Social Scientific Banditry*

Improving Experimental Designs through Adaptive Treatment Allocation

Drew Dimmery†

Early Draft: November 16, 2015

Abstract

In this paper, I discuss the use of adaptive treatment allocation based on the interim results of an experiment. I draw from the literature on multi-armed bandit algorithms and provide identification and consistency conditions for the estimation of treatment effects in adaptive experiments. I go on to consider how this might be used by social scientists and develop a method to generate a sequence of treatment assignment probabilities which will converge to an optimal treatment allocation in terms of asymptotic efficiency. This adaptive allocation method provides an estimator that is consistent and more efficient than static assignment. Methods are also assessed which optimize an outcome of interest (such as voting).

*Replication code available by request. All code written in R and Python and available on request from the author.

†New York University Department of Politics, 19 W. 4th St. Second Floor, New York, NY, 10012. Email: drewd@nyu.edu Web: http://drewdimmery.com

1 Introduction

The guiding principles of experimental design were layed out in large part in 1926 from R.A. Fisher working in Rothamsted, England (Fisher, 1926, 1935). These principles remain the foundation on which essentially all modern social scientific experimental designs are based (Gerber and Green, 2012). In particular, these designs begin with a simple scheme of randomization fixed at the beginning on an experiment, which is then used to justify a causal interpretation to resulting inferences. In this paper, I present results showing how, through the judicious use of machine learning, it is possible to improve on static, conventional designs through the use adaptive treatment allocation wherein the allocation scheme is allowed to evolve over time in response to observed data. The primary benefit of adaptive treatment allocation is in asymptotic guarantees of optimality (impossible with a conventional design). I present two such varieties of optimality.

This paper will be structured as follows: I will begin by discussing extant veins of thought on adaptive/sequential designs. I will then discuss the problems inherent in using adaptive treatment allocation. I will then present identification conditions which allow consistent estimation of treatment effects under generic adaptive allocation regimes. Next, I develop an adaptive allocation mechanism which has desirable optimality properties in terms of the variance for differences in means. This mechanism provides an asymptotic guarantee that it will outperform (in terms of variance) any conventional design. I follow up by demonstrating this with simulation studies, concluding by analyzing the likely improvements in variance to be expected in typical large scale political science field experiments.

2 Overview of Adaptive Design

The history of sequential design began, more or less, with Robbins (1952). In this paper and many of those which have followed in its footsteps, the ultimate goal is to ensure that of two treatments, a treatment assignment rule converges asymptotically to applying the "best" treatment almost surely. This has since spawned a robust literature on "multi-armed bandits": in reference to the "one-armed bandit" – a slot machine – wherein one is faced with an array of slot machines and one must choose which machine pays off at the highest

rate (and then play that machine as often as possible).

The multi-armed bandits literature has existed primarily in the field of computer science, as the goal has been to develop algorithms to better converge to the best treatment (or "arm" in the parlance of the literature). See, for example, a literature review by (Bubeck and Cesa-Bianchi, 2012) and the broader work on reinforcement learning (Sutton and Barto, 1998). The primary intuition behind these algorithms is that developing an algorithm to maximize a particular objective by playing a particular sequence of arms consists of two competing problems: exploration – in which one tries to learn the payoffs associated with particular arms – and exploitation – in which one plays the arms currently understood to be the best (in the sense of a given objective function).

By way of illustration of this tradeoff, I will provide an explanation of two algorithms approaching the bandit problem from two different directions. First, one can consider uniform exploration. Under a particular set of assumptions, uniform exploration is the optimal means by which to estimate the means of the distribution underlying each arm (Cochran, 1977). Thus, at the end of any given time-span, this method ensures that the experimenter knows the true means with great confidence. The problem, of course, is that this is very ineffective at exploitation. That is, if we wanted to play the arm with the largest expected value, a strategy of uniform exploration would only do so with probability of one over the number of arms. An easy to implement solution to this problem is the so-called ''epsilon-greedy'' algorithm. This algorithm allocates to arms with probability $1 - \epsilon$ on the arm currently believed to be best, and the remainder ϵ probability split equally among all other arms known to be inferior. When ϵ is very small, this algorithm will cheat almost entirely to exploitation. The difficulty, of course, is that many arms will not have their means estimated particularly well. This means that it is quite possible that this algorithm might mistakenly exploit the *wrong arm* for extended periods of time. Thus, exploration and exploitation are fundamentally at odds with one another.[1]

There are two classes of algorithms which I will describe in somewhat greater detail. These are Thompson sampling based approaches and Confidence Bound based approaches. The former is largely heuristic, but often performs quite reasonably. The latter comes

[1]For an accessible introduction to implementing and comparing some of these basic algorithms, see White (2012).

with particular theoretical guarantees, but may not perform quite as admirably in practice. There is an additional class of index-based methods which I will not dwell on here (Gittins, Glazebrook and Weber, 2011), as it is much less common both in the literature and in application. It also has undersirable properties as an 'incomplete learning' method.

Thompson Sampling has been acknowledged as a useful and successful heuristic for managing the tradeoff between exploration and exploitation. The basic idea of this approach is to construct a probability model of all arms, and then allocate among arms equal to the probability that a particular arm is the best arm (this allocation scheme has been referred to as 'random probability matching' in Scott (2010)). This approach has a long history, and ultimately dates back to Thompson (1935). This approach is one of the oldest to the bandit literature, and has been a very common approach used in practice, but only relatively recently was it shown to have good asymptotic properties (Agrawal and Goyal, 2012). Its practical effectiveness cannot be denied (Scott, 2010). Straightforward Thompson Sampling, however, requires the specification of a full parametric model connecting arms to outcomes which may be problematic. A related approach is to rely, instead, on a Bootstrap approximation to the relevant posterior and use that for determining the appropriate distribution with which to adjudicate how to assign treatment (Eckles and Kaptein, 2014).[2]

Perhaps the most rigorously studied of algorithms for multi-armed bandits are those of the UCB (upper confidence bound) variety. In broad strokes, these algorithms have been referred to as exploiting "optimism in the face of uncertainty". These were some of the first algorithms to be studied rigorously (Lai and Robbins, 1985; Lai, 1987; Agrawal, 1995). The main intuition of these approaches is that one should find an upper confidence bound for the mean of each arm, and allocate a given unit to the arm with the largest upper confidence bound. Then, in each period, the upper tail probability associate with that confidence bound should be adjusted downwards. By doing so, even arms which appear to be much worse than the best arm will still be played occasionally once the upper tail probability is sufficiently low. Since the best arm will have an upper confidence bound very close to its mean, the arms with more associated uncertainty (those that look suboptimal) will still be played occasionally. In theory, this approach makes a good tradeoff between

[2]These two approaches, Thompson Sampling and Bootstrap Thompson Sampling have been championed by researchers at Google and Facebook, respectively.

exploration and exploitation. In practice, however, it often does not perform as well as more heuristic methods. An array of advances in UCB algorithms have been made over time. Perhaps the most significant advancement was that of Auer, Cesa-Bianchi and Fischer (2002) showing that the logic of UCB approaches were not merely asymptotic, but provided similar optimality guarantees in finite samples.

