Special Issue Paper

Applied Stochastic Models in Business and Industry

Received 10 June 2014, Revised 23 December 2014, Accepted 24 December 2014 Published online 16 January 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2104

Multi-armed bandit experiments in the online service economy

Steven L. Scott[*†]

The modern service economy is substantively different from the agricultural and manufacturing economies that preceded it. In particular, the cost of experimenting is dominated by opportunity cost rather than the cost of obtaining experimental units. The different economics require a new class of experiments, in which stochastic models play an important role. This article briefly summarizes multi-armed bandit experiments, where the experimental design is modified as the experiment progresses to reduce the cost of experimenting. Special attention is paid to Thompson sampling, which is a simple and effective way to run a multi-armed bandit experiment.

Keywords: Thompson sampling; sequential experiment; Bayesian; reinforcement learning

1. Introduction

Service is the dominant sector of the US economy, accounting for roughly 80% of the US gross domestic product [1]. It is similarly important in other developed nations. Much service sector activity involves traditional retail and person-to-person services, but a growing fraction comes from technology companies like Google, Amazon, Facebook, Salesforce, and Netflix. These and similar companies provide Internet services related to search, entertainment, retail, advertising, and information processing under the 'software as a service' paradigm.

As with other industries, service can be improved through experimentation. Yet the cost structure of a typical service experiment is dramatically different from that of a manufacturing or agricultural experiment, particularly for online service. A traditional industrial experiment incurs costs mainly from acquiring or processing experimental units, such as growing crops on plots of land, destroying items taken from the production line, or paying subjects to participate as part of a focus group. By contrast, service experiments involve changing the service being provided to existing customers with whom the service provider would have engaged even if no experiment had taken place. Assuming the service can be modified with minimal expense (which is true for online services), then the cost of experimenting is dominated by the opportunity cost of providing sub-optimal service to customers.

A second distinction is that service providers are able to continually monitor the quality of the service they provide, for example, by keeping track of the number of clicks generated by an advertisement, or the frequency with which a software feature is used. This blurs the line separating the laboratory from the production line in a way that is difficult to imagine for manufacturing, as if the car could be improved after leaving the factory. Online service companies can conduct experiments faster and easier than ever before. Service providers can experiment continuously, perpetually improving different aspects of their offerings [2]. One impediment is that experiments are expensive. Dramatically increasing the frequency and scope of experiments requires a corresponding reduction in cost. Multi-armed bandits are a type of sequential experiment that is naturally aligned with the economics of the service industry. This article is a brief introduction to multi-armed bandit experiments, the 'Thompson sampling' heuristic for managing them, and some of the practical considerations that can arise during real-world applications. Section 2 describes multi-armed bandit experiments and reviews some of the techniques that have been developed to implement them. Section 3 discusses the particular method of Thompson sampling. Section 4 discusses some of the practical aspects of running a multi-armed bandit experiment in various contexts. Section 5 concludes.

Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, U.S.A.
**Correspondence to: Steven L. Scott, Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, U.S.A.*
† E-mail: stevescott@google.com

2. Multi-armed bandit experiments

A multi-armed bandit is a sequential experiment where the goal is to produce the largest reward. The typical setup considers K actions or 'arms'. Arm a is associated with an unknown quantity v_a giving the 'value' of that arm. The goal is to choose the arm providing the greatest value and to accumulate the greatest total reward in doing so. The name 'multi-armed bandit' is an allusion to a row of slot machines (colloquially known as 'one-armed bandits') with different reward probabilities. The job of the experimenter is to choose the slot machine with the highest probability of a reward.

A more formal description assumes that rewards come from a probability distribution $f_a(y|\theta)$, where a indexes the action taken (or arm played), y is the observed reward, and θ is a set of unknown parameters to be learned through experimentation. The value $v_a(\theta)$ is a known function of the unknown θ, so if θ were observed, the optimal arm would be known.

Consider a few examples for concreteness.

1. In the slot machine problem (the 'binomial bandit'), we have $\theta = (\theta_1, \ldots, \theta_K)$, a vector of success probabilities for K independent binomial models, with $v_a(\theta) = \theta_a$.
2. In a two-factor experiment for maximizing conversion rates on a website, suppose the factors are button color (red or blue) and button position (left or right). The experimental configuration can be expressed in terms of two dummy variables X_c (for button color) and X_p (for position). Then θ might be the set of logistic regression coefficients in the model

$$\text{logit } Pr(\text{conversion}) = \theta_0 + \theta_1 X_c + \theta_2 X_p + \theta_3 X_c X_p. \tag{1}$$

   The action a is isomorphic to the vector of design variables $\mathbf{x}_a = (1, X_c, X_p, X_c X_p)$, with $v_a(\theta) = \text{logit}^{-1}(\theta^T \mathbf{x}_a)$.
3. As a final example, one could model 'restless bandits' [3] by assuming that some or all the coefficients in Equation (1) were indexed by time in a Gaussian process, such as

$$\theta_{t+1} = \mathcal{N}(\theta_t, \Sigma_t). \tag{2}$$

There are many obvious generalizations, such as controlling for background variation (analogous to the 'blocking factors' in a traditional experiment) by including them as covariates in Equation (1) or (2), combining information from similar experiments using hierarchical models, or replacing binary rewards with small counts, continuous quantities, or durations.

