OPERATIONS RESEARCH
Vol. 55, No. 6, November–December 2007, pp. 1120–1135
ISSN 0030-364X | EISSN 1526-5463 | 07 | 5506 | 1120

DOI 10.1287/opre.1070.0427

A Learning Approach for Interactive Marketing to a Customer Segment

Dimitris Bertsimas
Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, dbertsim@mit.edu

Adam J. Mersereau
Kenan-Flagler Business School, University of North Carolina, Chapel Hill, North Carolina 27599, adam_mersereau@unc.edu

When a marketer in an interactive environment decides which messages to send to her customers, she may send messages currently thought to be most promising (exploitation) or use poorly understood messages for the purpose of information gathering (exploration). We assume that customers are already clustered into homogeneous segments, and we consider the adaptive learning of message effectiveness within a customer segment. We present a Bayesian formulation of the problem in which decisions are made for batches of customers simultaneously, although decisions may vary within a batch. This extends the classical multiarmed bandit problem for sampling one-by-one from a set of reward populations. Our solution methods include a Lagrangian decomposition-based approximate dynamic programming approach and a heuristic based on a known asymptotic approximation to the multiarmed bandit solution. Computational results show that our methods clearly outperform approaches that ignore the effects of information gain.

Subject classifications: dynamic programming/optimal control: relaxations, multiarmed bandit problem; marketing: advertising and media.
Area of review: Computing and Information Technologies.
History: Received December 2003; revisions received July 2005, August 2006, December 2006; accepted December 2006. Published online in *Articles in Advance* September 4, 2007.

1. Introduction

Marketers are increasingly capable of interacting with their customers directly. With this capability comes the challenge to gather and use evolving data to optimize the delivery of marketing messages. Due to scale and speed considerations, interactive marketing decisions must be automated, and there is a continuing need for sophisticated data-driven decision-support tools in this arena. Example contexts range from traditional target marketing contexts like direct mail and catalogs to Internet-based contexts like banner advertising and paid search advertising. We collectively refer to such contexts as interactive marketing environments.

In many such environments, a marketer may have insufficient data to accurately predict the effectiveness of available messages. Off-line experiments to collect data may prove both costly and slow, and may poorly reflect changing conditions in a real environment. For this reason, we propose methods that actively learn message effectiveness during real operations. With a limited number of available customer encounters, the marketer in such an environment faces the so-called "exploration versus exploitation" dilemma between using messages thought to be promising (exploitation) and using poorly understood messages for the purpose of data gathering (exploration). There has been very little work addressing this adaptive learning issue in the interactive marketing context, and there is no consensus on the proper modeling framework for approaching it.

Our model assumes that available customer encounters at each decision epoch have been segmented according to predetermined rules based on available data on the customer, for example, recency, frequency, and monetary value (RFM) variables, contact and response histories, or demographic attributes. We then assume that customers behave statistically homogeneously within a segment, and we select (and learn about) messages for each customer segment independently. Such a framework decouples the detailed modeling of customer attributes from the optimization of message selection. Recognizing that segmentation models have been well studied in many settings and are necessarily dependent on the data available and application, our focus in this paper is on the optimal allocation of messages to a segment of customer encounters.

The classic framework for studying the exploration-exploitation trade-off is the multiarmed bandit problem, an adaptive sampling problem in which a decision maker draws samples one-by-one from a set of available populations with the objective of maximizing the expected (or discounted expected) sum of the samples. In an interactive marketing environment, however, it is often necessary to make decisions for multiple customers at the same time due to simultaneous contact with many customers (as in tradi-

tional direct mailing), delayed responses, technology-based considerations favoring updating databases and generating decisions off-line, or systemwide global constraints.

The batched adaptive sampling problem we consider is difficult to solve but simple to state. We choose among a set of M marketing messages for N simultaneous customer encounters at a time. We assume that responses are independent Bernoulli random variables for which the success probability for each message is unknown but described by a prior distribution. The responses of the N customers encountered at stage t are assumed known before decisions must be made for stage $t+1$, and the objective is to maximize the expected number of successes over T decision stages. While related adaptive sampling problems have been considered for other applications, most notably for medical trials, previous methods are not scalable and general enough for use in an interactive marketing context.

This adaptive sampling problem can be straightforwardly formulated as a dynamic program. The dynamic programming formulation leads to some basic insights: that there is value to information and that the optimal reward is superadditive in N and T. Practically, however, the size of the state space prohibits an efficient exact solution for all but the smallest problem instances. We develop two approaches for its approximate solution. The first is an approximate dynamic programming algorithm that is based on a Lagrangian decomposition at each stage. The method performs well and can be readily extended to a multisegment problem in which the segments are linked through budget constraints and customer dynamics. Our second method extends an approximate policy for the multiarmed bandit problem due to Lai (1987). This heuristic is particularly efficient to compute and achieves very good results in our simulation study, although its extendibility is less clear.

As we point out in §2.1, there exists little academic work on adaptive learning in marketing contexts, even though the interactive marketing domain is dynamic and data-driven, and thus is a natural area to apply adaptive techniques. We believe that one reason for the void is the large gap between the relative simplicity of known, tractable adaptive learning models and the relative complexity (and scale) of many interesting interactive marketing problems. While our work attempts to narrow this gap, our customer behavior model remains simple compared with contemporary marketing research on modeling customer choice for descriptive purposes or for myopic optimization. The adaptive experimentation problem we consider is inherently more difficult, and we believe that some modeling sacrifice is therefore necessary. We also believe that the remaining modeling gap represents a challenging opportunity for future research.

A common customer choice modeling approach is the random utility framework (see, for instance, Ben-Akiva and Lerman 1985), in which a customer's action depends on a random utility function over choices or choice attributes, and the parameters of the utility function may include customer attributes and/or random terms to model the heterogeneity of customers (see Ansari and Mela 2003 for a contemporary model in the interactive marketing context). Such models imply complex dependencies between the unknown parameters. Bayesian approaches to fitting such choice models from data (see Rossi et al. 2005) yield complicated multidimensional posterior distributions that are typically estimated using numerical simulation and cannot be concisely parameterized.

On the other hand, known approaches to adaptive experimentation, specifically to the multiarmed bandit problem, nearly always assume that the parameters to be learned are statistically independent given the decision-maker's decisions. A significant benefit of this assumption is that it allows the decision maker to represent her state of knowledge compactly and update it easily, which in turn makes it tractable to fathom the future information impact of current decisions within a dynamic optimization algorithm. Thus, key assumptions ensuring tractability of our model are those that allow us to assume independence of the parameters to be learned.

Within a segment, our model requires that messages perform independently given the marketer's decisions: learning about one message does not inform her about the effectiveness of other messages. The independent parameters to be learned are the success probabilities of each message. We note that this assumption will be conservative when there are statistical dependencies among messages. That is, our approach will learn more slowly than if these dependencies were taken into account.

We also require that customers can be segmented at each stage into independent and internally homogeneous groups. Assuming independence among segments allows us to naturally decompose the problem by segment (at the expense of learning speed, if there are strong relationships between segments). Our basic formulation and our work in §3 are thus based on the important single-segment subproblem. Implicit in the assumption of homogeneous customer segments is that all customer-specific data has been adequately summarized by each customer's current segment label. For example, our optimization formulation does not keep track of which customers have been exposed to which messages, although we note that this can be managed by including contact history information in the segmentation model. In §4.3, we extend our model to account for the resulting migrations between segments.

An interesting special case of the segmented customer framework is when each segment consists of a single customer, and we learn the effectiveness of messages for each individual independently. Such a model, which assumes that customers are fully heterogeneous, can be handled by the methods we present. Learning may be slow in practice due to the large number of unknown parameters and limited encounters with each customer. Nevertheless, this special case suggests the segmented model as an approximation of customer heterogeneity, and that the trade-off between heterogeneity and learnability can be manipulated through the fineness of the segmentation.

A segmented customer model is natural in a number of practical contexts. In many applications, little data about individuals is known or available (e.g., due to constraints of the media or privacy considerations), and discrete bucketing of customers is appropriate and perhaps unavoidable. Examples include Internet applications such as paid search, in which customers are largely anonymous. Even for direct mailing campaigns where customer-specific data is available, it remains popular in practice to segment customers according to RFM variables and treat customers as homogeneous within the resulting discrete segments. Several previous studies of dynamic marketing optimization (e.g., Simester et al. 2006, Bitran and Mondschein 1996) assume homogeneous response within segments.

The contributions of this paper are the proposal of an adaptive sampling framework for interactive marketing environments, new methods for addressing a key subproblem of this framework, and associated insights. We note that the subproblem—an extension of the classic multiarmed bandit problem to batched decision making—is also of more general interest than the specific marketing application considered. Other example applications include medical trials (Cheng et al. 2002) and retail assortment (Caro and Gallien 2007).

The remainder of this paper is organized as follows. We examine related marketing and algorithmic literature in §2. Section 3 formulates the single-segment adaptive sampling problem we consider and presents some theoretical results about the optimal expected reward. We develop our approximate dynamic programming (ADP) method in §4, and our heuristic based on the asymptotic multiarmed bandit approximation of Lai (1987) in §5. Section 6 presents results of a simulation study to examine the performance of the proposed methods, and §7 points toward interesting future research directions.[1, 2]

2. Related Work

We organize relevant literature into two categories: research on marketing in interactive media, and general research on methods for adaptive sampling.

2.1. Relevant Work on Interactive Marketing

We classify the related work on decision making in interactive marketing into three sets: myopic models, dynamic models capturing customers' long-run profitability, and dynamic models incorporating learning effects.

A number of choice models have been proposed to capture detailed choice patterns in interactive marketing data. However, decisions produced by these models are myopic in that they maximize only immediate profits. Bult and Wansbeek (1995) develop a random utility model for sampling direct mail targets with the objective of maximizing profits from a single mailing. Rossi et al. (1996) use a Bayesian hierarchical model to model customer responses to point-of-sale grocery coupon offers. Their data model is sophisticated, accounting for both observed and unobserved sources of customer heterogeneity, and highlights computational advances allowing the application of Bayesian statistics to marketing (see Rossi et al. 2005). More recently, Ansari and Mela (2003) fit a hierarchical Bayesian model and apply deterministic optimization to the results for the purpose of designing customized e-mail offers. They optimize a single-period objective, ignoring their decisions' future impacts. In particular, active learning is not considered.

Using coarser models of customer choice, a number of studies have explored the impact of current decisions on a customer's long-run profit stream (i.e., "lifetime value"). Bitran and Mondschein (1996) introduce a dynamic programming-based mailing policy for the situation in which a direct mailer has limited capital. Simester et al. (2006) test a dynamic programming-based model with real catalog data. Both of these studies model customer behavior with discrete customer segments, based on RFM variables in the first and on a tree-based segmentation method in the second, within which response rates are assumed homogeneous. Gönül and Shi (1998) develop an estimable structural dynamic programming model of customer and direct mailer behavior, deriving a dynamic mailing policy. They simplify firmwide considerations so that each customer can be modeled independently, and they track customers using just two attributes, recency and frequency, due to difficulties inherent in high-dimensional dynamic programming.

