Efficient Sequences of Trials

Edith Cohen* Amos Fiat† Haim Kaplan†

An incompetent attorney can delay a trial for months or years. A competent attorney can delay one even longer.
— Evelle J. Younger (1918–1989), *LA Times, March 3rd, 1971.*

Abstract

We introduce a problem called *sequential trial optimization*, a generalization of the well studied set cover problem with a new objective function. We give a simple algorithm that achieves a constant factor approximation to this problem. Sequential trial optimization naturally arises in heterogeneous search environments such as peer to peer networks.

1 Introduction

We introduce a new problem called *sequential trial optimization*. This problem arises naturally in heterogeneous search environments, in medical diagnostic applications, and seems to be a very natural problem that may find additional, yet unknown, applications.

In this problem we are given a probability distribution over possible *queries* and a set of *trials*. For every trial and every query we are also given the success probability of the particular trial on the particular query. The optimization problem we consider is to construct a sequence of trials so that at least a f-fraction of the queries generated by the query distribution is answered by some trial in our sequence and that the expected number of trials performed is minimized.

We use the terminology "query" because of search applications where these queries may be the names of music titles, or some boolean predicate that evaluates to true when applied to the content (*e.g.*, author='Joyce' and title='Finnegans Wake' or title='Dubliners'). However, a query may also be a (physical) blood sample which is to be checked for some pathogen. In fact, a query may be some completely arbitrary problem.

For simplicity, we assume that the number of different queries is finite, but there is nothing inherent about the finiteness of the number of queries and our results easily generalize to the case of an infinite set of queries.

We use the terminology "trial" to capture the notion that we are applying a probabilistic process to the query that may (or may not) return the required answer (or a positive result in the medical diagnostic scenario). In the context of search applications, imagine that every trial is some different (possibly randomized) search heuristic. Alternately, trials may be medical diagnostic tools such as genotypic assays. If the queries are problem instances to be solved then the different trials may be different (possibly randomized) heuristics. Obviously, it is a reasonable goal to try to reduce the number of trials. More generally, it may be desired to associate different costs with different trials and to reduce the overall cost.

One very interesting aspect of this problem is that the probability that a certain trial will succeed on a given query, or on a query chosen from the query distribution Q, depends on the sequence of previous trials performed on this query. Thus, it is seemingly important to know the posterior probabilities for queries conditioned on sequences of previous trials.

Our main result is a polynomial time constant factor approximation to the sequential trial optimization problem. These results appear as Theorems 5.1 and 6.1. We also consider various generalizations such as weighted trials (minimizing the cost), and — more importantly — the scenario where the inputs are unknown.

In our applications, it is unreasonable to assume that much is known about the probability distribution over the queries, the success probabilities of any given trial for a specific query, and certainly not on the success probabilities of trials conditioned on the previous unsuccessful trials.

Nonetheless, we show that even without knowing any of these values, one can compute a constant factor approximation to the optimal (*NP*-complete) solution (which has all these inputs available).

1.1 Sequential Trials We now give a formal definition of the sequential trial optimization problem. Let $Q = (q_1, q_2, \dots, q_n)$ be a probability distribution over queries $1, 2, \dots, n$, where the probability that a query chosen at random from Q equals i is q_i. Let $T = \{t_1, t_2, \dots, t_m\}$ be a set of *trials* where $t(q)$, $t \in T$,

*AT&T research labs, 180 park ave. Florham park, NJ, USA. Email:`edith@research.att.com`.

†School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: `{fiat,haimk}@cs.tau.ac.il`.

$$M = \begin{pmatrix} \text{Prob}(t_1(1)=1) & \text{Prob}(t_1(2)=1) & \dots & \text{Prob}(t_1(n)=1) \\ \text{Prob}(t_2(1)=1) & \text{Prob}(t_2(2)=1) & \dots & \text{Prob}(t_2(n)=1) \\ \vdots & \vdots & & \vdots \\ \text{Prob}(t_m(1)=1) & \text{Prob}(t_m(2)=1) & \dots & \text{Prob}(t_m(n)=1) \end{pmatrix}.$$

Figure 1: The matrix M; each entry M_{ij} is the probability that trial i answers query j.

$q \in \{1, 2, \dots, n\}$, is a random variable getting values 0 and 1. We say that trial $t \in T$ *answers* query q, $1 \leq q \leq n$, if $t(q) = 1$. Thus $\text{Prob}(t(q) = 1)$ is the probability that trial t answers query q. Let M be the $m \times n$ matrix where $M_{ij} = \text{Prob}(t_i(j) = 1)$, an example of such a matrix appears in Figure 2.

It follows from our definitions that the success probability of trial t_j on a query drawn from Q is the inner product $(M_j, Q) = \sum_{k=1}^{m} M_{jk} q_k$, where M_j denoted the j'th row of the matrix M. Given a sequence of trials $\sigma = \langle \sigma_1, \sigma_2, \dots \rangle$ where $\sigma_i \in T$ and a query distribution Q, an experiment is a random choice of a query q according to the distribution Q and an application of the trials in σ to q. The outcome of an experiment for Q and σ is a pair (q, j) where q was chosen from Q and $1 \leq j \leq |\sigma|$ is derived as follows. Let $S \subset \{1, \dots, |\sigma|\}$ be a random variable, $S = \{i | \sigma_i(q) = 1\}$, we define $j = \min S$ if $S \neq \emptyset$ and $j = |\sigma|$ if $S = \emptyset$. Let $D(\sigma)$ be the distribution over outcomes defined by Q and σ.

If $S \neq \emptyset$ was say that the experiment was successful, if $S = \emptyset$ then the experiment is unsuccessful. We denote by $P(\sigma)$ the probability that an experiment is successful. This is the probability that sequentially trying $\sigma_1, \sigma_2, \dots$ to answer the (randomly chosen) query will succeed. We refer to the i^{th} of these trials as probing σ_i for Q.

We denote the prefix of length i of σ by $\text{pref}(i, \sigma)$. I.e. $\text{pref}(i, \sigma) = \langle \sigma_1, \sigma_2, \dots, \sigma_i \rangle$. It follows from our definitions that

$$\begin{aligned} P(\sigma) &= \sum_{i=1}^{|\sigma|} P(\text{pref}(i,\sigma)) - P(\text{pref}(i-1,\sigma)) \\ &= \sum_{i=1}^{|\sigma|} \big(M_{\sigma_i}\ , q * (\bar{1} - M_{\sigma_1}) * (\bar{1} - M_{\sigma_2}) * \cdots \\ &\qquad * (\bar{1} - M_{\sigma_{i-1}})\big) \end{aligned} \tag{1.1}$$

Where M_{σ_i} denotes the row in M which corresponds to trial σ_i, $\bar{1}$ denotes the all-one vector, and $v*u$ denotes the Hadamard product (component wise multiplication) of v and u. (Note that the sum above may be infinite.)

