Exceeding Expectations and Clustering Uncertain Data

Sudipto Guha*
Department of Computer and Information Sciences
University of Pennsylvania
Philadelphia, PA 19104-6389.
sudipto@cis.upenn.edu

Kamesh Munagala†
Department of Computer Science
Duke University
Durham, NC 27708-0129
kamesh@cs.duke.edu

ABSTRACT

Database technology is playing an increasingly important role in understanding and solving large-scale and complex scientific and societal problems and phenomena, for instance, understanding biological networks, climate modeling, electronic markets, etc. In these settings, uncertainty or imprecise information is a pervasive issue that becomes a serious impediment to understanding and effectively utilizing such systems. Clustering is one of the key problems in this context.

In this paper we focus on the problem of clustering, specifically the k-center problem. Since the problem is NP-Hard in deterministic setting, a natural avenue is to consider approximation algorithms with a bounded performance ratio. In an earlier paper Cormode and McGregor had considered certain variants of this problem, but failed to provide approximations that preserved the number of centers. In this paper we remedy the situation and provide true approximation algorithms for a wider class of these problems.

However, the key aspect of this paper is to devise general techniques for optimization under uncertainty. We show that a particular formulation which uses the contribution of a random variable above its expectation is useful in this context. We believe these techniques will find wider applications in optimization under uncertainty.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: [Non-numerical Algorithms and Problems]

General Terms

Approximation algorithms, Clustering

*Research supported in part by an Alfred P. Sloan Research Fellowship and an NSF CAREER Award CCF-0644119.

†Research supported by an Alfred P. Sloan Research Fellowsip and by NSF via a CAREER award and grant CNS-0540347.

PODS'09, June 29–July 2, 2009, Providence, Rhode Island, USA.

1. INTRODUCTION

Optimization problems arising in databases, for example, data integration, streaming, cluster computing and sensor network applications often involve parameters and inputs whose values are known only with some uncertainty. With the increasing use of information extraction techniques we would be gleaning information from larger and larger collections of data automatically, and that would imply developing tools and strategies for dealing with large collections of uncertain data efficiently. The notion of uncertainty is not new, and a long celebrated line of reasoning has suggested that the input be represented probability distributions. However, the sizes of the databases with probabilistic information has been unprecedented and is only growing. Developing formal methods for representation and optimization of probabilistic is an emerging challenge for database systems.

In this paper we focus on one of the core problems in representing a large collection of probabilistic input: namely clustering. The problem of clustering is not only central to data mining and visualization; but clustering is also an important optimization primitive that enables a host of other applications. Consider for example computing a join over two sets of data where each datapoint is a probabilistic distribution over strings. Clustering is an extremely useful preprocessing step in this case that cuts down on the number of pairs we may chose to consider. Given the centrality of the problem, it is surprising that there has been little formal investigation of clustering probabilitic data. The only prior work we are aware of, by Cormode and McGregor [6]: For independent distributions, in case of sums of (near) linear objective functions, such as the k-means and k-medians, they established that techniques such as linearity of expectation yield approximation algorithms. Interestingly, they showed that for thresholded functions such as the k-center problem, again for independent distributions, newer ideas are required and they succeeded in offering a bicriterion approximation for the *unassigned* case, where the assignment of points to the centers depends on the realization of the points. Their algorithm uses more than k centers and compares itself to an optimal clustering with k centers.

A natural question in this regard is to determine algorithms for k-center clustering for independent point clouds that: (a) do not violate the bound on the number of centers, and (b) are approximations even when the assignment of points to centers is done *before* the realization of the points (*assigned* version). We note that the latter is more reasonable since the assignment should be part of the output of the clustering procedure if the uncertainty is inherent to the

input. There is no prior result for this version of the problem (again formulated in [6]). Also note that since the k-center problem is already NP-Hard to approximate to a factor better than 1.8 even when the input is from a 2-dimensional plane [9], the best that can be hoped for is a constant factor approximation. We provide such an approximation algorithm that satisfies both (a) and (b), and is also an approximation to the unassigned version for which [6] only provides a bicriteria result.

More importantly, a related question in this regard is: What techniques are available to us to solve problems of this genre? The main focus of this paper is that we relate the thresholded "min max" objective to an optimization problem which behaves more closely to a "min sum" objective, via a novel metric truncation. The interesting twist in the connection is that for any independent collection of random variables, only the contributions from large (relative to the expectation) realizations are accounted. As a consequence, the uncertain k-center problem shows a behaviour which is similar to the k-median problem and the techniques for the latter can be applied to the former. This also has the added advantage of simplifying the space of objective functions one needs to consider in developing clustering algorithms for probabilistic data. To the best of our knowledge, the reduction of thresholded problems to min-sum problems in the context of clustering is novel.

We focus on independent distributions. This is quite justified, since Anthony *et al.* [1] show that for correlated distributions where the locations of the points are specified as scenarios, k-center clustering is as hard to approximate as the notorious densest subgraph problem, and is hence unlikely to have good approximation guarantees.

In a related thread, there has been ongoing investigations to quantify the benefit of selectively (due to resource constraints) resolving the uncertainty [21, 10, 12]. A natural question is approximation guarantees in this setting, and our techniques easily extend to this case.

Our Results: As indicated, our main result is to demonstrate that the thresholded "min-max" objective is related to the "min-sum" criterion, with only the larger values contributing. This technique is partly inspired by the $O(1)$ approximation algorithm for stochastic makespan scheduling on identical parallel machines, where the job sizes are independent random variables, due to Kleinberg, Rabani, and Tardos [20].

Using this approach, we provide a true contrant factor approximation for the k-center problem in *both* the unassigned and assigned versions, with independent distributions (point clouds) as input. We formally define the models in Section 2 and present the algorithms for the unassigned and assigned versions in Sections 3 and 4 respectively. Furthermore, the actual algorithms, shown in Figures 1 and 2, are extremely simple to describe: Perform parametric search on a truncated length function, using the primal dual algorithm for k-medians due to Jain and Vazirani [19] as a subroutine. Note that the primal-dual algorithm is combinatorial and simple to implement: In fact these same type of algorithms have been used in streaming settings as well and have been shown to perform well. We note that though our algorithms themselves are simple to describe, the analysis is not straightforward.

