Assessing Dataset Quality using Optimal Experimental Design for Linear Contextual Bandits

Matthew Jörke
Department of Computer Science
Stanford University
joerke@stanford.edu

Tong Mu
Department of Electrical Engineering
Stanford University
tongm@stanford.edu

Jonathan Lee
Department of Computer Science
Stanford University
jnl@stanford.edu

Emma Brunskill
Department of Computer Science
Stanford University
ebrun@cs.tanford.edu

Abstract

Practitioners are increasingly exploring the use of contextualized, data-driven decision policies in domains such as education, mobile health, behavioral science, or public policy. In these settings, it is common to gather initial pilot data to explore the potential benefit of new interventions, such as in the form of an A/B study. Estimating the benefits of future experimentation is important because additional data collection may incur significant operational costs, which must be weighed against the potential for learning a high-performing policy. Given a small amount of pilot data, we present a method in the linear contextual bandit setting for characterizing the quality of a dataset by computing the effective number of samples relative to minimax optimal batch exploration. When additional data collection is necessary, we extend existing algorithms for batch exploration and prove data-dependent reductions in sample complexity proportional to the quality of an initial dataset. In numerical experiments using simulated data, we illustrate both the benefit of our method in estimating the quality of the pre-existing data and how our exploration strategy can be used to efficiently gather additional data to find near-optimal policies.

Keywords: linear contextual bandits, optimal experimental design, pure exploration, linear bandits

1 Introduction

The use of machine learning methods to discover data-driven, contextual decision policies is increasingly receiving attention across a wide range of application domains [2, 5, 7, 8]. Currently, a hindrance to deploying data-driven methods in practice is that it is frequently necessary to characterize the expected benefit of future experimentation in advance. In real-world experiments, adding an additional 100 participants to a trial can require thousands of dollars and months of effort. Thus, researchers must weight the cost of collecting additional data against the potential for learning a better policy. In practice, it is exceedingly common to run small pilot studies before running a full-scale experiment. Such pilot data enables researchers to estimate how many additional participants their study will need to reach statistical significance for a given effect size (e.g., using a power analysis). However, it remains unclear how to design such forecasting methods for studies employing contextualized decision policies.

Moreover, most existing reinforcement learning algorithms are *adaptive* to past treatments and outcomes, presenting significant operational challenges to deploying these algorithms in practice. The ability to immediately update a policy after each step not only requires significant engineering overhead and personnel training, it may be impossible in longitudinal studies with delayed rewards or when treatments are assigned to multiple participants in parallel. It may also be necessary for policies to be audited for undesirable behavior (e.g., biased treatment allocation) prior to deployment, precluding the use of fully adaptive policies. Though it is often feasible to deploy stochastic, contextualized policies for data collection, practitioners require policies that are *non-adaptive* to incoming data for real-world experiments.

Recently, Zanette et al. [9] proposed the sampler-planner algorithm, which leverages offline state information to design a single, non-adaptive policy that learns an ϵ-optimal policy with optimal online sample complexity. The ability to leverage historical context information is often feasible in practice; for example, many organizations have access to demographic information about past study participants.

In this work, we aim to lower the barrier to conducting experiments using contextualized, data-driven decision policies in practice. We formalize the problem in the linear contextual bandit setting and present an extension to the sampler-planner algorithm that can be used to assess the quality of an initial dataset and guide future data collection to efficiently identify near-optimal policies. Specifically, we present a characterization of the quality of a pilot dataset by computing the effective number of samples relative to minimax optimal batch exploration. This *equivalent index* can be used to assess the efficiency of the initial data collection strategy and the potential benefit of optimal exploration for future data collection. We provide an initialization strategy that generates an exploration policy for future data collection and prove data-dependent reductions in sample complexity directly proportional to the quality of the pilot data. Lastly, we demonstrate the efficacy of our methods in numerical experiments on simulated data.