We will be interested in sequences σ such that $P(\sigma) \geq f$. Let $T^{f,\sigma}$ be the minimum ℓ such that $P(\text{pref}(\ell, \sigma)) \geq f$, define $T^{f,\sigma} = \infty$ if σ is infinite and $P(\sigma) = f$. We denote by $E(f, \sigma)$ the expected number of trials when we try to answer a query by first applying σ_1, if it fails then applying σ_2, up to $\sigma_{T^{f,\sigma}}$. More formally,

$$\begin{aligned} E(f,\sigma) &= \sum_{(q,i) \in D(\sigma), i < T^{f,\sigma}} i \cdot \text{Prob}((q,i)) + \\ &\qquad T^{f,\sigma} \cdot \sum_{(q,i) \in D(\sigma), i \geq T^{f,\sigma}} \text{Prob}((q,i)). \end{aligned}$$

($\text{Prob}((i,j))$ is the probability over $D(\sigma)$ of the outcome (i,j)). Note that an application of $\sigma_1, \dots, \sigma_{T^{f,\sigma}}$ on a query can be successful or unsuccessful, where the probability that it is successful is $\geq f$. We define

$$\begin{aligned} &E^+(f,\sigma) = \\ &= \sum_{i=1}^{T^{f,\sigma}-1} i \cdot (P(\text{pref}(i,\sigma)) - P(\text{pref}(i-1,\sigma))) + \\ &\quad T^{f,\sigma} \left(f - P(\text{pref}(T^{f,\sigma} - 1, \sigma))\right) \ . \end{aligned}$$

It now follows that $E(f,\sigma) = E^+(f,\sigma) + (1-f) \cdot T^{f,\sigma}$. What this means intuitively is that $E^+(f,\sigma)$ is the contribution to $E(f,\sigma)$ by the shortest successful outcomes (smallest number of trials) with total probability mass f.

Note that $E^+(f,\sigma)/f$ is the expected number of trials over outcomes (q,i) satisfying one of the following two conditions

- $i < T^{f,\sigma}$.
- $i = T^{f,\sigma}$ and $R = 1$ where R is an independent random $\{0,1\}$ variable with $\text{Prob}(R = 1) = (f - P(\text{pref}(T^{f,\sigma} - 1), \sigma))/(P(\text{pref}(T^{f,\sigma}, \sigma) - P(\text{pref}(T^{f,\sigma} - 1, \sigma))$.

Thus, $E^+(f,\sigma)/f$ is the expected number of trials over outcomes (q,i) where these outcomes are chosen according to a new distribution, $D^+(f,\sigma)$, that can be derived from the original distribution over outcomes, $D(\sigma)$, via these two conditions.

Given f, Q, and M we define the optimal sequence of trials, σ^{opt}, to be the sequence such that $E(f, \sigma^{opt}) \leq E(f,\sigma)$ for every sequence σ such that $P(\sigma) \geq f$.

We show that finding an optimal sequence is NP-hard. Our main result is a proof that the straightforward GREEDY algorithm produces a sequence whose expectation is within a constant factor of optimal. We

$$M = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad Q = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

For $\sigma = t_1, t_1, t_1, \ldots$, $P(\sigma) = \sum_{i=1}^{\infty}(2/3)^{i-1}/3 = 1$,$E(1,\sigma) = E^+(1,\sigma) = \sum_{i=1}^{\infty} i(2/3)^{i-1}/3 = 3$.
For $\sigma = t_2, t_3, t_4$, $P(\sigma) = 1$, $E(1,\sigma) = (1+2+3)/3 = 2$.

Figure 2: A Very Simple Example of the Sequential Trial Optimization Problem.

remark that Greedy for the sequential trial optimization problem is rather more complex than Greedy for set cover because the posterior probabilities change as previous trials fail.

In section 7 we indicate how to generalize our approach via sampling to work even when Q and M exist but are not given to us as an input.

1.2 Sequential Cover The input to the classical set cover problem is a set of elements, a collection of subsets over the elements, and the objective is to find the smallest number of subsets that cover the elements. This natural problem was extensively studied, and has numerous applications.

This problem is known to be NP-complete; but the simple Greedy algorithm, which iteratively selects the set which covers the largest number of yet uncovered elements obtains a logarithmic approximation ratio [1, 9]. In a seminal work, Feige showed that a better approximation ratio is not possible [6].

An important special case of the sequential trial optimization problem, closely related to set cover, is a problem we call *sequential cover*. This is a restriction of the trials to be deterministic, *i.e.*, restrict the probabilities $M_{ij} = \text{Prob}(t_i(j) = 1) \in \{0, 1\}$. This means that there is never any reason to choose a trial more than once. The optimal trial sequence for the more general formulation of trial sequence optimization may contain multiple occurrences of each trial and can also be infinite.

Sequential Cover Problem: Given a probability distribution $Q = (q_1, \ldots, q_n)$ over the elements of the base set $\{1, 2, \ldots, n\}$, a value $0 \le f \le 1$, and a collection of subsets $\{S_1, S_2, \ldots, S_m\}$, find a value $0 \le m' \le m$ and a sequence $S_{\pi(1)}, S_{\pi(2)}, \ldots, S_{\pi(m')}$ of m' sets that has the following properties:
Coverage:

$$\sum_{\ell \in \cup_{i=1}^{m'} S_{\pi(i)}} q_\ell \ge f$$

i.e., the probability that an element drawn according to Q is covered by at least one subset in $S_{\pi(1)}, \ldots, S_{\pi(m')}$ is at least f.
Efficiency: Among all sequences that fill the coverage requirement, find one that minimizes the expected size of the search, that is,

$$\sum_{\ell \in \cup_{i=1}^{m'} S_{\pi(i)}} q_\ell \min\{k \mid \ell \in S_{\pi(k)}\} + m' \sum_{\ell \notin \cup_{i=1}^{m'} S_{\pi(i)}} q_\ell$$

is minimized.

Generally, the performance here depends both on what subsets are used and their permutation. In contrast, the objective of the classic set cover problem is to minimize the total number of selected subsets m'.

It is not hard to see that like its classic counterpart, the sequential cover problem is NP-Complete. We show, however, that an appropriate version of the Greedy algorithm yields a constant approximation ratio to the more general sequential trial optimization problem (and thus to the sequential cover problem as well). This provides an interesting contrast to the known logarithmic lower bound [6] on the standard set cover problem.

