

# How to Probe for an Extreme Value\*

Ashish Goel<sup>†</sup>

Sudipto Guha<sup>‡</sup>

Kamesh Munagala<sup>§</sup>

## Abstract

In several systems applications, parameters such as load are known only with some associated uncertainty, which is specified, or modeled, as a distribution over values. The performance of the system optimization and monitoring schemes can be improved by spending resources such as time or bandwidth in *observing* or *resolving* the values of these parameters. In a resource-constrained situation, deciding which parameters to observe in order to best optimize the expected system performance (or in general, optimize the expected value of a certain objective function) itself becomes an interesting optimization problem.

In this paper, we initiate the study of such problems that we term “model-driven optimization”. In particular, we study the problem of optimizing the minimum value in the presence of observable distributions. We show that this problem is NP-HARD, and present greedy algorithms with good performance bounds. The proof of the performance bounds are via novel sub-modularity arguments and connections to covering integer programs.

---

\*A preliminary version of this paper [13] appeared in the Proceedings of the 25<sup>th</sup> ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), 2006.

<sup>†</sup>Departments of Management Science and Engineering and (by courtesy) Computer Science, Stanford University. Research supported by an NSF CAREER award and an Alfred P. Sloan faculty fellowship. Email: [ashishg@stanford.edu](mailto:ashishg@stanford.edu)

<sup>‡</sup>Department of Computer and Information Sciences, University of Pennsylvania. Supported in part by an Alfred P. Sloan Research Fellowship, by an NSF CAREER award, and by NSF award CCF-0430376. Email: [sudipto@cis.upenn.edu](mailto:sudipto@cis.upenn.edu)

<sup>§</sup>Department of Computer Science, Duke University, Durham NC 27708. Research supported in part by NSF Award CNS 0540347. Email: [kamesh@cs.duke.edu](mailto:kamesh@cs.duke.edu)

# 1 Introduction

Optimization problems arising in databases, streaming, cluster computing, and sensor network applications often involve parameters and inputs whose values are known only with some uncertainty. In many of these situations, the optimization can be significantly improved by *resolving* the uncertainty in the input before performing the optimization. For instance, a query optimizer often has the ability to observe characteristics in the actual data set, like selectivities, either via random sampling or by performing inexpensive filters [5, 3]. As another example, a system like Eddies [2] finds the best among several competing plans which are run simultaneously. Each plan’s running time is a distribution which is observed by executing the plan for a short amount of time. In all such examples, the process of resolving the uncertainty also consumes resources, e.g., time, network bandwidth, space, etc., which compete with finding the solution of the original problem.

Therefore, judiciously choosing which variables to observe, itself becomes an important problem in this context. In this paper, we study the optimization problem of finding the minimum, when the inputs are random variables and the values of one or more inputs can be observed and resolved by paying some cost. We show that even for this simplest of optimization problems, this choice becomes non-trivial and intractable, motivating the development of algorithmic techniques to address them.

We initiate the study of the following abstractly defined class of problems, that we term “model-driven optimization”:

**Problem 1.** *We are given the distributions of non-negative independent random variables  $\{X_i\}_{i=1}^n$ . Further, these random variables are observable: we can find the value of  $X_i$  by spending cost  $c_i$ . Given a budget  $C$  and an objective function  $f$ , can we choose a subset  $S$  of random variables to observe whose total observation cost is at most  $C$ , and optimize the expected value of the function  $f(S)$ ? Note that the function  $f$  is evaluated after the observations, and the expectation is over the outcome of the observations.*

In this paper, we focus on the function  $f(S) = \min_{i \in S} X_i$ , so that the goal is to choose a subset  $S$  whose observation cost is at most  $C$ , so that  $\mathbf{E}[\min_{i \in S} X_i]$  is minimized. We define this problem as the MINIMUM ELEMENT problem. In Section 1.3, we present a brief survey of our results [16, 15, 17] on other objective functions  $f$ .

As a motivating example, in the context of traditional and P2P networks, “multi-homing” schemes are becoming a common method for choosing communication routes and server assignments. These schemes [1, 18] probe and observe the current state of multiple routes before deciding the optimum (minimum latency) route to use for a specific connection. The distribution of the latency of each route is available *a priori*. The number of probes needs to be bounded since flooding the network is undesirable. Therefore the goal is to minimize the latency of the route found by a bounded number of probes. The mapping to our framework in this case is as follows:  $X_i$  are the distributions of route latencies. The probing cost  $c_i$  is a function of the delay and load incurred in detecting latency of route  $i$ . The budget  $C$  is the total load and processing time that can be incurred by the route choosing algorithm. Finally,  $f = \min_{i \in S} X_i$ , where  $S$  is the set of routes probed. This corresponds to the goal of choosing that set of routes to probe which minimizes the expected value of the smallest latency detected.

Note that if the minimum value among the probed set is more than the expected value of a variable that has not been probed, we would prefer to use that variable (the “backup”) as opposed to one of the probed values. We will not take this optimization into account while analyzing our algorithm. Refer Section 6 for the reason. We now show the benefit of probing with an example.

**Example 1.1.** *If all variables are Bernoulli  $B(1, p)$  (with any  $p$ ), the estimate of the **minimum** is  $p$  if only one probe is allowed, but is  $p^n \ll p$  if all nodes are probed. Probing can therefore yield*

an estimate which is **exponentially** smaller, which means that if there is a low utilization, we will very likely find it.

Note that the optimization chooses the final variable *after* the results of the probes are known. Therefore, the overall optimization is to choose  $S$  with cost at most  $C$ , so that  $\mathbf{E}[\min_{i \in S} X_i]$  is minimized. This is very different from optimizing  $\min_{i \in S} \mathbf{E}[X_i]$ , which is obtained when we choose the final variable *before* the results of the observations are known. The latter version is of course trivial to solve.

Consider the simple case when the probing costs of all nodes are equal. Let  $m$  denote the constraint on the number of variables that can be probed. It would appear that the optimal strategy would be to choose the  $m$  nodes with the smallest expected values. The example below shows that this need not be the case. Note further that our problem is not the same as minimizing residual entropy (Krause and Guestrin [25]) which minimizes uncertainty of the joint distribution – we are concerned with minimizing the uncertainty of an optimization that depends on the joint distribution, the minimum value. Minimizing residual entropy will most often involve probing a different set of variables than those required for estimating best the specific function at hand (see example below); therefore, the problems are orthogonal.

**Example 1.2.** *There are 3 distributions  $X_1, X_2$  and  $X_3$  and  $m = 2$ . Distribution  $X_1$  is 0 with probability  $\frac{1}{2}$  and 1 with probability  $\frac{1}{2}$ . Distribution  $X_2$  is 1 with probability  $\frac{1}{2}$  and 2 with probability  $\frac{1}{2}$ . Distribution  $X_3$  is 0 with probability  $\frac{1}{5}$  and 2 with probability  $\frac{4}{5}$ . Clearly,  $\mathbf{E}[X_1] < \mathbf{E}[X_2] < \mathbf{E}[X_3]$ . However, probing  $X_1, X_2$  yields an expected minimum value of 0.5, while probing  $X_1, X_3$  yields an expected minimum value of 0.4. Minimizing residual entropy [25] would also choose the sub-optimal set  $\{X_1, X_2\}$ .*

The simple strategy does not take into account the *shape* of the distributions. In fact, the MINIMUM-ELEMENT problem becomes NP-HARD for arbitrary distributions even with uniform costs. In fact, we show a stronger hardness result: it is NP-HARD to approximate the minimum value up to polynomial factors without exceeding the observation budget  $C$ . Hence the natural model to study this problem is from the perspective of *resource augmentation*: can we guarantee that we achieve the same solution as the optimum, while we pay a larger observation cost?

