A RANDOM ASSIGNMENT PROBLEM: SIZE OF NEAR MAXIMAL SETS AND CORRECT ORDER EXPECTATION BOUNDS

GILLES MORDANT

Abstract. We study the correlation structure arising in the random assignment problem under a Gaussian cost assumption. This study helps derive an upper bound on the expected log cardinality of near maximal sets. The latter bound helps improve the understanding of the peaks in a random array on a permutation tree and, in some sense, yields another quantitative indication similar to Aldous’ Asymptotical Essential Uniqueness property (Random Structures & Algorithms, 2001) for the random assignment problem. The study of the correlation structure also helps derive the correct order for the expectation of the minimal cost arising in the random assignment problem considered.

1. Introduction

Optimal assignment is a classical problem appearing in mathematics and computer sciences that also attracted probabilists’ attention. In the deterministic problem, given a cost matrix where each element $c(i,j)$ corresponds to the cost for person $i$ of performing task $j$, the objective is to find the overall cheapest cost that forms a bijection between tasks and people.

In the probabilistic version of the assignment problem, one considers a random cost matrix where all entries are independent and identically distributed. The goal is then to understand both the optimal cost and the optimal bijection.

One of the most important results for this problem was obtained nearly two decades ago by Aldous (2001). He showed that the expected cost for a cost matrix with Exp(1) entries converges to $\zeta(2)$. This result confirmed a conjecture by Mézard and Parisi (1987) that stood for years. In his paper, Aldous also proves that an Asymptotical Essential Uniqueness (AEU) property holds (see Theorem 2.1 below). This property can be understood as the fact that any near optimal configuration has a near optimal cost.

For space considerations, we refer to Aldous and Steele (2004) or Chatterjee (2019) for the history of the problem and more comprehensive literature reviews.

Date: May 1, 2020.

The author thanks S. Chatterjee for some insightful remarks. The comments by S. Kirilova, V. Plassier and J. Segers are also gratefully acknowledged.
2. Formalisation of the problem and main results

2.1. The problem. Formally, the stochastic matching problem involves a \( n \times n \) matrix with i.i.d. random entries \( c(i, j) \). The first object of study is the following random variable

\[
W_n := \min_{\sigma \in S_n} \sum_{i=1}^{n} c(i, \sigma(i)),
\]

where \( S_n \) is the set of all permutations of \( \{1, \ldots, n\} \). Also,

\[
\sigma^* := \arg\min_{\sigma \in S_n} \sum_{i=1}^{n} c(i, \sigma(i))
\]

and how \( \sigma^* \) compares to any other permutation in terms of overall cost are worth studying.

To the author’s knowledge, it is the first time this problem is studied under a Gaussian assumption (Assumption 1). Usually this problem is studied for non-negative random variables; still, we believe that studying the assignment problem in a different setting could lead to a better understanding of the problem as a whole. Also, owing to symmetry, studying the minimum is essentially the same as studying the maximum; only some quantities of interest will see their sign changed.

**Assumption 1.** Assume \( c \) to be a cost matrix with i.i.d. entries such that each \( c(i, j) \) is a centred Gaussian random variable with standard deviation \( 1/\sqrt{n} \).

Before stating the main results, observe that each permutation will select independent entries from the cost matrix and that the optimal assignment problem can be recast as an extreme value theory problem. Implicitly, there is a vector appearing in the right hand side of equation (1); we will denote it by \( g = (g_1, \ldots, g_n) \).

Indeed, denoting by \( \sigma_1, \sigma_2, \ldots \) different permutations, we can write

\[
g_u = \sum_{i=1}^{n} c(i, \sigma_u(i)), \quad 1 \leq u \leq n!.
\]

\( g \) is a vector of linear combinations of independent Gaussian random variables, it is thus multivariate Gaussian. Let us denote by \( r_{u,v} \) the elements of its correlation matrix.

2.2. Main results. Our first result concerns the study of the problem under the prism of equation (2), where the objective is to understand how a generic permutation differs from \( \sigma^* \). To the author’s knowledge, the only result on near-optimal permutations is given by the following theorem.