1.3 Applications Both sequential cover and the more general sequential trial optimization problems are appropriate to model problems when the solution constitutes a search, test, or trial sequence. The difference between the two is that in sequential cover all trials give deterministic results whereas in sequential trial optimization trials can be randomized. In both cases, we seek to find the trial sequence that provides an answer with minimum average number of trials.

One application for sequential cover is that of constructing optimal decision sequences, which are decision trees when all available tests provide a 1-sided decision, and thus, the output trees have the form of a long chain with leaves.

An interesting application of sequential trial optimization arises in peer to peer networks. Peer-to-peer protocols such as Gnutella [5] and FastTrack [11, 8, 7]) deploy "blind" search where the set of peers to which the search is propagated to is unrelated to the query. The performance of blind search can be modeled by probing hosts selected at random (say uniformly or in some weighted way to reflect skew) [4, 10]. The likelihood that a probe succeeds is captured by the fraction of the hosts that are able to answer the query.

One method to improve blind search is to associate hosts with some meaningful *search-rules* [2, 3]. Search rules can correspond to common property of the associated hosts (such as possession of particular item(s), geographic location, or a common interest area), and each host can belong to several search-rules. The hope is that search rules reflect more focused interest areas, so that searching within rules a host is associated with has increased likelihood of success. There is a requirement that hosts within each search-rule are connected in a way that allows to restrict the propagation of queries within a rule.

Optimizing the order in which the different search rules are used can be solved using our sequential trial optimization algorithm. Every search rule is a trial, and search queries are (naturally) queries. In [2] we show a truly dramatic improvement in search times, especially for rare items, by several orders of magnitude, through the use of an optimized order of search rules.

2 Preliminaries

We initially limit our discussion to the sequential trial optimization problem when $f = 1$. Note that in this case $E^+(f,\sigma) = E(f,\sigma)$ for every σ. To simplify our notation we will use $E(\sigma) = E(1,\sigma)$. In Section 3 we shall introduce the GREEDY algorithm and in Theorem 5.1 of Section 5 we prove that $E(\sigma^g) \leq O(1)E(\sigma^{opt})$. We can easily generalize this theorem to show that conditioned on successful applications, the greedy sequence approximates the optimal one to within a constant factor for any f, i.e. $E^+(f,\sigma^g) \leq O(1)E^+(f,\sigma^{opt})$. In Section 6 we indicate that this implies that also with respect to both successful and unsuccessful applications, greedy is a good approximation if we are willing to compromise a little the success probability, i.e.

$$E(f - \epsilon\min\{f,(1-f)\},\sigma^g) \leq O(1)E(f,\sigma^{opt}) \ .$$

Finally in Section 7 we indicate what to do in a scenario where q, M, and f are not known apriori.

Let $\sigma = \langle\sigma_1,\sigma_2,\ldots\rangle$ be a sequence of trials. We will use the following alternative definition of $E(\sigma)$

$$\begin{aligned} E(\sigma) &= 1 + (1-P(\sigma_1)) + (1-P(\text{pref}(2,\sigma))) + \cdots \\ &\quad + (1-P(\text{pref}(i,\sigma))) + \cdots \end{aligned} \tag{2.2}$$

We close this introductory section with the following easy hardness result.

LEMMA 2.1. *The sequential trial optimization problem is NP-complete.*

Proof. The *NP*-completeness of the sequential trial optimization problem follows from that of the more restricted *sequential cover* formulation. To prove that *sequential cover* is NP-complete we reduce the problem of exact cover by sets of size three (X3C) to it. Let S, such that $|S| = 3q$ and $S_1,\ldots,S_m$, where $|S_i| = 3$ be an input to X3S. It is not hard to see that there is a subset of q sets out of $S_1,\ldots,S_m$ that cover S exactly iff there is a sequence of sets out of $S_1,\ldots,S_m$ that cover S with expected search length $\sum_{i=1}^{q} 3i$ (for $f = 1$). □

3 The Greedy Search Sequence

We define a simple greedy algorithm that computes a sequence of trials σ^g tailored to be effective for a query drawn according to Q. We refer to this algorithm as GREEDY.

The first trial which GREEDY chooses, denoted by σ_1^g, is the trial that maximizes the success probability that we can achieve by doing a single experiment. In other words σ_1^g is the trial j which maximizes (M_j, q).

Assuming that the application of the first trial of GREEDY turns out to be unsuccessful we can update the query distribution Q to take into account the unsuccessful probe. For a trial j and a query distribution Q, we define the *posteriori* query distribution $u_j(Q)$ to be $< q_1', q_2', \ldots, q_n' >$, where q_k' is the probability that the query is for item k given that trial j failed.

$$\begin{aligned} q_k' &= \text{Prob}\,(\text{query for item k} \mid \text{trial } j \text{ unsuccessful}) \\ &= \frac{\text{Prob}\begin{pmatrix} (\text{query is for item k}) \cap \\ (\text{trial } j \text{ is unsuccessful}) \end{pmatrix}}{\text{Prob}\,(\text{trial } j \text{ is unsuccessful})} \\ &= \frac{q_k(1-M_{jk})}{1-(M_j,q)}. \end{aligned}$$

The second trial which GREEDY chooses, denoted by σ_2^g, is the trial that maximizes the success probability with respect to the query vector $u_{\sigma_1^g}(Q)$. In other words GREEDY picks as its second trial that trial j which maximizes $(M_j, u_{\sigma_1^g}(Q))$.

To define the ith trial picked by GREEDY we first extend the definition of the function u to sequences of trials as follows. Given a sequence of trials $\sigma = \sigma_1,\sigma_2,\ldots,\sigma_{|\sigma|}$, we define the posteriori query distribution $u_\sigma(Q)$ after $\sigma_1,\sigma_2,\ldots,\sigma_{|\sigma|}$ have failed, to be

$$u_\sigma(q) = u_{\sigma_{|\sigma|}}(u_{\sigma_{|\sigma|-1}}(\cdots u_{\sigma_2}(u_{\sigma_1}(Q))\cdots)).$$

The following lemma shows that the posterior query distribution is the same whatever is the order in which we fail on $\sigma_1, \sigma_2, \ldots, \sigma_{|\sigma|}$. We omit the proof which is easy. We also use the notation u_ε to denote the identity function where ε is the empty sequence.