## 1.1 Results

We introduce the problem of model driven optimization in the presence of observations, and particularly consider the MINIMUM-ELEMENT problem. We present natural algorithms that are easy to implement and provide strong evidence that these algorithms are likely to be the best possible. We note that naive greedy algorithms do not work and extra algorithmic techniques are required to augment the greedy algorithm.

Our first result shows a deep connection between the MINIMUM-ELEMENT problem and a certain type of covering integer program. We use this connection to show that it is NP-HARD to approximate the objective up to any polynomial factor without augmenting the cost budget. Consequently, we design algorithms that approximate the cost while achieving nearly optimal objective value. The algorithms we design yield a  $(1 + \epsilon)$  approximation to the optimal value by increasing the cost budget by a factor  $\mu + O(\log \frac{1}{\epsilon})$ . Our results show different values of  $\mu$  for different types of distributions.

In the most general case, we show that the function  $f(S) = \mathbf{E}[\min_{i \in S} X_i]$  is *log-concave*, i.e.,  $\log 1/f(S)$  is sub-modular (simply showing  $f(S)$  is submodular, yields a approximation ratio polynomial in  $m$ . Consequently a greedy algorithm gives  $\mu = O(\log \log \frac{\min_{i \in S^*} \mathbf{E}[X_i]}{\mathbf{E}[\min_{i \in S^*} X_i]})$  where  $S^*$  is the optimal solution.

We then show approximation algorithms whose bounds depend on entirely different parameters of the problem, as well as improved results for special cases. These are summarized below.

1. If the distributions are discrete over a domain of  $m$  arbitrary values, then  $\mu = O(\log m)$ . This uses the connection with covering programs mentioned above, and proceeds by using the rounding a covering program as the start solution to the greedy scheme.
2. If the distributions are over the domain  $\{0, 1, 2, \dots, M\}$ , then  $\mu = O(\log \log M)$ . This uses a scaling argument combined with the covering program.
3. For arbitrary uniform distributions,  $\mu = O(1)$ . To show this, we develop a novel characterization of how the minimum changes when uniform distributions are truncated. We show that a modified greedy algorithm that tries different truncated distributions in turn yields the desired result. The truncation argument is of independent interest.

In terms of techniques, we combine an involved sub-modularity argument along with the analysis of the best *fractional* solutions of covering problems. Although the analyses are complicated, the algorithms are natural.

## 1.2 Related Work

The notion of refining uncertainty has been considered in an *adversarial setting* by several researchers [27, 12, 22, 7]. In the adversarial model, the only prior information about an input is the lower and upper bounds on its value. The goal is to minimize the observations needed to estimate some function over these inputs *exactly*, and often negative results arise. The use of lower and upper bounds do not exploit the full power of models/samples/stochasticity of the data, i.e., the *distributions* of inputs. However to use the distributional information we must optimize the *expected value* of the function, which is also referred to as stochastic optimization.

More recently, significant attention has been devoted towards developing and using models and estimates of data based on prior knowledge, e.g., [9, 8, 11, 4, 5, 3, 29] among many others. Our work complements the body of research on maintenance of samples and estimates, and we show that judicious probing may yield exponentially better estimates.

Another line of work by Dean *et. al.*, [10] considers knapsack problem in the model that the job sizes are revealed only after the job has *irrevocably* placed in the knapsack. In the settings we described, this would imply that the decision to refine our estimate, i.e., probing, is equivalent to selecting the item in the final solution. This effectively disallows probing. In our model the choice of which variables to pack in the knapsack would be made *after* the observations. There also has been ongoing research in Multi-stage stochastic optimization [23, 14, 21, 19, 20, 28], however most of this literature also involves making irrevocable commitments.

## 1.3 Subsequent Results

The model-driven optimization framework can be defined for other objective functions  $f$ . Though this paper presents the best results known for MINIMUM-ELEMENT other work has considered different objective functions. First, when  $f$  is a single constraint packing problem such as knapsack with random profits which are observable, Guha and Munagala [15] show a 8 approximation based on rounding the solution of a natural linear program. The approximation ratio holds even when the optimal solution is allowed adaptive (*i.e.*, can be based on the results of previous observations). It further holds even when the hidden quantity is a distribution (instead of a single value) and a prior

on this distribution is specified as input. A special case of this result is when  $f(S) = \max_{i \in S} X_i$ , or the MAXIMUM ELEMENT problem. Since the MINIMUM and MAXIMUM element problems are equivalent from the point of view of exact solution, the NP-HARDNESS result we present below also shows that MAXIMUM ELEMENT is NP-HARD. However, from the point of view of approximation, the techniques and results for the MINIMUM and MAXIMUM element problems are very different.

In [17], Guha, Munagala, and Sarkar consider the Lagrangean version of the MAXIMUM ELEMENT problem, where the observations are adaptive and the goal is to maximize the expected difference between the maximum value and the observation cost. Note that there is no budget on this cost, instead it is part of the objective function. If the maximum value needs to be chosen from the observed distributions, this problem has an optimal solution. The same trivially holds for optimizing the sum of the minimum value and the observation cost. When an unobserved distribution can also be chosen as the maximum value, the Lagrangean version is shown to have a 1.25 approximation via a greedy algorithm that is analyzed by a non-trivial structural characterization of the optimal solution.

Finally, when  $f$  is geometric problem, such as  $K$ -median clustering or spanning tree construction, on independent (and observable) point clouds, Guha and Munagala [16] show constant factor approximation algorithms when the observations are adaptive. These algorithms are based on converting the observation problem into an *outlier* version of the problem. Similar results are also shown in [16] for the average completion time scheduling problem.

## 2 Minimum–Element: Preliminaries

We are given  $n$  independent non-negative random variables  $X_1, X_2, \dots, X_n$ . Assume that  $X_i$  has observation cost  $c_i$ . There is a budget  $C$  on the total cost. The goal is to choose a subset  $S$  of distributions with total cost at most  $C$  which minimizes  $\mathbf{E}[\min_{i \in S} X_i]$ . Without loss of generality, we can assume that  $c_i \leq C$  for all  $i$ . We further assume the distributions are specified as discrete distributions over  $m$  values,  $0 \leq a_1 \leq a_2 \leq \dots \leq a_m \leq M$ . Let  $p_{ij} = \Pr[X_i \geq a_j]$ . The input is specified as the  $p_{ij}$  and  $a_j$  values. Note that  $m$  is not very large since frequently the distribution is learned from a histogram/quantile.

**Some Notation:** Let  $a_0 = 0$  and  $a_{m+1} = M$ . For  $j = 0, 1, \dots, m$ , let  $l_j = a_{j+1} - a_j$ . We call  $I_j = [a_j, a_{j+1}]$  the  $j^{\text{th}}$  interval. This is illustrated in Figure 1. Recall that  $p_{ij} = \Pr[X_i \geq a_j]$ . We have  $\mathbf{E}[X_i] = \sum_{j=0}^{m-1} p_{ij} l_j$ . We define  $f(S) = \mathbf{E}[\min_{i \in S} X_i]$  for each subset  $S$  of variables. All logarithms are to base  $e$ . Let  $f(\Phi) = M$ .

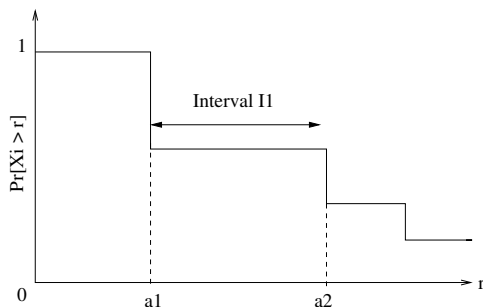


Figure 1: Notation used in MINIMUM–ELEMENT .

