Statistica Sinica Preprint No: SS-2020-0431	
Title	Bandit Theory: Applications to Learning Healthcare Systems and Clinical Trials
Manuscript ID	SS-2020-0431
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0431
Complete List of Authors	Michael Sklar, Mei-Chiung Shih and Philip Lavori
Corresponding Author	Michael Sklar
E-mail	sklarm@stanford.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

BANDIT THEORY: APPLICATIONS TO LEARNING HEALTHCARE SYSTEMS AND CLINICAL TRIALS

Michael Sklar[1], Mei-Chiung Shih[1,2], and Philip Lavori[1]

[1] *Stanford University*

[2] *VA Palo Alto Cooperative Studies Program Coordinating Center*

Abstract: **In recent years, statisticians and clinical scientists have defined two new approaches for studying the effects of medical practice, extending the "gold standard" classical randomized clinical trial to remedy some of its defects, improve its fit to clinical practice, and conform more closely to ethical principles. The Contextual Multi-Armed Bandit provides a natural statistical structure for a Learning Healthcare System, allowing the optimization of patient outcomes by adaptive assignment of treatments while building in experimental strength for accuracy in learning. The Sequential Multiple Assignment Randomized Trial has become the standard for comparing entire Dynamic Treatment Strategies for the management of chronic disease, which more closely match the goals and practice of clinicians. Theory and methods developed by Professor Tze Leung Lai over the course of his career turn out to be of central importance in bringing these two apparently different approaches to bear in efforts to improve clinical**

practice. We review these methods in this article.

Key words and phrases: clinical trials, medical and pharmaceutical statistics, sequential analysis and optimal stopping

1. Introduction

Throughout his career, Professor Tze Leung Lai has made major contributions to the design of randomized clinical trials. We note his creative work with many students and colleagues in group sequential stopping (Lai and Shih, 2004) and adaptive trials (Lai, Lavori, and Liao, 2014; Bartroff, Lai, and Shih, 2012; Lai and Liao, 2012), which have advanced the field of clinical trials. Here we focus on two areas of Lai's work that may have even greater impact on the future of clinical research, as the medical community grapples with the challenges of generating and applying knowledge at point of care in fulfillment of the concept of the "Learning Healthcare System (LHS)" (Chamberlayne, Green, Barer, Hertzman, Lawrence, and Sheps, 1998). "A learning healthcare system is one that is designed to generate and apply the best evidence for the collaborative healthcare choices of each patient and provider; to drive the process of discovery as a natural outgrowth of patient care; and to ensure innovation, quality, safety, and value in health care" (Olsen, Aisner, and McGinnis, 2007). The first branch of Lai's work discussed below deals with methods for incorporating true experimental strength into efforts to explore the comparative effects of different treatments while exploiting what is learned to improve outcomes in patients. The underlying idea of the "multi-armed bandit" (MAB) for clinical decision-making goes back almost a century (Thompson, 1933) and Lai made his foundational contributions to that subject at the start of his career (Lai and Robbins, 1985). Recent work on the contextual generalization of the MAB (CMAB) has brought the idea back to the fore, as a sound theoretical basis for the LHS enterprise. We surveyed the current literature on the use of the CMAB in clinical medicine as part of a demonstration project currently underway in the US Department of Veterans Affairs (VA) Cooperative Studies Program (CSP), and below we discuss the current "state of the art."

The other branch of Lai's work discussed below has been used to further the recent development of a natural framework for defining and comparing Dynamic Treatment Regimes (DTRs), also known as Adaptive Treatment Strategies (ATS), in the management of chronic disease using variants of the sequential multiple assignment randomized trial (SMART) (Lavori and Dawson, 2000, 2004; Murphy, 2005). This is again an attempt to bring experimental rigor to the study of clinical decision-making, taking full account of the inherently dynamic nature of ongoing clinical management of chronic disease.

We hope that the review and discussion in this celebratory paper will illustrate two of the less well-known but no less promising and valuable contributions made by Professor Lai in the course of his career, and perhaps encourage others to take up his ideas and carry them forward, as we have done.

2. The Multi-Armed Bandit Problem

The name "multi-arm bandit" suggests a row of slot machines, which in the 1930's were nicknamed "one-armed bandits." (Presumably the name is inspired by their pull-to-play levers and the often large house edge). For a gambler in an unfamiliar casino, the "multi-arm bandit problem" would refer to a particular challenge: to maximize their expected winnings over a total of T plays, moving between machines as they desire. The distribution of payouts from pulling each arm may be unknown and different for each machine. How should the gambler play? Study of the multi-arm bandit problem and its variants has led to foundational insights for problems in sequential sampling, sequential decisionmaking, and reinforcement learning.

Mathematical analysis of the multi-arm bandit problem has been motivated by medical applications since Thompson (1933) with different medical treatments playing the role of bandit machines. Subsequent theory has found wide application across disciplines including finance, recommender systems, and telecommunications (Bouneffouf and Rish, 2019). According to Whittle (1979) the bandit problem was considered by Allied scientists in World War II, but it "so sapped [their minds] that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage." It was Lai and Robbins (1985) who gave the first tractable asymptotically efficient solution.

