Active Learning for Personalizing Treatment

Kun Deng
Department of Statistics
University of Michigan
Email: kundeng@umich.edu

Joelle Pineau
Department of Computer Science
McGill University
Email: jpineau@cs.mcgill.ca

Susan Murphy
Department of Statistics
University of Michigan
Email: samurphy@umich.edu

***Abstract*—The personalization of treatment via genetic biomarkers and other risk categories has drawn increasing interest among clinical researchers and scientists. A major challenge here is to construct individualized treatment rules (ITR), which recommend the best treatment for each of the different categories of individuals. In general, ITRs can be constructed using data from clinical trials, however these are generally very costly to run. In order to reduce the cost of learning an ITR, we explore active learning techniques designed to carefully decide whom to recruit, and which treatment to assign, throughout the online conduct of the clinical trial. As an initial investigation, we focus on simple ITRs that utilize a small number of subpopulation categories to personalize treatment. To minimize the maximal uncertainty regarding the treatment effects for each subpopulation, we propose the use of a minimax bandit model and provide an active learning policy for solving it. We evaluate our active learning policy using simulated data and data modeled after a clinical trial involving treatments for depressed individuals. We contrast this policy with other plausible active learning policies. The techniques presented in the paper may be generalized to tackle problems of efficient exploration in other domains.**

I. Introduction

We propose to actively learn individualized treatment rules (ITRs), which are of increasing interest in the personalization of treatment. Formally, for each patient we have the pre-treatment observation variable $X \in \mathcal{X}$, summarizing various aspects of individual heterogeneity, treatment A taking values in a finite, discrete treatment space $\mathcal{A}$, and a real-valued response R (assuming large values are desirable). An ITR, denoted d, is a deterministic decision rule from $\mathcal{X}$ into the treatment space $\mathcal{A}$. We aim to construct this rule so as to maximize the future response R. From an AI perspective, learning ITRs is an example of reinforcement learning with a time horizon of 1 step, and clinical trials that aim to discover good ITRs represent the phase of exploration for learning a good policy.

Consider the following motivating example, in which one must decide between two actions (a_1 and a_2) for treating subjects from each of four subpopulations ($c_1 \sim c_4$). We consider perhaps the simplest ITR, which will assign each subpopulation to one of the two competing treatment options, ignoring other characteristics/covariates of the subjects. This setting, though simple, is not uncommon in many clinical trials [8]. For instance, a subpopulation type may correspond to patients with a particular Gene biomarker, and the two treatment options are the standard and alternative treatment option, or the top two choices of treatment for that disease. Further, we assume the mean response under each treatment option for subpopulation c_i is μ_{i1} and μ_{i2}. So the ITR looks like this:

$$d(c_i) = \begin{cases} a_1 & \text{if } \hat{\mu}_{\mathrm{i1}} - \hat{\mu}_{\mathrm{i2}} \geq 0 \\ a_2 & \text{if } \hat{\mu}_{\mathrm{i1}} - \hat{\mu}_{\mathrm{i2}} < 0 \end{cases} \quad \forall i \in \{1, 2, 3, 4\}$$

where the $\hat{\mu}_{i\cdot}$ are the estimates of $\mu_{i\cdot}$. We observe that the confidence in the correctness of an ITR lies in the variances of $\hat{\mu}_{i1} - \hat{\mu}_{i2}$, which we call the treatment effect for the ith subpopulation. Assuming the patients in the subpopulations respond independently, we have that the variance of the estimated treatment effect $\mathrm{Var}[\hat{\mu}_{i1} - \hat{\mu}_{i2}] = \mathrm{Var}[\hat{\mu}_{i1}] + \mathrm{Var}[\hat{\mu}_{i2}]$. In current trials for constructing ITRs, patients are generally recruited as they arrive, so that the number of patients in each subpopulation reflects their natural composition in the general population. However, this may not be the best utilization of the trial resources, and can be problematic, especially for cases where patients from certain subpopulations are rare.

We reiterate that our primary goal for running these trials is to quantify the treatment effects and their variances. More specifically, we would like to distribute the resources wisely so that the constructed ITR has a bounded uncertainty (or vice versa, high confidence) for each subpopulation. Since it may take significantly more resources to estimate the uncertainty of the estimated treatment effect for one subpopulation than for the others, enrolling patients in the trial as they arrive (hence with numbers of patients in each subpopulation reflecting the relative sizes of the subpopulations) may not yield an overall good ITR. As an extreme case, suppose that at the end of the trial we had only learned, with high confidence, that patients in subpopulation c_1 respond better to treatment a_1 than a_2, but we could not state anything about the relative importance of the treatments for the remaining subpopulations. In such case, it would seem the trial has at least partially failed in the task of constructing an overall good ITR. Furthermore, currently, most trials also use equal randomization of the treatment options for a subpopulation, which could not only waste trial resources of a particular treatment, but also run the risk of yielding highly variable responses for that subpopulation.

Three further characteristics of these clinical trials are the following. First, the trials are of relatively short duration compared to the pace of patient recruitment. Second, once a patient is recruited into the trial, the treatment and monitoring process is extremely costly. Third, usually the budget for the

clinical trial is specified a priori and thus will allow for the recruitment and treatment of, say , N subjects. For these reasons, we should carefully decide who to recruit and what treatment to assign in order to meet the goal of these trials.

We propose a minimax bandit model that intelligently recruits patient from different subpopulations and assigns them to different treatments in order to minimize the largest variance of the estimated treatment effects among the different subpopulations. In the language of reinforcement learning (RL), the action space is composed of (subpopulation, treatment) pairs. Each decision step corresponds to the recruitment and treatment of a new subject. Thus our problem can be viewed as a form of exploration in an online RL setting, with the horizon being N steps corresponding to N subjects.