### 3 NP–Hardness

We begin with a hardness result: It is NP–HARD to obtain a  $\text{poly}(m)$  approximation on the objective for MINIMUM–ELEMENT while respecting the cost budget, even for uniform costs. We therefore focus on approximation algorithms for this problem which achieve the optimal objective while augmenting the cost budget. Thus the approximation results are on the *cost* in this paper.

**Definition 3.1.** A *Covering Integer Program (CIP)* over  $n$  variables  $x_1, x_2, \dots, x_n$  (indexed by  $i$ ) and  $m$  constraints (indexed by  $j$ ) has the form

$$\begin{aligned} \min \sum_i c_i x_i \\ \text{subject to: } \quad A\vec{x} \geq \vec{b} \\ \vec{x} \in \{0, 1\}^n \end{aligned}$$

where  $c_i \in \mathbb{R}^+$  and  $A \in \mathbb{Z}^{m \times n}$ , i.e., the elements  $A_{ji}$  of the constraint matrix are non-negative integers. This is a generalization of SET–COVER where the matrix  $A$  is  $\{0, 1\}$  and  $c_i = 1$ . A CIP is defined to be column-monotone if  $A_{ji} \leq A_{(j+1)i}$  for all  $i$  and for all  $j < m$ . Without loss of generality, we can assume that  $b_j \in \mathbb{Z}^+$  and  $b_{j+1} \geq b_j$ .

Suppose we are given a column-monotone CIP with a cost budget  $C$  and our goal is to determine whether the optimum value of the CIP is less than  $C$  or more than  $rC$ ; if the optimum value lies between  $C$  and  $rC$  then either of the two answers is considered valid. We will relate the hardness of this decision problem (which we call  $r$ -GAP-CIP) to the hardness of approximating the MINIMUM–ELEMENT problem.

An  $(r, s)$ -approximation for the MINIMUM–ELEMENT problem violates the cost budget by at most  $r$  and obtains an objective function value (i.e. the expectation of the minimum) that is at most  $s$  times the optimum objective function value with the original cost budget.

**Lemma 3.2.** The  $r$ -GAP-CIP problem with polynomially bounded (in  $n$  and  $m$ ) coefficients  $A_{ji}$  reduces, in polynomial time, to the problem of obtaining an  $(r, \text{poly}(m))$ -approximation for MINIMUM–ELEMENT .

*Proof.* Fix any constant  $k$ , and let  $q = m^{k+1}$ . Define  $n$  distributions over values  $\mu_0, \mu_1, \dots, \mu_m$  where  $\mu_0 = 0$  and  $\mu_j - \mu_{j-1} = q^{b_j}$ . Distribution  $X_i$  has observation cost  $c_i$ , and  $\Pr[X_i \geq \mu_{j-1}] = q^{-A_{ji}}$ . Observe that column-monotonicity is crucial for this definition to correspond to a valid probability distribution; the requirement that  $A_{ji}$ 's be polynomially bounded is crucial for the reduction to be polynomial time. Let the cost budget for the MINIMUM–ELEMENT problem be  $C$ .

First assume that the original CIP has a solution  $x_1, x_2, \dots, x_n$  with cost at most  $C$ . Let  $S$  be the set  $\{i : x_i = 1\}$ . For the MINIMUM–ELEMENT problem, probe the variables  $X_i, i \in S$ . For any  $j$ ,

$$\Pr[\min_{i \in S} X_i \geq \mu_{j-1}] = \prod_{i \in S} \Pr[X_i \geq \mu_{j-1}] = q^{-\sum_{i \in S} A_{ji}} \leq q^{-b_j}.$$

Therefore,

$$\mathbf{E}[\min_{i \in S} X_i] = \sum_j (\mu_j - \mu_{j-1}) \Pr[\min_{i \in S} X_i \geq \mu_{j-1}] \leq m.$$

Now suppose that the original CIP has no solution of cost  $rC$  or less. Then for any index set  $S$  such that  $\sum_{i \in S} c_i \leq rC$ , there must be at least one constraint  $j$  such that  $\sum_{i \in S} A_{ji} \leq b_j - 1$ . Thus,  $\Pr[\min_{i \in S} X_i \geq \mu_{j-1}] = q^{-\sum_{i \in S} A_{ji}} \geq q^{1-b_j}$ . Now,

$$\mathbf{E}[\min_{i \in S} X_i] \geq (\mu_j - \mu_{j-1}) \Pr[\min_{i \in S} X_i \geq \mu_{j-1}] \geq q.$$

Thus, the problem of distinguishing whether the optimum value of the original CIP was less than  $C$  or more than  $rC$  has been reduced to the problem of deciding whether MINIMUM-ELEMENT has an optimum objective value  $\leq m$  with cost budget  $C$  or an optimum value  $\geq q$  with cost budget  $rC$ .

Since  $q/m = m^k$ , we have obtained a polynomial time reduction from  $r$ -GAP-CIP to the problem of obtaining an  $(r, m^k)$ -approximation of the MINIMUM-ELEMENT problem.  $\square$

**Theorem 3.3.** *It is NP-HARD to obtain any poly( $m$ ) approximation on the objective for MINIMUM-ELEMENT while respecting the cost budget.*

*Proof.* We reduce from the well-known NP-HARD problem of deciding if a set cover instance has a solution of value  $k$ . The SET COVER problem is the following: given a ground set  $U$  with  $m$  elements and  $n$  sets  $S_1, S_2, \dots, S_n \subseteq U$  over these elements, decide if there are  $k$  of these sets whose union is  $U$ .

Write this set cover instance as a CIP as follows. There is an row for each element and a column for each set.  $A_{ji} = 1$  if element  $j$  is present in set  $S_i$  and 0 otherwise. All  $b_j = 1$  and all  $c_i = 1$ . To make this column-monotone, set  $A_{ji} \leftarrow A_{ji} + j$  for each  $j, i$  and set  $b_j \leftarrow 1 + jk$ . Clearly, if there is a solution to this monotone instance of value  $k$ , this solution has to be feasible for the set cover instance and is composed of  $k$  sets. Conversely, if the set cover instance has a solution with  $k$  sets, the monotone CIP has a solution of value  $k$ . Since deciding if a set cover instance has a solution using  $k$  sets is NP-HARD, solving this class of 1-GAP-CIP instances is NP-HARD. By the proof of Lemma 3.2, this implies a  $(1, \text{poly}(m))$ -approximation to the MINIMUM-ELEMENT problem is NP-HARD.  $\square$

We have only been able to prove NP-Hardness of column-monotone CIPs, and so have not been able to fully exploit the approximation preserving reduction in Lemma 3.2. A hardness of approximating column-monotone CIPs will immediately lead to a stronger hardness result for the MINIMUM-ELEMENT problem via Lemma 3.2.

```

MINIMUM-ELEMENT ( $\tilde{C}$ )
/*  $\tilde{C} =$  Relaxed cost bound ( $\tilde{C} \geq C$ ). */
 $S \leftarrow \Phi$ .
While ( $\sum_{i \in S} c_i \leq \tilde{C}$ )
     $X_q \leftarrow \operatorname{argmin}_i \frac{\log f(S \cup \{X_i\}) - \log f(S)}{c_i}$ .
     $S \leftarrow S \cup \{X_q\}$ 
endwhile
Output  $S$ 

```

Figure 2: Greedy Algorithm for MINIMUM-ELEMENT .

## 4 Greedy Algorithm

The algorithm is described in Figure 2 and takes a relaxed cost bound  $\tilde{C} \geq C$  as parameter, and outputs a solution of cost  $\tilde{C}$ . As we discuss later, the parameter  $\tilde{C}$  trades-off in a provable fashion with value of the solution found. The algorithm uses the slightly unnatural function  $\log f(S)$  instead of the more natural function  $f(S)$ . As our analysis shows, this modification provably improves our

approximation bound. The analysis of this algorithm uses the theory of submodularity [26]. *Sub-modularity* is a discrete analogue of convexity which is the basis of many greedy approximation algorithms. We formally define sub-modularity next.

