Written Preliminary Examination Literature Review

Peter P. Norwood
Department of Statistics, North Carolina State University

Abstract

The exploration-exploitation trade-off is a fundamental problem in online learning and sequential experimentation. This trade-off describes the choice to exploit accruing information to gain some immediate reward or to explore the system in search of the optimal action. Special consideration is required when data are expensive since exploration can be costly or unethical. In this article, we (a) cover sequential decision making fundamentals like Multi-Armed Bandits and Markov Decision Processes, (b) discuss experimental design work that helps quantify information-gain in an experiment, and (c) review modern methodology that carefully considers if, when, and how to explore. A companion simulation study compares the modern methodology with classical methods in contextual-bandit scenarios. The review and simulation study highlight the opportunity to blend Reinforcement Learning and classical Design of Experiments to build more data-efficient sequential experimentation methods that simultaneously learn and optimize.

Keywords: Sequential Decision Making, Reinforcement Learning, Design of Experiments

1 Introduction

Sequential decision making in the face of uncertainty describes a variety of scenarios ranging from choosing a restaurant to searching for efficacious treatments in clinical trials. The fundamental trade-off is whether to exploit the previous information to choose an action that appears to be favorable or to explore other actions in search of the optimum. This review considers the statistical framework for sequential decision making and surveys methodology examining the exploration-exploitation trade-off. Particular attention is given to methodology that focuses on if, when, and how to explore.

Reinforcement Learning (RL) is a subfield of Machine Learning and Artificial Intelligence that formalizes sequential decision making. In their classic book on RL, Sutton and Barto (2018) first describe RL as learning what to do to maximize a numeric reward over time. Formalizing this process and developing methodology to find optimal strategies has tremendous scientific interest. Applications of RL are found in clinical trials (Williamson and Villar, 2019; Korn and Freidlin, 2017; Wathen and Thall, 2017; Berry, 2012), mobile health (mHealth) (Lei et al., 2017; Luckett et al., 2020; Tomkins et al., 2020), hiring practices (Li et al., 2020; Schumann et al., 2019), online advertising (Zeng et al., 2016; Nguyen-Thanh et al., 2019), and elsewhere.

Design of Experiments (DOE) is a subfield of Statistics as old as the field itself. Seminal contributions to science in the 20th century involve DOE work (Fisher, 1935; Snedecor and Cochran, 1937). A major topic of interest in DOE is optimal designs, first considered by Smith (1918). Optimal designs seek to maximize the information collected in an experiment given its constraints. Like RL, the ultimate goal of DOE is to maximize reward. With maximal information from an experiment, one can make the most certain and accurate decisions to produce higher reward.

In RL, heuristics are often used to determine if, when, and how to explore. Since DOE examines how best to explore, there is an opportunity to combine aspects from both fields to create more data-efficient RL algorithms. This work is important to research where data is expensive and exploration can be costly or unethical.

In this review, section 2 formalizes sequential decision making by covering Multi-Armed Bandits (MABs) and Markov Decision Processes (MDPs). Brief overviews of the methods to identify optimal strategies, Q and V-learning, conclude section 2. Section 3 provides a high-level overview of DOE topics that quantify information-gain. Modern methodology in sequential experimentation to balance exploration and exploitation is covered in section 4. The article concludes with a discussion about future research in section 5.

2 Sequential Decision Making Fundamentals

2.1 Multi-Armed Bandit Problems

A widely applicable sequential decision making problem is the MAB problem. In a MAB, the *agent*, or decision-maker, repeatedly makes the same decision: which of the K *actions* to choose (Sutton and Barto, 2018). The terms action and arm are used interchangeably when discussing MABs. The goal is to maximize cumulative *reward*, which is a function measuring the "goodness" of an action. Often, constraints like cost or number of total actions one can take need to be considered. A special case of MABs is *contextual bandits*. In contextual bandits, new subjects enter with their own context and an action must be chosen for them. The reward distribution depends on both the context and the action. An example of this is a clinical trial where the context is patient characteristics (e.g., age, BMI, etc.) and the action is the treatment.

MABs can be viewed as sequential experiments where the actions are experimental interventions and the goal is to find effective interventions. This framing is one reason why MABs are of such interest to statisticians, notably in the subfield of adaptive clinical trials (Williamson and Villar, 2019; Korn and Freidlin, 2017; Wathen and Thall, 2017; Berry, 2012).

Many heuristics exist to balance exploitation with exploration. We provide a high-level overview of three popular ones: ϵ-Greedy, Thompson sampling, and Upper Confidence Bound (UCB). These strategies to selection actions can be referred to as *policies* or *bandit algorithms*. In section 4 we discuss more sophisticated policies for MABs.

We use the following example to aid explanation. We model the t-th outcome with a linear model: $Y_t = \phi(\mathbf{X}_t, A_t)^T\theta^* + \varepsilon_t$. The reward is simply the outcome, where higher values are coded to be more favorable. Here, $\phi(\cdot)$ is a function that creates a linear relationship between the context $\mathbf{X}_t$, discrete action $A_t \in (1, \ldots, K)$, and the mean parameter vector $\theta^* \in \mathbb{R}^p$. Let $\widehat{\theta}_t$ be the ordinary least squares (OLS) estimator after t observations.

The first approach is ϵ-Greedy. Here, the policy is to explore with probability ϵ and choose the greedy (estimated optimal) action with probability $1 - \epsilon$. For the linear model example, the t-th action is selected as:

$$P(A_t = a) = \begin{cases} 1 - \epsilon, & \text{if } a = \arg\max_{a' \in (1,\ldots,K)} \phi(\mathbf{X}_t, A_t = a')^T \widehat{\theta}_{t-1} \\ \epsilon/(K-1), & \text{otherwise.} \end{cases} \tag{2.1}$$

The exploration probability need not be stationary. If ϵ decreases to zero with t, then ϵ-Greedy converges to a greedy policy. Similarly, an explore-then-commit strategy starts with a period of random exploration, then switches to a greedy policy after t_0 subjects

(Yekkehkhany et al., 2019).