Given a set of arms $k \in 1, ...K$, Lai and Robbins frame the question: How should we sample y_1, y_2,... sequentially from the K arms in order to achieve the greatest possible expected value of the sum $S_T = y_1 + ... + y_T$ as $T \to \infty$? They model each sample from arm k as an independent draw from a population Π_k from a family of densities f_{θ_k} indexed by parameter θ_k. Then, they formalize the space of (possibly random) strategies $\phi \in \Phi$, defining ϕ to be an *adaptive allocation rule* if it is a collection of random variables which makes the arm selection at each timestep, $\boldsymbol{\phi} := (\phi_1, \phi_2, ...\phi_T)$. Thus each ϕ_t is a random variable on $\{1, ..., K\}$ where the event $\{\phi_t = k\}$ ("arm k is chosen at time t") belongs to the σ-field generated by prior decisions and observations $(\phi_1, x_1, \phi_2, x_2, ...\phi_{t-1}, x_{t-1})$. In this framework, Lai and Robbins (1985) define the *cumulative regret* of an adaptive allocation rule which measures the strategy's expected performance against the best arm, equivalent to

$$R_T(\boldsymbol{\phi}, \boldsymbol{\theta}) := \sum_{t=1}^{T} \mu^*(\boldsymbol{\theta}) - \mathbb{E}\left[\mu(\theta_{\phi_t})\right]$$

where $\mu(\theta_k)$ is the expected value of arm k, and $\mu^*(\boldsymbol{\theta}) := \max_k\{\mu(\theta_k)\}$. Lai and Robbins (1985) give a strategy that achieves an expected cumulative regret of order $O(\log T)$ and provide a matching lower bound to show it is nearly optimal. This strategy creates an *upper confidence bound* (UCB) for each arm, where the estimated return is given a bonus for uncertainty. A simple example of a UCB is Auer et al. (2002)'s UCB1, which at round t picks the arm maximizing

$$\bar{y}_{k,t} + \sqrt{2\ln(t)/n_{k,t}}$$

where the rewards y_t are in $[0, 1]$, $\bar{y}_{k,t}$ is the average of observed rewards from arm k, and $n_{k,t}$ is the number of samples observed from arm k. Typically, upper confidence bounds are designed so that inferior arm(s) are discarded with minimal investment, and the best arm(s) are guaranteed to remain in play; a key contribution of Lai and Robbins (1985) was to show how such statements can be quantified using Chernoff bounds (or other concentration inequality arguments) and then converted into an upper bound on the cumulative regret. Their approach has been generalized and extended to yield algorithms and regret guarantees across a variety of applications, with upper confidence bounds acting as a guiding design principle.

The richness of the bandit problem has generated a multitude of other approaches. Adding to the above model a prior distribution for the arm parameters $\boldsymbol{\theta}$, the bandit problem can be framed as a Bayesian optimization over $\boldsymbol{\phi}$ to find the allocation strategy which minimizes the expected regret $\int R_T(\boldsymbol{\theta}, \boldsymbol{\phi}) dp\boldsymbol{\theta}$. This optimization can in principle be solved with dynamic programming (such as in Cheng and Berry (2007)); but dynamic programming does not scale well to large or complicated experiments, since the number of possible states explodes. Using results from Whittle (1980), Villar, Bowden, and Wason (2015) show how computation can be reduced considerably by framing the optimal solution as an index policy.

When solving for the optimal strategy is not feasible, the heuristic solution of Thompson sampling is a popular choice with good practical and theoretical performance (Chapelle and Li, 2011; Kaufmann, Korda, and Munos, 2012; Russo and Van Roy, 2016). The decision rule proposed by Thompson (1933) is an adaptive allocation rule where ϕ_t, given all data observed prior to time t, is nondeterministic and chooses arm k with probability equal to its posterior chance of being the best arm. That is, $\phi_t = k$

with probability $p_{k,t} := P_{\mathbb{F}_t}\{k^* = k\}$, where $P_{\mathbb{F}_t}$ is the posterior probability distribution given $(\phi_1, x_1, ...\phi_{t-1}, x_{t-1})$, and $k^* := \arg\max_k(\mu(\theta_k))$ is the index of the best arm (which is a random variable). If the best arm is not unique, the tie should be broken to ensure uniqueness of k^*. In fact, Thompson allocation can be performed with just one sample from the posterior $\mathbb{F}_t$, as shown in the following workflow:

Algorithm 1: Bayesian Workflow with Thompson Sampling

Assume a likelihood model parametrized by $\boldsymbol{\theta}$, such that $\boldsymbol{\theta}$ determines the arm means by $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \ldots, \mu_k(\boldsymbol{\theta}))$;
Assume a prior $\mathbb{F}_1$;
for *each sample* $t \in \{1, \ldots, T\}$ **do**
    Draw from the posterior a sample of the vector of arm means *sample* $\boldsymbol{\theta'} \sim \mathbb{F}_t$; *set* $\boldsymbol{\mu'} := (\mu_1(\boldsymbol{\theta'}), \ldots, \mu_k(\boldsymbol{\theta'}))$;
    Allocate to the arm corresponding to the largest entry of $\boldsymbol{\mu'}$: *set* $\phi_t := \text{argmax}_k\{\mu'_k\}$ (breaking ties at random);
    Receive from arm ϕ_t the next payoff x_t;
    Given the new observation, update posterior to $\mathbb{F}_{t+1}$
end

Exact sampling from the posterior is not always tractable. A popular technique for sampling the posterior approximately is Markov Chain

Monte Carlo (MCMC). The convergence properties of MCMC to the posterior distribution, and in particular the number of steps that must be run to achieve accurate sampling, are well-understood only in special cases (Diaconis, 2009; Dwivedi, Chen, Wainwright, and Yu, 2018). Where theory falls short, practitioners may appeal to a variety of diagnostics tools to provide evidence of convergence to the posterior (Roy, 2020).