LEMMA 3.1. *For any sequence $\sigma = \sigma_1, \sigma_2, \ldots, \sigma_{|\sigma|}$, and for any permutation τ of $1, \ldots, |\sigma|$*

$$u_\sigma(Q) = u_{\tau(\sigma)}(Q),$$

where $\tau(\sigma) = \sigma_{\tau(1)}, \sigma_{\tau(2)}, \ldots, \sigma_{\tau(|\sigma|)}$.

Given a trial sequence $\sigma = \sigma_1, \sigma_2, \ldots$, we denote by $\mathrm{pref}(i, \sigma)$ the prefix of length i of σ, for $i = 0$, we define $\mathrm{pref}(i, \sigma) = \varepsilon$, the empty sequence . Now we are ready to define the ith trial picked by GREEDY.

The ith probe which GREEDY picks, denoted by σ_i^g, is the trial that maximizes the success probability with respect to the query vector $u_{\mathrm{pref}_{(i-1,\sigma^g)}}(Q)$. In other words, GREEDY picks as its ith probe that trial j which maximizes $(M_j, u_{\mathrm{pref}_{(i-1,\sigma^g)}}(Q))$.

4 GREEDY has Worst-Case Approximation Ratio of at Least 2

In this section we show that the approximation ratio of GREEDY cannot be smaller than 2. This is stated in the following Theorem.

THEOREM 4.1. *For any ϵ, we can construct an instance of* sequential cover *where $E(\sigma^g) \geq (2 - \epsilon) E(\sigma^{\mathrm{opt}})$.*

Theorem 4.1 follows by selecting a "bad" instance that is guaranteed to exist by Corollary 4.1 of Lemma 4.1, where k is sufficiently large.

LEMMA 4.1. *For any integer k, the approximation ratio of* GREEDY *on instances of* sequential trial optimization *with $|\sigma^{opt}| \leq k$ is at least $2k/(k+1)$. (This is also true for instances with at most k queries).*

Proof. For any ϵ' and k we construct an instance of *sequential trial optimization* where there are k queries, $|\sigma^{\mathrm{opt}}| = k$, $E(\sigma^{\mathrm{opt}}) = (k+1)/2$ and $E(\sigma^g) \geq k - \epsilon'$.

Let $Q = (1/k)\bar{1}$ be the uniform query distribution. Our instance has $m = k + 1$ trials. For $1 \leq i \leq k$, the i^{th} trial succeeds with probability 1 on query i and fails with probability 1 on all other queries. The $(k+1)^{st}$ trial succeeds with probability $1/k + \epsilon''$ on every query.

The optimal strategy constitutes of some (arbitrary) permutation of the first k trials. There is equal likelihood of $1/k$ that the search succeeds in each step, and thus the expected number of trials is $\sum_{i=1}^{k} i/k = (1+k)/2$. The GREEDY strategy consists of trial $k+1$ repeated infinitely many times. It is not hard to see that the posterior query distribution remains the uniform vector after each unsuccessful trial, and thus GREEDY repeatedly selects the $(k+1)^{st}$ trial. The likelihood of success in the ith trial (given failures in all previous trials) is $1/k + \epsilon''$. Thus, the expected number of trials of GREEDY is $\sum_{i \geq 0} (1 - (1/k + \epsilon''))^i = k - \epsilon'$. Note that ϵ' can be made as small as we like by selecting sufficiently small ϵ''. Thus, we can have the approximation ratio arbitrarily close to $2k/(k+1)$. □

The proof of Lemma 4.1 can be modified to apply also to instances of *sequential cover*. The modified "bad" instance, however, has many more queries and trials. To obtain this instance we replace each query in the sequential trial optimization instance above by a sufficiently large set of queries. Let s_i be the ith set of queries. We have k subsets, where the kth subset succeeds for all elements in s_i. The optimal solution selects these k trials in any order. We also have additional subsets $j = 1, 2, \ldots$ where the jth subset contains $(1 - 1/k)^{j-1}/k + \epsilon''$ of the queries in each s_i. Different subsets contain different queries from each s_i. A simple calculation similar to the one above for *sequential trial optimization* shows that if the sets s_i are of sufficiently large size, we can have an approximation ratio arbitrarily close to $2k/(k+1)$. We thus obtain the following

COROLLARY 4.1. *For any integer k, the approximation ratio of* GREEDY *on* sequential cover *instances with $|\sigma^{\mathrm{opt}}| \leq k$ is at least $2k/(k+1)$.*

Last, we remark that in a similar manner, it is possible to "approximate" any instance of *sequential trial optimization* arbitrarily closely by a sufficiently large *sequential cover* instance. Thus, the GREEDY strategy admits the same approximation ratio on both formulations.

5 GREEDY is a Constant Factor Approximation to the Optimal Solution

We show that the average number of trials according to the sequence produced by GREEDY is within a small constant factor of the average number of trials that one can achieve with any other sequence.

THEOREM 5.1. *For any given $Q = (q_1, q_2, \ldots, q_n)$, and M, $E(\sigma^g) \leq \frac{4e}{e-1} E(\sigma^{\mathrm{opt}}) + O(1)$.*

To prove Theorem 5.1 we need the following lemma. The first lemma states that the earlier we perform a trial, the greater the increase in the success probability that we gain from the trial.

LEMMA 5.1. *Consider a trial sequence $\sigma = \langle \sigma_1, \sigma_2, \ldots \rangle$. Let $1 \leq i < j$ and consider the sequence $\hat{\sigma}$, $|\hat{\sigma}| = i$,* *where* $\mathrm{pref}(i-1,\hat\sigma) = \mathrm{pref}(i-1,\sigma)$ *and* $\hat\sigma_i = \sigma_j$. *Then,*

$$\begin{aligned} &P(\mathrm{pref}(j,\sigma)) - P(\mathrm{pref}(j-1,\sigma)) \\ &\quad\le\quad P(\mathrm{pref}(i,\hat\sigma)) - P(\mathrm{pref}(i-1,\hat\sigma)). \end{aligned}$$

Proof. Using equation (1.1) we get:

$$\begin{aligned} &P(\mathrm{pref}(j,\sigma)) - P(\mathrm{pref}(j-1,\sigma)) = \\ &= \big(M_{\sigma_j}\ , Q * (\bar 1 - M_{\sigma_1}) * (\bar 1 - M_{\sigma_2}) * \cdots * \\ &\quad (\bar 1 - M_{\sigma_{i-1}}) * (\bar 1 - M_{\sigma_i}) * \cdots * (\bar 1 - M_{\sigma_{j-1}})\big) \\ &\le \big(M_{\sigma_j}\ , Q * (\bar 1 - M_{\sigma_1}) * (\bar 1 - M_{\sigma_2}) * \cdots \\ &\quad *(\bar 1 - M_{\sigma_{i-1}})\big) \\ &= \big(M_{\hat\sigma_i}\ , Q * (\bar 1 - M_{\hat\sigma_1}) * (\bar 1 - M_{\hat\sigma_2}) * \cdots \\ &\quad (\bar 1 - M_{\hat\sigma_{i-1}})\big) \\ &= P(\mathrm{pref}(i,\hat\sigma)) - P(\mathrm{pref}(i-1,\hat\sigma)). \end{aligned}$$

The inequality on the second line follows since (a) For all k, $(\bar 1 - M_k)$ is a vector with all entries ≤ 1, and (b) $(x,y) \le (x,y')$ if $x_i \ge 0$, $y'_i \ge y_i \ge 0$. $\square$

The next lemma shows that the success probability of a trial sequence depends only on the trials and not on the order in which we carry them out.

