On Two-Stage Stochastic Minimum Spanning Trees*

Kedar Dhamdhere[1], R. Ravi[2], and Mohit Singh[2]

[1] Dept. of Computer Science, Carnegie Mellon University,
Pittsburgh PA 15213
kedar@cs.cmu.edu
[2] Tepper School of Business, Carnegie Mellon University,
Pittsburgh PA 15213
{ravi, mohits}@andrew.cmu.edu

Abstract. We consider the undirected minimum spanning tree problem in a stochastic optimization setting. For the two-stage stochastic optimization formulation with finite scenarios, a simple iterative randomized rounding method on a natural LP formulation of the problem yields a nearly best-possible approximation algorithm.

We then consider the Stochastic minimum spanning tree problem in a more general black-box model and show that even under the assumptions of bounded inflation the problem remains $\log n$-hard to approximate unless $P = NP$; where n is the size of graph. We also give approximation algorithm matching the lower bound up to a constant factor.

Finally, we consider a slightly different cost model where the second stage costs are independent random variables uniformly distributed between $[0, 1]$. We show that a simple thresholding heuristic has cost bounded by the optimal cost plus $\frac{\zeta(3)}{4} + o(1)$.

1 Introduction

Stochastic optimization refers to problems where the inputs have an uncertainty, usually modeled by giving a probability distribution over the inputs. A common framework for stochastic optimization problem is two-stage stochastic optimization with recourse [2]. The uncertainty in the variables is modeled by probability distribution over a set of *scenarios* one of which will emerge tomorrow. *Recourse* is the ability to take corrective action tomorrow when one of the scenarios emerges.

This framework is well suited to network design problems where, in practice, the network has to be designed depending on the future demand patterns. One of the basic problems in network design is the minimum cost spanning tree problem. In this paper, we consider various models of the Stochastic Minimum Spanning Tree problem and give near optimal approximation algorithms for them.

* Supported in part by NSF grant CCR-0105548 and ITR grant CCR-0122581 (The ALADDIN project).

M. Jünger and V. Kaibel (Eds.): IPCO 2005, LNCS 3509, pp. 321–334, 2005.

Related Work. Stochastic Programming is well studied field with vast literature [16]. Recently, there has been work in designing approximation algorithm for the problems [6, 7, 9, 12, 15].

The models considered here are usually the two stage stochastic optimization with recourse. However, the variables in each scenario are correlated; e.g., in the network design problems considered by [6, 7, 9, 12], the costs of all edges in each scenario increases by the same factor.

In [15], Shmoys et al consider the problem of Stochastic set-cover and show how a ρ-approximation for the deterministic version of the problem can be used to obtain a 2ρ-approximation to the Stochastic version of the set-cover problem. Although, the spanning tree problem can be formulated as a set cover problem (see Section 3), the techniques of [15] do not yield a solution as the set cover problem is exponential in size. Also, the deterministic spanning tree problem when formulated as a set cover problem has a integrality gap of 2 while stochastic version of the problem is $O(\log n)$-hard to approximate unless $P = NP$. Hence, it is unlikely that techniques discussed in [15] will be applicable to Stochastic MST problem.

2 Models and Our Results

2.1 Stochastic MST with Explicit Scenarios

A popular formulation for stochastic optimization problems explicitly lists the finite set of scenarios which can occur tomorrow. The stochastic version of the minimum spanning tree problem under this model can be stated as follows: In the first stage we are given a complete graph $G = (V, E)$ on n nodes and a non-negative cost function c^0 on edges. In the second stage we have k scenarios and probabilities p^i, $1 \leq i \leq k$ of their occurrences. Each scenario has a different cost function c^i on the edges. The problem is select a subset of edges E^0 in the first stage and then augment it with E^i in the i^{th} scenario to obtain a spanning tree in each of the k scenarios. The objective function to minimize is the expected cost of the edges chosen. Note that, this cost is $c^0(E^0) + \sum_{i=1}^{k} p^i c^i(E^i)$. Observe that in this model the costs of the edges in any scenario need not be correlated. Although, the deterministic version of the problem has linear-time randomized algorithm [10], the stochastic version of this problem is hard. It has been shown in [3, 5] that the problem is hard to approximate within a factor ofmin$\{\log n$, $\log k\}$, by a reduction from set-cover.

We give the following result for this model:

Theorem 1. *There exists a polynomial time randomized algorithm which returns a solution of expected cost* $O((\log n + \log k) \cdot \mathrm{OPT})$ *with high probability, where* OPT *is the cost of the optimum solution of the Stochastic spanning tree problem on graph* $G = (V, E)$, $|V| = n$ *with* k *scenarios. The running time is polynomial in* k *and* n.

Observe that, when $k = poly(n)$ our algorithm gives the best approximation up to a constant factor.

We formulate a simple linear program for the stochastic spanning tree problem. We then use an optimal fractional solution to randomly round each edge. The techniques used for proving Theorem 1 are adapted from Alon[1].

2.2 Stochastic MST in the Black-Box Model

In this model, the scenarios are not stated explicitly but we have access to a Black-box from which we can sample the second stage scenarios. The samples are drawn from the same probability distribution as the second stage scenarios. Let $\lambda = max_{1 \leq i \leq k, e \in E}\{\frac{c_e^i}{c_e^0}, \frac{c_e^0}{c_e^i}\}$ denote the maximum cost inflation factor. We show that even when λ is polynomial in $n = |V(G)|$ the problem is hard to approximate.

