both stochastic as well as adversarial settings [Abbasi-Yadkori et al., 2018, Lattimore and Szepesvari, 2019]. However, these algorithms typically personalize little if at all, and thus can result in high regret.

Other works focus on minimizing regret without considering testing hypotheses related to treatment effectiveness. While there exists a long history of optimizing bandits in RL (e.g. Abbasi-Yadkori et al. [2011], Agrawal and Goyal [2012]), perhaps most relevant are more recent works that can achieve optimal first order regret rates in highly stochastic, even adversarial settings [Lattimore and Szepesvari, 2019, Krishnamurthy et al., 2018, Greenewald et al., 2017]. Our approach also provides power guarantees in those challenging settings without significance increase in regret.

Finally, other works consider other simultaneous objectives. Degenne et al. [2019], Erraqabi et al. [2017] consider arm value estimation jointly with regret minimization. Nie et al. [2018], Deshpande et al. [2018], Hadad et al. [2019] consider how to accurately estimate the means or provide confidence intervals with data collected via adaptive sampling algorithms. At a high level, most similar to this work is that of Williamson et al. [2017], Villar et al. [2015] who assume stationary Bernoulli rewards. They consider the task of assigning treatments to N individuals so as to minimize regret (i.e., maximize success rate). They consider heuristic alternatives to improve power but not guarantee it. Our approach considers more general settings and provides theoretical guarantee for a stated power.

To our knowledge, ours is the first to consider how to accomplish two tasks: a sequential decision problem one per user with the goal to minimize regret during the study and to guarantee the power to detect a marginal (across the users) effect after the study is over. We guarantee the latter in a *non-stationary* and *stochastic* setting.

3 Notation, Model, and Statistical Setting

We consider a collection of histories $\{H_{nT}\}_{n=1}^{N}$ consisting of N users, each with T steps, where $H_{nt} = (C_{n0}, A_{n0}, R_{n0}, C_{n1}, A_{n1}, R_{n1} \ldots, C_{nt})$, $t \leq T$; C_{nt} denotes the context of user n at time step t, $A_{nt} \in \{0, 1\}$ denotes the binary action, and R_{nt} denotes the reward. The potential rewards are $(R_{nt}(0), R_{nt}(1))$. R_{nt} is a composite of the potential rewards and the action, A_{nt}: $R_{nt} = R_{nt}(A_{nt})$. For each user, a contextual bandit algorithm uses a policy π_t which is a function constructed from the user's prior data $H_{n,t-1}, A_{n,t-1}, R_{n,t-1}$, in order to select action A_{nt} based on the current context. (i.e. $p(A_{nt} = 1) = \pi_t(C_{nt})$). We write as π_{nt} for short in the following text.

In practice, it is common to make certain assumptions for efficient exploration and good performance when minimizing regret, but still desire to preserve sufficient power for later analysis even if those

assumptions are violated. In this section, we describe a very general setting for treatment effect analysis such that the environment can be stochastic, non-stationary, and future contexts can depend on past ones. In Section 5.2, we will consider a variety of additional assumptions that might be made by the regret minimization algorithm. For example, Action-Centered Thompson Sampling [Greenewald et al., 2017] and Semi-Parametric Contextual Bandit [Krishnamurthy et al., 2018] assume that the treatment effect only depends on the current context C_{nt} while our setting for power guarantees allows it to be a function of full history H_{nt}. We also allow reward noise to be correlated across time.

Finally, we will require policies to have action probabilities in some $[\pi_{\texttt{min}}, \pi_{\texttt{max}}]$ bounded away from 0 and 1. This policy class is preferred—and often required—by scientists who wish to preserve their ability to perform unspecified secondary analyses (e.g. Thomas and Brunskill [2016], Su et al. [2019]) and causal inference analyses (e.g. Boruvka et al. [2018]). We also run the algorithm for each user separately, as correctly accounting for treatment effect when combining data over users is nontrivial since users may enter the study at different times. Furthermore, some works have found that for online detection and prediction, user-specific algorithms work better than population-based algorithms [Dallery et al., 2013, Korinek et al., 2018, Albers et al., 2017].

Environment and Notation for Statistical Analyses We consider a semiparametric linear contextual bandit setting where the reward can be composed into an action-independent term and an action-dependent linear term. We assume the treatment effect satisfies

$$\mathbb{E}[R_{nt}(1)|H_{nt}] - \mathbb{E}[R_{nt}(0)|H_{nt}] = Z_t^\intercal(H_{nt})\delta_0, \tag{1}$$

where $Z_t(H_{nt})$ is a set of features that are a known function of history H_{nt}. Importantly, $Z_t(H_{nt})$ is independent of present action A_{nt} but may depend on prior actions. We assume that an expert defines what features of a history may be important for the reward but make *no* assumptions about how the history itself evolves. We assume the histories $\{H_{nt}\}_{n=1}^N$ are independent and identically distributed as we run algorithms on each user separately. However, there may be dependencies across time within a specific subject. Finally, we assume that $\text{Var}[R_{nt}(a)|H_{nt}] < \infty$ for $a \in \{0, 1\}$ and $t = 1, \ldots, T$. Denote the marginal reward averaged over the action as γ_{nt}, which can be a complex non-linear function. Thus, the potential reward can be written as

$$R_{nt}(a) = aZ_t^\intercal(H_{nt})\delta_0 + \gamma_t(H_{nt}) + \epsilon_{nt}.$$

where ϵ_{nt} is a noise term whose assumptions will be further specified later. In the following, we write $Z_t(H_{nt})$ as Z_{nt} and $\gamma_t(H_{nt})$ as γ_{nt} for short.

Hypothesis and Test Statistic We construct a test statistic that requires minimal assumptions to guarantee the desired Type 1 error rate and the desired power. A natural primary hypothesis concerns the treatment effect, here encoded by the value of δ_0 in Equation 1. Our goal is to test the null hypothesis: $H_0 : \delta_0 = 0$ and the alternate hypothesis: $H_1 : \delta_0 = \delta$. To test those hypotheses, we will construct a test statistic based on one used in multiple micro-randomized trials [Liao et al., 2016, Boruvka et al., 2018, Klasnja et al., 2019, Bidargaddi et al., 2018]. We first assume the model in Equation 1. Suppose the marginal reward can be written as:

$$\mathbb{E}[R_{nt}|H_{nt}] = B_{nt}^\intercal\gamma_0, \tag{2}$$

where B_{nt} is a vector of features constructed from H_{nt}. Note that the choice of B_{nt} can have an effect on the robustness of power guarantee (See Appendix Section A.4). Our estimated parameter $\hat{\delta}$ minimizes

$$L(\gamma, \delta) = \sum_{n=1}^{N}\sum_{t=1}^{T} \frac{(R_{nt} - B_{nt}^\intercal\gamma - (A_{nt} - \pi_{nt})Z_{nt}^\intercal\delta)^2}{\pi_{nt}(1 - \pi_{nt})}$$

The above loss function centers the action by $A_{nt} - \pi_{nt}$, which allows one to prove the asymptotic property of $\sqrt{N}(\hat{\delta} - \delta_0)$ even if the working model in Equation 2 is false [Boruvka et al., 2018].

Next, let $\theta = \begin{bmatrix}\gamma \\ \delta\end{bmatrix}$ and $X_{nt} = \begin{bmatrix}B_{nt} \\ (A_{nt} - \pi_{nt})Z_{nt}\end{bmatrix} \in \mathcal{R}^{(p+q)\times 1}$. The solution for $\hat{\theta}$ is given by

$$\hat{\theta} = \left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{X_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{R_{nt}X_{nt}}{\pi_{nt}(1-\pi_{nt})}\right) \tag{3}$$

and $\sqrt{N}(\hat{\theta} - \theta)$ is asymptotically normal with covariance defined by

$$\Sigma_\theta = \mathbb{E}\Big[\sum_{t=1}^{T}\frac{X_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]^{-1}\mathbb{E}\Big[\left(\sum_{t=1}^{T}\frac{\tilde{\epsilon}_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right)\left(\sum_{t=1}^{T}\frac{\tilde{\epsilon}_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right)\Big]\mathbb{E}\Big[\sum_{t=1}^{T}\frac{X_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]^{-1} \tag{4}$$

where p, q are the dimensions of B_{nt}, Z_{nt} respectively, and $\tilde{\epsilon}_{nt} = R_{nt} - X_{nt}^\intercal \hat{\theta}$.

Theorem 1. *Under assumptions in this section, and the assumption that matrices* $\mathbb{E}[\sum_{t=1}^T Z_{nt} Z_{nt}^\intercal]$, $\mathbb{E}\left[\sum_{t=1}^T \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right]$ *are invertible, the distribution of* $\sqrt{N}(\hat{\delta} - \delta_0)$ *converges, as* N *increases, to a normal distribution with* 0 *mean and covariance* $\Sigma_\delta = QW^{-1}Q$*, where* $Q = \mathbb{E}\left[\sum_{t=1}^T Z_{nt} Z_{nt}^\intercal\right]^{-1}$,

$$W = \mathbb{E}\Bigg[\sum_{t=1}^T \frac{(R_{nt} - X_{nt}^\intercal \theta^*)(A_{nt} - \pi_{nt}) Z_{nt}}{\pi_{nt}(1-\pi_{nt})} \sum_{t=1}^T \frac{(R_{nt} - X_{nt}^\intercal \theta^*)(A_{nt} - \pi_{nt}) Z_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Bigg].$$

for $\theta^* = \begin{bmatrix}\gamma^* \\ \delta_0\end{bmatrix}$*, and* $\gamma^* = \mathbb{E}\left[\sum_{t=1}^T \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right]^{-1} \mathbb{E}\left[\sum_{t=1}^T \frac{B_{nt} R_{nt}}{\pi_{nt}(1-\pi_{nt})}\right]$.

Proof. The proof is a minor adaptation of Boruvka et al. [2018]. See Appendix Section A.1. □

Recall that $\hat{\theta} = \begin{bmatrix}\hat{\gamma} \\ \hat{\delta}\end{bmatrix}$ and $\Sigma_\theta = \begin{bmatrix}\Sigma_\gamma & \Sigma_{\gamma\delta} \\ \Sigma_{\delta\gamma} & \Sigma_\delta\end{bmatrix}$. One can obtain $\hat{\delta}$ from $\hat{\theta}$, and Σ_δ from Σ_θ. To test the null hypothesis, one can use statistic $N\hat{\delta}{\Sigma_\delta}^{-1}\hat{\delta}$ which asymptotically follows a χ_p^2 where p is the number of parameters in δ_0. Under the alternate hypothesis $\delta_0 = \delta$, $N\hat{\delta}{\Sigma_\delta}^{-1}\hat{\delta}$ has an asymptotic non-central χ_p^2 distribution with degrees of freedom p and non-centrality parameter $c_N = N\delta{\Sigma_\delta}^{-1}\delta$.

4 Power Constrained Bandits

The asymptotic distribution for the estimator in Equation 3 depends on the policy π_{nt}. Intuitively, given N subjects and T times, we can imagine some minimum and maximum randomization probabilities $\pi_{\min}$ and $\pi_{\max}$ such that the experiment is sufficiently powered for the test above—that is, if we don't sufficiently explore, we won't be able to determine the treatment effect.

We first prove this intuition is true: for a fixed randomization probability $\pi_{nt} = \pi$, there exists a $\pi_{\min}$ and $\pi_{\max}$ ($\pi_{\min} \leq \pi_{\max}$) such that when π is $\pi_{\min}$ or $\pi_{\max}$, the experiment is sufficiently powered.

Theorem 2. *Assume that* ϵ_{nt} *is independent of* A_{nt} *conditional on* H_{nt} *(i.e.* $\mathbb{E}[\epsilon_{nt}|A_{nt}, H_{nt}] = 0$*) and* $Var(\epsilon_{nt}|H_{nt}) = \sigma^2$*. Let* α_0 *be the desired Type 1 error and* $1 - \beta_0$ *be the desired power. Set*

$$\pi_{min} = \frac{1 - \sqrt{1 - 4\triangle}}{2}, \quad \pi_{max} = \frac{1 + \sqrt{1 - 4\triangle}}{2}, \quad \triangle = \frac{\sigma^2 c_{\beta_0}}{N\delta_0^\intercal \mathbb{E}\left[\sum_{t=1}^T Z_{nt} Z_{nt}^\intercal\right]\delta_0}.$$

We choose c_{β_0} *such that* $1 - \Phi_{p;c_{\beta_0}}(\Phi_p^{-1}(1-\alpha_0)) = \beta_0$*, where* $\Phi_{p;c_{\beta_0}}$ *denotes the cdf of a non-central* χ^2 *distribution with d.f.* p *and non-central parameter* c_{β_0}*, and* Φ_p^{-1} *denotes the inverse cdf of a* χ^2 *distribution with d.f.* p*. For a given trial with* N *subjects each over* T *time units, if the randomization probability is fixed as* $\pi_{nt} = \pi_{min}$ *or* π_{max}*, the resulting power converges to* $1 - \beta_0$ *as* $N \to \infty$*.*

Proof. (Sketch) The rejection region for $H_0 : \delta_0 = 0$ is $\{N\hat{\delta}{\Sigma_\delta}^{-1}\hat{\delta} > \Phi_p^{-1}(1-\alpha_0)\}$, which results in the power of

$$1 - \beta_0 = 1 - \Phi_{p;c_N}(\Phi_p^{-1}(1-\alpha_0)) \tag{5}$$

where $c_N = N\delta_0^\intercal \Sigma_\delta^{-1}\delta_0$. The formula for Σ_δ is in Theorem 1, thus we only need to solve for $\pi_{\min}, \pi_{\max}$ when we substitute the expression for Σ_δ in c_N. Full analysis in Appendix A.2. □

In some cases, such as in the work of Liao et al. [2016], Z_{nt} may be available in advance of the study. In other cases, the study designer will need to specify a space of plausible models and determining the power for some fixed π will require finding the worst-case $\mathbb{E}[\sum_t Z_{nt} Z_{nt}^\intercal]$.

Next, we prove that as long as each randomization probability $\pi_{nt} \in [\pi_{\min}, \pi_{\max}]$, the power constraint will be met. Our proof holds for *any* selection strategy for π_{nt}, *including* ones where the policy is adversarially chosen to minimize power based on the subject's history H_{nt}. Having the condition across myraid ways of choosing π_{nt} is essential to guaranteeing power for any contextual bandit algorithm that can be made to produce clipped probabilities.

Theorem 3. *Given π_{min}, π_{max} we solved for above, if for all n and all t we have that $\pi_{nt} \in [\pi_{min}, \pi_{max}]$, then the resulting power will converge to a value no smaller than $1 - \beta_0$ as $N \to \infty$.*

Proof. (Sketch) The right hand side of Equation 5 is monotonically increasing with respect to c_N. The resulting power will be no smaller than $1 - \beta_0$ as long as $c_N \geq c_{\beta_0}$. This holds when $\pi_{nt} \in [\pi_{\min}, \pi_{\max}]$. Full proof in Appendix A.3. □

5 Regret with Power-Constrained Bandits

In Section 4, we provide an algorithm-agnostic way to guarantee a study's power constraints were met, under very general assumptions in Section 3. In practice, one often uses algorithms that make specific environment assumptions. Now, we consider how well we can do with respect to regret *under the bandit algorithm's environment assumptions*. Our goal is to preserve regret rates, now with respect to a clipped oracle whose action probabilities π_{nt} lie within $\pi_{\min}$ and $\pi_{\max}$. We study specific algorithms in which we can preserve regret rates with respect to a clipped oracle by simply clipping the action selection probability to lie in $[\pi_{\min}, \pi_{\max}]$. We also present general wrapper algorithms that allow us to adapt a large class of existing algorithms while preserving regret rates.