There are many other approaches to the bandit problem, including epsilon-greedy (Sutton, Barto, et al., 1998), knowledge gradient (Ryzhov, Powell, and Frazier, 2012), and information-directed sampling (Russo and Van Roy (2014a)).

3. Adaptive Randomization in a Learning Healthcare System

In an LHS, the arms of a multi-armed bandit are treatments and the rewards are patient outcomes. Thus, minimizing cumulative regret corresponds to maximizing patients' measured quality of care - a primary function of the LHS. However, typically there is a secondary goal of learning from a trial: useful takeaways may include confidence intervals for the treatment effects, developing a treatment guide, or making recommendations for non-participating patients in parallel with the trial.

The goals of regret minimization and knowledge generation, often framed as "exploitation vs. exploration," are indeed in fundamental conflict: Bubeck, Munos, and Stoltz (2011) formalized a notion of exploration-based experiments where recommendations are made outside the trial. They define the *simple regret* to be

$$r_T = \mu^* - \mu_{c,T}$$

where $\mu_{c,T}$ is the expectation of the recommended arm after round T and μ^* is the expectation of the best arm. Bubeck, Munos and Stoltz show that upper bounds on cumulative regret R_T lead to lower bounds on r_T, and vice versa. In this sense, algorithms that minimize cumulative regret occupy an extreme point of design space: they maximize welfare of trial patients, but sacrifice knowledge about the inferior treatments. At the other extreme point of design space, an ideal trial for knowledge generation with two arms of equal variance will split the sample sizes equally, consigning half of the patients to the inferior treatment.

Most practical implementations of adaptive randomization in clinical trials do not use bandit algorithms straight out of the box, but instead temper them slightly. A common prescription is to lead with a first phase of equal randomization. Or, allocation probabilities may be shrunk toward $1/K$ in some fashion. Wathen and Thall (2017) discusses the design options of restricting allocations to [.1, .9], leading with a period of equal randomization to prevent the algorithm from "getting stuck" on the worse arm, and altering Thompson Sampling to allocate with probability proportional to $p_{k,t}^c$ for $c \in (0, 1]$. Villar, Bowden, and Wason (2015) consider forced sampling of the control arm every $1/K$ patients. Kasy and Sautmann (2019) modify Thompson sampling to tamp down selection of the best arm(s), asymptotically leading to equal randomization between the best candidates. Lai, Liao, and Kim (2013) give a design which maintains a preferred set of arms, randomizing equally between them, and adaptively drops arms from this set at interim analyses. These various design choices and algorithmic tweaks are typically investigated and tuned with simulation. Even without explicit modification to the standard bandit approach, most medical applications will have *delay* between treatment assignment and observation of outcome; the resulting reduction in available information leads to more exploration for most algorithms.

There are many benefits to using nearer-to-equal randomization probabilities. First, balancing sample sizes between a pair of arms serves inference goals such as increased power of hypothesis tests, shorter confidence intervals, and more accurate future recommendations. Second, closer-to-equal randomization may improve the information for interim decisions such as early stopping and sample size re-estimation. Third, without tuning there may be an unacceptably high chance of sending a majority of patients to the wrong arm (Thall, Fox, and Wathen, 2015). Fourth, more equal randomization can help detect violated assumptions, such as time trends or model misspecification. Fifth, the possibility of violated assumptions suggests treating data as slightly less informative. Finally, probabilities nearer 1/2 are helpful for inverse-probability weighting and randomization tests.

On the other hand, when a treatment is strongly disfavored for a patient, ethical health care requires setting its randomization chance to 0. This may be achieved by thresholding allocation probabilities according to some rule, or suspending or dropping treatment arms at interim analyses. Furthermore, more equal randomization comes at an opportunity cost to the welfare of trial participants. Practical trial design in an LHS must seek a balance between these competing objectives of knowledge generation and participant welfare.

4. Inference for Multi-Armed Bandits in a Learning Healthcare System

The LHS may desire several forms of knowledge from an adaptive randomization trial, including confidence intervals for the outcomes of arms (and their differences), guarantees about selecting arms correctly, and recommendations for treatments in non-participating patients.

Frequentist inference under adaptive randomization designs can be challenging. Due to adaptive sampling, the distribution of standard estimates for the mean of an arm is typically non-Gaussian, and not pivotal with respect to the treatment effect. Concentration techniques for UCB's such as Chernoff bounds can be applied for confidence bounds which may hold uniformly over possible stopping times (Jamieson and Nowak, 2014; Zhao, Zhou, Sabharwal, and Ermon, 2016; Karnin, Koren, and Somekh, 2013). The concentration approach has been extended to FDR control with the always-valid p-values framework (Johari, Pekelis, and Walsh, 2015; Yang, Ramdas, Jamieson, and Wainwright, 2017), and self-normalization techniques from de la Peña, Lai, and Shao (2008) permit extensions to large classes of distributions. However, confidence intervals from concentration bounds may be conservative, slack by a constant or logarithmic factor of width.

In confirmatory trial design, adaptivity may be managed by dividing the trial into segments, each having constant randomization probabilities so that Gaussian theory can be used (with numerical integration for stopping boundaries to compute type I error and power at fixed alternatives). Lai, Liao, and Kim (2013) and Shih and Lavori (2013) show how to do this for their MAB-inspired designs. Alternatively, Korn and Freidlin (2011) suggest block-randomization and block-stratified analysis. Compared to the constantly changing allocation strategies of the standard bandit algorithms, discretization of strategy can come at a moderate or minimal cost, depending on the design and goals.