Rather than exploring randomly with some probability, Thompson sampling randomizes according to the probability that a action is optimal. This approach dates back to Thompson (1933) and is now widely popular, notably with adaptive clinical trials (Thall and Wathen, 2007; Wathen and Thall, 2017). Typically, Thompson sampling is a Bayesian approach. Based on the posterior distribution of θ,

$$P(A_t = a) = \mu \left[P \left\{ a = \underset{a' \in (1,\ldots,K)}{\arg\max}\, \phi(\mathbf{X}_t, A_t = a')^T \theta \right\} \right], \tag{2.2}$$

where μ is some function relating the randomization probability to the posterior probability. One example is to threshold the randomization probabilities to some maximum and minimum; e.g., if the posterior probability is 0.95, then then randomization probability is 0.90 (Zhang et al., 2020). Russo et al. (2017) provide an overview of Thompson sampling and its many applications.

The next approach is UCB, which selects the action with the highest upper confidence bound for the mean reward. An upper confidence bound is defined as the estimated reward plus some measure of uncertainty. A UCB policy for the linear model is,

$$A_t = \underset{a' \in (1,\ldots,K)}{\arg\max}\, \phi(\mathbf{X}_t, A_t = a')^T \widehat{\theta} + Z_\alpha \left\{ \phi(\mathbf{X}_t, A_t = a')^T \widehat{\Sigma}_t \phi(\mathbf{X}_t, A_t = a') \right\}^{1/2}, \tag{2.3}$$

where Z_α is the $1-\alpha$ quantile of the standard normal distribution and $\widehat{\Sigma}_t$ is the OLS estimator for the covariance matrix of $\widehat{\theta}_t$. Sutton and Barto (2018) suggest using $c\left\{t/N_t(a)\right\}^{1/2}$ as the measure of uncertainty, where c is some constant and $N_t(a)$ is the number of times action a has been selected.

A popular way to compare different MAB policies is through their *regret*, the expected

difference in reward from the true optimal action, and the action chosen. MAB policies that provide lower cumulative regret after T actions are more favorable. For the linear model, the cumulative regret is,

$$Regret(T) = \sum_{t=1}^{T} \max_{a' \in (1,\ldots,K)} \phi(\mathbf{X}_t, A_t = a')^T \theta^* - \phi(\mathbf{X}_t, A_t)^T \theta^*. \tag{2.4}$$

Regret bounds describe the asymptotic behavior of regret for different bandit policies. Analytical bounds on $Regret(T)$ can be derived and used to compare different policies. A policy is said to *rate-optimal* if it achieves an asymptotic minimal regret bound. For more information on regret bounds and rate optimality, see Lai and Robbins (1985); Slivkins (2019).

2.2 Markov Decision Processes

While MABs describe a specific sequential decision making problem, MDPs provide us the general notation and setup to formalize sequential decision making. We base this overview on both Alagoz et al. (2009) and Sutton and Barto (2018). In this section, we consider discrete-time MDPs, where decisions are made on discrete intervals (e.g., after the previous week's sales data is collected). For a more detailed summary of MDPs and their applications, see John et al. (1989).

In an MDP, the agent has $t = 0, 1, 2, \ldots$ decisions to make. At decision point t (sometimes referred to as a decision epoch), the agent knows the *state* $S_t \in \mathcal{S}$, which is all relevant information to the decision process. Based on S_t, the agent selects an action $A_t \in \mathcal{A}(s)$. The set $\mathcal{A}(s)$ is the set of possible actions based on the state s. Based on A_t and S_t, $R_t(a_t, s_t) \in \mathcal{R}$ is the reward, where $\mathcal{R}$ is the support of the reward's distribution.

The *dynamics* of the system is the following probability:

$$p_t(s', r|s, a) \triangleq P\left(S_{t+1} = s', R_t = r \mid S_t = s, A_t = a\right) \tag{2.5}$$

This distribution describes the probability of the state and reward based on the previous state and action. Based on the dynamics, we define the *transition probability* as $p_t(s'|s, a) \triangleq P\left(S_t = s' \mid S_{t-1} = s, A_{t-1} = a\right)$. The transition probability is the probability of the the state at decision point t, conditional on the previous state and action. It is quick to see that $p_t(s'|s, a) = \int_{r\in\mathcal{R}} p_t(s', r|s, a)dr$. The *Markov assumption* (also defined as the *Markov property*) states that dynamics are fully described by S_t and A_t. Conditional on (S_t, A_t), the dynamics are independent of all previous information.

The goal of sequential decision making is to maximize the cumulative reward. We define $G_t \triangleq \sum_{k\geq 0} \gamma^k R_{t+k+1}$ as the *return*, the sum of rewards after t. Here $\gamma \in [0, 1]$ is the *discounting factor* which discounts rewards over time. Discounting is especially popular in *infinite-horizon* MDPs ($T = \infty$), but can also be used in *finite-horizon* MDPs ($T < \infty$).

Like bandits, a policy, π, is the strategy for selecting actions based on the state. We define $\pi(a|s) \triangleq P^\pi(A_t = a|S_t = s)$, the probability we select $A_t = a$ given $S_t = s$ if we are following the policy π. Naturally, we are interested in the return if we follow a certain π. This introduces the *state-value function* for a policy, which is the expected return if we follow π after t, given the current state s: $V^\pi(s) \triangleq \mathbb{E}^\pi\left(G_t \mid S_t = s\right)$. Similarly, the *action-value function* for a policy π is the value of taking action a, then following π: $q^\pi(s, a) \triangleq \mathbb{E}^\pi\left(G_t \mid S_t = s, A_t = a\right)$.

A policy is the *optimal policy*, π_{opt}, if and only if $V^{\pi_{opt}}(s) \geq V^\pi(s)$ for any other π and all $s \in \mathcal{S}$. Similarly, a policy is optimal if and only if $q^{\pi_{opt}}(s, a) \geq q^\pi(s, a)$.