Additional advancements in the bandits literature have worked to use more information and to use information more effectively. For example, the contextual bandits literature added in contextual (covariate) information (Langford and Zhang, 2008) to inform a more fine-grained view of the optimality of arms. That is, perhaps the best arm is different for men and women, so different strategies are necessary for each. I will be implicitly in this world for much of this paper, insomuch as I consider strategies for individual strata defined by covariates. Recent work has focused on better utilizing information in this setting, for example through the use of a linear model (Filippi et al., 2010; Dani, Hayes and Kakade, 2008). Other work has sought to utilize more flexible models to share this information (Srinivas et al., 2009).

There is an additional literature examining adaptive survey sampling (Thompson, Seber et al., 1996). This literature tends to focus on things like adaptive cluster sampling, as it tends to consider the problem through a lens of sampling hard-to-reach populations (Seber and Salehi, 2012). Nevertheless, this literature examines how to draw coherent inferences about populations based on samples retrieved from an adaptive process, when one cannot completely ignore the sampling process in analysis.

This paper is devoted to the application of adaptive experimental design and, in particular, the lessons coming from the multi-armed bandits literature on the necessity to make a principled tradeoff between exploration and exploitation.

3 Sampling Setup

Define s_n to be an unordered sample of size n from the population $\boldsymbol{p}$.

$$\boldsymbol{s}_n = \{(x_{i_1}, d_{i_1}, y_{i_1}, i_1), \ldots, (x_{i_n}, d_{i_n}, y_{i_n}, i_n)\}$$

Where x_j is the observed covariate information (before treatment is applied to any unit) for a unit with label j, d_j is j's treatment information, y_j is j's observed outcome data, and j is just the unit's unique label. If a given unit was sampled more than once, it would only appear in s_n once. For notational simplicity, I assume that draws are without replacement (i.e. n represents the cardinality of s_n). Our infinite population, $\boldsymbol{p}$ is the set from which all elements of $\boldsymbol{s}_n$ are drawn. When the cardinality of the sample is not important, I may drop the subscript n for notational convenience. Also note that asymptotics assume that $s_n \subset s_{n+1}$ (since as we collect more data, we retain the data we previously collected).

Moreover, consider that y_j for some unit is a function of the covariates of that unit (observed x_j and unobserved u_j) and treatment d_j. The outcome, y_j, is a function of these variables, $y_j(d_j, x_j, u_j)$, as in the potential outcomes framework(Imbens and Rubin, 2015). The fundamental problem of causal inference in this context is that we wish to infer the unobserved additive causal effect of manipulating d_j to d'_j: $y_j(d_j, x_j, u_j) - y_j(d'_j, x_j, u_j)$. We can never observe this quantity, as we may only ever observe unit j with either d_j or d'_j and never both(Holland, 1986).

Instead, this paper will focus attention on the conditional average treatment effect, $\tau(x) = \mathbb{E}_{\boldsymbol{p}}[y_j(d_j, x_j, u_j) - y_j(d'_j, x_j, u_j)|x_j = x]$. In particular, note that this is the expectation over the entire population. Moreover, I will seek conditions under which sample statistics based on s_n can recover this population quantity. The next section will provide intuition about the specific goals of this search, as well as potential problems and solutions.

4 The Design-based approach

A sampling design is considered "design unbiased" when:

$$\mathbb{E}[f(\boldsymbol{s})|\boldsymbol{s}] = \mathbb{E}[f(\boldsymbol{p})]$$

where $f(\cdot)$ is some statistic from a set of data. On the left side, the expectation is conditional on the observed sample, while the right side is the expectation on the statistic in the population. The key takeaway from this definition is that no conditioning on information about the sampling design is necessary, only the sample itself.

Conventional designs with no adaptive allocation are design unbiased. This allows

for inference on population parameters without accounting for the sampling procedure. That is, it is not necessary to condition on the sampling *procedure* in order to estimate unbiased estimates for population quantities. Instead, it is merely necessary to examine the realized sample of data. This forms the basis of much model-agnostic inference in the social sciences, as no assumptions as to either the underlying population nor any model are necessary to define clear population estimands with feasible sample estimators.

This property is very desirable. For the estimation of a mean, for instance, this implies that an unweighted sample mean will be unbiased for the estimation of the population mean. When a sampling design does not have this property, however, it becomes necessary to use an estimator such as a Horvitz-Thompson or Hajek estimator in order to attain an unbiased estimate (Cochran, 1977). These methods, however, have a particular feature which may not be desirable: in order to ensure unbiasedness, different units are weighted differently according to their sampling probabilities. This is problematic, as units which were very likely to have been sampled contribute very little information to the final estimate. This, in turn, increases the variance of the sampling distribution of the resulting estimator, as it must rely more heavily on the few units which were unlikely to be sampled. If this is the only way to recover population quantities, this is fine, but it implies that it would be preferable to avoid using them when unnecessary.

Note that since the potential outcomes framework interprets causal effects of d on y as fundamentally a missing data problem(Rubin, 1978), the realized sample can be interpreted as a sample from a notional full population in which units are observed under every possible value of treatment. It is inference on this notional population which we imbue with the label of 'causal inference'.

4.1 Bias in Adaptive Designs

I will now provide two simulations as examples of the way in which adaptive designs are problematic when the ultimate goal is causal inference.

First, suppose that there is an adaptive allocation procedure which simply increases the probability of sampling from the Treatment group over time, while the conditional means of the potential outcomes also vary over time as seen in figure 1[3]. The reason

[3]All units are drawn from a normal distribution centered on the conditional mean shown in figure 1.

for bias in this case is clear. Treatment potential outcomes are sampled more for units with lower average potential outcomes, leading to a biased treatment effect. Put simply, there is a failure to sample uniformly over potential outcomes. This can be fixed by using an estimator which considers the sampling probabilities such as a Horvitz-Thompson or Hajek estimator of the group means. Figure 2 shows the sampling distributions of these estimators. In this plot, the conventional design is what the sampling distribution would look like if a conventional design allocating to each treatment group with fixed probability of one half. The important takeaway is that the Horvitz-Thompson estimator is a cure worse than the disease. As is omnipresent in statistics, there is a clear bias-variance tradeoff at play. While the Horvitz-Thompson estimator is unbiased, it's variance is very large. The Hajek estimator performs better than the Horvitz-Thompson, but retains a relatively larger variance than the uncorrected adaptive design.[4] This design admits a small variance of its sampling distribution. Indeed, its sampling distribution is essentially indistinguishable in shape from that of the conventional design. This reduced variance, however, comes at the cost of bias, as it is not centered at the true causal effect.