This formulation bears formal similarity with some of response adaptive trials [6]–[8], popular in cancer research, which also divide patients into groups. However our formal criterion (defined below) is different from these response adaptive trials. Response adaptive trials usually aim to place more patients on the better treatment based on patient responses already accrued in the trial. In reinforcement learning terms, our goal is more exploratory while the goal in response adaptive trials is more exploitative.

The rest of the paper is organized as follows. In Section II we give brief discussion for related work in active learning and budgeted bandit problems. In Section III we present the technical details of our methods and algorithms, followed by experimental results in Section IV. We conclude this paper with an extended discussion of potential issues and future work in Section V.

II. Related Work

Our problem is natually related to the famous "multi-armed bandit problem" by Robbins (1952), in which a gambler repeatedly chooses a slot machine to play, each with a different payoff. The gambler's goal is to maximize the total payoff over all pulls of all machines. The Multi-armed bandit problem is a prototypical example in RL that characterizes the necessities of a trade-off between exploration and exploitation, in an online decision process. A key difference between our work and the conventional multi-armed bandit problem is that in the latter, one tries to maximize the cumulative rewards over all pulls, whereas with our work, one simply wants to maximize the confidence of the resulting ITRs, which is a highly nonlinear reward. Problems with similar interests in only the "end results" as ours have been studied under the names of "budgeted" multi-armed bandit problems [9,17,21], where one tries to optimize a goal function, say picking a arm of a slot machine with maximal payoff, designing a classifier with minimal prediction loss, estimating quantities with minimal variances etc, *after* an experimentation phase that is typically constrained by a time or cost budget. For example, an application of "budgeted" multi-armed bandit problems in classification [17]–[21] is called budgeted learning , which allows the learner to request more complete *feature* information by "buying" the attributes (features), spending up to at most a fixed budget. When the budget is exhausted, the learner must output a classifier that can predict as accurately as possible. Budgeted learning is related to our problem in several ways. First, similar to the budget/cost constraint in these learning problems, the budget constraint in our problem is usually set upfront before the learning starts, so it is a "hard" constraint. For example, one of the requirements for a real clinical trial is to specify and justify the cost of the trial subject to specification of the minimum and maximum number of patients. Ideally, the budget should be set as low as possible to increase the chance of a trial obtaining funding, yet high enough to ensure that upon the completion of the trial, answers to important research questions can be drawn reliably from the results of the trial. Secondly, in budgeted learning, the reward is the performance of the *final* classifier once the budget is exhausted. This is a very different criterion from maximizing the total payoff of a conventional multi-armed bandit problem. Indeed, we formulated the objective function for our problem in a similar fashion since we primarily care about endpoint properties (e.g. the quality or variance of the estimated treatment effect) of the resulting treatment rules. Another example of budgeted bandit problem is described in Antos et al. [9]. There, a problem of learning the mean values of distributions associated with a finite number of options was considered, with a similar goal of reducing the variances of the estimated mean values. One major difference between their work and ours is that the criterion we consider imposes a group structure among the options for the same subpopulation (see below for specifics).

In the machine learning literature, active learning or adaptive sampling refers to a methodology for guiding the data acquisition, by parsimoniously querying some unknown aspects of existing data, or collecting new data based on the examples that the learning algorithm has seen so far. Traditionally, active learning [10,12]–[14,16,20,22]. has been studied in the context of supervised machine learning and classification problems. Active learning also has a long history in the statistics literature, which is generally referred to as optimal experimental design; see [11,23]–[25] for some recent work and review. The main objective of this line of research is to reduce the variance of prediction over parameter estimates, while controlling the bias of the prediction at the same time. Our problem shares some similarity with the above problems in that our goal is to minimize the maximal variance of the estimated treatment effects. The possibility of active learning also arises in other domains, such as in the unsupervised learning task of density estimation. For example, [15] presented a framework for actively learning parameters in Bayesian networks. It is assumed in this framework that some subset of the variables are controllable, so a query has the form of these variables taking specified values, and the result of such a query is a randomly sampled instance conditioned on the controlled variables. If we think of subpopulation type, treatment and response as variables in a Bayesian network, our problem could be be framed as such a parameter estimation problem. Finally, there are a number of RL approaches that also use

active learning such as [1]–[5]; the main difference is that these algorithms deal with infinite horizon and cumulative reward functions, and mostly focus on exploring either state or action, whereas we are interested in balancing exploration of actions over different subsets of the state space.

III. Methods and Algorithms

This problem is formally described as follows: there are C bandits (corresponding to the C subpopulations), each equipped with K arms (corresponding to K treatments; here $K = 2$). At each time step, corresponding to the recruitment of a patient, we are only allowed to pick one bandit. For that bandit, we need to further choose an arm to pull. There will be a total of N pulls overall. We assume that the response of the jth arm of the ith bandit, denoted as (i, j), follows the distribution D_{ij} , with mean μ_{ij} (corresponding to the primary response of treatment (i, j)) . At each time point n, the estimated mean response of arm (i, j) is $\hat{\mu}_{ij}^n$, and the error of the estimate is measured with the mean squared loss:

$$L_{ij}^n = E[(\hat{\mu}_{ij}^n - \mu_{ij})^2] = \mathrm{Var}[\hat{\mu}_{ij}^n].$$

