Simple Regret Minimization for Contextual Bandits Using Bayesian Optimal Experimental Design

Matthew Jörke JOERKE@STANFORD.EDU
Department of Computer Science
Stanford University

Jonathan Lee JNL@STANFORD.EDU
Department of Computer Science
Stanford University

Emma Brunskill EBRUN@CS.STANFORD.EDU
Department of Computer Science
Stanford University

Abstract

We study the best policy identification problem for contextual bandits through the lens of Bayesian optimal experimental design. Motivated by practical constraints when deploying contextual policies in real-world experiments, we focus on the non-adaptive setting in which a single exploration policy is used to collect data for the entire experiment. We demonstrate that common information-theoretic utilities can lead to suboptimal exploration in the presence of initial data and instead propose directly optimizing for the value of the resulting decision policy. To solve this optimal design problem, we derive a policy gradient algorithm that is able to learn good exploration policies in linear contextual bandit settings. Unlike existing algorithms, our Bayesian method is able to leverage prior reward information (e.g., pilot data or expert knowledge) for more efficient exploration. We evaluate our policy gradient algorithm on a linear benchmark task, demonstrating that our approach is able to identify optimal decision policies more efficiently than existing baselines.

1. Introduction

The use of machine learning methods to discover data-driven, contextual decision policies is receiving increasing attention across a wide range of application domains. In education, researchers have designed intelligent tutoring systems that adaptively assign students to pedagogical conditions that improve their learning outcomes [6, 34]. Mobile health researchers have explored personalized, adaptive interventions to support positive behavior change, such as promoting physical activity [27, 38] or preventing smoking relapse [7]. Econometricians have recently begun exploring adaptive treatment assignment algorithms for accelerating large-scale field experiments that inform public policy decisions [25]. For instance, researchers in public health may be interested in determining which interventions are most effective at increasing vaccinations [5, 33] or physical activity [32].

In these settings, the primary research interest is to identify a decision policy that assigns each unit to an optimal treatment condition as a function of their covariates, as opposed to accurately estimating the effect of one treatment compared to a control. The task of designing an experiment for this purpose is naturally modeled as a *best policy identification* problem, in which the objective is to design an experiment such that the likelihood of identifying an ε-optimal decision policy at the end of the study is maximized. This setting is known as pure exploration in the bandit literature and has been studied extensively in the multi-armed bandit [9, 19, 40] and linear bandit [14, 44, 45, 49] settings. Surprisingly, pure exploration for contextual bandits has received relatively little attention.

Currently, a hindrance to deploying data-driven methods in practice is that most existing algorithms for best policy identification are *adaptive* to past treatments and outcomes [30, 50]. 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 in studies with parallel treatment assignment. However, it is often feasible to deploy stochastic, contextualized policies for data collection, so long as they remain fixed throughout the study and are *non-adaptive* to incoming data. Moreover, existing non-adaptive algorithms (e.g., [50]) are insensitive to prior information that experimenters often have access to. For example, it is exceedingly common for researchers to run small pilot studies before launching a full-scale experiment. In principle, best policy identification algorithms ought to be able to make use of this information to more efficiently discover near-optimal treatments. Similarly, researchers may have knowledge (e.g., results from previous research) that can be encoded in a prior distribution over a model's parameters.

In this work, we aim to lower the barrier to conducting experiments using contextualized, data-driven decision policies in practice. We formalize the best policy identification problem as a Bayesian optimal experimental design (BOED) problem [10, 29] in the contextual bandit setting. In particular, we demonstrate that information-theoretic utilities from the BOED literature can lead to inefficient exploration and instead propose directly maximizing the expected value of the resulting decision policy. To solve this design problem, we propose a policy gradient algorithm inspired by REINFORCE [48] that can efficiently identify good exploration policies in settings where many global optimization methods fail. Unlike existing frequentist approaches, our Bayesian formulation naturally supports the presence of prior information and is able to exploit this information to more efficiently identify optimal policies. We demonstrate the benefit of our method in numerical experiments in simulated linear bandit setting.

2. Related Work

Pure Exploration in Bandits In the contextual bandit literature, most prior work optimizes for cumulative regret [1, 2, 18]. This objective is misaligned with our pure exploration setting, since suboptimal arms may still be informative for learning optimal policies. In the linear (i.e., non-contextual) bandit setting, a large body of work [11, 14, 24, 44, 45, 49] has studied the best-arm identification problem. Deshmukh et al. [12] investigated pure exploration in the contextual bandit setting and propose an adaptive, gap-based exploration algorithm.