Yet this is not the only way that an adaptive design may admit bias. To illustrate another way that adaptive designs are problematic, I turn to a different data generating process. Suppose that the previous example's problem of changing response surfaces does not exist, and both Treatment and Control are simply draws from a standard normal distribution. Now suppose that we fix an adaptive design wherein we stop sampling from Treatment if it looks sufficiently "bad" (that is, if the sample mean ever drops below a threshhold I set at 0.01). While this may sound like an unnatural design, it closely accords to adaptive designs which are commonly used in practice. For instance, medical trials may stop (and may be *required by law* to stop) when a treatment appears to be killing patients. The sampling distribution for the unweighted difference in means is shown for adaptive and conventional (no dropping of treatments) designs in figure 3. More troublingly, since this design simply stops sampling from Treatment when it drops below a threshhold, it admits inconsistency as well. No amount of sampling from Control can make up for the mistaken impression of the mean of the outcome under Treatment.

[4]Of course, the Hajek estimator also admits bias, unlike the Horvitz-Thompson.

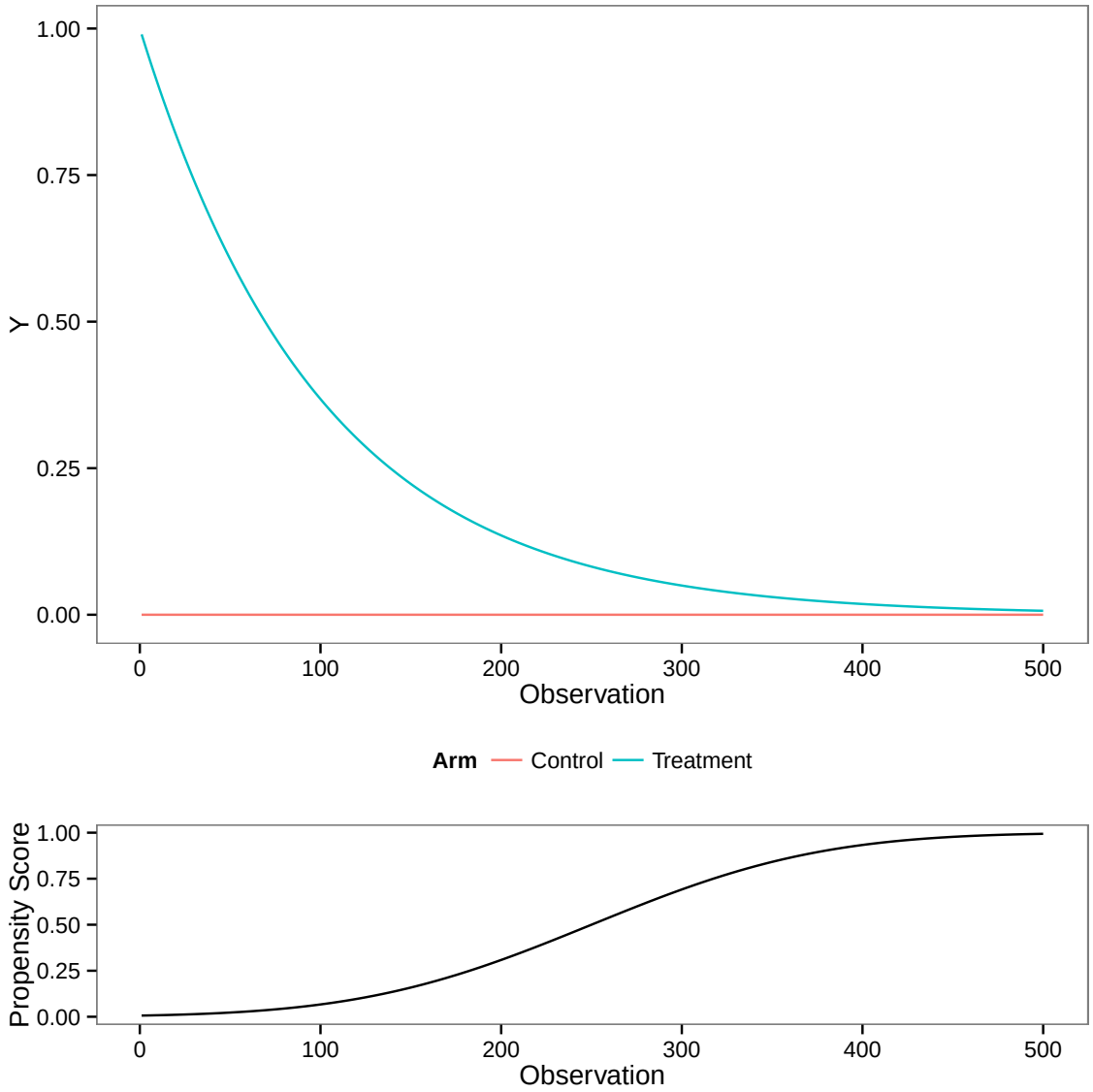

Figure 1: The top panel shows the conditional means of the response surface from which units are drawn in this simulation. While the control remains constant, the response to treatment varies systematically over time. The bottom panel shows how the probability of assignment to treatment varies over time.

4.2 Design Consistency

I now introduce a property of "design consistency". The previous section showed that adaptive designs can introduce bias and inconsistency to resulting design-independent estimators. However, in general, the history of statistics has shown that we have many reasons to expect estimators with small amounts of bias to nonetheless have more desirable asymptotic properties such as consistency and efficiency (Efron and Morris, 1975). Indeed, I will show that adaptive sampling designs that are design consistent are amenable to simple methods of estimation which are *almost surely* more efficient than any other weighted