For analyzing MAB designs with a constantly updating allocation strategy, a key idea for constructing valid frequentist p-values is the randomization test. The randomization test assumes the sharp null hypothesis that the treatment has exactly zero effect, and relies on probabilistic randomization in the allocation algorithm to generate power. In exchange, with other minimal assumptions, it grants valid p-values even in the presence of time trends and other confounders in the patient population (Simon and Simon, 2011). To form confidence intervals, a sharp additive model for the treatment effect may be considered. Confidence bounds then follow from inverting the randomization test, as in Ernst et al. (2004).

Another tool for constructing confidence intervals is hybrid resampling, by Lai and Li (2006). This procedure considers families of different shifts and scales of the observed data, and simulates via resampling to infer which distributions are consistent with the observed treatment effects. Lai and Li show that for group sequential trials, confidence intervals from hybrid resampling can have more accurate coverage than standard normal approximations.

Hadad, Hirshberg, Zhan, Wager, and Athey (2019) suggest a double-robust estimation approach. In addition to using an Augmented Inverse-Probability Weight (AIPW) model, they propose to further adaptively re-weight data to force the treatment effect estimate into an asymptotically Gaussian distribution. Double-robust estimation may help to correct for time trends or other confounding. Yet, data re-weighting comes at a price to efficiency, as pointed out by Tsiatis and Mehta (2003).

Finally, if one assumes a prior and enters the Bayesian framework, posterior inference is a highly flexible approach to analysis. Because Bayes' rule decouples the experimenter's allocation decisions from the rest of the likelihood, the standard Bayesian workflow can be applied to the data in hand without concern for the adaptivity of the design (Berger and Wolpert, 1988). Subject to typical caveats on prior selection and accurate posterior sampling, posterior inference can yield Bayes factors for testing, credible intervals for treatment effects, and decision analysis for treatment recommendations.

5. Contextual Multi-Armed Bandits and Personalized Medicine

For an LHS which continuously seeks to improve and personalize treatment, the important question is not *which* treatment is best, but *for whom* each treatment is best. To address this question, one must augment the bandit model with information about each patient. Calling this side information "covariates" or "*contexts*," one arrives at the "Contextual Multi-Armed Bandit" problem (CMAB).

Contextual multi-armed bandits have found great success in the internet domain, for problems such as serving ads, presenting search results, and testing website features. In contrast, applications in medicine have lagged (with the prominent exception of mobile health (Greenewald, Tewari, Murphy, and Klasnja, 2017; Xia, 2018)). Design of trials in an LHS brings new challenges to the CMAB framework, such as ethical requirements, small sample sizes (roughly 10^2 - 10^4 patients, in comparison to 10^4 - 10^9 clicks for internet applications), requirements for medical professionals to inspect and understand processes, feedback times, and demand for generalizable conclusions.

In the following section we will focus on correctly specified linear models. This assumption derives some justification from features of the LHS: assuming that covariates are continuous and low-dimensional, the patient population of greatest interest is expected to occupy a small region of the covariate domain, due to the systematic filtering of equipoise requirements and further shrinking of the population under experimental focus as "exploiting" increases. Additionally, the conditional expectation of the response is typically a smooth function of the covariates. Therefore, assuming both smoothness of conditional expectation and locality of the studied population, Taylor's theorem implies approximate correctness of the linear model. Similar arguments can be applied to logistic models and other smooth model classes.

5.1 Linear Models for the Reward

If at step t we observe a context vector x_t of length d, sample from arm $\phi_t = k$, and receive reward y_t, we may consider the following simple linear model for the expected reward:

$$E[y_t|x_t, \phi_t = k] = x_t^T \theta_k^*$$

Where θ_k^* is an unknown parameter vector of length d. The LinUCB algorithm of Li et al. (2010) brings the UCB of Lai and Robbins (1985) to this linear model. Assuming the linear model parameters are not shared between arms and that contexts do not depend on the arm chosen (see Li et al. (2010) for the general case,) they suggest to estimate θ_k^* for each arm by a ridge regression $\hat{\theta}_k$. That is, if $X_{k,t}$ is a design matrix whose rows are the contexts of the individuals previously assigned to arm k before time t and $Y_{k,t}$ is a vector of their rewards, the ridge estimator with tuning parameter λ is

$$\hat{\theta}_{k,t} = (X_{k,t}^T X_{k,t} + \lambda I_d)^{-1} X_{k,t}^T Y_{k,t}.$$

Next, Li et al. (2010) construct a UCB for the expected reward around the ridge regression prediction, suggesting the confidence interval

$$|x_t^T \hat{\theta}_{k,t} - x_t^T \theta_k^*| \leq \alpha \sqrt{x_t^T (X_{k,t}^T X_{k,t} + \lambda I_d)^{-1} x_t}$$

where λ is set to 1 and α is a tuning parameter. This confidence interval implicitly assumes a correctly specified linear model and independence of $Y_{k,t}$ given $X_{k,t}^T$, an assumption which is typically broken by the allocation mechanism unless (x_t, y_t) is i.i.d. for all t. Nevertheless, in analogy to the basic UCB algorithm, they propose the LinUCB algorithm which chooses the arm with highest UCB,

$$\phi_t^{UCB} := \arg\max_k \left\{ x_t^T \hat{\theta}_{k,t} + \alpha \sqrt{x_t^T (X_{k,t}^T X_{k,t} + \lambda I_d)^{-1} x_t} \right\}$$

.

