Approximation Algorithms - Tools and Techniques **Spring 2002**

Lecture 8 : Feb 6, 2002

Lecturer: Kamesh Munagala [1] *Scribe: Siddartha Y R*[2]

Analysis of Karp's Scheme for Euclidean TSP

In the previous lecture, we outlined the Karp's scheme for an n point Euclidean TSP and showed that the tour returned by it satisfies the relation

$$C_{KARP} \leq C^* + O\left(\sqrt{\frac{n}{s}}\right)$$

where C_{KARP} is the cost of the tour returned by Karp's Algorithm, C^* is the optimal TSP on the given problem instance and s was chosen as $s = \frac{\log n}{\log \log n}$. We will know show that this leads to a good approximation if the points are i.i.d., in $[0,1]^2$. In such a case, the expected value of the optimal TSP, C^* itself is large, and the approximation ratio for the Karp's Algorithm improves since

$$\begin{aligned} C_{KARP} &\leq C^* + O\left(\sqrt{\tfrac{n}{s}}\right) \\ \Rightarrow \quad E\left[C_{KARP}\right] &\leq E\left[C^*\right] + O\left(\sqrt{\tfrac{n}{s}}\right) \\ &\leq E\left[C^*\right]\left(1 + \frac{O\left(\sqrt{\frac{n}{s}}\right)}{E\left[C^*\right]}\right) \end{aligned}$$

Thus, by determining a bound on $E\left[C^*\right]$ under the above conditions, we obtain a probabilistic approximation ratio for Karp's algorithm.

Theorem 1 *If the n points of an instance of a Euclidean* TSP *are i.i.d. over $[0,1]^2$, then*

$$E\left[C^*\right] \geq \frac{\sqrt{n}}{O(1)}$$

[1]Dept of Comp. Sci., Stanford Univ.

[2]Dept of Comp. Sci. and Automation, Indian Institute of Science, Bangalore.
`siddu@csa.iisc.ernet.in`

Proof: We have an upper bound on C^* from C_{NAIVE}, the tour returned by the naive algorithm -

$$C^* \leq C_{NAIVE} \leq 2\sqrt{n} + O(1)\,;$$

which also implies

$$E[C^*] \leq E[C_{NAIVE}] \leq 2\sqrt{n} + O(1)\,.$$

Let $d(x_i)$ be the distance from x_i to it's nearest neighbour

$$d(x_i) = \min_{i \neq j} \|x_j - x_i\|.$$

This gives a lower bound on C^* as

$$C^* \geq \sum_{i=1}^{n} d(x_i)$$

and from the linearity of expectation, we also have

$$\begin{aligned} E[C^*] \geq & \quad \textstyle\sum_{i=1}^{n} E[d(x_i)] \\ \geq & \quad \textstyle\sum_{i=1}^{n} E[d(x_1)] \\ \geq & \quad nE[d(x_1)] \end{aligned} \tag{1}$$

where in the second step $d(x_i)$ has been replaced with $d(x_1)$, which we can assume w.l.o.g., to be the maximum among all $d(x_i)$,

From Lemma. 1 (See Box), we have

$$E[d(x_1] = \int_0^\infty Prob[d(x_1) \geq r]\, dr$$

Lemma 1 *For any nonnegative r.v. Y*

$$E[Y] = \int_0^\infty Prob[Y \geq y]\,dy$$

Proof: Define $F(y)$ as

$$F(y) = Prob[Y \geq y]; \quad f(y) = -F'(y)$$

Then, $E[y]$ is given by

$$\begin{aligned} E[Y] = & \int_0^\infty yf(y)dy \\ = & \int_0^\infty -yF'(y)dy. \end{aligned}$$

Integrating by parts, this gives

$$\begin{aligned} E[Y] = & \; -yF(y)]_0^\infty + \int_0^\infty F(y)dy \\ = & \; 0 - 0 + \int_0^\infty F(y)dy \\ = & \int_0^\infty Prob[Y \geq y]\,dy \end{aligned}$$

■

Consider $Prob[d(x_1) \geq r]$. This is the probability that all the other points lie outside a circle of radius r centred at x_1, as shown if Figure. 1. For a distribution of x_i satisfying our constraints, this may be upper bounded as follows

