Model-driven Optimization using Adaptive Probes

Sudipto Guha*   Kamesh Munagala†

Abstract

In several applications such as databases, planning, and sensor networks, parameters such as selectivity, load, or sensed values are known only with some associated uncertainty. The performance of such a system (as captured by some objective function over the parameters) is significantly improved if some of these parameters can be probed or observed. In a resource constrained situation, deciding which parameters to observe in order to optimize system performance itself becomes an interesting and important optimization problem. This problem is the focus of this paper. Unfortunately designing optimal observation schemes is NP-HARD even for the simplest objective functions, leading to the study of approximation algorithms.

One of the most important considerations in this framework is whether adaptivity is required for the observations. Adaptive observations introduce blocking or sequential operations in the system whereas non-adaptive observations can be performed in parallel. One of the important questions in this regard is to characterize the benefit of adaptivity for probes and observation.

We present general techniques for designing constant factor approximations to the optimal observation schemes for several objective functions which commonly arise in systems applications. The techniques we present also show constant factor upper bounds for the benefit of adaptivity of the observation schemes – we show that while probing yields significant improvement in the objective function, being adaptive about the probing is not beneficial beyond constant factors.

*Department of Computer and Information Sciences,University of Pennsylvania. Email: `sudipto@cis.upenn.edu`. Research supported in part by an Alfred P. Sloan Research Fellowship and by an NSF Award CCF-0430376.

†Department of Computer Science, Duke University, Durham NC 27708. Email: `kamesh@cs.duke.edu` Research supported in part by NSF CNS-0540347.

1 Introduction

Consider a measurement scenario such as sensor networks or spectrograms, where we have errors in estimation, but the distribution of the error is understood through models of the measuring instrument, historical data, regression, Kalman filters, etc. Suppose we have a few readings and measurements, and are asked to evaluate a function, e.g., the maximum temperature. Assume that we can repeat and refine a few of the measurements, i.e, "probe" a few sensors to make detailed measurements. Clearly it is inefficient for all nodes to make a detailed measurement. A natural question is which measurements do we refine, say if we wanted to refine only k of them and wanted to maximize the expected maximum value? In many situations, we can assume that the errors made by different instruments are independent of each other – even then the problem is NP–HARD [15].

The above is a typical example of a model driven optimization problem. This has gained significant currency in a variety of other research areas such as database query optimization and route selection in networks. In a database query optimization setting suppose the optimizer is presented with a set of sufficiently complicated and unrelated (independent) queries. The query-optimizer can estimate the resources needed by the queries from historical information, cached statistics, sampling of various sub-queries, or by performing inexpensive filters [3, 8, 9]. Subsequent to this estimation, the query optimizer schedules the tasks to optimize the throughput or the average completion times. However, this estimation process itself consumes resources such as time, network bandwidth, and space, and therefore the decision to choose the parameters to refine in estimate becomes a key optimization problem. A similar problem arises in networking, where the current state of multiple routes and servers can be probed and observed before deciding which route/server to use for a specific connection [1, 20, 19]. The above examples can also be extended to planning over a network, where we may wish to visit a sequence of nodes, about which we have imprecise information and can only refine a small number of them, to minimize total distance traveled – such problems arise in query processing in sensor networks [12].

The above problems of measurement refinement or optimizing independent query schedules can be formulated abstractly as follows. Let $X_1, X_2, \ldots, X_n$ be non-negative independent random variables, whose distributions are given as inputs (the distribution can be specified by samples and does not affect any result in this paper). There is a function $f(X_1, X_2, \ldots, X_n)$ on these variables which must be optimized. By spending probe cost c_i, the optimizer can find out the exact value of variable X_i. There is a budget $C(\geq \max_i c_i)$ on the total cost of observation. The probing policy chooses a subset of variables of total cost at most C to probe and observe. *Subsequent* to this, the optimization policy chooses a solution based on the outcome of the probes, which optimizes $\mathbf{E}[f(X_1, X_2, \ldots, X_n)|\{X_i, i \in S\}]$, where the expectation is over the distributions of the unobserved variables. We note that the solution is fixed *after* the outcomes of the probes, but *before* the realization of the unprobed variables. Therefore, given the set of probed variables and their outcomes, the solution is the same for *all* realizations of the unprobed variables. The final goal is to choose that S which optimizes $\mathbf{E}[f(X_1, X_2, \ldots, X_n)|S \text{ is probed}]$, where the expectation is now jointly over the outcomes of probes and the distributions of unprobed variables.

As an example, suppose $f = \max_i X_i$. Let $\mathbf{E}[X_i] = \mu_i$. If the probing policy chooses a subset S of variables to probe, after the probing, the optimization policy will choose that variable as the solution which corresponds to the maximum of $\max_{i \in S} X_i$ and $\max_{i \notin S} \mu_i$. Therefore, $\mathbf{E}[f|\{X_i, i \in S\}] = \max(\max_{i \in S} X_i, \max_{i \notin S} \mu_i)$, where the expectation on the LHS is over the outcomes of the unprobed variables. Note that the solution is chosen after $\{X_i, i \in S\}$ is probed and revealed, but before the values of $\{X_i, i \notin S\}$ is known. Averaging this over all outcomes of probing S, we have $\mathbf{E}[f|S \text{ is probed }] = g(S) = \mathbf{E}[\max(\max_{i \in S} X_i, \max_{i \notin S} \mu_i)]$. The goal is to design a probing policy that chooses S which maximizes $g(S)$, subject to the budget constraint $\sum_{i \in S} c_i \leq C$.

Several conceptual and systems issues arise immediately, the foremost of which is the benefit of being adaptive in the probing policy. The above formulation implicitly assumed that the probes are done in parallel upfront, *i.e.*, non-adaptively[1]. In contrast, adaptive observations are sequential, based on the outcomes of the previous probes. It is clear that adaptive strategies yield as good or better solutions than the non-adaptive counterpart. The first issue with an adaptive strategy is the complexity of expressing the solution; it is *a priori* not clear that the decision tree that encodes the optimum solution is polynomially bounded. The next interesting issue, which is a consequence of the adaptivity, is the budget. Note that while probing random variables, if a suitable value is found already, then the probing can be halted early, yielding significant cost savings. Thus it makes sense to assume that the budget is in expectation as well. This sets up two classes of problems, namely with *hard budgets* and with *soft budgets*, *i.e.*, budgets are in expectation.

In this paper, we design and analyze non-adaptive probing strategies for the large class of objective functions, and show that these strategies have a bounded *adaptivity gap*, i.e., there exists a non-adaptive strategy (as defined above) which is only a constant factor worse compared to the best adaptive probing strategy with soft budgets. Therefore for the problems we study, the non-adaptive and adaptive models with hard and soft budgets are all related to each other by constant factors. As a consequence, one of the main point we can establish is that although probing helps significantly, adaptive probing is no better than non-adaptive probing by more than a constant factor in many problems of interest. We focus on problems that have been considered in the known applications and literature to date, namely, knapsack, average completion time scheduling, metric clustering, spanning and Steiner trees. Furthermore, we demonstrate common techniques which we use in analyzing seemingly disparate problems.

Related Work: The adaptivity gap in the absence of probes has been considered in the literature earlier, notably in [10, 11, 30, 34] for Knapsack and scheduling problems. However, the model driven optimization problem is considerably different from the settings in those papers. In [10, 11, 30, 34] the optimum is allowed to decide on the next item to schedule based on the past, but once the next item is decided this is an irrevocable commitment. In contrast, in our problem, after probing we (as well as the optimum adversary) may choose *not to* include an item in the Knapsack, *arbitrarily (re)order* the schedule, come up widely different clustering depending on the outcome of the probed values. Hence, though we use linear programming based formulations to lower bound the adaptive strategy, we need completely different arguments and observation schemes in the absence of the irrevocable commitment. The classic stochastic optimization (non-adaptive, non-probing) versions of these scheduling problems were considered in [28, 16]. In an earlier paper [15], we presented non-adaptive probing schemes when the objective function f captures the minimum value and the Knapsack problems. In [19] we present adaptive probing strategies with soft budget constraints for the minimum value and maximum value objective functions . For correlated random variables, the problem of minimizing residual information is considered in [29].

The notion of refining uncertainty has been considered in an *adversarial setting* by several researchers [32, 13, 27, 4]. Here, 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 strong lower bounds arise. We also note that stochastic optimization problems were considered in [25, 21, 24, 33, 35, 22, 23]; these problems appear to be unrelated to adaptivity gaps.

1.1 Technical Contributions

We discuss several general technical issues for our model in Appendix A. We now outline our key contributions.

Packing problems: In this paper, we first focus on the adaptivity gap of KNAPSACK in the probing model. We show that the adaptivity gap for this problem is a constant factor, both when sizes are random

[1]Note that this notion of a non-adaptive policy is different from that in [10].

variables, and when profits are random variables. In a previous paper [15], we showed that in the probing model, the non-adaptive KNAPSACK problems can be reduced to a maximizing submodular functions, implying greedy constant factor approximation algorithms. However the same sub-modularity based proofs cannot be used to prove adaptivity gaps, since the adaptive optimal solution is a decision tree. One of the contributions of this paper (Section 2) is techniques for bounding the adaptivity gap by a constant factor (hence showing that the greedy non-adaptive algorithm can be used even in an adaptive setting).

Note that in our case, since the final solution involves a subset of the variables probed, the irrevocable decision LP bounds in [10] are rendered inapplicable. The irrevocable decision problems are quite different in nature; and to illustrate the difference in techniques, we consider the probing version of the stochastic knapsack problem [10] in Appendix C.

We show that a different way of formulating the problem yields the adaptivity gap – the *excess* of the profit to size ratio of a variable over average profit to size ratio of the non-adaptive optimal policy has a linear contribution to the adaptive optimal policy. This yields an approximate LP formulation which shows a constant factor adaptivity gap.

Scheduling and metric problems: For these problems, unlike packing problems such as knapsack, the solution has to be constructed over *all* the variables whether probed or unporbed. Now a different set of ideas are needed to obtain lower bounds. We show constant factor adaptivity gaps when the problems obey a "recombinant" property – for an arbitrary partitioning of the input variables, each of the parts have an induced solution with the property that the objective function values sum to no more than the respective quantities in the original problem; further, the two solutions can be combined without a significant increase in value of the resulting solution. Since encoding the value of the probed part is unwieldy, this partitioning is necessary to construct a solution that encodes just the unprobed part. The recombination shows a small adaptivity gap – the tricky part is to ensure that the interaction of the probed and unprobed parts can be bounded. This technique also yields non-adaptive algorithms in addition to the adaptivity gap proof.

For scheduling we consider the average completion time ($1||\sum_j w_j C_j$) and minimum makespan on identical machines; in both cases, the job sizes are random variables. We focus on the average completion time which illustrates the main issues (Section 3). A surprising feature of this algorithm is that *only the expected values* of job sizes are used in constructing the approximately optimal probing scheme. The makespan problem is discussed in Appendix E. For both these problems we design non-adaptive algorithm which shows that the adaptivity gap is a constant, while increasing the cost of probing by a small factor. The increase in probing cost is unavoidable in our technique which lower bounds the adaptive solutions by linear programs which have similar gaps. An interesting open question is to show a complexity result that the increase in budget is unavoidable.