LinUCB is easy to implement and has proven popular in applications, inspiring further improvements and competitors. Chu et al. (2011) analyze a theoretical fix to LinUCB and give a regret analysis for a modified algorithm of order $O\left(\sqrt{Td\ln^3(KT\ln(T)/\delta)}\right)$. They also give a nearly-matching general lower bound for the problem of order $\Omega\left(\sqrt{KT}\right)$.

Alternatively, Abbasi-Yadkori et al. (2011), working within a more general framework called "linear bandits" or "linear stochastic bandits," construct self-normalized confidence sets for the arm parameters. In the linear bandit, rather than choosing among a discrete set of arms, one chooses the context x_t from a set D_t, and the rewards are modeled as $y_t = x_t^T\theta^* + \eta_t$. Note that model (5.1) can be embedded within the linear bandit, by sufficiently increasing the dimension of x_t and θ^* and taking D_t as an appropriate finite set of K vectors. Abbasi-Yadkori et al. (2011) assume that, conditioned on data prior to time t, η_t is mean-zero and R-sub-Gaussian for some $R \geq 0$. Further, it is assumed that $\|\theta^*\|_2 \leq S$ for some $S \geq 0$. Then, defining X_t to be a $(t-1)\times d$ matrix whose rows consist of the contexts x_s^T for $s = 1,\ldots,t-1$, defining the reward vector Y_t to be a vector of length $(t-1)$ of the corresponding rewards y_s for $s = 1,\ldots,t-1$, and denoting $\bar{V}_t := \lambda I_d + X_t^T X_t$, for all $t \geq 1$, one may write the ridge estimator as

$$\hat{\theta}_t := \bar{V}_t^{-1} X_t^T Y_t.$$

Abbasi-Yadkori et al. (2011) then derive the confidence set:

$$C_t := \left\{ \|\hat{\theta}_t - \theta^*\|_{\bar{V}_t} \leq R\sqrt{2\log\left(\frac{\det(\bar{V}_t)^{1/2}\det(\lambda I_d)^{-1/2}}{\delta}\right)} + \lambda^{1/2}S \right\}$$

where $\|\cdot\|_{\bar{V}_t}$ is a matrix weighted 2-norm. The collection of these sets, $\mathbb{C} := \bigcap_{t\geq 1} C_t$, provides $1-\delta$ uniform confidence that $\theta^* \in \mathbb{C}$, regardless of an adaptive mechanism for context choice. Abbasi-Yadkori et al. (2011) leverage this confidence approach into a strategy which generalizes the UCB. They follow the underlying principle of "optimism in the face of uncertainty" to select the context

$$x_t := \arg\max_{x\in D_t} \max_{\theta\in C_t} x^T\theta$$

and prove regret guarantees for the linear bandit with this algorithm. For a K-arm trial designer, a key takeaway is that uniform confidence sets offer an approach to model inference (noting that practical use requires strong modeling assumptions, a choice of λ, and bounds for the unknown parameters R and S).

A different approach to the CMAB problem is to generalize the ϵ-greedy algorithm: periodic exploration can be used to estimate a model, and to verify that estimates based on adaptive data collection are not far off. Under the simple linear model (5.1), Goldenshluger and Zeevi (2013) propose to maintain two sets of linear model estimates: $\hat{\theta}_k^*$, estimated on a small amount of equal-randomized data, and $\tilde{\theta}_k^*$, based on all of the (adaptively allocated) data. If the estimated rewards from equal randomization $x_t^T \hat{\theta}_k^*$ are well-separated, the arm with the largest estimate is chosen. Else, the arm with the largest value of $x_t^T \tilde{\theta}_k^*$ is chosen. Under strong assumptions including $K = 2$ arms, i.i.d samples, and a *margin condition* which ensures that the decision boundary between the arms is sharp, i.e.

$$\mathbb{P}\left\{|(\theta_1^* - \theta_2^*)^T X_t| \leq \rho\right\} \leq L\rho, \forall \rho \in (0, \rho_0],$$

they derive a cumulative regret bounded by $O(d^3 \log T)$. Bastani and Bayati (2019) improve these bounds and extend this approach to high-dimensional sparse linear models using L^1 penalization. Bastani, Bayati, and Khosravi (2020) also show that under certain conditions, a pure-greedy approach can yield rate-optimal regret.

5.2 More General Models for the Reward

The Bayesian workflow for the MAB naturally extends to linear models and beyond. Russo and Van Roy (2014b) show that for several classes of well-specified Bayesian problems with contexts, Thompson sampling achieves near-optimal performance and behaves like a problem-adaptive UCB. A variety of competitive risk bounds have been proven for Thompson sampling (Agrawal and Goyal, 2012, 2013; Kaufmann, Korda, and Munos, 2012; Korda, Kaufmann, and Munos, 2013). In empirical studies Thompson sampling often outperforms competitors by a small margin (Scott, 2010; Chapelle and Li, 2011; Dimakopoulou et al., 2017).

An alternative for the jealous non-Bayesian is what we will call "pseudo-Thompson bootstrapping." Given a black box algorithm that models the outcomes under each arm, the idea is to bootstrap-resample data to generate variation in the model's estimates. Pretending that this resampling distribution is a posterior, one can drop the estimated "probabilities" of arm superiority into the Thompson rule and hope to recover its performance advantages. While this technique approximates Thompson sampling for some known cases (Eckles and Kaptein, 2014), its general theoretical properties remain unclear. The main appeal of the approach is to offer a wrapper for popular estimation algorithms for large datasets, including regression trees, random forests and neural networks (Elmachtoub et al., 2017; Osband et al., 2016)).

