

Randomized Approximation Algorithms

13.1 MAX-CUT

Consider the MAX-CUT problem, in which we are given a graph $G = (V, E)$ and wish to partition it into two parts so as to *maximize* the number of edges between the two parts. Previously, we showed that a greedy algorithm that iteratively considers each vertex and places it on the side with the fewest of its neighbours is a 2-approximation. We show here an alternate construction using randomization to achieve the same bound.

The algorithm is as follows. For each vertex, randomly assign it to one side or the other. To analyze this algorithm, we consider the expected number of edges cut, which equals

$$\sum_{e \in E} \Pr(e \text{ cut}) = \sum_{e \in E} 1/2 = m/2.$$

Thus at least half the edges are cut on average, and we have a randomized 2-approximation.

13.2 MAX-SAT

Consider a Boolean formula in conjunctive normal form (CNF), namely a formula expressed as the AND of a set of clauses, each of which is the OR of a set of literals, each of which is either a variable or the negation of a variable.

The satisfiability problem (SAT) is to determine whether or not a variable assignment exists that makes the formula true. This is one of the canonical NP-complete problems.

The *maximum satisfiability* problem (MAX-SAT) is to find a variable assignment that maximizes the number of true clauses.

One common variant of satisfiability is k SAT, in which each clause is restricted to be the OR of k literals, for fixed k . There is a straightforward polynomial time algorithm for 2SAT, but 3SAT is NP-complete.

We may similarly define MAX- k SAT to be MAX-SAT restricted to those formulas containing only clauses of size k . It turns out, however, that even MAX-2SAT is NP complete. We consider here, the approximation algorithm for MAX-3SAT consisting of making each variable true or false at random.

We have

$$\mathbb{E}[\text{number of clauses satisfied}] = \sum_{\text{clause } c} \Pr(c \text{ is satisfied}) \quad (13.1)$$

$$= \frac{7}{8}(\text{number of clauses}), \quad (13.2)$$

implying that we have an 8/7-approximation algorithm.

Note that a slight generalization of the above result implies a $\frac{1}{1-2^{-k}}$ -approximation algorithm for MAX- k SAT, which implies a 2-approximation for MAX-SAT.

We also note that it has been shown that the 8/7 approximation ration for MAX-3SAT is optimal unless P=NP.

13.3 LP relaxation of MAX-SAT

Consider the following integer linear program (ILP). Let binary variable z_j be 1 iff clause c_j is satisfied. Let binary variable y_i be 1 iff variable i is true. Let c_j^+ denote the set of variables appearing non-negated in clause c_j , and let c_j^- denote those variables that are negated. The problem is as follows:

$$\max \sum_j z_j \text{ such that } \forall j, \sum_{i \in c_j^+} y_i + \sum_{i \in c_j^-} (1 - y_i) \geq z_j, 0 \leq z_j \leq 1, 0 \leq y_i \leq 1.$$

It is fairly clear that this expresses the MAX-SAT problem if the variables are restricted to be integral. We consider the LP that results from relaxing this restriction.

We note that the solution we get in the relaxed problem may have fractional values for the variables y_i . Since these variables are meant to represent whether the variables x_i in the Boolean formula are true or not, we may have difficulty interpreting the LP solution as a variable assignment.

We introduce the technique of *randomized rounding*, a technique applicable to ILPs in general.

Given possibly fractional solutions y_i , randomized rounding sets $x_i = 1$ with probability y_i , and $x_i = 0$ with probability $1 - y_i$.

We note that a trivial consequence of this is that

$$\mathbb{E}[x_i] = y_i.$$

We now work to estimate the probability that a given clause is satisfied. From the constraints on the LP, we know that the expected number of satisfied variables in a clause c_j is at least z_j . Define a sequence β_k by

$$\beta_k = 1 - \left(1 - \frac{1}{k}\right)^k = \left(1, \frac{3}{4}, \dots \rightarrow 1 - \frac{1}{e} \approx .632\right).$$

Claim 1 Given a clause c_j of k variables, the probability that it is satisfied is at least $\beta_k z_j$.