The notion of limited adaptivity has received recent attention in the batch bandit literature, in which the goal is to learn a good policy with a limited number of updates [13, 22, 39, 51]. However, these algorithms are designed to minimize cumulative online regret, not policy suboptimality. Most relevant to our setting, Zanette et al. [50] recently proposed the Sampler-Planner algorithm, which leverages offline state information to design a single, non-adaptive exploration policy that learns an ϵ-optimal decision policy with $\widetilde{\Omega}(d^2/\varepsilon^2)$ online sample complexity, where d is the dimensionality.

In the Bayesian bandit setting, Russo [40] and Russo and Van Roy [41] present algorithms for best-arm identification. Though originally proposed for regret minimization, Russo and Van Roy's [41] information directed sampling objective can be adapted to the pure exploration setting (see their Proposition 9). This objective maximizes the mutual information between the data to be collected and the optimal arm. Similar information-theoretic acquisition functions are common in the field of Bayesian optimization [23, 46], where they are known as entropy search.

Bayesian Optimal Experimental Design Following Lindley [29], most work in BOED maximizes the expected information gain (EIG), which corresponds to the mutual information of the parameter and the data. Calculating the EIG is doubly intractable for general priors and likelihoods, and a large body of work has explored approximation methods [15, 16, 36, 47]. While the EIG is a natural objective for many scientific problems, it can yield sample inefficient designs for best policy identification problems because the EIG encourages *uniform* uncertainty reduction over the parameter [31]. Later, we will show that maximizing EIG can lead to suboptimal exploration in our contextual bandit setting. To compute the optimal design that maximizes the EIG, prior work has proposed a variety of approaches that rely on global optimization methods [3, 15, 26, 35, 35, 37]. There also exist methods that simultaneously estimate and optimize over the EIG [16, 20, 26]. In the

sequential design setting, several recent works have utilized policy gradient algorithms to compute optimal designs [4, 8, 17, 28, 43]. While this line of work bears similarity to our approach, we note that the methods above optimize for EIG and yield adaptive exploration policies.

3. Preliminaries

We consider a stochastic contextual bandit model where each context $s \in \mathcal{S}$ is independently sampled from a distribution ρ. We assume that ρ is known or that it can be approximated with a large set of offline context data $\mathcal{C} = \{s_i\}$.[1] 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 reward model $r_\theta(s, a)$, where θ is an unknown parameter with prior distribution $p(\theta)$. Upon choosing action a_t in state s_t, the environment reveals reward $r_t \sim r_{\theta^\star}(s_t, a_t)$, where $\theta^\star$ is the environment's true parameter. For example, a linear-Gaussian reward model follows $r_\theta(s_t, a_t) = \phi(s_t, a_t)^\top \theta + \eta_t$, where $\phi(s, a) : \mathcal{S} \times \mathcal{A}_s \to \mathbb{R}^d$ is a known feature extractor and $\eta_t \sim \mathcal{N}(0, \sigma^2)$. We define the expected reward for action a in state s as $\mu_\theta(s, a) = \mathbb{E}[r_\theta(s, a)]$, where the expectation is taken over the randomness in $r_\theta(s, a)$ for a fixed θ.

We define a policy $\pi \in \Pi$ to be a mapping from states $s \in \mathcal{S}$ to a probability distribution over the action space $\mathcal{A}_s$, i.e. $\Pi = \{\pi \mid \pi : \mathcal{S} \to \Delta_{\mathcal{A}_s}\}$. We say that π is *adaptive* if the distribution $a_t \sim \pi(\cdot|s_t)$ depends on 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 for all t. In this work, we differentiate between *exploration* policies π^e and *decision* policies $\widehat{\pi}$. Exploration policies are used to interact with the environment to collect a dataset $\mathcal{D}_N = \{(s_t, a_t, r_t)\}_{t=1}^N$, where $a_t \sim \pi^e(\cdot|s_t)$ and $r_t \sim r_{\theta^\star}(s_t, a_t)$. Given a dataset $\mathcal{D}_N$, we can update our posterior estimate of the expected reward, which we define as $\hat{\mu}_{\theta|\mathcal{D}_N}(s, a) = \mathbb{E}_{\theta|\mathcal{D}_N}[\mu_\theta(s, a)]$. We consider the class of decision policies that greedily choose the arm with the highest posterior expected reward: $\widehat{\pi}(s) = \text{argmax}_{a \in \mathcal{A}_s} \hat{\mu}_{\theta|\mathcal{D}_N}(s, a)$.