We next consider metric problems. We assume that the input nodes are discrete distributions of polynomial specification over points of the metric space. For these problem a small but important added twist is needed – we need to reformulate the problem on a different but related metric. We consider MST (Steiner Trees and TSP) and the k-Median problem. We discuss the k-Median problem and the general setup for these problems in Section 4. The MST problem is discussed in Appendix D. For the k-median problem an interesting issue comes to fore, which does not appear for deterministic input. It is well known that over any space, there exists a k-median solution which uses the input points and is at most twice the optimum. For distributional input we show that the adaptivity gap for obtaining exactly k medians is polynomially large, short of a polynomial blowup in the probing cost. However, if we restrict all solutions to use fixed points in the metric space as medians (as opposed to declaring an input node which could possibly be a distribution over points as a median) then the gap disappears! This exposes an interesting contrast in the problem based on which points are allowed to be medians. We expect these definitional issues to be of independent interest as more problems with distributions as input are investigated.

2 Adaptivity Gap for KNAPSACK

We show that for KNAPSACK, the ability to observe adaptively has only a constant factor benefit over observing non-adaptively. We discuss the version of knapsack where the profits are random variables; the version where the sizes are random variables is presented in Appendix C.

Item i has profit which is an independent random variable X_i, size $s_i \leq B$, and probing cost c_i. The knapsack capacity is B, and the probing budget is $C \geq \max_i c_i$. We will assume that s_i, B are (possibly exponential in magnitude) integers. The probing policy adaptively probes a subset of items to determine their exact profit. Subsequently, an item selection policy chooses a subset of probed and unprobed items to place in the knapsack so that the expected profit (over the distributions of the unprobed profit values) is maximized. Note that the selection policy will simply use profit $\mathbf{E}[X_i]$ for unprobed item i, while for a probed item, it uses the exact profit. The goal is to design an adaptive probing policy that maximizes the expected profit, where the expectation is over the outcome of the probes and over the distribution of the profit values of the unprobed items.

In what follows, we will show an existence result that the adaptivity gap is a constant, meaning that there is a non-adaptive probing policy that upfront chooses a subset of items to probe, which is a constant factor approximation to the best adaptive probing policy. Note that the benefit of probing itself is enormous, because even for finding the maximum element (each element of size 1, knapsack capacity 1), probing all items yields profit $\mathbf{E}[\max_i\{X_i\}]$, which can be $\Omega(n)$ times $\max_i \mathbf{E}[X_i]$, which is the expected profit of probing no item. However being adaptive does not give us significantly more power.

Greedy Algorithm: We first describe a greedy non-adaptive algorithm presented in [15] which is a constant factor approximation to the optimal non-adaptive probing policy. We first observe that after the probes are completed, in any scenario, either half of the profit arises from the probed variables or half of the profit arises from the unprobed variables. Thus we can compute two solutions: The first where the item selection policy is restricted to using only unprobed items (in this case, no probing is necessary, and the optimal solution is the solution to the deterministic knapsack problem where the item profits are $\mathbf{E}[X_i]$), and the solution where the selection policy is restricted to using just the probed items. We choose the better of the two solutions, losing a factor 2 in the approximation ratio. Since the former solution is simple to construct, we focus on the latter solution which is restricted to using probed items.

Consider the optimal fractional profit of non-adaptively probing set S and selecting items fractionally after the outcome of the probes. This is given by $f(S) = \mathbf{E}[\max_{\vec{y}\geq 0, y_i \leq 1, \sum_{i\in S} s_i y_i \leq B} \sum_{i \in S} X_i y_i]$. It is shown in [15] that (i) $f(S)$ is sub-modular (ii) it can be computed (approximately) efficiently and (iii) $f(S)$ has profit at most 2 times the expected profit of the integral solution (y_i restricted to 0/1). The greedy algorithm for sub-modular function maximization [31] therefore yields a constant factor approximation to the profit of the best probed solution, which implies a constant factor approximation to the profit of the best *non-adaptive* probing policy.

We will now show that the greedy algorithm designed in [15] is not only a $O(1)$ approximation to the best non-adaptive policy (as shown in [15]), but that it is also a constant factor approximation to the best *adaptive* observation policy! Our techniques for proving this "adaptivity gap" are completely different from the techniques in [15], since the sub-modularity based proof in [15] breaks down when a non-adaptive policy is analyzed against an adaptive policy. Our "adaptivity gap" proof holds against the optimal adaptive policy with either hard or soft budgets.

The adaptivity gap: Let f_X denote the probability density function (p.d.f.) of random variable X. Let X^* denote the fractional profit of the non-adaptive optimal solution and let S^* denote the set probed. That is, $X^* = \max_{S | \sum_{i \in S} c_i \leq C} f(S)$. The average profit to size ratio of the non-adaptive solution is $r^* = X^*/B$.

Definition 1. *Define $\mathbf{E}_q[X] = \int_{y=q}^{\infty} y f_X(y) dy$. This denotes the "excess" of random variable X above q.*

We first begin with a useful lemma showing that the expected maximum of a set of random variables is approximately linear in their "excess". The proof is in Appendix B.

Lemma 2.1. *(Implicit in [28]) For a set of independent random variables $S = \{X_1, X_2, \ldots X_k\}$, let $t = \mathbf{E}[\max_{i \in S} X_i]$, then $\sum_{i \in S} \mathbf{E}_{2t}[X_i] \le 2t$.*

Definition 2. *Define $R_i = X_i/s_i$. This random variable corresponds to the profit per unit size of item i.*

For a subset S of items, recall: $f(S) = \mathbf{E}[\max_{\vec{y} \ge 0, y_i \le 1, \sum_{i \in S} s_i y_i \le B} \sum_{i \in S} X_i y_i]$. Define $r(S) = f(S)/B$. The next lemma is the main technical lemma. We believe the techniques used in the proof of the lemma would be of interest to other packing problems as well.

Lemma 2.2. *For any subset S of items, $\sum_{i \in S} s_i \mathbf{E}_{2r(S)}[R_i] \le 5f(S)$.*

Proof. Suppose $\sum_{i \in S} s_i \mathbf{E}_{2r(S)}[R_i] > 5f(S)$. Define $L = 2f(S)/B = 2r(S)$. We have $\sum_{i \in S} s_i \mathbf{E}_L[R_i] > \frac{5}{2}BL$. We will show this implies $f(S) > BL/2$, which is a contradiction. In the rest of the proof we will show that if $\sum_{i \in S} s_i \mathbf{E}_L[R_i] > \frac{5}{2}BL$, then $f(S) > BL/2$.

We will show the above by creating feasible fractional solutions. For each i create s_i perfectly correlated copies of the random variable R_i. Denote these copies $R_{i1}, \ldots, R_{is_i}$. Let $\mathcal{R} = \cup_{i \in S} \cup_{j=1}^{s_i} R_{ij}$. Each variable R_{ij} correspond to a size unit. Note that we will be using these variables to show proof of existence only – the possibly exponential number of variables will not affect the algorithm. The fractional knapsack now corresponds to accommodating B of these variables. Let $\mathcal{R}_0$ denote the set of variables R_{ij} such that $\mathbf{E}_L[R_{ij}] > L$. Let $\mathcal{R}_1$ denote the set of variables R_{ij} such that $\mathbf{E}_L[R_{ij}] \le L$. We have two cases to consider: either $\sum_{R_{ij} \in \mathcal{R}_0} \mathbf{E}_L[R_{ij}] > BL/2$ or $\sum_{R_{ij} \in \mathcal{R}_1} \mathbf{E}_L[R_{ij}] > 2BL$.

Case 1, $\sum_{R_{ij} \in \mathcal{R}_0} \mathbf{E}_L[R_{ij}] > BL/2$: Suppose $|\mathcal{R}_0| \ge B$, then we can choose a subset of variables of cardinality exactly B, and name them P. Otherwise $P = \mathcal{R}_0$. The variables chosen in the subset P define a fractional knapsack solution. But for this fixed fractional solution $\mathbf{E}\left[\sum_{R_{ij} \in P} R_{ij}\right] = \sum_{R_{ij} \in P} \mathbf{E}[R_{ij}] \ge \sum_{R_{ij} \in P} \mathbf{E}_L[R_{ij}] > BL/2$. Now $f(S)$ will only do better than this fixed solution – because of the $\max$. Thus in this case we already have $f(S) > BL/2$.

Case 2, $\sum_{R_{ij} \in \mathcal{R}_1} \mathbf{E}_L[R_{ij}] > 2BL$: Split the knapsack into unit size bins $G_1, G_2, \ldots, G_B$ which are initially empty. Consider the variables $R_{ij} \in \mathcal{R}_1$ in increasing order of $\mathbf{E}_L[R_{ij}]$ (so that the variables corresponding to any item i are considered consecutively), and assign them to the bins in a round-robin fashion so that each item is assigned to the next bin in the round-robin order.

Consider the sum $\sigma_k = \sum_{R_{ij} \in G_k} \mathbf{E}_L[R_{ij}]$. Since the variables are being assigned in a round-robin fashion, we have $\max_{k,k'} |\sigma_k - \sigma_{k'}| \le L$ because after the first item is placed in $k' < k$, all subsequent pairs of items placed decrease the difference; an odd number helps lessen the difference even more. We also have $\sum_{k=1}^{B} \sigma_k > 2BL$. Therefore, $\sigma_k > L$ for all $k = 1, 2, \ldots, B$.

Since there are at most B variables $R_{ij} \in \mathcal{R}_1$ corresponding to any item i, the round robin scheme also ensures that the variables in a bin G_k correspond to distinct items. Thus the variables in a *single fixed bin* G_k are independent. For a fixed G_k, let $t_k = \mathbf{E}[\max_{R_{ij} \in G_k} R_{ij}]$. By Lemma 2.1, we have $\mathbf{E}[\max_{R_{ij} \in G_k} R_{ij}] \ge \frac{1}{2} \sum_{R_{ij} \in G_k} \mathbf{E}_{2t_k}[R_{ij}]$. Now if $2t_k \le L$ then we know that

$$\sum_{R_{ij} \in G_k} \mathbf{E}_{2t_k}[R_{ij}] \ge \sum_{R_{ij} \in G_k} \mathbf{E}_L[R_{ij}] = \sigma_k > L$$

. Therefore $t_k = \mathbf{E}[\max_{R_{ij} \in G_k} R_{ij}] > L/2$. Note that if $2t_k > L$ then $t_k > L/2$ trivially. Consider a fractional solution where each bin G_k chooses the maximum and the profit is $\max_{R_{ij} \in G_k} R_{ij}$. We have:

$$\max_{\vec{y} \ge 0, y_i \le 1, \sum_{i \in S} s_i y_i \le B} \sum_{i \in S} X_i y_i \ge \sum_{k=1}^{B} \max_{R_{ij} \in G_k} R_{ij}$$

Taking expectation, the left hand side yields $f(S)$. The right hand side by linearity of expectation is $\sum_{k=1}^{B} t_k > BL/2$. This proves the lemma. $\square$

Recall that $X^* = \max_{S|\sum_{i\in S} c_i \le C} f(S)$, S^* denotes the corresponding probed set, and $r^* = X^*/B$.