2 Setting

We consider a stochastic linear contextual bandit model where each context $s \in \mathcal{S}$ is sampled from a distribution μ. For each context s, a context-dependent action set $\mathcal{A}_s$ is made available to the learner. The bandit instance is defined by a known feature extractor $\phi(s,a) : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}^d$ and an unknown parameter $\theta^\star \in \mathbb{R}^d$. Upon choosing action $a \in \mathcal{A}_s$, the environment reveals a linear reward function $r(s,a) = \phi(s,a)^\top \theta^\star + \eta$ corrupted by mean-zero noise η. We define a policy π to be a mapping from states $s \in \mathcal{S}$ to a probability distribution over the action space $\mathcal{A}_s$. We say that π is adaptive if the distribution $a_t \sim \pi(s_t)$ is a function of the history $\mathcal{H}_{t-1} = \{(s_i, a_i, r_i)\}_{i=1}^{t-1}$ and that π is non-adaptive if this distribution is fixed.

In this work, we differentiate between *exploration* policies π^e and *exploitation* policies $\widehat{\pi}$. Exploration policies are used to interact with the environment to collect a dataset $\mathcal{D} = \{(s_t, a_t, r_t)\}_{t=1}^{N}$, where $a_t \sim \pi^e(s_t)$. Given a dataset $\mathcal{D}$, one can construct a regularized least-squares predictor $\hat{\theta}$ with regularization $\lambda > 0$, which defines an exploitation policy $\widehat{\pi}$.

$$\widehat{\Sigma} = \sum_{t=1}^{N} \phi(s_t, a_t)\phi(s_t, a_t)^\top + \lambda I, \quad \hat{\theta} = \left(\widehat{\Sigma}^{-1}\right) \sum_{t=1}^{N_{\text{init}}} \phi(s_t, a_t) \cdot r_t, \quad \widehat{\pi}(s) = \operatorname*{argmax}_{a \in \mathcal{A}_s} \phi(s,a)^\top \hat{\theta} \tag{1}$$

Unlike most prior work in the linear contextual bandit literature which aims to minimize cumulative regret [1, 4], exploration policies are not penalized for taking actions that incur large online regret. Instead, the objective is to collect an informative dataset $\mathcal{D}$ such that the resulting exploitation policy $\widehat{\pi}$ is near-optimal. Specifically, we optimize for the *suboptimality* $\mathbb{E}_{s\sim\mu}\left[\max_{a\in\mathcal{A}_s} \phi(s,a)^\top \theta^\star - \phi\left(s, \widehat{\pi}(s)\right)^\top \theta^\star\right]$ of an exploitation policy $\widehat{\pi}$ with respect to the optimal policy $\pi^\star$. This objective is analogous to simple regret in the multi-armed bandit literature [3]. By algebraic manipulation (see [9], pg. 16), one can show that a bound on the *maximum prediction error*, $\Delta(\widehat{\pi}) = \mathbb{E}_{s\sim\mu}\left[\max_{a\in\mathcal{A}_s}\left|\phi(s,a)^\top(\theta^\star - \hat{\theta})\right|\right] \leq \epsilon$, is sufficient for bounding the suboptimality by 2ϵ.

Algorithm 1 Planner

```
input: contexts $\mathcal{C}_{\text{offline}} = \{s_m\}_{m=1}^{M}$, $\mathcal{C}_{\text{eval}} = \{s_i\}_{i=1}^{N_{\text{eval}}}$,
    initial data $\mathcal{D}_{\text{init}} = \{(s_t, a_t, r_t)\}_{t=1}^{N_{\text{init}}}$, regularization $\lambda$
if $\mathcal{D}_{\text{init}} \neq \emptyset$ then
    $\widetilde{m} \leftarrow$ INITIALIZATIONSTRATEGY($\mathcal{D}_{\text{init}}, \mathcal{C}_{\text{offline}}, \mathcal{C}_{\text{eval}}$)
    $\Sigma_{\widetilde{m}+1} = \widehat{\Sigma}_{\text{init}}$
else
    $\widetilde{m} = 0; \Sigma_1 = \lambda I$
end if
for $m = \widetilde{m} + 1, 2, \dots M$ do
    if $\det(\Sigma_m) > 2\det(\Sigma_{\underline{m}})$ or $m = \widetilde{m} + 1$ then
        $\underline{m} \leftarrow m; \Sigma_{\underline{m}} \leftarrow \Sigma_m$
    end if
    Define $\pi_m : s \mapsto \text{argmax}_{a \in \mathcal{A}_s} \|\phi(s, a)\|_{\Sigma_{\underline{m}}^{-1}}$
    Receive context $s_m$ from $\mathcal{C}_{\text{offline}}$
    Define $\phi_m = \phi(s_m, \pi_m(s_m))$
    $\Sigma_{m+1} = \Sigma_m + \alpha \cdot \phi_m \phi_m^\top$
end for
output: policy mixture $\pi_{\text{mix}}$ of $\{\pi_{\widetilde{m}+1}, \dots, \pi_M\}$
```