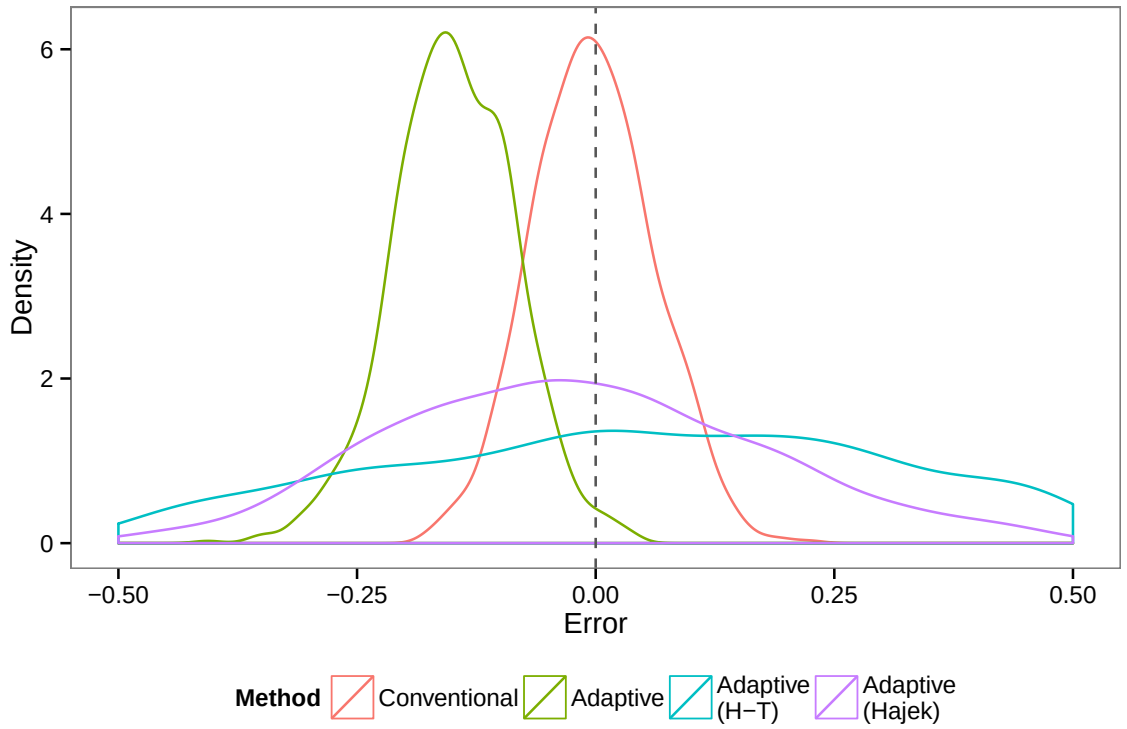

Figure 2: This figure shows the sampling distribution for three ways of estimating an average treatment effect. The conventional design (using a simple differences in means) is unbiased and clustered tightly around the true ATE. The differences in means under an adaptive design is similarly tightly clustered, but biased downwards. The Horvitz-Thompson estimator for the ATE under the adaptive design is unbiased, but highly variable. The Hajek estimator for the ATE is better, but still more variable than would would prefer.

estimator such as Hajek or Horvitz-Thompson.

Definition 1:

Design Consistency

$$\lim_{n\to\infty} \mathbb{E}[f(\boldsymbol{s}_n)|\boldsymbol{s}_n] = \mathbb{E}[f(\boldsymbol{p})]$$

In other words, this embeds the common understanding of a consistent estimator. While the statistic may have bias in small samples (as shown in the previous subsection), it is possible to reduce that bias to an arbitrarily small level given sufficient data. The next section will present conditions to ensure that an experimental design will be consistent by design, so that no adjustment for sampling scheme is necessary.

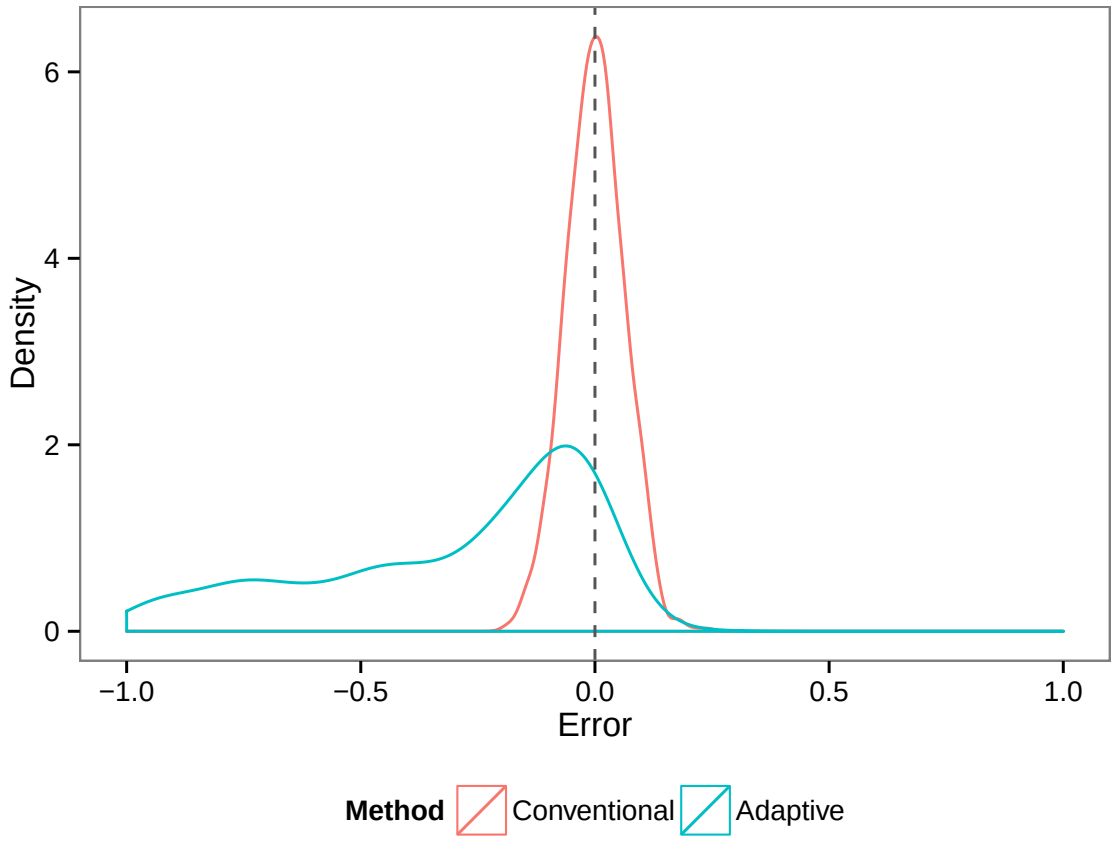

Figure 3: While the conventional design is tightly clustered around the ATE, the example adaptive design exhibits a clear downward bias from the ATE.

5 Identification in Adaptive Designs

For the following identification results, I presume only a dichotomous treatment, but this is only for ease of notation. Results hold for any finite number of discrete treatments. Moreover, similar results may be achieved for a continuous treatment when the goal is to estimate a dose-response function.

I begin by specifying necessary assumptions for my identification and consistency results.

Assumption 1:

No Self Selection - Treatment is assigned as a function of the pre-treatment covariates of all units, the vector of observed outcomes for (potentially) all other units as well as (potentially) a randomization device (call it V) independent of all other variables. That is, $D_i = g(\bar{X}_i, \bar{Y}_{i-1}, \bar{D}_{i-1}, V)$ where a variable's history, $\bar{Z}_i$, is the vector such that $(z_1, \ldots, z_i)$.

