Commun. Korean Math. Soc. **17** (2002), No. 2, pp. 321–330

ON THE FLUCTUATION IN THE RANDOM ASSIGNMENT PROBLEM

SUNGCHUL LEE* AND ZHONGGEN SU†

ABSTRACT. Consider the random assignment (or bipartite matching) problem with iid uniform edge costs $t(i,j)$. Let A_n be the optimal assignment cost. Just recently does Aldous [2] give a rigorous proof that $EA_n \to \zeta(2)$. In this paper we establish the upper and lower bounds for $\mathrm{Var}A_n$, i.e., there exist two strictly positive but finite constants C_1 and C_2 such that $C_1 n^{-5/2}(\log n)^{-3/2} \leq \mathrm{Var}A_n \leq C_2 n^{-1}(\log n)^2$.

1. Introduction

Suppose there are n workers available to fill n jobs, where each worker is to be assigned to exactly one job. Also suppose that we have a measurement $t(i,j)$ of how qualified the individual i is for each job j. An assignment of workers to jobs is then simply a permutation π on $\{1,2,\ldots,n\}$. For this job assignment π the cost is measured by $\sum_{i=1}^n t(i,\pi(i))$. The assignment problem is to find a job assignment π that costs less than any other job assignment. For this optimal job assignment π the cost A_n is given by

$$A_n = \min\Big\{\sum_{i=1}^n t(i,\pi'(i)) : \pi' \text{ a permutation on } \{1,2,\ldots,n\}\Big\}.$$

Received April 30, 2001.

2000 Mathematics Subject Classification: Primary 05C70, 60C05, 68W40; Secondary 90C27.

Key words and phrases: assignment problem, bipartite matching, conditional variance, probabilistic analysis of algorithms.

* This work was supported by the BK21 project of the Department of Mathematics, Yonsei University, the interdisciplinary research program of KOSEF 1999-2-103-001-5, and Com^{2}MaC in POSTECH.

† This work was supported by the BK21 project of the Department of Mathematics, Yonsei University, and by NSFC 10071072.

We can also think of the workers and jobs as vertices of a bipartite graph, with the $t(i,j)$ as the weights on the edges. Then, the assignment problem is to find a matching with the minimal total weight in this bipartite graph.

For the random assignment problem, we let $t(i,j)$ be iid uniform random variables on the unit interval $[0,1]$. It is traditional to use this uniform distribution. As Aldous [1] and Coppersmith and Sorkin [3] discussed, the exponential distribution with parameter 1 may be a better choice for several reasons. However, as Aldous [1] pointed out, since the density at 0 ultimately matters, these choices are asymptotically equivalent. So, in this paper we follow the traditional setting; the uniform distribution. Under this uniform setting the asymptotic behavior of EA_n has received a lot of attention. Using his objective method Aldous [1] showed that the limit of EA_n exists and it is the cost of an optimal bipartite matching on certain weighted infinite tree. Just recently does Aldous [2] identify the limit as $\zeta(2)$ by constructing the optimal bipartite matching on the infinite tree. See [12], [6], [3] for various results regarding the upper bounds of EA_n and [8], [5], [9] for the lower bounds.

It is natural to expect that $\mathrm{Var} A_n \approx \sigma^2/n$ for some $0 < \sigma < \infty$, and that the rescaled A_n has a normal limit; this is supported by the Monte-Carlo simulation. However there is no mathematical proof and this problem is largely open. Toward this problem, in this paper we establish the upper and lower bounds for the variance of A_n. Our main result is as follows.

THEOREM 1. *There exist strictly positive but finite constants C_1 and C_2 such that*

$$C_1 n^{-5/2}(\log n)^{-3/2} \leq \mathrm{Var} A_n \leq C_2 n^{-1}(\log n)^2.$$

In Frieze and Sorkin [4], they studied the relationship between the assignment problems and the asymmetric traveling salesman problems. One can use their result and the argument in this paper to obtain the upper and lower bounds for the variance of the asymmetric traveling salesman problems. We leave this to the interested reader.

The proof of the Theorem will be given in Section 2. The upper bound is a slight improvement to the known upper bound due to Talagrand [11] (see also Section 6.7 of Steele [10]). More specifically, Karp and Steele [7] showed that with high probability the greatest cost of an edge used in the optimal assignment is of order $(\log n)^2/n$. With this Talagrand [11] obtained a good concentration inequality for A_n and showed that $\mathrm{Var} A_n \leq C_3 n^{-1}(\log n)^4$. The present upper bound is an immediate consequence of the recent result of Frieze and Sorkin [4]; they showed that with high probability the greatest cost of an edge used in the optimal assignment is of order $\log n/n$. We sketch in Section 2 how to get the upper bound. Our main contribution is actually the lower bound. There was no lower bound available up to this point. In this paper we use the conditioning argument to establish the lower bound in Theorem 1.