Theorem 2. *Stochastic Spanning Tree problem on a graph $G = (V, E)$ with $n = |V(G)|$ with inflation λ is $O(\log n)$-hard to approximate unless $P = NP$ even when $\lambda = poly(n)$.*

We then give an approximation algorithm matching the lower bound proved in Theorem 2.

Theorem 3. *Given an instance of Stochastic Spanning Tree problem in the Black-box model with inflation factor λ, there exists a randomized algorithm which returns a solution of expected cost $O((\log n + \log \lambda) \cdot \mathrm{OPT})$ with high probability, where* OPT *is the cost of the optimum solution. The running time of the algorithm is polynomial in n and λ.*

Hence, if $\lambda = p(n)$ for some polynomial $p(n)$, we get an approximation algorithm optimal up to a constant factor. We reduce the Stochastic MST problem in the black-box model to the Stochastic MST with explicit scenarios by sampling $k = poly(n, \lambda)$ times and using these scenarios in the new problem constructed. We show that solving the Stochastic MST problem with these explicit scenarios gives a good solution to Stochastic MST problem in the Black-box model as well.

2.3 Stochastic MST with Independent Random Costs

A well-studied problem is to determine the cost of the minimum spanning tree of a complete graph when the edge costs are independent random variables uniformly distributed in $[0, 1]$. A classic result of Frieze [4] shows that the expected cost of the minimum spanning tree is $\zeta(3) + o(1)$. A natural stochastic version of the problem can be formulated when first-stage costs are given by a cost function c^0 and second-stage costs are independent random variables uniformly distributed between $[0, 1]$.

For this model, we analyze a thresholding heuristic. The thresholding heuristic with threshold α consists of excluding all edges with cost more than α in the first stage and constructing a minimum spanning tree over each of the resulting components. These components are then optimally joined in the second stage. We show that a thresholding heuristic gives a solution within $\frac{\zeta(3)}{4} + o(1)$ of the

optimum solution. We also show that if the optimum solution is 'tiny', then our solution is also small by showing that the expected cost of our solution is at most $\sqrt{3\zeta(3) \cdot \mathrm{OPT}} + o(1)$.

Theorem 4. *The thresholding heuristic with the threshold* $\frac{\zeta(3)}{n}$ *gives a solution whose expected cost is at most* $\mathrm{OPT} + \frac{\zeta(3)}{4} + o(1)$*, where* OPT *is the expected cost of the optimal solution. The expected cost is also bounded by* $\sqrt{3\zeta(3) \cdot \mathrm{OPT}} + o(1)$.

Another stochastic version of the problem, when both the first stage and second stage costs are independent random variables uniformly distributed between $[0, 1]$, was studied by Flaxman et al[3]. Flaxman et al[3] show that a thresholding heuristic gives a solution of cost $\zeta(3) - \frac{1}{2} + o(1)$, while $\frac{\zeta(3)}{2} + c$ for some $c > 0$ is a lower bound for the optimum. Surprisingly, the best threshold for their problem is $\frac{1}{n}$.

3 Stochastic MST with Explicit Scenarios

Linear Programming Formulation. The LP formulation for the two-stage Stochastic MST is the standard formulation enforcing the requirement that every non-trivial cut must be covered by an edge chosen in the first stage or, in each scenario, it must be covered with an edge chosen in the second stage. Although this LP formulation has an integrality gap of 2 in the deterministic setting, for stochastic version of the MST problem the optimal solution can be rounded to obtain a near optimal approximation algorithm. We call the following linear program the cut-cover LP.

$$
\begin{array}{lll}
 & \min \ \sum_e c_e^0 x_e^0 \ + \ \sum_e \sum_{i=1}^{k} p^i \, c_e^i x_e^i & \\
s.t. & & \\
 & \sum_{e \in \delta(S)} x_e^0 + x_e^i \ \geq \ 1 & \forall S \subset V, \ 1 \leq i \leq k \\
 & x_e^i \geq 0 & \forall e \in E, \ 0 \leq i \leq k
\end{array}
$$

We solve the the cut-cover linear programming formulation of the 2-stage spanning tree problem. Although, the linear program is not compact, an efficient separation subroutine can be supplied via an invocation to a min-cut subproblem. Alternately, equivalent compact reformulations can be readily obtained. (For eg. see the survey on spanning tree formulations by Magnanti and Wolsey [11]).

Rounding the Linear Program. Let $\hat{x}$ denote the optimal solution. We construct k forests, one for each scenario, by rounding the LP solution $\hat{x}$ in phases. Initially, all k forests have singleton components. In each phase, we pick edge e independently with probability $\hat{x}_e^0$ and include it in each of the k forests. Also for each i, $1 \leq i \leq k$, we pick edge e independently with $\hat{x}_e^i$ and include it in the i^{th} forest. Clearly, the expected cost of edges included in each phase is precisely OPT where OPT is the cost of the optimal LP solution $\hat{x}$. We argue that in one phase, for each of the k forests, the number of components decrease by a factor

of $\frac{9}{10}$ with probability at least $\frac{1}{2}$. We then argue that in $O(\log n + \log k)$ phases, with very high probability, each of the k forests will have just one component, i.e., a feasible solution for the 2-stage spanning tree problem and the expected cost of the solution is $O(\log n + \log k) \cdot \mathrm{OPT}$ as claimed. We elaborate on the above outline.