The loss for bandit i is measured by the summation of the squared losses:

$$L_i^n = \sum_{j=\{1,2\}} L_{ij}^n,$$

which is the variance of the estimated treatment effect for the i-th subpopulation, discussed earlier. (Note for convenience, our slight abuse of symbol j as a dummy variable for the summation index). The overall loss of an online active learning policy π is measured by the worst-case loss over the K bandits:

$$L^n(\pi) = \max_{1 \le i \le C} L_i^n.$$

We would like to design π such that the loss $L^n(\pi)$ is small. As mentioned earlier, a similar worst-case loss is considered by [9]; using our notation, the loss considered in [9] is $\max_{1 \le i \le C;\ j=1,2} L_{ij}^n$. However the loss $L^n(\pi)$ makes more sense in the context of the clinical trial since the focus is on a relative comparison between mean responses (e.g. the treatment effect), as opposed to the mean responses themselves. The $L^n(\pi)$ loss is motivated by our preference for an ITR whose uncertainty of treatment effect is bounded in the worst case over the different subpopulations. In the clinical setting, the ITR is most useful if it provides a high quality recommendation for each subpopulation.

To derive active learning policies for this problem, it is helpful to first consider an optimal "oracle" allocation policy O, that has been endowed with the knowledge of the variances σ_{ij}^2 of the responses. Note that if an arm (i, j) has been pulled T_{ij} times, then $\mathrm{Var}[\hat{\mu}_{ij}] = \frac{\sigma_{ij}^2}{T_{ij}}$. Recall that there are a total of N pulls. The loss of the optimal allocation policy O can be computed by solving the following convex optimization problem, ignoring integer constraints on T_{ij}:

$$\begin{aligned} \underset{T_{ij}}{\text{minimize}} \quad & \max_i \textstyle\sum_j \frac{\sigma_{ij}^2}{T_{ij}} \\ \text{s.t.} \quad & \textstyle\sum_i \sum_j T_{ij} = N \\ & T_{ij} \ge 0 \qquad i \in \{1, ..., C\}, j \in \{1, ..., K\} \end{aligned} \tag{1}$$

The optimal allocation for the ith bandit and jth arm, ignoring integer constraints, is

$$T_{ij}^* = \frac{\sigma_{ij} \sum_j \sigma_{ij}}{\sum_i (\sum_j \sigma_{ij})^2} N \tag{2}$$

To see this, note that the above convex optimization problem can be restated as:

$$\begin{aligned} \underset{T_{ij}, r}{\text{minimize}} \quad & r \\ \text{s.t.} \quad & \textstyle\sum_j \frac{\sigma_{ij}^2}{T_{ij}} \le r \qquad \forall i \in \{1, ..., C\} \\ & \textstyle\sum_i \sum_j T_{ij} = N \\ & -T_{ij} \le 0 \qquad \forall i \in \{1, ..., C\}, j \in \{1, ..., K\} \end{aligned}$$

The full Lagrangian is

$$\begin{aligned} L(T_{ij}, r, \lambda_{ij}, \alpha, \beta_{ij}) = r &+ \sum_i \lambda_i (\sum_j \frac{\sigma_{ij}^2}{T_{ij}} - r) \\ &+ \alpha (\sum_i \sum_j T_{ij} - N) \\ &+ \sum_i \sum_j \beta_{ij} (-T_{ij}) \end{aligned}$$

By the KKT condition, $\beta_{ij} = 0$ as T_{ij} has to be strictly positive ; $\frac{\partial L}{\partial T_{ij}} = 0$ yields

$$\alpha = \frac{\lambda_i \sigma_{ij}^2}{T_{ij}^2}$$

for all i and j; $\frac{\partial L}{\partial r} = 0$ yields

$$1 - \sum_i \lambda_i = 0,$$

and finally, $\alpha(\sum_i \sum_j T_{ij} - N) = 0$. It's easy to verify that T_{ij}^* in (2) and $r^* = \frac{\sum_i (\sum_j \sigma_{ij})^2}{N}$ are solutions.

Thus, the oracle would recruit T_{ij}^* subjects from subpopulation i and assign them treatment j. T_{ij}^* is proportional to the quantity $\sigma_{ij} \sum_j \sigma_{ij}$, which is an indicator of the uncertainties of both treatment (i, j) and subpopulation i. As a matter of fact, $\left\{ \frac{\sigma_{ij} \sum_j \sigma_{ij}}{\sum_i (\sum_j \sigma_{ij})^2}; i \in \{1, ..., C\}, j \in \{1, ..., K\} \right\}$ forms a proper probability distribution, so if an active learning algorithm samples according to this distribution at each of the N decision points, it will end up allocating in expectation T_{ij}^* for each arm (i, j). Of course, the values of σ_{ij} are unknown in an online setting, however they can be estimated via the sample standard deviation, $\hat{\sigma}_{ij}$, when there is sufficient data. Also, according to (2), when either σ_{ij} or $\sum_j \sigma_{ij}$ is large, arm (i, j) or bandit j should be pulled more often, which intuitively makes sense. Our proposed active learning policy called AREOA(Adaptive Randomization with Estimated