For finite MDPs (not to be confused with finite-horizon MDPs), the sets $\mathcal{S}_t, \mathcal{A}_t$, and

$\mathcal{R}_t$ all have finite cardinality. For finite MDPs, the *Bellman equation* (2.6) is useful for describing $V^{\pi}(s)$:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r \mid, a) \left\{ r + \gamma V^{\pi}(s') \right\}. \tag{2.6}$$

The Bellman equation is also useful for describing $V^{\pi_{opt}}$, specifically the *Bellman optimality equation* (2.7):

$$V^{\pi_{opt}}(s) = \max_{a \in \mathcal{A}(s)} \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r \mid s, a) \left\{ r + \gamma V^{\pi_{opt}}(S_{t+1}) \right\}. \tag{2.7}$$

See Sutton and Barto (2018) for these derivations.

2.3 Q-Learning

A popular approach for learning an optimal policy is *Q-learning*. Clifton and Laber (2020) describe Q-learning as the class of methods that focus on optimizing the *Q-function*. At decision point t, the Q-function models the return of each possible action, given the decision-maker chooses the optimal action at all subsequent decision points. The authors mention Bellman (1957), Watkins and Dayan (1992), and Murphy (2005) as seminal works in Q-learning.

For finite-horizon decision problems, Q-learning finds an optimal policy through backward-induction of the Q-functions. At the final decision point T, the Q-function is $Q_T(s, a) \triangleq \mathbb{E}\left(R_T \mid S_T = s, A_T = a\right)$. Then, for $t < T$, the Q-function is:

$$Q_t(s, a) \triangleq \mathbb{E}\left\{ R_t + \max_a \gamma Q_{t+1}(S_{t+1}, a) \mid S_t = s, A_t = a \right\}. \tag{2.8}$$

Sutton and Barto (2018) define $Q_t(\cdot)$ as the *learned* action-value function. The learned action-value function approximates the optimal action-value function, $q^{\pi_{opt}}(s, a)$. Given a class of possible Q-functions $\{\mathcal{Q}_t\}_{t=1}^T$, we can estimate the optimal Q-functions by minimizing the squared loss (or some other loss function):

$$\widehat{Q}_T = \underset{Q_T \in \mathcal{Q}_T}{\arg\min} \sum_{i=1}^{n} \{R_{T,i} - Q_T(S_{T,i}, A_{T,i})\}^2, \tag{2.9}$$

$$\ldots$$

$$\widehat{Q}_t = \underset{Q_t \in \mathcal{Q}_t}{\arg\min} \sum_{i=1}^{n} \left[\left\{ R_{t,i} + \max_a \widehat{Q}_{t+1}(S_{t+1,i}, A_{t+1,i} = a) \right\} - Q_t(S_{t,i}, A_{t,i}) \right]^2, \tag{2.10}$$

where i denotes the i-th subject in the data. Then, the optimal policy at time t is $\widehat{\pi}_{t,opt} \triangleq \arg\max_{a \in \mathcal{A}_t(s)} \widehat{Q}_t(S_t = s, A_t = a)$ and $\widehat{\pi}_{opt} = (\widehat{\pi}_{1,opt}, \ldots, \widehat{\pi}_{T,opt})$.

Q-learning in the finite-horizon context is popular in precision medicine literature. It is used to estimate optimal treatment regimes (Tsiatis et al., 2020; Schulte et al., 2014) and to make inference on optimal treatment regimes (Song et al., 2015; Chakraborty et al., 2013; Laber et al., 2014). Q-learning also serves as the basis for developing interpretable, list based policies (Zhang et al., 2018), adaptive randomization procedures for Sequential Multiple Assignment Randomized Trials (SMARTs) (Cheung et al., 2015), and sample-size calculations for SMARTs (Rose et al., 2019).

Q-learning can be extended to the infinite-horizon case. Given data with n subjects and up to T time points, Ertefaie and Strawderman (2018) propose positing one linear model

$Q(s, a, \theta) = \phi(s, a)^T\theta$ and solving the following estimating equation:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\left[\left\{R_{t,i} + \gamma \max_a \phi(S_{t+1,i}, A_{t+1,i} = a)^T\theta\right\} - \phi(S_{t,i}, A_{t,i})^T\theta\right]\phi(S_{t,i}, A_{t,i})\right) = 0.$$

The solution $\widehat{\theta}_n$ can then be used to determine an optimal policy. By leveraging the Markov property, the Q-function only depends on preceding information, so this method is tractable in the infinite-horizon case.

Infinite-horizon Q-learning is often used in online learning. In this context, decision-makers have the choice to select the optimal action or to explore. Naturally, exploration methods discussed in section 2.1 are used. For example, Clifton and Laber (2020) present a simple algorithm to blend Thompson sampling with Q-learning and cite other examples of exploration within online Q-learning.

2.4 V-Learning

An alternative to Q-learning in the infinite-horizon case is V-learning, proposed by Luckett et al. (2020). This approach is first proposed for mHealth applications where actions can be taken anywhere from a few times times a day to every minute. Like Q-learning, this approach also is applicable for both offline and online learning.

We first consider the offline case. Assume the data generating process satisfies the Markov property and typical causal inference assumptions (ignorability, consistency, positivity) hold. Let $\omega_t(a, s) = P(A_t = a \mid S_t = s)$ be the propensity score, which can be known or estimated from the data. Note the propensity is different than $\pi(a \mid s)$, the probability *under the policy*. Then, let $V^\pi(s, \theta)$ be some model for $V^\pi(s)$, the state-value function, and

$\nabla_\theta V^\pi(s,\theta)$ be its gradient. Based on this, consider the following estimating equation:

$$\Lambda_n^\pi(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\frac{\pi(A_{t,i},S_{t,i})}{\omega(A_{t,i},S_{t,i})}\left\{R_{t,i}+\gamma V^\pi(S_{t+1,i},\theta)-V^\pi(S_{t,i},\theta)\right\}\nabla_\theta V^\pi(S_{t,i},\theta). \quad (2.11)$$

Then, $\widehat{\theta}_n = \arg\min_\theta\left\{\Lambda_n^\pi(\theta)^T\Omega\Lambda_n^\pi(\theta)+\vartheta_n\mathcal{L}(\theta)\right\}$, where Ω is some positive definite matrix, $\mathcal{L}$ is a penalty function, and ϑ_n is a tuning parameter. The estimated *value* of a policy π is $\widehat{V}_n^\pi = \int V^\pi(s,\widehat{\theta}_n)ds$. The estimated optimal policy is then $\widehat{\pi} = \arg\max_{\pi\in\Pi}\widehat{V}_n^\pi$, where Π is the class of potential policies. As the authors note, the essence of V-learning is using the estimating equation 2.11 to estimate the value of a policy and then to maximize over possible policies.