Lemma 1. *The expected cost of the edges paid in any phase is at most* OPT.

Proof. An edge e is included in the first stage with probability $\hat{x}_e^0$ and in the i^{th} scenario with probability $\hat{x}_e^i$. Hence, the expected cost paid for including this edge is $c_e^0 \hat{x}_e^0 + \sum_{i=1}^k p^i c_e^i \hat{x}_e^i$. Hence, the total expected cost paid in any phase is $\sum_{e \in E}(c_e^0 \hat{x}_e^0 + \sum_{i=1}^k p^i c_e^i \hat{x}_e^i)$ which is exactly OPT. □

Let F_j^i denote the i^{th} forest after j phases. Let C_j^i denote the number of components in F_j^i. We call a phase j 'successful' for scenario i, if $C_j^i < 0.9 C_{j-1}^i$ or if $C_{j-1}^i = 1$. We now state a lemma from [1]. For completeness, we also include the proof in the Appendix.

Lemma 2 (Alon [1]). *For every i, j, the conditional probability that phase j is successful for scenario i, given any set of components in F_{j-1}^i, is at least $\frac{1}{2}$.*

Lemma 3. *After $t = (40 \log n + 16 \log k)$ phases, the probability that any of the k forests, F_t^i, $1 \le i \le k$, is not connected is at most $\frac{1}{(kn)^2}$.*

Proof. Observe that the i^{th} forest gets connected after at most $\log_{0.9} n < 10 \log n$ successful phases. Hence, if F_t^i is not connected, then there have been at most $10 \log n$ successful phases for the i^{th} scenario. The probability of this event is no more than the probability that we get at most $10 \log n$ heads in t independent tosses of a fair coin. Although, the event that a phase j is successful for scenario i is not independent for different j, we can apply the above bound since Lemma 2 gives a lower bound on success given any history. Using estimates for tails of Binomial distributions, this probability is at most

$$e^{-(10 \log n + 8 \log k)^2 / 2 \times (20 \log n + 8 \log k)} \le e^{-(10 \log n + 8 \log k)/4} \le \frac{1}{(kn)^2}.$$

Using union bound, we get that the probability that any of the k forests is not connected is at most $k \times \frac{1}{(kn)^2} = \frac{1}{kn^2}$. □

Hence, after $40 \log n + \log k$ phases with very high probability each of the k forests are connected and from Lemma 1, the expected cost of the solution is $(40 \log n + \log k) \cdot \mathrm{OPT}$ proving Theorem 1.

Remark 1. If the randomized rounding fails to connect any forest after $O(\log n + \log k)$ phases, then we use a simple k-approximation algorithm: for each scenario build a tree by running Kruskal's algorithm on edge costs equal to the minimum of first and second stage costs. This will guarantee that our algorithm always *builds a spanning tree; further, since the failure probability is at most $\frac{1}{(kn)^2}$ per scenario the expected cost of our solution is still bounded by $O(\log n + \log k)$ times the optimum.*

Remark 2. Theorem 1 shows that the cut-cover LP formulation has a integrality gap of $O(\log n)$ assuming $k = poly(n)$. Also, the reduction from set cover given by Gupta [5], yields examples where the cut-cover LP has a integrality gap of $\Omega(\log n)$. Hence, the integrality gap of the cut-cover LP formulation is $\theta(\log n)$.

4 Approximation Algorithm for Black-Box Model

In this section, we consider the black-box model for the Stochastic spanning tree problem. We assume that the inflation factor λ is bounded by some polynomial $p(n)$. Note that the problem still remains $O(\log n)$-hard to approximate unless P = NP (Theorem 2).

Let TRUE-LP denote the linear programming formulation of the stochastic spanning tree problem. Given any $\epsilon > 0, \delta > 0$, we sample from the black box, according to probability distribution ρ, the second stage scenarios $k = poly(n, \lambda, \frac{1}{\epsilon}, \frac{1}{\delta})$ times. We then formulate a new stochastic MST problem using these k samples as the second stage scenarios each occurring with probability $\frac{1}{k}$. We call the corresponding linear program of the new instance of the problem formed SAMPLE-LP.

Let $\hat{x}^0$ denote the first stage component of the optimal solution of SAMPLE-LP. We show that if $\hat{x}^0$ is used as the first stage solution for TRUE-LP, with probability $1 - \delta$, we can extend this solution in any scenario with total expected cost $(1 + \epsilon)\,\mathrm{OPT}$ where is the expected cost of the optimal solution of TRUE-LP. Hence, using the result of Theorem 1, we round this LP solution to get an integral solution of expected cost $O(\log n \cdot \mathrm{OPT})$. Observe that, to obtain the first stage solution we only need to know the first stage fractional variables $\hat{x}^0$ and need not know the second stage distribution or second stage solution.

Given any x^0, the first stage variables, let $f(x^0)$ denote the cost of extending x^0 to a feasible solution on a random sample according to probability distribution ρ and let $F(x^0) = \mathbf{E}\big[f(x^0)\big]$. Let $F_k(x^0)$ be the random variable denoting the average completion cost over k independently sampled scenarios, i.e., $F_k(x^0) = \frac{1}{k}\sum_{i=1}^{k} f_i(x^0)$ where $f_i(x^0)$ is the random variable denoting the cost of extending x^0 in the i^{th} sample. The key to bounding our sample size is Lemma 4. The techniques of [15, 14] also can also be used to prove Lemma 4, we here give a much simpler proof. For simplicity, we assume that probability distribution ρ is discrete and finite but the same argument also works for continuous distributions.