Vaswani et al. (2019) propose the RandUCB algorithm, which gives LinUCB nondeterministic allocation probabilities by perturbing the confidence bound randomly in a way that somewhat resembles bootstrapping. For the linear model, RandUCB can be viewed as a generalization of Thompson sampling under a Gaussian model. Vaswani et al. also prove competitive regret guarantees for RandUCB.

Finally, there are nonparametric methods that leverage smoothness of the expected response. Rigollet and Zeevi (2010) discretize space into buckets, and run MAB's on each of them independently. Lu et al. (2010) gives a contextual bandit which clusters data adaptively and provides guarantees under Lipschitz assumptions. Kim et al. (2020) perform local linear regression and pair it with ϵ-greedy randomization and arm elimination, meeting minimax lower bounds on regret under certain regularity conditions.

6. Dynamic Treatment Regimes

An LHS bears responsibility for patients over time as their clinical status and treatment needs evolve. Formalizing the notion of a complete care strategy, a Dynamic Treatment Regime (DTR) is a set of rules which dictates treatment decisions given a patient's history of covariates and prior treatment (Lavori and Dawson, 2004).

Thus, if a patient is observed at time-points τ_i when observations x_i will be recorded and treatment action a_i taken, a DTR is a function that maps $(\tau_i, x_{1:i}, a_{1:i-1})$ to a_i.

DTRs may be studied with a Sequential Multiple Assignment Randomized Trial (SMART), which begins with an initial treatment randomization and at each subsequent decision point re-randomizes patients among further treatment options. A SMART culminates in an outcome Y for each individual (which may be a function of $(x_1, \ldots, x_I)$) by which the treatments will be assessed. Lavori and Dawson (2007, 2008) construct confidence intervals for comparing DTRs based on their expected outcomes.

In the group sequential clinical trial setting, Zhong (2018) demonstrates asymptotic multivariate normal approximations to estimated outcomes and transition probabilities. Zhong proposes likelihood-based Wald tests with simultaneous coverage for comparing DTRs, and demonstrates the approach on adaptive play-the-winner designs.

Key challenges for design and analysis of SMARTs include incorporating patient covariates and handling estimation as the length of the decision tree grows, since the number of treatment strategies and possible patient histories explodes rapidly. Most SMARTS do not consider more than two decision points per patient.

One approach for handling patient covariates is Q-learning (Sutton et al., 1998; Murphy, 2005). Q-learning seeks to model the patient's expected final outcome, conditionally on taking action a_i given the history $(\tau_i, x_{1:i}, a_{1:i-1})$, and assuming optimal decisionmaking thereafter. This model is thus a function, $E[Y|\tau_i, x_{1:i}, a_{1:i}] \sim Q(\tau_i, x_{1:i}, a_{1:i})$, called the *Q-function*.

In order to estimate a Q-function, Q-learning alternates between model estimation of expected values of states and backward induction to select optimal actions, using a modified version of Bellman's inequality. Q-learning is therefore an approximate dynamic programming technique. It has been combined with a variety of modeling approaches including linear models (Murphy, 2005), regression trees (Ernst, Geurts, and Wehenkel, 2005), and kernels (Ormoneit and Sen, 2002). Chakraborty and Murphy (2014) discuss non-regular asymptotics for Q-learning with the linear model.

7. Recent Advances for Learning about DTRs

Intuitions and approaches from MAB and CMAB theory, including Lai and Robbins (1985), are proving fruitful in constructing algorithms with theoretical guarantees and good performance for analyzing DTRs.

In the general DTR setting, Zhang and Bareinboim (2019) use the UCB approach to motivate a Reinforcement Learning algorithm and derive regret guarantees. Following the techniques of Lai and Robbins (1985) and Auer, Jaksch, and Ortner (2009), at each time t they construct a uniform confidence set $\mathbb{M}_t$ with two-sided bounds on the final payouts and transitions, and then use the Bellman equation recursively to find an optimal DTR in $\mathbb{M}_t$. We note that this approach also permits confidence bands for the value of individual DTRs. Zhang and Bareinboim derive regret guarantees and further show that weak evidence from observational data collection can be used to narrow the range of possible transitions, thus narrowing $\mathbb{M}_t$ and improving performance.

Hu and Kallus (2020) analyze a two-stage DTR model. Assuming a linear model for Q-functions, they extend the contextual bandit approaches of Goldenshluger and Zeevi (2013) and Bastani, Bayati, and Khosravi (2020) to the two-stage two-treatment DTR setting, using a combination of unbiased estimates from a small sample and biased estimates from the full sample. They derive regret bounds under several margin conditions on the Q-functions, notably showing under a sharp margin condition a regret bound of order $O(d(\log d)^{2/3} \log T + (d \log d)^2)$. They demonstrate that their bounds have optimal dependence on T by applying lower bounds from contextual bandits.

Wang and Powell (2016) demonstrate an important connection between DTR's and contextual bandits in a Bayesian framework. They model binary outcomes using Bayesian generalized linear models and handle posterior computations using Laplace approximations. With quick recursive computation of the value function, they show how to collapse the DTR problem into a CMAB problem, where each decision point becomes a bandit sample, and payoffs are given by the change in the posterior expected value. This formulation naturally permits them to use Bayesian CMAB algorithms, including the Knowledge Gradient, Thompson sampling, and Greedy Bayes algorithms, for learning and executing DTRs.