In this paper, there are lots of strictly positive but finite constants whose specific values are not of interest. We denote them by C_i.

2. Proof of Theorem 1

Let's begin with the proof of the upper bound which is an immediate consequence of the recent result of Frieze and Sorkin [4] that with high probability the greatest cost of an edge used in the optimal assignment is of order $\log n/n$. With this we obtain a good concentration inequality for A_n and show that $\mathrm{Var} A_n \leq C_2 n^{-1}(\log n)^2$. We sketch this part.

The main idea of Frieze and Sorkin [4] is that if there are two sequences $i_1, i_2, \ldots, i_m$ and $i'_1, i'_2, \ldots, i'_m$ in $\{1, 2, \ldots, n\}$ such that (i_j, i'_{j-1}), $2 \leq j \leq m$, and (i_1, i'_m) are in an optimal assignment, then

$$\sum_{j=2}^{m} t(i_j, i'_{j-1}) + t(i_1, i'_m) \leq \sum_{j=1}^{m} t(i_j, i'_j). \tag{2.1}$$

Otherwise, it is better to use (i_j, i'_j), $1 \leq j \leq m$, than (i_j, i'_{j-1}), $2 \leq j \leq m$, and (i_1, i'_m) in an optimal assignment. They showed that with high probability for an optimal assignment π and for any i in $\{1, 2, \ldots, n\}$ there are two sequences $i_1, i_2, \ldots, i_m$ and $i'_1, i'_2, \ldots, i'_m$ in $\{1, 2, \ldots, n\}$ such that (i_j, i'_{j-1}), $2 \leq j \leq m$, and $(i_1, i_{m'})$ are in an optimal assignment π and that $i_1 = i$ and $i_{2m} = \pi(i)$. By (2.1), then we have

$$t(i, \pi(i)) \leq \sum_{j=1}^{m} t(i_j, i'_j) - \sum_{j=2}^{m} t(i_j, i'_{j-1}). \tag{2.2}$$

By their ingenious choice of two sequences the RHS of (2.2) is quite small with high probability.

Here is a sketch of their choice. Consider the bipartite graph $K_{n,n}$ on vertex sets $I = J = \{1, 2, \ldots, n\}$, in which each edge is assigned a cost

$t(i,j)$. Define for $I' \subset I$, the top-down short edges $TD(I')$ from I' by

$$TD(I') = \{(i,j) \ : \exists i \in I' \text{ such that } (i,j) \text{ is one of the 40 shortest arcs out of } i\}$$

and for $J' \subset J$, the down-top short edges $DT(J')$ from J' by

$$DT(J') = \{(i,j) \ : \exists j \in J' \text{ such that } (i,j) \text{ is one of the 40 shortest arcs into } j\}.$$

With these top-down short edges and down-top short edges we now define for $I' \subset I$, the neighborhood of I' by

$$N(I') = \{j : (i,j) \in TD(I')\}$$

and for $J' \subset J$, the neighborhood of J' by

$$N(J') = \{i : (i,j) \in DT(J')\}.$$

Then, they showed that for small I' and J' their neighborhoods are relatively large; with high probability for all $I' \subset I$ with $|I'| \leq n/5$ and $J' \subset J$ with $|T| \leq n/5$, we have $|N(I')| \geq 4|I'|$ and $|N(J')| \geq 4|J'|$. Under this good situation we use the following pigeon hole principle. We let $I_0 = \{i\}$ and we construct I_k by $I_k = \pi^{-1}(N(I_{k-1}))$ until $|I_{k_0-1}| > n/5$. We discard vertices from I_{k_0-1} to create a smaller set I'_{k_0-1} with $|I'_{k_0-1}| = [n/5]$ and we let $I'_{k_0} = \pi^{-1}(N(I'_{k_0-1}))$. By its construction we have $|I'_{k_0}| > n/2$. We do the same operation for $\pi(i)$. Let $J_0 = \{\pi(i)\}$ and we construct J_k by $J_k = \pi(N(J_{k-1}))$ until $|J_{k_1-1}| > n/5$. We discard vertices from J_{k_1-1} to create a smaller set J'_{k_1-1} with $|J'_{k_1-1}| = [n/5]$ and we let $J'_{k_1} = \pi(N(J'_{k_1-1}))$. By its construction we have $|J'_{k_1}| > n/2$. Since both $|I'_{k_0}|$ and $|J'_{k_1}|$ are larger than $n/2$, there must be some $i' \in I'_{k_0}$ with $\pi(i') \in J'_{k_1}$. Hence there are two sequences considered in (2.1). Moreover, the first few edges (i_j, i'_j) in the RHS of (2.1) is the top-down short edges and the rest is the down-top short edges. In addition, m is of order $\log n$.