Lemma 4. *Given any $\epsilon > 0$ and $\delta > 0$, if $k = \frac{\lambda^4}{\epsilon^2\delta}$, then $Pr[|F_k(x^0) - F(x^0)| < \epsilon \cdot \mathrm{OPT}] \leq \delta$. Here,* OPT *is the expected cost of the optimal solution.*

Proof. Clearly $\mathbf{E}\big[F_k(x^0)\big] = \frac{1}{k}\sum_{i=1}^{k} \mathbf{E}\big[f_i(x^0)\big] = \frac{1}{k}\sum_{i=1}^{k} F(x^0) = F(x^0)$. Also,

$$Var[F_k(x^0)] = \frac{1}{k^2}\sum_{i=1}^{k} Var[f_i(x^0)] = \frac{1}{k} Var[f(x^0)] \tag{1}$$

as $f^i(x^0)$ are independent random variables with distribution identical to $f(x^0)$.

Now,

$$Var[f(x^0)] = \sum_t \rho^t (c^t(x^0) - F(x^0))^2 \tag{2}$$

where ρ^t is the probability of the t^{th} scenario and $c^t(x^0)$ is the cost paid in this scenario to extend x^0 to a feasible solution. Let j denote the index for which $c^t(x^0)$ is minimum. As we can select exactly the edges selected in the j^{th} scenario in any other scenario to extend x^0 and inflation factor is bounded by λ, hence

$$c^t(x^0) \leq \lambda^2 c^j(x^0) \tag{3}$$

Also, if we do not select anything in the first stage our cost only goes up by a factor of λ and hence $\sum_t \rho^t c^t(x^0) \leq \lambda \cdot \mathrm{OPT}$. As $c^j(x^0)$ is minimum, this implies

$$c^j(x^0) \leq \lambda \cdot \mathrm{OPT} \tag{4}$$

From Equation (2), (3) and (4), we have

$$\begin{array}{rcl} Var[f(x^0)] & \leq & \sum_t \rho^t c^t(x^0)^2 \\ & \leq & (\lambda^2 c^j(x^0)) \sum_t \rho^t c^t(x^0) \\ & = & \lambda^2 c^j(x^0) \cdot \lambda \cdot \mathrm{OPT} \\ & \leq & \lambda^4 \, \mathrm{OPT}^2 \end{array} \tag{5}$$

Hence, using Equation (1),

$$Var[F_k(x^0)] = \frac{1}{k} Var[f(x^0)] = \epsilon^2 \delta \cdot \mathrm{OPT}^2 \tag{6}$$

Hence, by the Chebychev inequality we have,

$$Pr[|F_k(x^0) - F(x^0)| > \epsilon \cdot \mathrm{OPT}] \leq \frac{\epsilon^2 \delta \cdot \mathrm{OPT}^2}{\epsilon^2 \cdot \mathrm{OPT}^2} = \delta \tag{7}$$

□

Proof of Theorem 3. Let $\hat{x}^0$ denote the first stage component of the optimial solution of SAMPLE-LP formed using $k = \frac{\lambda^4}{\epsilon^2 \delta}$ samples. Also let $\bar{x}^0$ denote the first stage optimal solution to TRUE-LP. Using Lemma 4 twice, once for $\hat{x}^0$ and once for $\bar{x}^0$, we have with probability at least $1 - 2\delta$ that

$$|F_k(\hat{x}^0) - F(\hat{x}^0)| \leq \epsilon \cdot \mathrm{OPT}$$
$$|F_k(\bar{x}^0) - F(\bar{x}^0)| \leq \epsilon \cdot \mathrm{OPT}$$

As $\hat{x}^0$ is the optimal solution to SAMPLE-LP and $\bar{x}^0$ is the optimal solution to TRUE-LP, we have

$$F_k(\hat{x}^0) \leq F_k(\bar{x}^0),\ F(\bar{x}^0) \leq F(\hat{x}^0)$$

Using the above we have that

$$F(\hat{x}^0) \leq F(\bar{x}^0) + 2\epsilon \cdot \mathrm{OPT} \tag{8}$$

Hence, sampling $k = \frac{\lambda^4}{\epsilon^2\delta}$ times and solving the SAMPLE-LP gives first stage variables which can be extended to second stage with total cost at most $(1+\epsilon)\,\mathrm{OPT}$ with probability at least $1-2\delta$. Hence, now if we apply Theorem 1 to SAMPLE-LP, we obtain the first stage component of a solution of total cost $O((\log n + \log\lambda + \log\frac{1}{\epsilon\delta})\,\mathrm{OPT})$ with probability at least $1-\delta$ proving Theorem 3. □

5 Stochastic MST with Random Costs

In this section, we consider the two stage stochastic MST when the second stage costs are given by independent random variables uniformly distributed between $[0,1]$.

Thresholding. A thresholding heuristic with threshold α consists of removing all edges in the first stage with costs more than α and constructing a minimum cost spanning tree over each of the components formed. The components are then joined in an optimal manner in the second stage.

Observe that, for a forest F, the cost of extending F to a spanning tree in the second stage depends only on the sizes of the components of F. Let the components of F be $(C_1,\ldots,C_k)$. Construct an auxillary graph F' with k vertices v_i, $1 \le i \le k$ where v_i corresponds to component C_i of F. Include $|C_i|\cdot|C_j|$ edges between v_i and v_j, where the cost of edges are independent random variables distributed uniformly between $[0,1]$. Clearly, the cost of extending F to a spanning tree in the second stage is the cost of minimum spanning tree of F' which only depends on the sizes of the components of F.