Initial Data: In this work, we assume that the learner has access to an initial dataset $\mathcal{D}_{\text{init}} = \{(s_t, a_t, r_t)\}_{t=1}^{N_{\text{init}}}$. We assume that $\mathcal{D}_{\text{init}}$ was gathered using an exploration policy π^e_{init}, which may be adaptive or non-adaptive. We also assume the existence of a large collection of offline context data $\mathcal{C} = \{s_i\}$ (where $|\mathcal{C}| \gg N_{\text{init}}$), which is sufficiently large to be split into an evaluation set $\mathcal{C}_{\text{eval}}$ and offline training set $\mathcal{C}_{\text{offline}}$.

Given $\mathcal{D}_{\text{init}}$, one can construct the covariance matrix $\widehat{\Sigma}_{\text{init}}$, regularized least-squares predictor $\hat{\theta}_{\text{init}}$, and linear argmax policy $\widehat{\pi}_{\text{init}}$ as defined in Equation (1). Furthermore, $\mathcal{D}_{\text{init}}$ can optionally be used to generate a new exploration policy π^e_{after}. This policy is used to collect an additional dataset $\mathcal{D}_{\text{after}} = \{(s_t, a_t, r_t)\}_{t=1}^{N_{\text{after}}}$, where $a_t \sim \pi^e_{\text{after}}(s_t)$. Analogously to $\widehat{\pi}_{\text{init}}$, one can define the linear argmax policy $\widehat{\pi}_{\text{after}}$, where $\hat{\theta}_{\text{after}}$ is the regularized least-squares predictor learned from $\mathcal{D}_{\text{init}} \cup \mathcal{D}_{\text{after}}$. Our goal is design an exploration policy π^e_{after} given $\mathcal{D}_{\text{init}}$ such that the suboptimality of $\widehat{\pi}_{\text{after}}$ is minimized.

Assumptions: We assume that for each $(s_t, a_t, r_t) \in \mathcal{D}_{\text{init}} \cup \mathcal{D}_{\text{after}}$, the reward was generated using the true reward function. While we allow π^e_{init} to be adaptive, we assume that the noise η_t is 1-sub-Gaussian conditioned on the filtration $\mathcal{F}_{t-1}$ such that $\mathcal{H}_{t-1}$ is $\mathcal{F}_{t-1}$-measurable. Lastly, we assume that all features are bounded in ℓ_2 norm, $\forall (s, a) \in \mathcal{S} \times \mathcal{A}_s, \|\phi(s, a)\|_2 \leq 1$, and that all states in $\mathcal{C}$, $\mathcal{D}_{\text{init}}$, and $\mathcal{D}_{\text{after}}$ were sampled i.i.d. from the true distribution μ.

3 Algorithm

In this section, we present a different characterization of dataset quality via the equivalent index $\widetilde{n}$, which measures the number of samples of optimal exploration that would yield a policy that is at least as optimal as $\widehat{\pi}_{\text{init}}$. To calculate $\widetilde{n}$, we leverage the sampler-planner algorithm [9]. We briefly review the sampler-planner algorithm before outlining our equivalent index computation and initialization strategy.

3.1 Sampler-Planner Algorithm

Suppose for now that $\mathcal{D}_{\text{init}} = 0$. The sampler-planner algorithm [9] proceeds in two phases: (1) the *planner* (Algorithm 1) iterates over the offline context set $\mathcal{C}_{\text{offline}} = \{s_m\}_{m=1}^{M}$ for M steps and generates a mixture of policies $\pi_{\text{mix}} = \{\pi_m\}_{m=1}^{M}$ to be used as an exploration policy, and (2) the *sampler* executes this policy during online exploration for N steps, randomly sampling from π_{mix} for each online context. We enforce that $N \leq M$ and define the hyperparameter $\alpha = N/M$ to be the ratio of online to offline contexts.

