Approximation Methods for Determining Optimal Allocations in Response Adaptive Clinical Trials

Vishal Ahuja[1], John R. Birge[2] and Christopher Ryan[2]
[1]*The University of Chicago, Center on Aging at NORC, 1155 E 60th Street, Chicago, IL 60637, U.S.A.*
[2]*The University of Chicago Booth School of Business, 5807 S Woodlawn Ave., Chicago, IL 60637, U.S.A.*

Keywords: Adaptive Clinical Trials, Markov Decision Process, Grid Approximation, Approximate Dynamic Programming.

Abstract: Clinical trials have traditionally followed a fixed design, in which patient allocation to treatments is fixed throughout the trial and specified in the protocol. The primary goal of this static design is to learn about the efficacy of treatments. Response-adaptive designs, where assignment to treatments evolves as patient outcomes are observed, are gaining in popularity due to potential for improvements in cost and efficiency over traditional designs. Such designs can be modeled as a Bayesian adaptive Markov decision process (BAMDP). Given the forward-looking nature of the underlying algorithms which solve BAMDP, the problem size grows as the trial becomes larger or more complex, often exponentially, making it computationally challenging to find an optimal solution. In this study, we propose grid-based approximation to reduce the computational burden. The proposed methods also open the possibility of implementing adaptive designs to large clinical trials. Further, we use numerical examples to demonstrate the effectiveness of our approach, including the effects of changing the number of observations and the grid resolution.

1 INTRODUCTION

The costs of bringing a new drug to market have been estimated to be as high as $5 billion (Forbes, 2013). Clinical trials have been cited as a key factor in raising these costs; the total cost of a clinical trial can reach $300–$600 million (English et al., 2010), potentially an order of magnitude higher when including the value of remaining patent life. Consequently, drug manufacturers face pressure to produce conclusive results faster and reduce the number of subjects.

Traditional clinical trials follow a non-adaptive or *fixed* randomized designs, where patients are randomly assigned to treatments and are used widely. Although such designs provide a clean way of separating treatments and are well-understood by most practitioners, they are becoming increasingly costly and often end up producing inconclusive results. Consequently, regulatory bodies, such as the U.S. Food and Drug Administration, have recently encouraged the use of adaptive designs (FDA, 2010).

Response-adaptive designs for clinical trials, typically Bayesian in nature, are gaining in popularity. Such designs employ learn-and-confirm concepts, accumulating data on patient responses to make procedural modifications while the trial is still underway, increasing the likelihood of selecting the *right* treatment for the *right* patient population earlier in a drug development program. As a result, adaptive designs can potentially reduce costs and shorten overall development timelines significantly.

Bayesian adaptive designs are rooted in the *multi-armed bandit* problem that requires balancing reward maximization based on the knowledge already acquired with attempting new actions to further increase knowledge, commonly referred to as the *exploitation vs. exploration* tradeoff. Berry was one of the pioneers, who used this formulation in the clinical trials context (e.g. (Berry, 1978)).

Sequential allocation designs are the most common form of response-adaptive designs (e.g. (Berry and Fristedt, 1985)), where patients are treated one at a time (in a sequence), and each patient's responses is available before making an allocation decision for the next patient. (Ahuja and Birge, 2014) extends this model to incorporate simultaneous allocation of multiple patients and show that this results in an improved objective function value (e.g., expected patient successes) compared to naive implementation of sequential designs, thus substantially widening the potential

Ahuja V., R. Birge J. and Ryan C..
Approximation Methods for Determining Optimal Allocations in Response Adaptive Clinical Trials.
DOI: 10.5220/0004909504600465
In *Proceedings of the 3rd International Conference on Operations Research and Enterprise Systems* (ICORES-2014), pages 460-465
ISBN: 978-989-758-017-8

for applicability of such designs.