The multi-armed bandit problem is clearly driven by parameter uncertainty. If θ were known, then $v_a(\theta)$ would be known as well, and the optimal action would be clear. It is tempting to find an 'optimal' point estimate $\hat{\theta}$ and take the corresponding implied action $\hat{a} = \arg\max_a v_a(\hat{\theta})$. This is known as the 'greedy strategy', which has been well documented to underperform [4]. The problem with the greedy strategy is that estimation error in $\hat{\theta}$ can lead to an inferior choice of $\hat{a}$. Always acting according to $\hat{a}$ limits opportunities to learn that other arms are superior. To beat the greedy strategy, one must sometimes take different actions than those implied by $\hat{\theta}$. That is, one must experiment with some fraction of observations. The tension between following the (apparently) optimal $\hat{\theta}$ and experimenting in case $\hat{\theta}$ is wrong is known as the 'explore/exploit trade-off'. It is the defining characteristic of the multi-armed bandit problem.

Bandits have a long and colorful history. Whittle [5] famously quipped that

> ... [the bandit problem] was formulated during the [second world] war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage.

Many authors attribute the problem to Robbins [6], but it dates back at least to Thompson [7]. For very simple reward distributions such as the binomial, Gittins [8] developed an 'index policy' that produces an optimal solution under geometric discounting of future rewards. The Gittins index remains the method of choice for some authors today (e.g., [9]). Obtaining optimal strategies for more complex reward distributions such as Equations (1) and (2) is sufficiently difficult that heuristics are typically used in practice. Sutton and Barto [4] describe several popular heuristics such as ϵ-greedy, ϵ-decreasing, and softmax methods. All of these require arbitrary tuning parameters that can lead to inefficiencies. Auer *et al.* [10] developed a popular heuristic in which one selects the arm with the largest upper confidence bound (UCB) of $v_a(\theta)$. For independent rewards with no shared parameters, the UCB algorithm was shown to satisfy the optimal rate of exploration discovered by [11]. Agarwal *et al.* [12] used UCB to optimize articles shown on the main *Yahoo!* web page. More recently, attention has been given to a technique known as Thompson sampling. Chapelle and Li [13] produced simulations suggesting that Thompson sampling had superior regret performance relative to UCB. See [14] for comparisons to older heuristics. Thompson sampling is the method used for the remainder of this article.

3. Thompson sampling

Thompson sampling [7] is a heuristic for managing the explore/exploit trade-off in a multi-armed bandit problem. Let $\mathbf{y}_t$ denote the set of data observed up to time t and define

$$\begin{aligned} w_{at} &= Pr(a \text{ is optimal} | \mathbf{y}_t) \\ &= \int I(a = \arg\max v_a(\theta)) p(\theta|\mathbf{y}_t)\, d\theta, \end{aligned} \tag{3}$$

where $p(\theta|\mathbf{y}_t)$ is the Bayesian posterior distribution of θ given data observed up to time t. The Thompson heuristic assigns the observation at time $t+1$ to arm a with probability w_{at}. One can easily compute w_{at} from a Monte Carlo sample $\theta^{(1)}, \ldots, \theta^{(G)}$ simulated from $p(\theta|\mathbf{y}_t)$ using

$$w_{at} \approx \frac{1}{G} \sum_{g=1}^{G} I(a = \arg\max v_a(\theta^{(g)})). \tag{4}$$

Notice that the algorithm where one first computes w_{at} and then generates from the discrete distribution $w_{1t}, \ldots, w_{Kt}$ is equivalent to selecting a single $\theta_t \sim p(\theta|\mathbf{y}_t)$ and selecting the a that maximizes $v_a(\theta_t)$. Thus, Thompson sampling can be implemented using a single draw from $p(\theta|\mathbf{y})$, although computing w_{at} explicitly yields other useful statistics such as those described in Section 3.1.