Online V-learning is similar to online Q-learning. Rather than always following the estimated optimal policy, some exploration strategy is used to learn the system. An important feature of V-learning is the ability to consider randomized policies, particularly when deterministic policies are of scientific interest. As the authors note, using ϵ-Greedy introduces randomness into a deterministic policy and induces a randomized policy. For online learning with randomized policies, the estimating equation 2.11 can be replaced with,

$$\Lambda_n^\pi(\theta) = n^{-1}\sum_{i=1}^{n}\sum_{v=1}^{t}\frac{\pi(A_{v,i},S_{v,i})}{\widehat{\pi}_{n,v-1}(A_{v,i},S_{v,i})}\left\{R_{v,i}+\gamma V^\pi(S_{v+1,i},\theta)-V^\pi(S_{v,i},\theta)\right\}\nabla_\theta V^\pi(S_{v,i},\theta), \quad (2.12)$$

where $\omega(\cdot)$ is replaced with $\widehat{\pi}_{n,t}(a,S_{t+1})$, the probability of action assignment according to the estimated policy at time t from n subjects.

Since V-learning only requires modeling the policy and the state-action function, it is better suited than Q-learning for scenarios with small sample sizes and many decision points (mHealth falls into this category). Another key advantage is the ability to estimate

a randomized decision rule; as we have shown, this can be useful in online learning.

3 Design of Experiments

Since it is a fundamental subfield of Statistics, vast literature exists for DOE. In this section, rather than a high-level overview of the field, we will highlight some important ideas that are used for sequential decision making problems.

In this section, we frequently use the linear model example from section 2.1 to aid the explanation. With this linear model, under typical assumptions, the approximate large-sample covariance matrix of $\widehat{\theta}_t$ is proportional to $\mathbf{M}_t^{-1}$, where $\mathbf{M}_t = \sum_{i=1}^{t} \phi(\mathbf{X}_i, A_i)\phi(\mathbf{X}_i, A_i)^T$. Similar results exist for maximum likelihood estimators (MLE) in non-linear models.

3.1 Optimal Experimental Design

Optimal experimental design considers how to choose design points to achieve some goal (e.g., maximum reduction in uncertainty). It can be useful to define a *design* as follows (Fedorov, 1972; Chaloner and Verdinelli, 1995). Let $(\mathbf{X}_i, A_i) \in \mathcal{X}$ be a unique design point, where $\mathcal{X}$ is the set of all possible design points. After t subjects have completed the experiment, each $(\mathbf{X}_i, A_i)$ has been chosen t_i times each and there exist t^* unique points. Based on this, we can define $\eta_i = t_i/t$ and $\mathbf{M}_t = t\sum_{i=1}^{t^*} \eta_i \phi(\mathbf{X}_i, A_i)\phi(\mathbf{X}_i, A_i)^T$. Then, the design η is the allocation of design points according to each η_i. Furthermore, η can be viewed as a probability measure on $\mathcal{X}$.

Since $\mathbf{M}_t^{-1} \in \mathbb{R}^{p\times p}$, we denote $\psi(\eta)$ as the scalar optimality criterion used to measure the utility from a design. Commonly, $\psi(\eta)$ is a function relating to the variability of $\widehat{\theta}_t$, with smaller values indicating more reduction in variability. A design, η^{opt}, is an optimal design, if and only if $\psi(\eta^{opt}) \leq \psi(\eta)$ for all $\eta \in \mathcal{X}$.

Popular choices for $\psi(\eta)$ include the following (Fedorov and Leonov, 2019). The D-criterion is the determinant of $\mathbf{M}_t^{-1}$. The E-criterion is the maximum eigenvalue of $\mathbf{M}_t^{-1}$. The A-criterion is $\text{tr}\left\{\mathbf{A}\mathbf{M}_t^{-1}\right\}$, where $\mathbf{A}$ is the *utility matrix* and can be customized to fit the problem. A special case of the A-criterion is the C-criterion, which minimizes the variance of one linear combination, $\mathbf{c}^{\mathbf{T}}\theta^*$. Naturally, the C-criterion is $\mathbf{c}^T\mathbf{M}_t^{-1}\mathbf{c}$. These criteria generalize easily to non-linear models when the information matrix is used instead of $\mathbf{M}_t^{-1}$.

Often, the optimal design is decided before an experiment begins. For example, consider simple linear regression where $Y_t = \theta_0^* + \theta_1^* X_t + \varepsilon_t$ and $X_t \in [-1, 1] \cap \mathbb{R}$. With an even T total subjects, it is easy to verify that the D-optimal design is half the design points at -1 and the other half at 1: $\eta_1 = T/2, X_1 = -1, \eta_2 = T/2, X_2 = 1$. Software to find optimal designs in more complicated situations is available, notably the `OptimalDesign` package in `R` and `PROC OPTEX` in `SAS` (Radoslav Harman, 2019; SAS, 2014).

For sequential optimal designs, the next design point is chosen according to the optimality criterion. In fact, Wynn (1970) shows that sequential D-optimal designs and D-optimal designs are equivalent asymptotically. Sometimes the entire design point can be chosen, other times just the action. For example, in linear model example, the sequential D-optimal action is $\arg\min_{a\in(1,\dots,K)} \det\left\{\mathbf{M}_{t-1} + \phi(\mathbf{X}_t, A_t = a)\phi(\mathbf{X}_t, A_t = a)^T\right\}^{-1}$.