A major barrier to implementing adaptive designs in practice is computational. Bandit problems in clinical trials context are typically modeled as MDP's, where the solution is obtained by solving a finite-horizon dynamic program (Ahuja and Birge, 2014). However, the problem size increases exponentially as the number of time periods, patients, or treatment-outcome combinations increase, commonly referred to as the *curse of dimensionality* (Powell, 2007). As a result, a direct application of dynamic programming becomes computationally prohibitive and finding an optimal policy to this high-dimensional problem becomes challenging (Bertsimas and Mersereau, 2007).

Approximation techniques address this problem and allow users to find a solution by reducing the problem size and the associated computational burden. When the underlying problem is modeled as an MDP, the possible approximation techniques are generally collectively referred to as approximate dynamic programming (ADP). There currently exist several techniques to approximate the value function or state space or both. Although some techniques are more popular than others, ADP still remains more of an art than a science.

In this study, we use a grid to approximate the state space. We numerically evaluate the optimality loss or the loss in objective value with respect to fully enumerated solution, as a result of this approximation. Our proposed approach has implications for clinicians and policymakers interested in finding an efficient yet easily implementable design for large clinical trials, where currently existing adaptive designs either cannot be implemented or do not perform or scale well.

The rest of the paper is organized as follows. §2 provides an overview of the literature. §3 presents the model and the proposed approximation method. We present numerical results in §4. We conclude in §5.

2 LITERATURE

Several methods and techniques have been proposed in the literature for approximate solutions to large dynamic programs (see (Powell, 2007) for a discussion). Lagrangian decomposition-based ADP approach is one such method to approximate the value function (e.g. (Adelman and Mersereau, 2008)). The approach has been used to find approximate solution in interactive marketing (Bertsimas and Mersereau, 2007, e.g.) and retail assortment (Caro and Gallien, 2007, e.g.). Another method is to use polynomials, for example least squares approximation using Chebyshev polynomials (Judd, 1998). (Ahuja and Birge, 2014) is the only study that has used approximation in the context of adaptive designs for clinical trials; they use a truncated-horizon or limited-lookahead approximation method.

Grid-based methods are commonly used to approximate the state space. These techniques sample a finite number of points, called the grid, from the entire state space, compute the value of the points in the grid and approximate the values of the non-grid points via some form of interpolation (Sandikci, 2010).

There exists a rich literature on the grid-based approximation including notable studies within the operations management literature (Monahan, 1982; Lovejoy, 1991; Aviv and Pazgal, 2005), as well in the computer science literature (Hauskrecht, 1997; Zhou and Hansen, 2001). (Sandikci et al., 2013) is an example of a recent study that uses a grid-based approximation approach in a healthcare setting to approximate the position of the patient on the waiting list.

There are several approaches for grid-based approximation that depend on the grid construction choices, for example, uniform vs. non-uniform grid, fixed vs. variable resolution grid, etc. In general, all the corner points of the probability simplex are included in the grid since that eliminates the need to extrapolate (see (Sandikci, 2010) for a brief overview). In this paper, we use a fixed-resolution uniform grid since that allows for an efficient interpolation.

While grid-based approximation methods have been studied and implemented before, our contribution lies in the efficient use of such methods in the clinical trials context, specifically to the response-adaptive designs for clinical trials, thus widely broadening the practical applicability of such designs. In a later study, we provide bounds on optimality gap while noting that the solution obtained by approximation is a lower bound on the optimal solution obtained from a fully enumerated problem.

3 MODEL

We follow the *Bayes-adaptive Markov decision process* (BAMDP) model developed in (Ahuja and Birge, 2014). The state in the BAMDP model is a vector with dimension equal to the number of treatment-outcome combinations, also called *health conditions*. The state thus captures the information observed so far (history) and is used to derive the distributions that describe the uncertainty in the transition probabilities.

We first re-define the state in terms of fraction of patient observations within each health condition. Each state dimension then represents the fraction of patient observations observed so far in a given health

condition, where the fractions sum up to one.