**Definition 4.1.** A function  $g(S)$  defined on subsets  $S \subseteq U$  of a universal set  $U$  is said to be sub-modular if for any two sets  $A \subset B \subseteq U$  and an element  $x \notin B$ , we have  $g(A \cup \{x\}) - g(A) \geq g(B \cup \{x\}) - g(B)$ .

The key result in this section shows that the function  $\log \frac{1}{f}$  used by the greedy algorithm is sub-modular.

**Lemma 4.2.** Let  $f(S) = \mathbf{E}[\min_{i \in S} X_i]$ . Then, the function  $\log \frac{1}{f}$  is sub-modular.

*Proof.* Consider two sets of variables  $A$  and  $B = A \cup C$ , and a variable  $X \notin B$ . In order to prove the theorem, we need to show that  $\frac{f(A \cup \{X\})}{f(A)} \leq \frac{f(B \cup \{X\})}{f(B)}$ . We first define the following terms for each  $j = 0, 1, \dots, m-1$ :

1.  $\alpha_j = \Pr[(\min_{Y \in A} Y) \geq a_j] = \prod_{Y \in A} \Pr[Y \geq a_j]$ .
2.  $\beta_j = \Pr[(\min_{Y \in C} Y) \geq a_j] = \prod_{Y \in C} \Pr[Y \geq a_j]$ .
3.  $\gamma_j = \Pr[X \geq a_j]$ .

First note that the  $\alpha_j, \beta_j$  and  $\gamma_j$  values are non-negative and monotonically non-increasing with increasing  $j$ . Next, by the independence of the variables, we have:

$$\begin{aligned} f(A \cup \{X\}) &= \sum_{j=0}^{m-1} l_j \Pr[(X \geq a_j) \wedge (\min_{Y \in A} Y \geq a_j)] \\ &= \sum_{j=0}^{m-1} l_j \Pr[X \geq a_j] \Pr[(\min_{Y \in A} Y) \geq a_j] = \sum_{j=0}^{m-1} l_j \alpha_j \gamma_j \end{aligned}$$

Similarly,  $f(B) = \sum_{j=0}^{m-1} l_j \alpha_j \beta_j$  and  $f(B \cup \{X\}) = \sum_{j=0}^{m-1} l_j \alpha_j \beta_j \gamma_j$ .

Using the above, it follows that:

$$\frac{f(A \cup \{X\})}{f(A)} = \frac{\sum_j l_j \alpha_j \gamma_j}{\sum_j l_j \alpha_j} \quad \text{and} \quad \frac{f(B \cup \{X\})}{f(B)} = \frac{\sum_j l_j \alpha_j \beta_j \gamma_j}{\sum_j l_j \alpha_j \beta_j}$$

Therefore, we have:

$$f(A \cup \{X\})f(B) - f(B \cup \{X\})f(A) = \sum_{j < j'} l_j l_{j'} \alpha_j \alpha_{j'} (\gamma_j - \gamma_{j'}) (\beta_{j'} - \beta_j) \leq 0$$

The final inequality follows due to the monotonicity of both the  $\gamma_j$  and  $\beta_j$  values. The above implies  $\frac{f(A \cup \{X\})}{f(A)} \leq \frac{f(B \cup \{X\})}{f(B)}$ , which shows  $\log \frac{1}{f}$  is a sub-modular function.  $\square$

The connection between sub-modular functions and the greedy algorithm is captured by the following theorem [26]:

**Theorem 4.3** ([26]). Given a non-decreasing submodular function  $g(\cdot)$  on a universal set  $U$ , where each element  $i \in U$  has a cost  $c_i$ , and given a cost bound  $C \geq \max_i c_i$ , let  $S^* = \operatorname{argmax}\{g(S) \mid \sum_{i \in S} c_i \leq C\}$ . Consider the greedy algorithm that, having chosen a set  $T$  of elements, chooses the next one element  $i$  that maximizes the ratio  $\frac{g(T \cup \{i\}) - g(T)}{c_i}$ . Let  $g(\Phi)$  denote the initial solution. Then, for any  $\epsilon$ , the greedy algorithm using cost budget  $C \log \frac{g(S^*) - g(\Phi)}{\epsilon}$  finds a set  $T_\epsilon$  such that  $g(T_\epsilon) \geq g(S^*) - \epsilon$ .

Intuitively, sub-modularity ensures that current greedy choice has cost per unit increase in value of  $g()$  at most the corresponding value for the optimal solution. For MINIMUM-ELEMENT, let  $S^*$  denote the optimal solution using cost  $C$ .

**Theorem 4.4.** *Let  $V = \mathbf{E}[\min_{i=1}^n X_i]$ . The greedy algorithm for MINIMUM ELEMENT achieves a  $(1 + \epsilon)$  approximation to  $f(S^*)$  with cost parameter  $\tilde{C} = C(\log \log \frac{M}{V} + \log \frac{1}{\epsilon})$ .*

*Proof.* In the above theorem, set  $g = \log \frac{1}{f}$ . Also since  $f(\Phi) = M$ , we have  $g(\Phi) = \log \frac{1}{M}$ . Let  $S$  denote the greedy set. Suppose  $f(S) \leq (1 + \epsilon)f(S^*)$ . Therefore,  $g(S) \geq g(S^*) - \log(1 + \epsilon) \geq g(S^*) - \epsilon$ . Therefore,  $T_\epsilon = S$  in the above theorem, implying its cost  $\tilde{C} \leq C(\log \log \frac{M}{f(S^*)} + \log \frac{1}{\epsilon})$ . Since  $f(S^*) \geq V$ , we have  $\tilde{C} \leq C(\log \log \frac{M}{V} + \log \frac{1}{\epsilon})$ .  $\square$

**Note:** If we had used  $f()$  (which is submodular) as suggested by a naive greedy algorithm then we would have needed a worse cost of  $C(\log \frac{M}{V} + \log \frac{1}{\epsilon})$  to achieve a  $(1 + \epsilon)$  approximation to  $f(S^*)$ . Thus the improved analysis (and algorithm) was necessary. We next prove the following lower bound.

**Theorem 4.5.** *The analysis of the greedy algorithm is tight on the cost.*

*Proof.* There are  $K = \log \log M$  intervals of form  $I_i = [2^{2^i}, 2^{2^{i+1}}]$  for  $i = 1, 2, \dots, K$ . There are  $K$  distributions  $X_1, X_2, \dots, X_K$ . Distribution  $X_i$  ( $i = 1, 2, \dots, K$ ) takes value  $2^{2^i}$  with probability  $(1 - 2^{-2^{i+1} + \epsilon})$  and takes value  $2^{2^r}$  otherwise. Here,  $r = K + 1$ . There are two special distributions  $Y_1$  and  $Y_2$  such that  $\Pr[Y_* \geq 2^{2^i}] = 2^{-2^i + \epsilon}$ . All distributions have unit cost. The optimal solution chooses just these two distributions, so that its value is  $K$  and its cost is 2.

We claim that GREEDY first chooses  $X_K$ . If GREEDY chose any other distribution then on the last interval  $I_K$  the contribution to its expectation would be at least  $(2^{2^r} - 2^{2^{r-1}})2^{-2^{r-1} + \epsilon}$ . Since  $2^{x^2} > 2^{x+1}$  for  $x > 1$ , the contribution is at least  $2^{2^r - 1} 2^{-2^{r-1} + \epsilon} \geq 2^{2^{r-1} + \epsilon}$ .  $\mathbf{E}[X_K] \leq 2^\epsilon + 2^{2^{r-1}}$ , which is smaller.