3.2 Initialization Strategy

In this work, we extend the sampler-planner algorithm to take an initial dataset $\mathcal{D}_{\text{init}}$ as an input. Given $\mathcal{D}_{\text{init}}$, we define $\widetilde{m}$ to be the largest index m such that the empirical uncertainty over $\mathcal{C}_{\text{eval}}$ with respect to $\widehat{\Sigma}_{\text{init}}$ is less than or equal to the empirical uncertainty with respect to Σ_m.

$$\overline{u}(\widehat{\Sigma}) = \frac{1}{N_{\text{eval}}} \sum_{i=1}^{N_{\text{eval}}} \max_{a \in \mathcal{A}_{s_i}} \|\phi(s_i, a)\|_{\widehat{\Sigma}^{-1}}, \quad \widetilde{m} = \underset{m \in [M]}{\text{argmax}}\, m : \overline{u}(\widehat{\Sigma}_{\text{init}}) + 2\sqrt{\frac{\lambda \log(8M/\delta)}{2N_{\text{eval}}}} \leq \overline{u}(\Sigma_m) \tag{2}$$

The offset term results from Hoeffding's inequality and ensures that the inequality holds for the true expectations with high probability. $\widetilde{m}$ can be additionally interpreted as the index at which $\widehat{\pi}_{\text{init}}$ is at most as suboptimal as the planner's policy at stage $m = \widetilde{m}$. We define $\widetilde{n} = \alpha \cdot \widetilde{m}$, which places $\widetilde{m}$ on the same scale as N_{init} when $\alpha < 1$.

Algorithm 2 Initialization Strategy

```
input: $\mathcal{D}_{\text{init}} = \{(s_t, a_t, r_t)\}_{t=1}^{N_{\text{init}}}$, $\mathcal{C}_{\text{offline}} = \{s_m\}_{m=1}^{M}$, $\mathcal{C}_{\text{eval}} = \{s_i\}_{i=1}^{N_{\text{eval}}}$, regularization $\lambda$
Generate $\Sigma_1, ..., \Sigma_M$ using Algorithm 1 without $\mathcal{D}_{\text{init}}$    ▷ Generate reference covariance matrices
Let $\widetilde{m} = \text{argmax}_{m \in [M]}\ m : \overline{u}(\widehat{\Sigma}_{\text{init}}) + 2\sqrt{\frac{\lambda \log(8M/\delta)}{2N_{\text{eval}}}} \le \overline{u}(\Sigma_m)$    ▷ Compute equivalent index
output: equivalent index $\widetilde{m}$
```

Given a non-empty initial dataset $\mathcal{D}_{\text{init}} \neq 0$, the planner begins by running the initialization strategy (Algorithm 2), which computes the equivalent index as described above. The initialized planner then sets $\widehat{\Sigma}_{\text{init}}$ as its initial covariance matrix (as opposed to λI) and runs an identical procedure to the uninitialized case for $M - \widetilde{m}$ steps. This yields a policy mixture $\pi_{\text{mix}} = \{\pi_m\}_{m=\widetilde{m}+1}^{M}$ to be used as an exploration policy π^e_{after}. The sampler plays π_{mix} for $N - \widetilde{n}$ steps during online exploration, generating a dataset $\mathcal{D}_{\text{after}} = \{(s'_n, a'_n, r'_n)\}_{n=\widetilde{n}+1}^{N}$. In the following section, we evaluate the suboptimality of $\widehat{\pi}_{\text{after}}$, the policy learned from $\mathcal{D}_{\text{init}} \cup \mathcal{D}_{\text{after}}$.

4 Main Result

The following theorem provides a sample complexity bound on the initialized sampler-planner algorithm.