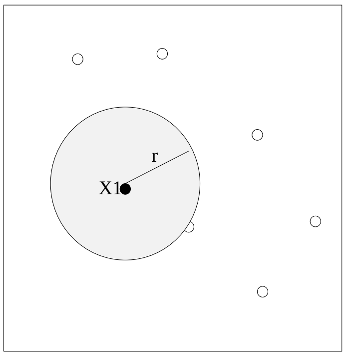

Figure 1: The distribution of $d(x_1)$

$$\begin{aligned} Prob\left[d(x_1) \geq r\right] = & \; Prob\left[x_2, \ldots, x_n \text{ do not lie in the shaded region}\right] \\ & = \left(1 - \tfrac{\pi r^2}{1}\right)^{n-1} \end{aligned}$$

Note that an $r > \frac{1}{\pi}$ will make the "probability" negative. Hence,

$$\begin{aligned} E\left[d(x_1)\right] = \int_0^{\infty} Prob\left[d(x_1) \geq r\right] dr \geq & \int_0^{\frac{1}{\pi}} Prob\left[d(x_1) \geq r\right] dr \\ = & \int_0^{\frac{1}{\pi}} \left(1 - \pi r^2\right)^{n-1} dr \end{aligned}$$

Lowering the upper limit on the integral still preserves the inequality. We choose to lower it to $\frac{1}{\sqrt{n\pi}}$ in order to simplify the evaluation.

$$E\left[d(x_i)\right] \geq \int_0^{\frac{1}{\sqrt{n\pi}}} \left(1 - \pi r^2\right)^{n-1} dr$$

From Figure. 2 the above term represents the area under the curve $\left(1 - \pi r^2\right)^{n-1}$ (A) bounded by $r = 0$ and $r = \frac{1}{\sqrt{n\pi}}$. But this is atleast the area under the st. line A' given by $(1 - \frac{1}{n})$. This area is given by

$$\left(1 - \frac{1}{n}\right)^{n-1} \cdot \frac{1}{\sqrt{n\pi}}$$

For large n, we then have the result

$$E\left[d(x_1)\right] \geq \frac{1}{e} \cdot \frac{1}{\sqrt{n\pi}}$$

Substituing this in Equation. 1, we have

$$\begin{aligned} E\left[C^*\right] \geq & \; nE\left[d(x_1)\right] \\ \geq & \; \tfrac{n}{e\sqrt{n\pi}} \\ \geq & \; \tfrac{\sqrt{n}}{O(1)} \end{aligned}$$

■

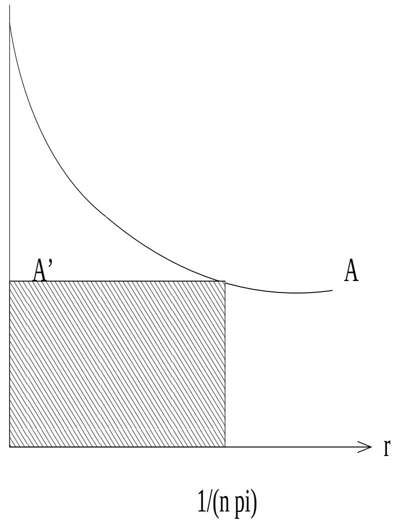

Figure 2: Approximating the area under $(1-\pi r^2)^{n-1}$ (A)

Probabilitiy	$\frac{1}{100}$	$\frac{99}{100}$
C_{KARP}	1000	100
C^*	1	100
$\frac{C_{KARP}}{C^*}$	1000	1

Table 1: Example distribution of TSP solutions.

From the above theorem, we obtain the result

$$\begin{aligned}\frac{E\left[C_{KARP}\right]}{E\left[C^*\right]} &\leq 1+O\left(\sqrt{\frac{n}{s}}\right)/E\left[C^*\right] \\ &\leq 1+O\left(\sqrt{\frac{n}{s}}\right)/\frac{\sqrt{n}}{O\left(1\right)} \\ &\leq 1+O\left(\frac{1}{\sqrt{s}}\right).\end{aligned}$$

We have an $(1+O\left(\frac{1}{\sqrt{s}}\right))$ expected approximation ratio for the Euclidean TSP. What does this tell us? Note that a good $\frac{E[C_{KARP}]}{E[C^*]}$ does not guarantee a good $E\left[\frac{C_{KARP}}{C^*}\right]$, which we would like to bound. An example will illustrate this. Consider a class of problems with the distribution shown in Table 1.

The excpected values may be computed to be $E\left[C_{KARP}\right] = 109$ and $E\left[C^*\right] \approx 99$. Thus a low value of $\frac{E[C_{KARP}]}{E[C^*]}(=\frac{109}{99})$ does not guarantee a good $E\left[\frac{C_{KARP}}{C^*}\right]$ $(= 10.99)$. We can better estimate the performanece of C_{KARP} by showing the result of Theorem. 2. Before that, A preliminary lemma

Lemma 2

$$Prob\left[C^* > \frac{\sqrt{n}}{1000}\right] \geq \frac{1}{10} - \frac{1}{2000} \geq frac111$$

where the $\frac{1}{11}$ can be driven can be driven up as required, by suitably choosing a larger constant instead of the 1000 .