In Lemma 6, we show that if we join k components then the expected average cost of each edge bought tomorrow is between $(\frac{k}{n})\frac{\zeta(3)}{n} + o(\frac{1}{n})$ and $\frac{\zeta(3)}{n} + o(\frac{1}{n})$. Hence, it is 'reasonable' not to buy any edge which costs more than $\frac{\zeta(3)}{n}$ and buy the edges which cost less than $\frac{\zeta(3)}{n}$ in the first stage itself.

First, we prove a lemma which shows how the cost of the second stage depends on the sizes of the components formed in the first stage. Note that the cost of extending a forest F to a spanning tree in the second is the exactly the cost of minimum cost spanning tree in the auxillary graph F' defined above.

Lemma 5. *Let G and G' be two forests, where the components of G are $(C_1, C_2, \ldots, C_k)$ and the components of G' are $(C_1\cup\{x\}, C_2\setminus\{x\}, C_3,\ldots,C_k)$ for some vertex $x \in C_2$. If $|C_1| \ge |C_2|$, then the expected cost of extending G to a spanning tree in the second stage is less than the cost of extending G' to a spanning tree.*

Proof. Let $\mathbf{E}[c(G)]$ denote the expected cost of extending G to a spanning tree in the second stage. Consider graph H which has components $(C_1, C_2', x, C_3, \ldots, C_k)$, where $C_2' = C_2 \setminus x$. If we connect up x to C_1 for free, then we get G', and if we connect x to C_2' for free, then we get G. Therefore,

$$\mathbf{E}[c(H)] \ge \mathbf{E}[c(G)] \qquad \text{and} \qquad \mathbf{E}[c(H)] \ge \mathbf{E}[c(G')]$$

We will show that the reduction in cost when connecting x to C_1 is larger. This will imply that $\mathbf{E}[c(G)] \leq \mathbf{E}[c(G')]$.

Let G_p denote the graph formed by including in G every edge independently with probability p. Similarly, define G'_p and H_p. In order to bound expected cost of spanning trees on G and G', we look at the expected number of connected components in G_p, G'_p and H_p. Let $\chi(G_p)$ and $\chi(G'_p)$ denote the number of connected components of G_p and G'_p respectively. We will show that $\mathbf{E}[\chi(G_p)] \leq \mathbf{E}\big[\chi(G'_p)\big]$. Note that this suffices in proving the above claim, since $\mathbf{E}[c(G)] = \int_0^1 (\mathbf{E}[\chi(G_p)] - 1)\, dp$.

Note that we can obtain G_p (resp. G'_p) from H_p by connecting x to C_1 (resp. C'_2) in H_p. We focus only on the number of connected components.

We use the *principle of deferred decision*. We first reveal the edges coming out of vertex x in H_p. If x has edges to both C_1 and C'_2 or to neither of them, then adding an edge from x to either C_1 or C'_2 and revealing rest of the edges in the graph gives the expected number of connected components.

Now observe that since $|C_1| > |C'_2|$, the probability that x is connected to component C_1 in H_p is reater than the probability that x is connected to C'_2. Therefore, G'_p formed by including x to C'_2 in H_p, has higher probability of reducing the number of connected components by 1 than in G_p, formed by including x in C_1. From this it follows that $\mathbf{E}[\chi(G_p)] \leq \mathbf{E}\big[\chi(G'_p)\big]$. This proves the lemma. □

The cost of extending a forest with k components to a spanning tree in the second stage can now be bounded as follows. The proof involving some careful calculations appears in the Appendix.

Lemma 6. *Given a forest F with k components, the expected cost of extending F to a spanning tree in the second stage is between $(\frac{k}{n})^2\zeta(3)+o(1)$ and $\frac{k}{n}\zeta(3)+o(1)$.*

Using this lemma, we now prove the main result of this section.

Proof of Theorem 4. Now, we prove Theorem 4 which bounds the expected cost of the solution returned by a thresholding heuristic. First, observe that the optimum solution must buy a forest in the first stage and join the components of the forest in the second stage. Suppose, the thresholding heuristic and the optimum solution buys k and opt_1 edges respectively, in the first stage. Then, they buy $n-k-1$ and $n-opt_1-1$ edges respectively, in the second stage. We form two cases depending on which of k and opt_1 is larger.

Case I ($k \leq opt_1$). Let the cost of the k edges bought by the thresholding heuristic be $c_1(k)$. List the opt_1 edges included in the optimum solution in the first stage in increasing costs. Clearly, cost of the first k edges in the order is at least $c_1(k)$ as we chose the cheapest k acyclic edges. Also, rest of the $opt_1 - k$ edges in the sequence must cost at least $\frac{\zeta(3)}{n}$ as any acyclic subgraph of edges with cost at most $\frac{\zeta(3)}{n}$ is of size at most k. If $\mathbf{E}[c(G)]$ denotes the expected cost of the thresholding heuristic and *Loss* denotes $\mathbf{E}[c(G)] - \mathrm{OPT}$,then by Lemma 6 we have,

$$\mathbf{E}[c(G)] \leq c_1(k) + \left(\frac{n-k-1}{n}\right)\zeta(3) + o(1)$$

$$\text{and} \quad \text{OPT} \geq c_1(k) + (opt_1 - k)\frac{\zeta(3)}{n} + \left(\frac{n-opt_1-1}{n}\right)^2\zeta(3) - o(1)$$

$$\Rightarrow \quad Loss \leq \left(\frac{n-opt_1-1}{n}\right)\zeta(3) - \left(\frac{n-opt_1-1}{n}\right)^2\zeta(3) + o(1)$$