5.1 Regret Rates with Probability Clipping

Here we describe cases where simply clipping action probabilities preserves regret rates, now with respect to a clipped oracle.

Action-Centered Thompson Sampling (ACTS). ACTS [Greenewald et al., 2017] already has optimal first order regret with respect to a clipped oracle in non-stationary, adversarial settings where the features and reward are a function of current context C_{nt} (rather than H_{nt}). They do not consider power; using our probabilities will result in optimal regret and satisfy required power guarantees.

Semi-Parametric Contextual Bandits (BOSE). BOSE [Krishnamurthy et al., 2018] has optimal first order regret with respect to a standard oracle in a non-stationary, adversarial setting. Like ACTS, features and rewards are functions of C_{nt}. They further assume noise term is action independent. In the two action case, BOSE will select actions with probability 0.5 or with probability 0 or 1. With probability clipping, the regret bound remains unaffected and the details are provided in Section 3.3 of [Krishnamurthy et al., 2018].

A More Subtle Case: Linear Stochastic Bandits (OFUL). Finally, consider the OFUL algorithm of Abbasi-Yadkori et al. [2011] which considers a linear assumption on the entire mean reward that $\mathbb{E}[R_{nt}|A_{nt} = a] = x_{t,a}^T \theta$ for features $(x_{t,0}, x_{t,1})$. To adapt OFUL to accommodate the clipped constraint, we make a slight modification to OFUL to ensure optimism under the constraint. Specifically we replace the criterion, $x_{t,a}^\intercal \theta$ by $\ell_t(a, \theta) = \mathbb{E}[x_{t,A_t^c}^\intercal \theta | A_t = a]$ where $A_t^c \sim \text{Bernoulli}(\pi_{\max}^a \pi_{\min}^{1-a})$. The construction of the confidence set remains the same. In Appendix A.5, we prove clipping preserves regret with respect to a clipped oracle.

5.2 General Power-Preserving Wrapper Algorithms

The above cases required a case-by-case analysis to determine whether clipping probabilities would preserve regret rates (now with respect to a clipped oracle). Now we describe how to adapt a wide variety of bandit algorithms in a way that (a) guarantees sufficient power and (b) preserves regret rates with respect to a clipped oracle.

Meta-Algorithm: Action-Flipping Denote the action probability given by a bandit algorithm $\mathcal{A}$ as $\pi_{\mathcal{A}}(C_{nt})$. Suppose we take the action outputted by any algorithm and flip it with some probability:

1. Given current context C_{nt}, algorithm $\mathcal{A}$ produces action probabilities $\pi_{\mathcal{A}}(C_{nt})$
2. Sample $A_{nt} \sim \texttt{Bern}(\pi_{\mathcal{A}}(C_{nt}))$.
3. If $A_{nt} = 1$, sample $A'_{nt} \sim \texttt{Bern}(\pi_{\max})$. If $A_{nt} = 0$, sample $A'_{nt} \sim \texttt{Bern}(\pi_{\min})$.
4. We perform A'_{nt} and receive reward R_{nt}.

5. The algorithm $\mathcal{A}$ stores the tuple C_{nt}, A_{nt}, R_{nt}. (Note that if A_{nt} and A'_{nt} are different, then, unbeknownst to the algorithm $\mathcal{A}$, a different action was actually performed.)
6. The scientist stores the tuple C_{nt}, A'_{nt}, R_{nt} for their analysis.

Let $A'_{nt} = G_(A_{nt})$ denote the stochastic transformation by which the wrapper above transforms the action A_{nt} from algorithm $\mathcal{A}$ to the new action A'_{nt}. Suppose that the input algorithm $\mathcal{A}$ had some regret rate $\mathcal{R}(T)$ for a set of environments Ω (e.g. assumptions on distribution of $\{C_{nt}, R_{nt}(0), R_{nt}(1)\}_{t=1}^{T}$). We prove under what conditions the altered version of algorithm $\mathcal{A}$, as described above, will achieve the same rate against a clipped oracle:

Theorem 4. *Given $\pi_{\min}$, $\pi_{\max}$ and a contextual bandit algorithm $\mathcal{A}$, assume that algorithm $\mathcal{A}$ has expected regret $\mathcal{R}(T)$ for any environment in Ω, with respect to an oracle $\mathcal{O}$. If there exists an environment in Ω such that the potential rewards, $R'_{nt}(a) = R_{nt}(G(a))$ for $a \in \{0, 1\}$, then the wrapper algorithm will (1) return a data set that satisfies the desired power constraints and (2) have expected regret no larger than $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$.*

Proof. (Sketch) Our wrapper algorithm makes the input algorithm $\mathcal{A}$ believe that the environment is more stochastic than it is. If algorithm $\mathcal{A}$ achieves some rate in this more stochastic environment, then it will be optimal with respect to the clipped oracle. Full proof in Appendix Section A.6. □

There exist many environments Ω which are closed under the reward transformation above, including Abbasi-Yadkori et al. [2011], Agrawal and Goyal [2012], Langford and Zhang [2007]. In Appendix Section A.6, we describe a large number of settings under which this wrapper could be used.

Meta-Algorithm: Selective Data Dropping The action flipping strategy above is simple, but it adds stochasticity into the agent's perceived environment. If this stochasticity is not desired, or for algorithms where the environmental class Ω is not closed under our stochastic transformation G, we provide another wrapper meta-algorithm that uses the following strategy:

1. Produce $\pi_{\mathcal{A}}(C_{nt})$ as before. If sampling $A_{nt} \sim \texttt{Bern}(\pi_{\mathcal{A}}(C_{nt}))$ would have produced the same action as sampling $A'_{nt} \sim \texttt{Bern}(\texttt{clip}(\pi_{\mathcal{A}}(C_{nt})))$ (see detailed algorithm description in Appendix A.8 as to how to do this efficiently), then perform A_{nt}; else perform A'_{nt}.
2. The algorithm $\mathcal{A}$ stores the tuple C_{nt}, A_{nt}, R_{nt} *if* A_{nt} was performed; else it stores nothing from that interaction.
3. The scientist *always* stores the tuple C_{nt}, A'_{nt}, R_{nt}

Theorem 5. *Given input $\pi_{\min}, \pi_{\max}$ and a contextual bandit algorithm $\mathcal{A}$. Assume algorithm $\mathcal{A}$ has a regret bound $\mathcal{R}(T)$ when one of the following holds for the setting $\mathcal{B}$: (1) under $\mathcal{B}$ the data generating process for each context is independent of history, or (2) under $\mathcal{B}$ the context depends on the history, and the bound $\mathcal{R}$ for algorithm $\mathcal{A}$ is robust to an adversarial choice of context.*

Then our wrapper algorithm will (1) return a dataset that satisfies the desired power constraints under the data generation process of Section 3 and (2) has expected regret no larger than $\mathcal{R}(\pi_{max}T) + (1 - \pi_{max})T$ if assumptions $\mathcal{B}$ are satisfied in the true environment.

We prove in Appendix Section A.8 that as long as the environment Ω remains closed when data are dropped, the expected regret rate is no worse than $R(\pi_{\max}T)$ with respect to a clipped oracle.

6 Experiments

We now demonstrate properties of our power-constrained bandits on several environments, ranging from standard semiparametric and adversarial settings to a realistic mobile health simulator.

6.1 Settings, Baselines, and Metrics

Standard Environments Our semiparametric contextual bandit (SCB) samples Z_{nt} and δ_0 uniformly from a sphere and ϵ_{nt} are i.i.d.. Our adversarial semiparametric (ASCB) setting is from Krishnamurthy et al. [2018]; it uses the nonparametric component of the reward γ_{nt} to corrupt the information the learner receives. Details in Appendix C.1, C.2.

Realistic Mobile Health Simulator Liao et al. [2016] introduced a mobile health simulator motivated by the HeartSteps mobile health study that used messages to increase the user's physical activity. Each simulated user n participates for 90 days, $A_{nt} = 1$ represents a message is delivered, and R_{nt} represents the square root of step count at day t. The γ_{nt} decreases linearly over time as people engage more at the start of the study. Z_{nt} is created by experts such that the treatment effect $Z_{nt}^{\intercal}\delta_0$ starts small at day 0, peaks at day 45, and decays to 0 at day 90 as people disengage. The ϵ_{nt} follows a AR(1) process. Details in Appendix C.3.

Baselines To our knowledge, bandit algorithms with power guarantees are novel. Thus, we compare our power-preserving strategies applied to various algorithms focused on minimizing regret: ACTS, BOSE described in Section 5.1 and linear Upper Confidence Bound (linUCB [Chu et al., 2011]) which is similar to OFUL in Section 5.1 but simpler to implement and more commonly used in practice. We also include a Fixed Policy ($\pi_{nt} = 0.5$ for all n, t), a clipped (power-preserving) oracle, and standard (non-power preserving) oracle. Algorithm details in Appendix B.

Metrics For each of the algorithms, we compute the Type 1 error, the power (under correct and incorrect specifications of the effect size and the reward mean structure) and the average return. We also compute the regret with respect to a clipped oracle (*reg$_c$*) as

$$\text{reg}_{\text{c}} = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{nt} + \pi_{nt}^{*} Z_{nt}^{\intercal}\delta_0 + \epsilon_{nt}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{nt}\right] \tag{6}$$

Hyperparameters All of the algorithms require priors or other hyperparameters, which are selected by maximizing the average return. The same parameter values are used in the adapted and non-adapted versions of the algorithms. We listed the hyperparameter settings in Appendix Table 1.

6.2 Results

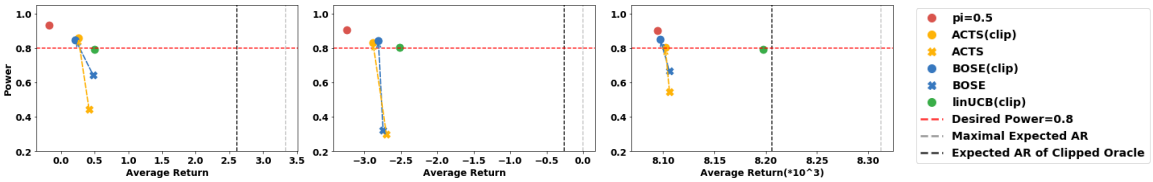
We generate $1,000$ simulation datasets for each experiment. We set the desired Type 1 error $\alpha_0 = 0.05$ and desired power $1 - \beta_0 = 0.8$. For the s^{th} simulation dataset, we estimate $\hat{\theta}^{(s)}$ using Equation 3. With all simulation datasets, we empirically estimate one $\hat{\Sigma}_\theta$ using Equation 4. Then we compute $\hat{\delta}^{(s)}$ and $\hat{\Sigma}_\delta$ from $\hat{\theta}^{(s)}$ and $\hat{\Sigma}_\theta$ respectively. The test statistics $\{N\hat{\delta}^{(s)\intercal}\hat{\Sigma}_\delta^{-1}\hat{\delta}^{(s)}\}_{s=1}^{1000}$ follow the distribution described in Section 3. We find the following main effects.

When there is no treatment effect, we recover the correct Type 1 error. Before power analysis, a basic but critical question is whether we achieve the correct Type 1 error when there is no treatment effect (see set-up in Appendix C). We confirm in Appendix Table 3 that Type 1 errors are near but slightly higher than 0.05. This makes sense as the estimated covariance $\hat{\Sigma}_\delta$ is biased downwards due to sample size [Mancl and DeRouen, 2001]; if needed, this could be controlled by various adjustments or by using critical values based on Hotelling's T^2 distribution instead of χ^2 distribution.

When there is a treatment effect, we recover the correct power if we guessed the effect size correctly. From Figure 1, we see that, as expected, Fixed Policy ($\pi = 0.5$) achieves the highest power because the exploration is maximal. Comparing the powers of non-clipped algorithms to those of clipped algorithms, we see that our clipping scheme is required to achieve the desired power as non-clipped algorithms are below the desired power level while clipped ones are above. Since linUCB selects between actions with probability π_{min} or π_{max}, the power is approximately 0.80. We cannot conduct statistical analyses on linUCB without clipping as our test requires a stochastic policy.

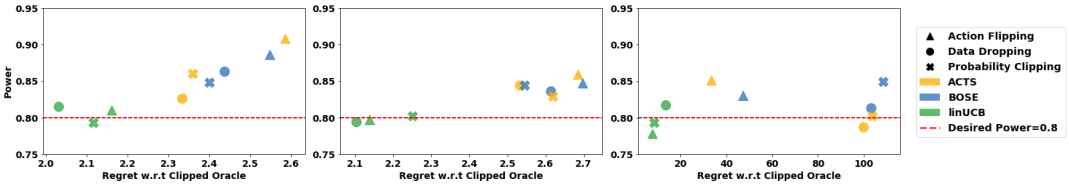
The power is reasonably robust to mis-estimated effect size and mis-specified mean reward. We consider the effect on the power when our guess of the effect size is overestimated ($Z_t\delta_{est} > Z_t\delta_0$) or underestimated ($Z_t\delta_{est} < Z_t\delta_0$). For all environments, we tested two different mis-estimated treatment effects, 1.1 times smaller and 1.1 times larger than the true effect (for each, the environment parameters and corresponding solved values for π_{min}'s and π_{max}'s are in Appendix Table 2). Appendix Figure 3 shows that underestimates of δ_0 result in more exploration, and thus higher power but lower return. Overestimates result in less exploration, lower power, and higher returns. linUCB is least robust to mis-estimated effect size as it drops greatest when the effect size is underestimated.

Secondly, we prove in Appendix A.4 that when the marginal reward model is mis-specified, the resulting power will decrease. The amount of decrease in power may vary. We experimentally confirm this in Appendix Figure 4, where we use $B_{nt} = 1$ as a bad approximation of the marginal reward structure. The figure shows that in SCB and ASCB, even with bad approximations, the resulting



(a) SCB (b) ASCB (c) Mobile Health

Figure 1: Average Return v.s. Resulting power: x-axis denotes average return and y-axis denotes the resulting power. Power tends to decrease as average return increases, though clipped linUCB preserves power with a stronger performance than the other baselines.



(a) SCB (b) ASCB (c) Mobile Health

Figure 2: Regret w.r.t clipped oracle v.s. Resulting power with different wrapper algorithms: x-axis denotes regret with respect to clipped oracle and y-axis denotes the resulting power.

powers are similar to those of correctly specified models. In mobile health, Fixed Policy ($\pi = 0.5$) performs most robustly while the other three suffers a drop in resulting power, with ACTS drops the most to a power above 0.5.

Different algorithms have different regrets, but all still converge as expected with respect to the clipped oracle. Fixed Policy ($\pi = 0.5$) has the lowest average return, as we see in Figure 1. LinUCB, which makes the strongest assumptions w.r.t. regret, has the highest average return among all algorithms. ACTS and BOSE performs similarly.

Overall, the regret of clipped algorithms with respect to a clipped oracle is roughly on the same scale as the regret of non-clipped algorithms with respect to a non-clipped oracle. (The distance between crosses(x) and the grey dashed line and the distance between circles(o) and the black dashed line are similar). The complete results of AR and regrets are listed in Appendix Table 5.