In other words, a new unit, i, is assigned to treatment based on the covariates of potentially every other previously observed units (including unit i), but only the history of treatment assignments and outcomes of earlier units (that is, excluding unit i). Crucially, this explicates the necessity that treatment be well-controlled by the experimenter, and that individuals may not choose their own treatment based on unobserved or unobservable characteristics.

Assumption 2:

SUTVA - The potential outcome of a unit i depends only on its own realization of treatment. That is, $Y_i = Y_i(d_i) \text{if} D_i = d_i$

This standard assumption implies that there is only one version of a given treatment, and that a unit's observed outcome is dependent on only their value of treatment (no spillover or interference). It should also be noted that this simplifies the discussion of causation, as it avoids the necessity of dealing with more complicated causal estimands, as in Robins, Hernan and Brumback (2000) or Blackwell and Glynn (2013). That is, SUTVA implies an exclusion restriction that only a unit's treatment status has the potential to induce a causal effect on that unit. Other units have no such potential. This can be thought of as a kind of Markovian assumption (combined with the assumption of no self selection), in which the state of the system (insomuch as it is collected in $g(\cdot)$) embeds all relevant aspects of the history of data.

Assumption 3:

Positivity - Any unit receiving any treatment is a random variable with strictly positive probability:

$$p(D_i = 1) \in (0, 1)$$

This assumption additionally requires the existence of a randomization mechanism, implying that there is a non-degenerate probability distribution over V in the assignment mechanism, g.

This suffices for our primary identification result, a simple application of the propensity score theorem (Rosenbaum and Rubin, 1983):

Theorem 1:

Any randomized adaptive design identifies the CATE when assumptions 1, 2 and 3 are met.

Proof.

$$\tau(x) \equiv \mathbb{E}[\boldsymbol{Y}(1) - \boldsymbol{Y}(0)|X = x] = \mu_1(x) - \mu_0(x)$$

By assumptions 1 and 2, we have strong ignorability:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i|(\bar{X}_i, \bar{D}_{i-1}, \bar{Y}_{i-1})$$

Notate $Z_i = (\bar{X}_i, \bar{D}_{i-1}, \bar{Y}_{i-1})$ so that $p(D_i = 1|\bar{X}_i, \bar{D}_{i-1}, \bar{Y}_{i-1}) = \mathbb{E}_V[g(\bar{X}_i, \bar{Y}_{i-1}, \bar{D}_{i-1}, V)] = e(Z_i)$

Since the distribution of $g(\cdot)$ is known by design (recall that V is the only random component of this function), this propensity score is known and controlled. This quantity, $e(Z_i)$ therefore, is a fixed, known quantity. It is not estimated.

Adding assumption 3, we have, further, the following:

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i|e(Z_i)$$

At this point, it is clear that we simply have an instance of unequal probability sampling of a given potential outcome surface.

For the dth surface:

$$\mathbb{E}[\boldsymbol{Y}(1)|D_i = d, X_i = x] = \mathbb{E}[\frac{1}{n(x)} \sum_{i:X_i=x} Y_i(d) \times 1(D_i = d)] = \sum_{i:X_i=x} Y_i(d) \times \mathbb{E}[1(D_i = d)] = \sum_{i:X_i=x} Y_i(d) \times e(Z_i)$$

Thus,

$$\mathbb{E}[\frac{1}{n(x)} \sum_{i:X_i=x} \frac{Y_i(d)}{e(Z_i)} \times 1(D_i = d)] = \mathbb{E}[\frac{1}{n(x)} \sum_{i:X_i=x} Y_i(d)] = \mu_d(x)$$

This demonstrates that the CATE is identified through a Horvitz-Thompson estimator. □

Note that it is not necessary for stability to hold in this case, as conditioning on the propensity score breaks the dependence between treatment and potential outcomes (within

covariate strata).

This depends crucially on the design ($e(Z_i)$), however. Due to this fact, the variance is greatly inflated, as this entails (potentially very) different weights on individual observations, which implies a larger resulting sampling variance of the estimate of the CATE. Since adaptive designs will tend to greatly reduce the frequency at which particular treatments are assigned, this will have the typical effect of vastly increasing the variance of estimates. In essence, adaptive designs often try to drive assignment probabilities to zero or one. The more effective at this they are, the larger the sampling variance of a Horvitz-Thompson style estimator.

Consider, for instance, that a unit treated when its probability of treatment is 0.01 would receive 50 times more weight than would a unit treated when its probability of treatment is .5. If these units are, in fact, simply random draws from a common distribution, the two observations each provide equal amounts of information about the underlying population. The differences in weights, however, widen the sampling distribution of resulting statistics. This helps to motivate an identification result which does not rely upon propensity scores.

I next provide an identification condition which resolves this difficulty, relying in part on intuition from Kasy (2013) that randomization of the treatment need not be the basis for causal inference. For this, I provide a weaker version of a positivity assumption:

Assumption 4:

Weak positivity

$$\lim_{n\to\infty} n_d(x) = \infty \quad \forall d \in \{0,1\} \text{ and } x \in \mathcal{X}$$

This assumption does not rely on randomization, but instead merely requires that, for each treatment arm d, the sample receiving that treatment goes to infinity. Each arm may approach infinity at a substantially different rate. Indeed, many adaptive designs will seek to ensure that this occurs.

Assumption 5:

Stability / i.i.d. Sampling: The joint distribution of the potential outcomes and covariate

vectors of each unit are independent and identically distributed.

$$p(\boldsymbol{Y}(0), \boldsymbol{Y}(1), \boldsymbol{X}) = \prod_{i=1}^{n} = p(Y_i(0), Y_i(1), X_i)$$

This assumption may be satisfied quite easily if units are sampled randomly from some large population. If z_i is the vector of potential outcomes and relevant covariates for unit i, then if z_i is drawn in a simple random sample from the large population of all units in the sampling frame, then this assumption will be satisfied by design.

The following provides a design consistency result for the conditional average treatment effect. With consistency of the CATE, inference on the ATE follows trivially through a matching estimator when the distribution of X is known. Note that thanks to assumption 5, the distribution of X in the sample follows that in the population.

Theorem 2:

When assumptions 1, 2, 5 and 4 are met and $E[\boldsymbol{Y}(d)] < \infty \ \ \forall d \in \{0, 1\}$, it is not necessary to condition on the design to identify the CATE at infinity:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} y_i \times 1(D = 1, X = x)}{\sum_{i=1}^{n} 1(D = 1, X = x)} - \frac{\sum_{i=1}^{n} y_i \times 1(D = 0, X = x)}{\sum_{i=1}^{n} 1(D = 0, X = x)} \to \tau(x)$$

Proof. We begin by examining the asymptotic behavior of the first term, the sample mean of Y under treatment.