Lemma 5.2. *Let* $\sigma = \langle \sigma_1, \sigma_2, \ldots, \sigma_{|\sigma|}\rangle$ *be a trial sequence, and let* τ *be a permutation of* $1,\ldots,|\sigma|$. *The success probability of* σ *is equal to the success probability of* $\tau(\sigma) = \sigma_{\tau(1)}, \sigma_{\tau(2)}, \ldots, \sigma_{\tau(\sigma)}$.

Proof. We show that the success probability of $\sigma =< \sigma_1, \sigma_2, \ldots, \sigma_{|\sigma|}\ >$ is equal to the success probability of $\sigma' = \langle \sigma_1, \sigma_2, \ldots, \sigma_{i-1}, \sigma_{i+1}, \sigma_i, \sigma_{i+2}, \ldots, \sigma_{|\sigma|}\rangle$. From this the lemma clearly follows. Using Equation (1.1) we have that

$$\begin{aligned} &P(\sigma) = \\ &\qquad (M_{\sigma_1}, Q) + (M_{\sigma_2}, Q * (\bar 1 - M_{\sigma_1})) \\ &\qquad + (M_{\sigma_3}, Q * (\bar 1 - M_{\sigma_1}) * (\bar 1 - M_{\sigma_2})) + \cdots \\ &\qquad + \big(M_{\sigma_i}, Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}})\big) \\ &\qquad + \big(M_{\sigma_{i+1}}, Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}}) \\ &\qquad *(\bar 1 - M_{\sigma_i})) + \ldots \\ &= \ldots + (M_{\sigma_i}, Q * (\bar 1 - M_{\sigma_1}) * (\bar 1 - M_{\sigma_2}) * \cdots \\ &\qquad *(\bar 1 - M_{\sigma_{i-1}})\big) + \\ &\qquad \big(M_{\sigma_{i+1}} * Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}}), \bar 1\big) - \\ &\qquad \big(M_{\sigma_{i+1}} * Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}}), M_{\sigma_i}\big) + \ldots \\ &= \ldots + \big(M_{\sigma_{i+1}}, Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}})\big) + \\ &\qquad \big(M_{\sigma_i} * Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}}), \bar 1\big) - \\ &\qquad \big(M_{\sigma_i} * Q * (\bar 1 - M_{\sigma_1}) * \cdots * (\bar 1 - M_{\sigma_{i-1}}), M_{\sigma_{i+1}}\big) + \ldots \\ &= P(\sigma') \end{aligned}$$

$\square$

As an immediate corollary of Lemma 5.2 we get that by additional trials inserted anywhere along the sequence we cannot decrease the success probability.

Corollary 5.1. *Consider a trial sequence* σ, *and let* $\hat\sigma$ *be a subsequence of* σ. *Then,* $P(\sigma) \ge P(\hat\sigma)$.

We define O_k to be the maximum success probability one can get with k trials, i.e. $O_k = \max\{P(\sigma) \mid |\sigma| \le k\}$. We also define G_k to be the success probability we get with the first k trials picked by greedy, i.e. $G_k = P(\mathrm{pref}(k, \sigma^g))$. The following two lemma are more specific to the proof of Theorem 5.1.

Lemma 5.3. *We have that* $G_0 = O_0 = 0$, *and for every* $i \ge 0$ *we have that*

$$G_{i+1} \ge G_i + \sup_k \frac{O_k - G_i}{k}.$$

Proof. Let $o_1, \ldots, o_k$ be the k trials that achieve success probability of O_k. Consider the sequence of $i+k$ trials where the first i trials are the first i trials of greedy, and the following k trials are $o_1, \ldots, o_k$. The probability of success after these $i+k$ trials is at least O_k. The probability of success after the first i trials is G_i. Therefore one of the last k trials increases the success probability by at least $(O_k - G_i)/k$.

Let ℓ be a trial which increases the success probability by at least $(O_k - G_i)/k$. Consider the trial sequence consisting of the first i trials of Greedy, followed by trial ℓ. By Lemma 5.1 if we perform trial ℓ immediately after the first i trials of greedy it increases the success probability by at least $(O_k - G_i)/k$. Therefore, by the definition of the greedy sequence, the $i+1$'st trial chosen by greedy increases the success probability by at least this amount. $\square$

Lemma 5.4. *For every* $k \ge 0$, $G_k \ge O_k(1 - \frac{1}{e})$.

Proof. From Lemma 5.3 we obtain that for every $i \ge 1$,

$$G_i \ge G_{i-1} + \frac{O_k - G_{i-1}}{k}\ .$$

By subtracting both sides from O_k we have that

$$O_k - G_i \le (O_k - G_{i-1}) \left(1 - \frac{1}{k}\right)\ .$$

using this equation for $i = k, k-1, \ldots, 1$ we obtain that

$$O_k - G_k \le (O_k - G_0) \left(1 - \frac{1}{k}\right)^k\ ,$$

and since $\left(1 - \frac{1}{k}\right)^k \le 1/e$ for $k \ge 1$ the lemma follows. $\square$

We now prove Theorem 5.1.

Proof. [of Theorem 5.1] Let n_1 be the minimum number of trials one needs to reach success probability which is at least $t_1 = 1/2$. Clearly it follows from our definitions that $O_{n_1} \geq 1/2$. By Lemma 5.4 we obtain that $G_{n_1} \geq O_{n_1}\rho \geq \frac{1}{2}\rho$ where $\rho = 1 - \frac{1}{e}$.

Now we consider the state of GREEDY after it performed its first n_1 trials and reached success probability of $G_{n_1} \geq \frac{1}{2}\rho$. We bound the number of additional trials GREEDY has to perform in order to increase its success probability to at least $\frac{1}{2}\rho + \frac{1}{2}\rho(1 - \frac{1}{2}\rho)$.