There can be trade-off between regret and the resulting power. Figure 1 also shows that the average return often increases as the power overall decreases. For example, Fixed Policy ($\pi = 0.5$) gives us the highest power but the lowest average return. Without probability clipping, ACTS and BOSE achieves higher average return but results in less power. Interestingly, clipped linUCB preserves the desired power guarantee while offering stronger performance than the other approaches.

All wrapper algorithms achieve good regret rate with slightly different trade-offs given the situation. Figure 2 shows that, for linUCB, all three strategies perform similarly in terms of power and regret. For ACTS, BOSE in the SCB, ASCB environments, action flipping results in most power and most regret as we have more exploration due to the forced stochasticity and a smaller perceived treatment effect in the modified environment (unlike dropping). In mobile health, we see the opposite. We speculate that some of these differences could be due to hyperparameter choices, which we only set for the original environment in order to be fair to all methods. In practice, results vary when we optimize hyperparameters for regret performance rather than setting them based on theoretical bounds; additional wrapper-specific hyperparameter search may allow for better performance.

7 Discussion & Conclusion

Our work provides a general approach to satisfy an important need for ensuring that studies are sufficiently powered while also personalizing for an individual. Our wrapper algorithms guarantee that power constraints are met without significant regret increase for a general class of algorithms; we

also provide stronger regret bounds for specific algorithms. Our results show that our algorithms meet their claims and are also robust to mis-specified models and mis-estimated effect size. In practice, how one chooses one wrapper algorithm over others depends on the designer's preference.

Finally, while we have focused on bandits in this work, we note that our power guarantees allow the feature Z_{nt} to be a function of full history H_{nt}; thus our results in Section 4 will give us power to identify marginal treatment effects *even if the environment is an MDP*. The action flipping strategy of Section 5.2 yields the following corollary to Theorem 4 (proof and details in Appendix A.7):

Corollary 1. *Given $\pi_{\min},\ \pi_{\max}$ and an MDP algorithm $\mathcal{A}$, assume that algorithm $\mathcal{A}$ has an expected regret $\mathcal{R}(T)$ for any MDP environment in Ω, with respect to an oracle $\mathcal{O}$. Under stochastic transformation G, if there exists an environment in Ω that contains the new transition probability function: $P'^{a}_{s,s'} = \left(\pi^{a}_{\min}\pi^{1-a}_{\max}P^{0}_{s,s'} + \pi^{1-a}_{\min}\pi^{a}_{\max}P^{1}_{s,s'}\right)$ then the wrapper algorithm will (1) return a data set that satisfies the desired power constraints and (2) have expected regret no larger than $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$.*

This is also an interesting direction for future work.

8 Broader Impact

Traditional statistical studies face the tension of treatment effect detection and better treatment assignment. Our work is extremely practical: it can be applied to a broad of statistical studies as we demonstrate that the study can be sufficiently powered with simple adaptations of existing algorithms. While we focus on derivations for a single power constraint, in settings where potential secondary analyses are known, one can seamlessly apply our methods to guarantee power for multiple analyses by considering the minimum $\pi_{\max}$ and maximum $\pi_{\min}$.

Additionally, for researchers who really care about maximizing treatment personalization, it may be possible to get better regrets if the clipping is allowed to change over time (but still be sufficiently bounded away from 0 and 1 to preserve the ability to perform post-hoc analyses).

References

Y. Abbasi-Yadkori, P. David, and C. Szepesvari. Improved algorithms for linear stochastic bandits. In *NIPS*, page 2312–232, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs, 2012.

Akshay Krishnamurthy, Zhiwei(Steven) Wu, and Vasilis Syrgkanis. Semiparametric contextual bandits. *arXiv preprint arXiv:1803.04204*, 2018.

Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A Murphy. Calculations for micro-randomized trials in mhealth. *Statistics in Medicine*, 35(12):1944–1971, 2016.

Predrag Klasnja, Eric B Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220, 2015.

JN Kramer, F Kunzler, V Mishra, B Presset, D Kotz, S Smith, U Scholz, and T Kowatsch. Investigating intervention components and exploring states of receptivity for a smartphone app to promote physical activity: Protocol of a microrandomized trial. *JMIR Res Protoc*, 8(1):e11540, 2019.

Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 189–203. Springer, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Proceedings of Machine Learning Research: 31st Annual Conference on Learning Theory*, page 1–32, 2018.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. preprint, 2019. URL `https://tor-lattimore.com/downloads/book/book.pdf`.

Kristjan Greenewald, Ambuj Tewari, Susan Murphy, and Predag Klasnja. Action centered contextual bandits. In *Advances in Neural Information Processing Systems*, pages 5977–5985, 2017.

Rémy Degenne, Thomas Nedelec, Clément Calauzènes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

A. Erraqabi, A. Lazaric, M. Valko, E. Brunskill, and Y.E. Liu. Trading off rewards and errors in multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Xinkun Nie, Xiaoying Tian, Jonathan Taylor, and James Zou. Why adaptively collected data have negative bias and how to correct for it. *International Conference on Artificial Intelligence and Statistics*, 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.

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.

S Faye Williamson, Peter Jacko, Sofía S Villar, and Thomas Jaki. A bayesian adaptive design for clinical trials in rare diseases. *Computational statistics & data analysis*, 113:136–153, 2017.

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

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *33rd International Conference on Machine Learning*, 2016.

Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *36th International Conference on Machine Learning*, 2019.

Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113 (523):1112–1121, 2018.

Jesse Dallery, Rachel N Cassidy, and Bethany R Raiff. Single-case experimental designs to evaluate novel technology-based health interventions. *Journal of medical Internet research*, 15(2):e22, 2013.

Elizabeth V Korinek, Sayali S Phatak, Cesar A Martin, Mohammad T Freigoun, Daniel E Rivera, Marc A Adams, Pedja Klasnja, Matthew P Buman, and Eric B Hekler. Adaptive step goals and rewards: a longitudinal growth model of daily steps for a smartphone-based walking intervention. *Journal of behavioral medicine*, 41(1):74–86, 2018.

David J Albers, Matthew Levine, Bruce Gluckman, Henry Ginsberg, George Hripcsak, and Lena Mamykina. Personalized glucose forecasting for type 2 diabetes using data assimilation. *PLoS computational biology*, 13(4), 2017.

P. Klasnja, S. Smith, N.J. Seewald, A. Lee, K. Hall, B. Luers, E.B. Hekler, and S.A Murphy. Efficacy of contextually-tailored suggestions for physical activity: A micro-randomized optimization trial of heartsteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019.

N Bidargaddi, D. Almirall, S.A. Murphy, I Nahum-Shani, M. Kovalcik, T. Pituch, H. Maaieh, and V. Strecher. To prompt or not to prompt? a micro-randomized trial of time-varying push notifications to increase proximal engagement with a mobile health application. *JMIR mHealth UHealth*, 6(11):e10123, 2018.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, pages 817–824. Citeseer, 2007.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Lloyd A Mancl and Timothy A DeRouen. A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57(1):126–134, 2001.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 56(4):931–954, 1988.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

A Proofs

A.1 Proof of Theorem 1

Theorem 6. *Under assumptions in Section 3 of main paper, and the assumption that matrices* $\mathbb{E}[\sum_{t=1}^{T} Z_{nt}Z_{nt}^{\intercal}]$, *and* $\mathbb{E}\left[\sum_{t=1}^{T} \frac{B_{nt}B_{nt}^{\intercal}}{\pi_{nt}(1-\pi_{nt})}\right]$ *are invertible, the distribution of* $\sqrt{N}(\hat{\delta}-\delta_0)$ *converges, as* N *increases, to a normal distribution with* 0 *mean and covariance* $\Sigma_\delta = Q^{-1}WQ^{-1}$, *where*

$$Q = \mathbb{E}\left[\sum_{t=1}^{T} Z_{nt}Z_{nt}^{\intercal}\right]^{-1},$$

$$W = \mathbb{E}\left[\sum_{t=1}^{T} \frac{\left(R_{nt} - X_{nt}^{\intercal}\begin{bmatrix}\gamma^* \\ \delta_0\end{bmatrix}\right)(A_{nt}-\pi_{nt})Z_{nt}}{\pi_{nt}(1-\pi_{nt})} \times \sum_{t=1}^{T} \frac{\left(R_{nt} - X_{nt}^{\intercal}\begin{bmatrix}\gamma^* \\ \delta_0\end{bmatrix}\right)(A_{nt}-\pi_{nt})Z_{nt}^{\intercal}}{\pi_{nt}(1-\pi_{nt})}\right],$$

where $\gamma^* = \mathbb{E}\left[\sum_{t=1}^{T} \frac{B_{nt}B_{nt}^{\intercal}}{\pi_{nt}(1-\pi_{nt})}\right]^{-1} \mathbb{E}\left[\sum_{t=1}^{T} \frac{B_{nt}R_{nt}}{\pi_{nt}(1-\pi_{nt})}\right]$ *and* $X_{nt} = \begin{bmatrix} B_{nt} \\ (A_{nt}-\pi_{nt})Z_{nt}\end{bmatrix}$.

Proof. Note that since the time series, $n = 1, \ldots, N$ are independent and identically distributed, Q, W, γ^* do not depend on n. Suppose the marginal reward can be written as

$$\mathbb{E}[R_{nt}|H_{nt}] = B_{nt}^{\intercal}\gamma_0 \tag{7}$$

the estimated effect $\hat{\delta}$ is the minimizer of the loss

$$L(\gamma, \delta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} \frac{\left(R_{nt} - B_{nt}^{\intercal}\gamma - (A_{nt}-\pi_{nt})Z_{nt}^{\intercal}\delta\right)^2}{\pi_{nt}(1-\pi_{nt})}$$

In the above loss function the action is centered by probability that the action is 1 (i.e., $A_{nt} - \pi_{nt}$); this is a classical orthogonalization trick used in both statistics and economics [Robinson, 1988, Boruvka et al., 2018]. This orthogonalization allows one to prove that the asymptotic (large N, fixed T) distribution of $\sqrt{N}(\hat{\delta} - \delta_0)$ is Gaussian even if the working model in Equation 7 is false (Boruvka et al. [2018]). A similar orthogonalization trick has been used in the bandit literature by Krishnamurthy et al. [2018], Greenewald et al. [2017] so as to allow a degree of non-stationarity.

Let, $\theta = \begin{bmatrix}\gamma \\ \delta\end{bmatrix}$, $X_{nt} = \begin{bmatrix} B_{nt} \\ (A_{nt}-\pi_{nt})Z_{nt}\end{bmatrix} \in \mathbb{R}^{(q+p)\times 1}$, where $q,\ p$ are the dimensions of B_{nt}, Z_{nt} respectively. Note X_{nt} is random because $B_{nt}, A_{nt}, \pi_{nt}, Z_{nt}$ depend on random history. The loss can be rewritten as

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} \frac{\left(R_{nt} - X_{nt}^{\intercal}\theta\right)^2}{\pi_{nt}(1-\pi_{nt})}$$

By solving $\frac{\partial L}{\partial \theta} = 0$, we have

$$\hat{\theta}_N = \left(\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{X_{nt} X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})} \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{R_{nt} X_{nt}}{\pi_{nt}(1-\pi_{nt})} \right)$$

where $\hat{\theta}_N$ denotes the estimate of θ with N samples. We drop the subscript N in the following text for short notation. Using the weak law of large numbers and the continuous mapping theorem we have that $\hat{\theta}$ converges in probability, as $N \to \infty$ to $\theta^* = \begin{bmatrix} \gamma^* \\ \delta^* \end{bmatrix}$ where

$$\theta^* = \left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{X_{nt} X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})} \right] \right)^{-1} \left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{R_{nt} X_{nt}}{\pi_{nt}(1-\pi_{nt})} \right] \right).$$

We then show that $\delta^* = \delta_0$ and γ^* is given by the statement in the theorem. One can do this directly using the above definition for θ^* or by noting that that $\mathbb{E}[\frac{\partial L}{\partial \theta}]|_{\theta=\theta^*} = 0$. We use the latter approach here. Recall all the time series are independent and identical; thus

$$\mathbb{E}\left[\frac{\partial L}{\partial \theta} \right] \Big|_{\theta=\theta^*} = \mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{R_{nt} - B_{nt}^\intercal \gamma^* - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^*}{\pi_{nt}(1-\pi_{nt})} \begin{bmatrix} B_{nt} \\ (A_{nt} - \pi_{nt}) Z_{nt} \end{bmatrix} \Bigg] = 0 \quad (8)$$

We first focus on the part with $(A_{nt} - \pi_{nt}) Z_{nt}$ which is related to δ^*

$$\mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{R_{nt} - B_{nt}^\intercal \gamma^* - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^*}{\pi_{nt}(1-\pi_{nt})} (A_{nt} - \pi_{nt}) Z_{nt} \Bigg] = 0$$

Note that given history H_{nt}, $A_{nt} \perp B_{nt}, Z_{nt}$. Thus, for all n, t,

$$\begin{aligned} \mathbb{E}\left[\frac{-B_{nt}^\intercal \gamma^* (A_{nt} - \pi_{nt}) Z_{nt}}{\pi_{nt}(1-\pi_{nt})} \right] &= \mathbb{E}\left[-B_{nt}^\intercal \gamma^* \mathbb{E}\left[\frac{A_{nt} - \pi_{nt}}{\pi_{nt}(1-\pi_{nt})} \middle| H_{nt} \right] Z_{nt} \right] \\ &= \mathbb{E}\left[-B_{nt}^\intercal \cdot 0 \cdot Z_{nt} \right] = 0 \end{aligned}$$

which leaves us with

$$\mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{R_{nt} - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^*}{\pi_{nt}(1-\pi_{nt})} (A_{nt} - \pi_{nt}) Z_{nt} \Bigg] = 0.$$

We rewrite $R_{nt} = R_{nt}(0) + [R_{nt}(1) - R_{nt}(0)] A_{nt}$. Note for all n, t,

$$\mathbb{E}\left[\frac{R_{nt}(0)(A_{nt} - \pi_{nt}) Z_{nt}}{\pi_{nt}(1-\pi_{nt})} \right] = \mathbb{E}\left[R_{nt}(0) \mathbb{E}\left[\frac{A_{nt} - \pi_{nt}}{\pi_{nt}(1-\pi_{nt})} \middle| H_{nt} \right] Z_{nt} \right] = 0.$$

Thus, we only need to consider,

$$\mathbb{E}\Bigg[\sum_{t=1}^{T} \frac{[R_{nt}(1) - R_{nt}(0)] A_{nt} - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^*}{\pi_{nt}(1-\pi_{nt})} (A_{nt} - \pi_{nt}) Z_{nt} \Bigg] = 0 \quad (9)$$

We observe that for all n, t,

$$\mathbb{E}\left[\frac{[R_{nt}(1) - R_{nt}(0)] \pi_{nt}}{\pi_{nt}(1-\pi_{nt})} (A_{nt} - \pi_{nt}) Z_{nt} \right] = 0. \quad (10)$$

Subtracting Equation 10 from Equation 9, we obtain

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \frac{[R_{nt}(1) - R_{nt}(0)](A_{nt} - \pi_{nt}) - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^*}{\pi_{nt}(1-\pi_{nt})} (A_{nt} - \pi_{nt}) Z_{nt} \right] &= 0 \\ \mathbb{E}\left[\sum_{t=1}^{T} \frac{[R_{nt}(1) - R_{nt}(0) - Z_{nt}^\intercal \delta^*](A_{nt} - \pi_{nt})^2 Z_{nt}}{\pi_{nt}(1-\pi_{nt})} \right] &= 0 \end{aligned}$$