The key idea behind the approximation approach is to cap the problem size by discretizing the fractions that form the component of each state, thus limiting the state space irrespective of the number of patients and time periods. Such a setup allows us to choose, ahead of time, a constant number of health states that are evaluated explicitly at each time period, thereby keeping the problem tractable and reducing the computational burden, often substantially. However, this leads to some optimality loss with respect to the solution obtained from a fully enumerated problem. Calculating theoretical bounds on the optimality loss is a subject of future work. The rest of the parameters and modeling assumptions remain the same as in (Ahuja and Birge, 2014).

3.1 General Model Specification

Let T be the trial length, n be the number of patients allocated per period in the trial, and $N = nT$ be the total number of patients (observations) in the trial. Let J and O be the set of treatments and outcomes, respectively. The corresponding set of health conditions, I, is then the Cartesian product $(J \times O)$.

The information state is a vector $\mathbf{h}_t \in \mathcal{H} \subseteq \mathbb{Z}^{|J|\times|O|}$, defined as $\mathbf{h}_t = (h_t^{1,1}, ..., h_t^{|J|,|O|})$, where $h_t^{j,o} \in \mathbb{Z}_+$ represents the the cumulative number of observed patients to date in health condition (j,o) at time $t \in \{0,1,..,T\}$, for all $j \in J$, $o \in O$, such that $\sum_{j\in J, o\in O} h_t^{j,o} = nt$.

The controls,$\mathbf{u}_t \in \mathcal{U} \subseteq \Re_+^{|J|}$ are defined as $\mathbf{u}_t = (u_t^1, ..., u_t^{|J|})$, where $u_t^j \in [0,1]$ is the probability of assigning a patient to treatment $j \in J$ at time $t \in \{0, ..., T-1\}$ such that $\sum_{j\in J} u_t^j = 1$. The set of decisions, $\mathbf{d_t}$, is random and obtained from the controls, are defined as $\mathbf{d}_t = (d_t^1, ..., d_t^{|J|})$. Here $d_t^j \in \mathbb{Z}_+$ is the number of patients assigned to treatment $j \in J$ such that $\sum_{j\in J} d_t^j = n$, $Pr(\mathbf{d}_t|n, \mathbf{u}_t) \sim Mu(\mathbf{d}_t; n; \mathbf{u}_t)$[1], and $\mathbb{E}d_t^j = nu_t^j$. Patients begin arriving at $t = 1$, and decisions for patients arriving at t are made at $t-1$, and no decision is made at $t = T$.

Finally, the probabilities are defined as $\mathbf{p}_t^j = (p_t^{j,1}, .., p_t^{j,|O|})$, where, $p_t^{j,o}$ represents the probability of observing outcome $o \in O$ at time $t+1$ given treatment $j \in J$ at time t. We assume a generalized multinomial likelihood on the transition to state $\mathbf{h}_{t+1}$ from state $\mathbf{h}_t$, given $\mathbf{p}_t$, and use a Dirichlet conjugate prior on $\mathbf{p}_t$ with hyperparameters $\alpha_t = (\alpha_t^{1,1}, ..., \alpha_t^{|J|,|O|})$ for $t \in \{0, ..., T\}$. If we denote the initial priors by $\alpha_0 = (\alpha_0^{1,1}, ..., \alpha_0^{|J|,|O|})$ and assume that the outcomes of patients in different health conditions are not informative of each other, then each $\alpha_t^{j,o}$ can be updated independently as follows: $\alpha_t^{j,o} = \alpha_0^{j,o} + h_t^{j,o}$, where $h_t^{j,o}$ captures all the (random) realizations from the past for that treatment-outcome combination.

Given the decision $\mathbf{d}_{t-1}$, the (random) outcomes are observed in the next period, captured in the vector $\mathbf{k}_t \in \mathcal{K} \subseteq \mathbb{Z}^{|J|\times|O|}$, that we define as the *physical state*, $\mathbf{k}_t^j = (k_t^{j,1}, .., k_t^{j,|O|})$. Here, $k_t^{j,o} \in \mathbb{Z}_+$ represents the number of observed patients in health condition (j,o) at a given time $t \in \{1, ..., T\}$, where the treatment $j \in J$ is given at time period $t-1$ and the outcome $o \in O$ is observed in time t, such that $\sum_{j\in J, o\in O} k_t^{j,o} = n$. The above definitions directly imply the following: for $t = 1$, $\mathbf{h}_t = \mathbf{k}_t$ and for $t = 2, ..., T$, $\mathbf{h}_t = \mathbf{h}_{t-1} + \mathbf{k}_t$.