3.2 Bayesian Experimental Design

As discussed in in Chaloner and Verdinelli (1995), Bayesian experimental design focusing on choosing η to maximize an expected utility function $U(\eta) = \int U(\theta, \eta, \mathbf{Y}) p(\theta|\mathbf{Y}, \eta) p(\mathbf{y}|\eta) d\theta d\mathbf{Y}$. A popular example of the utility function is the *Kullback-Leibler divergence* (KL-divergence)

between the prior ($\mathcal{P}_0$) and posterior ($\mathcal{P}_1$) distribution of θ:

$$D_{KL}(\mathcal{P}_0||\mathcal{P}_1) = \int \log\left\{\frac{p(\theta \mid \mathbf{Y}, \eta)}{p(\theta)}\right\} p(\mathbf{Y}, \theta \mid \eta) d\theta d\mathbf{Y}. \tag{3.1}$$

Based on this, we can define $U(\eta)$ as the expected *Shannon Information*:

$$U(\eta) = \int \log\left\{p(\theta \mid \mathbf{Y}, \eta)\right\} p(\mathbf{Y}, \theta \mid \eta) d\theta d\mathbf{Y}. \tag{3.2}$$

Since $p(\theta)$ does not depend on η, maximizing KL-divergence with respect to η is equivalent to maximizing Shannon Information. Intuitively, since KL-divergence measures the "distance" between two distributions, maximizing KL-divergence between the prior and posterior relates to maximizing information from the design.

To gain more intuition behind Bayesian optimal design, consider the example from Chaloner and Verdinelli (1995). Suppose the prior distribution for θ is $MVN(\theta_0, \mathbf{R})$ and $Y_t \mid \theta, \mathbf{X}_t, A_t \sim MVN\left\{\phi(\mathbf{X}_t, A_t)^T\theta, \sigma^2\right\}$. Then, the posterior distribution of θ is p-dimensional multivariate normal with covariance matrix proportional to $(t\mathbf{M}_t + \mathbf{R})^{-1}$. In this setup, the Shannon Information is:

$$U(\eta) = -\frac{p}{2}\left\{1 + \log(2\pi)\right\} + \frac{1}{2}\log\det\left\{\sigma^{-2}(t\mathbf{M}_t + \mathbf{R})\right\}.$$

Hence, the design that maximizes the KL-divergence and Shannon Information maximizes $\det(t\mathbf{M}_t + \mathbf{R})$, providing a clear connection to the D-optimal design.

Two other definitions of interest are the *entropy* and *mutual information*. For a probability distribution P, the entropy is $H(P) = -\sum_{x\in\mathcal{X}} P(x)\log\{P(x)\}$. For two random variables X_1, X_2, the mutual information is the KL-divergence between their joint distribution, $p(X_1, X_2)$, and the product of their marginal distributions, $p(X_1)p(X_2)$. If two

random variables are independent, then they share no information, hence their mutual information is zero.

4 Modern Methodology

4.1 Adaptive Optimization and D-Optimum Experimental Design

An important paper considering the balance between learning and optimization that uses optimal experimental design is Pronzato (2000). We are interested in choosing actions $\mathbf{A} \in \mathcal{X}$ to maximize a parametric reward function $R(\mathbf{A}, \theta^*)$. Pronzato considers the linear model setting where the p-dimensional action is the entire design: $Y_t = \phi(\mathbf{A}_t)^T\theta^* + \varepsilon_t$. We assume $\phi(\cdot)$ is continuous in $\mathbf{A}$ and ε_t is a martingale difference sequence with respect to an increasing sequence of σ-fields $\{F_t\}$, so $E(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0$ for all t. Let $\widehat{\theta}_t$ be the OLS estimator after t observations and $\mathbf{M}_t = \sum_{i=1}^{t} \phi(\mathbf{A}_t)\phi(\mathbf{A}_t)^T$ (similar to the previous linear model example). We assume a burn-in period of t_{min} of simple randomization, so $\mathbf{M}_{t_{min}}$ is positive definite and $\widehat{\theta}_t$ can be computed.

For $t > t_{min}$, to simultaneously learn θ^* and choose $\mathbf{A}$ to maximize the reward, Pronzato proposes choosing

$$\mathbf{A}_{t+1} = \arg\max_{\mathbf{A}\in\mathcal{X}} R(\mathbf{A}, \widehat{\theta}_t) + \frac{\alpha_t}{t} d_t(\mathbf{A}), \tag{4.1}$$

where $d_t(\mathbf{A}) = \phi(\mathbf{A})^T\mathcal{I}_t^{-1}\phi(\mathbf{A})$ and $\mathcal{I}_t = \mathbf{M}_t/t$. This represents a compromise between the greedy choice and the sequential D-optimal choice; the equivalence $\arg\max_{\mathbf{A}\in\mathcal{X}} d_t(\mathbf{A}) = \arg\min_{\mathbf{A}\in\mathcal{X}} \det \mathbf{M}_{t+1}^{-1}$ is a result of the matrix determinant lemma.

Pronzato considers how to choose α_t such that $\widehat{\theta}_t \xrightarrow{a.s.} \theta^*$ and $t^{-1}\sum_{i=1}^{t} R(\mathbf{A}_i, \theta^*) \xrightarrow{a.s.}$

$\max_{\mathbf{A}\in\mathcal{X}} R(\mathbf{A},\theta^*)$ as $t\to\infty$.

Theorem 1. *If $\phi(\mathbf{A})$ and $R(\mathbf{X},\theta)$ are continuous their arguments, $\mathcal{X}$ is compact, the second moment of ε_t is bounded, and α_t is chosen such that $(\alpha_t/t)\log(\alpha_t)$ is decreasing and $\alpha_t/(\log t)^{1+\delta}$ is increasing to ∞ for some positive δ, then $\widehat{\theta}_t \xrightarrow{a.s.} \theta^*$ and $t^{-1}\sum_{i=1}^{t} R(\mathbf{A}_i,\theta^*) \xrightarrow{a.s.} \max_{\mathbf{A}\in\mathcal{X}} R(\mathbf{A},\theta^*)$ as $t\to\infty$.*

This shows under mild assumptions for the model, design space, and reward function, a large class of α_t exists such that our OLS estimator converges to the truth and our design maximizes reward.

The idea of choosing $\mathbf{A}_{t+1} = \max_{\mathbf{A}\in\mathcal{X}} R(\mathbf{A},\widehat{\theta}_t) + d_t(\mathbf{A})$ where larger values of $d_t(\mathbf{A})$ correspond to more reduction in the variability of $\widehat{\theta}_t$ had been studied previously (Aström and Wittenmark, 1995; Wittenmark, 1975; Wittenmark and Elevitch, 1985). Pronzato improves on those methods by rigorously showing what type of penalty leads to convergence in parameter estimates and reward. Pronzato mentions adaptive control systems like chemical reactors and temperature regulation as applications of this work. This work also applies to other areas like mHealth and online advertising where self tuning algorithms that simultaneously learn parameters and maximize reward are useful.