Assumptions 5 and 4 show that $\sum_{i=1}^{n} 1(D = 1, X = x)$ will grow to infinity, and given the boundedness of the expectation of the potential outcomes under treatment, the law of large numbers implies that:

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} Y_i(d) \times 1(D_i = d, X_i = x)}{\sum_{i=1}^{n} 1(D_i = d, X_i = x)} = \mathbb{E}[\boldsymbol{Y}(d)|X_i = x] = \mu_d(x)$$

The same argument holds for Control, sufficing to show that the sample difference in means will be asymptotically valid for the CATE.

□

Since consistency is an asymptotic condition, then so, too, is identification for a differ-

ence in means estimator. Estimators which take into account the design may be used, but may have very undesirable properties in terms of variance. With sufficient sample sizes, design consistency provides a justification for retaining the more efficient design-independent estimator. Consider it a very simple bias/variance tradeoff. Admitting a bit of bias into estimation allows for greater efficiency.

Variance estimation proceeds along an elementary central limit theorem, providing for asymptotically normal inference, conditional on the existence of a finite second moment for each potential outcome surface:

$$\hat{\tau}(x) - \tau(x) \xrightarrow{\mathcal{D}} N\left(0, \frac{\sigma_1(x)}{n_1(x)} + \frac{\sigma_1(x)}{n_0(x)}\right)$$

where $\sigma_d(x)$ is the conditional variance of treatment d when $X = x$. Simply replacing these population quantities with their sample analogs provides an easy means of estimation and population level inference directly through the central limit theorem.

I next show that the ubiquitous UCB1 algorithm (**?**) meets the conditions for design consistency:

Theorem 3:

The strategy which, in each period plays $d^*(x_t)$ (where x_t is the covariate profile for the next period of play), allows asymptotically design free causal inference assuming 2 and 5 for a binary outcome.

$$d^*(x_t) = \arg\max_d \left[\hat{\mu}_d(x_t) + \sqrt{\frac{2\log n(x_t)}{n_d(x_t)}}\right)$$

Proof. Assumption 1 holds by design, and assumptions 2 and 5 are given.

All that remains is assumption 4 to show that UCB1 provides design consistent causal inference.

For any estimated gap $\hat{\Delta}(x)$ between the observed best d and second best d' arms for a given context x, it may be seen that there exists some n^*, such that d' will be played rather than d. This is the case when $\hat{\mu}_d(x) + \sqrt{\frac{2\log n(x)}{n_d(x)}} < \hat{\mu}'_d(x) + \sqrt{\frac{2\log n(x)}{n'_d(x)}}$. To simplify, d' will be played when: $\hat{\Delta}(x) < \sqrt{\frac{2\log n(x)}{n'_d(x)}} - \sqrt{\frac{2\log n(x)}{n_d(x)}}$. It then suffices to note that the second term

on the right hand side will converge to zero as the sample size increases, while the first term will diverge to infinity. That is, eventually the uncertainty in the estimate expressed on the right hand side will overwhelm the difference in estimated conditional means, implying that d' will be sampled for some n^*. This ensures that assumption 4 holds.

□

While this demonstrates that UCB1 allows for design consistent causal inference, note that it does *not* apply to methods based on the Gittins Index. As shown in Brezzi and Lai (2000), these approaches will be consistent for estimating the mean of *only one arm*. Thus, if causal inference is the goal, this class of methods is inadmissable if one demands the property of consistency (as one should).

5.1 Discussion

A comparison to static treatment assignment is useful here, to better understand the implications of the identifying conditions of adaptive experimentation. In particular, it is important to acknowledge the differences in the assumptions underlying the assignment mechanism. Typical accounts of the necessities for a strongly ignorable treatment assignment require three fundamental components: 1) individualistic assignment 2) probabilistic assignment and 3) unconfounded assignment (Imbens and Rubin, 2015). The first condition implies that a unit's assignment function depend only on it's own covariates and potential outcomes. The second that assignment be non-deterministic and occur with probability between zero and one. The final condition implies that assignment occur without reference to potential outcomes. All three of these conditions are relaxed through the above identification conditions. The first condition is relaxed by allowing assignment to depend on the covariates, treatment status and observed outcomes of prior units. The second is relaxed (in the second provided result) by allowing deterministic assignment to treatment. The third condition is more or less preserved, in the sense that a particular unit's assignment mechanism cannot depend on it's *own* potential outcomes. It's worth noting, however, that a unit's assignment can depend, in part, on the observed outcomes of previously observed units.

There are two ways that these identification conditions may be weakened. First, stability may be weakened if one is willing to allow for modeling of the changing response surfaces.

This, however, may lose some of the appealing aspects of an experiment. Continuing research looks into the possibility of conditioning on the changing design in a non-parametric manner by batching adaptive updates, providing a set of easy to analyze mini-experiments (Bakshy, Dimmery and White, 2015). This, of course, loses the continuous updating benefits of the adaptive experiment, however.

The logic of these results, however, provides some useful intuition as to how researchers should and should not apply adaptive experimentation techniques. If one desires consistent estimation, it is not acceptable to stop sampling any treatment arm at any point in time. It is useful to consider widely used approaches in practice and whether they meet these conditions.

5.2 Potential Difficulties

Post-treatment bias in the collection of covariate information can play a complicated role in the estimation of treatment effects in this context. If a covariate is not observed prior to the application of treatment for all units, then there is an additional necessary exclusion restriction that x_i cannot be effected by treatment for any $j < i$. If this is not the case, then stratification cells themselves are subject to treatment effects, leaving the estimand inherently ill-defined (and a moving target).

6 Optimal Design

I now turn to an application of adaptive experimental design to the optimal design of experiments. The Neyman allocation (proportional to standard deviation) is an optimal design, but depends on an **unknown quantity** (Cochran, 1977; Thompson, 1992). There have primarily been two solutions: use prior knowledge to determine these quantities, or use a two stage design to estimate them. Neither provide asymptotic guarantees of optimality. The status quo in political science often eschews even such two stage designs.

The two stage design was first proposed in Robbins (1952) which began much of the literature on the sequential design of experiments. I extend the intuition of this design to the continuously updating case in which one progresively uses the data coming in from an experiment to update the design of the experiment to ensure optimality.

This problem may be framed as an exploration (estimation of standard deviations)

versus exploitation (estimated Neyman allocation) tradeoff, which motivates an adaptive design. I propose just such an adaptive design which will use the following assignment probabilities to allocate a new unit to a treatment group:

$$p(\text{allocate to arm } d | X = x) \propto \hat{s}_d(x) + \epsilon \frac{\hat{s}_d(x)}{\sqrt{2(n_d(x) - 1)}} \tag{1}$$

I take $\epsilon = 3$. A one tailed Chebyshev's inequality makes this an approximate 90% upper confidence bound (approximate due to the $O(n^{-3/2})$ asymptotic approximation to the standard error of the standard deviation).