Let n_2 be the minimum number of trials it takes to reach success probability which is at least $t_2 = \frac{1}{2}\rho + \frac{1}{2}(1 - \frac{1}{2}\rho)$. Let σ be a sequence of n_2 trials that achieves such success probability. By Lemma 5.1 if we perform these n_2 trials after the first n_1 trials of GREEDY we also reach success probability that is at least t_2.

This means that if we consider the residual problem where the query distribution is $Q' = u_{\mathrm{pref}(n_1,\sigma^g)}(Q)$, the posterior query distribution after performing the first n_1 trials of greedy, then using at most n_2 trials we can reach success probability of at least $\frac{t_2 - G_{n_1}}{1 - G_{n_1}}$. In other words for this residual problem $O_{n_2} \geq \frac{t_2 - G_{n_1}}{1 - G_{n_1}}$. Using Lemma 5.4 we obtain that for the residual problem $G_{n_2} \geq \rho\frac{t_2 - G_{n_1}}{1 - G_{n_1}}$. Going back to our original problem its follows that greedy by performing n_2 additional trials increases its success probability to at least $G_{n_1} + \rho(t_2 - G_{n_1}) \geq \frac{1}{2}\rho + \frac{1}{2}\rho(1 - \frac{1}{2}\rho)$. That is,

$$G_{n_1+n_2} \geq \frac{1}{2}\rho + \frac{1}{2}\rho(1 - \frac{1}{2}\rho).$$

Generally, we define n_1 to be the minimum number of trials it takes to reach success probability which is at least $1/2$, and for $i \geq 2$ we define n_i to be the minimum number of trials it takes to reach success probability which is at least

$$t_i = \sum_{j=0}^{i-2} \frac{1}{2}\rho(1 - \frac{1}{2}\rho)^j + \frac{1}{2}(1 - \frac{1}{2}\rho)^{i-1} \ .$$

Using induction on i and an argument as for $G_{n_1+n_2}$ one can easily show that for every $i \geq 1$, $G_{n_1+n_2+\cdots+n_i} \geq g_i$ where $g_i = \sum_{j=0}^{i-1} \frac{1}{2}\rho(1 - \frac{1}{2}\rho)^j$.

To obtain an upper bound on $E(\sigma^g)$ we look at formula (2.2) for the expectation. From the lower bound on $G_{n_1+n_2+\cdots+n_i}$ it follows that we can upper bound the first n_1 terms in (2.2) by 1, the next n_2 terms by $1 - g_1$, etc. We obtain that

$$\begin{aligned} E(\sigma^g) &\leq n_1 + n_2(1-g_1) + \cdots + n_i(1-g_{i-1}) + \cdots \\ &= n_1 + n_2(1 - \frac{1}{2}\rho) + \cdots + n_i(1 - \frac{1}{2}\rho)^{i-1} \\ &\quad + \cdots \ , \end{aligned} \tag{5.3}$$

where the second inequality follows since $1 - g_i = (1 - \frac{1}{2}\rho)^i$ for $i \geq 1$.

We also derive the following lower bound on $E(\sigma^{opt})$. From the definition of n_1 follows that the success probability of σ^{opt} after each of the first $n_1 - 1$ trials is smaller than $t_1 = 1/2$. Similarly, from the definition of n_2 follows that the success probability of σ^{opt} after each of the following $n_2 - n_1$ trials is no larger than t_2, and in general the success probability of σ^{opt} after each of the experiments $n_{i-1}, n_{i-1}+1, \cdots n_i - 1$ is at most t_i. It follows using Equation (2.2) that

$$\begin{aligned} E(\sigma^{opt}) &\geq \frac{n_1}{2} + (n_2 - n_1)(1 - t_2) + \cdots + \\ &\quad (n_i - n_{i-1})(1 - t_i) + \cdots \\ &= \frac{n_1}{2} + (n_2 - n_1)\frac{1}{2}(1 - \frac{1}{2}\rho) + \cdots + \\ &\quad n_i\frac{1}{2}(1 - \frac{1}{2}\rho)^{i-1} + \cdots \ , \end{aligned} \tag{5.4}$$

where the second inequality follows since $1 - t_i = \frac{1}{2}(1 - \frac{1}{2}\rho)^{i-1}$ for $i \geq 1$. Rearranging terms in (5.4) and multiplying by 2 we obtain that

$$\begin{aligned} &2E(\sigma^{opt}) \geq \\ &\quad n_1(1 - (1 - \frac{1}{2}\rho)) + n_2((1 - \frac{1}{2}\rho) - (1 - \frac{1}{2}\rho)^2) \\ &\quad + \cdots + n_i((1 - \frac{1}{2}\rho)^{i-1} - (1 - \frac{1}{2}\rho)^i) + \cdots \end{aligned} \tag{5.5}$$

From the upper bound (5.3) and the lower bound (5.5) it now follows that

$$\begin{aligned} &E(\sigma^g) - 2E(\sigma^{opt}) \leq \\ &= n_1(1 - \frac{1}{2}\rho) + n_2(1 - \frac{1}{2}\rho)^2 + \cdots + \\ &\quad n_i(1 - \frac{1}{2}\rho)^i + \cdots \\ &= (1 - \frac{1}{2}\rho)\Big(n_1 + n_2(1 - \frac{1}{2}\rho) + \cdots + \\ &\quad n_i(1 - \frac{1}{2}\rho)^{i-1} + \cdots\Big) \\ &\leq \frac{1 - \frac{1}{2}\rho}{\frac{1}{2}\rho}\left(2E(\sigma^{opt}) + 1\right) \end{aligned} \tag{5.6}$$

Thus,

$$E(\sigma^g) \leq \left(2 + \frac{(2-\rho)}{\frac{1}{2}\rho}\right) E(\sigma^{opt}) + \frac{1 - \frac{1}{2}\rho}{\frac{1}{2}\rho}$$

Since $\rho = 1 - \frac{1}{e}$ our multiplicative factor is $\frac{4e}{e-1}$. $\square$

6 Performance of partial-success sequences

Recall that $E^+(f,\sigma)$ is the contribution to $E(f,\sigma)$ by the shortest successful outcomes with total probability mass f. We now study the relationship between $E^+(f,\sigma)$ and $E(f,\sigma)$. A straightforward generalization of Theorem 5.1 to $f < 1$ would give a sequence σ^g so that for any sequence σ with $P(\sigma) \geq f$

$$E^+(f,\sigma^g) \leq O(1) \cdot E^+(f,\sigma) \ . \tag{6.7}$$

We would like, however, to obtain a similar relation with respect to E rather than E^+. We obtain the following generalization of Theorem 5.1:

THEOREM 6.1. *For any given $Q = (q_1, q_2, \ldots, q_n)$, f, M, and ϵ, we can compute some prefix, $\sigma^{g'}$, of the greedy trial sequence, such that for any sequence σ with $P(\sigma) \geq f$*

$$E(f - \epsilon \min\{f, 1-f\}, \sigma^{g'}) \leq O(1/\epsilon) E(f,\sigma).$$

(Note that the constant in $O(1/\epsilon)$ is independent of f,σ, M, and Q.)