At this point, the contribution from the interval  $I_K$  to GREEDY is  $2^\epsilon$ . The contribution of the previous interval  $I_{K-1}$  is  $2^{2^{r-1}} \gg K$ . Clearly, the next distribution chosen by the greedy algorithm should reduce the contribution to this interval. Arguing inductively, GREEDY picks  $X_{K-1}$ ,  $X_{K-2}$ , and so on until it hits  $X_{K - \log \log K}$  in order to be competitive on the objective. Therefore greedy spends cost  $\Omega(K)$ .  $\square$

## 5 Improved Approximation Algorithms

The GREEDY algorithm has approximation ratio which depends (inversely) on the value of the optimal solution on that instance (a lower bound on which is  $V = \mathbf{E}[\min_{i=1}^n X_i]$ , which could be exponentially small). The approximation ratio increases as the value of the optimal solution reduces. However, the algorithm is desirable since the dependence is of the form  $\log \log V$ .

We now present an algorithm CIP-GREEDY whose approximation ratio of  $O(\log \frac{m}{\epsilon})$  depends just on  $m$ , the number of possible values that can be taken by the input discrete distributions. This algorithm is based on computing a good start solution for GREEDY using linear program rounding. This also has the advantage of yielding an algorithm for the case when the input distributions are integer valued over a small range – the approximation ratio in this case depends only on the range and not on the value,  $V$ . We finally show that if each distribution is uniform, the approximation on the cost is  $O(\log \frac{1}{\epsilon})$  in order to obtain a  $(1 + \epsilon)$  approximation on the minimum value. The approximation factor therefore becomes *independent* of the parameters of the distribution.

## 5.1 CIP-GREEDY Algorithm

We first consider the case where the  $X_i$  are discrete distributions over  $m$  values. We present a  $O(\log m)$  approximation (on the cost) by combining the greedy algorithm with column monotone CIPs. Let the number of distinct values taken by the discrete distributions be  $m$ , corresponding to the intervals  $I_1, I_2, \dots, I_m$ . Let  $l_j$  denote the length of the interval  $I_j$ . Let  $X^*$  denote the value of the optimal solution to MINIMUM-ELEMENT. Let  $a_{ji} = \log \frac{1}{p_{ij}}$ . The CIP-GREEDY algorithm runs in the following three steps.

**Step 1:** The first step of the algorithm defines the following integer program:

$$\begin{aligned} & \text{Minimize} && z \\ & \sum_{i=1}^n y_i a_{ji} &\geq & \log l_j - z \quad \forall j = \{1, 2, \dots, m\} \\ & \sum_{i=1}^n c_i y_i &\leq & C \\ & y_i &\in & \{0, 1\} \quad \forall i \in \{1, 2, \dots, n\} \end{aligned}$$

**Claim 5.1.**  $z = \log X^*$  is feasible for the above IP.

*Proof.* Consider the optimal solution  $S$  to MINIMUM-ELEMENT and set  $y_i = 1$  if  $i \in S$  and 0 otherwise. Clearly, we have  $\sum_i c_i y_i \leq C$ . Furthermore, the IP constraints together with the definition of  $X^*$  implies the following:

$$X^* = \mathbf{E}[\min_{i \in S} X_i] = \sum_{j=1}^m l_j \left( \prod_i p_{ij}^{y_i} \right) \geq \max_j \left( l_j \left( \prod_i p_{ij}^{y_i} \right) \right)$$

Taking logs, we have

$$\log X^* \geq \max_j \left( \log l_j - \sum_i y_i a_{ji} \right)$$

Therefore  $z = \log X^*$  satisfies all the constraints of the LP.  $\square$

The next claim follows using the same proof argument as the previous claim.

**Claim 5.2.** Consider any feasible solution to the above IP with  $z \leq \log X^*$ . Then the subset  $S$  corresponding to  $i$  s.t.  $y_i = 1$  is feasible and has  $\mathbf{E}[\min_{i \in S} X_i] \leq mX^*$ .

**Step 2:** Solve the linear relaxation of the above IP. Suppose the optimal solution has value  $z^*$ . From Claim 5.1, note that  $z^* \leq \log X^*$ . Re-write the IP as the following CIP:

$$\begin{aligned} & \text{Minimize} && \sum_i c_i y_i \\ & \sum_{i=1}^n y_i a_{ji} &\geq & \log l_j - z^* \quad \forall j = \{1, 2, \dots, m\} \\ & y_i &\in & \{0, 1\} \quad \forall i \in \{1, 2, \dots, n\} \end{aligned}$$

Note that if the  $y_i$  are allowed to be fractional in  $[0, 1]$ , then there exists a solution to the above CIP with objective value at most  $C$ . We now use the following proposition:

**Proposition 5.3.** [6, 24]. If a CIP with  $m$  constraints has a feasible solution, we can find a solution with approximation ratio  $O(\log m)$ .

The above implies that there is a solution to the above CIP which respects all the constraints and which has objective value (or cost)  $O(C \log m)$ . From Claim 5.2, this integer solution corresponds to a subset  $S^0$  such that  $\mathbf{E}[\min_{i \in S^0} X_i] \leq mX^*$ .

**Step 3:** Run the greedy algorithm in Figure 2 with initial solution  $S^0$  and with cost budget  $\tilde{C} = C(\log \log m + \log \frac{1}{\epsilon})$ . Let the final solution be  $S^f$ .

**Claim 5.4.**  $\mathbf{E}[\min_{i \in S^f} X_i] \leq (1 + \epsilon)X^*$ .

*Proof.* We repeat the proof of Theorem 4.4 with  $g(\Phi) = g(S^0) = -\log(mX^*)$  and  $g(S^f) = -\log((1 + \epsilon)X^*)$ .  $\square$

We finally have the following theorem:

**Theorem 5.5.** *For discrete distributions on  $m$  values, the CIP-GREEDY algorithm achieves a  $O(\log m + \log \frac{1}{\epsilon})$  approximation on the cost in order to find a solution of value at most  $(1 + \epsilon)X^*$ .*

## 5.2 Algorithm for Polynomial Input Domain

We now improve the above result when the domain of the distributions is the set  $\{0, 1, \dots, M-1\}$ . Assume w.l.o.g. that  $M$  is a power of 2. We first group the intervals so that the lengths are increasing in powers of 2, and the first interval has length 1. For each distribution  $X_i$ , construct a new distribution  $\tilde{X}_i$  over the domain  $R = \{0, 1, 2, 4, \dots, 2^{\log M}\}$  as follows: Set  $\tilde{p}_{i0} = p_{i0}$ , and for  $k \in R \setminus \{0\}$ , set  $\tilde{p}_{ik} = \sum_{j=\lceil k/2 \rceil}^k p_{ij}$ . Let  $X^*$  denote the value of MINIMUM-ELEMENT for the distributions  $X_i$  and let  $Y^*$  denote the corresponding value for the distributions  $\tilde{X}_i$ . It is easy to show that  $X^* \leq Y^* \leq 2X^*$ .

The algorithm is as follows: Run Steps (1) and (2) of the CIP-GREEDY algorithm using the distributions  $\tilde{X}_i$ . Setting  $m = \log M$ , it is easy to see that this produces a solution  $S^0$  so that  $\mathbf{E}[\min_{i \in S^0} X_i] \leq 2X^* \log M$  using cost  $O(C \log \log M)$ . Now run the greedy algorithm in Figure 2 using starting solution  $S^0$ , and using the original distributions  $X_i$ , with budget set to  $\tilde{C} = C(\log \log \log(2M) + \log \frac{1}{\epsilon})$ . Using the same proof as Theorem 4.4, this yields a solution of value  $(1 + \epsilon)X^*$ . We therefore have the following theorem:

**Theorem 5.6.** *In the MINIMUM-ELEMENT problem, when the domain of values is  $\{0, 1, \dots, M-1\}$ , then the approximation ratio (on the cost) of the modified CIP-GREEDY is  $O(\log \log M + \log \frac{1}{\epsilon})$ , and this achieves a solution of value  $(1 + \epsilon)X^*$ .*

**Remarks.** An interesting open question is to improve the above approximation factor to  $O(1)$  on the cost. The super-constant approximation ratio is in sharp contrast with the lack of even a NP-HARDNESS result for the case where  $M$  is polynomially bounded. (Note that the NP-HARDNESS proof required exponentially large values in the domain.)