All of the myopic and "lifetime value" studies discussed above ignore the impact of information gain on current decisions. We are aware of only two studies on adaptive experimentation in interactive marketing settings. In Ariely et al. (2002), choice models are assumed known for customer segments, but the customers' segment memberships are unknown. The paper of Gooley and Lattin (2000) discusses allocating marketing messages when customers are differentiated by covariates and customer contacts occur one-by-one. Through a data experiment, the authors provide evidence of the value of adaptive sampling in an interactive marketing environment. They also present ideas on an adaptive sampling approach when responses are modeled as linear functions of customer attributes. Their approach is not implemented or tested, and does not naturally extend to batched decision making or budget constraints.

2.2. Relevant Methodological Work

The basic single-segment adaptive sampling problem we consider relates closely to the so-called multiarmed bandit problem (see Gittins 1989), a prototypical resource allocation problem involving several projects or "arms," each of which is represented as a reward-generating Markov chain. At each stage, the decision maker chooses one arm to generate a reward and change state according to known transition probabilities. The highlight of the multiarmed bandit literature is the work of Gittins (see Gittins and Jones 1974

and Gittins 1979), which characterizes an optimal policy for the infinite-horizon version of the problem as comparing "indices" that can be independently and efficiently computed for each arm.

An important special case of this problem (see, for example, Berry and Fristedt 1985), also known as the multiarmed bandit problem, is when the decision maker has several available populations on which she has prior distributions, and she draws independent samples one-by-one from the populations, updating her prior distributions after each observation. The objective is to maximize expected (or discounted expected) responses. In this version of the problem, the populations correspond to "arms," and the Bayesian update at each stage can be thought of as a transition in a Markov chain. Much of the work on this problem has been in deriving asymptotic approximations to optimal policies. Important papers in this vein are those of Lai and Robbins (1985) and Lai (1987). The basic problem we consider can be viewed as a multiarmed bandit problem extended to the case of multiple samples at each stage. Anantharam et al. (1987) consider asymptotic approximations for the case of multiple plays at each stage, although they assume at most one sample can be taken from each population at each stage.

Related problems of sequential experimental design have been studied in the statistics and biostatistics communities, inspired by the problem of designing sequential medical trials. Much of the work has focused on few stages (typically two or three), few populations (typically no more than two), and fewer samples than considered here. Recent analytical research on two-stage experiments involving two populations can be found in Cheng et al. (2002). Hardwick and Stout (2002) solve two- and three-stage problems with two populations optimally using clever dynamic programming implementations. Sequential experimental design typically requires the decision maker to select the number of samples at each stage subject to an overall limit on total samples taken over the finite horizon. We have developed very similar methods to that discussed in §4 for this problem with promising results, but this paper deals with the case in which the number N of samples is exogenously fixed at each stage.

3. Batched Adaptive Sampling in a Single Segment

As discussed in §1, a customer segmentation model allows us to decouple the problem of modeling customer covariates from the problem of optimizing message selection, assuming that customer encounters are statistically homogeneous within a segment and that message effectiveness can be learned independently across segments. Given these assumptions, in this section we focus on a single segment of customers in isolation and formulate the problem of adaptive sampling of messages within a single segment. We also prove some analytical results on the behavior of the optimal expected rewards as a function of basic problem parameters.

3.1. Problem and Notation

We assume that a marketer must assign messages to N customer encounters in a segment at each stage with the overall objective of maximizing total rewards over a finite horizon. Immediate reward generated by a customer encounter is an independent Bernoulli random variable with probability p_m that depends on the message m sent to the customer in the current stage. We assume that the marketer does not know p_m, and has a beta prior distribution (see Drake 1967) on its true value. With parameters $s > 0$ and $f > 0$ integer, the beta density is given by

$$f_\beta(p; s, f) = \begin{cases} \dfrac{(s+f-1)!}{(s-1)!(f-1)!} p^{s-1}(1-p)^{f-1}, & 0 < p < 1, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

It can be shown that the beta distribution has the following properties:

(1) The beta distribution is a conjugate prior for the binomial distribution. That is, given a beta prior distribution beta(s^0, f^0) on the success probability p of a Bernoulli distribution, if we observe $s+f$ Bernoulli experiments that yield s successes and f failures, the Bayesian posterior distribution of p is beta(s^0+s, f^0+f).

(2) The uniform distribution on $[0, 1]$ can be written as beta$(1, 1)$.

The first of these properties implies that the beta distribution is particularly simple to update. The second leads to a practical interpretation of the beta distribution parameters in our problem. It is natural to assume that a marketer with no information about the effectiveness of her available messages has uniform priors on their success probabilities. Prior to stage zero, however, she may have gathered some information on the various messages by running off-line experiments. The parameters s_m^0 and f_m^0 of the prior distribution can be thought of as the number of successes (plus one) and the number of failures (plus one) observed in these off-line experiments.

We assume a binary reward structure for the sake of simplicity, and we note that this assumption can be relaxed to some extent. In §4.3, we relax it somewhat by introducing a reward for success that may vary (in a known way) across messages. An obvious way to model nonbinary responses would be to replace the beta/binomial framework we use with another conjugate model pair, say gamma/Poisson. Adapting our approaches to another conjugate model pair is conceptually straightforward but brings complexity by adding to the size of the state space.

We make use of the following notation:

- T: the finite horizon of the problem. We index discrete time stages $0, \ldots, T-1$ with the variable t.
- M: the number of messages available. We index the messages by m. We note that we may extend the notion of messages to encompass different channels, designs, or offers as suits the application. We assume that the messages are consistent over time, although one could model deterministic changes in M over time by adding constraints limiting when individual messages can be used.
- N: the total number of customers in the segment to be contacted at each stage. We can easily handle the case where *at most* N customers are contacted each stage by including a "do not send" option among the available messages. However, given nonzero response probabilities it will not be optimal to neglect an available customer encounter in the current problem.

We assume that the number of customer encounters N at each stage is constant in time and exogenous. Our formulation and methods can be easily extended to the case in which the number of customer encounters at each stage varies deterministically with t. The case in which the numbers of future customer encounters are exogenous but random is technically more complicated, although we suspect that replacing these random variables by their expectations would provide a reasonable approximation. As mentioned in §2.2, we have considered an alternate problem in which the decision maker may choose how many customers to contact each period, but must contact a total of $\overline{N}$ customers over the finite horizon. A similar decomposition method to that developed in §4 offers promising results, but for horizons longer than three periods it seems to offer little improvement over dividing contacts evenly among the available time periods.

- s_m^0, f_m^0: the parameters of the initial prior beta distribution on p_m. These numbers for each message m are to be specified for each problem instance as part of the problem data.
- x_m^t: the decision variable indicating the number of customers sent message m at stage t. The vector of decisions at stage t is denoted $\mathbf{x}^t = (x_1^t, \ldots, x_M^t)$. We note that the adequacy of this choice of decision variable follows from our assumption of homogeneous customer encounters within a segment.
- y_m^t: the random variable giving the number of customers sent message m in stage t who subsequently generate a "success."
- s_m^t, f_m^t: the parameters of the updated beta distribution on the success probability corresponding to message m. These numbers for each message m are maintained as state variables in the problem. We refer to the vectors of reward distribution parameters at stage t using boldface characters: $\mathbf{s}^t = (s_1^t, \ldots, s_M^t)$ and $\mathbf{f}^t = (f_1^t, \ldots, f_M^t)$. s_m^t and f_m^t accumulate the numbers of successes and failures observed so far. Thus, $\mathbf{s}^t = \mathbf{s}^0 + \sum_{j=0}^{t-1} \mathbf{y}^j$ and $\mathbf{f}^t = \mathbf{f}^0 + \sum_{j=0}^{t-1} (\mathbf{x}^j - \mathbf{y}^j)$. Also, we have $\sum_{m=1}^{M} (s_m^t - s_m^0 + f_m^t - f_m^0) = Nt$.

We assume that the problem is reformulated and resolved at each decision stage. Thus, more complicated information dynamics than we assume can be incorporated into the decisions passively via respecification of the message priors at each stage. If this is done, then our statistical assumptions can be seen as approximations for the purpose of fathoming future rewards.

3.2. Dynamic Programming Formulation

The exact solution of the problem is naturally approached via dynamic programming. Here we formulate the dynamic program by specifying the state, randomness, available controls, system dynamics, and reward structure of the system over which we are optimizing.

- *State*: The state of the system at stage t is the total number of successes and failures (plus the priors s_m^0 and f_m^0, respectively) observed so far for each of the messages, and is given by the vector $(\mathbf{s}^t, \mathbf{f}^t) = (s_1^t, \ldots, s_M^t, f_1^t, \ldots, f_M^t)$.
- *Control*: The system control is $\mathbf{x}^t = (x_1^t, \ldots, x_M^t)$, with x_m^t indicating the number of type m messages sent at stage t. We constrain $\sum_{m=1}^{M} x_m^t = N$ for all $t = 0, \ldots, T-1$, and $x_m^t \geqslant 0$, x_m^t integer for all $t = 0, \ldots, T-1$ and $m = 1, \ldots, M$.
- *Randomness*: y_m^t gives the number of successes resulting from sending x_m^t type m messages when the success probability p_m is distributed as beta(s_m^t, f_m^t). y_m^t is a beta-binomial random variable, with probability mass function (see, for example, Raiffa and Schlaiffer 1961, §7.11):

$$\begin{aligned}
&\Pr(y_m^t = y \mid x_m^t; s_m^t, f_m^t) \\
&= \int_0^1 \binom{x_m^t}{y} p^y (1-p)^{x_m^t - y} f_\beta(p; s_m^t, f_m^t)\, dp \\
&= \binom{x_m^t}{y} \frac{(s_m^t + y - 1)!}{(s_m^t - 1)!} \frac{(f_m^t + x_m^t - y - 1)!}{(f_m^t - 1)!} \frac{(s_m^t + f_m^t - 1)!}{(s_m^t + f_m^t + x_m^t - 1)!}, \\
&\qquad\qquad y = 0, \ldots, x_m^t. \quad (2)
\end{aligned}$$

As the y_m^t random variables are assumed independent of each other, the probability mass function for the vector $\mathbf{y}^t = (y_1^t, \ldots, y_M^t)$ is the product of the individual probability mass functions of its components.

- *System Dynamics*: The system state evolves as $\mathbf{s}^{t+1} = \mathbf{s}^t + \mathbf{y}^t$ and $\mathbf{f}^{t+1} = \mathbf{f}^t + \mathbf{x}^t - \mathbf{y}^t$.
- *Expected Rewards*: The total expected reward is the expected number of successes over T periods. The expected reward accruing at stage t is given simply by

$$\mathsf{E}\left[\sum_{m=1}^{M} y_m^t \,\middle|\, \mathbf{x}^t; \mathbf{s}^t, \mathbf{f}^t\right] = \sum_{m=1}^{M} x_m^t \left(\frac{s_m^t}{s_m^t + f_m^t}\right), \quad (3)$$

where we use the fact that $s_m^t/(s_m^t + f_m^t)$ is the expected value of the beta distribution with parameters s_m^t and f_m^t.