Since that given history H_{nt}, $A_{nt} \perp R_{nt}(0), R_{nt}(1), Z_{nt}$ and

$$\mathbb{E}\left[\frac{(A_{nt}-\pi_{nt})^2}{\pi_{nt}(1-\pi_{nt})}\middle| H_{nt}\right] = 1,$$

we have

$$\mathbb{E}\left[\sum_{t=1}^{T}(R_{nt}(1) - R_{nt}(0) - Z_{nt}^{\intercal}\delta^*)Z_{nt}\right] = 0$$

Solve for δ^*, by Equation 1 in the main paper ($\mathbb{E}\left[R_{nt}(1) - R_{nt}(0)|H_{nt}\right] = Z_{nt}^{\intercal}\delta_0$), we have

$$\mathbb{E}\left[\sum_{t=1}^{T} Z_{nt}Z_{nt}^{\intercal}\right](\delta_0 - \delta^*) = 0 \;\Rightarrow\; \delta^* = \delta_0.$$

Similarly, we can solve for γ^*. Focus on the part related to γ^* in Equation 8, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T}\frac{R_{nt} - B_{nt}^{\intercal}\gamma^* - (A_{nt}-\pi_{nt})Z_{nt}^{\intercal}\delta^*}{\pi_{nt}(1-\pi_{nt})}B_{nt}\Big] = 0.$$

Since for all n, t, $\mathbb{E}\left[\frac{(A_{nt}-\pi_{nt})}{\pi_{nt}(1-\pi_{nt})}\middle| H_{nt}\right] = 0$, we have

$$\mathbb{E}\Big[\sum_{t=1}^{T}\frac{(R_{nt} - B_{nt}^{\intercal}\gamma^*)B_{nt}}{\pi_{nt}(1-\pi_{nt})}\Big] = 0.$$

Hence,

$$\gamma^* = \left(\mathbb{E}\Big[\sum_{t=1}^{T}\frac{B_{nt}B_{nt}^{\intercal}}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1}\mathbb{E}\Big[\sum_{t=1}^{T}\frac{B_{nt}R_{nt}}{\pi_{nt}(1-\pi_{nt})}\Big].$$

Thus $\delta^* = \delta_0$ and γ^* is given by the theorem statement.

Now we provide a sketch of the proof that

$$\sqrt{N}(\hat{\delta} - \delta_0) \sim \mathcal{N}\left(0, \left(\mathbb{E}\Big[\sum_{t=1}^{T} Z_{nt}Z_{nt}^{\intercal}\Big]\right)^{-1} W \left(\mathbb{E}\Big[\sum_{t=1}^{T} Z_{nt}Z_{nt}^{\intercal}\Big]\right)^{-1}\right)$$

where

$$W = \mathbb{E}\Big[\sum_{t=1}^{T}\frac{(R_{nt} - X_{nt}^{\intercal}\theta^*)(A_{nt}-\pi_{nt})Z_{nt}}{\pi_{nt}(1-\pi_{nt})}\sum_{t=1}^{T}\frac{(R_{nt} - X_{nt}^{\intercal}\theta^*)(A_{nt}-\pi_{nt})Z_{nt}^{\intercal}}{\pi_{nt}(1-\pi_{nt})}\Big].$$

Given the facts that

1. $\frac{\partial L}{\partial \hat{\theta}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{R_{nt}-X_{nt}^{\intercal}\hat{\theta}}{\pi_{nt}(1-\pi_{n}t)}X_{nt} = 0$
2. $\mathbb{E}\Big[\frac{\partial L}{\partial \theta}\Big]_{\theta=\theta^*} = \mathbb{E}\Big[\sum_{t=1}^{T}\frac{R_{nt}-X_{nt}^{\intercal}\theta^*}{\pi_{nt}(1-\pi_{nt})}X_{nt}\Big] = 0$

We can now write

$$0 = \underbrace{\frac{\partial L}{\partial \hat{\theta}} - \mathbb{E}\Big[\frac{\partial L}{\partial \theta}\Big]_{\theta=\hat{\theta}}}_{\text{Term 1}} + \underbrace{\mathbb{E}\Big[\frac{\partial L}{\partial \theta}\Big]_{\theta=\hat{\theta}} - \mathbb{E}\Big[\frac{\partial L}{\partial \theta}\Big]_{\theta=\theta^*}}_{\text{Term 2}} \tag{11}$$

We first focus on Term 2.

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=1}^{T}\frac{R_{nt} - X_{nt}^{\intercal}\hat{\theta}}{\pi_{nt}(1-\pi_{nt})}X_{nt}\Big] - \mathbb{E}\Big[\sum_{t=1}^{T}\frac{R_{nt} - X_{nt}^{\intercal}\theta^*}{\pi_{nt}(1-\pi_{nt})}X_{nt}\Big] \\
=&\mathbb{E}\Big[\sum_{t=1}^{T}\frac{1}{\pi_{nt}(1-\pi_{nt})}\begin{bmatrix} B_{nt}B_{nt}^{\intercal} & B_{nt}Z_{nt}^{\intercal}(A_{nt}-\pi_{nt}) \\ B_{nt}^{\intercal}Z_{nt}(A_{nt}-\pi_{nt}) & Z_{nt}Z_{nt}^{\intercal}(A_{nt}-\pi_{nt})^2\end{bmatrix}\Big](\theta^* - \hat{\theta})
\end{aligned}$$

Note cross terms are 0 and $\mathbb{E}\left[\sum_{t=1}^{T} \frac{Z_{nt}Z_{nt}^\intercal (A_{nt}-\pi_{nt})^2}{\pi_{nt}(1-\pi_{nt})}\right] = \mathbb{E}\left[\sum_{t=1}^{T} Z_{nt}Z_{nt}^\intercal\right]$.
We have

$$-\mathbb{E}\left[\sum_{t=1}^{T}\begin{bmatrix}\frac{B_{nt}B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})} & 0\\ 0 & Z_{nt}Z_{nt}^\intercal\end{bmatrix}\right](\hat{\theta}-\theta^*) = \text{Term 2}.$$

We now look at Term 1. Define

$$u_N(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{R_{nt}-X_{nt}^\intercal\theta}{\pi_{nt}(1-\pi_{nt})}X_{nt} - \mathbb{E}\left[\sum_{t=1}^{T}\frac{R_{nt}-X_{nt}^\intercal\theta}{\pi_{nt}(1-\pi_{nt})}X_{nt}\right]$$

and note that Term 1 is $u_N(\hat{\theta})$ estimated with N samples. We again drop N for short. Now,

$$\begin{aligned} u(\hat{\theta})-u(\theta^*) &= -\left(\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}\frac{X_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})} - \mathbb{E}\left[\sum_{t=1}^{T}\frac{X_{nt}X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right]\right)(\hat{\theta}-\theta^*)\\ u(\hat{\theta}) &= -v(\hat{\theta}-\theta^*)+u(\theta^*)\end{aligned}$$

Plug Term 1 and Term 2 back into Equation 11 we have,

$$\mathbb{E}\left[\sum_{t=1}^{T}\begin{bmatrix}\frac{B_{nt}B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})} & 0\\ 0 & Z_{nt}Z_{nt}^\intercal\end{bmatrix} + v\right](\hat{\theta}-\theta^*) = u(\theta^*).$$

where by the weak law of large numbers v converges in probability to 0. Note $\mathbb{E}[u(\theta^*)]=0$. Apply central limit theorem on $\sqrt{N}u(\theta^*)$; that is as $N\to\infty$, $\sqrt{N}u(\theta^*)$ converges in distribution to $\mathcal{N}(0,\Sigma)$, where

$$\Sigma = \mathbb{E}\left[\sum_{t=1}^{T}\frac{(R_{nt}-X_{nt}^\intercal\theta^*)X_{nt}}{\pi_{nt}(1-\pi_{nt})}\sum_{t=1}^{T}\frac{(R_{nt}-X_{nt}^\intercal\theta^*)X_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right]$$

Denote the lower right matrix of Σ by W. Then

$$W = \mathbb{E}\left[\sum_{t=1}^{T}\frac{(R_{nt}-X_{nt}^\intercal\theta^*)(A_{nt}-\pi_{nt})Z_{nt}}{\pi_{nt}(1-\pi_{nt})}\sum_{t=1}^{T}\frac{(R_{nt}-X_{nt}^\intercal\theta^*)(A_{nt}-\pi_{nt})Z_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right]$$

and $\sqrt{N}(\hat{\delta}-\delta_0)\sim\mathcal{N}(0,\Sigma_\delta)$ where $\Sigma_\delta = \left(\mathbb{E}\left[\sum_{t=1}^{T}Z_{nt}Z_{nt}^\intercal\right]\right)^{-1} W \left(\mathbb{E}\left[\sum_{t=1}^{T}Z_{nt}Z_{nt}^\intercal\right]\right)^{-1}$.
Under the null hypothesis $H_0:\delta_0=0$, $N\hat{\delta}{\Sigma_\delta}^{-1}\hat{\delta}$ asymptotically follows χ^2 with degree of freedom p. Under the alternate hypothesis $H_1:\delta_0=\delta$, $N\hat{\delta}{\Sigma_\delta}^{-1}\hat{\delta}$ asymptotically follows a non-central χ^2 with degree of freedom p and non-central parameter $c_N = N(\delta^\intercal{\Sigma_\delta}^{-1}\delta)$. We estimate W by putting in sample averages and plugging in $\hat{\theta}$ as θ^*. □

Remark 1. *Suppose we make the further assumption that ϵ_{nt} is independent of A_{nt} conditional on H_{nt} ($i.e.\mathbb{E}[\epsilon_{nt}|A_{nt},H_{nt}]=0$) and that $Var(\epsilon_{nt}|H_{nt})=\sigma^2$. Then W can be further simplified as*

$$W = \mathbb{E}\left[\sum_{t=1}^{T}\frac{\sigma^2}{\pi_{nt}(1-\pi_{nt})}Z_{nt}Z_{nt}^\intercal\right] + \mathbb{E}\left[\sum_{t=1}^{T}\frac{(\gamma_{nt}+\pi_{nt}Z_{nt}^\intercal\delta_0 - B_{nt}^\intercal\gamma^*)^2 Z_{nt}Z_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\right],$$

Proof. Since in any cross term,

1. $\mathbb{E}[A_{nt}-\pi_{nt}|H_{nt}]=0$,
2. $Z_{nt}, Z_{nt'}, B_{nt}, B_{nt'}, \gamma_{nt}, \gamma_{nt'}, \epsilon_{nt'}, A_{nt'}, \pi_{nt'}$ are all determined by H_{nt} when $t'<t$,
3. and $\mathbb{E}[\epsilon_{nt}|A_{nt},H_{nt}]=0$,

we can rewrite W as

$$W = \mathbb{E}\Big[\sum_{t=1}^{T} \frac{(R_{nt} - X_{nt}^\intercal \theta^*)^2 (A_{nt} - \pi_{nt})^2 Z_{nt} Z_{nt}^\intercal}{\pi_{nt}^2 (1-\pi_{nt})^2}\Big].$$

As in the Remark suppose we make the further assumption that $Var(\epsilon_{nt}|H_{nt}) = \sigma^2$. Rewrite $R_{nt} - X_{nt}^\intercal \theta^* = R_{nt} - B_{nt}^\intercal \gamma^* - (A_{nt} - \pi_{nt}) Z_{nt}^\intercal \delta^* + \gamma_{nt} - \gamma_{nt} = \epsilon_{nt} + \gamma_{nt} - B_{nt}^\intercal \gamma^*$. Then,

$$\begin{aligned} W &= \mathbb{E}\Big[\sum_{t=1}^{T} \frac{\epsilon_{nt}^2 (A_{nt} - \pi_{nt})^2}{\pi_{nt}^2 (1-\pi_{nt})^2} Z_{nt} Z_{nt}^\intercal\Big] + \mathbb{E}\Big[\sum_{t=1}^{T} \frac{(\gamma_{nt} - B_{nt}^\intercal \gamma^* + \pi_{nt} Z_{nt}^\intercal \delta_0)^2 (A_{nt} - \pi_{nt})^2}{\pi_{nt}^2 (1-\pi_{nt})^2} Z_{nt} Z_{nt}^\intercal\Big] \\ &= \mathbb{E}\Big[\sum_{t=1}^{T} \frac{\sigma^2}{\pi_{nt}(1-\pi_{nt})} Z_{nt} Z_{nt}^\intercal\Big] + \underbrace{\mathbb{E}\Big[\sum_{t=1}^{T} \frac{(\gamma_{nt} + \pi_{nt} Z_{nt}^\intercal \delta_0 - B_{nt}^\intercal \gamma^*)^2 Z_{nt} Z_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]}_{\text{Term 2}}. \end{aligned} \tag{12}$$

Assuming the assumptions in the Remark, we have $\sqrt{N}(\hat{\delta} - \delta_0) \sim \mathcal{N}(0, \Sigma_\delta)$ where Σ_δ simplifies to

$$\Sigma_\delta = \mathbb{E}\left[\sum_t Z_{nt} Z_{nt}^\intercal\right]^{-1} W \mathbb{E}\left[\sum_t Z_{nt} Z_{nt}^\intercal\right]^{-1}. \tag{13}$$

where W is given in Equation 12. □

Remark 2. *Suppose the working model in Equation 7 is correct, then Σ_δ can be further simplified to*

$$\Sigma_\delta = \mathbb{E}\left[\sum_t Z_{nt} Z_{nt}^\intercal\right]^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{\sigma^2}{\pi_{nt}(1-\pi_{nt})} Z_{nt} Z_{nt}^\intercal\Big] \mathbb{E}\left[\sum_t Z_{nt} Z_{nt}^\intercal\right]^{-1}. \tag{14}$$

Proof. Recall that

$$\gamma^* = \left(\mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} R_{nt}}{\pi_{nt}(1-\pi_{nt})}\Big],$$

$R_{nt} = \gamma_{nt} + A_{nt} Z_{nt}^\intercal \delta_0 + \epsilon_{nt}$ and $\mathbb{E}[R_{nt}|H_{nt}, A_{nt}] = \gamma_{nt} + A_{nt} Z_{nt}^\intercal \delta_0$. Thus,

$$\begin{aligned} \gamma^* &= \left(\mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt}(\gamma_{nt} + A_{nt} Z_{nt}^\intercal \delta_0 + \epsilon_{nt})}{\pi_{nt}(1-\pi_{nt})}\Big] \\ &= \left(\mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt}(\gamma_{nt} + \pi_{nt} Z_{nt}^\intercal \delta_0 + \epsilon_{nt})}{\pi_{nt}(1-\pi_{nt})}\Big] \end{aligned}$$

where the last equality holds because $\mathbb{E}[\gamma_{nt} + A_{nt} Z_{nt}^\intercal \delta_0 | H_{nt}] = \gamma_{nt} + \pi_{nt} Z_{nt}^\intercal \delta_0$. Given the assumption that $\mathbb{E}[\epsilon_{nt}|A_{nt}, H_{nt}] = 0$, then for or all n, t,

$$\mathbb{E}\left[\frac{\epsilon_{nt} B_{nt}}{\pi_{nt}(1-\pi_{nt})}\right] = \mathbb{E}\left[\mathbb{E}\left[\epsilon_{nt}|H_{nt}, A_{nt}\right] \frac{B_{nt}}{\pi_{nt}(1-\pi_{nt})}\right] = 0$$

$$\text{and} \quad \gamma^* = \left(\mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt}(\gamma_{nt} + \pi_{nt} Z_{nt}^\intercal \delta_0)}{\pi_{nt}(1-\pi_{nt})}\Big].$$