TABLE I
DATASETS FOR THE EXPERIMENT

dataset	subpopulation/ treatments	distributions	means	variances
DS1	4/2	$\begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix}$	$\begin{pmatrix} 1 & 4 \\ 2 & 2 \\ 4 & 1 \\ 2 & 2 \end{pmatrix}$	$\begin{pmatrix} 1000 & 1000 \\ 100 & 100 \\ 100 & 100 \\ 100 & 100 \end{pmatrix}$
DS2	4/2	$\begin{pmatrix} .1 \\ .3 \\ .3 \\ .3 \end{pmatrix}$	$\begin{pmatrix} 1 & 4 \\ 2 & 2 \\ 4 & 1 \\ 2 & 2 \end{pmatrix}$	$\begin{pmatrix} 1000 & 1000 \\ 100 & 100 \\ 100 & 100 \\ 100 & 100 \end{pmatrix}$
DS3	8/2	$\begin{pmatrix} .125 \\ .125 \\ \dots \\ .125 \end{pmatrix}$	$\begin{pmatrix} 2 & 2 \\ 2 & 2 \\ \dots & \dots \\ 2 & 2 \end{pmatrix}$	$\begin{pmatrix} 5 & 5 \\ 10 & 10 \\ \dots & \dots \\ 640 & 640 \end{pmatrix}$
DS4	4/2	$\begin{pmatrix} .25 \\ .25 \\ .25 \\ .25 \end{pmatrix}$	$\begin{pmatrix} 1 & 4 \\ 2 & 2 \\ 4 & 1 \\ 2 & 2 \end{pmatrix}$	$\begin{pmatrix} 100 & 1000 \\ 100 & 1000 \\ 100 & 1000 \\ 100 & 1000 \end{pmatrix}$
DS-CBASP	3/2	$\begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix}$	$\begin{pmatrix} 10.9 & 16.2 \\ 9.3 & 19.4 \\ 12.9 & 15.8 \end{pmatrix}$	$\begin{pmatrix} 99.3 & 79.7 \\ 110.7 & 55.9 \\ 103.5 & 78.6 \end{pmatrix}$

Optimal Allocation), directly exploits this insight by using this criterion as the basis for selecting selecting subjects and assigning treatments.

IV. EXPERIMENTS

In this section we evaluate our proposed strategy, AREOA, on a number of synthetic datasets. AREOA is an ϵ-greedy active learning policy that is based on the oracle allocation T_{ij} derived above; here the frequency that the subpopulation and treatment (i, j) will be picked is roughly proportional to the $\hat{\sigma}_{ij} \sum_j \hat{\sigma}_{ij}$ where $\hat{\sigma}_{ij}$ are estimated standard deviation of the treatment response on subpopulation i under treatment j. Throughout the experiments, we fixed the tuning parameter $\epsilon = 0.1$ for AREOA.

We compare AREOA's performance with two alternative active learning policies. The first alternative, denoted AARandom, recruits subjects from the subpopulation according to the subpopulation fraction in the general population and assigns the treatment among uniformly at random. The second alternative, denoted GAFS-MAX, is an active learning policy proposed in [9]; this policy deterministically selects the next (subpopulation, treatment) pair with the highest estimated sample variance but forcing a revisiting of (subpopulation, treatment) pairs that haven't been visited for some time.

All of them start by first picking each arm of each bandit for a fixed number of times, defined by parameter B, and then proceed to a loop of actively selecting the next subpopulation and treatment pair (i, j) until the budget (e.g. N) runs out. Full algorithmic details of all three methods are provided in Figure 1.

To illustrate the different behaviors of the three active learning policies, we consider the five synthetic data sources described in table IV. The response for each subpopulation i under treatment j is modeled using a normal distribution $\mathcal{N}(\mu, \sigma)$. The means and the variances of these normal distributions are detailed in the table. The maximum budget N was set to be 200 in all experiments, each algorithm was repeated 100 times with different random initializations of the data sources, so the results reported below are averaged over 100 runs.

Choose each treatment for each subpopulation B times in the first $B \times C \times K$ trials
Set $T_{ij} = 1$ and $n = BCK + 1$
while $n < N$ **do**
  Compute the standard error estimate $\hat{\sigma}_{ij}^{(n)}$ for treatment (i, j) at time point n
  Option 1:AREOA.
  if $\sum_i \left(\sum_j \hat{\sigma}_{ij}^{(n)} \right)^2 \neq 0$ **then**
    Let $\tau_{ij} = \frac{\hat{\sigma}_{ij}^{(n)} \sum_j \hat{\sigma}_{ij}^{(n)}}{T_{ij} \sum_j T_{ij} Z}$ where Z is chosen such that $\sum_i \sum_j \tau_{ij} = 1$
  else
    let $\tau_{ij} = \frac{1}{CK}$
  end if
  Pick the next subpopulation and treatment pair (i, j) with probability $(1 - \epsilon)\tau_{ij} + \epsilon * \frac{1}{CK}$
  Option 2:AARandom.
  Randomly pick a subpopulation i according to its its composition in the general population, and then pick a treatment j uniformly at random.
  Option 3:GAFS-MAX.
  Assume some arbitrary but fixed ordering for the set of all (i, j) pairs.
  Let $U_n = \{(k, l) : T_{k,l}^n < \sqrt{n} + 1\}$
  Let

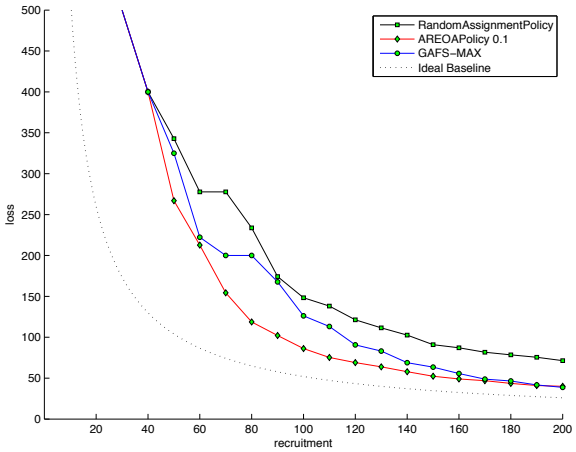
$$I_{n+1} = \begin{cases} \min U_n & \text{if } U_n \neq \varnothing \\ \arg\max \frac{(\hat{\sigma}_{ij}^{(n)})^2}{T_{ij}^n} & \text{otherwise} \end{cases}$$