The interesting aspect is that we solve the k-median problem on a truncated space which is *not a metric*; however, we show that triangle inequality holds in a certain relaxed sense, and the Jain-Vazirani algorithm yields an approximation using this relaxed notion (see Section 5). It is however not clear how to extend other combinatorial approaches, such as local search [2] to yield approximation guarantees in this setting. Our notion of relaxation is novel and different from that arising in say, squaring the metric (k-means clustering).

We finally mention in Section 6 that the result extends in a straightforward fashion to the model where we can resolve some of the uncertainty by "probing" or refining the input data. We have chosen to give simpler proofs with weakened constants and have not focused on optimizing them in this extended abstract.

Previous Results. As mentioned before, the unassigned version of the probabilistic k-center problem (with independent distributions) has a bicriteria approximation in [6]. The same paper presents constant approximations for the probabilistic k-medians problem. As we already noted, since the latter is a sum objective, it is simpler to deal with in a stochastic setting as opposed to the k-center problem, which has a max objective. Anthony *et al* [1] present algorithms for stochastic versions of the k-medians problem, and show that the correlated scenario version of k-center is possibly hard to approximate.

The problems we study fall in the broad class of stochastic optimization problems. Our solution technique is partly inspired by the algorithm for stochastic makespan from [20]. Stochastic scheduling also has a large body of work [7, 8, 11, 22, 24], but the techniques are unrelated. A large body of work has also addressed two-stage stochastic optimization, where the probability distributions resolve gradually, but the cost of constructing the solution increases as the distributions resolve. Various covering problems such as set cover and Steiner tree are shown to have constant approximations in this model [18, 13, 16, 23, 25, 14, 15]. However, the techniques needed for these problems is very different.

We finally note that the deterministic version of the k-center problem has a 2-approximation due to Hochbaum and Shmoys [17]. Similarly, the deterministic k-medians problem has several constant factor approximations, for instance, see [2, 3, 4, 19].

2. PROBLEM STATEMENTS

Let $\mathcal{P}$ be a finite metric space of m points with distance function $d : \mathcal{P} \times \mathcal{P} \to \Re$. There are n input "nodes" V, where node p_i follows an *independent* distribution $\mathcal{D}_i$ over $\mathcal{P}$. These (discrete) distributions are specified as input. Note that with some probability the node may not appear at all; this aspect does not affect our algorithms or analysis. We use "point" to mean deterministic member of $\mathcal{P}$, and "node" to mean the input point-clouds, or set V, that need to be clustered. In the ensuing discussion, we use symbols such as u, v to denote points in $\mathcal{P}$ and centers; and symbols such as i, j to denote input nodes in V. Let $\Delta = \max_{u,v \in \mathcal{P}} d(u, v)$ denote the diameter of $\mathcal{P}$.

In the k-center problem, the goal is to choose a subset $S \subset \mathcal{P}$ of exactly k points, which are called "centers". Cormode and McGregor [6] define two possible optimization problems:

Unassigned Version: In this version, in any realization

$\sigma : V \to \mathcal{P}$ of the values of the input nodes, each node $i \in V$ (which is now realized as $\sigma(i) \in \mathcal{P}$) is assigned by mapping $\nu : \mathcal{P} \to S$ to its closest center in S. The goal is to find S of size k that minimizes $\mathbf{E}_\sigma \left[\max_{i \in V} d(\sigma(i), \nu(i))\right]$, where the expectation is over the realizations σ drawn from $\prod_{i=1}^{n} \mathcal{D}_i$. In other words, minimize the expected value of the maximum distance a node needs to travel to reach a center, where the expectation is over all realizations of the nodes, and where the assignment to the center can depend on the realization.

Assigned Version: In this case, the output is a subset of centers, $S \subset \mathcal{P}$ of size k, as well as a mapping $\pi : V \to S$ of the input nodes V to the centers. In every realization $\sigma : V \to \mathcal{P}$ of the values of the input nodes, node $i \in V$ (which is now realized as $\sigma(i) \in \mathcal{P}$) is assigned to the same center $\pi(i) \in S$. The goal is to find S of size k that minimizes $\mathbf{E}_\sigma \left[\max_{i \in V} d(\sigma(i), \pi(i))\right]$, where the expectation is over the realizations σ drawn according to $\prod_{i=1}^{n} \mathcal{D}_i$. In other words, minimize the expected value of the maximum distance a node needs to travel, when the assignment of nodes to the centers cannot depend on the realization.

We note that the above problem formulations differ in whether the assignment of the nodes to the centers is done upfront before the node values are realized (*assigned* case), or whether this assignment is *a posteriori* depending on the realization (*unassigned case*). We show a $O(1)$ approximation for *both* these versions using the same solution technique.

3. UNASSIGNED VERSION

Recall that in this problem, the goal is to find a subset $S \subset \mathcal{P}$ of k centers, so that in realization $\sigma : V \to \mathcal{P}$, suppose $\nu(i)$ denotes the closest node in S to $\sigma(i)$, then the following quantity is optimized: $\mathbf{E}_\sigma \left[\max_{i \in V} d(\sigma(i), \nu(i))\right]$.

Let OPT denote the value of the optimal solution. The key to the algorithm is to define a k-median problem on a different, truncated length function, solve this approximately via the primal-dual algorithm in [19], and perform a binary search on the truncation. The final algorithm is presented in Figure 1. We gradually build towards this algorithm by first defining the metric truncation, then the k-median problem, and finally our algorithm.

3.1 Truncated Length Function

DEFINITION 1. *Define the "truncated" length function over $\mathcal{P} \times \mathcal{P}$ as $\mathcal{L}_T(u, v) = \max(d(u, v) - T, 0)$.*

Even though the function d defines a metric (*i.e.*, it satisfies triangle inequality), the function $\mathcal{L}_T$, though symmetric, is *not* a metric. However, this function satisfies the following *relaxed* triangle inequality.