The problem can in principle be solved using the following dynamic programming iteration. For the final stage, the

optimal strategy is to send all customers the message with the highest expected success probability:

$$J_{T-1}(\mathbf{s}^{T-1}, \mathbf{f}^{T-1}) = N \cdot \max\left\{ \frac{s_m^{T-1}}{s_m^{T-1} + f_m^{T-1}} \right\}. \tag{4}$$

For $t = T-2, \ldots, 0$,

$$\begin{aligned} J_t(\mathbf{s}^t, \mathbf{f}^t) = \max_{\mathbf{x}} & \sum_{m=1}^{M} \left(\frac{s_m^t}{s_m^t + f_m^t} \right) x_m^t \\ & + \mathsf{E}_{\mathbf{y}^t}[J_{t+1}(\mathbf{s}^t + \mathbf{y}^t, \mathbf{f}^t + \mathbf{x}^t - \mathbf{y}^t) \mid \mathbf{x}; \mathbf{s}^t, \mathbf{f}^t] \\ \text{s.t.} \quad & \sum_{m=1}^{M} x_m^t = N, \\ & x_m^t \geqslant 0, \quad m = 1, \ldots, M, \\ & x_m^t \text{ integer}, \quad m = 1, \ldots, M, \end{aligned} \tag{5}$$

where the expectation is with respect to M independent probability mass functions of the form given in Equation (2).

Discounting rewards geometrically in time with discount rate $0 \leqslant \alpha \leqslant 1$ requires a trivial modification to the formulation: we modify Equation (5) by adding the factor α in front of the expectation. This results in only minor modifications to the ADP methods. (We note that most of our theoretical results carry over, but that Propositions 3 and 4 in §3.3 will not generally hold as stated for the discounted case.) It is less clear, however, how to modify the heuristic of §5 to handle discounting.

3.3. The Value of Information

In this section, we show that the dynamic programming formulation of the single-segment problem implies there is value to information. That is, an optimal decision maker may expect larger expected rewards if allowed to observe the outcome of additional trials. This result has implications for the design of an interactive marketing campaign. Using the result, we prove that there are increasing returns to both stage size and problem horizon and that, if possible, the marketer should arrange a given number of customer encounters by maximizing the number of stages while minimizing the encounters per stage. In the interest of brevity, some of the proofs of this section's results are included in the online companion (Appendix A). An electronic companion to this paper is available as part of the on line version that can be found at http://or.journal.informs.org/.

We require the following lemma, which states that, for a fixed set of encounters $\mathbf{x}$, the a priori joint distribution of the outcomes does not change if the encounters are broken up into two sets that are performed in series.

LEMMA 1. *For any m, x_m^1, x_m^2, y, s_m, f_m, and letting y_m, y_m^1, and y_m^2 indicate the number of successes in $x_m^1 + x_m^2$, x_m^1, and x_m^2 encounters, respectively,*

$$\begin{aligned} &\Pr(y_m = y \mid x_m^1 + x_m^2; s_m, f_m) \\ &\quad = \sum_{y_m^1 = 0}^{y} \Pr(y_m^2 = y - y_m^1 \mid x_m^2; s_m + y_m^1, f_m + x_m^1 - y_m^1) \\ &\qquad \cdot \Pr(y_m^1 \mid x_m^1; s_m, f_m). \end{aligned}$$

PROOF. We omit the details of the proof. We note that the result can be verified by substituting expression (2) for the probabilities in the lemma statement, canceling terms, and applying Vandermonde's identity for binomial coefficients. □

The following result captures the intuitive notion that there is value to information. On an expected basis, the marketer should prefer more information to less.

PROPOSITION 1. *For all t, $\mathbf{s}^t$, $\mathbf{f}^t$, and fixed $\mathbf{x} \geqslant 0$,*

$$\mathsf{E}_{\mathbf{y}}[J_t(\mathbf{s}^t + \mathbf{y}, \mathbf{f}^t + \mathbf{x} - \mathbf{y}) \mid \mathbf{x}; \mathbf{s}^t, \mathbf{f}^t] \geqslant J_t(\mathbf{s}^t, \mathbf{f}^t).$$

PROOF. For expositional convenience, we define the reparameterized value function $\tilde{J}_t(\mathbf{n}^t, \mathbf{s}^t) = J_t(\mathbf{s}^t, \mathbf{n}^t - \mathbf{s}^t)$ for all t, $\mathbf{s}^t$, and $\mathbf{n}^t \geqslant \mathbf{s}^t$.

Consider stage $T-1$. For fixed $\mathbf{n}^{T-1}$, $\tilde{J}_{T-1}(\mathbf{n}^{T-1}, \mathbf{s}^{T-1}) = N \cdot \max_m\{s_m^{T-1}/n_m^{T-1}\}$ is the maximum over linear functions of $\mathbf{s}^{T-1}$, and thus is a convex function of $\mathbf{s}^{T-1}$. Jensen's inequality then gives, for fixed $\mathbf{x}$,

$$\begin{aligned} &\mathsf{E}_{\mathbf{y}}\left[\tilde{J}_{T-1}(\mathbf{n}^{T-1} + \mathbf{x}, \mathbf{s}^{T-1} + \mathbf{y}) \mid \mathbf{x}; \mathbf{n}^{T-1}, \mathbf{s}^{T-1}\right] \\ &\quad \geqslant \tilde{J}_{T-1}(\mathbf{n}^{T-1} + \mathbf{x}, \mathbf{s}^{T-1} + \mathsf{E}[\mathbf{y} \mid \mathbf{x}; \mathbf{n}^{T-1}, \mathbf{s}^{T-1}]) \\ &\quad = N \cdot \max_m \left\{ \frac{s_m^{T-1} + x_m s_m^{T-1}/n_m^{T-1}}{n_m^{T-1} + x_m} \right\} \\ &\quad = N \cdot \max_m \left\{ \frac{s_m^{T-1}}{n_m^{T-1}} \right\} \\ &\quad = \tilde{J}_{T-1}(\mathbf{n}^{T-1}, \mathbf{s}^{T-1}) \quad \text{for all } \mathbf{n}^{T-1}, \mathbf{s}^{T-1} \leqslant \mathbf{n}^{T-1}. \end{aligned}$$

Now consider stage $0 \leqslant t \leqslant T-2$. Assume that

$$\mathsf{E}_{\mathbf{y}}[\tilde{J}_{t+1}(\mathbf{n}^{t+1} + \mathbf{x}, \mathbf{s}^{t+1} + \mathbf{y}) \mid \mathbf{x}; \mathbf{n}^{t+1}, \mathbf{s}^{t+1}] \geqslant \tilde{J}_{t+1}(\mathbf{n}^{t+1}, \mathbf{s}^{t+1}) \quad \text{for all } \mathbf{n}^{t+1}, \mathbf{s}^{t+1} \leqslant \mathbf{n}^{t+1}, \mathbf{x},$$

and let

$$\mathbf{x}^* = \operatorname*{arg\,max}_{\mathbf{x}: \sum_m x_m = N} \left\{ \sum_m (s_m^t/n_m^t) x_m + \mathsf{E}_{\mathbf{y}}\left[\tilde{J}_{t+1}(\mathbf{n}^t + \mathbf{x}, \mathbf{s}^t + \mathbf{y}) \mid \mathbf{x}; \mathbf{n}^t, \mathbf{s}^t\right] \right\}.$$

Then, for fixed $\mathbf{x}$,

$$\begin{aligned} &\mathsf{E}_{\mathbf{y}}[\tilde{J}_t(\mathbf{n}^t + \mathbf{x}, \mathbf{s}^t + \mathbf{y}) \mid \mathbf{x}; \mathbf{n}^t, \mathbf{s}^t] \\ &= \mathsf{E}_{\mathbf{y}}\left[\max_{\mathbf{x}': \sum_m x'_m = N} \left\{ \sum_m \frac{s_m^t + y_m}{n_m^t + x_m} x'_m \right. \right. \\ &\qquad \left. \left. + \mathsf{E}_{\mathbf{y}'}[\tilde{J}_{t+1}(\mathbf{n}^t + \mathbf{x} + \mathbf{x}', \mathbf{s}^t + \mathbf{y} + \mathbf{y}') \mid \mathbf{x}'; \mathbf{n}^t + \mathbf{x}, \mathbf{s}^t + \mathbf{y}] \right\} \,\middle|\, \mathbf{x}; \mathbf{n}^t, \mathbf{s}^t \right] \end{aligned}$$

$$\geqslant \mathsf{E}_{\mathbf{y}}\Bigg[\sum_m \frac{s_m^t+y_m}{n_m^t+x_m}x_m^* + \mathsf{E}_{\mathbf{y}^*}[\tilde{J}_{t+1}(\mathbf{n}^t+\mathbf{x}+\mathbf{x}^*, \mathbf{s}^t+\mathbf{y}+\mathbf{y}^*) \mid \mathbf{x}^*; \mathbf{n}^t+\mathbf{x}, \mathbf{s}^t+\mathbf{y}] \mid \mathbf{x}; \mathbf{n}^t, \mathbf{s}^t\Bigg]$$

$$= \sum_m \frac{s_m^t}{n_m^t}x_m^* + \mathsf{E}_{\mathbf{y}}[\mathsf{E}_{\mathbf{y}^*}[\tilde{J}_{t+1}(\mathbf{n}^t+\mathbf{x}+\mathbf{x}^*, \mathbf{s}^t+\mathbf{y}+\mathbf{y}^*) \mid \mathbf{x}^*; \mathbf{n}^t+\mathbf{x}, \mathbf{s}^t+\mathbf{y}] \mid \mathbf{x}; \mathbf{n}^t, \mathbf{s}^t]$$

$$= \sum_m \frac{s_m^t}{n_m^t}x_m^* + \mathsf{E}_{\mathbf{y}^*}[\mathsf{E}_{\mathbf{y}}[\tilde{J}_{t+1}(\mathbf{n}^t+\mathbf{x}+\mathbf{x}^*, \mathbf{s}^t+\mathbf{y}+\mathbf{y}^*) \mid \mathbf{x}; \mathbf{n}^t+\mathbf{x}^*, \mathbf{s}^t+\mathbf{y}^*] \mid \mathbf{x}^*; \mathbf{n}^t, \mathbf{s}^t]$$

$$\geqslant \sum_m \frac{s_m^t}{n_m^t}x_m^* + \mathsf{E}_{\mathbf{y}^*}[\tilde{J}_{t+1}(\mathbf{n}^t+\mathbf{x}^*, \mathbf{s}^t+\mathbf{y}^*) \mid \mathbf{x}^*; \mathbf{n}^t, \mathbf{s}^t]$$

$$= \tilde{J}_t(\mathbf{n}^t, \mathbf{s}^t),$$

where the equality in the third-to-last line follows from Lemma 1 and the inequality on the second-to-last line follows from the induction assumption. The desired result then follows for all $t = T-1, \ldots, 0$ by induction. □

Proposition 1 can be generalized to the case in which the decisions are not fixed, but rather are chosen in a series of stages. We provide details in online Appendix A. Proposition 1 allows us to prove informative structural results about the optimal problem rewards. Let $J_t(\mathbf{s}^t, \mathbf{f}^t; N, T)$ indicate the optimal cost-to-go evaluated in stage t and state $(\mathbf{s}^t, \mathbf{f}^t)$ for a problem with horizon T and stage size N. The following result says that the optimal reward is superadditive in the stage size N, so that there are increasing returns to stage size. Intuitively, because information is valuable, a large pool of customer encounters, managed centrally, can give higher rewards than two smaller pools of experiments, managed independently. We include a proof in online Appendix A.