**Theorem 2.1** (Aldous, 2001). Consider \( c(i, j) \) to be i.i.d. \( \text{Exp}(1) \). For each \( 0 < \delta < 1 \), there exists \( \epsilon(\delta) > 0 \) such that if \( \sigma_n \) is the optimal permutation and \( \mu_n \) are
random permutations (depending on $c(i, j)$) such that $\mathbb{E}(n^{-1}|\{i : \sigma_n \neq \mu_n\}|) \geq \delta$, then

$$\liminf_{n \to \infty} \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} c(i, \mu_n(i))\right) \geq \frac{\pi^2}{6} + \epsilon(\delta).$$

Even though an approach to finding $\epsilon(\delta)$ is proposed in the same paper, the way of rigorously deriving such a function of $\delta$ is an open problem.

Pursuing the objective of understanding near-optimal permutations, one may also study the number of entries of $g$ that are above a certain threshold. This type of question is classical in statistics and probability. First, note that $g$ is a centered Gaussian vector with $\mathbb{E}(g_u^2) = 1$ for all $u$. With this notation, we can now define the object of interest.

**Definition 1.** For all $\epsilon > 0$, a near maximal set is given by

$$A(\epsilon) := \{u : g_u > (1 - \epsilon)m\},$$

where $m := \mathbb{E}(\max_u g_u)$.

The aim is thus to get the order of magnitude of the cardinality of this set. For this reason, S. Chatterjee proposed in his book *Superconcentration and related topics* (2014) to study

$$\mathbb{E}(\log|A(\epsilon)|; A(\epsilon) \neq \emptyset).$$

This is the object of the theorem hereafter, in the specific case of the random assignment problem. For convenience, we stick to the original definition of near maximal set and state the result in terms of maximum. The result for the minimisation problem easily follows from symmetry.

**Theorem 2.2.** For the vector $g$ defined above, it holds

$$\mathbb{E}(\log|A(\epsilon)|; A(\epsilon) \neq \emptyset) \leq \begin{cases} C'(n \log n)^{3/4} & \text{if } \epsilon < (2 \log(n!))^{-1/2} \\ C'' \sqrt{\epsilon(n \log n)} & \text{if } \epsilon > (2 \log(n!))^{-1/2} \end{cases}$$

for $C'$ and $C''$ two universal constants.

Also another natural question is to get sharp bounds on the expectation of the quantity of interest, $W_n$. A first result in this direction is given by the following theorem.

**Theorem 2.3.** Under Assumption 1, there exist $n_0$ such that $\forall n \geq n_0$,

$$C \sqrt{2 \log(n!)} \leq \mathbb{E}(-W_n) \leq \sqrt{2 \log(n!)},$$

for $C > 0$ a constant.

Surprisingly, we obtain that, even though some correlations are very close to one for the gaussian vector we study, the expectation has the same order as the
one that would arise setting all correlations to zero. Simulations suggest that $E(-W_n) \approx 0.9\sqrt{2\log(n!)}$.

3. Calculations and proofs

We now turn to the proofs of the results stated.

3.1. Dependence structure of $g$. Given two entries $g_1$ and $g_2$, one sees from the sum in equation (3) that the dependence will occur only through indices $(i, j)$ such that $\sigma_1(i) = \sigma_2(i) = j$, that is, through fixed points of the permutation from $\sigma_1$ to $\sigma_2$. As $n$ elements appear in each sum, denoting by $k(1, 2)$ the number of fixed points from $\sigma_1$ to $\sigma_2$, the correlation between two sums, say 1 and 2, is $r_{1,2} = \frac{k(1,2)}{n}$.

The following natural question is then: What can we say about those fixed points? Montmort (1708) was apparently the first to raise this question and proved that, setting $A(\sigma) = \{i : \sigma(i) = i\}, \sigma \in S_n$, it holds that $\frac{|\{\sigma : |A(\sigma)| = k\}|}{n!} \to \frac{1}{e \times k!}$, for $k$ being fixed as $n \to \infty$. More precisely, a simple inclusion-exclusion argument shows that $\frac{|\{\sigma : |A(\sigma)| = k\}|}{n!} = \frac{1}{k!} \sum_{l=0}^{n-k} \frac{(-1)^l}{l!}$.