We now present some lemmas that will be needed in the proof of the theorem. We start with the following easy claim

LEMMA 6.1. *For any sequence σ and $f \leq P(\sigma)$,*

$$E(f,\sigma) \geq (1 + (1-f)/f) E^+(f,\sigma) \ . \tag{6.8}$$

Proof. By its definition, $E^+(f,\sigma)/f$ is a weighted average of numbers of size at most $T^{f,\sigma}$, thus $E^+(f,\sigma) \leq fT^{f,\sigma}$. By definition we have $E(f,\sigma) - E^+(f,\sigma) = (1-f)T^{f,\sigma}$. Combining the above we obtain

$$E(f,\sigma) \geq E^+(f,\sigma) + \frac{1-f}{f} E^+(f,\sigma) \ ,$$

from which the claim easily follows. □

We will use the following lemma in the proof of Lemma 6.3. Let $\mathrm{Prob}_{D^+(f,\sigma)}(Z)$ denote the probability over the distribution $D^+(f,\sigma)$ that the boolean valued random variable Z is true.

LEMMA 6.2. *Let X be a random variable that gives the number of trials in outcomes chosen from the distribution $D^+(f,\sigma)$ and let $a > 0$, if $\mathrm{Prob}_{D^+(f,\sigma)}(X \geq \ell) \geq 1/a$ then $\ell \leq a\frac{E^+(f,\sigma)}{f}$.*

Proof. Recall that the expected number of trials over outcomes chosen from $D^+(f,\sigma)$ is $E^+(f,\sigma)/f$. Using the Markov inequality we get that

$$\mathrm{Prob}_{D^+(f,\sigma)} \left(X \geq a\frac{E^+(f,\sigma)}{f} \right) \leq 1/a \ . \tag{6.9}$$

By reversing the condition, we obtain that for any ℓ, if $\mathrm{Prob}_{D^+(f,\sigma)}(X \geq \ell) \geq 1/a$ then $\ell \leq a\frac{E^+(f,\sigma)}{f}$, or equivalently if $\mathrm{Prob}_{D^+(f,\sigma)}(X \leq \ell) \leq 1 - 1/a$ then $\ell \leq a\frac{E^+(f,\sigma)}{f}$. □

It remains to prove the following lemma.

LEMMA 6.3. *Given f, a query distribution vector Q, a matrix M, and a trial sequence σ where $P(\sigma) \geq f$, then for all $\delta > 0$ there exists a trial sequence σ', a prefix of σ, such that*

$$E(f-\delta, \sigma') \leq (1 + (1-f)/\delta) E^+(f,\sigma) \ .$$

Proof. Let $\sigma' = \mathrm{pref}(\ell,\sigma)$ for the minimal ℓ such that

$$P(\mathrm{pref}(\ell,\sigma)) \geq f - \delta. \tag{6.10}$$

By definition, all outcomes in $D(\sigma)$ with at most $\ell - 1$ trials lie in the probability subspace $D^+(f,\sigma) \subset D(\sigma)$. The measure of the subspace $D^+(f,\sigma)$ is f whereas by definition, the measure of all outcomes with at most $\ell - 1$ trials is at most $f - \delta$. It follows, that the measure of outcomes in $D(\sigma)$ that lie in $D^+(f,\sigma)$ and have at least ℓ trials is at least δ. Thus, at least δ/f fraction of the outcomes in the probability subspace $D^+(f,\sigma)$ have at least ℓ trials. Stated formally, we obtain

$$\mathrm{Prob}_{D^+(f,\sigma)}(X \geq \ell) \geq \delta/f$$

(where the random variable X is defined as in the statement of Lemma 6.2.) By invoking Lemma 6.2 we get that

$$\ell \leq E^+(f,\sigma)/\delta \ .$$

We now get that

$$\begin{aligned}
E(f-\delta,\sigma') &\leq E^+(f-\delta,\sigma') + (1-(f-\delta))\ell \\
&\leq E^+(f,\sigma) - \delta\ell + (1-(f-\delta))\ell \\
&= E^+(f,\sigma) + (1-f)\ell \\
&\leq E^+(f,\sigma) + (1-f)\frac{E^+(f,\sigma)}{\delta} \\
&= (1 + \frac{1-f}{\delta}) E^+(f,\sigma) \ .
\end{aligned}$$

□

We are now ready to present the proof of the theorem:

Proof. [of Theorem 6.1] By applying Lemma 6.3 to the greedy sequence σ^g and by substituting $\delta = \epsilon \min\{f, 1-f\}$, we obtain that there exists a prefix $\sigma^{g'}$ of σ^g such that $P(\sigma^{g'}) \geq f - \epsilon \min\{f, 1-f\}$ and

$$\begin{aligned}
&E(f - \epsilon\min\{f,1-f\}, \sigma^{g'}) \leq \\
&\left(1 + \frac{1-f}{\epsilon \min\{f, 1-f\}}\right) E^+(f,\sigma^g) \ .
\end{aligned} \tag{6.11}$$

By substituting Inequality (6.7) in the RHS of Inequality (6.8) we obtain

$$(6.12) \quad (1 + \frac{1-f}{f})E^+(f, \sigma^g) \leq O(1)E(f, \sigma) \ .$$

We now substitute Inequality (6.11) in the RHS of Inequality (6.12) to eliminate $E^+(f, \sigma^g)$ and obtain

$$E(f - \epsilon \min\{f, 1-f\}, \sigma^{g'}) \leq O(1)\frac{1 + \frac{1-f}{\epsilon \min\{f,1-f\}}}{1 + \frac{1-f}{f}}E(f, \sigma) \ .$$

To conclude observe that for $0 < f < 1$

$$\frac{1 + \frac{1-f}{\epsilon \min\{f,1-f\}}}{1 + \frac{1-f}{f}} = O(1/\epsilon) \ .$$

(It is easy to see the above by considering separately the cases $f \leq 1/2$ and $f \geq 1/2$.) □

Theorem 6.1 states that if we are willing to slightly compromise the success probability by $\epsilon \min\{f, 1-f\}$ then an appropriate greedy sequence approximates the optimal sequence with respect to E. Note that the amount by which we compromise f tends to 0 when f approaches 1. At the limit we obtain Theorem 5.1.