The Thompson heuristic strikes an attractive balance between simplicity, generality, and performance. Perhaps most importantly, it is easy to understand. If arm a has a 23% chance of being the best arm, then it has a 23% chance of attracting the next observation. This statement obscures technical details about how the 23% is to be calculated or why the two probabilities should match, but it contains an element of 'obviousness' that many people find easy to accept. If desired, tuning parameters can be introduced into Thompson sampling. For example, one could assign observations with probability proportional to w_{at}^{γ}. Setting $\gamma < 1$ makes the bandit less aggressive, increasing exploration. Setting $\gamma > 1$ makes the bandit more aggressive. Of course, exploration can also be encouraged by introducing a more flexible reward distribution, for example, replacing the binomial with the beta-binomial. Section 4 discusses reward distributions that can be used for controlling exploration, which can be more effective than tuning parameters. Modifying the reward distribution to slow convergence highlights a second feature of Thompson sampling, which is that it can be applied generally. The only requirement is the ability to simulate θ from $p(\theta|\mathbf{y}_t)$, which can be performed using standard Bayesian methods for a very wide class of reward distributions.

Thompson sampling handles exploration gracefully. Let $a_t^* = \arg\max_a w_{at}$ denote the arm with the highest optimality probability given data to time t. The fraction of data devoted to exploration at time $t+1$ is $1 - w_{a_t^* t}$, which gradually diminishes as the experiment evolves. Thompson sampling not only manages the overall fraction of exploration but also manages exploration at the level of individual arms. Clearly, inferior arms are explored less frequently than arms that might be optimal, which has two beneficial implications. First, it improves the economic performance of the experiment and offers a better experience to customers who would have been assigned to inferior arms. Second, it produces greater sample sizes among arms near the top of the value scale, which helps distinguish the best arms from the merely good ones. Thompson sampling tends to shorten experiments while simultaneously making them less expensive to run for longer durations.

The randomization aspect of Thompson sampling is an often overlooked advantage. Real online experiments can involve many (hundreds or thousands) of visits to a website before updating can take place. It is generally preferable to update as soon as possible, but updates may be delayed for technical reasons. For example, the system that logs the results of site visits might be different from the system that determines which version of the site should be seen. It can take some time (several minutes, hours, or perhaps a day) to collect logs from one system for processing by another, and during that time a high-traffic site might attract many visitors. Thompson sampling randomly spreads observations across arms in proportion to w_{at} while waiting for updates. Nonrandomized algorithms pick a single arm, making the same 'bet' for each experimental unit, which substantially increases the variance of the rewards. (Rewards are great if you bet on the right arm and terrible if you bet on the wrong one.) Randomization also offers a source of pure variation that can help ensure causal validity.

Although Thompson sampling does not explicitly optimize any specific criterion, there are mathematical and empirical results showing that it tends to beat other heuristics. Chapelle and Li [13] produced a highly cited simulation study suggesting that Thompson sampling outperformed UCB in the case of the binomial bandit. May *et al.* [15] showed that Thompson sampling is a consistent estimator of the optimal arm, and over the life of the experiment, 'almost all' (in a probabilistic sense) of the time is spent on the optimal arm. Kaufmann *et al.* [16] showed that Thompson sampling for the binomial bandit satisfies the optimal bound of Lai and Robbins [11]. Bubeck and Liu [17, 18] established regret bounds for Thompson

sampling in the case of independent arms with rewards in [0, 1]. Russo and Van Roy [19] established bounds on expected regret in a much broader setting, arguing that Thompson sampling will never do worse, in expectation, than a well-tuned UCB algorithm. A second paper by Russo and Van Roy [20] showed that it is possible to improve the convergence rate of Thompson sampling, in particular stylized settings, by introducing a form of one-step look-ahead. However, attaining the improved rate comes at the cost of either restricting the reward distribution to be an exponential family model with a conjugate prior or else forward simulating the model at a computational cost many times that of the Thompson algorithm.

3.1. Using regret to end experiments

The methods used to compute w_{at} for Thompson sampling can also produce a reasonable method of deciding when experiments should end. Let θ_0 denote the true value of θ, and let $a^* = \arg\max_a v_a(\theta_0)$ denote the arm that is truly optimal. The *regret* from ending the experiment at time t is $v_{a^*}(\theta_0) - v_{a_t^*}(\theta_0)$, which is the value difference between the truly optimal arm and the arm that is apparently optimal at time t. In practice, regret is unobservable, but we can compute its posterior distribution. Let $v_*(\theta^{(g)}) = \max_a v_a(\theta^{(g)})$ where $\theta^{(g)}$ is a draw from $p(\theta|\mathbf{y}_t)$. Then,

$$r^{(g)} = v_*(\theta^{(g)}) - v_{a_t^*}(\theta^{(g)})$$

is a draw from posterior distribution of regret. Note the distinction: $v_*(\theta^{(g)})$ is the maximum value available within Monte Carlo draw g, while $v_{a_t^*}(\theta^{(g)})$ is the value (again in draw g) for the arm deemed best across all Monte Carlo draws. Their difference is often 0 but is sometimes positive.