Lemma 2.3. $\sum_{i\in S} s_i \mathbf{E}_{2r^*}[R_i] \le 5X^*$ *for any set* S *with* $\sum_{i\in S} c_i \le C$.

Proof. Since $f(S) \le X^*$ and $r(S) \le r^*$, from the previous lemma we have:

$$\sum_{i\in S} s_i \mathbf{E}_{2r^*}[R_i] \le \sum_{i\in S} s_i \mathbf{E}_{2r(S)}[R_i] \le 5f(S) \le 5X^*$$

$\square$

We now show that the optimal non-adaptive solution yields a constant factor approximation to the optimal adaptive strategy with either hard or soft budgets.

Theorem 2.4. *The expected value of the non-adaptive optimal observation strategy is a constant factor approximation to the value of the optimal adaptive strategy.*

Proof. Consider any adaptive algorithm. Let Z_i be random variable denoting whether X_i is observed, and let $z_i = \mathbf{E}[Z_i]$. Since the expected probe cost is at most C, we have $\sum_i c_i z_i \le C$.

Define $Y_i = X_i$ if $\frac{X_i}{s_i} \ge 2r^*$, and 0 otherwise. Therefore, $\mathbf{E}[Y_i] = s_i \mathbf{E}_{2r^*}[R_i]$. Let $\mathcal{Y} = \sum_i Z_i Y_i$ be the sum of profits of probed variables whose profit to size ratio is larger than $2r^*$. By the independence of Z_i and X_i we have, $\mathbf{E}[\mathcal{Y}] = \sum_i z_i s_i \mathbf{E}_{2r^*}[R_i]$. The z_i values therefore define a fractional knapsack instance. The best $0/1$ setting of the z_i yields $\sum_i z_i s_i \mathbf{E}_{2r^*}[R_i] \le 5X^*$ by the previous lemma. Since the best fractional solution can have at most twice this value (assuming all $c_i \le C$), we have $\mathbf{E}[\mathcal{Y}] \le 10X^*$.

Let OPT denote the value of the adaptive optimal solution. The total profit it can derive from items with profit to size ratio at most $2r^*$ is at most $2X^*$. Therefore, $OPT \le 2X^* + \mathbf{E}[\mathcal{Y}]$. Therefore, the expected value of the adaptive optimal solution is at most $12X^*$. $\square$

3 Average Completion Time Scheduling

We consider the weighted completion time problem ($1||\sum w_j C_j$) of scheduling jobs on a single processor to minimize the sum of the weighted completion times. All jobs are released at time $t = 0$ and there are no deadlines or precedence constraints.

In the probing model the sizes (or processing times) of jobs $J_1, \ldots, J_n$ are distributed according to independent random variables $X_1, X_2, \ldots, X_n$ respectively. The weight of job J_i is w_i, which is *not* a random variable. Let $\mathbf{E}[X_i] = \mu_i$. Each variable X_i corresponding to the size of job J_i has probing cost c_i; probing yields its exact value. Let C denote the (soft) budget on probing cost.

The solution is a strategy for adaptively probing a subset the jobs so that the expected weighted completion time of scheduling *all* the jobs after the outcome of the probes is known, is minimized. This expectation is over the outcome of the probes, and over the distribution of the processing times of the unprobed jobs. We note that the scheduling policy fixes the ordering of all jobs after the results of the probes are known, but before the sizes of the unprobed jobs are revealed. Therefore, the scheduling policy will simply order the jobs in non-decreasing order of the ratio of processing time to weight (*Smith's Rule*) – this processing time is exactly known for probed jobs, and is the expected processing time for unprobed jobs.

Roadmap: We show that there is a one-shot (non-adaptive) probing scheme which is a constant factor approximation to the optimal adaptive probing scheme. We will write an LP over job *pairs* to bound the contribution of only the unprobed part of the solution. Therefore for the variables we do not probe, their contribution would be a lower bound of the optimum. We show via a recombinant property in the problem structure that this lower bound is sufficient to achieve a constant factor non-adaptive approximation to the optimal adaptive solution.

The Recombinant Property: First consider n *deterministic* jobs. Let $l_1, l_2, \ldots, l_n$ denote the job-lengths, and $w_1, w_2, \ldots, w_n$ denote the job weights. Let $\alpha_i = \frac{l_i}{w_i}$. By Smith's rule, the optimal solution sorts the jobs in increasing order of α_j and schedules in this order. The completion time of the optimal ordering can therefore be written as $\sum_{i=1}^{n} \sum_{j \geq i} w_i w_j \min(\alpha_i, \alpha_j)$. This summation is over *job pairs*. We now show an important *recombinant property* of this summation. The proof is in Appendix B.

Lemma 3.1. *For any partitioning of n deterministic jobs into two disjoint sets A and B, we have:*

$$\sum_{i,j \in A, j \geq i} w_i w_j \min(\alpha_i, \alpha_j) + \sum_{i,j \in B, j \geq i} w_i w_j \min(\alpha_i, \alpha_j) \geq \sum_{i \in A, j \in B} w_i w_j \min(\alpha_i, \alpha_j)$$

In the above lemma, the LHS represents the contribution to the optimal completion time which arises from job pairs *within* A and within B. The RHS represents contributions of job pairs such that one of the jobs is in A and the other in B. The above shows that the interaction term across the two sides of any partition, can be bounded by the sum of the interactions within each side. This property will be crucial in our analysis. Though our analysis is for stochastic processing times (while the above lemma is for deterministic processing times), we will apply this lemma sample by sample to achieve the desired bounds.

3.1 Algorithm and Adaptivity Gap

We now show that the above recombinant property implies the following: The adaptivity gap of a non-adaptive algorithm based on solving a simple non-stochastic *outlier* version of the problem is a constant factor. In this outlier version, we first assume all jobs have (deterministic) processing times equal to their expected values, and choose a subset of jobs of total probing cost at most C to be "ignored" so as to minimize the weighted completion time of scheduling the remaining jobs. Note that no scheduling decisions are made yet; the goal of this step is to simply compute the subset to "ignore". Next, the probing policy probes the subset of jobs which we "ignored" in the previous step. Finally, the scheduling policy schedules all the probed and unprobed jobs using Smith's Rule.

Algorithm. More formally, the following non-adaptive algorithm has a constant adaptivity gap. The LP in the first step encodes the outlier version of the problem, where all jobs have processing times equal to their expected values, and the goal is to discard a set of jobs of cost at most C such that the weighted completion time of the remaining jobs is minimized.

1. Define the following LP and solve it

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sum_{j > i} e(i,j) \min\{w_j \mu_i, w_i \mu_j\} + \sum_{i=1}^{n} (1 - z_i) w_i \mu_i \\ & \sum_{i=1}^{n} z_i c_i \leq C \\ & e(i,j) + z_i + z_j \geq 1 \;\; \text{for all } i, j > i \\ & e(i,j), z_i \geq 0 \;\; \text{for all } i, j \in \{1, 2, \ldots, n\} \end{aligned}$$

2. Round the LP solution as follows: Let $S_1 = \{i | z_i \geq \frac{1}{3}\}$, and $e(i,j) = 1$ if $i \notin S_1$ and $j \notin S_1$. Let S_2 be the set of jobs not in S_1. The set S_1 is the jobs which are "ignored" in the outlier solution.

3. **Probe** the jobs $\{J_i | i \in S_1\}$. Let $f_i = \frac{X_i}{w_i}$ if $i \in S_1$ (where X_i is the realized size of J_i) and let $f_i = \frac{\mu_i}{w_i}$ for $i \in S_2$. Sort J_i in increasing order of f_i and output the schedule.

Analysis. Let OPT_{LP} denote the value of the optimal solution of the linear program. Let $ADAPT$ denote the value of the optimal adaptive policy. The next lemma is proved in Appendix B.

Lemma 3.2. $OPT_{LP} \leq ADAPT$.

The next lemma follows easily from the description of the rounding scheme.

Lemma 3.3. *The cost of probing set S_1 is at most $3C$. Furthermore, the weighted completion time of the unprobed set S_2 is $\sum_{i,j \in S_2, j \geq i} \min(w_i\mu_j, w_j\mu_i) \leq 3 \cdot OPT_{LP} \leq 3 \cdot ADAPT$.*

We now use the recombinant property (Lemma 3.1) to show that the above non-adaptive has a constant adaptivity gap against the optimal (hard or soft) budget adaptive strategy.

Theorem 3.4. *The non-adaptive strategy is a 8 approximation to the optimal (expected cost C) adaptive strategy, and spends cost $3C$.*

Proof. Consider any scenario σ of values of variables in set S_1. Let $P_\sigma = \sum_{i,j \in S_1, j \geq i} \min(w_j X_i, w_i X_j)$; let $Q = \sum_{i,j \in S_2, j \geq i} \min(w_j\mu_i, w_i\mu_j)$ and let $R_\sigma = \sum_{i \in S_1, j \in S_2} \min(w_j X_i, w_i\mu_j)$. The expected value of the non-adaptive solution is $\mathbf{E}[P] + Q + \mathbf{E}[R]$ where the expectation is over all scenarios σ. From Lemma 3.3, we have $Q \leq 3 \cdot ADAPT$. We also have $\mathbf{E}[P] \leq ADAPT$, since the adaptive solution obtains the best value from S_1 by observing all the variables. Using Lemma 3.1, we have $P_\sigma + Q \geq R_\sigma$ in all scenarios σ. Therefore, $\mathbf{E}[P] + Q \geq \mathbf{E}[R]$. The value of the non-adaptive solution is therefore at most $2(\mathbf{E}[P] + Q) \leq 8 \cdot ADAPT$. □

4 K-median Clustering

In this problem, we are given a metric space with point set $\mathcal{P}$, which defines a distance function l. The input is a set of nodes V, where the location of node i follows an independent distribution X_i over $\mathcal{P}$. Distribution $i \in V$ has probe cost c_i. The goal is to design an adaptive policy to probe the nodes which spends expected cost at most C. After probing, the algorithm opens K centers and assigns all probed and unprobed nodes to some center so that the expected distance cost (or *value*) of the clustering is minimized. This expectation is over the locations of the unprobed nodes. Note that after probing, the center selection and assignment policy assigns an unprobed node i to that open center w which minimizes the expected distance $\mathbf{E}[l(X_i, w)]$ where the expectation is over the random variable X_i. The goal is to design a probing policy whose resulting expected distance cost (or value) of K-median clustering is minimized, where the expectation is over the outcomes of the probes and the locations of the unprobed nodes.