LEMMA 3.1. *For any points $u, v, w \in \mathcal{P}$, we have $\mathcal{L}_T(u, v) + \mathcal{L}_T(v, w) \geq \mathcal{L}_{2T}(u, w)$.*

PROOF. First note that $d(u, v) + d(v, w) \geq d(u, w)$ by the triangle inequality. Therefore,

$$\begin{aligned} & \mathcal{L}_T(u, v) + \mathcal{L}_T(v, w) \\ = \; & \max(d(u, v) - T, 0) + \max(d(v, w) - T, 0) \\ \geq \; & \max(d(u, v) + d(v, w) - 2T, 0) \\ \geq \; & \max(d(u, w) - 2T, 0) = \mathcal{L}_{2T}(u, w) \end{aligned}$$

□

DEFINITION 2. *For any point $v \in \mathcal{P}$, define its "weight" $w_v = \Pr\left[\cup_{i \in V} (\mathcal{D}_i = v)\right] \in [0, 1]$. Let X_u be the $0/1$ random variable denoting whether at least one node in V is mapped to $u \in \mathcal{P}$. Then $\mathbf{E}[X_u] = w_u$*

3.2 Truncated k-Medians

The crux of our algorithm is to define the following "truncated" k-median problem that we denote $\mathbf{UnMed}(T)$.

DEFINITION 3. *The problem $\mathbf{UnMed}(T)$ is defined over the graph $\mathcal{P}$ with edge lengths $\mathcal{L}_T$. Choose a subset S of k medians, and find an assignment $\pi : \mathcal{P} \to S$ which optimizes the cost:*

$$\mathcal{C}(S, T) = \sum_{u \in \mathcal{P}} w_u \mathcal{L}_T\left(u, \pi(u)\right)$$

Define $\mathcal{C}^*(T)$ denote $\min_{S \subset \mathcal{P}, |S| = k} \mathcal{C}(S, T)$, the optimal cost. Note that $\mathcal{C}^*(T)$ is non-increasing in T.

Though the problem $\mathbf{UnMed}(T)$ is defined over a non-metric space, note that $\mathcal{L}_T$ satisfies the relaxed triangle inequality from Lemma 3.1. This implies that any k-median approximation algorithm that starts with an approximate LP solution to this problem and performs a limited reassignment of the points to medians will yield a relaxed approximation. In particular, if the primal-dual algorithm of Jain and Vazirani [19] is run on the graph $\mathcal{P}$ with length function $\mathcal{L}_T$, it can be shown to achieve the following guarantee. *Note that the guarantee relaxes the value of T to $9T$.* We present details in Section 5.

THEOREM 3.2. *The primal-dual algorithm in [19] when run using lengths $\mathcal{L}_T$ finds a set of medians $S \subset \mathcal{P}$ of size k and a mapping $\pi : \mathcal{P} \to S$ that satisfies:*

$$\sum_{u \in \mathcal{P}} w_u \mathcal{L}_{9T}\left(u, \pi(u)\right) = \mathcal{C}(S, 9T) \leq 6\mathcal{C}^*(T)$$

DEFINITION 4. *Define*

$$\mathbf{PrimalDual}(T) = \sum_{u \in \mathcal{P}} w_u \mathcal{L}_{9T}\left(u, \pi(u)\right)$$

as the value of the solution $(S.\pi)$ for $\mathbf{UnMed}(T)$ found by the primal-dual algorithm (Theorem 3.2), when its value is computed using the truncated length function $\mathcal{L}_{9T}$.

3.3 Bounds on OPT using $\mathbf{UnMed}(T)$

Before presenting the final algorithm, we ask how does $\mathbf{UnMed}(T)$ relate to the optimal probabilistic k-center objective OPT? We first show a bound in the easier direction relating the two problems:

LEMMA 3.3. *For any set of k centers $S \subset \mathcal{P}$ and mapping $\pi : \mathcal{P} \to S$, let $\sigma : V \to \mathcal{P}$ denote the realizations of V, and in each such realization, let ν denote the mapping of $\sigma(i) \in \mathcal{P}$ to center $\nu(i) \in S$ that is induced by π. Then:*

$$\mathcal{C}(S, T) \geq \mathbf{E}_\sigma \left[\max_{i \in V} \mathcal{L}_T(\sigma(i), \nu(i))\right]$$

PROOF. The quantity $\mathcal{C}(S,T)$ has the following interpretation: Let X_u be the 0/1 random variable denoting whether at least one node in V is mapped to $u \in \mathcal{P}$. We have $\mathbf{E}_\sigma[X_u] = w_u$. Therefore:

$$\begin{aligned}\mathcal{C}(S,T) &= \sum_{u\in\mathcal{P}} w_u \mathcal{L}_T(u,\pi(u)) \\ &= \mathbf{E}_\sigma\left[\sum_{u\in\mathcal{P}} X_u \mathcal{L}_T(u,\pi(u))\right] \\ &\geq \mathbf{E}_\sigma\left[\max_{u\in\mathcal{P}} X_u \mathcal{L}_T(u,\pi(u))\right] \\ &= \mathbf{E}_\sigma\left[\max_{i\in V} \mathcal{L}_T(\sigma(i),\nu(i))\right]\end{aligned}$$

□

The crux of our analysis will be the strong Lemma 3.5 that relates OPT with the solution of the truncated k-medians problem. Before presenting the statement, we will present an important ingredient in the form of a general probability lemma, due to due to Kleinberg, Rabani, and Tardos [20]. We prove a slightly weaker version (using $T/3$ instead of $T/2$) here for completeness.

CLAIM 3.4 ([20]). *For a set Q of independent random variables, where $X_u \sim$ Bernoulli$(1,w_u)$ and $s_u \geq T$, suppose we have $\mathbf{E}[\max_{u\in Q} s_u X_u] < \frac{T}{3}$. Then $\sum_{u\in Q} w_u s_u < T$.*

PROOF. First observe that if $\sum_{u\in Q} w_u \geq \frac{2}{3}$, then

$$\mathbf{E}[\max_{u\in Q} X_u] \geq 1 - e^{-2/3} > \frac{1}{3}$$

Since $s_u \geq T$, we have

$$\mathbf{E}[\max_{u\in Q} X_u s_u] \geq T\mathbf{E}[\max_{u\in Q} X_u] > \frac{T}{3}$$

which is a contradiction. Therefore, $\sum_{u\in Q} w_u \leq \frac{2}{3}$.