  Choose option I_{n+1} and update T_{ij}^{n+1} accordingly
  $n = n + 1$
end while

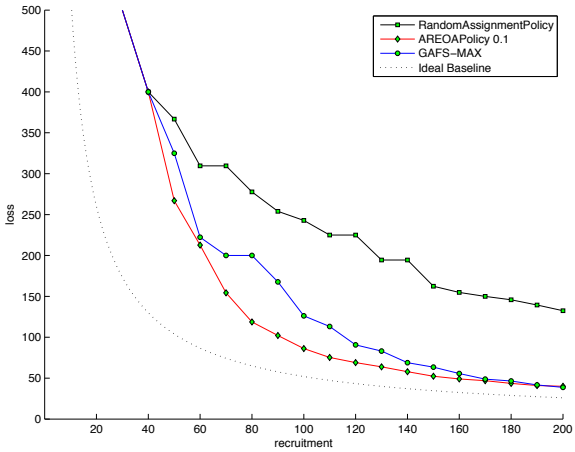
Fig. 1. Algorithm Framework

We have chosen these settings to demonstrate the behaviors of the various policies under different scenarios. For dataset DS1, subpopulation c_1 has a large treatment effect variance relative to the other subpopulations. For dataset DS2, the subpopulation distribution is non-uniform in that subpopulation c_1 is rare compared to the other subpopulations. For dataset DS3, we consider a scenario where there are 8 subpopulations with moderate to large variances across subpopulations. For dataset DS4, we consider a scenario where all subpopulations have the same variance in treatment effect, but within each subpopulation there is a large difference in variance between the estimated mean responses to each treatment. For dataset DS-CBASP, the mean and variances are taken from a real clinical trial for chronic depression [32]. In this case, the treatment variances of each subpopulations are very similar.

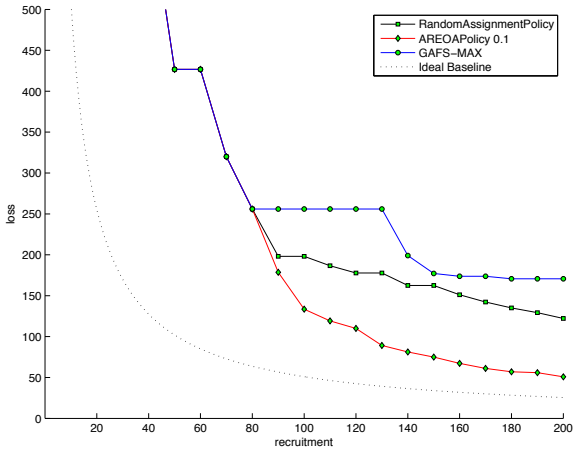
For each dataset, we first plotted the loss of a policy as the number of recruited subjects increases (Figure 2). In this set of plots, we fixed the number of initial recruitments per treatment to $B = 5$. In the figures, the dotted lines at the bottom correspond to the optimal loss of the oracle allocation