Next, note that in both the above cases, the bottleneck in the approximation ratio is in solving the covering program. In particular, replacing the GREEDY algorithm in each case by the *naive* greedy algorithm that adds the variable  $\arg \min_i \frac{f(S \cup \{X_i\}) - f(S)}{c_i}$  at each step, would also yield the same approximation ratio. This strongly suggests that the  $m$ -constraint COLUMN-MONOTONE CIPs that arise in these settings have a better approximation ratio than  $O(\log m)$ . Note that we have only been able to show NP-HARDNESS for these problems, and the hardness of approximation proofs for general CIPs do not carry over to COLUMN-MONOTONE CIPs.

### 5.3 Uniform Distributions

We now show that a slightly modified greedy algorithm actually performs much better when each distribution  $X_i$  is uniform in the range  $[a_i, b_i]$ . The problem with using the greedy algorithm directly is that initially, the algorithm could make a sequence of wrong choices which are costly to rectify. We show that if the distributions are truncated near the optimal solution value, such a problem cannot arise, and therefore the greedy algorithm performs much better. To get around the issue of not knowing the optimal solution value, we try all possible truncations in powers of 2. In order to describe the new algorithm, we first define the *truncation* of a distribution:

**Definition 5.7.**  $X^t$  is a truncation of  $X$  at point  $t$  if  $\Pr[X^t = t] = \Pr[X \geq t]$ , and  $\Pr[\tilde{X} = r] = \Pr[X = r]$  for  $r < t$ .

The algorithm is presented in Figure 3. It clearly has polynomial running time since it tries values of  $t$  in powers of 2, so that the number of tries is polynomial in the bit complexity of the input. As before, let  $S^*$  denote the optimal solution and  $X^* = \mathbf{E}[\min_{i \in S^*} X_i]$ . We will show the following theorem:

**Theorem 5.8.** For  $\epsilon \leq 0.1$ , the modified GREEDY algorithm yields a solution of value  $(1 + 9\epsilon)X^*$  for uniform distributions using cost  $2C \log \frac{1}{\epsilon}$ .

MODIFIED GREEDY ( $C, \epsilon$ )      /\*  $\epsilon \leq 0.1$  \*/

$\tilde{C} \leftarrow 2C \log \frac{1}{\epsilon}$   
 $t \leftarrow \min_i \mathbf{E}[X_i] / \epsilon$   
**While**  $t \geq \mathbf{E}[\min_i X_i]$  **do**:  
     $S^t \leftarrow$  GREEDY solution on  $X_1^t, \dots, X_n^t$  with cost bound  $\tilde{C}$ .  
     $t \leftarrow t/2$ .  
**endwhile**  
Output  $S^t$  with minimum  $\mathbf{E}[\min_{i \in S^t} X_i]$ .

Figure 3: Modified Greedy Algorithm for Uniform Distributions.

The crux of the proof is the following lemma, which in effect states that truncation around  $X^*/\epsilon$  preserves the expected minimum of all solutions whose original minimum was close to  $X^*$ .

**Lemma 5.9.** Let  $t = \alpha \frac{X^*}{\epsilon}$  for  $\alpha \in [1/2, 1]$ . For any set  $S$  with  $\mathbf{E}[\min_{i \in S} X_i] = q \geq X^*$ , either  $\mathbf{E}[\min_{i \in S} X_i^t] \geq 1.2X^*$  or  $\mathbf{E}[\min_{i \in S} X_i^t] \geq (1 - 7\epsilon)q$ .

We prove this lemma later. Now, using this lemma, we complete the proof of Theorem 5.8.

*Proof.* (of Theorem 5.8) During some point in the execution of the algorithm, we will have  $t = \alpha \frac{X^*}{\epsilon}$  for  $\alpha \in [1/2, 1]$ . Consider this value of  $t$ . For any set  $S$  with cost at most  $C$ , we have  $\mathbf{E}[\min_{i \in S} X_i] \geq X^*$ . Applying the previous lemma, it is clear that  $\mathbf{E}[\min_{i \in S} X_i^t] \geq (1 - 7\epsilon)X^*$ .

Note now that  $\mathbf{E}[X_i^t] \leq X^*/\epsilon$  for all  $i$ . Therefore, repeating the proof of Theorem 4.4 with  $g(\Phi) = -\log(X^*/\epsilon)$ ,  $g(S^*) = -\log((1 - 7\epsilon)X^*)$ , and  $g(S^t) = -\log((1 + \epsilon)X^*)$  yields  $\tilde{C} \leq C(\log \log \frac{1+7\epsilon}{\epsilon} + \log \frac{1}{\epsilon}) \leq 2C \log \frac{1}{\epsilon}$  for  $\epsilon = 0.1$ . This yields a set  $S^t$  such that  $\mathbf{E}[\min_{i \in S^t} X_i^t] \leq (1 + \epsilon)X^* < 1.2X^*$  for  $\epsilon = 0.1$ .

Now, either  $\mathbf{E}[\min_{i \in S^t} X_i] \leq X^*$ , in which case we are done; or we apply the previous lemma to show that  $\mathbf{E}[\min_{i \in S^t} X_i] \leq (1 + 7\epsilon)\mathbf{E}[\min_{i \in S^t} X_i^t] \leq (1 + 7\epsilon)(1 + \epsilon)X^* \leq (1 + 9\epsilon)X^*$ .  $\square$

**Proof of Lemma 5.9.** Let  $X_i = \text{Unif}[a_i, b_i]$ . Now define a new r.v.  $Y_i = a_i + \text{Exponential}(1/(b_i - a_i))$ , which is  $a_i$  plus an exponential distribution with rate  $\lambda_i = 1/(b_i - a_i)$ . Define:

$$G_S(r) = \Pr[\min_{j \in S} Y_j \geq r] = \prod_{j \in S: a_j < r} \exp((a_j - r)\lambda_j)$$

**Lemma 5.10.** *For all  $r$ , we have:*

$$G_S(r) \geq \frac{\int_{x=r}^{\infty} G_S(x) dx}{\int_{x=0}^{\infty} G_S(x) dx}$$

*Proof.* Let  $S = \{1, 2, \dots, k\}$  and let  $a_1 \leq \dots \leq a_k$  and set  $a_{k+1} = \infty$ . We can assume without loss of generality that all  $a_j < r$ . If not, set  $\lambda_j = 0$  if  $a_j > r$  and prove as below. Now, setting  $\lambda_j$  back to its original value reduces both the numerator and denominator of the RHS of the lemma by the same amount, and hence only reduces the RHS. Therefore, the claim will still hold for the original  $\lambda_j$ . We therefore assume  $a_j < r$  for all  $j = 1, 2, \dots, k$ .