Let $Q_u = Q \setminus \{u\}$. Let $\mathcal{E}_u$ denote the event that $X_u = 1$, and $X_v = 0$ for all $v \in Q_u$. Let $\Phi_u = \sum_{v\in Q_u} X_v$, so that:

$$\mathbf{E}[\Phi_u] = \sum_{v\in Q_u} w_v \leq \frac{2}{3}$$

Also note that $\mathcal{E}_u$ corresponds to $X_u = 1$ **and** $\Phi_u = 0$. By Markov's inequality, we have:

$$\Pr[\Phi_u = 0] = \Pr[\Phi_u < 1] \geq 1 - \mathbf{E}[\Phi_u] \geq \frac{1}{3}$$

By the independence of X_u and $\cup_{v\in Q_u} X_v$, we have

$$\begin{aligned}\Pr[\mathcal{E}_u] &= \Pr[X_u = 1 \wedge \Phi_u = 0] \\ &= \Pr[X_u = 1]\Pr[\Phi_u = 0] \\ &\geq \frac{1}{3}\mathbf{E}[X_u] = \frac{w_u}{3}\end{aligned}$$

The following inequalities now complete the proof.

$$\frac{T}{3} > \mathbf{E}\left[\max_{u\in Q} X_u s_u\right] \geq \mathbf{E}\left[\sum_{u\in Q} s_u \mathcal{E}_u\right] \geq \sum_{u\in Q} s_u \frac{w_u}{3}$$

□

We now use the above ingredient to formulate the important bound connecting the truncated k-median objective with OPT. The bound in the following lemma is illustrated in Figure 3(a).

LEMMA 3.5. $\mathcal{C}^*(0) \geq OPT \geq \frac{1}{3}\max\{T \mid \mathcal{C}^*(T) \geq T\}$.

PROOF. When $T = 0$ in Lemma 3.3, there is no truncation, so that the minimum of the RHS over all S, π is precisely OPT, while the minimum of the LHS is $\mathcal{C}^*(0)$. Therefore we have $\mathcal{C}^*(0) \geq OPT$.

We will now show that if $\mathcal{C}^*(T) \geq T$, then $OPT \geq \frac{T}{3}$. Towards this end, define a slightly different truncated length function $\mathcal{L}'_T$ over $\mathcal{P}\times\mathcal{P}$ as follows: $\mathcal{L}'_T(u,v) = d(u,v)$ if $d(u,v) \geq T$, and $\mathcal{L}'_T(u,v) = 0$ otherwise. Note that $\mathcal{L}'_T(u,v) \geq \mathcal{L}_T(u,v)$.

We will prove by contradiction. Suppose $OPT < \frac{T}{3}$. Recall that X_u be the 0/1 random variable denoting whether at least one node in V is mapped to $u \in \mathcal{P}$. We have $\mathbf{E}_\sigma[X_u] = w_u$. Let S denote the subset chosen by the optimal k-center solution, and let π denote the mapping of points in $\mathcal{P}$ to the closest center in S. We have:

$$\begin{aligned}\frac{T}{3} &> OPT = \mathbf{E}\left[\max_{u\in\mathcal{P}} X_u d(u,\pi(u))\right] \\ &\geq \mathbf{E}\left[\max_{u\in\mathcal{P}} X_u \mathcal{L}'_T(u,\pi(u))\right]\end{aligned}$$

Let Q denote the subset of $\mathcal{P}$ so that for $u \in Q$, $\mathcal{L}'_T(u,\pi(u)) = d(u,\pi(u))$. Note that for $u \notin S$, we have $\mathcal{L}'_T(u,\pi(u)) = 0$, and for $u \in Q$, we have $\mathcal{L}'_T(u,\pi(u)) \geq T$. Let $s_u = \mathcal{L}'_T(u,\pi(u))$ for $u \in Q$. Therefore, $s_u \geq T$.

From the above, we have $\mathbf{E}[\max_{u\in Q} X_u s_u] < \frac{T}{3}$. By Claim 3.4, this implies $\sum_{u\in Q} w_u s_u < T$. Therefore, we have

$$\begin{aligned}T &> \sum_{u\in Q} w_u s_u = \mathbf{E}\left[\sum_{u\in\mathcal{P}} X_u \mathcal{L}'_T(u,\pi(u))\right] \\ &\geq \sum_{u\in\mathcal{P}} w_u \mathcal{L}_T(u,\pi(u)) = \mathcal{C}(S,T) \geq \mathcal{C}^*(T)\end{aligned}$$

This contradicts our assumption that $\mathcal{C}^*(T) \geq T$, which proves the lemma. □

3.4 Probabilistic k-center Algorithm

With the above ingredients in place, the algorithm in Figure 1 is a $O(1)$ approximation to the unassigned version of the probabilistic k-center problem.

LEMMA 3.6. *The algorithm in Figure 1 terminates in polynomial time.*

PROOF. We can assume that all $d(u,v)$ are non-zero, if we collapse all points at distance 0 together as preprocessing. Note this implies that the Diameter $\Delta > 0$ as well as $C^*(0) > 0$.

Consider $T = \Delta$. We have $\mathcal{L}_\Delta(u,v) = 0$. Therefore, $\mathbf{PrimalDual}(\Delta) = 0 < 6\Delta$.

Consider

$$T \leq \min\left((1-9\epsilon)\frac{C^*(0)}{6}, \epsilon \min_{u,v} d(u,v)\right)$$

we have $\mathcal{L}_{9T}(u,v) \geq (1-9\epsilon)d(u,v)$, which implies

$$\mathbf{PrimalDual}(T) \geq (1-9\epsilon)\mathcal{C}^*(0) \geq 6T$$

Since the T is decreased in powers of $(1-\epsilon)$, this shows a polynomial running time in input size and $\frac{1}{\epsilon}$. □

Let T^* denote the value of T at which the algorithm stops. Let ALG denote the value of the probabilistic k-center solution found by the algorithm. The proof of the following lemma is also illustrated in Figure 3(b).

Algorithm for Probabilistic k-center: Unassigned Version

$\epsilon > 0$ is a small constant.