Case II ($k > opt_1$). Let the cost of the opt_1 edges bought by the optimum solution be $c_1(opt_1)$. Order the k edges bought by the thresholding heuristic in increasing order of costs. The cost of the first opt_1 edges in this order is at most $c_1(opt_1)$ as we buy the cheapest acyclic subgraph of size opt_1. Also, each of the last $k - opt_1$ edges in the order costs at most $\frac{\zeta(3)}{n}$ as we threshold at $\frac{\zeta(3)}{n}$. Hence, using Lemma 6, we get

$$\mathbf{E}[c(G)] \leq c_1(opt_1) + (k - opt_1)\frac{\zeta(3)}{n} + \left(\frac{n-k-1}{n}\right)\zeta(3) + o(1)$$

$$\text{and} \quad \text{OPT} \geq c_1(opt_1) + \left(\frac{n-opt_1-1}{n}\right)^2\zeta(3) - o(1)$$

$$\Rightarrow \quad Loss \leq \left(\frac{n-opt_1-1}{n}\right)\zeta(3) - \left(\frac{n-opt_1-1}{n}\right)^2\zeta(3) + o(1)$$

Now, $(\frac{n-opt_1-1}{n})\zeta(3) - (\frac{n-opt_1-1}{n})^2\zeta(3) \leq \frac{\zeta(3)}{4}$ for any value of opt_1. Hence, we get that

$$\mathbf{E}[c(G)] \leq \text{OPT} + \frac{\zeta(3)}{4} + o(1)$$

for both of the cases, proving the first claim in the theorem.

Now observe that $\text{OPT} \leq \zeta(3)$ and consequently, each of the terms in the expression of OPT must be less than $\zeta(3)$. Using this, in case I, we have that

$$\begin{aligned}
\mathbf{E}[c(G)] &\leq c_1(k) + \left(\frac{n-k-1}{n}\right)\zeta(3) + o(1) \\
&= c_1(k) + \left(\frac{opt_1-k}{n}\right)\zeta(3) + \left(\frac{n-opt_1-1}{n}\right)\zeta(3) + o(1) \\
&\leq \sqrt{3\left(c_1(k)^2 + \left(\left(\frac{opt_1-k}{n}\right)\zeta(3)\right)^2 + \left(\left(\frac{n-opt_1-1}{n}\right)\zeta(3)\right)^2\right)} + o(1) \\
&\leq \sqrt{3(\zeta(3)c_1(k) + \zeta(3)(opt_1-k)\frac{\zeta(3)}{n} + \zeta(3)(\frac{n-opt_1-1}{n})^2\zeta(3))} + o(1) \\
&\leq \sqrt{3\zeta(3)\cdot\text{OPT}} + o(1)
\end{aligned}$$

The second step of the inequality follows from the fact that $\sqrt{3(a^2+b^2+c^2)} \geq a+b+c$ for any reals a, b, c.

In the second case, we have that

$$\begin{aligned}
\mathbf{E}[c(G)] &\leq c_1(k) + \left(\frac{k - opt_1}{n}\right)\zeta(3) + \left(\frac{n - opt_1 - 1}{n}\right)\zeta(3) \ + \ o(1) \\
&= c_1(opt_1) + \left(\frac{n - opt_1 - 1}{n}\right)\zeta(3) \ + \ o(1) \\
&\leq \sqrt{2\left(c_1(opt_1)^2 + \left(\left(\frac{n - opt_1 - 1}{n}\right)\zeta(3)\right)^2\right)} \ + \ o(1) \\
&\leq \sqrt{2\left(\zeta(3)c_1(opt_1) + \zeta(3)\left(\frac{n - opt_1 - 1}{n}\right)^2\zeta(3)\right)} \ + \ o(1) \\
&\leq \sqrt{2\zeta(3)\cdot \mathrm{OPT}} \ + \ o(1)
\end{aligned}$$

Here, the second step follows from $\sqrt{2(a^2+b^2)} \geq a + b$ for any reals a, b. Hence, $c(T) \leq \sqrt{3\zeta(3)\cdot \mathrm{OPT}} + o(1)$ as claimed.