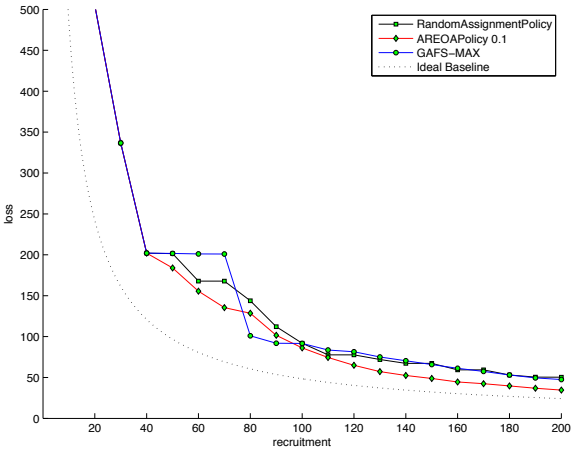
(a) DS1



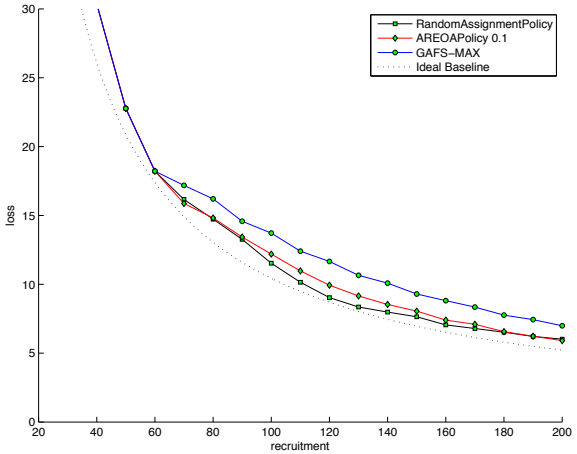
(b) DS2



(c) DS3

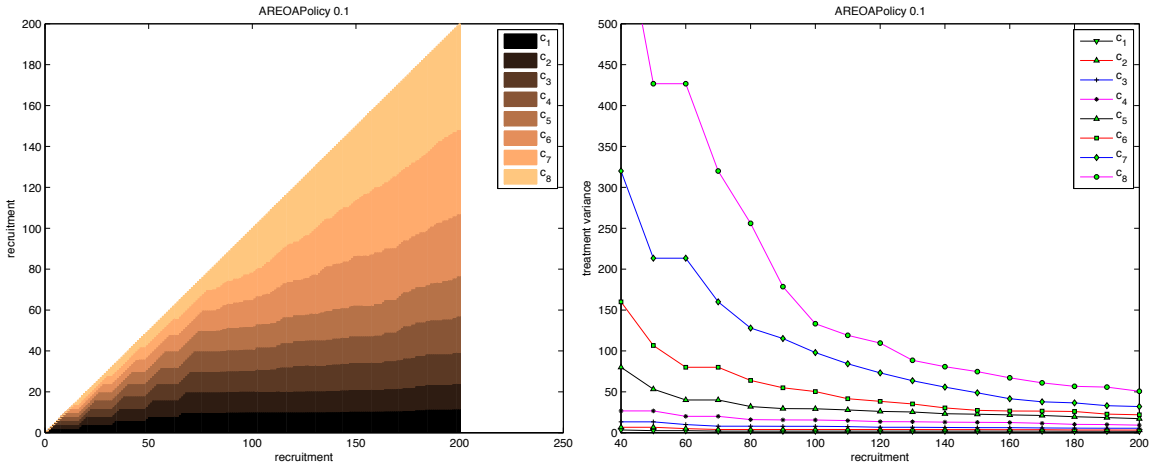


(d) DS4



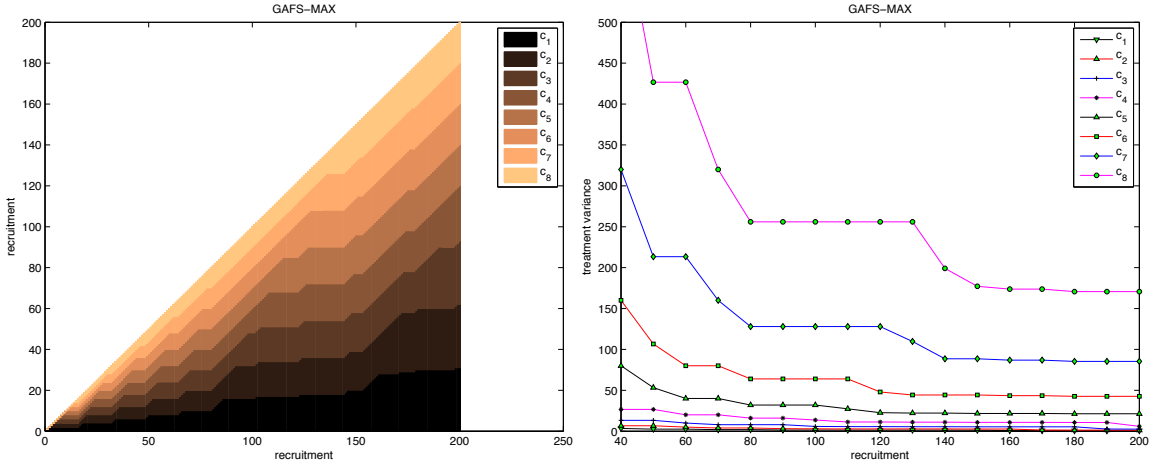
(e) DS-CBASP

Fig. 2. Simulation results

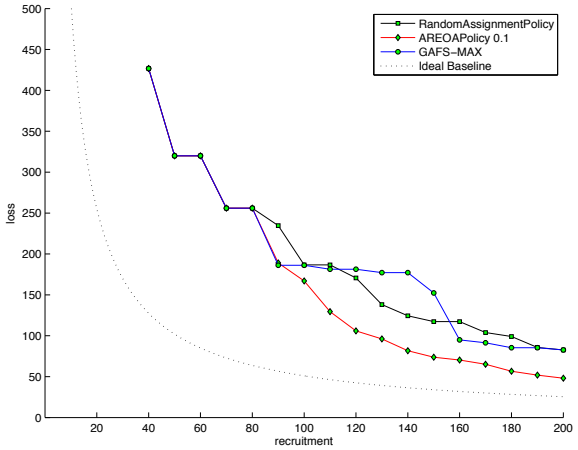


(a) subpopulation allocations (b) subpopulation variances

Fig. 3. Algorithm AREOA on DS3



(a) subpopulation allocations (b) subpopulation variances



(c) DS3 with subpopulation ordering reversed

Fig. 4. Algorithm GAFS-MAX on DS3, subfigure (a), (b) are the subpopulation allocations and variances, subfigure (c) uses the same dataset as DS3, with ordering of the subpopulations reversed

policy (i.e. $\frac{\sum_i (\sum_j \sigma_{ij})^2}{N}$ with N varying from 1 to 200).

As shown in Figure 2, for DS1, both AREOA and GAFS-MAX converge, but AREOA is able to utilize the data more efficiently. For DS2, due to the nonuniform distribution of the subpopulation distribution, random assignment performs significantly worse than algorithms that make use of the estimated treatment variances. For DS3, the performance of GAFS-MAX is worse than random, which is suspicious; we discuss this further below. For DS4, we notice that even in cases where there are no significant differences in treatment variance across subpopulations, AREOA still performs quite well; note that as the budget is spent, AREOA approaches the baseline (optimal oracle allocation) slightly more quickly than the other two active learning policies. We included the dataset DS-CBASP, as there are no big differences in estimated variances across subpopulations, and across treatments, for which the random allocation policy AARandom is perhaps the right choice; we observe that AREOA converges slowly at the beginning but approaches the baseline as quickly as AARandom as the budget runs out. It is reassuring to see that AEROA performs in a reasonable manner, even for cases which it wasn't specifically designed to handle.