1. $T \leftarrow \Delta$. Recall Δ is the diameter of $\mathcal{P}$.
2. **while** (**PrimalDual**$(T) \leq 6T$) do: /* **PrimalDual**(T) (Def. 4) is the approximate solution to **UnMed**(T) */
   $T \leftarrow T(1-\epsilon)$.
   endwhile
3. **Output** the k median solution corresponding to **PrimalDual** $\left(\frac{T}{1-\epsilon}\right)$.

Figure 1: Algorithm for Probabilistic k-center: Unassigned Version.

LEMMA 3.7. *For $\epsilon < 1$ being a sufficiently small constant, we have $ALG \leq 15(1+2\epsilon)T^*$*

PROOF. Let $T' = T^*/(1-\epsilon)$ and let S^* denote the solution corresponding to **PrimalDual**(T'), *i.e.*, the k centers the algorithm outputs. From the execution of the algorithm: **PrimalDual**$(T') = \mathcal{C}(S^*, 9T') \leq 6T'$.

We now bound ALG as follows. For any realization of node values $\sigma : V \to \mathcal{P}$, let ν (which depends on σ) denote the closest center in S^* to the node.

$$\begin{aligned} ALG &= \mathbf{E}_\sigma \left[\max_i d(\sigma(i), \nu(i))\right] \\ &\leq 9T' + \mathbf{E}_\sigma \left[\max_i \mathcal{L}_{9T'}(\sigma(i), \nu(i))\right] \\ &\leq 9T' + \mathcal{C}(S^*, 9T') \leq 15T' \leq 15(1+2\epsilon)T^* \end{aligned}$$

The second inequality follows from Lemma 3.3. □

THEOREM 3.8. *The algorithm in Figure 1 is a $O(1)$ approximation to the unassigned version of probabilistic k-center.*

PROOF. The proof is a simple combination of Lemmas 3.5 and 3.7. From the stopping condition, **PrimalDual**$(T^*) > 6T^*$. But from Theorem 3.2, we have **PrimalDual**$(T^*) \leq 6\mathcal{C}^*(T^*)$, which implies $\mathcal{C}^*(T^*) > T^*$. From Lemma 3.5, this implies $OPT \geq \frac{T^*}{3}$. But from Lemma 3.7, we have $ALG \leq 15(1+2\epsilon)T^*$, which implies a $O(1)$ approximation. □

4. ASSIGNED VERSION

Again, the key to the algorithm is to define a k-median problem on a different, truncated length function. Note that in the assigned case, the goal is to output not only a set of centers S of size k, but also a mapping π from input nodes $i \in V$ to centers $u \in S$. Regardless of where this node is realized, it is mapped to the same center. Let OPT denote the value of the optimal solution.

For simplicity of exposition, we will assume the case that a node does not materialize does not arise; the case where a node can vanish with some probability is an easy extension, and omitted.

4.1 Truncated Length Function

In this case, we will work on the bipartite graph on $\mathcal{P} \cup V$, suitably defining a truncated distance between $i \in V$ and $u \in \mathcal{P}$. Let $\sigma : V \to \mathcal{P}$ denote realizations of nodes, so that node $i \in V$ is mapped to $\sigma(i) \in \mathcal{P}$.

DEFINITION 5. *Define $\rho_T(i,u)$ over $V \times \mathcal{P}$ as $\rho_T(i,u) = \mathbf{E}_\sigma [\mathcal{L}_T(\sigma(i), u)]$.*

As before, the length function ρ_T, though symmetric, is *not* a metric. However, as before, the crucial point is that this function satisfies the following *relaxed* triangle inequality. The proof now follows by linearity of expectation on Lemma 3.1.

LEMMA 4.1. *For any points $i, j \in V$ and $u, v \in \mathcal{P}$, we have $\rho_T(i,u) + \rho_T(j,u) + \rho_T(j,v) \geq \rho_{3T}(i,v)$.*

4.2 Truncated k-Medians

As before, we define the following "truncated" k-median problem that we denote **AsgMed**(T).

DEFINITION 6. *The problem **AsgMed**(T) is defined over the bipartite graph $V \times \mathcal{P}$. Choose a subset $S \subset \mathcal{P}$ of k medians, and find an assignment $\pi : V \to S$ which optimizes the cost:*

$$\mathcal{C}(S,T) = \sum_{i \in V} \rho_T(i, \pi(i))$$

Define $\mathcal{C}^*(T)$ denote $\min_{S \subset \mathcal{P}, |S|=k, \pi} \mathcal{C}(S,T)$, i.e.*, the optimal cost. $\mathcal{C}^*(T)$ is non-increasing in T.*

As before, though the problem **AsgMed**(T) is defined over a non-metric space, since ρ_T satisfies the relaxed triangle inequality from Lemma 4.1, we have the following:

THEOREM 4.2. *The primal-dual algorithm in [19] finds a set $S \subset \mathcal{P}$ of k medians, and mapping $\pi : V \to S$ that satisfies:*

$$\sum_{i \in V} \rho_{9T}(i, \pi(i)) = \mathcal{C}(S, 9T) \leq 6\mathcal{C}^*(T)$$

DEFINITION 7. *Define*

$$\textbf{PrimalDual2}(T) = \sum_{i \in V} \rho_{9T}(i, \pi(i))$$

*as the value of the solution $(S.\pi)$ for **AsgMed**(T) found by the primal-dual algorithm (Theorem 4.2), when its value is computed using the truncated length function ρ_{9T}.*

4.3 Bounding OPT using **AsgMed**(T)

Before presenting the final algorithm, we relate the k-median objective to the optimal probabilistic k-center objective OPT. We first bound the easier direction:

LEMMA 4.3. *For any set of k centers $S \subset \mathcal{P}$, and mapping $\pi : V \to S$, we have:*

$$\mathcal{C}(S,T) \geq \mathbf{E}_\sigma \left[\max_{i \in V} \mathcal{L}_T(\sigma(i), \pi(i))\right]$$

where $\sigma : V \to \mathcal{P}$ denotes the realizations of nodes.