References

1. N. Alon, *A note on network reliability,* Discrete Probability and Algorithms, D. Aldous et al., eds., IMA Volumes in mathematics and its applications, Vol. 72, Springer Verlag (1995), 11-14.
2. J. Birge and F. Louveaux. *Introduction to Stochastic Programming.* Springer, 1997.
3. A D Flaxman, A M Frieze and M Krivelevich. *On the random 2-stage minimum spanning tree.* Preprint (2004).
4. A M Frieze. On the value of a random minimum spanning tree problem, *Discrete Applied Mathematics 10 (1985) 47-56.*
5. Anupam Gupta. *Personal Communication (2004).*
6. Anupam Gupta, R. Ravi, Amitabh Sinha. *Boosted Sampling: Approximation Algorithms for Stochastic Optimization.* In Proceedings of the thirty-sixth annual ACM symposium on Theory of computing, 2004.
7. Anupam Gupta, R. Ravi, Amitabh Sinha. *An edge in time saves nine: LP rounding Approximation Algorithms for Stochastic Network Design.* In the Proceedings of 45th Annual IEEE Symposium on Foundations of Computer Science, 2004.
8. U. Feige. *A threshold of* $\ln n$ *for approximating set cover.* J. ACM 45, 634-652,1998.
9. Nicole Immorlica, David Karger, Maria Minkoff, Vahab S. Mirrokni. *On the Costs and Benefits of Procastination: Approximation Algorithms for Stochastic Combinatorial Optimization Problems*, Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms. Pages: 691 - 700, 2004.
10. David Karger, Philip Klein, Robert Tarjan. *A randomized linear-time algorithm to find minimum spanning trees*, Journal of the ACM, Volume 42, Issue 2, Pages: 321 - 328, 1995.
11. T. L. Magnanti, L. A. Wolsey. *Optimal Trees, Handbook in OR and MS*, Volume 7, Eds M.O. Ball et al, 503-615, 1995.
12. R. Ravi, Amitabh Sinha. *Hedging Uncertainty: Approximation Algorithms for Stochastic Optimization Problems.* In Proceedings of the 10th International Conference on Integer Programming and Combinatorial Optimization (IPCO), 2004.

13. Ran Raz, S. Safra. *A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP.* Proceedings of the twenty-ninth annual ACM symposium on Theory of computing, pages 475-484, 1999.
14. A. Shapiro. *Monte Carlo sampling approach to stochastic programming.* Proceedings of 2003 MODE-SMAI Conference Pau, France,pages 65-73, March 27-29, 2003
15. David B. Shmoys, Chaitanya Swamy. *Stochastic Optimization is (Almost) as easy as Deterministic Optimization* Proceedings of 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS'04), October 17 - 19, 2004 Rome, Italy
16. M. H. van der Vlerk. *Stochastic programming bibliography.* World Wide Web, http://mally.eco.rug.nl/spbib.html, 1996-2003.

Appendix A: Hardness of Stochastic MST with Bounded Inflation

In this section, we prove Theorem 2 stating that Stochastic MST with polynomial inflation is $O(\log n)$-hard to approximate unless $NP \subseteq \text{DTIME}(n^{O(\log \log n)})$. We give an approximation preserving transformation from the Set Cover problem to the Stochastic MST with inflation $\lambda = p(n)$ for some polynomial $p(n)$.

Given an instance of set cover problem, i.e, a set $U = [n]$ and collection of subsets $S_1, \ldots, S_m$, we construct an instance of stochastic spanning tree problem with $m+n+1$ vertices. Let the vertices be $\{v_1, \ldots, v_n, s_1, \ldots, s_m, r\}$. The vertex v_i corresponds to element $i \in U$ and s_j corresponds to subset S_j.

The first stage cost of edge (r, s_j) is 1, for each $1 \leq j \leq m$ and cost of all other edges is mn. See Figure 1(a). The edges not present have cost mn.

There are n scenarios, one corresponding to each element $i \in U$, each having a probability $\frac{1}{n}$ of appearance in the second stage. In the i^{th} scenario, each edge of cut separating $T_i = \{v_i\} \cup \{s_j : i \in S_j\}$ from $G \setminus T_j$ has cost mn and remaining edges have cost $\frac{1}{mn}$. See Figure 1(b). The edges absent have cost mn.

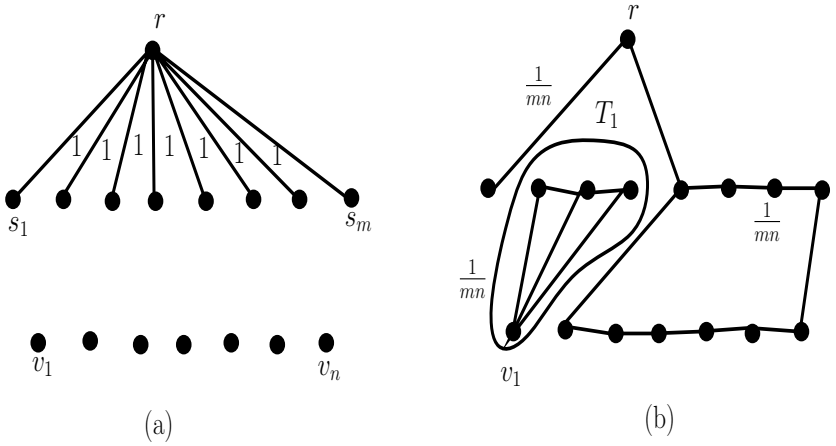

Fig. 1. (a) First stage costs (b) Second stage scenario corresponding to v_1

It is clear that the inflation factor of any edge is bounded by m^2n^2. Given any set cover C of size k, choose edges (r, v_j) such that $S_j \in C$ in the first stage. In the i^{th} scenario, this first stage solution can be extended to a spanning tree with edges of cost $\frac{1}{mn}$. Hence, there exists a feasible solution of stochastic spanning tree of total cost at most $k + \frac{1}{n}$. Given a feasible solution to the stochastic spanning tree of total cost $c \leq m + 1$, we construct a feasible set cover of cost between $c - \frac{1}{n}$ and c. Clearly, the solution contains no edge of cost mn. Let $C = \{S_j : (r, s_j)$ is selected in the first stage$\}$. In the i^{th} scenario, we cannot buy any edge of cost mn. Hence, the i^{th} element must have been covered by some set in C. Hence, the sets selected form a set cover. Also, the total cost paid in the second stage is at most $\frac{1}{n}$. Hence, the size of C is at least $c - \frac{1}{n}$. The cost of the set cover is exactly the first stage cost paid in the solution.