Focusing further on the performance of AREOA with dataset DS3, we see in Figure 3 how the allocations and variances of each subpopulation evolve as the budget is spent. We can clearly see that the recruitment rates across subpopulations correlate with the subpopulation variance, and thus also the amount of exploration needed. For the same dataset, and method GAFS-MAX, we plotted the allocations and variances of each subpopulation in Figure 4. As is shown in subfigure (a), (b), there are many plateauing regions for certain high variance subpopulations, which are severely under-explored, due to the fact that resources (allocations) have been devoted to other subpopulations. We think there are a couple of causes behind this behavior. First, when there are many subpopulations, GAFS-MAX still will spend time revisiting subpopulations whose treatment variance is well estimated. This can be problematic, especially if the budget is small and there exists some subpopulations that have much larger variances and thus still need more exploration. Another problem with using GAFS-MAX in our setting is the fixed ordering for picking the next rarely visited arm, which may also delay a high variance arm being revisited in the short term. To confirm that the algorithmic behavior of GAFS-MAX is dependent on the ordering of subpopulations, we reversed their ordering in DS3, and we see now in subfigure (c) of Figure 4 that the performance of GAFS-MAX has changed significantly.

One of the key parameters of our framework, common to all three methods considered, is the initial number of allocations for each treatment per subpopulation B. We varied B from 1 to 10 and we confirmed that in general, most results presented above are robust to the choice of B, however in some cases using an initial sample that is too small may lead to problems. As shown in Figure 5, when using $B = 2$ for the DS4 dataset, the AREOA approach tends to start converging too early and converge slowly (compare this to the Figure 2(d), where $B = 5$). It is standard practice in any adaptive trial to reserve a significant portion (up to 50%) of the trial population to be fully randomized in order to avoid such problems.

V. Discussion

We presented an active learning problem for use in clinical trials that aims to construct an informative, yet well-balanced, ITR. In this section, we discuss open issues and possible future work.

1. We demonstrated the potential of an active learning policy (AREOA) in comparison with a completely randomized patient–treatment allocation policy, and with the GAFS-MAX active learning policy from the bandit literature. Our AREOA strategy is shown to be consistent empirically, but a formal proof that it will asymptotically converge to the optimal is required. We could also consider more advanced algorithms, for example algorithms that take into account the knowledge of total budget. Another possible improvement is to develop a strategy that further considers the heterogeneity within each subpopulation by considering bandit with covariates. This would require some prior knowledge, or parametric modeling, to capture how these covariates are relevant to the treatment response.

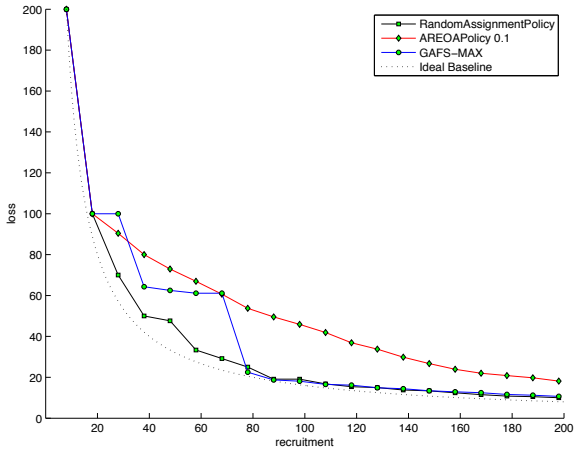
2. When there are more than two arms ($K >= 3$), our original formulation of minimizing maximal variances across subpopulations require generalization. The problem is that our approach relies on pairwise comparisons, which do not translate readily to multi-arm cases. In cases where there is a control arm (e.g. a well known standard treatment) for each bandit, whose mean and variance is given, then perhaps we can proceed by bounding the maximal uncertainty of the treatment effects compared to this arm. This has not been explored in detail. Another possibility, which is closer to what we want, is to minimizing the maximal variances of the (true) top 2 treatments, albeit their indentity needs to be discovered by the algorithm as well with high probability.

3. If the variance in one of the subpopulations is much larger than the variance in the other subpopulations, an active learning policy such as AREOA would devote the most resources to this high variance subpopulation. This behavior is expected, as it is due to the goal of minimizing the maximal treatment variances. For cases where the discrepancy in treatment variance between subpopulations is extremely high, this could turn out to be problematic; ideally, we may prefer a stopping rule for excessively high variance subpopulations. This stopping rule would tell us to "give up" on minimizing the variance of the treatment effect for these subpopulations. Note, however, since the variance of the estimated mean is decreasing at rate $1/n$, that the subpopulation variance would have to be grossly bigger than the other subpopulation variances. Also, if we know from prior knowledge that certain subpopulations have high variance treatment effect, another possible solution is to reweight the importance of the corresponding terms in objective function by linear or logarithmic scaling. For instance, (1) could become $\max_i \sum_j \alpha_{ij} \frac{\sigma_{ij}^2}{T_{ij}}$, with

α_{ij} being the scaling factors. This variation can also be used to accommodate nonuniform cost models where the costs of treating certain subpopulations are significantly different from treating other subpopulations.

4. The above remark motivates a different choice of reward function. For each subpopulation, we want to know whether a treatment is significantly better than the others. Once we can answer this question (accept or reject the null hypothesis) with high probablity, the trial should stop for that subpopulation. In other words, a more important reward pertains to the type 1 and type 2 error rates, instead of just minimizing the variance. In this regard, we think a bayesian formulation that incrementally updates our belief regarding these questions would be natural.