For communication purposes, it is helpful that the units of regret are the units of value (e.g., dollars, clicks, or conversions). The distribution of regret can be summarized by an upper quantile, such as the 95th percentile, to give the 'potential value remaining' (PVR) in the experiment. PVR is the value per play that might be lost if the experiment ended at time t. Because businesses experiment in the hope of finding an arm that provides greater value, a sensible criterion for ending the experiment is when PVR falls below a threshold of practical significance.

The PVR statistic handles ties gracefully. If there are many arms in the experiment, there can be several that give essentially equal performance. Ties between arms can easily happen as part of a multifactor experiment with one or more irrelevant factors. Experiments ended by the PVR criterion naturally produce two sets of arms: one set that is clearly inferior and a second containing arms that are nearly equivalent to one another. Any arm from the set of potential winners can be chosen going forward. If desired, one may use w_{at} as a guide for choosing among the potential winners, but subjective preference (e.g., for an existing version) may be used as well.

Note that regret can also be defined as a percentage change from the current apparently optimal arm, so that draws from the posterior are given by

$$\rho^{(g)} = \frac{v_*(\theta^{(g)}) - v_{a_t^*}(\theta^{(g)})}{v_{a_t^*}(\theta^{(g)})}, \tag{5}$$

which is unit free. If the experimenter is unwilling to specify a definition of 'practical significance', then an experimental framework can use an arbitrary operating definition such as $\rho < 0.01$.

4. Practical applications of Thompson sampling

This Section discusses some of the practicalities associated with Thompson sampling in applied problems. It begins with a simulation study showing the gains from Thompson sampling relative to traditional experiments, before turning to a set of useful generalizations. It ends with a second, more realistic simulation that compares the models discussed later in the text.

4.1. A/B testing

Among Internet companies, the term 'A/B testing' describes an experiment comparing a list of alternatives along a single dimension, which statisticians would call the 'one-way layout'. An A/B test often involves only two alternatives, but the term is sometimes sloppily applied to experiments with multiple alternatives. A canonical example of A/B testing is website optimization, where multiple versions of a website are constructed, traffic is randomly assigned to the different versions, and counts of conversions are monitored to determine which version of the site performs the best. A 'conversion' is an action designated by the site owner as defining a successful visit, such as making a purchase, visiting a particular section of the site, or signing up for a newsletter.

Consider two versions of a website that produce conversions according to independent binomial distributions. Version A produces a conversion 0.1% of the time ($p = 0.001$), and version B produces conversions 0.11% of the time ($p = 0.0011$). To detect this difference using a traditional experiment with 95% power under a one-sided alternative, we would need roughly 4.5 million observations (2,270,268 in each arm). If we decreased the power to 0.5, we would need slightly more than 1.1 million (567,568 in each arm). The regret for each observation assigned to version A is $0.0011 - 0.001 = 0.0001$, so for every 10,000 observations assigned to version A, we lose one conversion.

Figure 1 shows the results from simulating the multi-armed bandit process 100 times under the conditions described earlier. Each simulation assumes the site receives 100 visits per day. The true success probabilities are held constant across all simulation runs. Observations are assigned to arms according to the Thompson procedure with updates occurring once per day (so there are 100 observations per update). Each simulated experiment ended when the PVR statistic in Equation (5) fell below $\rho = 1\%$. The simulation found the correct version in 84 out of 100 runs. Figure 1(a) shows that the number of observations required to end the experiment is highly variable but substantially less than the numbers obtained from the power calculations for the traditional experiments. Figure 1(b) compares the number of lost conversions, relative to the optimal policy of always showing version B, for the bandit and for the traditional experiments with power 0.5, 0.84 (the realized power of this simulation), and 0.95. Roughly two-thirds of the simulation runs produced single digits of lost conversions compared with hundreds of lost conversions for the traditional experiment with comparable power. Given that our fictitious 'site' generates about one conversion every 10 days, 100 conversions are a staggering difference. The savings are partly due to shorter experiment times and partly due to the experiments being less expensive while they are run. Both factors are important. Under the 95% power calculation, the experiment would take roughly 125 years, while 29 of 100 bandit simulations finished within 1 year. Both are impractically long, but the bandit offers at least some chance of completing the experiment.

Detecting small differences is a hard problem, made harder when the baseline probabilities are small as well. Scott [21] gives similar results for an easier setting with true success rates of 4% and 5%, and for settings with multiple arms. When there is a difference to be found, the multi-armed bandit approach is dramatically more efficient at finding the best arm than traditional statistical experiments, and its advantage increases as the number of arms grows [22].