The entries of the transition matrix at time $t \in \{0, ..., T-1\}$, $P_t(\mathbf{h}_{t+1}|\mathbf{h}_t, \mathbf{d}_t, \alpha_0)$, representing the probability of transitioning to state $\mathbf{h}_{t+1}$, given $\mathbf{h}_t$, $\mathbf{d}_t$, and α_0, is then defined as follows:

$$P_t(\mathbf{h}_{t+1}|\mathbf{h}_t, \mathbf{d}_t, \alpha_0) = \prod_{j\in J} Pr(\mathbf{k}_{t+1}^{j,\cdot}|\mathbf{h}_t^{j,\cdot}, d_t^j, \alpha_0^{j,\cdot}) = \prod_{j\in J} \int_0^1 Pr(\mathbf{k}_{t+1}^{j,\cdot}|d_t^j, \mathbf{p}_t^{j,\cdot}) g(\mathbf{p}_t^{j,\cdot}|\mathbf{h}_t^{j,\cdot}, \alpha_0^{j,\cdot}) dp_t^{j,\cdot}, \quad (1)$$

if $d_t^j \in \mathbb{Z}$ and $k_{t+1}^{j,o} \leq d_t^j$ for all $j \in J$, $o \in O$, and 0 otherwise. Here, $Pr(\mathbf{k}_{t+1}^j|d_t^j, \mathbf{p}_t^j) = Pr(k_{t+1}^{j,1}, ..., k_{t+1}^{j,|O|}; d_t^j; p_t^{j,1}, .., p_t^{j,|O|})$ is the multinomial likelihood or the marginal joint distribution of observing $k_{t+1}^{j,1}, ..., k_{t+1}^{j,|O|}$ outcomes from d_t^j patients given that the probability of observing these outcomes is $p_t^{j,1}, .., p_t^{j,|O|}$, respectively, and $g(\mathbf{p}_t^j|\mathbf{h}_t^j, \alpha_0^j) = g(\mathbf{p}_t^j|\alpha_t^j) = g(p_t^{j,1}, .., p_t^{j,|O|}; \alpha_t^{j,1}, ..., \alpha_t^{j,|O|})$ is the pdf for the Dirichlet distribution.

Finally, the reward, R_t, is defined for each objective function as follows: (a) Patient Health: $R_T = 0$ and $R_t = \mathbf{r}^T \mathbf{k}_{t+1}$ $\forall t \in \{0, .., T-1\}$, where $\mathbf{r} \subseteq \Re^{|J|\times|O|}$, and (b) Learning: $R_T = \max_{j\in J} Pr\{p_T^j(\tilde{o}|\mathbf{h}_T) > \max_{j'\in J\setminus\{j\}} \{p_T^{j'}(\tilde{o}|\mathbf{h}_T)\}\}$ and $R_t = 0$ $\forall t \in \{0, .., T-1\}$, where $\tilde{o} \in O$ is the desired outcome.

The entire formulation is a dynamic program, in which the objective is to maximize the expected value function (V_t) that captures expected total reward and solves the Bellman equation as follows:

$$V_t(\alpha_t, \beta_t) = \max_{\mathbf{u}_t} \{R_t + \mathbb{E}_{\mathbf{k}_{t+1}}[V_{t+1}(\alpha_{t+1}, \beta_{t+1})]\}. \quad (2)$$

[1] Mu denotes multinomial distribution.