PROOF. The quantity $\mathcal{C}(S,T)$ has the following interpretation: Let X_u be the 0/1 random variable denoting whether at least one node in V is mapped to $u \in \mathcal{P}$. We have $\mathbf{E}_\sigma[X_u] = w_u$. Therefore:

$$\begin{aligned}\mathcal{C}(S,T) &= \sum_{i\in V} \rho_T(i,\pi(i)) = \mathbf{E}_\sigma\left[\sum_{i\in V} \mathcal{L}_T(\sigma(i),\pi(i))\right] \\ &\geq \mathbf{E}_\sigma\left[\max_{i\in V} \mathcal{L}_T(\sigma(i),\pi(i))\right]\end{aligned}$$

□

As before, the crux of our analysis will be the following stronger lemma that relates OPT with the solution of the truncated k-medians problem.

LEMMA 4.4. $\mathcal{C}^*(0) \geq OPT \geq \frac{1}{3}\max\{T \mid \mathcal{C}^*(T) \geq T\}$.

PROOF. When $T = 0$ in Lemma 4.3, there is no truncation, so that the minimum of the RHS over all S is precisely OPT. Therefore we have $\mathcal{C}^*(0) \geq OPT$.

We will now show that if $\mathcal{C}^*(T) \geq T$, then $OPT \geq \frac{T}{3}$. Towards this end, as before, define $\mathcal{L}'_T$ over $\mathcal{P}\times\mathcal{P}$ as follows: $\mathcal{L}'_T(u,v) = d(u,v)$ if $d(u,v) \geq T$, and $\mathcal{L}'_T(u,v) = 0$ otherwise. Note that $\mathcal{L}'_T(u,v) \geq \mathcal{L}_T(u,v)$. Define random variable $R(i,u) = d(\sigma(i),u)$, and define $R_T(i,u) = \mathcal{L}'_T(\sigma(i),u)$, where $\sigma : V \to \mathcal{P}$ is the realization of the nodes. Since $\mathcal{L}'$ dominates $\mathcal{L}$, we have $\mathbf{E}_\sigma[R_T(i,u)] \geq \rho_T(i,u)$.

We will prove by contradiction. Suppose $OPT < \frac{T}{3}$. Let (S,π) denote the subset and mapping chosen by the optimal k-center solution. We have:

$$\begin{aligned}\frac{T}{3} &> OPT = \mathbf{E}\left[\max_{i\in V} R(i,\pi(i,S))\right] \\ &\geq \mathbf{E}\left[\max_{i\in V} R_T(i,\pi(i,S))\right]\end{aligned}$$

Let X_i denote the random variable $R_T(i,\pi(i,S))$. The above shows $\mathbf{E}[\max_{i\in V} X_i] < \frac{T}{3}$. Note further that each X_i is either 0 or takes on a value larger than T. We will now show this implies $\sum_i \mathbf{E}[X_i] < T$.

We express the distribution X_i as the maximum of a collection of Bernoulli distributions as follows. Suppose distribution X_i takes on values $T \leq s_1 < s_2 < \cdots < s_m$ with probabilities $p_{1i}, p_{2i}, \ldots, p_{mi}$. Create m independent Bernoulli variables $Y_{i1} = B(s_m, p_{mi})$, $Y_{i2} = B(s_{m-1}, \frac{p_{(m-1)i}}{1-p_{mi}})$, $Y_{i3} = B(s_{m-2}, \frac{p_{(m-2)i}}{1-p_{(m-1)i}-p_{mi}})$ and so on. We have $X_i = \max_{j=1}^m Y_{ij}$. Therefore, we have $\mathbf{E}[\max_{i\in V, j\in\{1,\ldots,m\}} Y_{ij}] < \frac{T}{3}$. From Claim 3.4, we have $\sum_{i\in V}\sum_{j=1}^m \mathbf{E}[Y_{ij}] < T$. However, we also have:

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

Therefore, $\sum_{i\in V}\mathbf{E}[X_i] < T$. Finally, we have

$$\begin{aligned}T &> \sum_{i\in V}\mathbf{E}\left[R_T(i,\pi(i,S))\right] \geq \sum_{i\in V}\rho_T(i,\pi(i,S)) \\ &= \mathcal{C}(S,T) \geq \mathcal{C}^*(T)\end{aligned}$$

This contradicts our assumption that $\mathcal{C}^*(T) \geq T$, which proves the lemma. □

4.4 Probabilistic k-center Algorithm

With the above ingredients in place, the algorithm in Figure 2 is a $O(1)$ approximation to the assigned version of the probabilistic k-center problem. The proof of the next lemma is identical to Lemma 3.6.

LEMMA 4.5. *The algorithm in Figure 2 terminates in polynomial time.*

Let T^* denote the value of T at which the algorithm stops. Let ALG denote the value of the probabilistic k-center solution found by the algorithm. We have:

LEMMA 4.6. *For $\epsilon < 1$ being a sufficiently small constant, we have $ALG \leq 15(1+2\epsilon)T^*$*

PROOF. Let $T' = T^*/(1-\epsilon)$ and let (S^*,π) denote the solution corresponding to **PrimalDual2**(T'), *i.e.*, the k centers the algorithm outputs and the mapping. From the execution of the algorithm, we have: **PrimalDual2**$(T') = \mathcal{C}(S^*, 9T') \leq 6T'$.

We now bound ALG as follows. Recall $\sigma : V \to \mathcal{P}$ is the realization of node values, so that node i is realized as point $\sigma(i) \in \mathcal{P}$

$$\begin{aligned}ALG &= \mathbf{E}_\sigma\left[\max_i d(\sigma(i),\pi(i))\right] \\ &\leq 9T' + \mathbf{E}_\sigma\left[\max_i \mathcal{L}_{9T'}(\sigma(i),\pi(i))\right] \\ &\leq 9T' + \mathcal{C}(S^*, 9T') \\ &\leq 15T' \leq 15(1+2\epsilon)T^*\end{aligned}$$

The second inequality follows from Lemma 4.3. □

THEOREM 4.7. *The algorithm in Fig. 2 is a $O(1)$ approximation to the assigned version of k-center.*

PROOF. The proof is a simple combination of Lemmas 4.4 and 4.6. From the stopping condition: **PrimalDual2**$(T^*) > 6T^*$. But from Theorem 4.2, we have **PrimalDual2**$(T^*) \leq 6\mathcal{C}^*(T^*)$, which implies $\mathcal{C}^*(T^*) > T^*$. From Lemma 4.4, this implies $OPT > \frac{T^*}{3}$. But from Lemma 4.6, we have $ALG \leq 15(1+2\epsilon)T^*$, which implies a $O(1)$ approximation. □