A referee has pointed out that there are two senses in which the preceding simulation is unfair. The first is that bandits' advantages may be overstated because the bandit is able to use sequential information while the traditional experiment does not. A more balanced comparison would pit the multi-armed bandit against the sequential probability ratio test, which would also greatly reduce the cost of the experiment. However, the simulation is instructive because the traditional one-way layout is currently viewed as the 'best practice' by the A/B testing community. The second source of unfairness understates the bandits' advantage, because the classical design assumes the true value of the alternative hypothesis, which is almost always unknown. Classical (or sequential probability ratio test) designs based on hypothesis testing become even more awkward when there are multiple arms. As the number of arms grows, the null hypothesis that all arms are equally effective becomes increasingly dubious, and the need to specify a specific alternative for each arm becomes an increasing burden. By eschewing hypothesis testing as a foundation for experimental design, the multi-armed bandit sidesteps such artificial technical issues.

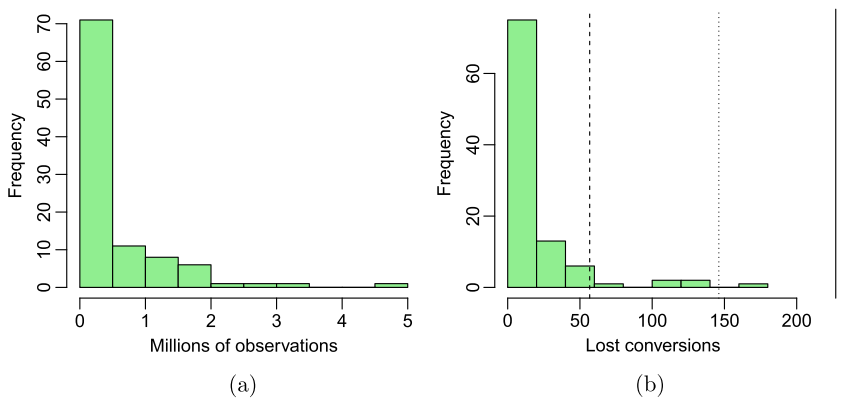

Figure 1. (a) Histogram of the number of observations required to end the experiment in 100 runs of the binomial bandit described in Section 4.1. (b) The number of conversions lost during the experiment period. The vertical lines show the number of lost conversions under the traditional experiment with 95% (solid), 50% (dashed), and 84% (dotted) power.

4.2. Contextual information

The multi-armed bandit can be sensitive to the assumed model for the rewards distribution. The binomial model assumes all observations are independent with the same success probability. This assumption can fail if the arms are exposed to subpopulations with different performance characteristics at different times. For companies with an international Web presence, this can appear as a temporal effect as Asian, European, and American markets become active during different times of the day. It can also happen that people exhibit different browsing modes on different days of the week, for example, by researching an expensive purchase during lunch hours at work but buying on the weekends.

Now suppose an experiment has two arms, A and B. Arm A is slightly better during the week when people browse but tend not to buy. Arm B is much better during the weekend, when people buy. With enough traffic it is possible for the binomial model to conclude that A is the superior arm prior to seeing any weekend traffic. This is a possibility regardless of whether the experiment is run as a bandit or as a traditional experiment, but the bandit experiment is more susceptible because it will be run for a shorter period of time.

There are two methods that can be used to guard against the possibility of being misled by distinct subpopulations. If the specific subpopulations are known in advance, or if a proxy for them such as the geographically induced temporal patterns is known, then the binomial model can be modified to a logistic regression

$$\text{logit}(p_{a|\mathbf{x}}) = \beta_{0a} + \beta^T \mathbf{x}, \tag{6}$$

where $\mathbf{x}$ is a set of variables describing the context of the observation, $p_{a|\mathbf{x}}$ is the success probability if arm a is played during context $\mathbf{x}$, β_{0a} is an arm-specific coefficient (with one arm's coefficient set to zero), and β is a set of coefficients for the contextual data to be learned as part of the model. For the purposes of running the experiment, the value function may be taken to be $v_a(\theta) = \beta_{0a}$ because the linear term does not impact the order of the actions. (Though for computing summaries such as PVR, one should compute value by averaging Equation (6) over the distribution of $\mathbf{x}$.) Model (6) will give less 'credit' to an arm that produces a conversion inside a context where conversions are plentiful and more credit in contexts where conversions are rare. This can slow down the bandit and help prevent spurious convergence to an arm based on a lucky run during a conversion-rich period.