Proof: Let p_1 and p_2 be defined as

$$p_1 = Prob\left[C^* < \frac{\sqrt{n}}{1000}\right];\ p_2 = Prob\left[C^* \geq \frac{\sqrt{n}}{1000}\right].$$

Since from NaiveTSP, $2\sqrt{n}$ is an upperbound on C^*,

$$\begin{aligned} E\left[C^*\right] \leq & \; p_1 \tfrac{\sqrt{n}}{1000} + p_2 2\sqrt{n} \\ \leq & \; \tfrac{\sqrt{n}}{5}\left(\tfrac{p_1}{200} + 10p_2\right). \end{aligned}$$

Since $p_1 + p_2 = 1$, p_1 is atmost 1 and $E\left[C^*\right] \geq \frac{\sqrt{n}}{O(1)}$, we can approximate the above equation as

$$\begin{aligned} & E\left[C^*\right] && \leq \frac{\sqrt{n}}{5}\left(\frac{1}{200} + 10p_2\right) \\ \Rightarrow & \quad 1 && \leq \frac{1}{200} + 10p_2 \\ \Rightarrow & \quad p_2 && \geq \frac{1}{10} - \frac{1}{2000} \geq \frac{1}{11} \end{aligned}$$

and we are done. ■

Theorem 2 *For the conditions outlined in Theorem. 1*

$$Prob\left[\frac{C_{KARP}}{C^*} \leq 1 + O\left(\frac{1}{\sqrt{s}}\right)\right] \geq \frac{1}{11}.$$

Proof:

$$
\begin{aligned}
&Prob\left[\frac{C_{KARP}}{C^*} \le 1 + O\left(\frac{1}{\sqrt{s}}\right)\right] \\
&\qquad = Prob\left[C^*\left(1 + O\left(\tfrac{1}{\sqrt{s}}\right)\right) \ge C_{KARP}\right] \\
&\qquad = Prob\left[C^*\left(1 + O\left(\tfrac{1}{\sqrt{s}}\right)\right) \ge \tfrac{\sqrt{n}}{\sqrt{s}}\right] \\
&\qquad = Prob\left[C^* \ge \frac{\sqrt{n}}{\sqrt{s}\left(1+O\left(\frac{1}{\sqrt{s}}\right)\right)}\right]
\end{aligned}
$$

where, in the second step, we have C_{KARP} is alteast $\frac{\sqrt{n}}{\sqrt{s}}$. For a suitably large n (and thereby s), this has the form of p_2 of previous lemma and the result follows. ■

Stronger results than the above discussion are known.

Result 1 *There exists β such that*

$$\lim_{n\to\infty} \frac{E\left[C^*\right]}{\sqrt{n}} = \beta.$$

and further

Result 2 *There exists β such that*

$$Prob\left[\lim_{n\to\infty} \frac{E\left[C^*\right]}{\sqrt{n}} = \beta\right] = 1.$$

These in turn imply

$$Prob\left[\lim_{n\to\infty} \frac{C_{KARP}}{C^*} = 1\right] = 1.$$

Discussion

We have seen an instance of an average case analysis of an algorithm, one which uses a particular probability distribution over the problem instances. Typically, algorithms are developed from hueristics, tested and found to work well in practise, and an average case analysis is performed to demonstrate the "goodness" of the algorithm. Algorithms such as Spectral Graph Partitioning, Simulated Annealling and the Simplex Method for Linear Programming are examples where effectiveness demonstrated in practise led to work on average case analysis.

In constrast, to design an algorithm by assuming distributions on the input, and analytically developing the algorithm to perform well over such inputs is usually not a good idea. Often, an analytically defined random distribution on input possess properties "convenient" for an algorithm to exploit, that real-life random inputs don't. An illustrative example would be graph connectivity. A result by Erdos and Renyi shows that the class $G(n,p)$ of random graphs over n nodes, where each edge is included independently with probability p has a useful property that, as $n \to \infty$:

Result 3 *If* $p \geq \frac{\log n}{n} + \frac{c}{n}$, *then* G *is connected with high probability -*

$$\lim_{n \to \infty} Prob\,[G(n,p) \text{ is connected}] = e^{-e^{-c}}.$$

A randomised algorithm based on such a result would not work well in practice unless we know with certainity that the graphs are chosen uniformly at random from $G(n,p)$.