Unlike the traditional bandit setting, exploration policies are not penalized for taking actions that incur large online regret. Instead, the objective is to collect an informative dataset $\mathcal{D}_N$ such that the resulting decision policy $\widehat{\pi}$ is near-optimal. Specifically, we optimize for *simple regret* [9], defined as the expected suboptimality of a decision policy $\widehat{\pi}$ with respect to the optimal policy $\pi^\star(s) = \text{argmax}_{a \in \mathcal{A}_s} \mu_{\theta^\star}(s, a)$.

$$\mathbb{E}_{s \sim \rho}\left[V^{\pi^\star}(s) - V^{\widehat{\pi}}(s)\right] = \mathbb{E}_{s \sim \rho}\left[\max_{a \in \mathcal{A}_s} \mu_{\theta^\star}(s, a) - \mu_{\theta^\star}(s, \widehat{\pi}(s))\right] \tag{1}$$

Initial Data We allow for the possibility that the learner has access to an initial dataset $\mathcal{D}_{N_1} = \{(s_t, a_t, r_t)\}_{t=1}^{N_1}$. Given $\mathcal{D}_{N_1}$, one can generate a new exploration policy π_2^e, which is used to collect an additional dataset $\mathcal{D}_{N_2} = \{(s_t, a_t, r_t)\}_{t=1}^{N_2}$, where $a_t \sim \pi_2^e(\cdot|s_t)$. Defining the decision policy $\widehat{\pi}_2(s) = \text{argmax}_{a \in \mathcal{A}_s} \hat{\mu}_{\theta|\mathcal{D}_1\mathcal{D}_2}(s, a)$, the objective is to design π_2^e such that the suboptimality of $\widehat{\pi}_2$ is minimized. In this work, we primarily focus on the two-stage procedure described above. However, this setting can be extended to an arbitrary number of sequential experiments.

4. Simple Regret Minimization Using BOED

We formulate the best policy identification problem inspired by a Bayesian optimal experimental design (BOED) framework [10]. First, assume that the prior $p(\theta)$ and reward model $p(r_t|s_t, a_t, \theta)$ are known and well-specified. If initial data $\mathcal{D}_{\text{init}}$ is provided, the prior can be refined to $p(\theta|\mathcal{D}_{\text{init}})$. Under the contextual bandit model described in Section 3, the likelihood is given by

$$p\left(\mathcal{D}_N = \{(s_t, a_t, r_t)\}_{t=1}^N \mid \theta, \pi^e\right) = \prod_{t=1}^N \rho(s_t) \cdot \pi^e(a_t|s_t) \cdot p(r_t|s_t, a_t, \theta) \tag{2}$$

[1] This assumption is frequently tenable in practice. For example, researchers often have access to participant demographic information prior to running an experiment.

Algorithm 1: Policy Gradient Optimization for Best Policy Identification

if $\mathcal{D}_{\text{init}} \neq \varnothing$ **then** let $p(\theta) \leftarrow p(\theta|\mathcal{D}_{\text{init}})$
while π^e_τ is not converged **do**
  sample $\theta^{(i)} \overset{iid}{\sim} p(\theta)$ for $i = 1, ..., K$
  sample $\mathcal{D}^{(i)}_N \overset{iid}{\sim} p(\mathcal{D}_N|\theta^{(i)}, \pi^e_\tau)$ for $i = 1, ..., K$
  compute $U(\theta^{(i)}, D^{(i)}_t)$ for $i = 1, ..., K$, $t = 0, ..., N$
  let $\nabla_\tau \hat{\ell}(\tau) = \frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{N} \nabla_\tau \log\left(\pi^e_\tau(a_t|s_t)\right) \left(U(\theta^{(i)}, \mathcal{D}^{(i)}_N) - U(\theta^{(i)}, \mathcal{D}^{(i)}_{t-1})\right)$
  $\tau \leftarrow \tau + \eta \nabla_\tau \hat{\ell}(\tau)$
end