Theorem 1. *Let $\mathcal{D}_{init}$ be an arbitrary dataset satisfying the assumptions in Section 2 and let $\widetilde{m}$ be its equivalent index (Eq. 2). Consider running Alg. 1 initialized with $\widehat{\Sigma}_{init}$ for $M - \widetilde{m}$ offline iterations, where $M = \widetilde{\Omega}(\frac{d^2\beta}{\lambda\epsilon^2})$, and running the sampler for $N - \widetilde{n}$ online iterations, where $N = \widetilde{\Omega}(\frac{d\beta}{\epsilon^2})$, with regularization $\lambda \in (\Omega(\ln(d/\delta), d]$. For any $\epsilon \le 1$, with probability at least $1 - \delta$ the suboptimality of the greedy policy $\widehat{\pi}_{after}$ satisfies $\mathbb{E}_{s\sim\mu}\left[\max_{a\in\mathcal{A}_s} \phi(s,a)^\top\theta^\star - \phi(s, \widehat{\pi}_{after}(s))^\top\theta^\star\right] \le \epsilon$*

Due to space this proof is omitted. Our main contribution is in proving that our initialization strategy induces the same upper bound on the planner's cumulative uncertainty as in the uninitialized setting, after which our proof closely follows the structure of [9]. While many possible initialization strategies are feasible (e.g., comparing determinants or matrix norms), the main implication of this theorem is that $\widetilde{m}$ and $\widetilde{n}$ precisely characterize the reduction in sample complexity induced by $\mathcal{D}_{\text{init}}$. For a dataset of high quality (i.e., a dataset that yields a policy that close to ϵ-optimal), $\widetilde{m}$ will be large because fewer samples are needed to reach ϵ-optimality. For datasets of low quality, $\widetilde{m}$ will be low. In the uninitialized setting ($\mathcal{D}_{\text{init}} = \emptyset$), the sample complexity is identical to that of [9].

The ratio of $\widetilde{n}/N_{\text{init}}$ can also be used as a heuristic to assess the sample efficiency of π^e_{init} relative to the optimal exploration strategy. For example, suppose that π^e_{init} pulls arms uniformly at random (i.e., a randomized controlled trial), as is standard for most studies. If $\widetilde{n}/N_{\text{init}}$ is less than 1, this implies that random allocation requires significantly more samples to learn an ϵ-optimal policy than optimal exploration. This also implies that there is significant benefit to collecting additional data using the initialized sampler-planner: optimal exploration can greatly reduce the number of samples, and thus the amount of money or resources required to learn an ϵ-optimal policy. If $\widetilde{n}/N_{\text{init}} \approx 1$, then there is little benefit to deploying optimal exploration strategies over random assignment. This situation might arise when the context space has little structure or the optimal treatment is easy to identify.

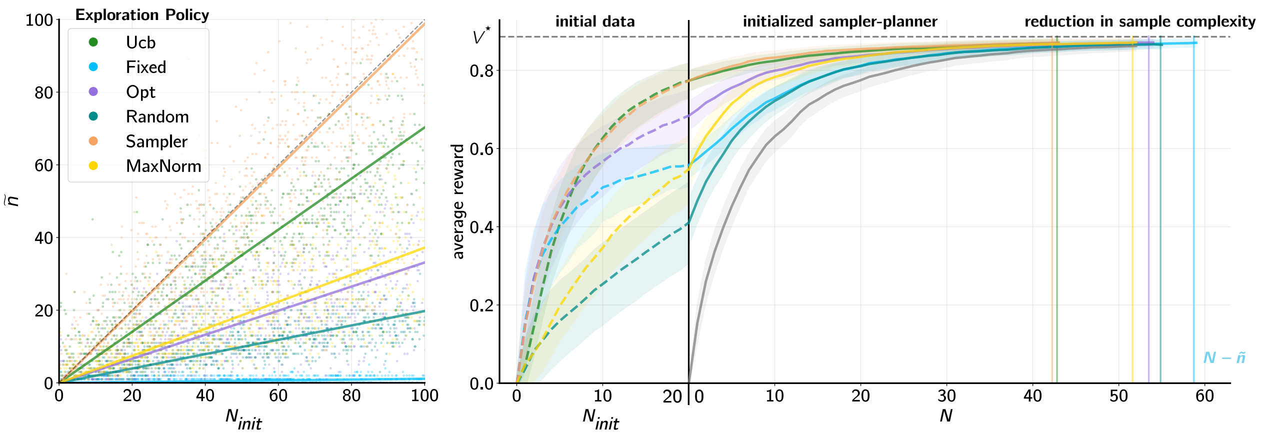

Figure 1: Numerical experiments using simulated data. **Left:** Assessing the quality of initial data. **Right:** Future exploration with data-dependent reductions in sample complexity.