In Lemma 7 of Frieze and Sorkin [4], they showed that with high probability

$$Z_1 = \max\left\{\sum_{l=0}^{k} t(i_l, j_l) - \sum_{l=0}^{k-1} t(i_{l+1}, j_l)\right\}$$

is of order $\log n/n$ where the maximum is over any sequences i_0, j_0, i_1, $j_1, \ldots, i_k, j_k$ with top-down short edges $(i_l, j_l) \in TD(\{i_l\})$ leaving from

i_l and $k \leq [3\log_4 n]$, and similarly

$$Z_2 = \max\left\{ \sum_{l=0}^{k} t(i_l, j_l) - \sum_{l=0}^{k-1} t(i_{l+1}, j_l) \right\}$$

is of order $\log n/n$ where the maximum is over any sequences i_0, j_0, i_1, $j_1, \ldots, i_k, j_k$ with down-top short edges $(i_l, j_l) \in DT(\{j_l\})$ leaving from j_l and $k \leq [3\log_4 n]$. With these Z_1 and Z_2, we can easily see that the RHS of (2.2) is bounded by $Z_1 + Z_2$. Hence, by (2.2) with high probability $t(i, \pi(i))$ is of order $\log n/n$.

We can quantify the meaning of "with high probability" and "of order $\log n/n$", and the following is the precise statement we need. We skip its proof.

PROPOSITION 1. *Let T_{max} be the maximum cost of an edge used in an optimal assignment. Then,*

$$P(T_{max} \geq C_4 \frac{\log n}{n}) \leq C_5 n^{-3}.$$

As a direct consequence of Proposition 1, we can obtain the following concentration inequality for A_n around its median. Since one can prove it in a similar way to Talagrand [11], we skip its proof.

PROPOSITION 2. *Let $m(A_n)$ be the median of A_n. Then*

$$P(|A_n - m(A_n)| \geq t) \leq C_6 \exp\Big(- C_7 \frac{nt^2}{(\log n)^2}\Big) + C_8 n^{-3}.$$

PROOF OF THE UPPER BOUND. By Proposition 2, we have

$$\begin{aligned} E(A_n - m(A_n))^2 &= 2\int_0^n sP(|A_n - m(A_n)| \geq s)ds \\ &\leq 2\int_0^n s\Big(C_6 \exp(-C_7 \frac{ns^2}{(\log n)^2}) + C_8 n^{-3}\Big)ds \\ &\leq C_9 \frac{(\log n)^2}{n}. \end{aligned}$$

Since $\mathrm{Var} A_n \leq E(A_n - m(A_n))^2$, the upper bound follows. □

To establish the lower bound for $\mathrm{Var} A_n$, we use the conditioning argument. We first recall a basic fact on the conditional variance.

LEMMA 1. *Let X be in $L^2(\Omega, \mathcal{F}, P)$. Then, for any sub-field $\mathcal{F}'$*

$$\mathrm{Var}(X) \geq E\,\mathrm{Var}(X|\mathcal{F}')$$

where $\mathrm{Var}(X|\mathcal{F}') = E(X^2|\mathcal{F}') - (E(X|\mathcal{F}'))^2$. *In particular, if E_1, E_2, $\ldots$, E_m are mutually disjoint, then*

$$\mathrm{Var}(X) \geq \sum_{k=1}^{m} \mathrm{Var}(X|E_k)P(E_k).$$

PROOF OF THE LOWER BOUND. For $1 \leq i \leq n$ we let $\pi^{(1,i)}$ be a one to one mapping from $\{2, \ldots, n\}$ to $\{1, 2, \ldots, n\}\backslash\{i\}$, and we denote by $A_{n-1}^{(1,i)}$ the optimal cost of workers $\{2, \ldots, n\}$ assigned to jobs $\{1, 2, \ldots, n\}\backslash\{i\}$, i.e.,

$$A_{n-1}^{(1,i)} = \min_{\pi^{(1,i)}} \sum_{j=2}^{n} t(j, \pi^{(1,i)}(j)).$$

Then, we can rewrite A_n as

$$A_n = \min_{1 \leq i \leq n} (t(1,i) + A_{n-1}^{(1,i)}).$$

It is simple but important to note that $t(1,i)$, $1 \leq i \leq n$, are iid, $\{t(1,i), 1 \leq i \leq n\}$ is independent of $\{A_{n-1}^{(1,i)}, 1 \leq i \leq n\}$, and $A_{n-1}^{(1,i)}$ has the same distribution as A_{n-1}.