We wish to learn an exploration policy $\pi^e \in \Pi$ that minimizes the simple regret (Equation 1) in expectation over $p(\theta, \mathcal{D}_N|\pi^e)$. This is equivalent to maximizing the expected value of the learned decision policy $\widehat{\pi}$. This motivates a utility function based on the expected decision policy's value, $U_{\text{value}}(\theta, \mathcal{D}_N) = \mathbb{E}_{s\sim\rho}\left[\mu_\theta(s, \widehat{\pi}(s))\right]$ and the resulting optimization problem is given by

$$\pi^e_\star = \underset{\pi^e \in \Pi}{\text{argmax}} \; \mathbb{E}_{\theta, \mathcal{D}_N|\pi^e}\left[U_{\text{value}}(\theta, \mathcal{D}_N)\right] \tag{3}$$

This differs from standard BOED in two important ways. First, our design variable is a stochastic policy π^e (i.e., a function), not a single decision or sequence of decisions. This distinction motivates a different class of solution methods for computing Equation (3). Second, and perhaps even more importantly, our utility function directly minimizes expected simple regret, in contrast to most BOED work that focuses on expected information gain.

4.1 Solving for the Optimal Policy

The expected utility $\mathbb{E}_{\theta, \mathcal{D}_N|\pi^e}\left[U_{\text{value}}(\theta, \mathcal{D}_N)\right]$ generally does not admit a closed form expression and must be approximated. In principle, one could approximate the expected utility using samples from $p(\theta)$ and $p(\mathcal{D}_N|\theta, \pi^e)$ and directly apply any global optimization technique (e.g., grid search, simulated annealing, CMA-ES, Bayesian optimization) to solve for an approximate optimizer. However, there are several challenges with this approach. Practical problems of interest frequently involve differentiating between actions with small gaps, meaning that the true differences in expected utility may be very small for two candidate optima. This necessitates a prohibitively large number of samples such that the Monte Carlo sampling error is less than the true differences in utility. However, even when the utility can be reasonably approximated or the optimizer is robust to noisy objectives, $U_{\text{value}}(\theta, \mathcal{D}_N)$ only provides a global weight on the quality of an entire dataset, but does not reliably indicate which actions in a dataset $\mathcal{D}_N$ contributed to an increase in utility.

Instead, we propose parameterizing the policy class and employing policy gradient algorithms. Suppose that the policy class Π can be parameterized by τ, e.g. through a tabular parameterization or a neural network. Using the log-derivate trick, we can calculate the gradient of the expected utility with respect to τ.

$$\nabla_\tau \, \mathbb{E}_{\theta, \mathcal{D}_N|\pi^e_\tau}\left[U(\theta, \mathcal{D}_N)\right] = \mathbb{E}_{\theta, \mathcal{D}_N|\pi^e_\tau}\left[U(\theta, \mathcal{D}_N) \nabla_\tau \log p(\mathcal{D}_N|\theta, \pi^e_\tau)\right] \tag{4}$$

The variance of this estimator can be further reduced using common strategies from the policy gradient literature [21], including rearranging summation to remove independent terms (much like REINFORCE [48]) and subtracting a baseline utility. The estimator belongs to a class of generalized advantage estimators [42] using $U(\theta, D_t)$ as the reward at time t. We now prove that this advantage form of Equation (4) is an unbiased estimator of the gradient, both for our simple regret utility, as well as other alternate utility functions:

Theorem 1. *Let $U(\theta, \mathcal{D})$ be an arbitrary utility function. Let $U(\theta, \mathcal{D}_0)$ be the baseline utility given no additional information, where $\mathcal{D}_0 = \varnothing$. Then, the following yields an unbiased estimator of the gradient.*

$$\nabla_\tau \mathbb{E}_{\theta, \mathcal{D}_N | \pi^e_\tau} \left[U\left(\theta, \mathcal{D}_N\right) \right] = \mathbb{E}_{\theta, \mathcal{D}_N | \pi^e_\tau} \left[\sum_{t=1}^{N} \nabla_\tau \log \left(\pi^e_\tau(a_t | s_t) \right) \cdot \Big(U(\theta, \mathcal{D}_N) - U(\theta, \mathcal{D}_{t-1}) \Big) \right] \tag{5}$$

The proof is omitted for space and is nearly identical to the derivation of REINFORCE [48] with a baseline. This gradient estimator can be approximated using sample averages for use in policy optimization with stochastic gradient descent, as shown in Algorithm 1. In numerical experiments, we find the additional variance reduction from using the advantage utility form to be crucial for good performance even in low-dimensional settings.