Proposition 2. *For any initial state* $\mathbf{s}^0$, $\mathbf{f}^0$, *horizon* T, *and stage sizes* N_A *and* N_B,

$$J_0(\mathbf{s}^0, \mathbf{f}^0; N_A+N_B, T) \geqslant J_0(\mathbf{s}^0, \mathbf{f}^0; N_A, T) + J_0(\mathbf{s}^0, \mathbf{f}^0; N_B, T).$$

Similarly, the value of information implies that the optimal reward is also superadditive in the horizon T. See online Appendix A for a proof.

Proposition 3. *For any initial state* $\mathbf{s}^0$, $\mathbf{f}^0$, *stage size* N, *and horizons* T_A *and* T_B,

$$J_0(\mathbf{s}^0, \mathbf{f}^0; N, T_A+T_B) \geqslant J_0(\mathbf{s}^0, \mathbf{f}^0; N, T_A) + J_0(\mathbf{s}^0, \mathbf{f}^0; N, T_B).$$

Finally, for a fixed total number TN of customer encounters, higher rewards are possible if the marketer can increase the number of stages while decreasing the stage size. We include a proof in online Appendix A.

Proposition 4. *For any initial state* $\mathbf{s}^0$, $\mathbf{f}^0$, *stage size* N, *and horizon* T, *and assuming integer* $\gamma > 1$ *such that* T/γ *is integer,*

$$J_0(\mathbf{s}^0, \mathbf{f}^0; N, T) \geqslant J_0(\mathbf{s}^0, \mathbf{f}^0; \gamma N, T/\gamma).$$

4. Solution via Approximate Dynamic Programming

While the dynamic programming formulation provides a method for solving the single-segment problem of §3 exactly, the direct application of dynamic programming is computationally prohibitive for problems of reasonable size. (For a small problem with $N = 10$, $T = 10$, and $M = 3$, the dynamic program has on the order of 10^8 states. With $N = 10$, $T = 5$, and $M = 10$, the number of states jumps to 10^{34}.) In particular, the number of states grows exponentially with the number of messages. To mitigate this state explosion, we investigate an approximation technique that uses Lagrange multipliers to decompose the problem by message.

We note that there is a literature on the decomposition of large dynamic programming problems for approximate solutions. Meuleau et al. (1998) is an example from the reinforcement learning community. The use of Lagrangian relaxation for decomposition of dynamic programs is investigated by Castañon (1997), Yost and Washburn (2000), and in the PhD dissertation of Hawkins (2003). Our algorithm is closely related to the ones proposed by these authors and to Whittle's (1988) proposed heuristic for the restless bandit problem. Subsequent to working versions of our paper, Caro and Gallien (2007) and Adelman and Mersereau (2007) have further investigated the Lagrangian approach.

The approximation method is motivated by the observation that the dynamic programming problem formulated in Equation (5) would be separable by message if not for the constraint $\sum_{m=1}^M x_m^t = N$. We add the redundant constraint $x_m^t \leqslant N$ for all m and replace the constraint $\sum_{m=1}^M x_m^t = N$ with a Lagrangian term in the objective function. We assume a constant Lagrange multiplier λ^t across all states at stage t. This gives a new value function which is a function of a vector of Lagrange multipliers $\boldsymbol{\lambda} = (\lambda^0, \ldots, \lambda^{T-1})$. We can write the dynamic programming iteration for the relaxed problem as follows:

$$J_{T-1}^{\boldsymbol{\lambda}}(\mathbf{s}^{T-1}, \mathbf{f}^{T-1}) = N\lambda^{T-1} + N\sum_{m=1}^M \left(\max\left\{0, \frac{s_m^{T-1}}{s_m^{T-1}+f_m^{T-1}} - \lambda^{T-1}\right\}\right). \quad (6)$$

For $t = T-2, \ldots, 0$,

$$J_t^{\boldsymbol{\lambda}}(\mathbf{s}^t, \mathbf{f}^t) = \max_{\mathbf{x}^t} \lambda^t\left(N - \sum_{m=1}^M x_m^t\right) + \sum_{m=1}^M \left(\frac{s_m^t}{s_m^t+f_m^t}\right)x_m^t + \mathsf{E}_{\mathbf{y}^t}[J_{t+1}^{\boldsymbol{\lambda}}(\mathbf{s}^t+\mathbf{y}^t, \mathbf{f}^t+\mathbf{x}^t-\mathbf{y}^t) \mid \mathbf{x}^t; \mathbf{s}^t, \mathbf{f}^t]$$
$$\text{s.t.} \quad 0 \leqslant x_m^t \leqslant N, \quad m = 1, \ldots, M, \quad (7)$$
$$x_m^t \text{ integer}, \quad m = 1, \ldots, M.$$

We observe first that the relaxed value function is separable by message at stage $T-1$, and that if the problem

is separable for some $t+1 \leqslant T-1$, then it is separable for t. Thus, we can write the relaxed value function for all $t=0,\ldots,T-1$ as

$$J_t^{\boldsymbol{\lambda}}(\mathbf{s}^t,\mathbf{f}^t) = N\sum_{\tau=t}^{T-1}\lambda^\tau + \sum_{m=1}^{M}\hat{J}_{t,m}^{\boldsymbol{\lambda}}(s_m^t,f_m^t), \tag{8}$$

where

$$\hat{J}_{T-1,m}^{\boldsymbol{\lambda}}(s_m^{T-1},f_m^{T-1}) = N\cdot\max\left\{0,\frac{s_m^{T-1}}{s_m^{T-1}+f_m^{T-1}}-\lambda^{T-1}\right\} \tag{9}$$

and

$$\begin{aligned}\hat{J}_{t,m}^{\boldsymbol{\lambda}}(s_m^t,f_m^t) = \max_x\ & \left(\frac{s_m^t}{s_m^t+f_m^t}-\lambda^t\right)x \\ & +\mathsf{E}_y\left[\hat{J}_{t+1,m}^{\boldsymbol{\lambda}}(s_m^t+y,f_m^t+x-y)\,|\,x;s_m^t,f_m^t\right] \\ \text{s.t.}\quad & 0\leqslant x\leqslant N, \\ & x \text{ integer},\end{aligned} \tag{10}$$

for $t=T-2,\ldots,0$.

The solution to the relaxed problem will not generally be feasible for the problem of Equation (5). To generate a feasible stage zero allocation, we consider the following constrained problem that uses the relaxed value functions to approximate the future implications of current actions:

$$\begin{aligned}\bar{J}_0^{\boldsymbol{\lambda}}(\mathbf{s}^0,\mathbf{f}^0) = \max_{\mathbf{x}^0}\ & \sum_{m=1}^{M}\left(\frac{s_m^0}{s_m^0+f_m^0}\right)x_m^0 \\ & +\mathsf{E}_{\mathbf{y}^0}\left[J_1^{\boldsymbol{\lambda}}(\mathbf{s}^0+\mathbf{y}^0,\mathbf{f}^0+\mathbf{x}^0-\mathbf{y}^0)\,|\,\mathbf{x}^0;\mathbf{s}^0,\mathbf{f}^0\right] \\ \text{s.t.}\quad & \sum_{m=1}^{M}x_m^0 = N, \\ & x_m^0\geqslant 0,\quad m=1,\ldots,M, \\ & x_m^0 \text{ integer},\quad m=1,\ldots,M.\end{aligned} \tag{11}$$

The problem gives, for any choice of $\boldsymbol{\lambda}$, a feasible stage zero allocation of messages. Similar formulations could potentially be used to generate feasible message allocations at stages $t=1,\ldots,T-2$ as well. We instead decide the message allocation at each stage t once the state $(\mathbf{s}^t,\mathbf{f}^t)$ becomes known, by solving (11) for a reduced horizon of $T-t$.

Implementation of the approximate dynamic programming method reduces to two methodological components: selecting the Lagrange multipliers $\boldsymbol{\lambda}$, and solving the subproblems. We consider these two components in §§4.1 and 4.2, respectively.

We add that Castañon (1997) motivates a similar decomposition approach in a way that leads to an interesting interpretation (see also Adelman and Mersereau 2007). The alternate motivation for the approach is to replace the constraints $\sum_{m=1}^{M}x_m^t = N$ by relaxed constraints $\mathsf{E}[\sum_{m=1}^{M}x_m^t]=N$, and then solve the resulting problem exactly using the method of Lagrange multipliers to accomplish the maximization.

4.1. Selecting $\boldsymbol{\lambda}$

We consider two different methods for selecting $\boldsymbol{\lambda}$. We can show that $\bar{J}_0^{\boldsymbol{\lambda}}(\mathbf{s}^0,\mathbf{f}^0)$ is convex as a function of $\boldsymbol{\lambda}$ and is an upper bound for the true value function $J_0(\mathbf{s}^0,\mathbf{f}^0)$ (see Appendix B in the online companion). The first method makes use of these observations, seeking the $\boldsymbol{\lambda}$ that gives the tightest possible value function bound. Our decision at state $(\mathbf{s}^0,\mathbf{f}^0)$ will then be a feasible $\mathbf{x}^0$ arising from the following minimization problem:

$$\min_{\boldsymbol{\lambda}}\bar{J}_0^{\boldsymbol{\lambda}}(\mathbf{s}^0,\mathbf{f}^0). \tag{12}$$

The convexity of $\bar{J}_0^{\boldsymbol{\lambda}}(\mathbf{s}^0,\mathbf{f}^0)$ implies that this minimum can be found efficiently. We consider simplifying the computation further by assuming a constant multiplier, $\lambda^1=\cdots=\lambda^{T-1}=\lambda$, as an approximation. With this assumption, choosing λ requires a one-dimensional numerical optimization. Following the interpretation introduced at the end of the previous section, we note that assuming time-invariant multipliers is equivalent to further relaxing the constraints $\mathsf{E}[\sum_{m=1}^{M}x_m^t]=N$ (for all $t>0$) by replacing them with a single constraint $\mathsf{E}[\sum_{\tau=1}^{T-1}\sum_{m=1}x_m^\tau]=N(T-1)$. The assumption seems reasonable in our case because the nature of the constraint $\sum_m x_m^t=N$ does not change over time nor do we expect the expected rewards to vary greatly across time periods. We provide evidence in the online companion (Appendix C) that the restriction to time-invariant multipliers does not significantly affect the quality of the policy.

The second method follows the development of Castañon (1997). First, we define for all m the function

$$\begin{aligned}&\Phi_{0,m}^{\boldsymbol{\lambda}}(s_m^0,f_m^0,x_m^0) \\ &\quad=\left(\frac{s_m^0}{s_m^0+f_m^0}-\lambda^0\right)x_m^0 \\ &\qquad+\mathsf{E}_y[\hat{J}_{1,m}^{\boldsymbol{\lambda}}(s_m^0+y,f_m^0+x_m^0-y)\,|\,x_m^0;s_m^0,f_m^0],\end{aligned} \tag{13}$$

which gives the value of sending x_m^0 type m messages at stage zero in the relaxed subproblem. Restricting ourselves to time-invariant λ as above, we seek a λ that induces a set of solutions to the problems $\{\max_{x_m^0\geqslant 0}\Phi_{0,m}^{\lambda}(s_m^0,f_m^0,x_m^0)\}$ that is also feasible with respect to the constraint $\sum_{m=1}^{M}x_m^0=N$ in the original problem. We note that it may not be possible to exactly satisfy the constraint in this way, in which case we estimate a suitable λ using binary search (typically seven iterations in our implementations), then use the resulting λ in problem (11) to generate a feasible stage zero allocation. We have observed numerically that $\Phi_{0,m}^{\lambda}(s_m^0,f_m^0,x_m^0)$ appears roughly concave. For large problems, we approximate it with its concave envelope in our implementation to aid computation. We approximate the objective function of (11) similarly for large problems.