5. PRIMAL-DUAL ALGORITHM

We now give a sketch of Theorems 3.2 and 4.2. In both cases, the algorithms **PrimalDual**(T) and **PrimalDual**2(T) are *exactly* the primal-dual algorithm of Jain and Vazirani [19]. This algorithm works on a bipartite graph where one side A are the demand nodes, and the other side B are possible centers, and there are edges with lengths between these sides. In the **UnMed**(T) case, the bipartite graph has $A = \mathcal{P}$ and $B = \mathcal{P}$, where the edge length between $u \in A$ and $v \in B$ is $\mathcal{L}_T(u,v)$. Node $u \in A$ has demand w_u from Def. 2. In the case of **AsgMed**(T), the bipartite graph has $A = V$, and $B = \mathcal{P}$, where all demands are one, and the edge length between $i \in A$ and $v \in B$ is $\rho_T(i,v)$.

The description of the Jain-Vazirani algorithm *does not* need these distances to obey a metric; the metric property is only used in the analysis, particularly in triangle inequalities in Lemmas 5 and 10 of [19]. However, Lemmas 3.1 and 4.1 imply these inequalities are preserved if we successively relax T. We mention how this fact affects their key theorems;

Algorithm for Probabilistic k-center: Assigned Version

$\epsilon > 0$ is a small constant.

1. $T \leftarrow \Delta$. Recall Δ is the diameter of $\mathcal{P}$.
2. **while** (**PrimalDual2**$(T) \leq 6T$) do: /* **PrimalDual2**(T) (Def. 7) is the approximate solution to **AsgMed**(T) */
   $T \leftarrow T(1-\epsilon)$.
   endwhile
3. **Output** the k median solution corresponding to **PrimalDual2** $\left(\frac{T}{1-\epsilon}\right)$.

Figure 2: Algorithm for Probabilistic k-center: Assigned Version.

the details are straightforward by re-working the triangle inequalities in Lemmas 5 and 10 of [19].

We focus on **AsgMed**(T); the description for **UnMed**(T) is similar. First, let $O_T(\lambda)$ denote the value of the optimal solution to **AsgMed**(T) where there is no bound on the number of medians, but instead a cost λ to open a median. This is the *facility location* problem. In Theorem 7 of [19], they construct a $S \subseteq B$ and a mapping $\pi : A \rightarrow S$, so that:

$$\sum_{i \in A} \rho_{3T}(i, \pi(i)) + 3\lambda|S| \leq 3O_T(\lambda)$$

Note that this uses the triangle inequality over three edges, so that T needs to be relaxed to $3T$.

Next, they perform a binary search on λ to identify two close-by values where the corresponding sets S_1 and S_2 satisfy $S_1 \leq k$ and $S_2 > k$. They finally combine these to obtain a feasible solution S^* and mapping π^* with exactly k medians. In Lemma 11 of [19], they show the following:

$$\textbf{PrimalDual2}(T) = \sum_{i \in A} \rho_{9T}(i, \pi^*(i)) \leq 6\mathcal{C}^*(T)$$

The proof again needs a triangle inequality over three edges, so the $3T$ is relaxed to $9T$. This completes a sketch of the analysis. We re-iterate that the algorithm itself is identical to that in [19], and not described.

We note that the algorithm in [19] combines a greedy procedure with parametric search, and is hence combinatorial and efficient. Interestingly, we note that local search schemes such as [2] use the triangle inequality several times on any node, and hence do not yield approximation guarantees for truncated metrics.

6. OBSERVABLE DISTRIBUTIONS

An immediate extension is to study the benefit of *adaptive observations* in this context [12]. In this class of problems, a subset of nodes of cost at most C can be resolved prior to the optimization; the goal is to find the adaptive resolution procedure that optimizes the expected value of the subsequent optimization. We sketch a constant factor approximation for this version.

Formally, suppose distribution $i \in V$ could be observed and resolved, but there is a budget C on the total number of nodes that can be observed. Any observation policy adaptively observes a subset $\mathcal{M}$ of input nodes of size at most C. Let γ denote the outcome of the observations, and let $\mathcal{A}_\gamma$ denote the resulting problem instance where the nodes $i \in \mathcal{M}$ have resolved to deterministic points according to their respective $\mathcal{D}_i$, and the nodes in $V \setminus \mathcal{M}$ retain their original distributions. Subsequent to this, the probabilistic k-center problem is solved on this instance; let $\mathcal{K}_\gamma$ denote the optimal value of this solution. The goal is to find an adaptive observation policy with at most C observations, which minimizes $\mathbf{E}_\gamma[\mathcal{K}_\gamma]$. Denote this OPT. Note that the final expectation is over the outcomes γ of observations. In the ensuing discussion, we focus on the assigned center case; the unassigned case is similar.

For the probabilistic k-medians problem, a constant factor approximate observation scheme is presented in [12], that blows up the number of observations by a constant factor. This scheme can be directly adapted to probabilistic k-center. We present an outline below; since the details are very similar to that in [12], we only present the high-level ideas.

First observe that in any adaptive observation policy, for any outcome γ of the observations, at most a subset of nodes of size C has been observed. Therefore, the $\mathcal{K}_\gamma$ is at least the value of the probabilistic k-center solution on the unobserved nodes. Since this is true for all outcomes, we must have that OPT is at least the value of the optimal probabilistic k-center solution that can omit a subset of at most C nodes. Call this the *outlier* problem. This problem can be solved to a constant approximation by blowing up the outlier cost. This yields the following overall policy:

Outlier Problem: Solve the *outlier* version of **AsgMed**(T). In this problem (defined using the notation in Section 4.1), the goal is to choose the optimal subset $\mathcal{M} \subseteq V$ of size at most C to discard, so that the problem **AsgMed**(T) on nodes $V \setminus \mathcal{M}$ and possible center locations $\mathcal{P}$ has minimum k-median objective:

$$\Pi(\mathcal{M}) = \min_{S \subset \mathcal{P}, |S|=k, \pi: V \rightarrow S} \sum_{i \in V \setminus \mathcal{M}} \rho_T(i, \pi(i))$$

The primal-dual algorithm extends to an algorithm for the outlier problem [5], which yields a constant factor approximation, with T and outlier bound C relaxed by a constant factor. Let $\eta(T)$ denote the value of this relaxed approximation (analogous to **PrimalDual2**(T)).