5 Experiments

In this section, we report on numerical experiments using simulated data. We use the same simulator as [9], which is designed such that the context space contains structure to be exploited by optimal exploration. A random policy is unlikely to perform well since half of the actions lead to features that are zero. A policy which only chooses actions by the largest norm is also likely to underperform since certain action have large norms but lead to small rewards.

We run two experiments: (1) we compute $\widetilde{n}$ as a function of N_{init} for various initial exploration policies to test our method's ability to assess initial dataset quality and (2) we initialize the sampler-planner with various initial datasets, testing the extent to which our data-dependent reductions in sample complexity hold in practice. In our experiments, we reference the following exploration policies π^e_{init}: Sampler: the (uninitialized) sampler-planner using $N = M = 2N_{\text{init}}$; UCB: standard LinUCB [6]; Fixed: a policy that always return a fixed action ($a_t = 1$); Opt: an oracle policy that chooses the action which yields the highest online reward; Random: a policy that chooses each action uniformly at random; MaxNorm: a policy that chooses the action that yields the largest ℓ_2 norm.

In Figure 1 (left), we explore how $\widetilde{n}$ changes as a function of N_{init} and π^e_{init}. More specifically, we generate $\mathcal{D}_{\text{init}}$ for various sizes of $N_{\text{init}} \in [0, 100]$ using all exploration policies π^e_{init} listed above. We plot the equivalent index $\widetilde{n}$ (y-axis) as a function of N_{init} (x-axis) and include linear trend-lines for each policy. We find that exploration policies of higher quality are assigned a greater equivalent index $\widetilde{n}$. Policies that explore at an optimal or near-optimal rate (e.g., Sampler or UCB) yield a higher equivalent index for the same N_{init} as policies that explore at suboptimal rates (e.g., Random or Fixed).

In Figure 1 (right), we verify the results of Theorem 1, namely that $\widetilde{n}$ characterizes the data-dependent reductions in sample complexity when collecting additional data using our initialization strategy. For each exploration policy, we generate an initial dataset $\mathcal{D}_{\text{init}}$ of size $N_{\text{init}} = 20$, which is plotted on the left half of the figure. We run the initialized sampler-planner for $N - \widetilde{n}$ steps with $N = 60$, plotted on the right half of the figure. In gray, we plot the uninitialized sampler-planer for reference and average values of $N - \widetilde{n}$ are plotted as vertical lines. We find that the reduction in sample complexity is proportional to quality of $\mathcal{D}_{\text{init}}$: for initial datasets of high quality (e.g., Sampler or UCB), we require a few more than $N - N_{\text{init}}$ samples to compute a near-optimal policy. For datasets of low quality (e.g., Random or Fixed), we find that they require closer to N samples to compute a near-optimal policy.

6 Conclusion

In this work, we present a novel extension to the sampler-planner algorithm [9] which accounts for the presence of initial data. We provide a characterization of initial dataset quality via an equivalent index computation and prove that our initialization strategy leads to reductions in sample complexity that directly proportional to the quality of the initial dataset. In future work, we aim to explore how initial reward information can be used to change the planner's policy and to study our algorithm's performance under reward misspecification.

References

[1] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR, 2013.

[2] A. Banerjee, A. G. Chandrasekhar, S. Dalpath, E. Duflo, J. Floretta, M. O. Jackson, H. Kannan, F. N. Loza, A. Sankar, A. Schrimpf, et al. Selecting the most effective nudge: Evidence from a large-scale experiment on immunization. Technical report, National Bureau of Economic Research, 2021.

[3] S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

[4] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214. JMLR Workshop and Conference Proceedings, 2011.

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

[6] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[7] P. Liao, K. Greenewald, P. Klasnja, and S. Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.

[8] T. Mu, S. Wang, E. Andersen, and E. Brunskill. Automatic adaptive sequencing in a webgame. In *International Conference on Intelligent Tutoring Systems*, pages 430–438. Springer, 2021.

[9] A. Zanette, K. Dong, J. Lee, and E. Brunskill. Design of experiments for stochastic contextual linear bandits. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.