We consider two variants of the problem. In the first variant, we assume that the center selection policy is restricted to opening centers from a set $S \subseteq \mathcal{P}$. This means the centers can only be chosen from points of the underlying metric space. Therefore, for an unprobed node i assigned to a center $w \in \mathcal{P}$, the expected distance cost is $\mathbf{E}[l(X_i, w)]$, where the expectation is over the random variable X_i. In the second variant, the centers are allowed to be *input nodes*, and therefore distributions. In this latter variant, an unporbed node j can be opened as a center after probing a set of nodes. Suppose an unprobed node i is assigned to this center, then the expected distance cost is $\mathbf{E}[l(X_i, X_j)]$, where the expectation is over *both* the random variables X_i and X_j.

We present a constant factor adaptivity gap for the former variant. We then show that the adaptivity gap is polynomially large for the latter variant. This shows a fundamental difference in the two variants.

4.1 Fixed Centers

We first consider the first variant where the centers can only be points from $S \subseteq \mathcal{P}$. We define a new metric space d over the points $\mathcal{P} \cup V$. For $i, j \in V$, define $d(i,j) = \mathbf{E}[l(X_i, X_j)]$ where the expectation is over the random variables X_i and X_j. For $i \in \mathcal{P}, j \in V$, define $d(i,j) = \mathbf{E}[l(i, X_j)]$, and for $i, j \in \mathcal{P}$, define $d(i,j) = l(i,j)$. The function d defines a metric space on the nodes $V \cup \mathcal{P}$ (refer Claim D.1 for a proof).

As before, we consider the outlier version of this problem: The goal is to find the subset $T \subseteq V$ of nodes of total probing cost at most C such that the cost of K-median clustering of the remaining nodes $V \setminus T$ in metric space d is minimized. The linear relaxation of this problem is the following:

$$\text{Minimize} \quad \sum_{i \in V, s \in S} d(i,s) x(i,s)$$

$$\begin{array}{rcll}
\sum_{i \in V} c_i z_i & \leq & C & \\
x(i,s) & \leq & y_s & \forall i \in V, s \in S \\
\sum_{s \in S} x(i,s) + z_i & \geq & 1 & \forall i \in V \\
\sum_{s \in S} y_s & \leq & K & \\
z_i, y_s, x(i,s) & \in & [0,1] & \forall i \in V, s \in S
\end{array}$$

Lemma 4.1. *The optimal solution to the linear program provides a lower bound on the value of the adaptive optimal probing and center selection policy. (Refer Appendix B for proof).*

Algorithm.

1. Round the solution of the above LP to construct a K-median solution with total cost of the outliers is $5C$ and the distance cost at most a factor of 5 of the distance cost in the linear program. This is achieved by the rounding procedure in [7]. Let the set of outliers be T and let $V \setminus T$ denote the remaining set of nodes.

2. Probe the set T. Let $T_\sigma \subseteq \mathcal{P}$ denote the realization of these nodes. Construct the K-median solution on the points $V \setminus T \cup T_\sigma$ in metric space d using any approximation algorithm (either the 4-approximation in [5, 26, 6] which work against the LP bound, or the $(3+\epsilon)$ approximation in [2]).

Analysis. Let $ADAPT$ denote the value of the optimal adaptive K-median solution. The following lemma is immediate from the description of the algorithm and Lemma 4.1.

Lemma 4.2. *There exists a polynomial time computable K-median solution restricted to the unprobed node-set $V \setminus T$ of value at most $5 \cdot ADAPT$.*

Using Lemma 4.2, we can see that the non-adaptive strategy can spend probe cost $5C$ and find a solution with $2K$ centers, K for the probed variables and K for the unprobed, of expected value at most $(8+\epsilon) \cdot ADAPT$. We reduce this to K centers below. The proof involves constructing a K-median solution that fractionally maps centers in the non-adaptive solution to centers in the optimal solution. This is done for every scenario of the realization of the locations of all the nodes, and averaged over the scenarios. The key technical hurdle is that the K-median solutions are over nodes which are distributions, and the optimal solution and the non-adaptive solutions could probe different sets of these nodes. We therefore need to construct fractional solutions for analysis which are *independent* of the distributions. The technical details are in Appendix B.

Theorem 4.3. *The non-adaptive solution constructed above has probing cost $5C$ and has value at most $29 \cdot ADAPT$.*

4.2 Arbitrary Centers

Consider now the case where the centers themselves are allowed to be input nodes, and therefore distributions. After probing, the center selection policy could decide to open an unprobed input node as a center, and assign probed and unprobed nodes to this center. The distance cost between the center and the assigned node is the expected distance between them, where the expectation is taken over possible locations of the center and the assigned node. We note that the linear programming relaxation is still a lower bound on the adaptive optimal solution. We omit the proof of the following.

Theorem 4.4. *The previous algorithm yields a non-adaptive probing strategy of cost $5C$, using $2K$-medians, an $(8+\epsilon)$-approximation to the adaptive optimal solution using K medians.*

We show that the above is the best possible in the following sense. We show below that any non-adaptive algorithm has polynomially large adaptivity gap on *both* distance and probing cost if it is restricted to opening at most $(1+\epsilon)K$ centers for some constant $\epsilon > 0$. Intuitively what fails is the following: In the proof for fixed centers, we mapped the unprobed nodes of the non-adaptive solution to a set of nodes in $\mathcal{P}$, and constructed a fractional solution using these locations for these nodes. This ensures that the metric space is over $\mathcal{P}$ and independent of distributions. In the current setting, an unprobed node could be mapped to an unprobed center. Therefore across scenarios, the metric space over the locations of the nodes changes.

Theorem 4.5. *The adaptivity gap for K-medians when centers can be input nodes is polynomially large on both distance and probing cost.*

Proof. Consider M distinct copies (at a mutual distance of at least M^3L from each other, $L = M^2$) of the following 2-dimensional node set. In copy m, there are $r+1$ "cheap" nodes $X_{1m}, X_{2m}, \ldots$ which cost 1 to probe. Distribution X_{im} is 0 with probability $1/2$ and $i+1$ otherwise. In addition, there are pairs of nodes which are well-separated from other pairs by a large distance L. Pair j corresponds to two distributions: Y_{jm} and Z_{jm}. Y_{jm} is $(L+jL, 1)$ with probability $1-\log t/t$ and $(L+jL, 0)$ with probability $\log t/t$. Z_{jm} is $(L+jL, -1)$ with probability $1-\log t/t$ and $(L+jL, 0)$ with probability $\log t/t$. These distributions are "expensive" with probing cost $(r+1)M$. Again note that the nodes for each m are far removed from the nodes corresponding to other m. Let $K = (2t+r)M$. For each m, the adaptive solution will place $2t+r$ medians. Assume $r > 2\log Mt$, and $M \ll t^2$.

The adaptive algorithm probes the "cheap" distributions X_*. For some m, if X_{*m} resolve to $r+1$ values distinct points (this happens with probability $1/2^{r+1}$ for a particular m and therefore $\frac{M}{2^{r+1}}$ overall), then probe *all* the "expensive" distributions Y_* and Z_*. If at most r distinct values are observed for all m, then do not probe further, since there is a k median solution of value 0 in which every one of the expensive nodes, and the realized locations of the cheap nodes is a median. The expected probing cost of this scheme is at most $M(r+1) + (r+1)M \cdot Mt \cdot \frac{M}{2^{r+1}} \leq 2M(r+1)$.

We now analyze the expected distance. For a certain m, when the expensive nodes Y_{*m} and Z_{*m} are probed, with probability $1-(1-\log t/t)^{2t} \geq 1-1/t^2$ some pair collides on the x-axis. In this case, the k medians solution has distance cost 0, else it has a distance cost of at most 2. The expected distance cost is therefore at most $(1/2^{r+1}) \cdot (1/t^2) \cdot 2$ for each m. Therefore, the overall expected distance cost is $\frac{M}{t^2 2^r}$.

Any non-adaptive probing scheme must probe at least one expensive distribution in each copy, else the distance cost is at least $\frac{1}{2^{r+2}} \gg \frac{M}{t^2 2^r}$ in that copy (in the case where the cheap distributions resolve to distinct values w.p. $1/2^{r+1}$, the distance cost will be at least 0.5). Therefore, the probing cost needed is $(r+1)M^2$, which implies that unless the probing cost is a factor M larger than the adaptive scheme, the distance cost must be $\frac{t^2}{M}$ times larger. Therefore, no non-trivial adaptivity gap is possible. □

Acknowledgments: We would like to thank Saswati Sarkar, Sariel Har-Peled, Ashish Goel, Shivnath Babu, Utkarsh Srivastava, 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] V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. Comput.*, 33(3):544–562, 2004.

[3] S. Babu and P. Bizarro. Proactive reoptimization. In *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, 2005.

[4] M. Charikar, R. Fagin, J. Kleinberg, P. Raghavan, and A. Sahai. Querying priced information. In *Proc. of the Annual ACM Symp. on Theory of Computing*, 2000.

[5] M. Charikar and S. Guha. Improved combinatorial algorithms for the facility location and k-median problems. In *FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, 1999.

[6] M. Charikar, S. Guha, É. Tardos, and D. B. Shmoys. A constant-factor approximation algorithm for the k-median problem (extended abstract). In *STOC '99: Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 1–10, 1999.

[7] M. Charikar, S. Khuller, D. M. Mount, and G. Narasimhan. Algorithms for facility location problems with outliers. In *SODA '01: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 642–651, 2001.

[8] F. Chu, J. Halpern, and J. Gehrke. Least expected cost query optimization: What can we expect? In *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. In *Proc. of the Annual Symp. on Foundations of Computer Science*, 2004.

[11] B. C. Dean, M. X. Goemans, and J. Vondrák. Adaptivity and approximation for stochastic packing problems. In *SODA '05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 395–404, 2005.

[12] 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.

[13] T. Feder, R. Motwani, R. Panigrahy, C. Olston, and J. Widom. Computing the median with uncertainty. *SIAM J. Comput.*, 32(2), 2003.

[14] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[15] 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.

[16] A. Goel and P. Indyk. Stochastic load balancing and related problems. In *Proc. of the Annual Symp. on Foundations of Computer Science*, 1999.

[17] M. X. Goemans and D. P. Williamson. A general approximation technique for constrained forest problems. In *SODA '92: Proceedings of the third annual ACM-SIAM symposium on Discrete algorithms*, pages 307–316, 1992.

[18] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. *Proceedings of FOCS*, pages 359–366, 2000.

[19] S. Guha, K. Munagala, and S. Sarkar. Jointly optimal probing and transmission strategies for multi-channel wireless systems. In *Proc. of the Conference on Information Sciences and Systems*, 2006.

[20] 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.

[21] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: Approximation algorithms for stochastic optimization. In *Proc. of the Annual ACM Symp. on Theory of Computing*, 2004.

[22] A. Gupta and M. Pál. Stochastic steiner trees without a root. *Proc. of ICALP*, pages 1051–1063, 2005.