4.2 Learning to Optimize Via Information-Directed Sampling

The next paper we discuss is Russo and Van Roy (2018), which introduces *Information-Directed Sampling* (IDS). IDS is a general design principle for online optimization problems like MABs. With multiple $A_t \in \mathcal{A}(s)$ to choose from, IDS quantifies the information gain from selecting A_t. Based on Bayesian experimental design, they define information gain as the mutual information between the true optimal action and the next observation. Then, actions are sampled based on the ratio of their squared expected regret and their

information gain – the cost per bit of information gained. The policy with the lowest cost per bit of information gained is the selected policy.

Russo and Van Roy consider a range of Bayesian online optimization problems where $Y_t(a) \in \mathcal{Y}$ is the random outcome at decision-point t for an action a. The state at decision t is all accruing information from the experiment, $S_t = \{A_1, Y_1(A_1), \ldots, A_{t-1}, Y_1(A_{t-1})\}$. We define $\pi_t(a \mid s) = P^{\pi}(A_t = a \mid S_t = s)$. The goal is to find a policy that minimizes the the expected regret. Let A^* denote the optimal action, which yields zero regret.

Let $p_t(a) = P(A^* = a \mid S_t)$ denote the posterior distribution of A^*. Russo and Van Roy define information gain as the mutual information of $p_t(a)$ and $p_{t+1}(a)$. This can be expressed as the expected reduction in entropy of the posterior distribution of A^* due to observing $Y_t(a)$: $g_t(a) \triangleq \mathbb{E}\left\{H\left\{p_t(a)\right\} - H\left\{p_{t+1}(a)\right\} \mid S_t, A_t = a\right\}$. Additionally, let $\Delta_t(a) \triangleq \mathbb{E}\left\{R_t(A^*) - R_t(a)\right\}$ denote the instantaneous regret if $A_t = a$.

Let $\pi_t(\mathcal{S}_t) \in \mathcal{D}(\mathcal{A})$ denote a probability distribution over $\mathcal{A}$, where $\mathcal{D}(\mathcal{A})$ is the set of all probability distributions over $\mathcal{A}$. For a policy π, the expected information gain is $g_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a|S_t) g_t(a)$. Similarly, $\Delta_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a|S_t)\Delta_t(a)$ denotes the expected regret if π is followed. The IDS policy minimizes the ratio of squared regret and information gain. That is,

$$\pi_t^{IDS} \in \underset{\pi \in \mathcal{D}(\mathcal{A})}{\arg\min} \left\{ \Psi_t(\pi) \triangleq \frac{\Delta_t(\pi)^2}{g_t(\pi)} \right\}, \tag{4.2}$$

where $\Psi_t(\pi)$ is the *information ratio*. A policy with a small information ratio could have a low regret, high information gain, or ideally a combination of the two.

Computing $\Psi_t(\pi)$ is problem specific and likely depends on the form of the posterior distribution, numerical integration, or approximation. Russo and Van Roy provide algorithms to compute the information ratio in beta-Bernoulli, independent Gaussian, and linear ban-

dits. In the simulation section, we propose two algorithms to calculate the information ratio for a contextual linear bandit.

Russo and Van Roy argue that IDS improves on other heuristics like Thompson sampling and UCB because (a) it is able to collect information on other actions through the selected action, (b) it can collect feedback that enables higher reward later, and (c) it avoids selecting actions which fail to provide information. They supply practical examples where Thompson sampling or UCB fails to satisfy (a), (b), or (c). Additionally, simulation experiments show IDS outperforms the state of the art methods (including Thompson sampling and UCB) in beta-Bernoulli, independent Gaussian, and linear bandit scenarios. In all three scenarios, IDS has lower mean regret compared to all other methods considered.

4.3 Mostly Exploration-Free Algorithms for Contextual Bandits

The previous two works mostly consider how to explore. Now, we discuss Bastani et al. (2017), which studies *if* exploration is necessary in contextual bandits. They show if sufficient randomness exists in the context, then a greedy algorithm has enough *natural exploration* to outperform algorithms with forced exploration. Due to this, the *Greedy-First* (GF) algorithm is presented. GF begins as a greedy algorithm. After selecting A_t, GF checks to see if sufficient information is gathered from the greedy algorithm. If yes, the next action is chosen greedily. If no, GF switches to a different algorithm with forced exploration.

A key definition is that a random variable $\mathbf{Z} \in \mathbb{R}^p$ has *covariate diversity* if there exists a constant λ_0 such that for every $\mathbf{u} \in \mathbb{R}^p$,

$$\lambda_{min}\left(\mathbb{E}_{\mathcal{Z}}\left[\mathbf{Z}\mathbf{Z}^T\ \mathbb{I}\{\mathbf{Z}^T\mathbf{u} \geq 0\}\right]\right) \geq \lambda_0, \tag{4.3}$$

where λ_{min} denotes the minimum eigenvalue, and $\mathbb{I}(\cdot)$ is the indicator function.

The problem formula is the same contextual bandit previously discussed. Here, the contexts $\mathbf{X}$ are i.i.d. from some distribution. In the case where $A_t \in \{0, 1\}$, Bastani et al. demonstrate when $\mathbf{X}$ has covariate diversity, greedy optimization has enough natural exploration to ensure parameter estimates converge to their true values. First, in the two-arm case, notice we can represent $Y_t = \phi(\mathbf{X}_t, A_t)^T\theta^* = (1-A_t)\mathbf{X}_t^T\theta_0^* + A_t\mathbf{X}_t^T\theta_1^*$. Due to this, $\mathbf{M}_t$ can be represented as a block diagonal matrix with $\mathbf{M}_{t,0}, \mathbf{M}_{t,1}$ on the diagonal. Hence, we can decompose the OLS estimator into two groups, $\widehat{\theta}_t = (\widehat{\theta}_{t,0}, \widehat{\theta}_{t,1})$. This formulation plus the next lemma imply that when the context has covariate diversity, each action will be estimated optimal infinitely often.