We now provide a simple example which demonstrates that some such compromise is necessary, *i.e.*, that $E(f, \sigma^g)$ can be arbitrarily far from $E(f, \sigma^{opt})$: Let $Q = (f/2, f/2, (1-f))$, and

$$\begin{aligned} M_1 &= (1/2 + \epsilon, 1/2 + \epsilon, 0) \\ M_2 &= (1, 0, 0) \\ M_3 &= (0, 1, 0) \ . \end{aligned}$$

The optimal sequence σ^{opt} will include t_2 and t_3 in some order and has $E^+(f, \sigma^{opt}) = 3f/2$ and $E(f, \sigma^{opt}) = 3f/2 + 2(1-f) = 2 - f/2$. The greedy sequence consists of infinite repetitions of t_1, and has $E^+(f, \sigma^g) < 2f$ but $E(f, \sigma^g)$ is unbounded. Each finite prefix $\sigma^{g'}$ of the greedy sequence has $P(\sigma^{g'})$ strictly less than f.

7 Sampling

It is important to consider the case when Q, M are not given apriori. We show that if we are able to sample queries from Q and apply trials to them we can still generate a sequence of trials that approximates the performance of the optimal sequence. The main idea here is to use sampling to learn the appropriate distributions. Because the Greedy strategy is a good approximation, it means that we can build the trial sequence in a greedy fashion, greatly reducing the sampling we have to do. Our sampling strategy is described next.

Assume we have already generated a prefix $\sigma_1, \ldots, \sigma_{j-1}$ of the sequence. We pick trial σ_j as follows. We sample a set of S queries from Q. On each query $q \in S$ we apply $\sigma_1, \ldots, \sigma_{j-1}$. If all these applications of $\sigma_1, \ldots, \sigma_{j-1}$ failed to answer q then we apply every possible trial in T to q and increment the score of each trial that succeeds. After performing this process for every $q \in S$ we pick σ_j to be the trial with maximum score.

Let σ^{opt} be the optimal sequence for some fraction f, and let $\ell = E^+(f, \sigma^{opt})/f$. By Lemma 6.2 we know that $P(\mathrm{pref}(\lceil 2l/\epsilon \rceil, \sigma)) \geq f - \epsilon/2$.

From Lemma 5.3 it follows that if $\sigma_1, \ldots, \sigma_{j-1}$ does not reach success probability $f - \epsilon$ then the best trial to apply next increases the success probability of $\sigma_1, \ldots, \sigma_{j-1}$ by at least

$$p = \frac{(f - \epsilon/2) - (f - \epsilon)}{\lceil 2\ell/\epsilon \rceil} = \frac{\epsilon^2}{4\ell} \ .$$

It follows from the standard Chernoff bound that if $|S| = O(\frac{1}{p\delta^2} \log m)$ then the trial which we pick to be σ_j increases the success probability by at least $(1-\delta)p$ with high probability. Thus we see that via sampling we are able to simulate with high probability an "approximate" greedy algorithm

Our Greedy algorithm picked the next trial to be the one that generates maximum increase in its overall success probability. An *approximate greedy* algorithm picks at each step a trial that increases the success probability by at least $(1-\delta)$ fraction of the increase of the best trial. We can generalize our analysis of Greedy to capture approximate greedy as well with a penalty to the approximation factor proportional to $\frac{1}{\delta}$.

We can now summarize the following:

Theorem 7.1. *Let $\ell = E^+(f, \sigma^{opt})$. By sampling $O(\frac{\ell}{\epsilon^2\delta^2} \log m)$ queries in each step and applying the algorithm described above to pick the next trial we can simulate an approximate greedy algorithm. This approximate greedy algorithm picks a trial that increases the success probability to within a factor of $(1-\delta)$ of the increase of the best trial. Let σ be the sequence generated by this algorithm. Then $E^+(f - \epsilon, \sigma) \leq O(1) \cdot E^+(f, \sigma^{opt})$ where the approximation factor is inversely proportional to δ.*

8 Further Generalizations

It is not hard to see that our analysis could be generalized for the case where the trials are weighted, with the goal of minimizing the expected sum of weights of trials that we apply to a randomly chosen query. The Greedy algorithm in this case is the algorithm that chooses the trial with the highest incremental success rate per unit of weight.

The analysis now requires several changes where we replace the O_k and G_k with O_{w_k} and G_{w_k}, where the value w_k is the cumulative weight of the first k trials chosen by greedy.

In our discussion a trial can either succeed or fail and when it fails we try another trial. A possible generalization is to consider the case where a trial can give a false result both when it succeeds and when it fails, and in both cases we may try another trial to increase our confidence in the outcome.

For example in the medical diagnostic setting we would like to allow, per query, some probability of a false positive and some probability of a false negative, conditioned on previous positives and negatives of various trials. The goal is to come up with a diagnosis that holds with some probability above a preset threshold, while optimizing the sequence of trials to come up with this diagnosis.

We leave this generalization as an open problem. A natural greedy approach here is to choose the trial that has the largest mutual information with our query, conditioned on the previous results of previous trials.

References

[1] V. Chvátal. A greedy heuristic for the set-covering problem. *Math. of Oper. Res.*, 4:233–235, 1979.

[2] E. Cohen, A. Fiat, and H. Kaplan. Associative search in peer to peer networks: Harnessing latent semantics. Manuscript, 2002.

[3] E. Cohen, A. Fiat, and H. Kaplan. A case for associative peer to peer overlays. In *Proceedings of HotNets-I, ACM SIGCOMM*, 2002.

[4] E. Cohen and S. Shenker. Replication strategies in unstructured peer-to-peer networks. In *Proceedings of the ACM SIGCOMM'02 Conference*, 2002.

[5] Open Source Community. Gnutella. In *http://gnutella.wego.com/*, 2001.

[6] U. Feige. A threshold of $\ln n$ for approximating set cover. *J. Assoc. Comput. Mach.*, 45:634–652, 1998.

[7] KaZaA file sharing network. KaZaA. In *http://www.kazaa.com/*, 2002.

[8] Morpheus file sharing system. Morpheus. In *http://www.musiccity.com/*, 2002.

[9] L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13:383–390, 1975.

[10] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. In *Proceedings of the 16th annual ACM International Conference on supercomputing*, 2002.

[11] FastTrack Peer-to-Peer technology company. FastTrack. In *http://www.fasttrack.nu/*, 2001.