We have found that the choice of method for choosing $\boldsymbol{\lambda}$ does not appear to impact the policy performance significantly (see Appendix C in the online companion), and thus

we mainly use the second method with time-invariant λ in our implementation, as it is conceptually and computationally straightforward.

Finally, we note that the algorithm is only of interest for $0 < \lambda < 1$. In fact, for λs outside this range, the proposed algorithm coincides with a greedy (or play-the-leader) algorithm that sends the message with best expectation to all N customer encounters in the segment.

PROPOSITION 5. *For* $\lambda = \lambda^0 = \cdots = \lambda^{T-1} \leqslant 0$ *or* $\lambda = \lambda^0 = \cdots = \lambda^{T-1} \geqslant 1$, *problem* (11) *is solved by the greedy solution*:

$$x_m^0 = \begin{cases} N & \text{for } m = \arg\max_{n \in \{1,\ldots,M\}} \left\{ \dfrac{s_n^0}{s_n^0 + f_n^0} \right\}, \\ 0 & \text{otherwise}. \end{cases} \tag{14}$$

PROOF. Consider the case $\lambda \leqslant 0$, and observe that $s_m^t/(s_m^t + f_m^t) \geqslant 0$ for all t, m; hence, $s_m^t/(s_m^t + f_m^t) - \lambda \geqslant 0$. Equation (9) gives $\hat{J}^{\boldsymbol{\lambda}}_{T-1,\,m}(s_m^{T-1}, f_m^{T-1}) = N(s_m^{T-1}/(s_m^{T-1} + f_m^{T-1}) - \lambda)$. Suppose that for some $t < T-1$, we have $\hat{J}^{\boldsymbol{\lambda}}_{t+1,\,m}(s_m^{t+1}, f_m^{t+1}) = N(T-t-1)(s_m^{t+1}/(s_m^{t+1} + f_m^{t+1}) - \lambda)$ for any s_m^{t+1} and f_m^{t+1}. Then, in the evaluation of Equation (10), we have

$$\begin{aligned} &\mathsf{E}_y\left[\hat{J}^{\boldsymbol{\lambda}}_{t+1,\,m}(s_m^t + y, f_m^t + x - y) \mid x; s_m^t, f_m^t\right] \\ &\quad = N(T-t-1) \cdot \mathsf{E}_y\left[\frac{s_m^t + y}{s_m^t + f_m^t + x} - \lambda \,\middle|\, x; s_m^t, f_m^t \right] \\ &\quad = N(T-t-1)\left(\frac{s_m^t + \mathsf{E}[y \mid x; s_m^t, f_m^t]}{s_m^t + f_m^t + x} - \lambda \right) \\ &\quad = N(T-t-1)\left(\frac{s_m^t + (s_m^t/s_m^t + f_m^t)x}{s_m^t + f_m^t + x} - \lambda \right) \\ &\quad = N(T-t-1)\left(\frac{s_m^t}{s_m^t + f_m^t} - \lambda \right). \end{aligned} \tag{15}$$

Thus, $x = N$ is optimal in the problem of Equation (10), and we have $\hat{J}^{\boldsymbol{\lambda}}_{t,\,m}(s_m^t, f_m^t) = N(T-t)(s_m^t/(s_m^t + f_m^t) - \lambda)$. Induction gives us this result for all $t < T$. Applying Equation (15) for $t = 0$ gives us that the objective function of problem (11) depends on the variables x_m^0 only through the term $\sum_{m=1}^{M}(s_m^0/(s_m^0 + f_m^0))x_m^0$. Problem (11) is thus a linear optimization over a simplex constraint, and is optimized by the greedy solution.

For $\lambda \geqslant 1$, we have $s_m^t/(s_m^t + f_m^t) - \lambda \leqslant 0$. A simple inductive argument gives us that problem (10) is maximized by $x = 0$, giving $\hat{J}^{\boldsymbol{\lambda}}_{t,\,m}(s_m^t, f_m^t) = 0$ for all t, m. Problem (11) becomes

$$\begin{aligned} N\sum_{\tau=1}^{T-1}\lambda^\tau + \max \sum_{m=1}^{M}\left(\frac{s_m^0}{s_m^0 + f_m^0}\right)x_m^0 \\ \text{s.t.} \quad \sum_{m=1}^{M} x_m^0 = N, \\ x_m^0 \geqslant 0, \quad m = 1, \ldots, M, \\ x_m^0 \text{ integer}, \quad m = 1, \ldots, M, \end{aligned} \tag{16}$$

which is clearly optimized by the greedy solution. □

This result underscores the importance of selecting λ to obtain good policies for the adaptive sampling problem, and gives us bounds on the set of interesting λs to be chosen using the two methods discussed in this section.

4.2. Solving the Subproblems

Consider the subproblem in Equations (9) and (10), and observe that at stage t, $(s_m^t - s_m^0)$ and $(f_m^t - f_m^0)$ may each range from 0 to Nt, such that $0 \leqslant s_m^t - s_m^0 + f_m^t - f_m^0 \leqslant Nt$. Thus, the number of states at stage t is $O(t^2N^2)$, and the total number of states in the problem is $O(T^3N^2)$. For a full backward induction, we must evaluate $O(N)$ possible decisions x_m^t and $O(N)$ possible outcomes y_m^t for each decision. Thus, solution of the subproblem for each message requires on the order of T^3N^4 operations using a full backward induction, and is difficult to solve for reasonable-sized problems. For this reason, we look to approximate solutions of the subproblems so that the algorithm is scalable to medium and large problems.

First, we observe that the subproblems are simplified greatly for limited time horizons, and that the benefits of exploration intuitively diminish as time progresses and information is accumulated. Thus, we solve the subproblems using a limited lookahead horizon H. At the end of the lookahead horizon, we estimate the value function as if there were one final stage with $N(T-H)$ customers to contact. That is, we estimate the value function at time H as

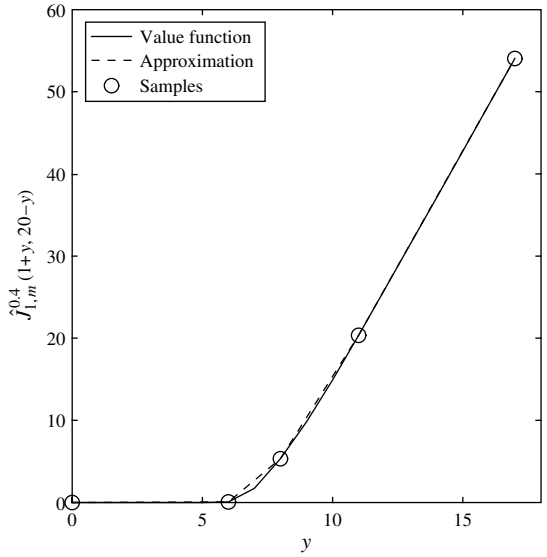
$$\begin{aligned} &\hat{J}^{\boldsymbol{\lambda}}_{H,\,m}(s_m^H, f_m^H) \\ &\quad = N(T-H)\left(\max\left\{0, \frac{s_m^H}{s_m^H + f_m^H} - \lambda^H\right\}\right). \end{aligned} \tag{17}$$

At stages $t = 1, \ldots, H-1$, the value functions are computed as in Equation (10). In our computational results, we typically use $H = 2$. Section 6 includes some computational results supporting this choice.

Another approximation we employ is to allocate messages to customers in blocks, typically of size $B = N/10$. This approximation leads to fewer solutions to consider at each stage, and allows us to consider only states for which $(s_m^t + f_m^t - s_m^0 - f_m^0)$ is divisible by the block size.

A further approximation we consider relies on a central idea of approximate dynamic programming (see Bertsekas and Tsitsiklis 1996), namely, a functional approximation of the value function in the state space. Figure 1 is a plot of a typical subproblem value function. As a function of the number of successes y observed in some stage of the problem, the value functions are everywhere approximately zero, everywhere approximately linear and increasing, or of the shape represented in Figure 1. We observe that a very close approximation can be made via linear interpolation among very few carefully chosen samples. In Figure 1, we demonstrate by comparing the true value function with an approximation formed using only five samples.

Figure 1. Illustration of an approximation of a value function using linear interpolation.



Note. Plot is as a function of the number of successes y out of $x = 17$ trials in one decision stage, given $s^0_m = 1$, $f^0_m = 3$, and $T = 5$.

Finally, we observe that for large s^t_m and f^t_m, the probability mass function of y^t_m has very little significant support outside a small range. In computing the expectations in Equation (10), we use this fact to limit our computational efforts to values of y^t_m with significant nonzero probabilities.

4.3. Extensions to Multiple Segments

Up until this point, we have considered the adaptive sampling problem within a single segment. To handle a K-segment model, a simple approach would be to replicate K independent single-segment models. However, simultaneously considering multiple segments brings about distinct challenges. In this section, we illustrate two extensions to our basic approximate dynamic programming model for use in a multisegment framework: addition of systemwide budget constraints, and accounting for customer lifetime value when the customers may migrate among segments. Either of these additions destroys the independence of customer segments, thus we must consider all segments together when we include these considerations.

Incorporating Systemwide Constraints. An interactive marketing campaign may realistically face budget or other systemwide constraints on the set of messages sent out each period. In addition, the messages may incur different profits if accepted. Extension of the approximate dynamic programming algorithm to these considerations is straightforward.

In particular, label the various customer segments $1, \ldots, K$ and let r_{km} be the expected profit if message m elicits a positive response from a segment k customer. Let b_{km} (for all segments $k = 1, \ldots, K$ and messages $m = 1, \ldots, M_k$) indicate the cost coefficients (assumed nonnegative) and let D indicate the right-hand side of a budget constraint $\sum_{k=1}^{K}\sum_{m=1}^{M} b_{km}x^t_m = D$ on the overall set of messages sent out in a single period. In addition, for the development in this section, we introduce multisegment decision variables x^t_{km}, state variables s^t_{km} and f^t_{km}, and available customer encounters N_k per stage for each segment $k = 1, \ldots, K$. We assume that the problem is feasible (for example, by including a cost-free default message).