This will asymptotically converge to the Neyman allocation as $n(x) \to \infty$, providing a guarantee of asymptotic optimality for the estimation of treatment effects.

This allocation rule is by necessity design consistent for the estimation of treatment effects whenever the true standard deviations of every strata/treatment combination is non-zero. This may be readily seen insomuch as the sample standard deviation is necessary non-negative, implying that all allocation probabilities are themselves non-zero.

7 Simulations

I now perform simulations to explore the additonal effective sample sizes possible through the use of adaptive treatment assignment.

My metric of interest will be the "compound efficiency" between two different sampling designs, d and d'. This is essentially a size independent measure of effective sample size (design effect). That is, if the compound efficiency is 0.80, then that implies that the variance of the adaptive design provides around 20% more effective sample size (in terms of the variance of the resulting estimator) for a given number of units sampled.

$$\text{Compound efficiency}_{d,d'} = \frac{\sum_{t \in \mathcal{T}} \text{var}(\bar{x}_{t,d})}{\sum_{t \in \mathcal{T}} \text{var}(\bar{x}_{t,d'})}$$

That is, compound efficiency measures the relative sum of the variances for the sample mean ($\bar{x}_{t,d}$) of each arm (t) for a given sampling design (d).

In all of the following simulations, the DGP for arms $k \in \{1, 2, 3, 4\}$ is $X_k \sim \mathcal{N}(1, 1)$.

The remaining arm has standard deviation of 1, 2 or 4 (uniform, double and quadruple, respectively).

For the set of simulations in figure 4, the base design (d') is static uniform sampling over arms. Thus, all compound efficiency scores can be interpreted as the improvement relative to this baseline when using adaptive Neyman sampling. The results here demonstrate the non-maleficense of an adaptive design. Even in the case on the far left, in which the base design is, in fact, exactly correct (the optimal design is uniform assignment), the adaptive design does not perform significantly worse. When the true optimal allocation diverges from uniform, my method can be seen to lead to increases in compound efficiency of twenty to fourty percent.

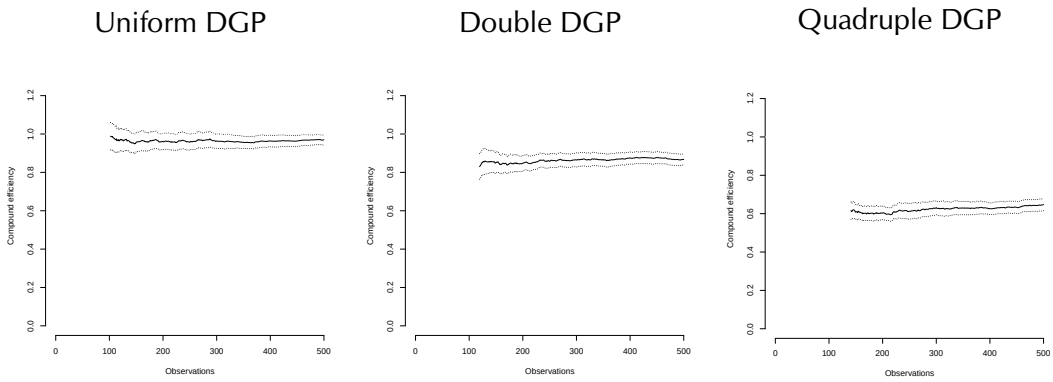

Figure 4: Compound efficiency of my adaptive design relative to a uniform design.

In the set of simulations in figure 5, 5 units are selected from each arm in the first stage and used to estimate the standard deviation. These standard deviations are then used to provide a static approximation to the Neyman allocation. The key takeaway from this set of simulations is that even when a pilot study provides unbiased estimates of the Neyman allocation, there are still significant gains to be made by updating adaptively. In these simulations, gains are on the order of ten to twenty percent.

7.1 Empirical Simulation

Simulations in figure 6 from the observed variation in strata in existing studies. Examined are Arceneaux (2007), Gerber, Green and Larimer (2008) and Bolsen, Ferraro and Miranda (2014). The base design in each of these charts is from the authors' allocation to strata and to treatment groups. That is, I assume here that it is possible to adaptively allocate units both to treatment groups and to stratification cells. In practice, this would mean

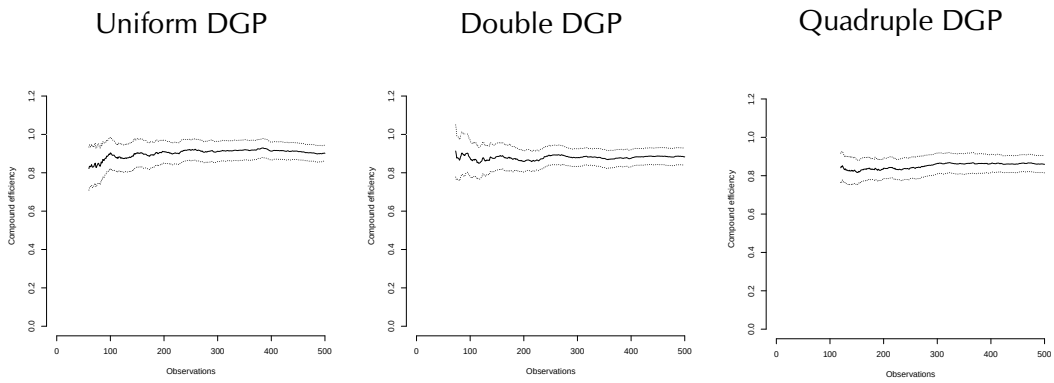

Figure 5: Compound efficiency of my adaptive design relative to a simple two stage design.

that individual units could be chosen to enter the experiment based on their pre-existing covariates. I do not mean to suggest that it would be possible to alter the attributes of these individuals. In effect, this is merely asking, at each stage of the experiment, what kinds of people would be most useful to inform the eventual treatment effect estimates.

This simulation demonstrates the types of gains feasible for real world studies. That is, what sorts of gains in effective sample size are possible by using adaptive allocation rather than the static allocation (by the criteria used by the original authors)? The results show that gains of around ten percent are reasonable on even the most well-designed of studies, and that gains of twenty percent may often be reasonable. In this particular study (Bolsen, Ferraro and Miranda, 2014), the dependent variable was water usage, a continuous measure. Since an unbounded variable provides greater potential for variation across strata, there is greater opportunity for gains to be made through judicious designs.