Lemma 1. *Assume there exists a positive constant x_{max} such that the $p_{\mathbf{X}}$ has no support outside the ball of radius x_{max} and there exists a constant b_{max} such that $||\theta_j||_2 \leq b_{max}$ for all $j \in K$ where $K = dim(\mathcal{A})$. If $\mathbf{X}$ has covariate diversity, then for λ_0 in (4.3) and any $\mathbf{u} \in \mathbb{R}^p, P_{\mathbf{X}}\left\{\phi(\mathbf{X}, A)^T\mathbf{u} \geq 0\right\} \geq \lambda_0/x_{\max}$.*

When taking $\mathbf{u} = \left(\widehat{\theta}_{t,0}, \widehat{\theta}_{t,1}\right)^T$ or $\mathbf{u} = \left(\widehat{\theta}_{t,1}, \widehat{\theta}_{t,0}\right)^T$, we see for all t, there is a non-zero probability that each arm will be the greedy choice when $K = 2$. This implies each action will be selected infinitely often as t approaches infinity.

Based on the previous result, it is shown that $\lambda_{min}\mathbf{M}_{t,0}$ and $\lambda_{min}\mathbf{M}_{t,1}$ both grow linearly with t under greedy optimization. This is fundamental to showing convergence of $\widehat{\theta}_t$, our next result.

Theorem 2. *Assume the conditions for Lemma 1, $Y_t = \phi(\mathbf{X}_t, A_t)^T\theta^* + \varepsilon_t$ where ε_t is σ-subgaussian, $\theta^* \in \mathbb{R}^p$, and $A_t \in \{0, 1\}$. Let $C_2 = \lambda^2/(2p\sigma^2 x_{max})$ and let n be an upper bound on how often an action can be selected at time t. Then, under greedy optimization, for all $\lambda, \chi > 0$, $P\left\{||\widehat{\theta}_{t,i} - \theta_i^*||_2 > \chi \text{ and } \lambda_{min}\mathbf{M}_{t,i} \geq \lambda t\right\} \leq 2pe^{-C_2 t^2\chi^2/n}$.*

Theorem 2 tells us that under common bandit and linear model assumptions, if the context has sufficient randomness, then greedy optimization leads to consistent OLS estimators and is a rate optimal solution. Based on these results, the authors present the GF algorithm. For the first t_0 decisions, GF selects the greedy action. After the $t \geq t_0$ action is selected, if $\lambda_{min}\mathbf{M}_t > \lambda_0 t/4$, then the greedy algorithm continues for $t+1$. If $\lambda_{min}\mathbf{M}_t \leq \lambda_0 t/4$, GF switches to some other bandit algorithm with forced exploration. It is reasonable to run a burn-in period of simple randomization until $\mathbf{M}_t$ is non-singular before starting GF. In GF, t_0 and λ_0 are user defined parameters. The authors suggest using the data-driven $\widehat{\lambda}_0 = (2t_0)^{-1}\lambda_{min}\mathbf{M}_{t_0}$ for λ_0.

The authors show similar convergence guarantees for generalized linear models and provide a GF algorithm in that case. In multiple simulation experiments, the authors show greedy optimization and GF outperform (measured by minimizing regret) variants of UCB, Thompson sampling, and ϵ-Greedy when the context has covariate diversity. When covariate diversity fails, GF has similar performance to the competitors and greedy optimization under-performs.

In this paper, Bastani et al. ponder if we even should explore. They then prove that greedy optimization is preferable to forced exploration in certain contextual bandit scenarios. Since this is not the case in all scenarios, they provide the simple and intuitive GF algorithm that automatically checks to see if exploration is needed. This work has considerable implications for situations where exploration is costly.

5 Conclusion

This review covers the fundamentals of sequential decision making and optimal experimental design, then presents three modern methods which consider if, when, and how to

explore. Plenty of future work comparing novel sequential experimentation methods exists, notably in situations where exploration is expensive and minimizing regret is not the only goal (e.g., clinical trials for potentially life-saving therapies). Additionally, studying different exploration methods in online Q and V-learning is relevant.

For the three proposed methods, immediate future research ideas are the following. For the method proposed in Pronzato (2000), developing data-driven methods to determine α_t that still satisfy the convergence assumptions is of interest. For IDS, generalizing the methodology to consider other definitions of information gain is intriguing. Defining information gain through optimal experimental design is an obvious approach to consider. If covariate diversity is satisfied, looking at the efficiency of estimators based on data from greedy optimization versus other algorithms is particularly interesting and relevant to these high-cost situations.

In the simulation section, we compare these three methods among others in a contextual bandit scenario. We propose a variant of IDS based on D-optimality, address the efficiency of parameter estimates, and examine how well we can assign interventions to new, out-of-experiment subjects based on the information collected through different policies. In this study, the methods that consider if, when, and how to explore perform well minimizing regret, learning parameters, and assigning out-of-experiment subjects effective interventions.

6 BibTeX

References

Oguzhan Alagoz, Heather Hsu, Andrew J. Schaefer, and Mark S. Roberts. Markov decision processes: A tool for sequential decision making under uncertainty. *Medical Decision Making*, 30(4):474–483, 2009. doi: 10.1177/0272989x09353194.

Karl Johan Aström and Björn Wittenmark. *Adaptative control*. Addison Wesley, 1995.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits, 2017.

Richard Bellman. *Dynamic programming*. Princenton University Press, 1957.

Donald Berry. Adaptive clinical trials in oncology. *Nat Rev Clinical Oncology*, 9:199–207, 2012.

Bibhas Chakraborty, Eric B. Laber, and Yingqi Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3): 714–723, 2013. doi: 10.1111/biom.12052.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. ISSN 08834237. URL `http://www.jstor.org/stable/2246015`.

Ying Kuen Cheung, Bibhas Chakraborty, and Karina W. Davidson. Sequential multiple assignment randomized trial (smart) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, 71(2):450–459, 2015. doi: 10.1111/biom.12258. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12258`.

Jesse Clifton and Eric B. Laber. Q-learning: Theory and applications. *Annual Review of Statistics and Its Applications*, 7(1):279–301, Mar 2020.