8. Discussion and Conclusion

In this paper we have reviewed the current literature bearing on the use of adaptive randomization, as represented by the CMAB, as a natural model for an LHS that seeks to add experimental strength to its portfolio of learning methodologies. Lai's contributions to the theory and practice of bandits goes back almost 40 years, and the basic insights have also led to many innovations in diverse areas, such as finance, internet commerce, and medical drug development.

This past year, the world has been engulfed in a global pandemic, and many of the ideas reviewed here have taken on much greater significance. Governments are learning how to control the outbreak with a combination of non-pharmacologic interventions ranging from masks and "social distancing" to rapid learning by intensivists faced with a virus having pleiotropic clinical effects, development of new therapeutics and repurposing of existing drugs, to ambitious programs of vaccine development. They are also dealing with the economic disruption, both directly from the pandemic and indirectly from the control efforts.

The limitations of both observational, non-experimental approaches and conventional randomized clinical trials have been cast in sharp relief. Each scientific specialty has begun to propose ways to make its responses better and faster "next time around." Virologists propose getting a jump on both therapeutic drug and vaccine development, even in advance of knowing the identity of the new pandemic agent. Ecologists and wildlife conservationists urge a greatly expanded global project to survey likely animal sources of the next spillover event and target those agents that are likely to pose a substantial global threat. Clinical scientists and trialists seek to create pre-formed platforms for rapid testing of the drugs, non-pharmacologic interventions, and vaccines that will be proposed. In this area, innovative experimental design will be critical, and as statisticians are recruited to help prepare for the next emergency, they will find as we have that the work of Tze Leung Lai will provide a sturdy basis and flexible but reliable framework for their efforts.

9. Acknowledgements

Michael Sklar acknowledges support of this research by the National Science Foundation under DMS-1811818.

References

Abbasi-Yadkori, Y., D. Pál, and C. Szepesvári (2011). Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 24*, pp. 2312–2320. Curran Associates, Inc.

Agrawal, S. and N. Goyal (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1.

Agrawal, S. and N. Goyal (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning 47*(2-3), 235–256.

Auer, P., T. Jaksch, and R. Ortner (2009). Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pp. 89–96.

Bartroff, J., T. L. Lai, and M.-C. Shih (2012). *Sequential experimentation in clinical trials: design and analysis*, Volume 298. Springer Science & Business Media.

Bastani, H. and M. Bayati (2019). Online decision making with high-dimensional covariates. *Operations Research*.

Bastani, H., M. Bayati, and K. Khosravi (2020). Mostly exploration-free algorithms for contextual bandits. *Management Science*.

Berger, J. O. and R. L. Wolpert (1988). *The likelihood principle.* IMS.

Bouneffouf, D. and I. Rish (2019). A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*.

Bubeck, S., R. Munos, and G. Stoltz (2011). Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science 412*(19), 1832–1852.

Chakraborty, B. and S. A. Murphy (2014). Dynamic treatment regimes. *Annual review of statistics and its application 1*, 447–464.

Chamberlayne, R., B. Green, M. L. Barer, C. Hertzman, W. J. Lawrence, and S. B. Sheps (1998). Creating a population-based linked health database: a new resource for health services research. *Canadian Journal of Public Health 89*(4), 270–273.

Chapelle, O. and L. Li (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257.

Cheng, Y. and D. A. Berry (2007). Optimal adaptive randomized designs for clinical trials. *Biometrika 94*(3), 673–689.

Chu, W., L. Li, L. Reyzin, and R. Schapire (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214.

Diaconis, P. (2009). The markov chain monte carlo revolution. *Bulletin of the American Mathematical Society 46*(2), 179–205.

Dimakopoulou, M., Z. Zhou, S. Athey, and G. Imbens (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Dwivedi, R., Y. Chen, M. J. Wainwright, and B. Yu (2018). Log-concave sampling: Metropolis-hastings algorithms are fast! In *Conference on Learning Theory*, pp. 793–797.

Eckles, D. and M. Kaptein (2014). Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*.

Elmachtoub, A. N., R. McNellis, S. Oh, and M. Petrik (2017). A practical method for solving contextual bandit problems using decision trees. *arXiv preprint arXiv:1706.04687*.

Ernst, D., P. Geurts, and L. Wehenkel (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research 6*(Apr), 503–556.

Ernst, M. D. et al. (2004). Permutation methods: a basis for exact inference. *Statistical Science 19*(4), 676–685.

Goldenshluger, A. and A. Zeevi (2013). A linear response bandit problem. *Stochastic Systems 3*(1), 230–261.

Greenewald, K., A. Tewari, S. Murphy, and P. Klasnja (2017). Action centered contextual bandits. In *Advances in neural information processing systems*, pp. 5977–5985.

Hadad, V., D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey (2019). Confidence intervals for policy evaluation in adaptive experiments. *arXiv preprint arXiv:1911.02768*.

Hu, Y. and N. Kallus (2020). Dtr bandit: Learning to make response-adaptive decisions with low regret. *arXiv preprint arXiv:2005.02791*.

Jamieson, K. and R. Nowak (2014). Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE.

Johari, R., L. Pekelis, and D. J. Walsh (2015). Always valid inference: Bringing sequential analysis to a/b testing. *arXiv preprint arXiv:1512.04922*.

Karnin, Z., T. Koren, and O. Somekh (2013). Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pp. 1238–1246.

Kasy, M. and A. Sautmann (2019). Adaptive treatment assignment in experiments for policy choice.