When the working model in Equation 7 is true, we have

$$\mathbb{E}[R_{nt}|H_{nt}] = \mathbb{E}[\mathbb{E}[R_{nt}|H_{nt}, A_{nt}]|H_{nt}] = \gamma_{nt} + \pi_{nt} Z_{nt}^\intercal \delta_0 = B_{nt}^\intercal \gamma_0$$

and thus

$$\gamma^* = \left(\mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}^\intercal}{\pi_{nt}(1-\pi_{nt})}\Big]\right)^{-1} \mathbb{E}\Big[\sum_{t=1}^{T} \frac{B_{nt} B_{nt}}{\pi_{nt}(1-\pi_{nt})} \gamma_0\Big] = \gamma_0.$$

Additionally, Term 2 of Equation 12 is equal to 0. When the working model is false, Term 2 is positive semidefinite and $\hat{\delta}$ will likely have inflated covariance matrix. Assuming the working model is correct and assuming the assumptions in the Remark, we simply have Σ_δ stated in the Remark □

A.2 Proof of Theorem 2

Theorem 7. *Assume the working model in Equation 7 is correct. Suppose the desired Type 1 error is α_0 and the desired power is $1-\beta_0$. Set*

$$\pi_{\min}=\frac{1-\sqrt{1-4\triangle}}{2},\ \pi_{\max}=\frac{1+\sqrt{1-4\triangle}}{2},\ \triangle=\frac{\sigma^2 c_{\beta_0}}{N\delta_0^\intercal\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\delta_0}.$$

We choose c_{β_0} such that $1-\Phi_{p;c_{\beta_0}}(\Phi_p^{-1}(1-\alpha_0))=\beta_0$, where $\Phi_{p;c_{\beta_0}}$ denotes the cdf of a non-central χ^2 distribution with d.f. p and non-central parameter c_{β_0}, and Φ_p^{-1} denotes the inverse cdf of a χ^2 distribution with d.f. p. For a given trial with N subjects each over T time units, if the randomization probability is fixed as $\pi_{nt}=\pi_{\min}$ or $\pi_{\max}$, the resulting power converges to $1-\beta_0$ as $N\to\infty$.

Proof. Suppose, the working model is correct (that is, Term 2 of Equation 12=0), the effect size is correctly guessed (that is, $\delta=\delta_0$), and the patient is given treatment with a fixed probability at every trial. i.e. $p(A_{nt}=1)=\pi$.
According to Section A.1, $N\hat{\delta}\hat{\Sigma}^{-1}\hat{\delta}$ will asymptotically follows a non-central χ^2 with degree of freedom p and non-central parameter $c_N=N(\delta_0^\intercal{\Sigma_\delta}^{-1}\delta_0)$, and thus it will result in power,

$$1-\Phi_{p;c_N}(\Phi_p^{-1}(1-\alpha_0)) \tag{15}$$

Note function 15 is monotonically increasing w.r.t c_N. If we want the desired power to be asymptotically $1-\beta_0$, we need $c_N=c_{\beta_0}$. If the working model is right, plug Equation 14 into c_N, we then have

$$\begin{aligned}
c_N =& c_{\beta_0}\\
N(\delta_0^\intercal{\Sigma_\delta}^{-1}\delta_0) =& c_{\beta_0}\\
N\delta_0^\intercal\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\frac{\sigma^2}{\pi_{nt}(1-\pi_{nt})}\right]^{-1}\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\delta_0 =& c_{\beta_0}\\
\frac{N\pi(1-\pi)}{\sigma^2}\delta_0^\intercal\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\delta_0 =& c_{\beta_0}\\
\pi(1-\pi) =& \triangle,
\end{aligned} \tag{16}$$

where $\triangle$ is given by the statement in the theorem. Solving the quadratic function 16 gives us $\pi=\frac{1\pm\sqrt{1-4\triangle}}{2}$ and theorem is proved. We let $\pi_{\min}=\frac{1-\sqrt{1-4\triangle}}{2}$ and $\pi_{\max}=\frac{1+\sqrt{1-4\triangle}}{2}$. Note that $\pi_{\min}$ and $\pi_{\max}$ are symmetric to 0.5. Also note that N needs to be sufficiently large so that there exists a root for function 16. □

A.3 Proof Theorem 3

Theorem 8. *Given the values of $\pi_{\min},\pi_{\max}$ we solved in Theorem 7, if for all n and all t we have that $\pi_{nt}\in[\pi_{\min},\pi_{\max}]$, then the resulting power will converge to a value no smaller than $1-\beta_0$ as $N\to\infty$.*

Proof. Recall function 15 is monotonically increasing w.r.t c_N. To ensure the resulting power is no smaller than $1-\beta_0$, we need

$$c_N=N\delta_0^\intercal\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\frac{\sigma^2}{\pi_{nt}(1-\pi_{nt})}\right]^{-1}\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\delta_0\geq c_{\beta_0}.$$

We rewrite some of the terms for notation simplicity. Let $b=\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal\right]\delta_0$. Note b is a vector and $b\in\mathcal{R}^{p\times 1}$, where p is the dimension of Z_{nt}. Let $\Sigma=\mathbb{E}\left[\sum_{t=1}^T Z_{nt}Z_{nt}^\intercal a_{nt}\right]$ where

$a_{nt} = \frac{1}{\pi_{nt}(1-\pi_{nt})}$. Hence, we have $c_N(a_{nt}) = \frac{N}{\sigma^2} b^\intercal \Sigma^{-1} b$

$$\begin{aligned}
\frac{\partial c_N}{\partial a_{nt}} &= tr\left(\left(\frac{\partial c_N}{\partial \Sigma^{-1}}\right)^\intercal \frac{\partial \Sigma^{-1}}{\partial a_{nt}}\right) \\
&= \frac{N}{\sigma^2} tr(bb^\intercal \times -\Sigma^{-1}\frac{d\Sigma}{da_{nt}}\Sigma^{-1}) \\
&= \frac{N}{\sigma^2} tr(-bb^\intercal \Sigma^{-1}\mathbb{E}[Z_{nt}Z_{nt}^\intercal]\Sigma^{-1}) \\
&= -\frac{N}{\sigma^2}(b^\intercal \Sigma^{-1})\mathbb{E}[Z_{nt}Z_{nt}^\intercal](\Sigma^{-1}b)
\end{aligned}$$

Since $Z_{nt}Z_{nt}^\intercal$ is semi-positive definite, $\mathbb{E}[Z_{nt}Z_{nt}^\intercal]$ is semi-positive definite. Thus $\frac{\partial c_N}{\partial a_{nt}} \leq 0$ and c_N is non-increasing w.r.t a_{nt}. As long as we have

$$\frac{1}{\pi_{nt}(1-\pi_{nt})} \leq \frac{1}{\pi_{\min}(1-\pi_{\min})} = \frac{1}{\pi_{\max}(1-\pi_{\max})},$$

we will have that $c_N \geq c_{\beta_0}$.
Since for all n, t and $\pi_{nt} \in [\pi_{\min}, \pi_{\max}]$, we have

$$\pi_{nt}(1-\pi_{nt}) \geq \pi_{\min}(1-\pi_{\min}) = \pi_{\max}(1-\pi_{\max}),$$

and hence

$$\frac{1}{\pi_{nt}(1-\pi_{nt})} \leq \frac{1}{\pi_{\min}(1-\pi_{\min})} = \frac{1}{\pi_{\max}(1-\pi_{\max})}.$$

Thus, $c_N \geq c_{\beta_0}$. The power constraint will be met. □

A.4 The Effect of Model Mis-specification on Power

Corollary 2. *When the marginal reward structure is incorrect ($B_{nt}\gamma_0 \neq \gamma_{nt} + \pi_{nt}Z_{nt}^\intercal\delta_0$), the resulting power will converge to a value less than the desired power $1-\beta_0$ as $N \to \infty$.*

Proof. When the construction model is not correct, the estimator $\hat{\gamma}$ will be biased and now Term 2 in W (Equation 12) is non-zero. Using the same notation in Section A.3, $c_N = \frac{N}{\sigma^2} b^\intercal \Sigma^{-1} b$, we now have

$$\Sigma = \mathbb{E}\left[\sum_{t=1}^{T} Z_{nt}Z_{nt}^\intercal a_{nt}(1+c_{nt})\right], \text{ where } c_{nt} = \frac{(\gamma_{nt} + \pi_{nt}Z_{nt}^\intercal\delta_0 - B_{nt}^\intercal\gamma^*)^2}{\sigma^2}$$

Following similar derivation in Section A.3, we have

$$\frac{\partial c_N}{\partial c_{nt}} = -\frac{N}{\sigma^2}(b^\intercal \Sigma^{-1})\mathbb{E}[Z_{nt}Z_{nt}^\intercal a_{nt}](\Sigma^{-1}b)$$

Since $a_{nt} > 0$, $\frac{\partial c_N}{\partial c_{nt}} < 0$. Thus c_N is monotonically decreasing w.r.t c_{nt}. Hence, when the reward mean structure is incorrect, the noncentral parameter c_N will decrease and thus, power will be less than $1-\beta_0$. □

A.5 Regret Bound of OFUL with clipping

In this section, we prove that with probability clipping, OFUL will maintain the same regret rate with respect to a clipped oracle. The clipped OFUL algorithm is given in Algorithm 1. The proof below is separate for each subject; thus for simplicity we drop the subscript n (e.g. use R_t instead of R_{nt}). We also only assume that $0 < \pi_{\min} \leq \pi_{\max} < 1$, that is, we do not require the sum, $\pi_{\min} + \pi_{\max} = 1$. As we have binary actions, we can write Abbasi-Yadkori et al.'s decision set as $D_t = \{x_{t,0}, x_{t,1}\}$; the second subscript denotes the binary action and x denotes a feature vector for each action.

Clipped OFUL uses a two-step procedure to select the (binary) action in D_t. It first selects an optimistic A_t in step 4. However, instead of implementing A_t, it implements action A_t^c where

Algorithm 1 Clipped OFUL (Optimism in the Face of Uncertainty)

1: Input: $\pi_{\text{max}}, \pi_{\text{min}}$
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Observe context features for each possible action: $\{x_{t,1}, x_{t,0}\}$
4: $(A_t, \tilde{\theta}_t) = \arg\max_{(a,\theta)\in\{0,1\}\times C_{t-1}} \ell_t(a, \theta)$
5: Play $A_t^c \sim$ Bernoulli $(\pi_{\text{max}}^{A_t}\pi_{\text{min}}^{1-A_t})$ and observe reward $R_t(A_t^c)$
6: Update confidence set C_t
7: **end for**

$A_t^c \sim \texttt{Bern}(\pi_{max}^a \pi_{min}^{1-a})$ given $A_t = a$. This means that X_t in Abbasi-Yadkori et al. [2011] becomes x_{t,A_t^c} in clipped OFUL.

We use notations and assumptions similar to Abbasi-Yadkori et al. [2011]. Let $\{F_t\}_{t\geq 1}$ be a filtration, the error terms, $\{\eta_t\}_{t\geq 1}$ be a real-valued stochastic process, the features, $\{X_t\}_{t\geq 1}$ be a $\mathbb{R}^d$-valued stochastic process. η_t is F_t measurable and X_t is F_{t-1} measurable. Further assume that $||X_t||_2 \leq L$ for a constant L. Define $V = \lambda I \in \mathbb{R}^{d\times d}$ with $\lambda \geq 1$. The observed reward is assumed to satisfy

$$R_t = \theta_*^\intercal X_t + \eta_t$$

for an unknown $\theta_* \in \mathbb{R}^d$. The error term η_t is assumed to be conditionally σ-sub-Gaussian for a finite positive constant σ. This implies that $\mathbb{E}[\eta_t|F_{t-1}] = 0$ and $Var[\eta_t|F_{t-1}] \leq \sigma^2$. The coefficient satisfies $||\theta_*||_2 \leq S$ for a constant S. Lastly assume that $|\max\{\theta_*^\intercal x_{t,1}, \theta_*^\intercal x_{t,0}\}| \leq 1$.

Under these assumptions, Theorems 1, 2, Lemma 11 of Abbasi-Yadkori et al. [2011] as well as their proofs remain the same with X_t defined as x_{t,A_t^c}. Theorem 2 concerns construction of the confidence set. Neither Theorems 1, 2 or Lemma 11 concern the definition of the regret and only Theorem 3 and its proof need be altered to be valid for clipped OFUL with the regret against a clipped oracle.

Define

$$\ell_t(a, \theta) = a[\pi_{\max}\theta^\intercal x_{t,1} + (1-\pi_{\max})\theta^\intercal x_{t,0}] + (1-a)[\pi_{\min}\theta^\intercal x_{t,1} + (1-\pi_{\min})\theta^\intercal x_{t,0}].$$

Below it will be useful to note that $\ell_t(a, \theta) = \mathbb{E}[\theta^\intercal x_{t,A_t^c}|A_t = a, F_{t-1}]$.

First we define the clipped oracle. Recall the oracle action is $A_t^* = \arg\max_a \theta_* x_{t,a}$. It is easy to see that $A_t^* = \arg\max_a \mathbb{E}[\theta_*^\intercal x_{t,A_t^{c*}}|A_t^* = a, F_{t-1}]$ for $A_t^{c*} \sim$ Bernoulli$(\pi_{\max}^a \pi_{\min}^{1-a})$. The clipped oracle action is A_t^{c*}. Note that $\mathbb{E}[\theta_*^\intercal x_{t,A_t^{c*}}|A_t^* = a, F_{t-1}] = \ell_t(a, \theta_*)$. So just as A_t^* maximizes $\ell_t(a, \theta_*)$, in clipped OFUL the optimistic action, A_t, similarly provides an $\arg\max$ of $\ell_t(a, \theta)$; see line 4 in Algorithm 1.