5. Finally, the problem that we discussed is a one-stage RL problem, which doesn't involve state changes. In the literature, "dynamic treatment regimes" or "adaptive treatment strategies" [28]–[31] naturally generalize the idea of ITRs to multiple stages by constructing a sequence of decision rules, one for each disease stage. How to extend the ideas of this paper to such time-varying settings is also an open question.



(a) DS4

Fig. 5. Simulation results for DS4, when $B = 2$

REFERENCES

[1] S. Thrun, "Efficient exploration in reinforcement learning (Technical Report CS-92-102)." Carnegie Mellon University, Pittsburgh, PA, 1992.

[2] M. Strens, "A Bayesian framework for reinforcement learning," in machine learning-international workshop then conference, pp. 943-950, 2000.

[3] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," Machine Learning, vol. 49, no. 2, pp. 209-232, 2002.

[4] R. I. Brafman and M. Tennenholtz, "R-max-a general polynomial time algorithm for near-optimal reinforcement learning," The Journal of Machine Learning Research, vol. 3, pp. 213-231, 2003.

[5] A. Epshteyn, A. Vogel, and G. DeJong, "Active reinforcement learning," in Proceedings of the 25th international conference on Machine learning, pp. 296-303, 2008.

[6] X. Zhou, S. Liu, E. S. Kim, R. S. Herbst, and J. J. Lee, "Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine," Clinical Trials, vol. 5, no. 3, pp. 181-193, 2008.

[7] S. Biswas, D. D. Liu, J. J. Lee, and D. A. Berry, "Bayesian clinical trials at the University of Texas M. D. Anderson Cancer Center," Clinical Trials, vol. 6, no. 3, pp. 205-216, Jun. 2009.

[8] J. J. Lee, Xuemin Gu, and Suyu Liu, "Bayesian adaptive randomization designs for targeted agent development," Clinical Trials, vol. 7, no. 5, pp. 584-596, Oct. 2010.

[9] A. Antos, V. Grover, and C. Szepesvri, "Active learning in Multi-armed Bandits," in Proceedings of the 19th international conference on Algorithmic Learning Theory, pp. 287-302, 2008.

[10] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," Machine Learning, vol. 15, no. 2, pp. 201-221, 1994.

[11] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," Statistical Science, vol. 10, no. 3, pp. 273-304, Aug. 1995.

[12] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the Query by Committee algorithm," Mach. Learn., vol. 28, no. 2, pp. 133-168, 1997.

[13] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in Proceedings of the ninth ACM international conference on Multimedia, pp. 107-118, 2001.

[14] G. Xiao, F. Southey, R. C. Holte, and D. F. Wilkinson, "Software testing by active learning for commercial games," in AAAI, pp. 898-903, 2005.

[15] S. Tong and D. Koller, "Active learning for parameter estimation in bayesian networks," in In NIPS, pp. 647-653, 2001.

[16] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," J. Mach. Learn. Res., vol. 5, pp. 255-291, 2004.

[17] O. Madani, D. J. Lizotte, and R. Greiner, "The budgeted multi-armed bandit problem," in Learning theory: 17th annual Conference on Learning Theory, COLT 2004, vol. 3120, pp. 643–645, 2004.

[18] A. Kapoor and R. Greiner, "Reinforcement learning for active model selection," in Proceedings of the 1st international workshop on Utility-based data mining, pp. 17-23, 2005.

[19] H. Raghavan, O. Madani, and R. Jones, "Active learning with feedback on features and instances," J. Mach. Learn. Res., vol. 7, pp. 1655-1686, 2006.

[20] K. Deng, C. Bourke, S. Scott, J. Sunderman, and Y. Zheng, "Bandit-based algorithms for budgeted learning," in Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 463-468, 2008.

[21] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in Proceedings of the 20th international conference on Algorithmic learning theory, pp. 23-37, 2009.

[22] B. Settles, "Active learning literature survey." University of Wisconsin-Madison, 2009.

[23] V. Fedorov, "Optimal experimental design," Wiley Interdisciplinary Reviews: Computational Statistics,2010.

[24] A. I. Schein, "Active learning for logistic regression," University of Pennsylvania, 2005.

[25] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in Proceedings of the 23rd international conference on machine learning, pp. 1081-1088, 2006.

[26] L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference." Springer, 2003.

[27] S. A. Murphy, "Optimal dynamic treatment regimes," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 65, no. 2, pp. 331-355, 2003.

[28] R. Dawson and P. W. Lavori, "Placebo-free designs for evaluating new mental health treatments: the use of adaptive treatment strategies," Statistics in medicine, vol. 23, no. 21, pp. 3249-3262, 2004.

[29] S. A. Murphy, K. G. Lynch, D. Oslin, J. R. McKay, and T. TenHave, "Developing adaptive treatment strategies in substance abuse research," Drug and Alcohol Dependence, vol. 88, pp. S24, 2007.

[30] A. Guez, R. D. Vincent, M. Avoli, and J. Pineau, "Adaptive treatment of epilepsy via batch-mode reinforcement learning," in Proceedings of the Twentieth Innovative Applications of Artificial Intelligence Conference, pp. 1671-1678, 2008.

[31] J. K. Wathen, P. F. Thall, J. D. Cook, and E. H. Estey, "Accounting for patient heterogeneity in phase II clinical trials," Statistics in Medicine, vol. 27, no. 15, pp. 2802-2815, Jul. 2008.

[32] M. B. Keller et al., "A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression," New England Journal of Medicine, vol. 342, no. 20, p. 1462, 2000.