3.2 Grid-based Approximation of the State Space

We approximate the state space using a uniform grid, where each grid point, that we call *grid state*, represents a health state $\tilde{\mathbf{h}}_t \in \tilde{\mathcal{H}} \subseteq \Re^{|J|\times|O|}$, defined as

$$\tilde{\mathbf{h}}_t = (\tilde{h}_t^{1,1}, ..., \tilde{h}_t^{|J|,|O|}),$$

where $\tilde{h}_t$ has the same dimensionality as h_t and $\tilde{\mathcal{H}}$ is the *approximate* state space.

The number of grid points at each time is a function of the grid resolution, q^s, where a higher resolution implies a finer grid and a larger state space. In this paper, we use a fixed-resolution grid, implying that the number of states at each time period are the same but note that it is easy to incorporate variable-resolution grid, one that varies with time. Each grid state can then be described in terms of q^s as follows:

$$\tilde{h}_t^{j,o} = \frac{x}{q^s},$$

where $x \in \{0,1,2,...,q^s\}$. q^s provides a lever for adjusting the granularity of the fraction that we can use to modify how refined (big) or coarser (small) the state space is. In other words, q^s allows us to tradeoff between a close approximation (and hence a higher objective value) and the computational burden imposed as a result.

A direct consequence of using grid-based approximation is grid state transitions may not belong to the grid state space and requires approximation. To illustrate, suppose the state to which $\tilde{\mathbf{h}}_t$ transitions to at time $t+1$ is denoted by $\mathbf{h}_{t+1} = (h_{t+1}^{1,1}, ..., h_{t+1}^{|J|,|O|})$, where $h_{t+1}^{j,o} = \dfrac{nt\tilde{h}_t^{j,o} + k_{t+1}^{j,o}}{n(t+1)}$. If $h_{t+1} \in \tilde{\mathcal{H}}$, then there is no need to approximate the state (and consequently $\mathbf{V}_{t+1}$) as we have an exact match. However, if $h_{t+1} \notin \tilde{\mathcal{H}}$ we interpolate the value function, as defined in §3.3. The optimal solution is still obtained by solving the Bellman equation in (2).

3.3 Value Function Interpolation

We estimate the value function of this transition state, $\mathbf{h}_t$, by combining values at neighboring states (called vertices of the simplex) to obtain an approximation. For an n-dimensional state, this implies taking linear combinations of the values at grid points of the simplex that surround the state whose value needs to be approximated. This leads to to a linear system with $n+1$ equations. We formulate this interpolation problem as a linear program (LP), where the objective is to maximize the sum of rewards, as shown below.

$$\begin{aligned} max \quad & \lambda_t{}^T \mathbf{V_t}, \\ \text{s.t. } \mathbf{h}_t &= \sum_{k=1}^{|\mathcal{H}|} \lambda_t^k \tilde{\mathbf{h}}_t^k, \\ \sum_{k=1}^{|\mathcal{H}|} \lambda_t^k &= 1, \\ \lambda &\geq 0. \end{aligned}$$

Here, $\mathbf{h}_t$ represents the state whose value function needs to be approximated using grid states at time t, $\tilde{\mathbf{h}}_t^k$ represents the k^{th} state amongst the set of grid states, $\mathbf{V_t}$ represents the associated set of (known) value function of the grid states, and λ_t are the coefficients that the LP solves for. The constraints and the relationship $0 \leq \tilde{h}_t^{j,o} \leq 1$ ensures that all corner points of the simplex are included amongst the grid states. A consequence of this approximation is the potential loss in optimality, which we discuss further in the numerical results (see §4).

The model works as follows. First consider the terminal period, T, where no decision needs to be made. For the second to last time period, since the transitions happen into the terminal stage, there is no more ambiguity. The value function is simply a *dot* product of the state and the corresponding reward vector representing the value of being in that state. However, for a given state in any other time period, $\tilde{\mathbf{h}}_t$, $t \in \{1,..,T-2\}$, the state to which it transitions to may not belong to the grid state space, in which case it needs to be approximated as defined above.