We denote by λ_k (assumed constant in time, as discussed in §4.1) the Lagrange multiplier corresponding to the constraint on the number of customer encounters in segment k. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$. We can also apply Lagrangian relaxation to the budget constraint, yielding the following approximate formulation:

$$J^{\mu,\boldsymbol{\lambda}}_t(\mathbf{s}^t_1, \ldots, \mathbf{s}^t_K, \mathbf{f}^t_1, \ldots, \mathbf{f}^t_K) = (T-t)\left(\mu D + \sum_{k=1}^{K}\lambda_k N_k\right) + \sum_{k=1}^{K}\sum_{m=1}^{M_k}\hat{J}^{\mu,\lambda_k}_{t,k,m}(s^t_{km}, f^t_{km}), \quad (18)$$

where

$$\hat{J}^{\mu,\lambda_k}_{T-1,k,m}(s^{T-1}_{km}, f^{T-1}_{km}) = N_k \cdot \max\left\{0, r_{km}\frac{s^{T-1}_{km}}{s^{T-1}_{km}+f^{T-1}_{km}} - \lambda_k - \mu b_{km}\right\} \quad (19)$$

and

$$\begin{aligned}\hat{J}^{\mu,\lambda_k}_{t,k,m}(s^t_{km}, f^t_{km}) &= \max_x \left(r_{km}\frac{s^t_{km}}{s^t_{km}+f^t_{km}} - \lambda_k - \mu b_{km}\right)x \\ &\quad + \mathsf{E}_y[\hat{J}^{\mu,\lambda_k}_{t+1,k,m}(s^t_{km}+y, f^t_{km}+x-y) \mid x; s^t_{km}, f^t_{km}] \\ &\quad \text{s.t.} \quad 0 \leqslant x \leqslant N_k, \\ &\qquad x \text{ integer},\end{aligned} \quad (20)$$

for $t = T-2, \ldots, 0$.

The subproblems of this problem can be solved using the same procedures used to solve the subproblems of the single-segment problem. It remains to choose appropriate multipliers μ and $\lambda_1, \ldots, \lambda_K$. These can be chosen by minimizing the quantity $J^{\mu,\boldsymbol{\lambda}}_t(\mathbf{s}^t_1, \ldots, \mathbf{s}^t_K, \mathbf{f}^t_1, \ldots, \mathbf{f}^t_K)$ using subgradient methods, or, analogously to the approach we used in §4.1, we can perform a line search to select μ so that the budget constraint is made binding.

Accounting for Migrating Customers. Here we consider adaptive sampling in a multisegment system in which customers are assumed to migrate among segments according to message-dependent transition probabilities. Such an extension is relevant when the customer segmentation depends on past customer behavior data (such as RFM variables). In such systems, message choices can impact customers' future states as well as immediate rewards and the decision maker's information state.

We assume that when a customer in segment k at stage t is shown message m, he then migrates with probability P_{kml} to segment l at stage $t+1$. The migration probability matrices are assumed known with certainty. While we recognize that it is not likely to be the case that transition probabilities are known while the response rates are not, we make this assumption as an approximation for tractability. Here the problem data include the problem horizon T, the initial allocation $N_1^0, \ldots, N_K^0$ of customers to segments, the beta distribution priors $(\mathbf{s}_1^0, \ldots, \mathbf{s}_K^0, \mathbf{f}_1^0, \ldots, \mathbf{f}_K^0)$, and the transition matrices P_{kml}.

The exact solution of this problem requires a state space that includes both the vector $\{N_1^t, \ldots, N_K^t\}$ indicating the number of customers in each segment at stage t and the vectors $\mathbf{s}^t$ and $\mathbf{f}^t$ parameterizing the updated distributions of p_{km} for each segment k and message m. We implicitly assume in this section that the N_k^t are known going forward (in fact, constant in our implementation), and we account for migration through a value function adjustment at the initial time stage.

Suppose that we have at our disposal estimates $H_t(k)$ of the value of having a single customer in segment k from stage t onwards and that the total future value of a set of customers can be estimated by summing the individuals' values. As the true value function corresponding to an exact dynamic programming formulation cannot be decomposed by customer, this is an approximation of the true value of customers in the system. Nevertheless, given these estimates, we can approximately account for the migration of a single customer from segment k to l at stage t by adding to the no-migration value function the term $[H_{t+1}(l) - H_{t+1}(k)]$. Given this adjustment, we may write a decomposed dynamic programming iteration that approximately accounts for the effect of customer migration as follows:

$$J_{t,k}^{\lambda_k}(\mathbf{s}_k^t, \mathbf{f}_k^t) = N_k^t(T-t)\lambda_k + \sum_{m=1}^{M_k} \hat{J}_{t,k,m}^{\lambda_k}(s_{km}^t, f_{km}^t), \tag{21}$$

where

$$\hat{J}_{T-1,k,m}^{\lambda_k}(s_{km}^{T-1}, f_{km}^{T-1}) = N_k^{T-1} \cdot \max\left\{0, \frac{s_{km}^{T-1}}{s_{km}^{T-1} + f_{km}^{T-1}} - \lambda_k\right\} \tag{22}$$

and

$$\begin{aligned}\hat{J}_{t,k,m}^{\lambda_k}(s_{km}^t, f_{km}^t) &= \max_{x_{km}^t}\left(\frac{s_{km}^t}{s_{km}^t + f_{km}^t} - \lambda_k + \sum_{l=1}^{K} P_{kml}[H_{t+1}(l) - H_{t+1}(k)]\right)x_{km}^t \\ &\quad + \mathsf{E}_y[\hat{J}_{t+1,k,m}^{\lambda_k}(s_{km}^t + y, f_{km}^t + x_{km}^t - y) \mid x_{km}^t; s_{km}^t, f_{km}^t] \\ &\quad \text{s.t.} \quad 0 \leqslant x_{km}^t \leqslant N_k^t, \\ &\qquad\quad x_{km}^t \text{ integer},\end{aligned} \tag{23}$$

for $t = T-2, \ldots, 0$.

It remains to estimate the values $H_{t+1}(k)$, $k = 1, \ldots, K$. While there are several potential estimates, a conceptually and computationally simple one is to approximate $H_{t+1}(k)$ by the value of a single migrating customer in segment k at stage $t+1$ assuming that the reward probabilities p_{km} are fixed at their current ($t = 0$) expected values. These values can be efficiently computed ahead of time by solving a simple dynamic program with K states at each stage.

5. A Heuristic Based on a Bandit Approximation

We return to the single-segment problem of §3. In addition to the dynamic programming-based heuristic, we have designed a heuristic for the single-segment problem based on the asymptotic approximation of Lai (1987) for the finite-horizon multiarmed bandit problem, in which samples are drawn from unknown populations one at a time with the objective of maximizing the sum of rewards. The problem Lai considers is a special case of our problem with $N = 1$. We develop a new heuristic, which we call "Interval," for the problem of §3, by extending Lai's method to the case of batched decision making. We note that the Interval method we develop does not appear to naturally extend to the multisegment features introduced in §4.3.

We briefly summarize Lai's method using terminology and notation that fit with our discussion. The reader is referred to the original paper (Lai 1987) for a more complete treatment. Lai's paper considers a relatively simple heuristic for allocating customers one at a time to M messages assumed to have reward densities of the form $f(y; p_m)$, where the p_ms are unknown parameters belonging to some set $\mathcal{P}$. He develops an allocation rule that he proves is asymptotically optimal from a Bayesian perspective (for priors on p_m in the set $\mathcal{P}$ that meet certain technical conditions) as $T \to \infty$.

For the case in which we are interested, $f(y; p_m)$ is the Bernoulli distribution parameterized by an unknown success probability p_m for each message m. Suppose that we are ready to make an allocation decision for stage t. Let $\hat{p}_{m,r_{t,m}}$ indicate the maximum likelihood estimator of p_m given past observations, and let $r_{t,m}$ indicate the number of customers contacted with message m by time t. Lai's allocation rule is then to send the next customer the message m with the highest upper confidence bound $U_{m,r_{t,m}}$:

$$U_{m,r_{t,m}} = \inf\left\{p\colon p \geqslant \hat{p}_{m,r_{t,m}} \text{ and } I(\hat{p}_{m,r_{t,m}}, p) \geqslant \frac{\gamma(r_{t,m}/T)}{r_{t,m}}\right\}, \tag{24}$$

where $I(\hat{p}, p)$ is the so-called Kullback-Liebler information number, given in the Bernoulli case by $I(\hat{p}, p) = \hat{p} \cdot \log(\hat{p}/p) + (1-\hat{p}) \cdot \log((1-\hat{p})/(1-p))$, and γ is a nonnegative function satisfying certain technical conditions (namely, $\sup_{t \geqslant a} \gamma(t)/t < \infty$ for all $a > 0$, $\gamma(t) \sim \log t^{-1}$ as $t \to 0$, and $\gamma(t) \geqslant \log t^{-1} + \xi \log\log t^{-1}$ as $t \to 0$ for some ξ).

Intuitively, we can think of the "confidence bound" $U_{m,r_{t,m}}$ as an inflated version of the estimator $\hat{p}_{m,r_{t,m}}$, where the adjustment decreases with the number $r_{t,m}$ of customers already contacted with message m. Thus, the rule favors messages with high success probability estimates and also messages on which we have little accumulated experience.

Lai's method is developed for the case in which one customer is contacted at a time ($N = 1$). Furthermore, his allocation rule is not developed specifically for a Bayesian formulation of the problem. (He assumes only that the true parameters p_m are drawn from some known set.) Thus, we have developed the following heuristic based on Lai's method to be relevant for comparison with the method developed in §4.

First, we must calibrate the parameters $r_{t,m}$ and $\hat{p}_{m,t_{t,m}}$ of Lai's method to somehow reflect the available prior information at stage t. As discussed in §3.1, from the Bayesian perspective we may imagine that our prior distribution beta(s_m^t, f_m^t) on p_m at stage t has been developed starting with a uniform$(0, 1)$ (=beta$(1, 1)$) prior, then observing $s_m^t - 1$ successes and $f_m^t - 1$ failures on message m. Alternatively, we may view these observations (summed across all messages) as the first $\sum_{m=1}^{M}(s_m^t + f_m^t - 2)$ customer encounters in a hypothetical finite adaptive sampling problem involving a total of $\bar{N} = N(T-t) + \sum_{m=1}^{M}(s_m^t + f_m^t - 2)$ customer encounters. We can thus calibrate Lai's method by applying it to this larger hypothetical problem, assuming it has observed $r_{t,m} = s_m^t + f_m^t - 2$ experiments on message m (including $s_m^t - 1$ successes) and has the maximum likelihood estimator $\hat{p}_{m,r_{t,m}} = (s_m^t - 1)/(s_m^t + f_m^t - 2)$ of the true probability p_m.

In stage t, we are to contact a total of N customers before observing the associated outcomes. Arbitrarily number the N customers to be contacted in the current stage $n = 1, \ldots, N$ and set $\bar{r}_{t,m,1} = r_{t,m}$. For each customer $n = 1, \ldots, N$, we send the message $m^* = \arg\max_m\{\bar{U}_{m,\bar{r}_{t,m,n}}\}$, where

$$\bar{U}_{m,\bar{r}_{t,m,n}} = \inf\left\{p:\ p \geqslant \hat{p}_{m,r_{t,m}} \text{ and } I(\hat{p}_{m,r_{t,m}}, p) \geqslant \frac{\gamma(\bar{r}_{t,m,n}/\bar{N})}{\bar{r}_{t,m,n}}\right\}. \quad (25)$$

We also increment the $\bar{r}_{t,m,n}$ variables as follows:

$$\bar{r}_{t,m,n+1} = \begin{cases} \bar{r}_{t,m,n+1} = \bar{r}_{t,m,n} + 1 & \text{if } m = m^*, \\ \bar{r}_{t,m,n+1} = \bar{r}_{t,m,n} & \text{if } m \neq m^*. \end{cases} \quad (26)$$

Thus, our modification makes decisions about customers one at a time using the upper confidence intervals $\bar{U}_{m,\bar{r}_{t,m,n}}$, adjusting the width of the interval (but not its midpoint) after each decision. We will see in §6 that this heuristic performs quite well in our simulations.