4.2 Alternate Utility Functions

Having introduced our policy gradient algorithm for computing an exploration policy that minimizes Bayesian simple regret, we now frame this choice within the broader context of alternative utility functions and demonstrate the formal benefits of using this approach.

Expected Information Gain In the BOED literature, it is common to maximize the expected information gain (EIG), which corresponds to the mutual information between θ and $\mathcal{D}_N$.

$$\mathbb{E}_{\theta, \mathcal{D}_N | \pi^e}[U_{\text{EIG}}(\theta, \mathcal{D}_N)] = \mathcal{I}(\theta; \mathcal{D}_N) = H(\theta) - H(\theta | \mathcal{D}_N) \tag{6}$$

This objective encourages *uniform* reduction in uncertainty over each dimension in θ. As captured by the following lemma, it is possible to show that Zanette et al.'s [50] algorithm for simple regret minimization is equivalent to a greedy approximation to the N-step Bayesian optimal design problem that maximizes the EIG.

Lemma 1. *Consider a greedy approximation to the N-step optimal design problem maximizing the expected information gain: $\pi_t = \text{argmax}_{\pi^e \in \Pi} \mathbb{E}_{\theta, \mathcal{D}_t | \mathcal{D}_{t-1}}[U_{\text{EIG}}(\theta, \mathcal{D}_t)]$ for $t = 1, ..., N$. When $\theta \sim \mathcal{N}(0, \lambda^{-1} I)$, $r_\theta(s_t, a_t) = \phi(s_t, a_t)^\top \theta + \eta_t$, and $\eta_t \sim \mathcal{N}(0, 1)$, we recover the Sampler-Planner policy from Zanette et al. [50].*

This result follows directly from a known equivalence between Bayesian D-optimal designs and maximum EIG designs in the linear-Gaussian setting [10, 29], since $H(\theta | \mathcal{D}_N) \propto \log \det(\Sigma_N)$. The Sampler-Planner always chooses the action a_t that greedily maximizes $\det(\Sigma_{t+1})$.

However, we now show that methods that compute exploration policies to maximize expected information gain about the underlying problem parameters can be significantly inefficient for optimizing simple regret. As captured by the following theorem, the Sampler-Planner's policy, which uniformly reduces uncertainty over each dimension, can explore at a suboptimal rate by wasting samples on dimensions that do not help discriminate between arms.

Theorem 2. *There exists an $O(d)$-dimensional linear bandit instance and initial dataset $\mathcal{D}$ where the optimal exploration policy can obtain an $O(\varepsilon)$ upper bound on simple regret using $\widetilde{O}(1/\varepsilon^2)$ samples with probability at least $1 - \delta$. For the same instance, the Sampler-Planner exploration policy requires $\widetilde{O}(d/\varepsilon^2)$ samples to guarantee a simple regret of at most $O(\varepsilon)$ with probability at least $1 - \delta$.*

A proof will be provided in a longer version of this paper.

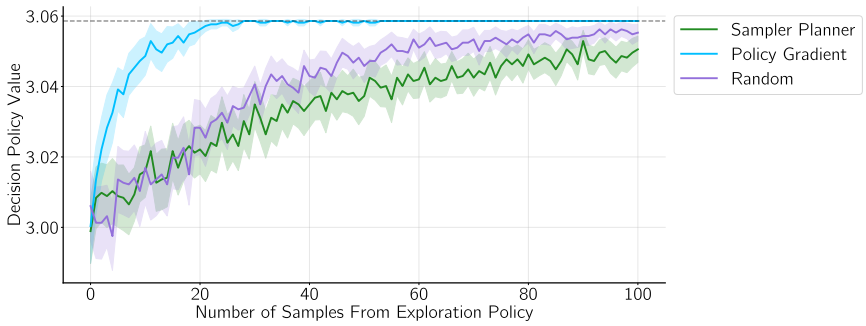

Figure 1: Linear Benchmark Task. We plot the value of a decision policy (y-axis) learned from N samples of exploration data (x-axis). The shaded region corresponds to ± 1.96 standard deviation computed across 100 trials.