If the important contexts are unknown, then one may choose to assume that contexts occur as random draws from a distribution of contexts, such as in the beta-binomial hierarchical model,

$$\begin{aligned} \theta_{at} &\sim Be(\alpha_a, \beta_a) \\ y_t | a &\sim Bin(\theta_{at}). \end{aligned} \tag{7}$$

The model parameters here are $\theta = \{\alpha_a, \beta_a : a = 1, 2, \ldots, K\}$, with value function $v_a(\theta) = \alpha_a/(\alpha_a + \beta_a)$. Model (7) will continue exploring for several update periods, even if arm a is dominant during the early periods, to guard against the possibility that the early advantage was simply the result of random variation at the θ_{at} level. As mentioned in Section 3, the random effect distribution in Equation (7) can be more effective at slowing down an aggressive bandit than simply adjusting the optimality probabilities with the binomial bandit.

4.3. Personalization

The models from Equations (6) and (7) purposefully omit interactions between contextual variables and experimental factors because they are intended for situations where the experimental goal is to find a global optimum. Interactions between contextual and design variables allow for the possibility that the optimal arm may differ by context. For example, one version of a page may perform better in Asia while another performs better in Europe. Depending on the granularity of the contextual data available, this approach may be used to personalize results down to the individual level. Similar approaches, though not necessarily using Thompson sampling, are being applied to personalized medicine [23].

4.4. Multivariate testing

Internet companies use the term 'multivariate testing' to describe an experiment with more than one experimental factor. Statisticians would call this the 'multiway layout' or simply a 'designed experiment'. Scott [14] showed how a multifactor experiment can be handled using Thompson sampling with a probit regression model analogous to Equation (1).

Fractional factorial designs (e.g., [24]) are a fundamental tool for handling traditional multifactor experiments. Fractional factorial designs work by finding a minimal set of design points (i.e., rows of a design matrix) that allow a prespecified set of main effects and interactions to be estimated in a linear regression. The equivalent for a multi-armed bandit experiment is to specify a set of main effects and interactions to be estimated as part of the reward distribution. Spike and slab priors can be used to select which interactions to include [25].

For manufacturing experiments, each distinct configuration of experimental units involves changing the manufacturing process, which is potentially expensive. Thus, having a design matrix with a minimal number of unique rows is an important aspect of traditional fractional factorial experiments. With online service experiments, configurations can often be generated programmatically, so that it is theoretically possible to randomize over all potential combinations of experimental factors. The restricted model helps the bandit because there are fewer parameters to learn, but the constraint of minimizing the number of design points is usually no longer necessary.

It is worth noting that the currently accepted 'best practice' among A/B testing frameworks is that experiments should be confined to one factor at a time, on the basis that changing multiple factors will make it difficult to determine which factors are affecting the outcome measured by the experiment. Trained statisticians will recognize this argument as incorrect. Montgomery [26, page 4] points out that 'one-factor-at-a-time experiments are always less efficient than other methods based on a statistical approach to design'. For problems where there are no important interactions, multivariate experiments allow one to simultaneously measure each factor with a single collection of experimental units. To obtain the same precision with standard A/B tests requires many similarly sized experiments. The limitations of A/B testing are more obvious when there are significant interactions between factors, because interactions clearly cannot be measured using single-factor experiments. Readers unfamiliar with these ideas should consult Montgomery [26], Box *et al.* [24], or any number of other excellent books on experimental design.

4.5. Large numbers of arms

Many online experiments involve online catalogs that are too large to implement Thompson sampling as described earlier. For example, when showing ads for commodity consumer electronics products (e.g., 'digital camera'), the number of potential products that could be shown is effectively infinite. With infinitely many arms, attempting to find the best arm is hopeless. Instead, the decision problem becomes whether or not you can improve on the arms that have already been seen. This problem has been studied in [27], among others.

One approach to the problem is as follows. Consider the infinite binomial bandit problem, and assume that the arm-specific success probabilities independently follow $\theta_a \sim Be(\alpha, \beta)$. Suppose K_t arms have been observed at time t. One can proceed by imagining a $K_t + 1$ armed bandit problem, where success probabilities for K_t of the arms are determined as in Section 4.1, but success probabilities for the 'other arm' are sampled from the prior. If the 'other arm' is selected by the Thompson heuristic, then sample an as-yet unseen arm from the catalog.

A similar hierarchical modeling solution can be applied to context-dependent or multivariate problems by assuming that the coefficients of the design variables with many levels are drawn from a common distribution, such as a Gaussian or t. The amount of exploration can be increased by increasing the number of draws from the prior at each stage.