For notational convenience, denote  $G_S$  by  $G$ . First, we have:

$$\int_{x=r}^{\infty} G(x) dx = \frac{\exp(\sum_{j=1}^k (a_j - r)\lambda_j)}{\sum_{j=1}^k \lambda_j} = \frac{G_S(r)}{\sum_{j=1}^k \lambda_j}$$

By piece-wise integration, we also have:

$$\int_{x=0}^{\infty} G(x) dx = a_1 + \sum_{j=1}^k \left( \frac{\exp(\sum_{i=1}^j a_i \lambda_i) (\exp(-a_j \sum_{i=1}^j \lambda_i) - \exp(-a_{j+1} \sum_{i=1}^j \lambda_i))}{\sum_{i=1}^j \lambda_i} \right)$$

To show the lemma, we thus have to prove:

$$h(a_1, a_2, \dots, a_k) = a_1 \sum_{j=1}^k \lambda_j + \sum_{j=1}^k \left( \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^j \lambda_i} \cdot e^{\sum_{i=1}^j a_i \lambda_i} (e^{-a_j \sum_{i=1}^j \lambda_i} - e^{-a_{j+1} \sum_{i=1}^j \lambda_i}) \right) \geq 1$$

We will argue that  $h(\cdot)$  achieves its minimum is when  $a_1 = a_2 = \dots = a_k = 0$ . First:

$$\frac{\partial h}{\partial a_k} = \exp\left(\sum_{i=1}^{k-1} a_i \lambda_i\right) \exp\left(a_k \sum_{i=1}^{k-1} \lambda_i\right) \frac{\lambda_k \sum_{i=1}^{k-1} \lambda_i}{\sum_{i=1}^k \lambda_i} \geq 0 \text{ if } a_k \geq 0$$

Therefore,  $h(\cdot)$  is minimized by setting  $a_k = a_{k-1}$ . For this setting, we can inductively argue that  $h(\cdot)$  is minimized by setting  $a_{k-1} = a_{k-2}$ , and so on. This argument finally sets all the  $a_i$  equal to  $a_1$ . For this setting,  $h$  is clearly minimized when  $a_1 = 0$ . In this case, we have  $h = 1$ .  $\square$

**Lemma 5.11.** *For every set  $S$ , we have:  $\mathbf{E}[\min_{i \in S} X_i] \geq 0.324 \times \mathbf{E}[\min_{i \in S} Y_i]$ .*

*Proof.* Let  $S = \{1, 2, \dots, k\}$  and let  $a_1 \leq \dots \leq a_k$ . Let  $X = \min(X_1, X_2, \dots, X_k)$  and  $Y = \min(Y_1, Y_2, \dots, Y_k)$ . Note that  $Y_i$  stochastically dominates  $X_i$  for all  $i$ , and hence  $Y$  stochastically dominates  $X$ . Let  $F(x) = \Pr[X \geq x]$ . The stochastic domination implies  $G(x) \geq F(x)$  for all  $x$ .

Let  $\lambda = \sum_{i=1}^k \lambda_i$ . Let  $r = \frac{1}{2\lambda}$ . First, suppose  $a_1 = a_2 = \dots = a_k = 0$ , so that  $\lambda_i = 1/b_i$ . It is easy to show that the area under  $G(x)$  in the range  $x \in [0, r]$  is:

$$\int_{x=0}^r G(x) dx = \mathbf{E}[Y] \left(1 - \frac{1}{\sqrt{e}}\right)$$

Now, for any  $x \in [0, r]$ , the ratio of  $F(x)$  to  $G(x)$  is:

$$\frac{F(x)}{G(x)} = \frac{\prod_{i=1}^k (1 - x\lambda_i)}{\exp(-x \sum_i \lambda_i)}$$

This is a decreasing function of  $x$ , so that the minimum is achieved when  $x = r$ . At this point:

$$\frac{F(r)}{G(r)} = \frac{\prod_{i=1}^k (1 - \frac{\lambda_i}{2 \sum_i \lambda_i})}{\exp(-0.5)} \geq \frac{\sqrt{e}}{2}$$

Since  $F(x) \geq G(x) \frac{\sqrt{e}}{2}$  for all  $x \in [0, r]$ , we have:

$$\mathbf{E}[X] = \int_{x=0}^{\infty} F(x) dx \geq \int_{x=0}^r F(x) dx \geq \frac{\sqrt{e}}{2} \int_{x=0}^r G(x) dx = \frac{\sqrt{e}}{2} \mathbf{E}[Y] \left(1 - \frac{1}{\sqrt{e}}\right) = 0.324 \mathbf{E}[Y]$$

Suppose now that the  $a_i$  are general. Intuitively, this case should only be better. Formally, we start with all  $a_i = 0$ . For every distribution  $X_i$  and  $Y_i$  in turn, we increase the  $a_i$  to its correct value and consider the ratio  $\frac{\mathbf{E}[X]}{\mathbf{E}[Y]}$ . We claim that the ratio only goes up. Let  $f_i(x) = \Pr[X_i \geq x]$  and  $g_i(x) = \Pr[Y_i \geq x]$ . We have  $F(x) = \prod_{i=1}^k f_i(x)$ , and  $G(x) = \prod_{i=1}^k g_i(x)$ . Consider the distributions  $X_1, Y_1$ , and suppose  $a_1$  is increased from 0 to  $q$ . Note that the old  $f_1(x)/g_1(x)$  is the new  $f_1(x+q)/g_1(x+q)$ . Since  $f_1(x)/g_1(x)$  is non-increasing in  $x$  as shown above, when  $a_1$  is increased, for a given  $x$ , the ratio  $h_1(x) = f_1(x)/g_1(x)$  does not decrease. Further, as  $a_1$  is increased, both  $f_1(x)$  and  $g_1(x)$  do not decrease. Now consider:

$$\frac{\mathbf{E}[X]}{\mathbf{E}[Y]} = \frac{\int h_1(x) g_1(x) A(x) dx}{\int g_1(x) B(x) dx}$$

where  $A(x) = \prod_{i=2}^k f_i(x)$ , and  $B(x) = \prod_{i=2}^k g_i(x)$ , which do not change as  $a_1$  is increased. As  $a_1$  is increased, first, fix  $h_1(x)$ ; the ratio  $\frac{\mathbf{E}[X]}{\mathbf{E}[Y]}$  does not decrease as  $g_1(x)$  is non-decreasing. Since  $h_1(x)$  is also non-decreasing, the overall ratio cannot decrease. This proves the claim.  $\square$

**Lemma 5.12.** *Let  $X = \min_{i \in S} X_i$  and let  $F_S(r) = \Pr[X \geq r]$ . We have:*

$$\mathbf{E}[X] F_S(r) \geq 0.324 \int_{x=r}^{\infty} F_S(x) dx$$

*Proof.* Let  $Y = \min_{i \in S} Y_i$ . From Lemma 5.10, we have:

$$G_S(r) \geq \frac{\int_{x=r}^{\infty} G_S(x) dx}{\int_{x=0}^{\infty} G_S(x) dx} \quad \text{or} \quad \mathbf{E}[Y] \geq \frac{\int_{x=r}^{\infty} G_S(x) dx}{G_S(r)}$$

Combining this with the previous lemma:

$$\mathbf{E}[X] \geq 0.324 \mathbf{E}[Y] \geq 0.324 \frac{\int_{x=r}^{\infty} G_S(x) dx}{G_S(r)}$$

We showed in the previous lemma that  $\frac{F(x)}{G(x)}$  is a non-increasing function of  $x$ . Therefore,

$$\frac{F_S(r)}{G_S(r)} \geq \frac{\int_{x=r}^{\infty} F_S(x) dx}{\int_{x=r}^{\infty} G_S(x) dx}$$

Multiplying the above two inequalities proves the lemma.  $\square$

To complete the proof of Lemma 5.9, consider any set  $S$  and  $q = \mathbf{E}[\min_{i \in S} X_i] \geq X^*$ . Let  $t = \alpha \frac{X^*}{\epsilon}$  for  $\alpha \in [1/2, 1]$ . Define  $F_S$  as in Lemma 5.12. We split the analysis into two cases :

**Case 1:** If  $F_S(t) \geq 2.4\epsilon$ , then,

$$\mathbf{E}[\min_{i \in S} X_i^t] = \int_{x=0}^t F_S(x) dx \geq t F_S(t) \geq \alpha \frac{X^*}{\epsilon} 2.4\epsilon \geq 1.2X^*$$

where we have used  $\alpha \geq 1/2$ .