5. Experiments

We evaluate our algorithm on a benchmark task inspired by the linear bandit literature. This task is a generalization of the simulator first introduced in Soare et al. [44] to the contextual setting. Specifically, we consider a problem in $S = 5$ states with $K = 50$ actions in $d = 51$ dimensions. In each state i, actions 1 and 2 are near-optimal with a small gap in the pivot dimension $i + 1$. Each state contains $S - 1 = 4$ informative actions that can resolve uncertainty about all pivot dimensions in other states, but not dimension $i + 1$. Each state also contains 45 "bad" actions that have large norm but lead to zero reward and do not resolve uncertainty over any pivot dimension. This might reasonably correspond to an experiment with many possible treatments, a select few of which are good, some of which are informative, and most of which are ineffective.

In Figure 1, we compare three exploration policies: (1) our policy gradient algorithm (Algorithm 1), (2) the Sampler-Planner [50], and (3) uniform random exploration. We consider the setting with 5 states, each with 2 near-optimal actions, 4 informative actions, and 45 bad actions. Each policy is initialized with four samples of near-optimal actions. We plot the value of a decision policy learned from N samples of data collected using a given exploration policy in addition to the four samples of initial data, averaged over 100 trials. We find that the policy gradient algorithm is able to learn an optimal decision policy with far fewer samples than either baselines. This is because random exploration and the Sampler-Planner both waste samples pulling uninformative actions, while the policy gradient algorithm concentrates its budget on informative actions.

6. Conclusion & Future Work

In this work, we have studied the best policy identification problem in the contextual bandit setting using a BOED framework. This Bayesian reinterpretation yielded insights into existing approaches and our preliminary numerical experiments show that our policy gradient algorithm is able to efficiently discover near-optimal decision policies.

There are a number of promising directions for future work. First, further theoretical analysis is necessary to characterize the sample complexity of $\pi^e_\star$ in the non-adaptive setting. We are also interested in studying other information-theoretic utility functions, e.g., one which maximizes the mutual information between the optimal arm and the dataset. Second, while our BOED framework is general enough to account to arbitrary likelihood functions and priors, our policy gradient algorithm is computationally expensive when the posterior $p(\theta|\mathcal{D}_N)$ is difficult to approximate. It may be possible to amortize this computation upfront using approximation techniques Foster et al. [15, 17]. It may also be possible to approximate more complex reward functions with Gaussian processes. The last and perhaps most important future direction is to extend the approach to assess the empirical benefits of direct simple regret minimization in more complex environments.

Acknowledgements Research reported in this paper was sponsored in part by NSF grant 2112926.

References

[1] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

[3] B. Amzal, F. Y. Bois, E. Parent, and C. P. Robert. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical association*, 101(474):773–785, 2006.

[4] Y. Ashenafi, P. Pandita, and S. Ghosh. Reinforcement learning based sequential batch-sampling for bayesian optimal experimental design. *arXiv preprint arXiv:2112.10944*, 2021.

[5] 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.

[6] J. Bassen, B. Balaji, M. Schaarschmidt, C. Thille, J. Painter, D. Zimmaro, A. Games, E. Fast, and J. C. Mitchell. Reinforcement learning for the adaptive scheduling of educational activities. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[7] S. L. Battalio, D. E. Conroy, W. Dempsey, P. Liao, M. Menictas, S. Murphy, I. Nahum-Shani, T. Qian, S. Kumar, and B. Spring. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.

[8] T. Blau, E. Bonilla, A. Dezfouli, and I. Chades. Optimizing sequential experimental design with deep reinforcement learning. *arXiv preprint arXiv:2202.00821*, 2022.

[9] 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.

[10] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.

[11] R. Degenne, P. Ménard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

[12] A. A. Deshmukh, S. Sharma, J. W. Cutler, M. Moldwin, and C. Scott. Simple regret minimization for contextual bandits. *arXiv preprint arXiv:1810.07371*, 2018.

[13] H. Esfandiari, A. Karbasi, A. Mehrabian, and V. Mirrokni. Regret bounds for batched bandits. *arXiv preprint arXiv:1910.04959*, 2019.

[14] T. Fiez, L. Jain, K. G. Jamieson, and L. Ratliff. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems*, 32:10667–10677, 2019.

[15] A. Foster, M. Jankowiak, E. Bingham, P. Horsfall, Y. W. Teh, T. Rainforth, and N. Goodman. Variational bayesian optimal experimental design. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] A. Foster, M. Jankowiak, M. O'Meara, Y. W. Teh, and T. Rainforth. A unified stochastic gradient approach to designing bayesian-optimal experiments. In *International Conference on Artificial Intelligence and Statistics*, pages 2959–2969. PMLR, 2020.