In our implementation, we make use of the same choice of γ (called g in his paper) that Lai identifies while developing his own computational results. We also make use of a recursive algorithm Lai presents for identifying the message with the highest (or nearly highest) confidence bound without numerically computing the confidence bounds themselves. For an exact specification of this algorithm, see the original paper (Lai 1987, pp. 1,110–1,111).

The resulting "Interval" method is a new heuristic for the problem of §3 that we have developed as a computationally attractive alternative to the approximate dynamic programming-based method developed in §4 and for comparison. We note that it is not clear how to rigorously extend the Interval method to the multisegment considerations of §4.3.

6. Computational Results

In this section, we present some computational results comparing the methods developed in this paper with each other and with various benchmarks. For purposes of comparison, we consider only the single-segment problem of §3 (instead of the extended problems of §4.3) because it captures the core adaptive sampling trade-off and because both the ADP and Interval methods have been developed for this problem.

We consider a number of benchmarks and heuristics for comparison in our simulation study. The algorithms considered are as follows:

- **Ideal:** This algorithm sends to all customers the message with the highest true success probability. As the p_m are unknown to the decision maker, this algorithm is not implementable in practice but gives an upper bound on the performance of implementable methods. It is included as a benchmark.
- **Exact:** This algorithm uses backward induction to exactly solve the formulation of §3.2, and is only practical for problems of very limited size.
- **Greedy:** This algorithm sends all customers the message with the highest expected purchase probability given the current information. In the case of a tie, Greedy divides the customers equally among the tied messages. This algorithm is also commonly referred to as the "play-the-leader" rule in the multiarmed bandit literature.
- **GGreedy:** This algorithm sends all customers at a given stage the message with the highest Gittins index, computed using an infinite horizon and discount rate of 0.90. Our procedure for computing Gittins indices for this problem follows the discussion in the first chapter of Gittins (1989). In the case of a tie, GGreedy divides the set of customers equally among the tied messages.
- **Interval:** The heuristic based on the asymptotic approximation of the multiarmed bandit problem, as discussed in §5.
- **ADP:** The approximate dynamic programming algorithm using the decomposition idea outlined in §4. For the results presented in this section, we approximate the subproblems using the ideas presented in §4.2. For these problems, we typically use $H = 2$ and $B = N/10$, and we approximate the value functions using linear interpolation among 10 samples.

To compare the decisions produced by the various methods with an optimal solution, we require problems sufficiently small to be solvable using the exact backward induction specified in §3.2. We have considered several problems with two messages, stages numbering up to eight, and customers numbering up to 12. For these small examples, we have found the results for Greedy, ADP, and Interval to be indistinguishable from the results of Exact in most of the simulations. From these results, it is difficult to draw conclusions differentiating the methods' performances. We look to larger examples, for which we can judge the methods' performances relative to one another but not relative to an optimal strategy.

Our simulation study is designed as follows. For each simulation run, we select the numbers $\bar{s}$ and $\bar{f}$ to serve as parameters of a true beta distribution from which we choose true message success probabilities. To generate problem instances, true message success probabilities p_m, $m = 1, \ldots, M$, are drawn from a beta distribution with parameters $\bar{s}$ and $\bar{f}$. Then, we simulate a_m preliminary experiments on each message to strengthen the priors before we record the results.

As input to the algorithms, the prior distribution is then the beta distribution with parameters s_m^0 and f_m^0, where s_m^0 is set as $\bar{s}$ plus the number of successes from the preliminary experiments on message m, while f_m^0 is set as $\bar{f}$ plus the number of failures from the a_m preliminary message m experiments. This method of generating prior distributions reflects a hypothetical situation in which the decision maker initially begins with rough (but statistically consistent) priors on p_m, then develops beliefs on p_m based on a few off-line preexperiments. We look at examples with constant a_m across messages, which reflect situations in which the amount of prior information is the same for all messages, and with a_m randomly chosen, which represents potentially more interesting cases in which the amount of available information varies across messages.

Table 1 represents numbers of successes averaged over 2,000 simulated problems with batch sizes up to 1,000 per stage. We denote those table entries in boldface whose value cannot be statistically distinguished at a 95% confidence level from the best of Greedy, GGreedy, Interval, and ADP. For Greedy, Interval, and ADP, we present the same results in terms of regret (defined as the difference between successes under Ideal and successes under the method of interest) in Table 2. The last two columns of this table give the percent reduction in measured regret achieved by ADP relative to Greedy and by ADP relative to Interval.

We observe from Tables 1 and 2 that the ADP and Interval methods perform similarly to each other and better than the Greedy and GGreedy methods over a wide range of problems. In most cases, the advantage of the ADP and Interval methods is statistically significant. In addition, Table 2 shows that they achieve a sizeable percentage reduction in average regret versus the other methods. The ADP and Interval methods give the largest improvements in those cases where we might have expected the value of adaptive sampling to be the greatest, namely, where there is little prior information available.

To further examine the effects of the problem parameters on the performance of the methods, we plot in Figure 2 the average performance of the methods as functions of the problem parameters. The first plot, Figure 2(a), plots regret of the methods versus batch size N for a problem with 10 messages. Message success probabilities p_m were chosen randomly from the distribution beta$(2, 8)$, and the number of extra prior samples a_m was chosen uniformly on $[20, 40]$. Customer sets of size 1,000, 500, 250, and 100 were tried, with the number of stages chosen so that the total number of customers was 2,000 for each simulation run. The performances of the adaptive methods generally improve as the number of customers is decreased and the number of stages is increased. This matches the behavior of the optimal solution, as proven in Proposition 4. We observe that the ADP method outperforms Interval for large N and $T \leqslant 10$. It underperforms Interval for relatively large T and small N, which is likely due to the limited lookahead approximation in the ADP subproblem computations.

Table 1. Average number of successes for a variety of large adaptive sampling problems.

					Average successes				
$\bar{s}$, $\bar{f}$	T	N	M	a_m	Ideal	Greedy	GGreedy	Interval	ADP
2, 8	10	50	10	$U[0, 20]$	204.78	185.63	187.48	**190.12**	**189.78**
2, 8	10	100	10	$U[0, 20]$	412.68	373.64	378.63	**388.87**	**388.60**
2, 8	10	100	10	20	418.44	399.46	400.99	**403.94**	403.40
2, 8	10	100	10	$U[0, 50]$	411.06	388.05	390.82	394.55	**395.04**
2, 50	10	100	8	$U[30, 60]$	82.80	73.92	74.24	**74.83**	**74.71**
4, 100	10	100	6	$U[100, 200]$	64.64	**60.36**	**60.43**	**60.37**	60.10
1, 3	6	200	10	20	708.80	691.01	691.67	693.58	**694.06**
1, 3	6	200	10	$U[0, 20]$	711.65	665.64	669.17	**682.52**	**682.66**
1, 3	6	200	10	$U[0, 50]$	712.80	687.99	690.16	**695.21**	**695.73**
2, 50	6	200	8	$U[30, 60]$	98.89	87.40	87.95	**88.75**	**88.61**
4, 100	6	200	6	$U[100, 200]$	77.64	**72.33**	**72.37**	**72.28**	**72.24**
2, 50	5	1,000	8	$U[30, 60]$	406.28	362.29	363.53	**374.14**	373.36
4, 100	5	1,000	6	$U[100, 200]$	323.58	305.07	306.00	**307.03**	**306.99**

Table 2. Average regret for a variety of large adaptive sampling problems.

					Average regret			% Improvement	
$\bar{s}, \bar{f}$	T	N	M	a_m	Greedy	Interval	ADP	ADP vs. Greedy	ADP vs. Interval
2, 8	10	50	10	$U[0, 20]$	19.15	**14.66**	**15.00**	21.7	−2.3
2, 8	10	100	10	$U[0, 20]$	39.04	**24.07**	**23.81**	38.3	−1.1
2, 8	10	100	10	20	18.97	**14.50**	15.03	20.8	−3.7
2, 8	10	100	10	$U[0, 50]$	23.01	16.51	**16.02**	30.4	3.0
2, 50	10	100	8	$U[30, 60]$	8.88	**7.97**	**8.09**	8.9	−1.5
4, 100	10	100	6	$U[100, 200]$	**4.28**	**4.27**	4.54	−6.0	−6.4
1, 3	6	200	10	20	17.80	15.22	**14.74**	17.1	3.2
1, 3	6	200	10	$U[0, 20]$	46.01	**29.13**	**28.99**	37.0	0.5
1, 3	6	200	10	$U[0, 50]$	24.81	**17.59**	**17.06**	31.2	3.0
2, 50	6	200	8	$U[30, 60]$	11.49	**10.14**	**10.27**	10.6	−1.3
4, 100	6	200	6	$U[100, 200]$	**5.31**	**5.36**	**5.40**	−1.8	−0.8
2, 50	5	1,000	8	$U[30, 60]$	44.00	**32.14**	32.93	25.2	−2.4
4, 100	5	1,000	6	$U[100, 200]$	18.50	**16.55**	**16.59**	10.4	−0.3

Figure 2(b) illustrates the improvement in the performances of the adaptive methods as the priors are strengthened. The points plotted represent the average number of successes for a problem with message success probabilities chosen from beta$(2, 8)$. We make decisions for 400 customers at each of the five stages. The number a_m of extra pre-experiments is fixed at 0, 5, 10, 20, 50, and 100. This plot reveals that all of the methods benefit from additional prior information, in keeping with Proposition 1, which implies that information has value given the optimal policy. The performance gaps between the various methods narrow as the amount of prior information is increased, supporting the intuition that intelligent adaptive sampling has less impact the more accurate prior information is available. For the problems presented, the ADP and Interval methods perform similarly over a wide range of available prior information, both offering substantial benefits over the two greedy methods.

We conclude from the computational results presented in this section that the ADP and Interval methods perform similarly to each other and at least as well as the other methods over a wide range of problems, and achieve significant gains when the amount of prior information is low. For certain problem instances, the Greedy method may perform sufficiently well to be the preferred method due to its simplicity, but we expect adaptive sampling to be particularly beneficial in marketing situations with many customers and in very dynamic and changing environments in which short planning horizons are appropriate. The Interval heuristic based on Lai's (1987) algorithm is particularly fast to compute and performs quite well over a wide range of parameters. Thus, this heuristic may be appropriate when computational speed is at a premium. The fact that the ADP approach is based on a mathematical programming formulation makes it extendable. Relevant extensions have been developed in §4.3.

We have implemented versions of the ADP method that select λ by minimizing $\bar{J}_0^{\lambda}(\mathbf{s}^0, \mathbf{f}^0)$ and that relax the computational assumption that $\boldsymbol{\lambda}$ is constant over time. Experiments on a few problems have shown that the method we use for selecting $\boldsymbol{\lambda}$ performs better than the minimization method and indistinguishably as well as the version without the assumption of constant $\boldsymbol{\lambda}$. Details can be found in the online companion (Appendix C) to this paper.