4 NUMERICAL RESULTS

In this section, we perform numerical analyses under various scenarios to demonstrate how the policy derived from the grid-based approximation approach, π_{AO}, compares with the optimal policy. Our choice of optimal policy for the case of multiple patients is the *Jointly Adaptive* policy of (Ahuja and Birge, 2014), that we denote as π_{JO}. Unless otherwise stated, we make the following assumptions. We consider two treatments, henceforth referred to as treatments A and B, and two mutually exclusive outcomes, namely success (s) and failure (f) as defined earlier. This implies the following: $J = \{A,B\}$, $O = \{s,f\}$, and $I = \{As, Af, Bs, Bf\}$. It follows then that $\tilde{\mathbf{h}}_t = (\tilde{h}_t^{As}, \tilde{h}_t^{Af}, \tilde{h}_t^{Bs}, \tilde{h}_t^{Bf})$ for all $t \in \{1,..,T\}$. Consequently, the assumed distribution that is used to derive transition probabilities reduces to a beta-binomial model with a beta prior distribution and a binomial likelihood resulting in a beta posterior distribution. We define additional terms as follows: $\alpha_t^{j,s} = \alpha_t^j$,

$\alpha_t^{j,f} = \beta_t^j$, $\alpha_t = (\alpha_t^A, \alpha_t^B)$, $\beta_t = (\beta_t^A, \beta_t^B)$ $p_t^{j,s} = p_t^j$ and $p_t^{j,f} = 1 - p_t^j$.

The prior distribution on the probability of success with treatment j at time t is then given as $g(p_t^j) \sim Beta(\alpha_t^j, \beta_t^j)$ and $\mathbb{E}p_t^j = \frac{\alpha_t^j}{\alpha_t^j + \beta_t^j}$. Given that the likelihood of observing k_{t+1}^j successes out of d_t^j is Binomial, i.e. $Pr(k_{t+1}^j | d_t^j, p_t^j) \sim Bin(k_{t+1}^j; d_t^j; p_t^j)$, the posterior distribution of p_{t+1}^j is given as $g(p_{t+1}^j) \sim Beta(\alpha_t^j + k_{t+1}^j, \beta_t^j + d_t^j - k_{t+1}^j)$. The joint posterior probability distribution is then the product of individual probabilities. In the absence of any knowledge of treatment efficacy, a commonly assumed starting prior is non-informative, i.e., $(\alpha_0^j, \beta_0^j) = (1,1)$ for all $j \in J$, equivalent to a uniform[0,1] distribution. Finally, the rewards are defined for each objective function. For health, following existing literature (e.g., (Berry, 1978)), $\mathbf{r} = (1,0,1,0)$, implying a reward of 1 for success and 0 for failure.

For numerical illustration, we only consider the patient health objective and further let $\mathcal{S}_t$ denote the value function (V_t) for this objective.

4.1 Calculating Performance of approximately Optimal Policy

The comparison is between $\mathcal{S}^{\pi_{JO}}$ and $\mathcal{S}^{\pi_{AO}}$, where calculation of $\mathcal{S}^{\pi_{JO}}$ has been defined in (Ahuja and Birge, 2014). However, a meaningful comparison requires the application of approximately optimal policy to the problem instance where no approximation is done that we call a *fully enumerated* problem and whose state space we denote as $\widehat{\mathcal{H}}$. In other words, we first calculate π_{AO} by solving the Bellman equation (using grid-based approximation), given in (2) and then apply it to the fully enumerated problem.

Given that in general, the approximate state space is smaller than the fully enumerated space, application of π_{AO} to $\widehat{\mathcal{H}}$ requires finding the grid-state in $\tilde{\mathcal{H}}$, say $\tilde{\mathbf{h}}_t$ that is "closest" to the fully-enumerated state in $\widehat{\mathcal{H}}$, say $\widehat{\mathbf{h}}_t$ and then applying $\pi_{AO}(\tilde{\mathbf{h}}_t)$ to $\widehat{\mathbf{h}}_t$. To find the grid-state that is closest to the fully-enumerated state, we use nearest-neighbor interpolation, one that minimizes L_1 *norm*.[2]

We compare the two policies under multiple scenarios that vary in the number of patient observations (N) and starting priors, measured as parameters of beta distribution, (α_0^j, β_0^j), $j \in \{A,B\}$. We used 91 unique combinations of starting priors, same as used in (Ahuja and Birge, 2014).