Proof: To simplify the argument, we assume that all literals are positive (non-negated). We note that the probability that c_j is satisfied is the complement of the probability that none of its variables is true, which is

$$\prod_{i \in c_j} (1 - y_i).$$

We note that the y_i s are constrained by

$$\sum_{i \in c_j} y_i \geq z_j.$$

Thus appealing to inequality tricks (Jensen's inequality and the fact that the logarithm function is convex) we conclude that this probability is minimized when

$$y_i = z_j/k.$$

Thus we can lower bound the probability of c_j being satisfied by

$$1 - \left(1 - \frac{z_j}{k}\right)^k \geq \beta_k z_j,$$

as desired. ■

We now observe that this claim implies that this algorithm is a $\frac{1}{1-1/e} = \frac{e}{e-1}$ -approximation: since the sequence β_k is always at least $1 - \frac{1}{e}$, the expected number of satisfied clauses is at least $(1 - \frac{1}{e}) \sum z_j$, which is thus within $(1 - \frac{1}{e})$ factor of optimal since the solution to the relaxed problem $\sum z_j$ is at least as good as the solution to the ILP.

13.4 Combined algorithm

We introduce a third algorithm to approximate MAX-SAT that consists of running the previous two algorithms and picking the solution that maximizes the number of satisfied clauses. While randomized rounding provides a solution at least half as good as OPT, and LP relaxation provides a solution within $1 - \frac{1}{e}$ factor, the combined algorithm will provide a solution within 3/4 of optimal.

We note that the above algorithm performs at least as well as the algorithm where we pick randomly between the two choices suggested by randomized rounding and LP relaxation. Letting k_j denote the number of variables in clause j , we have that the expected number of variables satisfied in the random solution is at least

$$\frac{1}{2} \left[\sum_j (1 - 2^{-k_j}) z_j + \sum_j (\beta_{k_j}) z_j \right] = \sum_j \frac{1}{2} (1 + \beta_{k_j} - 2^{-k_j}) \geq \frac{3}{4} \sum_j z_j.$$

Thus this is a 4/3-approximation, as desired.

13.5 Set Cover

As before, in the *set cover* problem we are given sets S_i containing elements in $\{1\dots n\}$ and asked to find the minimum number of sets such that their union “covers” all the elements. We have already seen a greedy algorithm that produces an $O(\log n)$ -approximation. We present an LP relaxation algorithm that also provides an $O(\log n)$ -approximation.

Consider the following ILP. Let $y_i = 1$ iff S_i is in the cover. Find

$$\min \sum y_i, \text{ such that} \quad (13.3)$$

$$\sum_{S_i \ni j} y_i \geq 1 \quad (13.4)$$

$$0 \leq y_i \leq 1. \quad (13.5)$$

Observe that when y_i is restricted to be integral this captures the set cover problem. Now consider the result of optimizing the LP relaxation. If we try the above randomized rounding technique of putting set S_i in our cover with probability y_i , then we end up with the expected number of sets covering any point j to be 1. However, the set cover problem requires that *all* sets be covered, which is a considerably stronger statement. Thus regular randomized rounding might fail to give a solution to set cover.

We remedy this problem by modifying the randomized rounding scheme as follows: for some fixed $\alpha \geq 1$, choose set S_i with probability αy_i . Note that the expected number of sets chosen is now $\alpha \sum y_i \leq \alpha |\text{OPT}|$, so that we will have an α -approximation if the sets actually do cover $1\dots n$.

In order to show that the chosen sets do provide a cover with high probability, we introduce the *Chernoff bound*, another fundamental technique in the probabilistic method.

Theorem 1 (Chernoff Bound) *Let $X = X_1 + \dots + X_n$ be a sum of indicator (0-1) variables that are independent. Denote $E[X]$ by μ . Then*

$$\Pr(X \leq (1 - \epsilon)\mu) \leq e^{-\epsilon^2 \mu / 2}.$$

And if $\epsilon \leq 2e - 1$, then

$$\Pr(X \geq (1 + \epsilon)\mu) \leq e^{-\epsilon^2 \mu / 4},$$

while if $\epsilon > 2e - 1$, then

$$\Pr(X \geq (1 + \epsilon)\mu) \leq e^{-(1+\epsilon)\mu}.$$

To apply this theorem, we note that the number of sets covering an element j is the sum of independently random indicator variables, with mean α . We need to bound the probability that this random variable is 0. We apply the Chernoff bound with $\epsilon = 1$, which states that the probability that no set covers j is at most

$$e^{-\epsilon^2 \mu / 2} = e^{-\alpha / 2}.$$

Applying the union bound, we have that the probability that *any* element is uncovered is at most

$$ne^{-\alpha / 2}.$$

Setting $\alpha = 3 \log_2 n$, this probability becomes at most $1/\sqrt{n} \ll 1$. Thus we have a Monte Carlo $O(\log n)$ approximation, as desired.