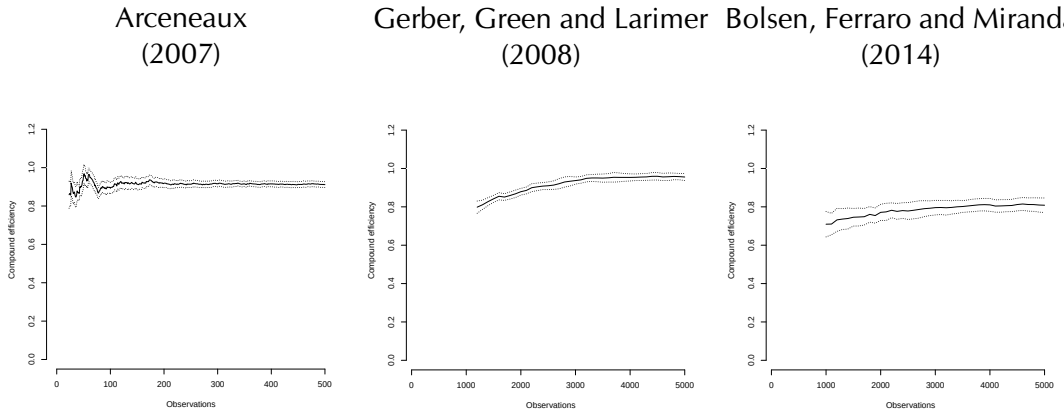

Figure 6: Compound efficiency of my adaptive design relative to three existing large-scale field experiments.

8 Conclusion

In this paper, I have discussed adaptive treatment allocation through a lens of the multi-armed bandit problem and show how it can provide insights relevant to social scientists. By providing identification conditions for these experiments, I showed that experimenters need not fear controlled adaptation of treatment allocations. I demonstrated how some algorithms for determining treatment assignment from the bandit literature are consistent for the estimation of treatment effects, and provided a new method which brings with it asymptotic guarantees of optimality in the estimation of treatment effects.

References

Agrawal, Rajeev. 1995. "Sample mean based index policies with O (log n) regret for the multi-armed bandit problem." *Advances in Applied Probability* pp. 1054–1078.

Agrawal, Shipra and Navin Goyal. 2012. "Further optimal regret bounds for Thompson sampling." *arXiv preprint arXiv:1209.3353* .

Arceneaux, Kevin. 2007. "I'm Asking for Your Support: The Effects of Personally Delivered Campaign Messages on Voting Decisions and Opinion Formation." *Quarterly Journal of Political Science* 2(1):43–65.
URL: *http://dx.doi.org/10.1561/100.00006003*

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

Bakshy, Eytan, Drew Dimmery and John Myles White. 2015. "Design-based Adaptive Experimentation." CODE@MIT.

Blackwell, Matthew and A Glynn. 2013. How to Make Causal Inferences with Time-Series Cross-Sectional Data. In *Conference on Political Methodology*.

Bolsen, Toby, Paul J. Ferraro and Juan Jose Miranda. 2014. "Are Voters More Likely to Contribute to Other Public Goods? Evidence from a Large-Scale Randomized Policy Experiment." *American Journal of Political Science* 58(1):17–30.
URL: *http://dx.doi.org/10.1111/ajps.12052*

Brezzi, Monica and Tze Leung Lai. 2000. "Incomplete Learning from Endogenous Data in Dynamic Allocation." *Econometrica* 68(6):1511–1516.
URL: *http://dx.doi.org/10.1111/1468-0262.00170*

Bubeck, Sébastien and Nicolo Cesa-Bianchi. 2012. "Regret analysis of stochastic and nonstochastic multi-armed bandit problems." *Foundations and Trends in Machine Learning* 5:1–122.

Cochran, William G. 1977. *Sampling techniques*. John Wiley & Sons.

Dani, Varsha, Thomas P Hayes and Sham M Kakade. 2008. Stochastic Linear Optimization under Bandit Feedback. In *COLT*. pp. 355–366.

Eckles, Dean and Maurits Kaptein. 2014. "Thompson sampling with the online bootstrap." *arXiv preprint arXiv:1410.4009* .

Efron, Bradley and Carl Morris. 1975. "Data Analysis Using Stein's Estimator and its Generalizations." *Journal of the American Statistical Association* 70(350):311–319.

Filippi, Sarah, Olivier Cappe, Aurélien Garivier and Csaba Szepesvári. 2010. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*. pp. 586–594.

Fisher, R.A. 1926. "The Arrangement of Field Experiments." *Joiurnal of the Ministry of Agriculture of Great Britain* 33:503–513.

Fisher, Ronald A. 1935. "The Design of Experiments.".

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gerber, Alan S., Donald P. Green and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102:33–48.

Gittins, John, Kevin Glazebrook and Richard Weber. 2011. *Multi-armed bandit allocation indices*. John Wiley & Sons.

Holland, Paul W. 1986. "Statistics and causal inference." *Journal of the American statistical Association* 81(396):945–960.

Imbens, Guido W and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kasy, Maximilian. 2013. "Why experimenters should not randomize, and what they should do instead.". Please email me for password to the MATLAB files, which generate optimal designs for your data.

Lai, Tze Leung. 1987. "Adaptive Treatment Allocation and the Multi-Armed Bandit Problem." *The Annals of Statistics* 15(3):1091–1114.

Lai, Tze Leung and Herbert Robbins. 1985. "Asymptotically efficient adaptive allocation rules." *Advances in applied mathematics* 6(1):4–22.

Langford, John and Tong Zhang. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*. pp. 817–824.

Robbins, Herbert. 1952. "Some aspects of the sequential design of experiments." *Bulletin of the American Mathematical Society* 58(5):527–535.

Robins, James M., Miguel Angel Hernan and Babette Brumback. 2000. "Marginal structural models and causal inference in epidemiology." *Epidemiology* 11:550–560.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Rubin, Donald. 1978. "Bayesian inference for causal effects: The role of randomization." *The Annals of Statistics* .

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

Seber, George AF and Mohammad Mahmoud Salehi. 2012. *Adaptive sampling designs: inference for sparse and clustered populations*. Springer.

Srinivas, Niranjan, Andreas Krause, Sham M Kakade and Matthias Seeger. 2009. "Gaussian process optimization in the bandit setting: No regret and experimental design." *arXiv preprint arXiv:0912.3995* .

Sutton, Richard S and Andrew G Barto. 1998. *Reinforcement learning: An introduction*. Vol. 1 MIT press Cambridge.

Thompson, Steven K. 1992. *Sampling*. Wiley.

Thompson, Steven K, George Arthur Frederick Seber et al. 1996. *Adaptive sampling*. Wiley New York.

Thompson, William R. 1935. "On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation." *The Annals of Mathematical Statistics* 6(4):pp. 214–219.

White, John. 2012. *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.".