The time t regret against the clipped oracle is given by $r_t = \ell_t(A_t^*, \theta_*) - \ell_t(A_t, \theta_*)$. In the proof to follow it is useful to note that r_t can also be written as $r_t = \mathbb{E}[\theta_*^\intercal x_{t,A_t^{c*}}|A_t^*, F_{t-1}] - \mathbb{E}[\theta_*^\intercal x_{t,A_t^c}|A_t, F_{t-1}]$. In the following we provide an upper bound on the expected regret, $\mathbb{E}\left[\sum_{t=1}^n r_t\right]$.

$$\begin{aligned}
r_t &= \ell_t(A_t^*, \theta_*) - \ell_t(A_t, \theta_*) \\
&\leq \ell_t(A_t, \tilde{\theta}_t) - \ell_t(A_t, \theta_*) \text{ (by line 4 in Alg. 1)} \\
&= \mathbb{E}[\tilde{\theta}_t^\intercal x_{t,A_t^c}|A_t, F_{t-1}] - \mathbb{E}[\theta_*^\intercal x_{t,A_t^c}|A_t, F_{t-1}] \text{ (by line 5 in Alg. 1)} \\
&= \mathbb{E}[(\tilde{\theta}_t - \theta_*)^\intercal x_{t,A_t^c}|A_t, F_{t-1}].
\end{aligned}$$

Thus we have that

$$\mathbb{E}[r_t] \leq \mathbb{E}[(\tilde{\theta}_t - \theta_*)^\intercal x_{t,A_t^c}] = \mathbb{E}[(\tilde{\theta}_t - \theta_*)^\intercal X_t]$$

with the second equality holding due to the definition of X_t. The proof of Theorem 3 in Abbasi-Yadkori et al. [2011] provides a high probability upper bound on $(\tilde{\theta}_t - \theta_*)^\intercal X_t$. In particular the proof shows that with probability at least $(1-\delta)$, for all $n \geq 1$,

$$\begin{aligned}
\sum_{t=1}^n (\tilde{\theta}_t - \theta_*)^\intercal X_t &\leq 4\sqrt{nd\log(\lambda + nL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log(1/\delta) + d\log(1 + nL/(\lambda d))}\right) \\
&\leq 4\sqrt{nd\log(\lambda + nL/d)}\left(\lambda^{1/2}S + R\sqrt{2\log(1/\delta)} + R\sqrt{d\log(1 + nL/(\lambda d))}\right)
\end{aligned}$$

since for $x > 0$, $\sqrt{1+x} \leq 1 + \sqrt{x}$.

Let $a_n = 4\sqrt{nd\log(\lambda + nL/d)}$, $b_n = \lambda^{1/2}S + R\sqrt{d\log(1 + nL/(\lambda d))}$ and $c = R\sqrt{2}$. We have $P\left[\sum_{t=1}^{n}(\tilde{\theta}_t - \theta_*)^\intercal X_t \geq a_n(b_n + c\sqrt{\log(1/\delta)}\right] \leq \delta$. Let $v = a_n\left(b_n + c\sqrt{\log(1/\delta)}\right)$ then solving for δ one obtains $\delta = \exp\left\{-\left(v - b_n a_n\right)^2/(a_n c)^2\right\}$. Thus $P\left[\sum_{t=1}^{n}(\tilde{\theta}_t - \theta_*)^\intercal X_t \geq v\right] \leq \exp\left\{-\left(v - b_n a_n\right)^2/(a_n c)^2\right\}$.

Recall that for any random variable, Y, $\mathbb{E}[Y] \leq \int_0^\infty P[Y > u]du$. Thus

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} r_t\right] &= \mathbb{E}\left[\sum_{t=1}^{n}(\tilde{\theta}_t - \theta_*)^\intercal X_t\right] \\
&\leq \int_0^\infty \exp\left\{-\left(v - b_n a_n\right)^2/(a_n c)^2\right\}dv \\
&\leq a_n c\sqrt{\pi} \\
&= 4R\sqrt{2\pi nd\log(\lambda + nL/d)}.
\end{aligned}$$

Thus the expected regret up to time n is of order $O(\sqrt{n})$ up to terms in $\log(n)$ for clipped OFUL.

A.6 Action Flipping Wrapper Algorithm

In this section, we provide full analyses of the action flipping wrapper algorithm described in Section 5.2 in the main paper. We first prove that the wrapper algorithm can be applied to a large class of algorithms and achieves good regret rate with respect to a clipped oracle and then we listed common algorithms on which the wrapper algorithm can be used. The proof below will drop the subscript n since the algorithm is for each user separately.

Meta-Algorithm: Action-Flipping (Restated)

1. Given current context C_t, algorithm $\mathcal{A}$ produces action probabilities $\pi_{\mathcal{A}}(C_t)$
2. Sample $A_t \sim \texttt{Bern}(\pi_{\mathcal{A}}(C_t))$.
3. If $A_t = 1$, sample $A'_t \sim \texttt{Bern}(\pi_{\max})$. If $A_{nt} = 0$, sample $A'_t \sim \texttt{Bern}(\pi_{\min})$.
4. We perform A'_t and receive reward R_t.
5. The algorithm $\mathcal{A}$ stores the tuple C_t, A_t, R_t. (Note that if A_t and A'_t are different, then, unbeknownst to the algorithm $\mathcal{A}$, a different action was actually performed.)
6. The scientist stores the tuple C_t, A'_t, R_t for their analysis.

Theorem 9. *Given $\pi_{\min}$, $\pi_{\max}$ and a contextual bandit algorithm $\mathcal{A}$, assume that algorithm $\mathcal{A}$ has expected regret $\mathcal{R}(T)$ for any environment in Ω, with respect to an oracle $\mathcal{O}$. If there exists a new environment in Ω such that the potential rewards, $R'_t(a) = R_t(G(a))$ for $a \in \{0, 1\}$, then the wrapper algorithm will (1) return a data set that satisfies the desired power constraints and (2) have expected regret no larger than $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$.*

Proof. **Satisfaction of power constraints:** Note that in step 6, we store the transformed action A'_t, thus we need to compute $\pi'_{\mathcal{A}}$. From step 3, we see that we can write the transformed probability $\pi'_{\mathcal{A}}$ as follows:

$$\pi'_{\mathcal{A}} = p(A'_t = 1) = \pi_{\mathcal{A}}\pi_{\max} + (1 - \pi_{\mathcal{A}})\pi_{\min}. \tag{17}$$

Since $\pi_{\max} - \pi'_{\mathcal{A}} = (\pi_{\max} - \pi_{\min})(1 - \pi_{\mathcal{A}}) \geq 0$ and $\pi'_{\mathcal{A}} - \pi_{\min} = (\pi_{\max} - \pi_{\min})\pi_{\mathcal{A}} \geq 0$, it follows that $\pi'_{\mathcal{A}} \in [\pi_{\min}, \pi_{\max}]$. Thus,by Theorem 8 ,the power constraint is met.

Regret with respect to a clipped oracle: Under the wrapper algorithm, A_t is transformed by the stochastic mapping G and the potential rewards can be written as $R'_t(a) = R_t(G(a))$ for $a \in \{0, 1\}$. And by assumption there is an environment in Ω with these rewards. Further algorithm $\mathcal{A}$ has regret rate no greater than $\mathcal{R}(T)$ with respect to an oracle $\mathcal{O}$ on the original environment. The expected reward of an oracle on the new environment is the same as the expected reward of the wrapper

algorithm applied to the oracle on the original environment, i.e. $\mathbb{E}[R_t(\mathcal{O}')] = \mathbb{E}[R_t(G(\mathcal{O})]$. Thus, we can equivalently state that the algorithm resulting from transforming A_t by G has expected regret bound $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$. □

For sure, we should ask what collections of environments Ω are closed under the reward transformation above. In the following, we characterize properties of Ω satisfying Theorem 9.

Corollary 3. *For a stochastic contextual bandit, the following environment class has the closure property assumed by Theorem 9 under the action-transforming operation G - that is, for all environments in Ω, the potential rewards $\{R_t(1), R_t(0)\}$ transforms to $\{R_t(G(1)), R_t(G(0)\}$, which are still in Ω:*

1. *$R_t(a) \leq L$, where L is a constant.*
2. *$R_t - \mathbb{E}[R_t|A_t, C_t]$ is σ-sub-Gaussian*

Proof. Condition 1. above clearly holds for $R_t(G(a))$ as $G(a) \in \{0, 1\}$. Now, under the stochastic mapping G on actions, the new reward is

$$\begin{aligned} R'_t = R_t(G(A_t)) =&[A_t G(1) + (1 - A_t)G(0)]R_t(1) \\ &+ [A_t(1 - G(1)) + (1 - A_t)(1 - G(0))]R_t(0) \end{aligned}$$

and the new reward function is given by:

$$\begin{aligned} \mathbb{E}[R'_t|C_t, A_t] =&[A_t\pi_{\max} + (1 - A_t)\pi_{\min}]\mathbb{E}[R_t(1)|C_t] \\ &+ [A_t(1 - \pi_{\max}) + (1 - A_t)(1 - \pi_{\min})]\mathbb{E}[R_t(0)|C_t]. \end{aligned}$$

Since $A_t, G(0), G(1)$ are binary, and the set of sub-Gaussian random variables is closed under finite summation, Condition 2. still holds albeit with a different constant σ. □

Next, we discuss how Corollary 3 applies to a set of common algorithms. In the derivations of regret bounds for these algorithms, in addition to the environmental assumptions outlined in Corollary 3, each derivation makes further assumptions on the environment. We discuss how each set of assumptions is preserved under the closure operation defined by our stochastic transformation G.

Remark 3. *LinUCB [Abbasi-Yadkori et al., 2011], SupLinUCB [Chu et al., 2011], SupLinREL [Auer, 2002] and TS [Agrawal and Goyal, 2012] further assume that the reward takes the form of $\mathbb{E}[R_t(a)|C_{t,a}] = C_{t,a}^\intercal\theta$ (Note θ here is different from θ in Section 4 in main paper). They assume that $\|\theta\| \leq S_1$, $\|C_{t,a}\| \leq S_2$. Under G,*

$$\mathbb{E}[R_t(G(a))|C_t] = \pi_{\min}^a\pi_{\max}^{1-a}C_{t,0}^\intercal\theta + \pi_{\min}^{1-a}\pi_{\max}^a C_{t,1}^\intercal\theta.$$

$\{\theta, \pi_{\min}^a\pi_{\max}^{1-a}C_{t,0} + \pi_{\min}^{1-a}\pi_{\max}^a C_{t,1}\}$ are still bounded but may with different constant.

Differently, ϵ-greedy [Langford and Zhang, 2007] assumes the learner is given a set of hypothesis $\mathcal{H}$ where each hypothesis h maps a context C_t to an action A_t. The goal is to choose arms to compete with the best hypothesis in $\mathcal{H}$. It assumes that there is a distribution on $P \sim (C_t, R_t)$, which remains true but now with a different distribution $P' \sim (C'_t, R'_t)$ under G. Langford and Zhang derived the regret bounds when the hypothesis space is finite $|\mathcal{H}| = m$ with an unknown expected reward gap. Let $R(h)$ be the expected total reward under hypothesis h and $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$. Without loss of generality, assume $R(h_1) \geq R(h_2) \geq \ldots R(h_m)$ and $R(h_1) \geq R(h_2) + \triangle$ where $\triangle$ is the unknown expected reward gap, $\triangle > 0$. Under G, the hypothesis space needs to change accordingly to $\mathcal{H}'$ where each new hypothesis h' maps the new context $C'_{t,a} = \pi_{\min}^a\pi_{\max}^{1-a}C_{t,0} + \pi_{\min}^{1-a}\pi_{\max}^a C_{t,1}$ to a new action A'_t; however, the hypothesis space size remains the same, $|\mathcal{H}'| = m$. And without loss of generality, we can reorder $R(h')$ so that $R(h'_1) \geq R(h'_2)$, thus the environment is closed under G.

A.7 Action Flipping Wrapper Algorithm in MDP Setting

In this sectoin, we prove that our action flipping strategy can also be applied to an MDP setting since our test statistic allows the features to depend on the full history. We again drop n for convenience. A MDP M is defined with a set of finite states $\mathcal{S}$ and a set of finite actions A. An environment for an MDP is defined by the initial state distribution $S_0 \sim P_0$, the transition probability $P_{s,s'}^a$ and the reward which is a function of current state, action and next state, $R_t = r(S_t, A_t, S_{t+1})$.

Again we use potential outcome notation; this notation is coherent with the standard MDP notation and allows us to make the role of the stochastic transformation, G, clear. At time t, given the current state S_t, the algorithm selects the action a and transits to the next state S_t with transition probability $P^a_{s,s'} = P(S_{t+1} = s' | S_t = s, A_t = a)$. The observed reward is $R_t(A_t)$ and the expected reward given a state-action pair is $\mathbb{E}[R_t(A_t)|S_t = s, A_t = a] = \sum_{s'} P^a_{s,s'} r(s, a, s')$.

Recall that the set of environments is denoted by Ω. At state S_t, an algorithm $\mathcal{A}$ maps the history for each user up to time t: $H_t = (\{S_j, A_j, R_j\}_{j=t}^{t-1}, S_t)$ to a probability distribution over action space A. As before the wrapper algorithm makes the input algorithm $\mathcal{A}$ believe that it is in an environment more stochastic than it truly is (particularly the distribution of S_{t+1} is more stochastic). Intuitively, if algorithm $\mathcal{A}$ is capable of achieving some rate in this more stochastic environment, then it will be optimal with respect to the clipped oracle.

Corollary 4. *Given $\pi_{\min},\ \pi_{\max}$ and an MDP algorithm $\mathcal{A}$, assume that algorithm $\mathcal{A}$ has an expected regret $\mathcal{R}(T)$ for any MDP environment in Ω, with respect to an oracle $\mathcal{O}$. Under stochastic transformation G, if there exists an environment in Ω that contains the new transition probability function: $P'^a_{s,s'} = \left(\pi^a_{\min}\pi^{1-a}_{\max}P^0_{s,s'} + \pi^{1-a}_{\min}\pi^a_{\max}P^1_{s,s'}\right)$ then the wrapper algorithm will (1) return a data set that satisfies the desired power constraints and (2) have expected regret no larger than $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$.*

Proof. The proof of satisfaction of power constraints follows as in Theorem 9.

Regarding regret: Under the wrapper algorithm, the action A_t is transformed by the stochastic mapping G, which only impacts the next state S_{t+1}. The new transition probability function $P'^a_{s,s'}$ can be written as $\left(\pi^a_{\min}\pi^{1-a}_{\max}P^0_{s,s'} + \pi^{1-a}_{\min}\pi^a_{\max}P^1_{s,s'}\right)$. And by assumption there is an environment in Ω with this probability transition function. Recall that the reward is a deterministic function of the current state, the action and the next state. Further recall that $\mathcal{A}$ has regret rate no greater than $\mathcal{R}(T)$ with respect to an oracle $\mathcal{O}$ on the original environment. Thus the expected reward of an oracle on this environment is the same as the expected reward of the wrapper algorithm applied to the oracle on the original environment, i.e. $\mathbb{E}[R_t(G(\mathcal{O}))|S_t, A_t] = \mathbb{E}[R_t(\mathcal{O}')|S_t, A_t]$. Thus, we can equivalently state that the algorithm resulting from transforming A_t by G has expected regret bound $\mathcal{R}(T)$ with respect to a clipped oracle $\mathcal{O}'$. □

A.8 Data-Dropping Power-Preserving Wrapper Algorithm

In this section, we give full analyses of the data-dropping wrapper algorithm which can also be used for power preserving purpose. The algorithm implementation is given in Algorithm 2. The wrapper takes as input a contextual bandit algorithm $\mathcal{A}$ and pre-computed $\pi_{\min},\ \pi_{\max}$ $(\pi_{\max} + \pi_{\min} = 1)$ computed from Theorem 7. The input algorithm $\mathcal{A}$ can be stochastic or deterministic. Conceptually, our wrapper operates as follows: for a given context, if the input algorithm $\mathcal{A}$ returns a probability distribution over choices that already satisfies $\pi_{\mathcal{A}} \in [\pi_{\min}, \pi_{\max}]$, then we sample the action according to $\pi_{\mathcal{A}}$. However, if the maximum probability of an action exceeds $\pi_{\max}$, then we sample that action according to $\pi_{\max}$.