Ashkan Ertefaie and Robert L Strawderman. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977, 2018. doi: 10.1093/biomet/asy043.

Fedorov and Leonov. *Optimal Design for Nonlinear Response Models*. CRC Press, 2019.

V. Fedorov. *The Theory of Optimal Experiments*. Academic Press, 1972.

R.A. Fisher. *The design of experiments. 1935*. Oliver and Boyd, Edinburgh, 1935.

F. John, J. E. Marsden, and L. Sirovich. *Adaptive Markov Control Processes*. Springer New York, 1989.

Edward L. Korn and Boris Freidlin. Adaptive clinical trials: Advantages and disadvantages of various adaptive design elements. *JNCI: Journal of the National Cancer Institute*, 109 (6), 2017. doi: 10.1093/jnci/djx013.

Eric B. Laber, Daniel J. Lizotte, Min Qian, William E. Pelham, and Susan A. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225–1272, 2014. doi: 10.1214/14-ejs920.

T.l Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

Huitian Lei, Ambuj Tewari, and Susan A. Murphy. An actor-critic contextual bandit algorithm for personalized mobile health interventions, 2017.

Danielle Li, Lindsey R Raymond, and Peter Bergman. Hiring as exploration. Working Paper 27736, National Bureau of Economic Research, August 2020. URL http://www.nber.org/papers/w27736.

Daniel J. Luckett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 115(530): 692–706, 2020. doi: 10.1080/01621459.2018.1537919. URL https://doi.org/10.1080/01621459.2018.1537919.

Susan A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6(37):1073–1097, 2005. URL http://jmlr.org/papers/v6/murphy05a.html.

Nhan Nguyen-Thanh, Dana Marinca, Kinda Khawam, David Rohde, Flavian Vasile, Elena Simona Lohan, Steven Martin, and Dominique Quadri. Recommendation system-based upper confidence bound for online advertising. 2019.

Luc. Pronzato. Adaptive optimization and d-optimum experimental design. *Ann. Statist.*, 28(6):1743–1761, 2000. doi: 10.1214/aos/1015957479. URL https://projecteuclid.org/euclid.aos/1015957479.

Lenka Filova Radoslav Harman. *OptimalDesign: A Toolbox for Computing Efficient Designs of Experiments*, 2019.

Eric J. Rose, Eric B. Laber, Marie Davidian, Anastasios A. Tsiatis, Ying-Qi Zhao, and Michael R. Kosorok. Sample size calculations for smarts, 2019.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. doi: 10.1287/opre.2017.1663.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *CoRR*, abs/1707.02038, 2017. URL http://arxiv.org/abs/1707.02038.

SAS. *SAS/QC 13.2 User's Guide*, 2014.

Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 29(4):640–661, 2014. ISSN 08834237, 21688745. URL http://www.jstor.org/stable/43288503.

Candice Schumann, Zhi Lang, Jeffrey S. Foster, and John P. Dickerson. Making the cut: A bandit-based approach to tiered interviewing. *CoRR*, abs/1906.09621, 2019. URL http://arxiv.org/abs/1906.09621.

Aleksandrs Slivkins. Introduction to multi-armed bandits, 2019.

Kirstine Smith. On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polymonial Function and its Constants and the Guidance They Give Towards a Proper Choice of the Distribution of Observations. *Biometrika*, 12(1-2):1–85, 11 1918. ISSN 0006-3444. doi: 10.1093/biomet/12.1-2.1. URL https://doi.org/10.1093/biomet/12.1-2.1.

George W. Snedecor and William G. Cochran. *Statistical methods*. Iowa State Univ. Press, 1937.

Rui Song, Weiwei Wang, Donglin Zeng, and Michael R. Kosorok. Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 2015. doi: 10.5705/ss.2012.364.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Peter Thall and Kyle Wathen. Practical bayesian adaptive randomization in clinical trials. *European Journal of Cancer*, 43(5):859–866, 2007.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Sabina Tomkins, Peng Liao, Predrag Klasnja, and Susan Murphy. Intelligentpooling: Practical thompson sampling for mhealth, 2020.

Anastasios A. Tsiatis, Marie Davidian, Shannon T. Holloway, and Eric B. Laber. *Dynamic treatment regimes statistical methods for precision medicine*. London, 2020.

J Kyle Wathen and Peter F Thall. A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials*, 14(5):432–440, 2017. doi: 10.1177/1740774517692302. URL `https://doi.org/10.1177/1740774517692302`. PMID: 28982263.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4): 279–292, 1992. doi: 10.1007/bf00992698.

S. Faye Williamson and Sofía S. Villar. A response-adaptive randomization procedure for multi-armed clinical trials with normally distributed outcomes. *Biometrics*, 76(1): 197–209, 2019. doi: 10.1111/biom.13119.

B. Wittenmark and C. Elevitch. An adaptive control algorithm with dual features. *IFAC Proceedings Volumes*, 18(5):587–592, 1985. doi: 10.1016/s1474-6670(17)60623-2.

Björn Wittenmark. An active suboptimal dual controller for systems with stochastic parameters. *Automatic Control Theory and Application*, 3:13–19, 1975.

Henry P. Wynn. The sequential generation of d-optimum experimental designs. *Ann. Math. Statist.*, 41(5):1655–1664, 10 1970. doi: 10.1214/aoms/1177696809. URL https://doi.org/10.1214/aoms/1177696809.

Ali Yekkehkhany, Ebrahim Arian, Mohammad Hajiesmaili, and Rakesh Nagi. Risk-averse explore-then-commit algorithms for finite-time bandits, 2019.

Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. page 2025–2034, 2016. doi: 10.1145/2939672.2939878. URL https://doi.org/10.1145/2939672.2939878.

Kelly W. Zhang, Lucas Janson, and Susan A. Murphy. Inference for batched bandits, 2020.

Yichi Zhang, Eric B. Laber, Marie Davidian, and Anastasios A. Tsiatis. Interpretable dynamic treatment regimes. *Journal of the American Statistical Association*, 113(524): 1541–1549, 2018. doi: 10.1080/01621459.2017.1345743. URL https://doi.org/10.1080/01621459.2017.1345743. PMID: 30774169.