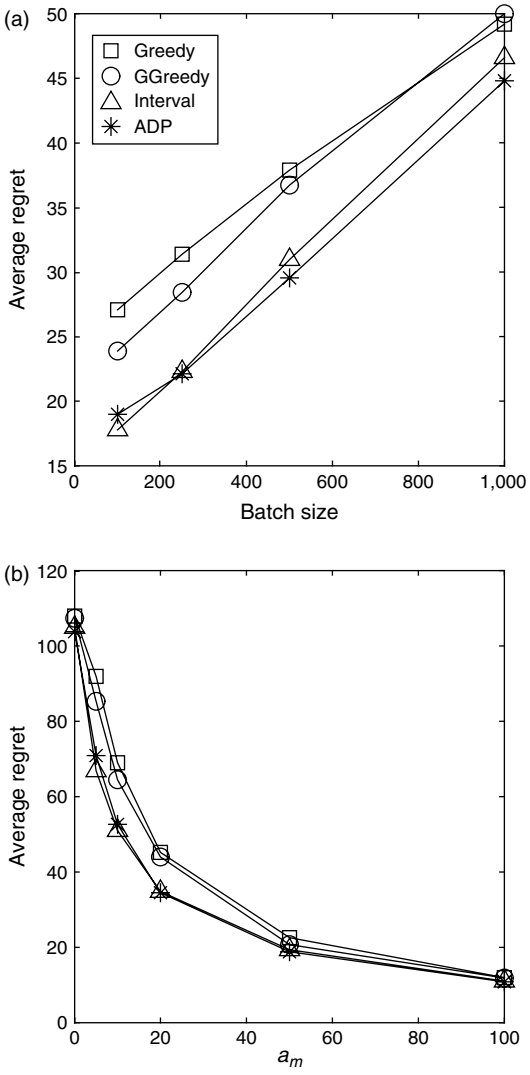
We have also investigated the sensitivity of the ADP algorithm performance to choices of the lookahead horizon H and the decision block size B. The choice of B appears to have little impact on the results, and we have generally chosen $B = N/10$, which seems to provide good results relatively efficiently. Table 3 gives results for various choices of H for selected parameters, averaged over 2,000 problem runs. We see that while $H = 2$ seems to consistently outperform $H = 1$, little or no improvement is noted for $H > 2$. In the interest of simplicity and computation time, we have used $H = 2$ in the other results reported in this paper.

Finally, we give some evidence of the computational effort required for the various algorithms. Table 4 gives processor times required per decision for several of the algorithms discussed in the previous section for selected parameters, averaged over 2,000 runs. We note that all of the solution methods discussed in this paper are capable of generating decisions for the problems mentioned in no more than a few seconds.

7. Directions for Future Research

To our knowledge, we are among the first to propose a workable solution to the problem of adaptive experimentation in an interactive marketing context. While we believe this is an important problem, solving it comes at a price. As mentioned in §1, there is a gap between the relative

Figure 2. Impact of parameter changes on average regret. (a) Batch size indicated on x-axis, and time horizon varied as $N = 2{,}000/T$. (b) Number of extra preexperiments a_m for each message is fixed at the number indicated on the x-axis.



simplicity of our model of customer behavior and the relative sophistication of state-of-the-art customer response models used in marketing research for descriptive and myopic optimization purposes. It remains to be seen if the benefits of active learning in a real-world implementation justifies the modeling sacrifice, and it remains a challenge to develop rigorous adaptive learning technologies for more complicated and realistic models of customer behavior, for example, allowing dependence among messages, dependence among customer segments, and customer heterogeneity. In short, we believe that there is much room for interesting future research in this area.

We feel that research in this area remains challenging for a few fundamental reasons. The first is that Bayes Rule tends to generate intricate statistical dependence even in relatively simple statistical models. We managed this complexity using the beta/binomial conjugacy and assuming independence among segments and messages, but learning models that deal more directly with parameter dependence are of interest. Contemporary Bayesian customer response models (e.g., Rossi et al. 2005) typically make use of simulation-based techniques, like Markov Chain Monte Carlo, to generate Bayes posteriors. Such techniques are computationally intensive and the results are not easily parameterized. Hence, it remains a challenge to interface such techniques tractably with dynamic optimization formulations. We believe that approximations, either of the customer behavior model or of optimality or both, are necessary. Our ADP approach shows that problem decomposition is a powerful approximation tool, although we also believe that functional and distributional approximations of the Bayes updates warrant investigation.

Possible alternatives to the Bayes solution include non-Bayesian approaches or approaches that seek a weaker form of optimality (e.g., asymptotic optimality). We note that a portion of the past research on the multiarmed bandit problem and its variants has sought optimality in an asymptotic sense. Our adaptation of the asymptotically optimal sampling method of Lai (1987) is fast to compute and yields rewards similar to the ADP approach. However, it remains an interesting topic of research to adapt such techniques, themselves based on relatively simple problem formulations, to constraints and customer dynamics that may be important features of real interactive marketing problems.

A second challenge in studying adaptive experimentation in marketing contexts is in testing the methodologies. Testing an adaptive learning model requires not just data but also decision-making control in a real-world system. While examples of successful field testing exist (e.g., Simester et al. 2006), it is understandable why marketers have been reluctant to cede this control to academic researchers.

Table 3. Simulation results comparing different choices of lookahead horizon H for the ADP algorithm.

$\bar{s}$, $\bar{f}$	T	N	a_m	Ideal	Greedy	Interval	ADP $H=1$	ADP $H=2$	ADP $H=3$	ADP $H=4$
2, 8	10	100	$U[0, 20]$	412.68	373.64	388.86	388.17	388.60	388.38	388.49
2, 8	10	50	$U[0, 20]$	204.78	185.63	190.12	189.62	189.78	189.76	189.72
4, 100	5	1,000	$U[100, 200]$	323.58	305.07	307.03	306.80	306.99	307.29	307.20
2, 50	6	200	$U[30, 60]$	98.89	87.40	88.75	88.52	88.61	88.81	88.79

Table 4. Computation times per stage for the various methods.

$\bar{s}$, $\bar{f}$	T	N	a_m	Interval	ADP $H=1$	ADP $H=2$	ADP $H=3$	ADP $H=4$
2, 8	10	100	$U[0, 20]$	<0.005	0.01	0.69	1.92	3.37
4, 100	5	1,000	$U[100, 200]$	0.02	0.01	0.53	1.26	1.97

Notes. Numbers represent average CPU time in seconds on an Intel Xeon 2.4 GHz computer. The Greedy algorithm was found to take negligible time per stage (<0.005 seconds).

8. Electronic companion

An electronic companion to this paper is available as part of the online version that can be found at http://or.journal.informs.org/.

Endnotes

1. The online companion includes proofs of the propositions in §3.3, proofs of the structural properties of $\bar{J}_0^{\boldsymbol{\lambda}}(\mathbf{s}^0, \mathbf{f}^0)$ claimed in §4.1, and computational evidence supporting the choice of method for selecting $\boldsymbol{\lambda}$ discussed in §4.1.
2. Working versions of this paper bore the titles "Adaptive Interactive Marketing to a Customer Segment" and "A Learning Approach to Customized Marketing."

Acknowledgments

This work represents part of the second author's Ph.D. thesis at the Operations Research Center, Massachusetts Institute of Technology. The second author thanks the Graduate School of Business, University of Chicago, for support. Both authors thank three anonymous referees, an associate editor, and area editor Hemant Bhargava for helpful comments and suggestions.

References

Adelman, D., A. Mersereau. 2007. Relaxations of weakly coupled stochastic dynamic programs. *Oper. Res.* Forthcoming.

Anantharam, V., P. Varaiya, J. Walrand. 1987. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part I: I.I.D. rewards. *IEEE Trans. Automat. Control* **32**(11) 968–976.

Ansari, A., C. F. Mela. 2003. E-customization. *J. Marketing Res.* **40**(2) 131–145.

Ariely, D., G. Bitran, P. R. Oliveira. 2002. Design to learn: Customizing services when the future matters. Working paper, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA.

Ben-Akiva, M., S. R. Lerman. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Berry, D., B. Fristedt. 1985. *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, UK.

Bertsekas, D., J. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Bitran, G., S. Mondschein. 1996. Mailing decisions in the catalog sales industry. *Management Sci.* **42**(9) 1364–1381.

Bult, J. R., T. Wansbeek. 1995. Optimal selection for direct mail. *Marketing Sci.* **14**(4) 378–394.

Caro, F., J. Gallien. 2007. Dynamic assortment with demand learning for seasonal consumer goods. *Management Sci.* **53**(2) 276–292.

Castañon, D. 1997. Approximate dynamic programming for sensor management. *Proc. 36th IEEE Conf. Decision and Control*, San Diego, CA, 1202–1207.

Cheng, Y., F. Su, D. Berry. 2002. Asymptotic optimal group sequential strategies in two-armed bandit problems. Technical Report UTMDABTR-001-02, M.D. Anderson Cancer Center, University of Texas, Austin, TX.

Drake, A. W. 1967. *Fundamentals of Applied Probability Theory*. McGraw Hill, New York.

Gittins, J. 1979. Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B* **41** 148–164.

Gittins, J. 1989. *Multiarmed Bandit Allocation Indices*. Wiley, Chichester, UK.

Gittins, J., D. M. Jones. 1974. A dynamic allocation index for the sequential design of experiments. J. Gani, ed. *Progress in Statistics*. North-Holland, Amsterdam, The Netherlands, 241–266.

Gönül, F., M. Z. Shi. 1998. Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Sci.* **44**(9) 1249–1262.

Gooley, C., J. Lattin. 2000. Dynamic customization of marketing messages in interactive media. Research Paper 1664, Graduate School of Business, Stanford University, Stanford, CA.

Hardwick, J., Q. Stout. 2002. Optimal few-stage designs. *J. Statist. Plan. Infer.* **104** 121–145.

Hawkins, J. 2003. A Lagrangian decomposition approach to weakly coupled dynamic optimization problems and its applications. Ph.D. thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Lai, T. L. 1987. Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**(3) 1091–1114.

Lai, T. L., H. Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6** 4–22.

Meuleau, N., M. Hauskrecht, K.-E. Kim, L. Peshkin, L. Kaelbling, T. Dean, C. Boutilier. 1998. Solving very large weakly coupled Markov decision processes. *Proc. 15th National Conf. Artificial Intelligence*, Madison, WI. American Association for Artificial Intelligence, Menlo Park, CA, 165–172.

Raiffa, H., R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA.

Rossi, P. E., R. McCulloch, G. M. Allenby. 1996. The value of purchase history data in target marketing. *Marketing Sci.* **15**(4) 430–444.

Rossi, P. E., R. McCulloch, G. M. Allenby. 2005. *Bayesian Statistics and Marketing*. Wiley, New York.

Simester, D., P. Sun, J. N. Tsitsiklis. 2006. Dynamic catalog mailing policies. *Management Sci.* **52**(5) 683–696.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. *A Celebration of Applied Probability. J. Appl. Probab.* **25A** 287–298.

Yost, K. A., A. R. Washburn. 2000. The LP/POMDP marriage: Optimization with imperfect information. *Naval Res. Logist.* **47** 607–619.