The key to guaranteeing good regret with this wrapper for a broad range of input algorithms $\mathcal{A}$ is in deciding what information we share with the algorithm. Specifically, the sampling approach in lines 9-22 determines whether the action that was ultimately taken would have been taken absent the wrapper; the context-action-reward tuple from that action is only shared with the input algorithm $\mathcal{A}$ if $\mathcal{A}$ would have also made that same decision. This process ensures that the input algorithm $\mathcal{A}$ only sees samples that match the data it would observe if *it* was making all decisions.

Now, suppose that the input algorithm $\mathcal{A}$ was able to achieve some regret bound $\mathcal{R}(T)$ with respect to some setting $\mathcal{B}$ (which, as noted before, may be more specific than that in Section 3 in main paper). The wrapped version of input $\mathcal{A}$ by Algorithm 2 will achieve the desired power bound by design; but what will be the impact on the regret? We prove that as long as the setting $\mathcal{B}$ allows for data to be dropped, then an algorithm that incurs $\mathcal{R}$ regret in its original setting suffers at most $(1 - \pi_{\max})$ linear regret in the clipped setting. Specifically, if an algorithm $\mathcal{A}$ achieves an optimal rate $O(\sqrt{T})$ rate with respect to a standard oracle, its clipped version will achieve that optimal rate with respect to the clipped oracle.

Algorithm 2 Data-Dropping Power-Preserving Wrapper Algorithm

Input: $\pi_{\min}$, $\pi_{\max}$, Algorithm $\mathcal{A}$
for $t = 1, 2, \ldots$ **do**
  Observe context C_t and outputs $\pi_{\mathcal{A}}(C_t)$ each action
  if $\pi_{\max} \leq_a \{\pi_{\mathcal{A}}(a)\} \leq \pi_{\max}$ **then**
    $A_t \sim \pi_{\mathcal{A}}$ {Sample action}
    Observe R_t
    Update Algorithm $\mathcal{A}$ with (C_t, A_t, R_t)
  else
    $u \sim \texttt{unif}(0, 1)$
    $A_t^* = \arg\max_a \pi_{\mathcal{A}}(a)$
    if $u \leq \pi_{\max}$ or $u > \max_a\{\pi_{\mathcal{A}}(a)\}$ **then**
      if $u \leq \pi_{\max}$ **then**
        $A_t = A_t^*$
      else
        $A_t = \arg\min\{\pi_{\mathcal{A}}(a)\}$
      end if
      Observe R_t
      Update Algorithm $\mathcal{A}$ with (C_t, A_t, R_t) {Both approaches agree on action}
    else
      $A_t = \arg\min\{\pi_{\mathcal{A}}(a)\}$
      Observe R_t {Do not give data to $\mathcal{A}$}
    end if
  end if
end for

Theorem 10. *Assume as input $\pi_{\max}$ and a contextual bandit algorithm $\mathcal{A}$. Assume algorithm $\mathcal{A}$ has a regret bound $\mathcal{R}(T)$ under one of the following assumptions on the setting $\mathcal{B}$: (1) $\mathcal{B}$ assumes that the data generating process for each context is independent of history, or (2) $\mathcal{B}$ assumes that the context depends on the history, and the bound $\mathcal{R}$ for algorithm $\mathcal{A}$ is robust to an adversarial choice of context.*

Then our wrapper Algorithm 2 will (1) return a dataset that satisfies the desired power constraints and (2) has expected regret no larger than $\mathcal{R}(\pi_{\max}T) + (1 - \pi_{\max})T$ if assumptions $\mathcal{B}$ are satisfied in the true environment.

Proof. **Satisfaction of power constraints:** By construction our wrapper algorithm ensures that the selected actions always satisfy the required power constraints.

Regret with respect to a clipped oracle: Note that in the worst case, the input algorithm $\mathcal{A}$ deterministically selects actions A_t, which are discarded with probability $1 - \pi_{\max}$. Therefore if running in an environment satisfying the assumptions $\mathcal{B}$ of the input algorithm $\mathcal{A}$, our wrapper could suffer at most linear regret on $T(1-\pi_{\max})$ points, and will incur the same regret as the algorithm $\mathcal{A}$ on the other points (which will appear to algorithm $\mathcal{A}$ as if these are the only points it has experienced).

Note that since the wrapper algorithm does not provide all observed tuples to algorithm $\mathcal{A}$, this proof only works for assumptions $\mathcal{B}$ on the data generating process that assumes the contexts are independent of history, or in a setting in which $\mathcal{A}$ is robust to adversarially chosen contexts. □

Essentially this result shows that one can get robust power guarantees while incurring a small linear loss in regret (recall that $\pi_{\max}$ will tend toward 1, and $\pi_{\min}$ toward 0, as T gets large) if the setting affords additional structure commonly assumed in stochastic contextual bandit settings. Because our wrapper is agnostic to the choice of input algorithm $\mathcal{A}$, up to these commonly assumed structures, we enable a designer to continue to use their favorite algorithm—perhaps one that has seemed to work well empirically in the domain of interest—and still get guarantees on the power.

Corollary 5. *For algorithms $\mathcal{A}$ that satisfy the assumptions of Theorem 10, our wrapper algorithm will incur regret no worse than $O(\mathcal{R}(\pi_{max}T))$ with respect to a clipped oracle.*

Proof. Recall that a clipped oracle policy takes the optimal action with probability $\pi_{\max}$ and the other action with probability $1 - \pi_{\max}$. By definition, any clipped oracle will suffer a regret of $(1 - \pi_{\max})T$. Therefore relative to a clipped oracle, our wrapper algorithm will have an regret rate $O(\mathcal{R}(\pi_{\max}T))$ that matches the regret rate of the algorithm in its assumed setting when the true environment satisfies those assumptions. This holds for algorithms $\mathcal{A}$ satisfying the assumptions of Theorem 10. □

B Descriptions of Algorithms

Below, we provide pseudocode of all the algorithms we used for reference. All the algorithms listed below is for each user n and we drop subscript n for simplicity.

B.1 Fixed Randomization with $\pi = 0.5$

Algorithm 3 Fixed Randomization with $\pi = 0.5$

for $t = 1, 2, \cdots, T$ **do**
  $A_t \sim \texttt{Bern}(0.5)$
  Observe R_t
end for

B.2 ACTS

Algorithm 4 Clipped ACTS(Action Centered Thompson Sampling)

Input: $\sigma^2,\ \pi_{\min},\ \pi_{\max}$
$b = 0, V = I, \hat{\delta} = V^{-1}b, \hat{\Sigma} = \sigma^2 V^{-1}$
for $t = 1, 2, \cdots, T$ **do**
  Observe C_t
  if $1 - \phi_{C_t^\intercal \hat{\delta}, C_t^\intercal \hat{\Sigma} C_t}(0) < \pi_{\min}$ **then**
    $\pi_t = \pi_{\min}$
    $A_t \sim \texttt{Bern}(\pi_t)$
  else if $1 - \phi_{C_t^\intercal \hat{\delta}, C_t^\intercal \hat{\Sigma} C_t}(0) > \pi_{\max}$ **then**
    $\pi_t = \pi_{\max}$
    $A_t \sim \texttt{Bern}(\pi_t)$
  else
    $\tilde{\delta} \sim \mathcal{N}(C_t^\intercal \hat{\delta}, C_t^\intercal \hat{\Sigma} C_t)$
    $A_t = \arg\max(0, C_t^\intercal \tilde{\delta})$
    $\pi_t = 1 - \phi_{C_t^\intercal \hat{\delta}, C_t^\intercal \hat{\Sigma} C_t}(0)$
  end if
  Observe R_t
  Update $V = V + (1 - \pi_t)\pi_t C_t C_t^\intercal, b = b + (A_t - \pi_t)R_t C_t, \hat{\delta} = V^{-1}b$
end for

B.3 BOSE

Algorithm 5 Clipped BOSE (Bandit Orthogonalized Semiparametric Estimation)

Input: $\pi_{\text{min}},\ \pi_{\text{max}},\ \eta$
$b = 0, V = I, \hat{\delta} = V^{-1}b$
for $t = 1, 2, \cdots, T$ **do**
  Observe C_t
  if $C_t^\intercal \hat{\delta} > \eta C_t^\intercal V^{-1} C_t$ **then**
    $\pi_t = \pi_{\max}$
  else if $-C_t^\intercal \hat{\delta} > \eta C_t^\intercal V^{-1} C_t$ **then**
    $\pi_t = \pi_{\min}$
  else
    $\pi_t = 0.5$
  end if
  $A_t \sim \texttt{Bern}(\pi_t)$ and observe R_t
  Update $V = V + (A_t - \pi_t)^2 C_t C_t^\intercal, b = b + (A_t - \pi_t)R_t C_t, \hat{\delta} = V^{-1}b$
end for

B.4 linUCB

Algorithm 6 linUCB(linear Upper Confidence Bound)

Input: $\pi_{\text{min}},\ \pi_{\text{max}},\ \eta$
$b = 0, V = I, \hat{\theta} = V^{-1}b$
for $t = 1, 2, \cdots, T$ **do**
  Observe $C_t(0),\ C_t(1)$
  For $a \in \{0, 1\}$, compute $\mathcal{L}_{t,a} = C_t(a)\hat{\theta} + \eta\sqrt{C_t^\intercal(a)V^{-1}C_t(a)}$
  $A_t^* = \arg\max_a \mathcal{L}_{t,a}$
  $\pi_t = \pi_{\text{min}}^{1-A_t^*}\pi_{\text{max}}^{A_t^*}$
  $A_t \sim \texttt{Bern}(\pi_t)$ and observe R_t
  Update $V = V + C_tC_t^\intercal, b = b + C_tR_t, \hat{\theta} = V^{-1}b$
end for

C Environments

C.1 Semi-parametric Contextual Bandit(SCB)

In this environment, for each user n, at each round, a feature Z_{nt} is independently drawn from a sphere with norm 0.4. Specifically:

$$\begin{aligned} R_{nt}(A_{nt}) &= \gamma_{nt} + A_{nt}Z_{nt}^\intercal\delta + \mathcal{N}(0, \sigma^2) \\ \delta &= [0.382, -0.100, 0.065] \quad (\|\delta\| = 0.4) \\ \gamma_{nt} &= \frac{1}{900}t - 0.05 \\ \sigma^2 &= 0.25 \\ A_{nt} &\in \{0, 1\} \end{aligned}$$

C.2 Adversarial Semi-parametric Contextual Bandit(ASCB)

The adversarial semi-parametric contextual bandit is similar to SCB except that in each round, γ_{nt} is chosen by an adaptive adversary. We specifically used the adversary introduced in [Abbasi-Yadkori et al., 2018]. The environment is defined as follows:

$$\begin{aligned} R_{nt}(A_{nt}) &= \gamma_{nt} + A_{nt}Z_{nt}^\intercal\delta + \mathcal{N}(0, \sigma^2) \\ Z_{nt} &= [-0.5, 0.3 \cdot (-1)^t, (t/100)^2] \\ \delta &= [0.2, 0.2, 0.2] \\ \gamma_{nt} &= -\max(0, A_{nt}Z_{nt}^\intercal\delta) \\ \sigma^2 &= 0.25 \\ A_{nt} &\in \{0, 1\} \end{aligned}$$

C.3 Mobile Health Simulator

The mobile health simulator, which mimics the data generation process of a mobile application to increase users' physical activities, was originally developed in [Liao et al., 2016]. In this environment, the effect changes over time but is still independent across days. The noise terms are correlated and follows Gaussian AR(1) process. The response to the binary action $A_t \in \{0, 1\}$ is R_t, which is

interpreted as $\sqrt{\text{Step count on day } t}$.

$$\begin{aligned} R_{nt} &= A_{nt} Z_{nt}^T \delta + \alpha(t) + \frac{\sigma(t)}{\sqrt{2}} \epsilon_{nt} \\ \epsilon_{nt} &= \phi \epsilon_{n,t-1} + e_{nt} \\ e_{nt} &\sim \mathcal{N}(0,1) \\ \epsilon_0 &\sim \mathcal{N}\left(0, \frac{1}{1-\phi^2}\right) \\ A_{nt} &\in \{0,1\} \end{aligned}$$

Note $Var(\epsilon_{nt}) = \frac{1}{1-\phi^2}$ for all t. One can choose $\phi = 1/\sqrt{2}$. The features are

$$Z_{nt} = [1, \frac{t-1}{45}, \left(\frac{t-1}{45}\right)^2]^\intercal \tag{18}$$

The $\alpha(t)$ represents the square root of the step-count under no action $A_t = 0$. Let $\alpha(t)$ vary linearly from 125 at $t = 0$ to 50 at $t = T$. The $\sigma^2(t)$ is the residual variance in step count. We set $\sigma(t) = 30$. For δ_0, under null hypothesis, $\delta_0 = \mathbf{0}$. Under alternate hypothesis, $\delta^{(0)} = 6$. There is no effect at $T = 90$ and peak effect at $T = 21$. By solving the system, we have $\delta^\intercal = [6.00\ , -2.48\ , -2.79]$, $\bar{\delta} = \sum_{t=1}^{T} Z^T \delta \approx 1.6$.

C.4 Environmental Set-up for Type 1 error Experiment

For all environments, to verify Type 1 error is recovered, during simulation, we set $\delta_0 = \mathbf{0}$ where $\mathbf{0}$ is a zero vector. When solving for $\pi_{\text{min}},\ \pi_{\text{max}}$, we used δ values specified in the above sections.

C.5 Environmental Set-up for Robustness Test of Treatment Effect Estimation

To study the impact of the estimated effect size, we tested two different types of mis-estimation: underestimation and overestimation of the average treatment effect. For the experiment purpose, the guessed size of each dimension d is set as $\delta_{est}^{(d)} = \delta^{(d)}/1.1$ (underestimation) and $\delta_{est}^{(d)} = \delta^{(d)} \times 1.1$ (overestimation) while the effect size of the simulation environment remains as $\delta_0 = \delta$.

D Experiment Settings

For all environments, we use $N = 20$ subjects and $T = 90$ trajectory length. In the regret minimization algorithm, C_{nt} is set as Z_{nt}.

D.1 Identifying optimal hyperparameters

For all algorithms, the hyperparameters are chose by maximizing the average return over $1,000$ individuals. The prior of the ACTS algorithm is set as $\mathcal{N}(0, \sigma_0^2)$ and σ_0^2 is chosen between $[0.05, 0.5]$ for SCB and ASCB, and between $[50, 150]$ for the mobile health simulator. The parameter η of BOSE is chosen between $[0.1, 2.0]$ for SCB and ASCB, and between $[10, 150]$ for the mobile health simulator. The hyperparameter η of linUCB is chosen between $[0.01, 0.25]$ for SCB and ASCB, and between $[10, 100]$ for the mobile health simulator. (Note: in reality, we would not be able to repeatedly run experiments of 1000 individuals to find the optimal hyperparameters; we do this to give the baseline versions of the algorithms their best chance for success.) The optimal hyperparameters, that is, those that minimize empirical regret, are listed below:

Table 1: Optimal hyperparameter chosen for a given pair of an algorithm and an environment

	SCB	ASCB	Mobile Health Simulator
ACTS	0.15	0.05	60
BOSE	0.2	0.2	120
linUCB	0.03	0.02	95

D.2 Solved π_{min}, π_{max}

Table 2 lists solved π values given a pair of an environment and a guessed effect size. We see the smaller in magnitude of δ_{est}, the closer $\pi_{\min}, \pi_{\max}$ are to 0.5, which results in more exploration. The larger in magnitude of δ_{est}, the further $\pi_{\min}, \pi_{\max}$ are from 0.5 exploration, which results in less exploration.