[23] A. Gupta, M. Pál, R. Ravi, and A. Sinha. What about wednesday? Approximation algorithms for multistage stochastic optimization. *Proc. of APPROX-RANDOM*, pages 86–98, 2005.

[24] 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.

[25] 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.

[26] K. Jain and V. V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *J. ACM*, 48(2):274–296, 2001.

[27] S. Khanna and W-C. Tan. On computing functions with uncertainty. In *Proc. of the ACM Symp. on Principles of Database Systems*, 2001.

[28] J. Kleinberg, Y. Rabani, and É. Tardos. Allocating bandwidth for bursty connections. *SIAM J. Comput*, 30(1), 2000.

[29] 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.

[30] R. H. Mohring, A. S. Schulz, and M. Uetz. Approximation in stochastic scheduling: the power of LP-based priority policies. *J. ACM*, 46(6):924–942, 1999.

[31] 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.

[32] C. Olston. *Approximate Replication*. PhD thesis, Stanford University, 2003.

[33] D. Shmoys and C. Swamy. Stochastic optimization is (almost) as easy as discrete optimization. In *FOCS '04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science*, pages 228–237, 2004.

[34] M. Skutella and M. Uetz. Scheduling precedence-constrained jobs with stochastic processing times on parallel machines. In *SODA '01: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 589–590, 2001.

[35] C. Swamy and D. B. Shmoys. Sampling-based approximation algorithms for multi-stage stochastic. In *FOCS '05: Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 357–366, 2005.

A On the Technical Difficulty of Model-driven Optimization

Obtaining convex relaxations: First consider the technical difficulty in expressing *even the non-adaptive versions* of the problems as a convex optimization problem. Let us begin from any convex stochastic optimization problem over random variables $\vec{X}$, e.g., $\min\{\mathbf{E}[\vec{X} \cdot \vec{y}] \mid \vec{y} \in \mathcal{P}\}$ where $\mathcal{P}$ is a convex set. By linearity of expectation, the above is no harder than deterministic convex optimization, since the objective is simply $\sum_{i=1}^{n} \mathbf{E}[X_i]y_i$. In a non-probing model the vector $\vec{y}$ is chosen *before* the values of $\vec{X}$ are revealed, and thus expressing the problem as a convex program is straightforward. In the probing model, the vector $\vec{y}$ is chosen *after* the values of variables in a set S are probed and revealed. The goal now is to find the best set S^* of variables to probe such that (i) $\sum_{i \in S^*} c_i \leq C$, and (ii) we minimize $\mathbf{E}[\min\{\sum_{i \in S^*} X_i y_i + \sum_{i \notin S^*} \mu_i y_i \mid \vec{y} \in \mathcal{P}\}]$ where $\mu_i = \mathbf{E}[X_i]$. The most natural relaxation for the problem is to have an indicator variable z_i which is 1 if X_i is probed, and 0 otherwise. The relaxation yields:

$$\text{Minimize } \mathbf{E}[\min\{\sum_{i=1}^{n} (X_i z_i + \mu_i(1 - z_i)) \cdot y_i \mid \vec{y} \in \mathcal{P}\}]$$

$$\begin{array}{rcl} \sum_i c_i z_i & \leq & C \\ 0 \leq z_i & \leq & 1 \quad \forall i = 1, 2, \ldots, n \end{array}$$

The above formulation is clearly non-linear. Unfortunately, *it is also concave in general*, even in the simple case where $\mathcal{P} = \{\sum_i y_i \geq 1, \vec{y} \geq 0\}$, which encodes finding the minimum element. In view of the concavity, we cannot expect general convex programming relaxation based approach to easily solve our problem. In general, these problems become hard to approximate – when $\mathcal{P}$ encodes an arbitrary set of linear covering constraints, we now show an approximation preserving reduction from set cover.

Theorem A.1. *If the convex set $\mathcal{P}$ encodes m arbitrary linear covering constraints, then no sub-polynomial approximation factor on the objective is possible in polynomial time unless the probing cost is inflated by a factor of $\Omega(\log m)$, assuming $NP \not\subseteq DTIME(m^{O(\log \log m)})$.*

Proof. We reduce from SET COVER. Given a set cover instance with n sets $Q_1, Q_2, \ldots, Q_n$ and m elements $e_1, e_2, \ldots, e_m$ where the goal is to decide if there is a set cover of size k, construct a stochastic fractional weighted set cover instance:

$$\min \ \mathbf{E}[\sum_{i=1}^{n} X_i y_i + \gamma y_U]$$

$$\begin{array}{rcl} \sum_{i: e_j \in Q_i} y_i + y_U & \geq & 1 \quad \forall e_j, j = 1, 2, \ldots, m \\ y_i, y_U & \geq & 0 \quad \forall i = 1, 2, \ldots, n \end{array}$$

The weights X_i in the objective are stochastic and can be probed at cost c_i. Note that in the absence of probing, such a fractional set cover instance is polynomial time solvable. The X_i are i.i.d. Bernoulli variables which are W with probability p and 0 otherwise. Define $\gamma = Wp$. Set $p = \frac{1}{2^n}$ so that $W = \gamma 2^n$. Choose $\gamma = n^2$. Note that there is a set U which contains all elements, and has weight γ which is deterministic. The probing budget is $C = k$. Let V denote the collection of all sets.

First consider designing a non-adaptive probing scheme. Any scheme chooses a subset $S \subseteq V$ of sets to probe to resolve their weight. With probability at least $(1 - 1/2^n)^n$, all weights resolve to 0. If these sets cover all elements, then the weight of the objective for the fractional set cover instance is 0. With the remaining probability the fractional solution sets $y_U = 1$, and therefore its objective value is γ. Therefore, if S defines a set cover, then the expected objective value is $\approx (n\gamma)/2^n \ll \gamma$. If the sets in S do not define a set cover, then the expected weight for any choice of the vector $\vec{y}$ is at least γ. Since approximating the size of the optimal set cover is $\Omega(\log m)$ hard [14], designing a non-adaptive scheme with any non-trivial guarantee on the objective has to incur probing cost $\Omega(k \log m) = \Omega(C \log m)$.

Next consider adaptive observation strategies. Consider the event that all observed weights are 0. Let S_0 denote the collection of sets whose weights are probed, conditioned on this event. The probability of this event is at least $p_0 = (1 - 1/2^n)^n$. Therefore, the expected probing cost of this strategy is $\approx |S_0|$. Let b_0 denote objective value of the solution in this event. If this event does not occur (which is with probability at most $\frac{n}{2^n}$, the objective value of the strategy is at most γ. The expected value of the strategy is therefore at most $\approx b_0 + n\gamma/2^n$. If $b_0 = 0$ (so that S_0 is a set cover), this cost is $\approx n\gamma/2^n \ll \gamma$, else it is at least γ, since $b_0 \geq \gamma$ in this case. Therefore, an adaptive strategy cannot have a better approximation guarantee than a non-adaptive strategy, since it would imply an algorithm for choosing S_0 so that $b_0 = 0$, which corresponds to S_0 being a set cover. $\square$

In view of the above observations, one of the major challenges is in capturing ideal abstractions that can be expressed and rounded easily – this is one of the main contributions of our work.

Adaptivity gaps: Even the simplest covering problems in the probing model have significant adaptivity gaps. In particular, the adaptivity gap of finding the minimum element (even up to any polynomial factor) is polynomial. This can be seen with n i.i.d. variables $X_i = B(1, \frac{1}{2})$ with unit probing costs. The optimal adaptive algorithm probes all variables in arbitrary order, stopping when any variable is observed to be 0. The expected value is $\frac{1}{2^n}$ and the expected cost is 2. The non-adaptive algorithm that spends cost k achieves an objective of $\frac{1}{2^k}$, implying an exponentially large gap. For a version of the minimum element problem an adaptive solution is presented in [19]. We again note that our definition of non-adaptive means *parallel* or one-shot observations, and is *different* from the definition of non-adaptive in [10]. The difference in definitions is intrinsic to the two models. The adaptive solution we present in [19] would be ”non-adaptive” according to the definition of [10], since the *order* of probing the variables can be fixed in advance.

B The Omitted Proofs

Lemma B.1. *(Lemma 2.1) For a set $S = \{X_1, X_2, \dots X_n\}$, let $t = \mathbf{E}[\max_{i \in S} X_i]$. Then, $\sum_{i \in S} \mathbf{E}_{2t}[X_i] \leq 2t$.*

Proof. If all the variables are independent Bernoulli so that $X_i = B(s_i, p_i)$ is s_i with probability p_i and 0 with probability $1 - p_i$, Kleinberg, Rabani, and Tardos [28] show that $\sum_i \mathbf{E}_{2t}[X_i] \leq 2t$. For completeness, we re-prove Lemma 3.3 in [28].

Fix any number L such that $\sum_i \mathbf{E}_L[X_i] > L$. We will show that $t > \frac{L}{2}$. This would mean that setting $L = 2t$, we must have $\sum_i \mathbf{E}_{2t}[X_i] \leq 2t$. For the proof, continuously reduce the p_i values until for the new random variables $\tilde{X}_i$, we have $\sum_i \mathbf{E}_L[\tilde{X}_i] = L$. We will prove that $\mathbf{E}[\max_i \tilde{X}_i] \geq L/2$. Since $t = \mathbf{E}[\max_i X_i] > \mathbf{E}[\max_i \tilde{X}_i]$, this would imply $t > L/2$. Let q_i denote the new probability of s_i in random variable $\tilde{X}_i$ (we have $q_i \leq p_i$).

Let $S' = \{i | s_i \geq L\}$. Since $\sum_{i \in S'} q_i s_i = L$ and $s_i \geq L$ for $i \in S'$, we have $\sum_{i \in S'} q_i \leq 1$. Re-number the variables so that $s_k \geq s_{k-1} \geq \cdots \geq s_1 \geq L$ belong to set S'. Let $x_i = \sum_{j \geq i} q_j$. Let $\mathcal{E}_i$ denote the event that at least one of the random variables $\{\tilde{X}_j\}_{j\geq i}$ is in "on" state (that is, does not have value 0). Let $y_i = \Pr[\mathcal{E}_i]$. We have $x_i \leq 1$ for all $i = 1, 2, \ldots, k$ since $\sum_{i\in S} q_i \leq 1$.

For any $i = 1, 2, \ldots, k$, set $q = \frac{\sum_{j\geq i} q_j}{k-i+1}$. Note that $x_i \leq 1$, so we have

$$\begin{array}{rcccl} y_i & = & 1 - \Pi_{j\geq i} 1 - (1-q_j) & \geq & 1-(1-q)^{(\frac{1}{q}\sum_{j\geq i} q_j)} \\ & \geq & 1 - \exp(-\sum_{j\geq i} q_j) & \geq & 1 - e^{-x_i} \geq \frac{x_i}{2} \end{array}$$

We therefore have

$$\begin{array}{rclcl} t > \mathbf{E}[\max_i \tilde{X}_i] & \geq & \sum_{i=1}^k s_i \times \Pr[\tilde{X}_i \text{ is on } \wedge \textbf{ not } \mathcal{E}_{i+1}] & & \\ & \geq & \sum_{i=1}^k s_i (y_i - y_{i+1}) & = & \sum_{i=1}^k y_i (s_i - s_{i-1}) \\ & \geq & \frac{1}{2}\sum_{i=1}^k x_i (s_i - s_{i-1}) & \geq & \frac{1}{2}\sum_{i=1}^k q_i s_i = \frac{L}{2} \end{array}$$

This proves that $t > \frac{L}{2}$, which completes the proof for the Bernoulli case.