4.6. Example

Consider an experiment involving a button's position and color, the font used on the button, and the background color it is set against. We imagine these factors have 2, 3, 4, and 5 levels, with specific values listed in Table I to make the story concrete. Suppose the page behaves differently on weekends and week days, with conversion rates determined by the logistic regression coefficients listed in Table II. Conversions are generally more common on weekends (thus the larger intercept). Certain fonts (e.g., courier and comic sans) perform better on weekends, as does the background color orange. However, the optimal button color is blue in both cases, as is the optimal background color. Finally, the success probability is set to zero if the button color and the background color are the same, which can be viewed as an interaction between button color and background color not listed in Table II.

Figure 2(a) shows the distribution of success probabilities for the 120 arms in this simulation, while Figure 2(b) shows the cumulative regret obtained by modeling this scenario using four different reward distributions. The 'binomial bandit' is the model from Section 4.1. The 'beta-binomial bandit' is the model from Equation (7). The 'fractional factorial' model is the logistic regression from Section 4.4, ignoring any potential weekend/weekday effect. The 'contextual bandit' augments the logistic regression from the fractional factorial bandit with a linear logistic term for modeling the weekend/weekday

Table I. Factors and levels for the hypothetical example in Section 4.6.

Factor	Levels
Button position	Left, right
Button color	Red, blue, green
Font	Times, arial, courier, comic sans
Background color	Red, orange, yellow, green, blue

Table II. True logistic regression coefficients used in Section 4.6.

Coefficient	Weekend	Weekday
Intercept	−2	−3
Button.position.right	0.02	0.02
Button.color.blue	0.2	0.2
Button.color.green	−0.1	−0.1
Font.arial	−0.3	−0.3
Font.courier	−0.2	−0.5
Font.comic.sans	1.2	0.7
Background.color.orange	−0.3	−1.2
Background.color.yellow	−0.7	−0.7
Background.color.green	0.8	0.8
Background.color.blue	1.4	1.4

The simulation also includes a constraint forcing the success probability to zero if the button color matches the background color.

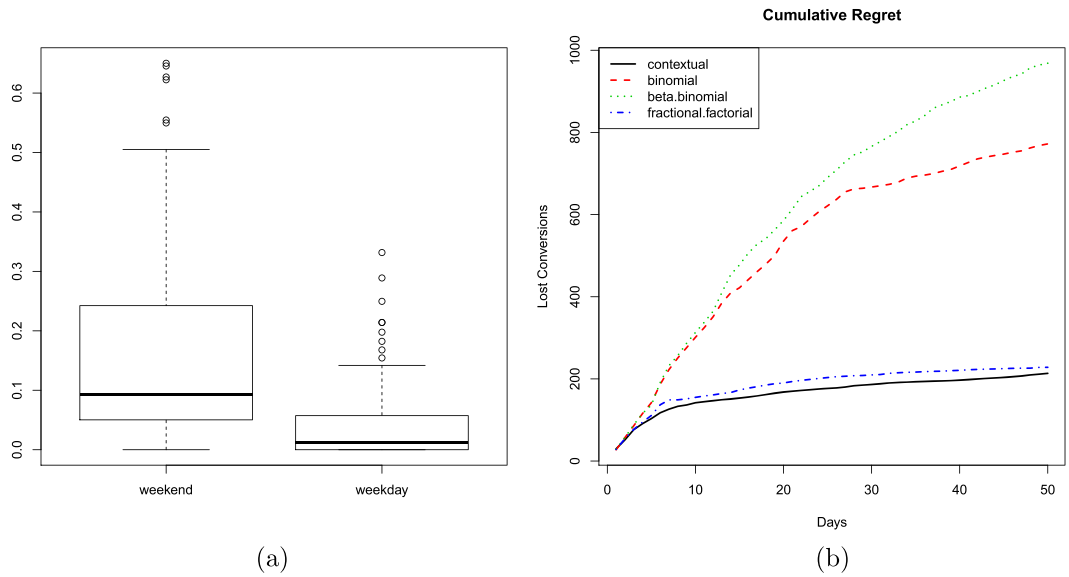

Figure 2. (a) The distribution of true success probabilities, across arms, for the simulation in Section 4.6. (b) The cumulative regret from Thompson sampling assuming four different reward distributions.

effect (but does not allow for day-of-week interactions with experimental factors). The fractional factorial and contextual bandits perform similarly here because here the optimal arm (blue background, red button, on the right, with comic sans font) is the same on weekends and week days, but controlling for the weekend/weekday variation allows the contextual bandit to find the optimal combination slightly faster. The two bandits that consider the multivariate structure of the experiment significantly outperform the two that do not. The preceding simulation was run many times under different seeds. The size of the regret differences between the arms varied from run to run, but the overall ranking of the four algorithms was the same each time.