Kaufmann, E., N. Korda, and R. Munos (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pp. 199–213. Springer.

Kim, D. W., T. L. Lai, and H. Xu (2020). Multi-armed bandits with covariates: Theory and applications.

Korda, N., E. Kaufmann, and R. Munos (2013). Thompson sampling for 1-dimensional exponential family bandits. In *Advances in neural information processing systems*, pp. 1448–1456.

Korn, E. L. and B. Freidlin (2011). Outcome-adaptive randomization: is it useful? *Journal of Clinical Oncology 29*(6), 771.

Lai, T. L., P. W. Lavori, and O. Y.-W. Liao (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary clinical trials 39*(2), 191–200.

Lai, T. L. and W. Li (2006). Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. *Biometrika 93*(3), 641–654.

Lai, T. L. and O. Y.-W. Liao (2012). Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. *Sequential analysis 31*(4), 441–457.

Lai, T. L., O. Y.-W. Liao, and D. W. Kim (2013). Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemporary clinical trials 36*(2), 651–663.

Lai, T. L. and H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics 6*(1), 4–22.

Lai, T. L. and M.-C. Shih (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika 91*(3), 507–528.

Lavori, P. W. and R. Dawson (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A (Statistics* *in Society) 163*(1), 29–38.

Lavori, P. W. and R. Dawson (2004). Dynamic treatment regimes: practical design considerations. *Clinical trials 1*(1), 9–20.

Lavori, P. W. and R. Dawson (2007). Improving the efficiency of estimation in randomized trials of adaptive treatment strategies. *Clinical Trials 4*(4), 297–308.

Lavori, P. W. and R. Dawson (2008). Adaptive treatment strategies in chronic disease. *Annu. Rev. Med. 59*, 443–453.

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

Lu, T., D. Pál, and M. Pál (2010). Contextual multi-armed bandits. In *Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics*, pp. 485–492.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine 24*(10), 1455–1481.

Olsen, L., D. Aisner, and J. M. McGinnis (2007). Institute of medicine roundtable on evidence-based medicine: The learning healthcare system. In *Workshop Summary*.

Ormoneit, D. and Ś. Sen (2002). Kernel-based reinforcement learning. *Machine learning 49*(2-3), 161–178.

Osband, I., C. Blundell, A. Pritzel, and B. Van Roy (2016). Deep exploration via bootstrapped

DQN. In *Advances in neural information processing systems*, pp. 4026–4034.

Peña, V. H., T. L. Lai, and Q.-M. Shao (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Rigollet, P. and A. Zeevi (2010). Nonparametric bandits with covariates. In *23rd COLT*, pp. 54–66.

Roy, V. (2020). Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application 7*, 387–412.

Russo, D. and B. Van Roy (2014a). Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pp. 1583–1591.

Russo, D. and B. Van Roy (2014b). Learning to optimize via posterior sampling. *Mathematics of Operations Research 39*(4), 1221–1243.

Russo, D. and B. Van Roy (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research 17*(1), 2442–2471.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier (2012). The knowledge gradient algorithm for a general class of online learning problems. *Operations Research 60*(1), 180–195.

Scott, S. L. (2010). A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry 26*(6), 639–658.

Shih, M.-C. and P. W. Lavori (2013). Sequential methods for comparative effectiveness experiments: Point of care clinical trials. *Statistica Sinica 23*(4), 1775–1791.

Simon, R. and N. R. Simon (2011). Using randomization tests to preserve type i error with response adaptive and covariate adaptive randomization. *Statistics & probability letters 81*(7), 767–772.

Sutton, R. S., A. G. Barto, et al. (1998). *Introduction to reinforcement learning*, Volume 135. MIT press Cambridge.

Thall, P., P. Fox, and J. Wathen (2015). Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology 26*(8), 1621–1628.

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

Tsiatis, A. A. and C. Mehta (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika 90*(2), 367–378.

Vaswani, S., A. Mehrabian, A. Durand, and B. Kveton (2019). Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*.

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

Wang, Y. and W. Powell (2016). An optimal learning method for developing personalized treatment regimes. *arXiv preprint arXiv:1607.01462*.

Wathen, J. K. and P. F. Thall (2017). A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clinical Trials 14*(5), 432–440.

Whittle, P. (1979). Discussion of dr gittins' paper. *Journal of the Royal Statistical Society 41*, 164–177.

Whittle, P. (1980). Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society: Series B (Methodological) 42*(2), 143–149.

Xia, I. (2018). *The Price of Personalization: An Application of Contextual Bandits to Mobile Health.* Ph. D. thesis, Harvard University.

Yang, F., A. Ramdas, K. G. Jamieson, and M. J. Wainwright (2017). A framework for multi-a (rmed)/b (andit) testing with online fdr control. In *Advances in Neural Information Processing Systems*, pp. 5957–5966.

Zhang, J. and E. Bareinboim (2019). Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pp. 13401–13411.

Zhao, S., E. Zhou, A. Sabharwal, and S. Ermon (2016). Adaptive concentration inequalities for sequential decision problems. In *Advances in Neural Information Processing Systems*, pp. 1343–1351.

Zhong, X. (2018). *Design and Analysis of Sequential Multiple Assignment Randomized Trial for Comparing Multiple Adaptive Interventions.* Ph. D. thesis, Mailman School of Public Health, Columbia University.

Stanford University

E-mail: sklarm@stanford.edu

VA Palo Alto Cooperative Studies Program Coordinating Center

E-mail: mei-chiung.shih@va.gov