For general distributions, we reduce to the Bernoulli case. We express the distributions as the maximum of a collection of Bernoulli distributions as follows. Suppose distribution X_i takes on values $s_1 < s_2 < \cdots < s_m$ with probabilities $p_{1i}, p_{2i}, \ldots, p_{mi}$. Create m independent Bernoulli variables $Y_{i1} = B(s_m, p_{mi})$, $Y_{i2} = B(s_{m-1}, \frac{p_{(m-1)i}}{1-p_{mi}})$, $Y_{i3} = B(s_{m-2}, \frac{p_{(m-2)i}}{1-p_{(m-1)i}-p_{mi}}), \ldots$. We have $X_i = \max_{j=1}^m Y_{ij}$. Therefore, $\mathbf{E}[\max_{i\in S, j\in\{1,\ldots,m\}} Y_{ij}] = t$. From the Bernoulli case, we have $\sum_{i\in S}\sum_{j=1}^m \mathbf{E}_{2t}[Y_{ij}] \leq 2t$.

Clearly,

$$\sum_{j=1}^m \mathbf{E}_q[Y_{ij}] = s_m p_{mi} + s_{m-1}\frac{p_{(m-1)i}}{1-p_{mi}} + \ldots \geq s_m p_{mi} + s_{m-1} p_{(m-1)i} + \ldots = \mathbf{E}_q[X_i]$$

Therefore, $\sum_{i\in S} \mathbf{E}_{2t}[X_i] \leq 2t$. □

Lemma B.2. *(Lemma 3.1) For any partitioning of the jobs into two disjoint sets A and B, we have:*

$$\sum_{i,j\in A, j\geq i} w_i w_j \min(\alpha_i, \alpha_j) + \sum_{i,j\in B, j\geq i} w_i w_j \min(\alpha_i, \alpha_j) \geq \sum_{i\in A, j\in B} w_i w_j \min(\alpha_i, \alpha_j)$$

Proof. Let $\gamma_{ij} = \min(\alpha_i, \alpha_j)$. First, suppose all $\gamma_{ij} = 1$. We have:

$$\begin{array}{rrcl} & (\sum_{i\in A} w_i - \sum_{j\in B} w_j)^2 & \geq & 0 \\ \Rightarrow & 2(\sum_{i,j\in A, j\geq i} w_i w_j + \sum_{i,j\in B, j\geq i} w_i w_j) & \geq & 2(\sum_{i\in A} w_i)(\sum_{j\in B} w_j) \\ \Rightarrow & \sum_{i,j\in A, j\geq i} w_i w_j + \sum_{i,j\in B, j\geq i} w_i w_j & \geq & \sum_{i\in A, j\in B} w_i w_j \end{array}$$

We next consider the case of general α_i. We will prove this by induction on the number of jobs (the base case being trivial). Let $i^* = \text{argmin}_i \alpha_i$. For each job i, let $\beta_i = \alpha_i - \alpha_{i^*}$, and let $\delta_{ij} = \min(\beta_i, \beta_j)$. We have:

$$w_i w_j \gamma_{ij} = w_i w_j (\alpha_{i^*} + \delta_{ij})$$

From the proof of the $\gamma_{ij} = 1$ case, we have:

$$\alpha_{i^*}(\sum_{i,j\in A, j\geq i} w_i w_j + \sum_{i,j\in B, j\geq i} w_i w_j) \geq \alpha_{i^*} \sum_{i\in A, j\in B} w_i w_j$$

The set of jobs with non-zero β values is strictly smaller than n. Let Z be the set of jobs with $\beta_i = 0$. By the inductive hypothesis we have:

$$\sum_{i,j\in A\setminus Z, j\geq i} w_i w_j \delta_{ij} + \sum_{i,j\in B\setminus Z, j\geq i} w_i w_j \delta_{ij} \geq \sum_{i\in A\setminus Z, j\in B\setminus Z} w_i w_j \delta_{ij}$$

Adding the previous two inequalities, we have the proof of the lemma. □

Lemma B.3. *(Lemma 3.2)* $OPT_{LP} \leq ADAPT$.

Proof. Given a strategy of optimum, denote $b(i, j)$ be the probability that between i and j, the optimum solution probes i first (including the event that it does not probe j at all). Likewise define $b(j, i)$ to be the probability that j is probed before i (or only j was probed). Define $n(i, j)$ to be the probability that neither i nor j is probed. Thus $b(i, j) + b(j, i) + n(i, j) = 1$. Let event $\varphi(i), \varphi(j)$ be the probabilities with which i and j were probed respectively. Clearly $\varphi(i) \geq b(i, j)$ and $\varphi(j) \geq b(j, i)$.

The adaptive optimal algorithm obeys $\sum_i \varphi(i)c_i \leq C$. Thus the quantities $\varphi(i), n(i, j)$ satisfy the set of equations:

$$\sum_i z_i c_i \leq C$$
$$e(i, j) + z_i + z_j \geq 1 \text{ for all } i, j$$

with $z_i = \varphi(i)$ and $e(i, j) = n(i, j)$.

The solution of the optimum strategy is lower bounded by

$$\sum_{j>i} n(i, j) \min\{w_i\mu_j, w_j\mu_i\} + \sum_i (1 - \varphi(i))w_i\mu_i$$

This follows from the fact that in the scenarios in which neither i nor j is probed, the best strategy for the optimum is to schedule the jobs according to their expectation. Thus the LP is a lower bound of the optimum adaptive policy. □

Lemma B.4. *(Lemma 4.1) The optimal solution to the linear program provides a lower bound on the value of the adaptive optimal solution.*

Proof. Let Φ_i be an indicator random variable which is 1 if node i is probed by the adaptive optimal solution, and 0 otherwise. Let $\phi_i = \mathbf{E}[\Phi_i]$. Clearly, $\sum_{i\in V} c_i\phi_i \leq C$.

Let $\Gamma(i, s)$ be an indicator variable which is set to 1 if the optimal solution does not probe node i and assigns this node to center s. Let $\gamma(i, s) = \mathbf{E}[\Gamma(i, s)]$. If a node i is not probed, the distance to center s is $d(i, s)$. Therefore, the cost of the adaptive optimal solution is at least $\sum_{i\in V, s\in S} d(i, s)x(i, s)$.

Let $R(s) = \max_{i\in V} \Gamma(i, s)$, and let $\rho(s) = \mathbf{E}[R(s)]$. In any solution, we always have $\sum_{s\in S} \Gamma(i, s) + \Phi_i \geq 1 \quad \forall i \in V$. This statement is trivial if node i is probed and $\Phi_i = 1$. Else, node i has to be assigned to some center in S, and hence the inequality follows. Taking expectations, we have $\sum_{s\in S} \gamma(i, s) + \phi_i \geq 1 \quad \forall i \in V$.

Similarly, since $R(s) \geq \Gamma(i, s)$, we have $\rho(s) \geq \gamma(i, s) \quad \forall s \in S, i \in V$.

We also have $\sum_{s\in S} R(s) \leq K$, since centers with $R(s) = 1$ are definitely used by the solution. Taking expectations, we have $\sum_{s\in S} \rho(s) \leq K$. Setting $x(i, s) = \gamma(i, s)$, $z_i = \phi_i$, and $y_s = \rho(s)$ completes the proof. □

Theorem B.5. *(Theorem 4.3) The non-adaptive observation strategy for K-medians has probing cost $5C$ and has value at most $29 \cdot ADAPT$.*

Proof. We construct the non-adaptive solution in two phases: first, construct K medians on the unprobed nodes $V \setminus T$ using the outlier algorithm. Now move the unprobed nodes to the corresponding centers, and construct a K-median solution on the realization of the nodes in T and the nodes in $V \setminus T$ located at their respective medians.

We will proceed by considering scenarios of the values of all nodes. Fix some sample σ of the locations of all nodes. Let A_σ denote the value of the optimal adaptive solution given this scenario. Therefore, $\mathbf{E}[A_\sigma] = ADAPT$ where the expectation is over σ.

Focus now on the non-adaptive solution. For the scenario σ, let U_σ denote the distance value of the first phase which clusters just the unprobed nodes. Therefore, $\mathbf{E}[U_\sigma] \leq 5 \cdot ADAPT$ from Lemma 4.2.

Fix the realization r of the nodes in the probed set T. Let $\mathcal{F}_r$ denote the set of samples σ corresponding to this realization of T. For scenario $\sigma \in \mathcal{F}_r$, let $P_\sigma \subseteq \mathcal{P}$ denote the set of realized nodes for T, along with the nodes of $V \setminus T$ located at their corresponding medians constructed in the first phase. Note that the set P_σ is the same for all $\sigma \in \mathcal{F}_r$; therefore, the non-adaptive solution constructs a unique K-median solution for all these scenarios in the second phase. We now construct a fractional K-median solution on P_σ for $\sigma \in \mathcal{F}_r$.

In scenario σ send each probed node (in set T) to its assigned center in the optimal solution A_σ. Similarly, send each unprobed node (which is located at its assigned center in the solution U_σ) back to its realized location in scenario σ and from there to its assigned center in A_σ. This yields a mapping from the nodes P_σ to K medians in scenario σ. The distance value of this mapping is at most $A_\sigma + U_\sigma$ by triangle inequality. Note that the distances in this mapping are distances between points in $\mathcal{P}$, and do not involve the distributional nodes V. This yields a valid K-median solution on P_σ.

Since there is a feasible K-median solution for each $\sigma \in \mathcal{F}_r$, these when averaged over $\sigma \in \mathcal{F}_r$ define a fractional K-median solution for the set of points P_σ. Note again that $P_\sigma \subseteq \mathcal{P}$ is the same set of points for all $\sigma \in \mathcal{F}_r$. Therefore, there is an integer K-median solution on P_σ of value 4 times this fractional value. The non-adaptive algorithm pays at most this integer value for the second phase for scenarios $\sigma \in \mathcal{F}_r$.

The expected value of this second phase solution over all realizations r of nodes in T is therefore at most $4\mathbf{E}[A_\sigma + U_\sigma] \leq 24 \cdot ADAPT$. The overall value the non-adaptive algorithm is the sum of the values of the two phases, and is therefore at most $29 \cdot ADAPT$. $\square$

C Knapsack: Random Item Sizes

We consider the Knapsack problem where the profits are deterministic values t_i and the sizes are random variables X_i. This generalizes the stochastic knapsack problem in [10] to include probing. Item i has observation cost c_i and the knapsack capacity is B. The "soft" budget on the probing cost is given by $C \geq \max_i c_i$.