Given $A_{n-1}^{(1,1)}, A_{n-1}^{(1,2)}, \ldots, A_{n-1}^{(1,n)}$, we rearrange them in the increasing order. In other words, we find a permutation σ on $\{1, 2, \ldots, n\}$ such that $A_{n-1}^{(1,\sigma(1))} \leq A_{n-1}^{(1,\sigma(2))} \leq \cdots \leq A_{n-1}^{(1,\sigma(n))}$. Now, we let $T_i = A_{n-1}^{(1,\sigma(i))}$. Then,

$$A_n = \min_{1 \leq i \leq n} (t(1,i) + A_{n-1}^{(1,i)}) = \min_{1 \leq i \leq n} (X_i + T_i)$$

where $X_i = t(1, \sigma(i))$ are iid uniform on $[0,1]$ and independent of T_j. Therefore, from now on instead of looking at $\mathrm{Var} A_n$ we look at $\mathrm{Var} \min_{1 < i < n} (X_i + T_i)$. For $1 \leq m \leq n$ we define m mutually disjoint events as follows.

$$\begin{aligned}
&E_1 = \Big\{X_1 \le \frac{1}{n}, X_2 \ge \frac{1}{n}, \dots, X_n \ge \frac{1}{n}\Big\},\\
&E_2 = \Big\{X_2 \le \frac{1}{n}, X_1 \ge \frac{1}{n} + T_2 - T_1, X_3 \ge \frac{1}{n}, \dots, X_n \ge \frac{1}{n}\Big\},\\
&\dots\dots\\
&E_m = \Big\{X_m \le \frac{1}{n}, X_1 \ge \frac{1}{n} + T_m - T_1, \dots,\\
&\qquad X_{m-1} \ge \frac{1}{n} + T_m - T_{m-1}, X_{m+1} \ge \frac{1}{n}, \dots, X_n \ge \frac{1}{n}\Big\}.
\end{aligned}$$

Then, $E_1, E_2, \dots, E_m$ are mutually disjoint, and

$$\begin{aligned}
&P(E_1) = \frac{1}{n}(1-\frac{1}{n})^{n-1},\\
&P(E_2) = \frac{1}{n}(1-\frac{1}{n})^{n-2}E(1-\frac{1}{n}-(T_2-T_1)),\\
&\dots\dots\\
&P(E_m) = \frac{1}{n}(1-\frac{1}{n})^{n-m}E\prod_{l=1}^{m-1}(1-\frac{1}{n}-(T_m-T_l)).
\end{aligned} \tag{2.3}$$

Most importantly, on E_k we have

$$A_n = \min_{1\le i\le n}(X_i+T_i) = X_k + T_k. \tag{2.4}$$

Thus, by Lemma 1 and (2.4)

$$\begin{aligned}
\mathrm{Var}A_n &= \mathrm{Var}(\min_{1\le i\le n}(X_i+T_i))\\
&\ge \sum_{k=1}^{m}\mathrm{Var}(\min_{1\le i\le n}(X_i+T_i)|E_k)P(E_k)\\
&= \sum_{k=1}^{m}\mathrm{Var}(X_k+T_k|E_k)P(E_k).
\end{aligned}$$

Thus, since X_i and T_j are independent, by (2.3)

$$\begin{aligned}\mathrm{Var}A_n &\geq \sum_{k=1}^{m}\mathrm{Var}(X_k+T_k|E_k)P(E_k)\\ &\geq \sum_{k=1}^{m}\mathrm{Var}(X_k|E_k)P(E_k)\\ &= \sum_{k=1}^{m}\mathrm{Var}(X_k|X_k\leq\frac{1}{n})P(E_k)\\ &= \frac{1}{12n^2}\sum_{k=1}^{m}P(E_k)\\ &= \frac{1}{12n^2}\frac{1}{n}\sum_{k=1}^{m}(1-\frac{1}{n})^{n-k}E\prod_{l=1}^{k-1}(1-\frac{1}{n}-(T_k-T_l)).\end{aligned} \tag{2.5}$$