Table 2: Solved π_{min},π_{max} given a pair of an environment and a guessed effect size

	$\delta_{est} < \delta$		$\delta_{est} = \delta$		$\delta_{est} > \delta$	
	$\pi_{\min}$	$\pi_{\max}$	$\pi_{\min}$	$\pi_{\max}$	$\pi_{\min}$	$\pi_{\max}$
SCB	0.288	0.712	0.216	0.784	0.168	0.882
ASCB	0.301	0.699	0.225	0.775	0.174	0.826
Mobile Health Simulator	0.335	0.665	0.243	0.757	0.187	0.813

E Additional Results

E.1 Type 1 Error

Table 3: Type 1 error with 2 standard error($\hat{\alpha}_0 \pm 2\sqrt{\hat{\alpha}_0(1-\hat{\alpha}_0)/N}$ where $N = 1000$). We see some Type 1 errors are close to $\alpha_0 = 0.05$ while some are larger than 0.05 but not significantly.

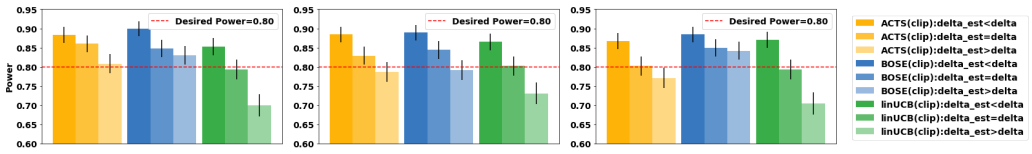
SCB					
Fix $\pi = 0.5$	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.060 ± 0.015	0.074 ± 0.017	0.054 ± 0.014	0.068 ± 0.016	0.060 ± 0.015	0.060 ± 0.015
ASCB					
Fix $\pi = 0.5$	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.065 ± 0.016	0.054 ± 0.014	0.052 ± 0.014	0.090 ± 0.018	0.054 ± 0.014	0.062 ± 0.015
Mobile Health Simulator					
Fix $\pi = 0.5$	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.055 ± 0.014	0.058 ± 0.016	0.049 ± 0.014	0.062 ± 0.015	0.069 ± 0.016	0.052 ± 0.014

As Table 3 suggests, when there is no treatment effect, we recover the correct Type 1 error. We see some Type 1 errors are close to $\alpha_0 = 0.05$ while some are larger than 0.05 but not significantly. This is likely caused by small sample variation in the estimated covariance matrix. BOSE, especifically, drops a large portion of data on which it is certain about the treatment effect. In practice, small sample corrections are used to adjust the estimated covariance matrix [Liao et al., 2016].

E.2 Power

Table 4 shows that when there is a treatment effect, we recover the correct power if we guessed the effect size correctly.

E.3 Mis-estimation of Effect Size and Mis-specification of the Reward Model

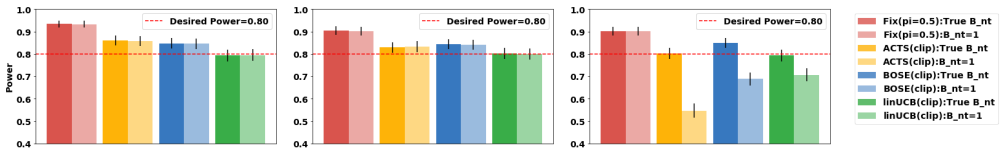


(a) SCB (b) ASCB (c) Mobile Health

Figure 3: Effect of mis-estimated treatment effect on power: In general, when $Z_t\delta_{est} < Z_t\delta_0$, power is higher and when $Z_t\delta_{est} > Z_t\delta_0$, power is lower. ACTS and BOSE are more robust to effect mis-specification.

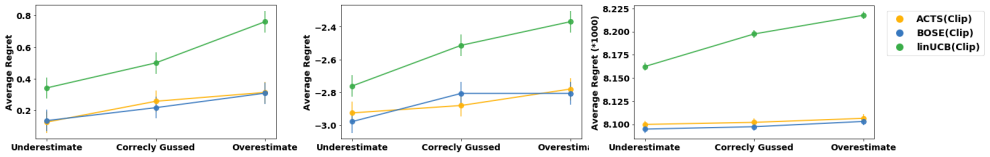
Table 4: Resulting power with 2 standard error ($\hat{\beta}_0 \pm 2\sqrt{\hat{\beta}_0(1-\hat{\beta}_0)/N}$ where $N = 1000$). With probability clipping, the correct power $\beta_0 = 0.80$ is recovered while without clipping, sufficient power is not guaranteed.

SCB					
Fix $\pi = 0.5$	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.934 ± 0.016	0.442 ± 0.031	0.860 ± 0.022	0.643 ± 0.030	0.848 ± 0.023	0.793 ± 0.026
ASCB					
Fix π=0.5	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.905 ± 0.019	0.299 ± 0.029	0.829 ± 0.024	0.321 ± 0.030	0.844 ± 0.023	0.802 ± 0.025
Mobile Health Simulator					
Fix $\pi = 0.5$	ACTS	ACTS (clip)	BOSE	BOSE (clip)	linUCB(clip)
0.902 ± 0.019	0.547 ± 0.031	0.802 ± 0.025	0.667 ± 0.030	0.849 ± 0.023	0.793 ± 0.026



(a) SCB (b) ASCB (c) Mobile Health

Figure 4: Effect of wrong reward model on power: Powers is robust to reward model mis-specification in SCB and ASCB where the bar heights are similar. In mobile health, the highest drop is with ACTS which is less than 0.3.



(a) SCB (b) ASCB (c) Mobile Health

Figure 5: Effect of mis-estimated treatment effect on average return: In general, when $Z_t\delta_{est} < Z_t\delta_0$, average return is lower and when $Z_t\delta_{est} > Z_t\delta_0$, average return is higher.

Combining Figure 3 and Figure 5, we see overestimation of the treatment effect results in higher average return but lower power while underestimation results in lower average return but higher power. Figure 4 shows that the resulting power is fairly robust to model mis-specification in SCB and ASCB. Although in mobile health, the resulting power drops, the largest amount of decrease does not exceed 0.3 (ACTS).

E.4 Average Return & Regret

Table 5 lists the complete results of AR, *reg* and *reg*$_c$. We compute reg as

$$\text{reg}_\text{c} = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{nt} + \max(0, Z_{nt}^\intercal \delta_0) + \epsilon_{nt}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{nt}\right]$$

and reg$_\text{c}$ as

$$\text{reg}_\text{c} = \mathbb{E}\left[\sum_{t=1}^{T} \gamma_{nt} + \pi_{nt}^* Z_{nt}^\intercal \delta_0 + \epsilon_{nt}\right] - \mathbb{E}\left[\sum_{t=1}^{T} R_{nt}\right]$$

where $\pi_{nt}^* = \text{argmax}_{\pi_{nt} \in [\pi_{\min}, \pi_{\max}]} \pi_{nt} Z_{nt}^\intercal \delta_0$.

The regret metrics allow us to see how the returns of our adapted algorithms compare against the best possible rate we could achieve (against the clipped oracle); they also highlight the cost of clipping.

Table 5: Average Return, *reg*, reg_c with 2 standard error for $N = 1000 \times 20$ simulated users: For all algorithms, clipping decreases the average return. Different algorithms have different regrets, but the difference between the regret (reg) and the clipped regret (reg_c) are relatively constant across algorithms within each environment.

SCB							
	Fix $\pi = 0.5$	ACTS	ACTS(clip)	BOSE	BOSE(clip)	linUCB	linUCB(clip)
AR	-0.172 ± 0.068	0.419 ± 0.069	0.256 ± 0.068	0.477 ± 0.070	0.216 ± 0.068	1.243 ± 0.069	0.500 ± 0.068
reg	3.509 ± 0.068	2.918 ± 0.069	3.080 ± 0.068	2.859 ± 0.070	3.121 ± 0.068	2.093 ± 0.069	2.837 ± 0.068
reg_c	2.788 ± 0.068	2.197 ± 0.069	2.360 ± 0.068	2.139 ± 0.070	2.401 ± 0.068	1.373 ± 0.069	2.117 ± 0.068
ASCB							
	Fix $\pi = 0.5$	ACTS	ACTS(clip)	BOSE	BOSE(clip)	linUCB	linUCB(clip)
AR	-3.245 ± 0.068	-2.697 ± 0.069	-2.882 ± 0.068	-2.743 ± 0.072	-2.808 ± 0.068	-1.655 ± 0.066	-2.514 ± 0.066
reg	3.245 ± 0.068	2.697 ± 0.069	2.882 ± 0.068	2.743 ± 0.072	2.808 ± 0.068	1.655 ± 0.066	2.514 ± 0.066
reg_c	2.983 ± 0.068	2.435 ± 0.069	2.620 ± 0.068	2.481 ± 0.072	2.546 ± 0.068	1.394 ± 0.066	2.252 ± 0.066
Mobile Health Simulator							
	Fix $\pi = 0.5$	ACTS	ACTS(clip)	BOSE	BOSE(clip)	linUCB	linUCB(clip)
AR($\times 10^3$)	8.095 ± 0.004	8.106 ± 0.004	8.102 ± 0.004	8.106 ± 0.004	8.097 ± 0.004	8.295 ± 0.004	8.197 ± 0.004
reg($\times 10^3$)	0.218 ± 0.004	0.206 ± 0.004	0.210 ± 0.004	0.207 ± 0.004	0.215 ± 0.004	0.017 ± 0.004	0.115 ± 0.004
reg_c($\times 10^3$)	0.111 ± 0.004	0.100 ± 0.004	0.104 ± 0.004	0.100 ± 0.004	0.109 ± 0.004	-0.089 ± 0.004	0.009 ± 0.004

Table 5 shows that different algorithms have different regrets, but all still converge as expected with respect to the clipped oracle.

E.5 Comparison of Wrapper Algorithms

The full results of action-flipping/ data-dropping/ probability-clipping wrapper algorithms are listed in Table 6 and Table 7.

Table 6: Resulting power with 2 standard error ($\beta_0 \pm 2\sqrt{\hat{\beta}_0(1-\hat{\beta}_0)/N}$ where $N = 1000$): All wrapper algorithms satisfied the power constraints.

SCB			
ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
0.442 ± 0.031	0.908 ± 0.018	0.826 ± 0.024	0.860 ± 0.022
BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
0.643 ± 0.030	0.886 ± 0.020	0.863 ± 0.022	0.848 ± 0.023
linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
-	0.810 ± 0.025	0.815 ± 0.025	0.793 ± 0.026
ASCB			
ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
0.299 ± 0.029	0.859 ± 0.022	0.844 ± 0.023	0.829 ± 0.024
BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
0.321 ± 0.030	0.847 ± 0.023	0.836 ± 0.023	0.844 ± 0.023
linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
-	0.797 ± 0.025	0.794 ± 0.026	0.802 ± 0.025
Mobile Health Simulator			
ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
0.547 ± 0.031	0.851 ± 0.023	0.787 ± 0.026	0.802 ± 0.025
BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
0.667 ± 0.030	0.830 ± 0.024	0.813 ± 0.025	0.849 ± 0.023
linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
-	0.778 ± 0.026	0.817 ± 0.024	0.793 ± 0.026

Table 7: Average Return, *reg*,*reg*$_c$ with 2 standard error for $N = 1000 \times 20$ simulated users:All wrapper algorithms achieve good regret rate with slightly different trade-offs given the situation.

SCB				
	ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
AR	0.419 ± 0.069	0.030 ± 0.068	0.282 ± 0.067	0.256 ± 0.068
reg	2.918 ± 0.069	3.307 ± 0.068	3.054 ± 0.067	3.080 ± 0.068
reg$_c$	2.197 ± 0.069	2.586 ± 0.068	2.334 ± 0.067	2.360 ± 0.068
	BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
AR	0.477 ± 0.070	0.068 ± 0.067	0.179 ± 0.069	0.216 ± 0.068
reg	2.859 ± 0.070	3.269 ± 0.067	3.158 ± 0.069	3.121 ± 0.068
reg$_c$	2.139 ± 0.070	2.548 ± 0.067	2.437 ± 0.069	2.401 ± 0.068
	linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
AR	1.243 ± 0.069	0.455 ± 0.068	0.585 ± 0.068	0.500 ± 0.068
reg	2.093 ± 0.069	2.882 ± 0.068	2.752 ± 0.068	2.837 ± 0.068
reg$_c$	1.373 ± 0.069	2.161 ± 0.068	2.031 ± 0.068	2.117 ± 0.068
Adversarial Semi-parametric Contextual Bandit				
	ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
AR	-2.697 ± 0.069	-2.947 ± 0.067	-2.795 ± 0.068	-2.882 ± 0.068
reg	2.697 ± 0.069	2.947 ± 0.067	2.795 ± 0.068	2.882 ± 0.068
reg$_c$	2.435 ± 0.069	2.686 ± 0.067	2.533 ± 0.068	2.620 ± 0.068
	BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
AR	-2.743 ± 0.072	-2.960 ± 0.067	-2.876 ± 0.068	-2.808 ± 0.068
reg	2.743 ± 0.072	2.960 ± 0.067	2.876 ± 0.068	2.808 ± 0.068
reg$_c$	2.481 ± 0.072	2.698 ± 0.067	2.614 ± 0.068	2.546 ± 0.068
	linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
AR	-1.655 ± 0.066	-2.401 ± 0.066	-2.366 ± 0.067	-2.514 ± 0.066
reg	1.655 ± 0.066	2.401 ± 0.066	2.366 ± 0.067	2.514 ± 0.066
reg$_c$	1.394 ± 0.066	2.139 ± 0.066	2.104 ± 0.067	2.252 ± 0.066
Mobile Health Simulator				
	ACTS	ACTS (flip)	ACTS (drop)	ACTS (clip)
AR($\times 10^3$)	8.106 ± 0.004	8.172 ± 0.004	8.106 ± 0.004	8.102 ± 0.004
reg($\times 10^3$)	0.206 ± 0.004	0.140 ± 0.004	0.206 ± 0.004	0.210 ± 0.004
reg$_c$($\times 10^3$)	0.100 ± 0.004	0.034 ± 0.004	0.100 ± 0.004	0.104 ± 0.004
	BOSE	BOSE (flip)	BOSE (drop)	BOSE (clip)
AR($\times 10^3$)	8.106 ± 0.004	8.159 ± 0.004	8.103 ± 0.004	8.097 ± 0.004
reg($\times 10^3$)	0.207 ± 0.004	0.154 ± 0.004	0.210 ± 0.004	0.215 ± 0.004
reg$_c$($\times 10^3$)	0.100 ± 0.004	0.047 ± 0.004	0.103 ± 0.004	0.109 ± 0.004
	linUCB	linUCB(flip)	linUCB(drop)	linUCB(clip)
AR($\times 10^3$)	8.295 ± 0.004	8.198 ± 0.004	8.192 ± 0.004	8.197 ± 0.004
reg($\times 10^3$)	0.017 ± 0.004	0.114 ± 0.004	0.120 ± 0.004	0.115 ± 0.004
reg$_c$($\times 10^3$)	-0.089 ± 0.004	0.008 ± 0.004	0.014 ± 0.004	0.009 ± 0.004