In the absence of probing, the problem formulation is exactly the same as in [10]. Call this problem $\mathcal{U}$. The solution is an adaptive policy to place items into the knapsack, when the exact size of an item is revealed *after* it is placed into the knapsack, and the goal is to choose that placement policy which maximizes the expected profit.

In the presence of probing, our goal is to design an adaptive policy that can possibly probe an item to determine its exact size *before* deciding to place it in the knapsack, subject to the expected probing cost being at most C. If an unprobed item is placed in the knapsack, there is no probing cost incurred; however, the exact size of the item is revealed *after* it is placed in the knapsack. Therefore the policy makes joint adaptive decisions both about probing the items, as well as about their placement (with or without probing) into the knapsack. We assume that the probing decisions need not be coupled with the placement decisions – a policy can probe item i, followed by item j, and then decide to place unprobed item k in the knapsack, and finally, based on the knowledge of the exact sizes of items i, j, and k, decide to place item i in the knapsack. Let the optimal such policy be $\mathcal{OPT}$. Note that if the probing costs of the items are infinitely large, the placement policy is restricted to using only unprobed variables; therefore, optimal policy is the same as that for $\mathcal{U}$ (or stochastic knapsack [10]).

For a set of probed items S define the best expected profit for packing a subset of S into the knapsack to be:

$$g(S) = \mathbf{E}[\max_{Q \subseteq S, \sum_{i \in Q} X_i \leq B} \sum_{i \in Q} t_i]$$

If the placement policy is restricted to using only probed variables, the formulation is as follows: The adaptive algorithm now adaptively probes a set S of items. Subject to the constraint $\mathbf{E}[\sum_{i \in S} c_i] \leq C$, the goal is to maximize $\mathbf{E}[g(S)]$. Here, the expectation is on the adaptive choice of S and the random variables probed. Call this problem formulation $\mathcal{P}$.

Consider the algorithm $\mathcal{A}$ that chooses the better of two policies: The optimal policy for $\mathcal{P}$ (which restricts the placement policy to use only probed items), and the optimal policy for $\mathcal{U}$ (which restricts the placement policy to use only unprobed variables). The latter policy is of course is the same as the optimal policy for stochastic knapsack [10].

Lemma C.1. *$\mathcal{A}$ is a $1/2$-approximation to the optimal adaptive joint probing and placement policy, $\mathcal{OPT}$.*

Proof. Consider any leaf node in the optimal decision tree. Half the profit at that leaf node is derived from either probed or unprobed items. If half the optimal expected profit is derived from unprobed items, then the optimal algorithm that uses only unprobed variables yields at least this expected profit. If half the optimal profit is derived from probed items, the optimal strategy that uses only probed items yields at least this profit. This is easy to see by performing the probes and ignoring the items after probing, but conditioning the remaining actions on the results of the probes. We omit the details. $\square$

For $\mathcal{U}$, the algorithm in [10] yields a $1/4$-approximation. We now focus on designing an adaptive algorithm for $\mathcal{P}$.

We note that the definition of ”non-adaptive” is different for the probing and the placement policies. The approximate placement policy for $\mathcal{U}$ is non-adaptive according to the definition of [10], since the *order* in which items are placed into the knapsack is fixed. A probing policy is non-adaptive if the probes are done in parallel (or one-shot). We show below that the optimal adaptive probing policy for $\mathcal{P}$ can be approximated to a constant factor by a *non-adaptive* probing policy, that chooses upfront a set of items to probe, and based on the result of the probes (now, all item sizes are known exactly), chooses a subset of these to place in the knapsack. The placement policy involves solving a deterministic knapsack instance given the outcome of the probing policy. Therefore, we show that the overall joint probing and placement policy is non-adaptive by losing a constant factor on $\mathcal{OPT}$.

C.1 Restriction to Probed Items

We now show that for the problem $\mathcal{P}$, any adaptive probing policy has at most a constant factor benefit over the best non-adaptive (one-shot) probing policy. This holds for adaptive policies with either hard or soft budgets.

Non-adaptive Greedy Algorithm We first show a greedy algorithm which is a constant factor approximation to the optimal *non-adaptive* probing policy for $\mathcal{P}$. Let S^* denote $\text{argmax}_{S \mid \sum_{i \in S} c_i \leq C} g(S)$ denote the optimal non-adaptive solution. Denote by $f(S)$ the fractional profit that is obtained by probing the set S and packing a subset in the knapsack. Formally,

$$f(S) = \mathbf{E}[\max_{\vec{y} \geq 0, y_i \leq 1, \sum_{i \in S} X_i y_i \leq B} \sum_{i \in S} t_i y_i]$$

We can assume w.l.o.g. that our distributions satisfy $X_i = \infty$ if $X_i > B$. Therefore, in any scenario of values of sizes of items, the integer profit is at least half the fractional profit. Therefore, $g(S) \geq 0.5 f(S)$. We show in [15] that f is submodular. This defines a natural greedy algorithm which is a constant factor approximation to the best non-adaptive solution S^*.

We now show that the greedy non-adaptive algorithm is also a constant factor approximation to the best adaptive probing policy.

Adaptivity Gap: Let $T^* = \max\{f(S) | \sum_{i \in S} c_i \leq C\}$. Split the items into two classes based on whether $t_i \leq 2T^*$ or not. We bound the "excess" over the non-adaptive solution for each set of items using separate techniques.

First consider the class with $t_i \leq 2T^*$ denoted G_1. Let $R_i = t_i$ if $\frac{X_i}{t_i} \leq \frac{B}{2T^*}$ and 0 otherwise. Intuitively, $R_i > 0$ only for items with profit to size ratio at least 2 times the non-adaptive optimal profit to size ratio. We use Chernoff bounds in this case.

Lemma C.2. *$\sum_{i \in S \cap G_1} \mathbf{E}[R_i] \leq 10T^*$ for any set S such that $\sum_{i \in S} c_i \leq C$.*

Proof. Suppose $\sum_{i \in S \cap G_1} \mathbf{E}[R_i] \geq 10T^*$. By Chernoff bounds, $\Pr[\sum_{i \in S \cap G_1} R_i \leq 2T^*] \leq e^{-\frac{10T^*}{4t_{\max}}} < 0.5$. In the event that $\sum_{i \in S \cap G_1} R_i \geq 2T^*$, each item with $R_i > 0$ has profit to size ratio at least $\frac{2T^*}{B}$. A (possibly fractional) subset of these items of profit exactly $2T^*$ therefore has total size at most B, which is feasible. Since the event $\sum_{i \in S \cap G_1} R_i \geq 2T^*$ happens with probability greater than 0.5, the expected profit from S is greater than T^*, contradicting the optimality of T^*. □

For the other class with $t_i \geq 2T^*$ denoted G_2, we need a different bounding technique. Define $D_i = t_i$ if $X_i \leq B$ and 0 otherwise. Therefore, $D_i > 0$ implies the item by itself fits into the knapsack. We have the following lemma, which is based on Lemma 2.1.

Lemma C.3. *$\sum_{i \in S \cap G_2} \mathbf{E}[D_i] \leq 2T^*$ for any set S with $\sum_{i \in S} c_i \leq C$.*

Proof. Consider the solution that probes set S and fits the item with largest D_i in the knapsack. The profit of this scheme is $\mathbf{E}[\max_{i \in S \cap G_2} D_i]$. Therefore, $\mathbf{E}[\max_{i \in S \cap G_2} D_i] \leq T^*$. By using Definition 2 and applying Lemma 2.1, we have $\sum_{i \in S} \mathbf{E}_{2T^*}[D_i] \leq 2T^*$. Observing that $t_i \geq 2T^*$ and D_i is a Bernoulli variable which is 0 or t_i, we have $\mathbf{E}_{2T^*}[D_i] = \mathbf{E}[D_i]$. Therefore, $\sum_{i \in S \cap G_2} \mathbf{E}[D_i] \leq 2T^*$. □

Theorem C.4. *The non-adaptive optimal solution is a constant factor approximation to the value of any adaptive observation strategy.*

Proof. Let Z_i denote the event that the adaptive algorithm observes the size of item i and let $z_i = \mathbf{E}[Z_i]$. Let random variable $\mathcal{Y} = \sum_{i \in G_1} Z_i R_i + \sum_{i \in G_2} Z_i D_i$ denote the total profit of items with $R_i > 0$ and $D_i > 0$ observed by the algorithm. We have $\mathbf{E}[\mathcal{Y}] = \sum_{i \in G_1} z_i \mathbf{E}[R_i] + \sum_{i \in G_2} z_i \mathbf{E}[D_i]$. We also have $\sum_i c_i z_i \leq C$. Therefore, using Lemma C.2 and Lemma C.3, we have $\mathbf{E}[\mathcal{Y}] \leq 2(10T^* + 2T^*) = 24T^*$.

The items with $R_i = 0$ cannot account for more than $2T^*$ of the profit in any scenario, given the upper bound on their profit to size ratio of $2T^*/B$. The items with $D_i = 0$ do not fit into the knapsack anyway. Therefore, expected profit of the adaptive algorithm is at most $\mathbf{E}[\mathcal{Y}] + 2T^* \leq 26T^*$. □

D Minimum Spanning Trees

We consider the minimum spanning tree problem – the algorithms extend to Steiner trees and Metric Traveling salesman problem naturally. The input is a collection of n nodes. The location of node i is an independent random variable X_i, which is a distribution over points $\mathcal{P}$ in space. Let $\mathcal{V}$ denote the set of random variables. The exact location of node i is determined by spending probing cost c_i. The goal is to design an adaptive probing scheme which minimizes the expected cost of connecting the nodes by a spanning tree, subject to the constraint that this decision tree has expected probing cost at most C. Let l denote the distances in the underlying metric space.

Roadmap: We will follow the same recombinant strategy used earlier; and as in the k-Median problem we would have to express an LP over a different metric space. Let $D(i,j) = l(X_i, X_j)$ denote the random variable corresponding to the distance between X_i and X_j. Let $d(i,j) = \mathbf{E}[D(i,j)]$ where the expectation is over the random variables X_i and X_j.

Claim D.1. $d(i,j) + d(j,k) \geq d(i,k)$.

Proof. For any realization of the values of X_i, X_j, and X_k, we have $D(i,j) + D(j,k) \geq D(i,k)$. Taking expectations over the random choices of X_i, X_j, and X_k, the claim follows. $\square$

Define the following metric space: The vertices are the points $\mathcal{M} = \mathcal{P} \cup \mathcal{V}$. The distance metric is d. For $i, j \in \mathcal{V}$, $d(i,j) = \mathbf{E}[D(i,j)] = \mathbf{E}[l(X_i, X_j)]$. For $i, j \in \mathcal{P}$, $d(i,j) = l(i,j)$. For $i \in \mathcal{P}, j \in \mathcal{V}$, define $d(i,j) = \mathbf{E}[l(i, X_j)]$.