Hence, the hardness of the approximation result for the set cover [8, 13] imply that the stochastic spanning tree is at least $O(\log n)$-hard to approximate unless $P = NP$.

Appendix B: Proofs

Proof of Lemma 2. If F^i_{j-1} is connected, then the claim is trivial. Given some forest F^i_{j-1}, shrink each component to a singleton vertex and call the constructed multi-graph H. Consider any vertex v in H. An edge e is not included in F^i_j with probability $(1-x^0_e)(1-x^i_e)$. Let $\delta(v)$ denote the neighborhood of node v. Hence, the probability that v remains isolated is

$$\prod_{e\in\delta(v)} (1 - x^0_e)(1 - x^i_e) \leq exp(- \sum_{e\in\delta(v)} (x^0_e + x^i_e))$$

Using the fact that $\sum_{e\in\delta(v)}(x^0_e + x^i_e) \geq 1$, we get that the probability that v is isolated is at most $\frac{1}{e}$. Using linearity of expectation, the expected number of isolated vertices in H is $|H|/e$, and hence with probability at least $\frac{1}{2}$, the number of isolated vertices is less than $2|H|/e$. Hence, the number of connected components in F^i_j is at most

$$\frac{2|H|}{e} + \frac{1}{2}(|H| - 2|H|/e) = (\frac{1}{2} + \frac{1}{e})|H| < 0.9|H|$$

Using that $|H| = C^i_{j-1}$, we get the desired result.

Proof of Lemma 6. Let $\mathbf{E}[c(F)]$ denote the expected cost of extending the forest F to a spanning tree. By Lemma 5, we get that $\mathbf{E}[c(F)]$ is minimum when the k components of F are of equal sizes and it is maximum when all except one component are singletons. We calculate the expected costs of joining k components in either cases and get appropriate bounds as claimed.

Consider the case when F has one component of size $n - k + 1$ and rest $k - 1$ components are singleton vertices. Then, the expected cost of joining the components optimally is

$$\mathbf{E}[c(F)] = \int_{p=0}^{1} \mathbf{E}[\kappa(F_p) - 1]\, dp \tag{9}$$

where F_p is the graph formed when each edge is included in F, independently, with probability p and $\kappa(F_p)$ denotes the connected component of F_p. Now, using the ideas from [3], we need to consider components of size at most $(\log n)^2$. There are at most $\frac{n}{(\log n)^2}$ components of size greater than $(\log n)^2$ and each can be joined with probability $1 - o(1)$ with an edge of cost at most $\frac{\log n}{n}$. Hence, the total cost of joining such components is $o(1)$. Also, we can assume that $n - k + 1 \geq (\log n)^2$, otherwise the claim holds trivially as otherwise $\frac{k-1}{n}\zeta(3) + o(1) \geq \frac{n-(\log n)^2}{n}\zeta(3) + o(1) = \zeta(3) + o(1)$. Hence, we assume that the that the largest component of F is not included in a component of size more than $(\log n)^2$. Using the above observations we have that,

$$\begin{aligned}
\mathbf{E}[c(F)] &= \int_{p=0}^{1} \sum_{j=1}^{(\log n)^2} \binom{k-1}{j} j^{j-2}\, p^{j-1}(1-p)^{j(n-j)+O(j^2)} dp \;+\; o(1) \\
&= \sum_{j=1}^{(\log n)^2} \binom{k-1}{j} j^{j-2} \int_{p=0}^{1} p^{j-1}(1-p)^{jn}\, dp \;+\; o(1) \\
&= \sum_{j=1}^{(\log n)^2} \binom{k-1}{j} j^{j-2} \frac{(j-1)!(jn)!}{(j+jn)!} + o(1) \\
&= \sum_{j=1}^{(\log n)^2} \frac{(k-1)^j}{j!} j^{j-2} \frac{(j-1)!}{j^j n^j} + o(1) \\
&= \sum_{j=1}^{(\log n)^2} \left(\frac{k-1}{n}\right)^j \frac{1}{j^3} + o(1) \\
&\leq \frac{k-1}{n} \sum_{j=1}^{(\log n)^2} \frac{1}{j^3} + o(1) = \frac{k}{n}\zeta(3) + o(1)
\end{aligned}$$

Now consider the case when all k components are nearly of equal sizes. Assume for the sake of simplicity that all the components are of size $\frac{n}{k}$. Then F is equivalent to a multi-graph on k vertices, each vertex representing a component in F. The multi-graph has $(\frac{n}{k})^2$ edges between any two vertices. We replace the $(\frac{n}{k})^2$ edges edges with a single edge whose cost is a random variable defined to be the minimum of $(\frac{n}{k})^2$ independent random variables uniformly distributed between $[0, 1]$. Observe that the new random variable has is distributed around the origin like a uniform random variable between $[0, (\frac{k}{n})^2]$. Hence, using the result by Frieze [4], the cost of the minimum cost spanning tree is $(\frac{k}{n})^2\zeta(3)$. The argument goes through even in the case when k does not divide n and all components differ in sizes by at most one and the same bound holds. The calculations are tedious and are omitted.