5. Conclusion

Multi-armed bandit experiments can be a substantially more efficient optimization method than traditional statistical experiments. The Thompson sampling heuristic for implementing multi-armed bandit experiments is simple enough to allow flexible reward distributions that can handle the kinds of issues that arise in real applications.

Business and science have different needs. The classical theory of the design of experiments was created to address uncertainty in scientific problems, which is a naturally conservative enterprise. Business decisions tend to be more tactical than scientific decisions, and there is a greater cost to inaction. Classical design of experiments has been the dominant theory of industrial experiments partly because agriculture and manufacturing happen to have economic structures that align with scientific conservatism. A type I error is costly for manufacturing or agriculture because it means a potentially expensive change to a production environment with no accompanying benefit. In the service economy, type I errors are nearly costless, so artificially emphasizing them over expensive type II errors makes little sense. When paired with the fact that the proportion of type I errors is bounded by the proportion of true null hypotheses, the significance versus power framework underlying traditional statistical experiments seems a poor fit to the modern service economy. By explicitly optimizing value, multi-armed bandits match the economics of the service industry much more closely than traditional experiments, and should be viewed as the preferred experimental framework. This is not to say that there is no room for traditional experiments, but their use should be limited to high-level strategic decisions where type I errors are truly important.

Acknowledgements

I thank two anonymous referees who made helpful comments on an earlier draft of this article.

References

1. World Bank, 2013. http://data.worldbank.org/indicator/NV.SRV.TETC.ZS/countries [Accessed on 2014].
2. Varian HR. Computer mediated transaction. *American Economic Review: Papers and Proceedings* 2010; **100**:1–10.
3. Whittle P. Restless bandits: activity allocation in a changing world. *Journal of Applied Probability* 1988; **25A**:287–298.
4. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA, 1998.
5. Whittle P. Discussion of "bandit processes and dynamic allocation indices". *Journal of the Royal Statistical Society, Series B: Methodological* 1979; **41**:165.
6. Robbins H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* 1952; **58**:527–535.
7. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 1933; **25**:285–294.
8. Gittins JC. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B: Methodological* 1979; **41**: 148–177.
9. Hauser JR, Urban GL, Liberali G, Braun M. Website morphing. *Marketing Science* 2009; **28**:202–223.
10. Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 2002; **47**:235–256.
11. Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 1985; **6**:4–22.
12. Agarwal D, Chen BC, Elango P. Explore/exploit schemes for Web content optimization. *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, IEEE Computer Society Washington, DC, USA, 2009,1–10.
13. Chapelle O, Li L. An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems (NIPS)*, 2011.
14. Scott SL. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 2010; **26**:639–658. (with discussion).
15. May BC, Korda N, Lee A, Leslie DS. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 2012; **13**:2069–2106.
16. Kaufmann E, Korda N, Munos R. Thompson sampling: an asymptotically optimal finite time analysis. *International Conference on Algorithmic Learning Theory*, 2012. http://link.springer.com/chapter/10.1007/978-3-642-34106-9_18.
17. Bubeck S, Liu CY. A note on the Bayesian regret of Thompson sampling with an arbitrary prior, 2013a. arXiv preprint arXiv:1304.5758.
18. Bubeck S, Liu CY. Prior-free and prior-dependent regret bounds for Thompson sampling. In *Advances in Neural Information Processing Systems 26*, Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds). Curran Associates, Inc., 2013b; 638–646.
19. Russo D, Van Roy B. Learning to optimize via posterior sampling. *Mathematics of Operations Research* 2014; **39**(4):1221–1243.
20. Russo D, Van Roy B. Learning to optimize via information directed sampling. *Computing Research Repository* 2014; **abs/1403.5556**. http://arxiv.org/abs/1403.5556 [Accessed on 2014].
21. Scott SL. Google Analytics help page, 2012. https://support.google.com/analytics/answer/2844870?hl=en.
22. Berry DA. Adaptive clinical trials: the promise and the caution. *Journal of Clinical Oncology* 2011; **29**:606–609.
23. Chakraborty B, Moodie EEM. *Statistical Methods of Dynamic Treatment Regimes*. Springer: New York, 2013.
24. Box GE, Hunter JS, Hunter WG. *Statistics for Experimenters*. Wiley: New York, 2005.
25. Box G, Meyer RD. Finding the active factors in fractionated screening experiments. *Journal of Quality Technology* 1993; **25**:94–94.
26. Montgomery DC. *Design and Analysis of Experiments* (5th edn). Wiley: New York, 2001.
27. Berry DA, Chen RW, Zame A, Heath DC, Shepp LA. Bandit problems with infinitely many arms. *Annals of Statistics* 1997; **25**:2103–2116.