We again show that an outlier strategy achieves a constant factor approximation to the best adaptive solution. Define the following "prize-collecting" Steiner problem over the points $\mathcal{V}$ in the metric space d: There is a cost of c_i for every node $i \in \mathcal{V}$. The goal is to choose a set of nodes $S \subseteq \mathcal{V}$ with total cost at most C to discard such that the value of the Steiner tree on the nodes in $\mathcal{V} \setminus S$ is minimized.

The following is the well-known linear relaxation of this problem. Define a variable z_i which is 1 of node $i \in \mathcal{V}$ is discarded and 0 otherwise. For $i, j \in \mathcal{M}$, define an indicator variable $x(i,j)$ if the MST includes an edge between these two nodes.

$$\text{Minimize} \quad \sum_{i,j \in \mathcal{M}} d(i,j) x(i,j)$$

$$\begin{array}{rcll} \sum_{k \in \mathcal{V}} c_k z_k & \leq & C & \\ \sum_{i \in S, j \notin S} x(i,j) + z_k + z_l & \geq & 1 & \forall S \subseteq \mathcal{M}, k \in S \cap \mathcal{V}, l \in \mathcal{V} \setminus S \\ z_k, x(i,j) & \in & [0,1] & \forall k \in \mathcal{M}, i, j \in \mathcal{V} \end{array}$$

Lemma D.2. *The value of the optimal adaptive strategy is at least the optimal value of the above linear program.*

Proof. Define a random variable Φ_k which is 1 if the optimal solution probes node $k \in \mathcal{V}$, and zero otherwise. We have $\mathbf{E}[\Phi_k] = \phi_k$. We have $\sum_{k \in \mathcal{V}} c_k \phi_k \leq C$ since the expected probing cost of the adaptive solution is at most C.

For any leaf l of the optimal decision tree, a certain set $S_l \subseteq \mathcal{V}$ of nodes are unprobed, and the locations of nodes in $\mathcal{V} \setminus S_l$ are known precisely. Let $P_l \subseteq \mathcal{P}$ denote the set of these locations. The solution T_l constructed by the adaptive solution for this leaf node l is the MST on the nodes $S_l \cup P_l$ in the metric space d. This tree is a Steiner tree on the unprobed nodes S_l.

Each leaf node l of the optimal decision tree generates one such Steiner tree T_l on the unprobed nodes S_l. For $i, j \in \mathcal{M}$, let $\Gamma(i,j)$ be a random variable which is 1 if the Steiner tree has a link between node i and node j, and 0 otherwise. Let $\gamma(i,j) = \mathbf{E}[\Gamma(i,j)]$.

The value of the Steiner tree is $\sum_{i,j} d(i,j) \Gamma(i,j)$. Therefore, the optimal adaptive solution is lower bounded by $\sum_{i,j} d(i,j) \gamma(i,j)$.

If two nodes $k, l \in \mathcal{V}$ are both unprobed, then $\Phi_k = \Phi_l = 0$ and the Steiner tree must connect k and l. Therefore, $\sum_{i \in S, j \notin S} \Gamma(i,j) + \Phi_k + \Phi_l \geq 1$ for all sets $S \subseteq \mathcal{M}$ such that $k \in S$ and $l \notin S$. This condition is trivially true if one of the nodes is probed. Therefore, it is true always. By linearity of expectation,

$$\sum_{i \in S, j \notin S} \gamma(i,j) + \phi_k + \phi_l \ \geq \ 1 \ \ \forall S \subseteq \mathcal{M}, k \in S \cap \mathcal{V}, l \in \mathcal{V} \setminus S$$

Setting $x(i,j) = \gamma(i,j)$ and $z_k = \phi_k$ now completes the proof. $\square$

Algorithm.

1. Round the above LP into an integer solution using the primal-dual algorithm of Goemans and Williamson [17]. This yields a Steiner tree on $S^* \subseteq \mathcal{V}$ which is a 4 approximation to the objective value such that the cost of the nodes in $\mathcal{V} \setminus S^*$ is at most $4C$.

2. Probe the variables corresponding to the nodes in $\mathcal{V} \setminus S^*$.

3. The locations for nodes in $\mathcal{V} \setminus S^*$ are known exactly and are now a subset $P^* \subseteq \mathcal{P}$. Output the MST on $S^* \cup P^*$ in metric space d.

Analysis. We now show that the non-adaptive algorithm described above has a constant factor adaptivity gap. Let $ADAPT$ denote the expected value of the optimal adaptive solution.

Lemma D.3. *The expected value of the MST on the probed set $\mathcal{V} \setminus S^*$ is at most $2 \cdot ADAPT$.*

Proof. In any realization of the values of nodes in $\mathcal{V} \setminus S^*$, the adaptive solution defines a Steiner tree connecting these nodes. Ignoring Steiner nodes at most doubles the cost of the tree. Therefore, the proof follows. □

Lemma D.4. *The value of the MST on the unprobed set S^* has value at most $8 \cdot ADAPT$.*

Proof. The rounding scheme returns a Steiner tree on nodes S^* in metric d whose value is 4 times the value of the fractional solution, whose value in turn is at most $ADAPT$ by Lemma D.2. Since the spanning tree on these nodes has value at most twice the value of the Steiner tree, the claim follows. □

Theorem D.5. *The non-adaptive algorithm has value at most $11 \cdot ADAPT$ while spending probing cost $4C$.*

Proof. The value of the non-adaptive algorithm is at most the expected value of the MST on the probed set plus the value of the MST on the unprobed set plus the expected cost of connecting some probed and unprobed nodes. The former two values add up to at most $10 \cdot ADAPT$ by Lemmas D.3 and D.4.

We now show that the expected value of connecting some pair of probed and unprobed nodes is at most $ADAPT$. Fix any node $i \in \mathcal{V} \setminus S^*$ and a node $j \in S^*$. In any scenario of the locations of these nodes, the value of the optimal solution is at least the distance between these nodes. This is true whether the nodes are probed or not. Therefore, the expected value of connecting i and j is at most $ADAPT$. This completes the proof. □

E Minimum MAKESPAN Scheduling

We now consider the problem of minimizing the makespan on identical parallel machines. In this problem, there are m identical machines, and n jobs, whose sizes are random variables, $X_1, X_2, \ldots, X_n$. We can probe job i by spending cost c_i, and find the exact value of its processing time. Given a bound on the total query cost C, the goal is to find the subset of variables to probe so that the expected value of the makespan (load on the most loaded machine) is minimized. We show a constant factor adaptivity gap for Bernoulli distributions; the proof for general distributions is similar. Let V denote the set of all jobs.

Roadmap: In the formulation below, we focus on minimizing the makespan of the unprobed jobs subject to the budget constraint on the cost of the probed jobs. The reason is the following. The probed and unprobed jobs can be scheduled separately to a factor 2 loss in approximation ratio. Also, the makespan of the probed jobs can be approximated to arbitrary accuracy with respect to the load of the optimal adaptive algorithm in any realization of their values. Therefore, the *outlier* problem which minimizes the makespan of just the unobserved jobs only loses a factor of 2 in the approximation ratio. For these jobs, we use the lower bounds developed by Kleinberg, Rabani, and Tardos [28], using Lemma 2.1 for large jobs, and a notion of "effective" size for small jobs.

In previous applications of the outlier technique, we showed that the adaptive solution is a lower bound to the LP relaxation to the outlier problem. However, showing that the optimal solution is feasible for the linear relaxation of the outlier problem is difficult in this case. We therefore use a simpler approach of blowing up the outlier cost by a factor of two, and arguing that the optimal *integer* solution to the new outlier problem is an approximate lower bound to the optimal adaptive solution. This illustrates that any outlier algorithm with approximation guarantees against the corresponding integer optimum can also be used for all the problems we discussed, provided the probing cost is a priori blown up by a factor of 2.

LP formulation. Each variable X_i is of the form $B(s_i, p_i)$, meaning that it has size s_i with probability p_i and size 0 otherwise (Bernoulli variables). The goal is to find a set T of jobs with probing cost at most $2C$ such that the expected makespan of scheduling $V \setminus T$ is minimized.

Assume the optimal load to be 1; this is achieved by guessing the optimal value and scaling. Split the jobs into "small" and "large" sizes, depending on the value of s_i. If $s_i > 1$, the job is large, and it is small otherwise.

We now use the lower bounds developed in [28]: For the jobs to be scheduled with load at most 1, the following must hold:

1. The sum of the expected sizes of the large jobs should be at most a constant.
2. The sum of the *effective sizes* of the small jobs averaged over the m machines should be at most a constant. The *effective size* of a job X_i, denoted $G(i)$ is equal to $\frac{\log \mathbf{E}[m^X]}{\log m}$.

Let A be the set of "small" jobs, and B be the set of large jobs. Let y_i denote whether job i belongs to T and z_i denote whether it is assigned to the set $V \setminus T$. The optimal load of scheduling $V \setminus T$ is at most 1 iff the following constraints are feasible.

$$\begin{array}{rcll} \sum_{i \in A} G(i) z_i & \leq & 18m & \\ \sum_{i \in B} p_i s_i z_i & \leq & 1 & \\ \sum_{i \in A \cup B} c_i y_i & \leq & 2C & \\ y_i + z_i & = & 1 & \forall i \in A \cup B \\ y_i, z_i & \in & \{0, 1\} & \forall i \in A \cup B \end{array}$$

Non-adaptive Algorithm. We solve the above LP for various guesses of the optimal load (in increasing order of value) until it is feasible. The rounding scheme is simple. Set $y_i = 1$ if it is greater than half in the fractional solution. This does not violate any constraint by more than a factor of 2, which implies (by [28]) that there is a solution whose expected load is at most a constant, and has outlier cost $4C$. The non-adaptive algorithm probes this set and schedules all the jobs using the scheduling algorithm for stochastic jobs in [28]. This yields a constant factor approximation to the makespan of the optimal non-adaptive algorithm.

Theorem E.1. *The above algorithm spends cost $4C$ and yields a constant factor approximation to the expected makespan of the best adaptive observation scheme.*

Proof. Let X^* denote the value of the optimal expected makespan on a set $V \setminus T$ of jobs conditioned on the probing cost of the set T of jobs is at most $2C$. The rounding procedure described above produces a solution whose expected makespan is $O(X^*)$ by spending probing cost at most $4C$.

Let indicator $\Phi_i = 1$ if job i is probed by the optimal adaptive solution. Let $\phi_i = \mathbf{E}[\Phi_i]$. We have $\sum_{i \in A \cup B} c_i \phi_i \leq C$. This implies $\Pr[\sum_{i \in A \cup B} c_i \Phi_i \leq 2C] \geq 0.5$. For the events where $\Pr[\sum_{i \in A \cup B} c_i \Phi_i \leq 2C]$, the expected makespan of the adaptive solution conditioned on that event is at least X^*. Therefore, the expected makespan of the adaptive optimum is at least $\frac{X^*}{2}$. This completes the proof. $\square$