[2] L_2 or the *Euclidean norm* yeilds similar results.

Table 1 lists the expected proportion of successes for all 91 combinations of starting priors under both policies ($\frac{\mathcal{S}^{\pi_{JO}}}{N}$, $\frac{\mathcal{S}^{\pi_{AO}}}{N}$) when $q^s = 12$ under various scenarios. For comparison purposes, we also list the expected proportion of successes under the fixed design (π_{EA}) as well as the following heuristics - *Greedy* (π_{Gr}), *GGreedy* (π_{GG}), *UCB1* (π_{UC}), and *BK* (π_{BK}), where the policies have been defined in (Ahuja and Birge, 2014). Comparison with the fixed design other heuristics provides a measure of the performance of approximation algorithm, where we note that the heuristics may not be feasible for large problem sizes. We note from the table that π_{AO} improves patient successes compared to fixed designs in most of the cases, although some heuristics such as π_{Gr} provide a superior performance.

The following quantity provides a measure of loss in optimality (using expected proportion of successes) as a result of using the approximation approach: $\delta_{AO} := \frac{\mathcal{S}^{\pi_{JO}} - \mathcal{S}^{\pi_{AO}}}{\mathcal{S}^{\pi_{JO}}}$. Figure 1 shows how δ_{AO} varies with the number of time periods (alternately, N) and the grid resolution (q^s) when the initial priors are assumed to follow a uniform[0,1] distribution.

Observations from the figure, include, first, δ_{AO} is increasing in N but decreasing in q^s, both of which make sense and are expected. The increase of δ_{AO} in N is expected because a bigger problem size (a function of N) increases optimality loss. The decrease of δ_{AO} in q^s also makes sense because a higher q^s creates a finer grid with more grid states that can be used for approximating the true state, thus minimizing opportunities for optimality loss. We note that δ_{AO} can be substantial but given that we are comparing the two policies for small problem sizes, where calculating exact optimal solution is feasible, this may not be as surprising. It is worth reiterating that this comparison is only possible for states for which it is computationally feasible to solve the fully enumerated problem.

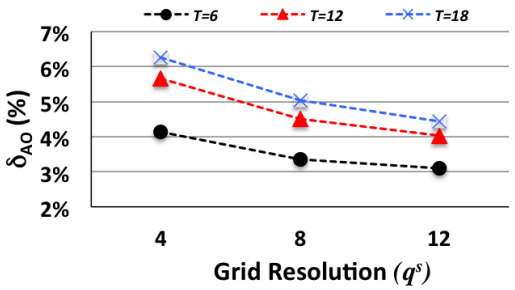

Figure 1: δ_{AO} as a function of q^s and T; $n = 4$ and $(\alpha_0^A, \beta_0^A) = (\alpha_0^B, \beta_0^B) = (1,1)$.

Table 1: Expected proportion of successes for a variety of problem scenarios when $q^s = 12$.