[17] A. Foster, D. R. Ivanova, I. Malik, and T. Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *International Conference on Machine Learning*, pages 3384–3395. PMLR, 2021.

[18] D. Foster and A. Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

[19] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[20] T. Goda, T. Hironaka, W. Kitade, and A. Foster. Unbiased mlmc stochastic gradient-based optimization of bayesian experimental designs. *SIAM Journal on Scientific Computing*, 44(1):A286–A311, 2022.

[21] E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.

[22] Y. Han, Z. Zhou, Z. Zhou, J. Blanchet, P. W. Glynn, and Y. Ye. Sequential batch learning in finite-action linear contextual bandits. *arXiv preprint arXiv:2004.06321*, 2020.

[23] P. Hennig and C. J. Schuler. Entropy search for information-efficient global optimization. *Journal of Machine Learning Research*, 13(6), 2012.

[24] Y. Jedra and A. Proutiere. Optimal best-arm identification in linear bandits. *arXiv preprint arXiv:2006.16073*, 2020.

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

[26] S. Kleinegesse and M. U. Gutmann. Efficient bayesian experimental design for implicit models. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 476–485. PMLR, 2019.

[27] 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.

[28] V. Lim, E. Novoseller, J. Ichnowski, H. Huang, and K. Goldberg. Policy-based bayesian experimental design for non-differentiable implicit models. *arXiv preprint arXiv:2203.04272*, 2022.

[29] D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

[30] T. Matsushima, H. Furuta, Y. Matsuo, O. Nachum, and S. Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint arXiv:2006.03647*, 2020.

[31] V. Mehta, B. Paria, J. Schneider, S. Ermon, and W. Neiswanger. An experimental design perspective on model-based reinforcement learning. *arXiv preprint arXiv:2112.05244*, 2021.

[32] K. L. Milkman, D. Gromet, H. Ho, J. S. Kay, T. W. Lee, P. Pandiloski, Y. Park, A. Rai, M. Bazerman, J. Beshears, et al. Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889): 478–483, 2021.

[33] K. L. Milkman, M. S. Patel, L. Gandhi, H. N. Graci, D. M. Gromet, H. Ho, J. S. Kay, T. W. Lee, M. Akinola, J. Beshears, et al. A megastudy of text-based nudges encouraging patients to get vaccinated at an upcoming doctor's appointment. *Proceedings of the National Academy of Sciences*, 118(20), 2021.

[34] 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.

[35] P. Müller. Simulation based optimal design. *Handbook of Statistics*, 25:509–518, 2005.

[36] J. I. Myung, D. R. Cavagnaro, and M. A. Pitt. A tutorial on adaptive design optimization. *Journal of mathematical psychology*, 57(3-4):53–67, 2013.

[37] D. J. Price, N. G. Bean, J. V. Ross, and J. Tuke. An induced natural selection heuristic for finding optimal bayesian experimental designs. *Computational Statistics & Data Analysis*, 126:112–124, 2018.

[38] M. Rabbi, M. H. Aung, M. Zhang, and T. Choudhury. Mybehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 707–718, 2015.

[39] Y. Ruan, J. Yang, and Y. Zhou. Linear bandits with limited adaptivity and learning distributional optimal design. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–87, 2021.

[40] D. Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

[41] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[42] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[43] W. Shen and X. Huan. Bayesian sequential optimal experimental design for nonlinear models using policy gradient reinforcement learning. *arXiv preprint arXiv:2110.15335*, 2021.

[44] M. Soare, A. Lazaric, and R. Munos. Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 27:828–836, 2014.

[45] C. Tao, S. Blanco, and Y. Zhou. Best arm identification in linear bandits with linear dimension dependency. In *International Conference on Machine Learning*, pages 4877–4886. PMLR, 2018.

[46] J. Villemonteix, E. Vazquez, and E. Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

[47] B. T. Vincent and T. Rainforth. The darc toolbox: automated, flexible, and efficient delayed and risky choice experiments using bayesian adaptive design. *PsyArXiv. October*, 20, 2017.

[48] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[49] L. Xu, J. Honda, and M. Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

[50] 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.

[51] Z. Zhang, X. Ji, and Y. Zhou. Almost optimal batch-regret tradeoff for batch linear contextual bandits, 2021.