Next, we choose m and estimate the RHS of (2.5). Define

$$G_n=\cap_{i=1}^{n}\{|A_{n-1}^{(1,i)}-m(A_{n-1}^{(1,i)})|\leq s_{n-1}\}$$

where $m(A_{n-1}^{(1,i)})$ is the median of $A_{n-1}^{(1,i)}$ and where s_n is specified below On G_n, since $m(A_{n-1}^{(1,i)})=m(A_{n-1}^{(1,j)})$, we have $|A_{n-1}^{(1,i)}-A_{n-1}^{(1,j)}|\leq 2s_{n-1}$ Since T_i are just the renumbering of $A_{n-1}^{(1,j)}$, on G_n

$$|T_i-T_j|\leq 2s_{n-1}. \tag{2.6}$$

Therefore, by (2.5) and (2.6)

$$\begin{aligned}\mathrm{Var}A_n &\geq \frac{1}{12n^2}\frac{1}{n}\sum_{k=1}^{m}(1-\frac{1}{n})^{n-k}E\prod_{l=1}^{k-1}(1-\frac{1}{n}-(T_k-T_l))\\ &\geq \frac{1}{12n^2}\frac{1}{n}\sum_{k=1}^{m}(1-\frac{1}{n})^{n-k}E\prod_{l=1}^{k-1}(1-\frac{1}{n}-(T_k-T_l))1(G_n)\\ &\geq \frac{1}{12n^2}\frac{1}{n}\sum_{k=1}^{m}(1-\frac{1}{n})^{n-k}(1-\frac{1}{n}-2s_{n-1})^{k-1}P(G_n)\\ &\geq \frac{m}{12n^3}(1-\frac{1}{n})^{n}(1-\frac{1}{n}-2s_{n-1})^{m}P(G_n).\end{aligned} \tag{2.7}$$

Since $A_{n-1}^{(1,i)}$ has the same distribution as A_{n-1}, by Proposition 2

$$\begin{aligned}
&P(|A_{n-1}^{(1,i)} - m(A_{n-1}^{(1,i)})| \geq s_{n-1}) \\
&= P(|A_{n-1} - m(A_{n-1})| > s_{n-1}) \\
&\leq C_6 \exp\Big(- C_7 \frac{(n-1)s_{n-1}^2}{(\log(n-1))^2}\Big) + C_8(n-1)^{-3}.
\end{aligned}$$

Now, we take $s_n = C_{10}n^{-1/2}(\log n)^{3/2}$. Then, we have $P(G_n) \to 1$. After the choice of s_n, we take $m = [1/2s_{n-1}]$. With this choice of m, $(1-1/n)^n \to e^{-1}$, $(1-1/n-2s_{n-1})^m \to e^{-1}$, and hence the lower bound follows from (2.7). □

References

[1] D. J. Aldous, *Asymptotics in the random assignment problem*, Probab. Theory Relat. Fields **93** (1992), 507–534.

[2] ———, *The $\zeta(2)$ limit in the random assignment problem*, Random Struct. Alg., to appear, 2001.

[3] D. Coppersmith and G. B. Sorkin, *Constructive bounds and exact expectations for the random assignment problem*, Random Struct. Alg. **15** (1999), 113–144.

[4] A. M. Frieze and G. B. Sorkin, *The probabilistic relationship between the assignment and asymmetric traveling salesman problems*, Proceedings of SODA, (2001), 652–660.

[5] M. X. Goemans and M. S. Kodiallam, *A lower bound on the expected cost of an optimal assigenemnt*, Math. Oper. Res. **18** (1993), 267–274.

[6] R. M. Karp, *An upper bound on the expected cost of an optimal assignment*, Discrete Algorithms and Complexity: Proceedings of the Japan-U.S. joint seminar, 1-4. Academic Press, 1987.

[7] R. M. Karp and J. M. Steele, *Probabilistic analysis of heuristics*, The Traveling Salesman Problem, 181-205. John Wiley and Sons, 1985.

[8] A. J. Lazarus, *The assignment problem with uniform* $(0,1)$ *cost matrix*, B.A. thesis, Department of Mathematics, Princeton University, Princeton, NJ, 1979.

[9] B. Olin, *Asymptotic properties of random assignment problems*, Ph.D. thesis, Kungl Tekniska Högskolan, Stockholm, Sweden, 1992.

[10] J. M. Steele, *Probability Theory and Combinatorial Optimization*, SIAM, 1997.

[11] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Publ. Math. IHES. **81** (1995), 73–205.

[12] D. W. Walkup, *On the expected value of a random assignment problem*, SIAM J. Comput. **8** (1979), 440–442.

Sungchul Lee
Department of Mathematics
Yonsei University
Seoul 120-749, Korea
E-mail: sungchul@yonsei.ac.kr

Zhonggen Su
Department of Mathematics
Zhejiang University
Hangzhou 310028, P. R. China
E-mail: zgsu@mail.hz.zj.cn

and

Department of Mathematics
Yonsei University
Seoul 120-749, Korea
E-mail: zgsu2001@yahoo.com