					FIXED	ADAPTIVE		HEURISTICS			
(α^A,β^A)	(α^B,β^B)	n	T	N	Equal	Jointly	Approximately	Greedy	GGreedy	UCB1	BK
					(π_{EA})	(π_{JA})	(π_{RA})	(π_{Gr})	(π_{GG})	(π_{UC})	(π_{BK})
(1,1)	(1,1)	4	6	24	0.5000	0.6132	0.5862	0.6127	0.6127	0.6096	0.6049
(1,1)	(1,1)	4	12	48	0.5000	0.6333	0.5978	0.6321	0.6321	0.6250	0.6232
(2,1)	(1,4)	4	12	48	0.6667	0.6691	0.6464	0.6686	0.6410	0.6339	0.6611
(1,4)	(1,4)	4	6	24	0.2000	0.2495	0.2393	0.2482	0.2482	0.2476	0.2414
(1,4)	(1,4)	4	12	48	0.2000	0.2614	0.2427	0.2603	0.2603	0.2574	0.2531
(4,4)	(4,4)	4	6	24	0.5000	0.5480	0.5402	0.5479	0.5479	0.5434	0.5438
(4,4)	(4,4)	4	12	48	0.5000	0.5614	0.5457	0.5608	0.5608	0.5502	0.5547
(4,1)	(1,4)	4	12	48	0.8000	0.8001	0.7928	0.8000	0.7642	0.7593	0.7958
(4,1)	(4,1)	4	6	24	0.8000	0.8516	0.8372	0.8497	0.8497	0.8470	0.8476
(0.5,0.5)	(6,6)	4	12	48	0.5000	0.6405	0.6024	0.6355	0.6292	0.6233	0.6313

5 FUTURE WORK

In the near-term, we aim complete this work grid-based approximation methods. While the numerical results provide a sense of the optimality loss with respect to optimal solution, work is underway to establish theoretical bounds on optimality loss. Further, we plan to perform numerical analyses to demonstrate the magnitude of computational burden that can be reduced by implementing our proposed method. We also plan to compare our approach with other approximation approaches that have been proposed in the literature. While this study is focused on clinical trials, the methods and solution proposed here are relevant in other contexts such as simultaneous learning about multiple marketing messages where the set of possible actions may be very large.

REFERENCES

Adelman, D. and Mersereau, A. J. (2008). Relaxations of weakly coupled stochastic dynamic programs. *Operations Research,* 56(3):712727.

Ahuja, V. and Birge, J. (2014). Fully adaptive designs for clinical trials: Simultaneous learning from multiple patients. *Working paper available at SSRN:* http://ssrn.com/abstract=2126906.

Aviv, Y. and Pazgal, A. (2005). A partially observed markov decision process for dynamic pricing. *Management Science,* 51(9):14001416.

Berry, D. (1978). Modified two-armed bandit strategies for certain clinical trials. *Journal of the American Statistical* Association, 73(362):pp. 339345.

Berry, D. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments. *Chapman and Hall London.*

Bertsimas, D. and Mersereau, A. (2007). A learning approach for interactive marketing to a customer segment. *Operations Research,* 55(6):11201135.

Caro, F. and Gallien, J. (2007). Dynamic assortment with demand learning for seasonal consumer goods. *Management Science,* 53(2):276.

English, R., Lebovitz, Y., Griffin, R., et al. (2010). Transforming Clinical Research in the United States: Challenges and Opportunities: *Workshop Summary. National Academies Press*.

FDA (2010). Adaptive design clinical trials for drugs and biologics. *Guidance for Industry.*

Forbes (2013). The cost of creating a new drug now 5 billion, pushing big pharma to change.

Hauskrecht, M. (1997). Incremental methods for computing bounds in partially observable markov decision processes. In *Proceedings of The National Conference on Artificial Intelligenc*e, pages 734739. Citeseer.

Judd, K. (1998). Numerical Methods in Economics. *The MIT press.*

Lovejoy, W. (1991). Computationally feasible bounds for partially observed markov decision processes. Operations research, 39(1):162175.

Monahan, G. (1982). State of the arta survey of partially observable markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):116.

Powell, W. (2007). Approximate Dynamic Programming: Solving the curses of dimensionality. *Wiley-Interscience.*

Sandikci, B. (2010). Reduction of a pomdp to an mdp. Wiley Encyclopedia of Operations Research and Management Science.

Sandikci, B., Maillart, L. M., Schaefer, A. J., and Roberts, M. S. (2013). Alleviating the patients price of privacy through a partially observable waiting list. Management Science.

Zhou, R. and Hansen, E. (2001). An improved grid-based approximation algorithm for pomdps. In *International Joint Conference on Artificial Intelligence,* volume 17, pages 707716. Citeseer