**Case 2:** If  $F_S(t) \leq 2.4\epsilon$ , by Lemma 5.12, we have  $\int_{x=t}^{\infty} F_S(x) dx \leq 3.1 \mathbf{E}[\min_{i \in S} X_i] F_S(t) \leq 7\epsilon q$ . Therefore,

$$\mathbf{E}[\min_{i \in S} X_i^t] = \int_{x=0}^t F_S(x) dx = \mathbf{E}[\min_{i \in S} X_i] - \int_{x=t}^{\infty} F_S(x) dx \geq q(1 - 7\epsilon)$$

This completes the proof of Lemma 5.9 and hence of Theorem 5.8.  $\square$

## 6 The Mixed Model

So far the only variables we were allowed to use in our solution were the ones that we observed. In general, our solution can use both probed and unprobed variables: If the minimum value among the probed set is larger than the expected value of a variable that has not been probed, we would prefer to use that variable as opposed to one of the probed values.

We show that the restriction of using only the probed set does not significantly alter the problem:

**Theorem 6.1.** *In order to achieve the same (or better) objective value for MINIMUM-ELEMENT, the solution that uses only probed variables probes at most one more variable than the solution that is allowed to use unprobed variables.*

*Proof.* Consider the optimal solution in the mixed model. Suppose it probes set  $S^*$  and let  $X^*$  denote the variable not in  $S^*$  with the smallest expectation. The strategy is to probe  $S^*$  and if the minimum value observed is larger than  $\mathbf{E}[X^*]$ , output  $X^*$ . The value of the solution is given by the expression  $\mathbf{E}[\min(\min_{Y \in S^*} Y, \mathbf{E}[X^*])]$ . Consider now the solution that probes  $S^* \cup \{X^*\}$ . The value of this solution is  $\mathbf{E}[\min(\min_{Y \in S^*} Y, X^*)]$ . It is easy to see that this value is smaller than the value of the optimal strategy for the mixed model.  $\square$

## 7 Conclusions

We have presented a framework (along with simple greedy algorithms) for studying the cost-value trade-off in resolving uncertainty based on the objective function being optimized. This paradigm will increasingly play a role in model-driven optimization in sensor networks and other complex distributed systems.

In the context of MINIMUM-ELEMENT our work presents interesting open questions. First, there is a huge gap between the lower bounds and approximation ratios we show. Can this gap be closed? In particular, can logarithmic hardness be shown for the general case, and NP-HARDNESS for the case where the domain is restricted to be poly-bounded. Furthermore, can the algorithms be extended to the case where the observations are adaptive, *i.e.*, based on the results of previous observations? As future work, we also plan to enhance the model with correlated random variables, other metrics measuring the trade-off (like the difference between value and cost), observing time-evolving processes, and optimizing more complex objective functions.

**Acknowledgments:** We thank Shivnath Babu, Utkarsh Srivastava, Sampath Kannan and Brian Babcock for helpful discussions.

## References

- [1] A. Akella, B. M. Maggs, S. Seshan, A. Shaikh, and R. K. Sitaraman. A measurement-based analysis of multihoming. In *ACM SIGCOMM Conference*, pages 353–364, 2003.
- [2] R. Avnur and J. M. Hellerstein. Eddies: Continuously adaptive query processing. In *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 261–272, 2000.
- [3] B. Babcock and S. Chaudhuri. Towards a robust query optimizer: A principled and practical approach. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 119–130, 2005.
- [4] B. Babcock and C. Olston. Distributed top-k monitoring. In *SIGMOD '03: Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 28–39, 2003.
- [5] S. Babu and P. Bizarro. Proactive reoptimization. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2005.
- [6] R. Carr, L. Fleischer, V. Leung, and C. Phillips. Strengthening integrality gaps for capacitated network design and covering problems. In *Proc. of the Annual ACM-SIAM Symp. on Discrete Algorithms*, 2000.
- [7] M. Charikar, R. Fagin, J. Kleinberg, P. Raghavan, and A. Sahai. Querying priced information. *Proc. of the Annual ACM Symp. on Theory of Computing*, 2000.
- [8] F. Chu, J. Halpern, and J. Gehrke. Least expected cost query optimization: What can we expect? *Proc. of the ACM Symp. on Principles of Database Systems*, 2002.
- [9] F. Chu, J. Halpern, and P. Seshadri. Least expected cost query optimization: An exercise in utility. In *Proc. of the ACM Symp. on Principles of Database Systems*, 1999.
- [10] B. Dean, M. Goemans, and J. Vondrák. Approximating the stochastic knapsack problem: The benefit of adaptivity. *Proc. of the Annual Symp. on Foundations of Computer Science*, 2004.
- [11] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *Proc. of the 2004 Intl. Conf. on Very Large Data Bases*, 2004.
- [12] T. Feder, R. Motwani, R. Panigrahy, C. Olston, and J. Widom. Computing the median with uncertainty. *SIAM J. Comput.*, 32(2), 2003.
- [13] A. Goel, S. Guha, and K. Munagala. Asking the right questions: Model-driven optimization using probes. In *Proc. of the 2006 ACM Symp. on Principles of Database Systems*, 2006.
- [14] A. Goel and P. Indyk. Stochastic load balancing and related problems. *Proc. of the Annual Symp. on Foundations of Computer Science*, 1999.
- [15] S. Guha and K. Munagala. Approximation algorithms for budgeted learning problems. In *Proc. ACM Symp. on Theory of Computing*, 2007.
- [16] S. Guha and K. Munagala. Model-driven optimization using adaptive probes. In *Proc. ACM-SIAM Symp. on Discrete Algorithms*, 2007.

- [17] S. Guha, K. Munagala, and S. Sarkar. Information acquisition and exploitation in multi-channel wireless systems. *Submitted to IEEE Transactions on Information Theory*, 2007.
- [18] P. K. Gummadi, H. V. Madhyastha, S. D. Gribble, H. M. Levy, and D. Wetherall. Improving the reliability of internet paths with one-hop source routing. In *6th Symposium on Operating System Design and Implementation (OSDI)*, pages 183–198, 2004.
- [19] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: Approximation algorithms for stochastic optimization. *Proc. of the Annual ACM Symp. on Theory of Computing*, 2004.
- [20] A. Gupta, R. Ravi, and A. Sinha. An edge in time saves nine: LP rounding approximation algorithms for stochastic network design. In *Proc. of the Annual Symp. on Foundations of Computer Science*, pages 218–227, 2004.
- [21] N. Immorlica, D. Karger, M. Minkoff, and V. Mirrokni. On the costs and benefits of procrastination: Approximation algorithms for stochastic combinatorial optimization problems. In *Proc. of the Annual ACM-SIAM Symp. on Discrete Algorithms*, 2004.
- [22] S. Khanna and W-C. Tan. On computing functions with uncertainty. In *Proc. of the ACM Symp. on Principles of Database Systems*, 2001.
- [23] J. Kleinberg, Y. Rabani, and É. Tardos. Allocating bandwidth for bursty connections. *SIAM J. Comput.*, 30(1), 2000.
- [24] S. Kolliopoulos and N. Young. Tight approximation results for general covering integer programs. *Proc. of the Annual Symp. on Foundations of Computer Science*, 2001.
- [25] A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. *Twenty-first Conference on Uncertainty in Artificial Intelligence (UAI 2005)*, 2005.
- [26] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Math Programming*, 14(1):265–294, 1978.
- [27] C. Olston. *Approximate Replication*. PhD thesis, Stanford University, 2003.
- [28] D. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as discrete optimization. *Proc. of the Annual Symp. on Foundations of Computer Science*, 2004.
- [29] A. Silberstein, R. Braynard, C. Ellis, K. Munagala, and J. Yang. A sampling based approach to optimizing top-k queries in sensor networks. In *Proc. of the Intl. Conf. on Data Engineering*, 2006.