Binary Search: Perform binary search to find the T^* for which $\eta(T^*) = T^*$. The same analysis as the non-outlier case shows that this yields a subset $\mathcal{M}^*$ with $|\mathcal{M}^*| = O(C)$, and a subset S^* of centers with $|S^*| = k$ with an assignment π, so that the value of the probabilistic k-center solution on $V \setminus \mathcal{M}$ is within a constant factor of the value of the optimal probabilistic k-center solution that is allowed to omit a subset of at most C nodes (whose value in turn is at most OPT from the above discussion).

Final Policy: Observe the nodes in $\mathcal{M}$. Subsequently, find a 2-approximate deterministic k-center solution [17] on

the observed nodes (denote the set of centers S'); and output $S^* \cup S'$ as the set of centers.

The final step yields $2k$ centers. Note that by the above argument, the value of the k-center solution on the unobserved nodes using centers S^* is $O(OPT)$. Further, in any realization of the nodes in $\mathcal{M}$, their k-center clustering is within factor 2 of best possible, so that the expected value of clustering these observed nodes over all possible realizations is at most $2OPT$. Therefore, the expected value of the entire policy is $O(OPT)$, which yields:

THEOREM 6.1. *The* outlier *algorithm yields a constant factor approximation to the probabilistic assigned (resp. unassigned) k-center problem when a set of at most C input nodes can be observed and resolved prior to the optimization. The solution relaxes the number of observed nodes by a constant factor, and uses $2k$ centers.*

7. CONCLUSIONS

In this paper, we have presented an intuitive approximation algorithm for the probabilistic k-center problem, that does not violate the number of centers. The algorithm uses a fairly general and powerful metric truncation paradigm to reduce the max version (k-center) to a sum version (k-median) on a different length function, for which we use an existing efficient primal-dual algorithm [19]. The interesting aspect of this k-median problem is that it is not over a metric space, but satisfies a novel relaxed notion of a metric. We note that our constant factor is deliberately large to keep the exposition simple.

We conclude with some open questions. Though the correlated scenario version of this problem is possibly hard to approximate [1], it would be nice to extend our results to more reasonable models of correlation. It would also be interesting to explore non-center based clustering methods, such as spectral approaches and hierarchical clustering in this model, as well as techniques for robustness (good with high probability) as opposed to optimizing expected value. Finally, the truncation technique is a simple and powerful method, that as already found applications to scheduling [20] and clustering (this paper). It would be interesting to explore other problems on probabilistic databases for which this technique applies.

8. REFERENCES

[1] B. M. Anthony, V. Goyal, A. Gupta, and V. Nagarajan. A plant location guide for the unsure. *Proceedings of SODA*, pages 1164–1173, 2008.

[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] 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.

[4] 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.

[5] 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.

[6] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. *Proceedings of PODS*, pages 191–200, 2008.

[7] 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.

[8] 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.

[9] Tomás Feder and Daniel Greene. Optimal algorithms for approximate clustering. In *Annual ACM Symp. on Theory of Computing*, pages 434–444, 1988.

[10] A. Goel, S. Guha, and K. Munagala. How to probe for an extreme value. *ACM Trans. Algorithms (to appear)*, 2008. Preliminary version appeared in Proc. of the ACM Symp. on Principles of Database Systems (PODS), 2006.

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

[12] S. Guha and K. Munagala. Model-driven optimization using adaptive probes. *CoRR, arXiv:0812.1012*, 2008. Preliminary version appeared in SODA 2007.

[13] 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.

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

[15] 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.

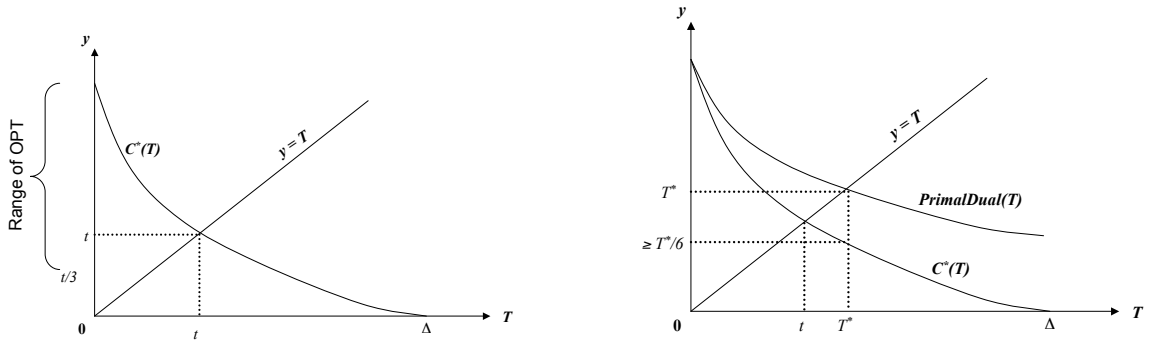
[16] A. Gupta, R. Ravi, and A. Sinha. An edge in time saves nine: LP rounding approximation algorithms for stochastic network design. In *Proc. of the Annual Symp. on Foundations of Computer Science*, pages 218–227, 2004.

[17] D. Hochbaum and D. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of Operations Research*, 10(2):180–184, May 1985.

[18] 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.

[19] 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.

[20] J. Kleinberg, Y. Rabani, and É. Tardos. Allocating bandwidth for bursty connections. *SIAM J. Comput*,



(a) (b)

Figure 3: (a) The bounds obtained using Lemma 3.5. Note that the function $\mathcal{C}^*(T)$ is monotonically non-increasing in T. (b) The point T^* obtained by the algorithm in Figure 1.

30(1), 2000.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] C. Swamy and D. B. Shmoys. Sampling-based approximation algorithms for multi-stage stochastic. In *FOCS*, pages 357–366, 2005.