**Remark:** A closer look indicates that, for $n$ sufficiently large, more than 99% of the entries of $g$ have a correlation lower than $4/n$ with $g_k$ for any $k$.

3.2. Near maximal sets. To prove the main result, we base ourselves on Theorem 3.1. Before stating the latter, let us introduce a few elements.

For all $1 \leq u \leq n!$ and $\delta > 0$, introduce $B(u, \delta) := \{v : E(g_ug_v) > 1 - \delta\}$ and further define $V(\delta) = \max_u |B(u, \delta)|$.

**Theorem 3.1** (Chatterjee). With the above notation and assumptions, we have $E(\log|A(\epsilon)|; A(\epsilon) \neq \emptyset) \leq \inf_{\delta \in (0,1)} \left( \log V(\delta) + \frac{C \max \{\epsilon m^2, m\}}{\delta} \right)$, for $C$ a constant.

This theorem and its proof can be found in Section 12.3 in Chatterjee (2014).

With this result at hand, we can turn to the proof of the main result.

---

1 Simulations based on 5000 replications for $n$ between 5 and 300.
Proof of Theorem 2.2. From the discussion in Section 3.1 regarding the correlations, we have

\[ |B(u, \delta)| = n! \sum_{k>(1-\delta)n}^{n} \frac{1}{k!} \sum_{\ell=0}^{n-k} \frac{(-1)^\ell}{\ell!}, \quad \forall 1 \leq u \leq n! \]

when we consider the Gaussian vector \( g \). It follows that

\[ V(\delta) \leq C_1 \frac{n!}{((1-\delta)n)!}, \]

for \( C_1 \) a constant, and thus

\[ \mathbb{E}(\log |A(\epsilon)|; A(\epsilon) \neq \emptyset) \leq \inf_{\delta \in (0,1)} \left( \log C_1 + \log \frac{n!}{((1-\delta)n)!} + \frac{C \max\{\epsilon m^2, m\}}{\delta} \right). \]

Further note that by Stirling’s approximation

\[ \log \frac{n!}{((1-\delta)n)!} = n \log n - n - (1-\delta)n \log ((1-\delta)n) + (1-\delta)n + O(\log n) + O(\delta \log n) \]

\[ = n[(\delta - 1) \log(1 - \delta) - \delta] + \delta n \log n + O(\log n) \]

\[ \leq \delta n \log n + O(\log n), \]

where in the third line we used that \((\delta - 1) \log(1 - \delta) - \delta \leq 0\), for all \( \delta \in (0,1) \).

Using the fact that \( m \leq \sqrt{2 \log(n!)} \) (Chatterjee, 2014, Proposition A.3), we get

\[ \mathbb{E}(\log |A(\epsilon)|; A(\epsilon) \neq \emptyset) \leq \inf_{\delta \in (0,1)} \left( C_2 + \delta n \log n + \frac{C \max\{2\epsilon \log(n!), \sqrt{2 \log(n!)}\}}{\delta} + O(\log n) \right). \]

It remains to distinguish the two cases.

If \( \epsilon < (2 \log(n!))^{-1/2} \), choose \( \delta = (n \log n)^{-1/4} \).

If \( \epsilon > (2 \log(n!))^{-1/2} \), set \( \delta = \sqrt{\epsilon} \). The claim follows. \( \square \)

3.3. Bounds on the expectation. We now turn to the proof of the second main result.

Proof of Theorem 2.3. The upper bound is classic and follows from (A.3) as in previous proof. For the lower bound, we use a classical argument relying on Payley–Zygmund inequality. It follows the lines of Theorem 8.3 in Chatterjee (2014); still, we reproduce it here for completeness. Also, in the case that interests us, more
attention must be paid to constants. Following the reference stated, we define
\[ N := |\{u : g_u \geq \sqrt{2 \log(n!)}\}|. \]
Using the Mill’s ratio lower bound, we have
\[ \mathbb{E}N \geq \frac{1}{\sqrt{2} \sigma} n! \sqrt{\frac{2 \log(n!)}{1 + 2 \log(n!)}}. \]

There thus exists an (unimportant) \( n_0 \), such that for all \( n \geq n_0 \) it holds
\[ \mathbb{E}N \geq \frac{3}{10} \frac{1}{\sqrt{2 \log(n!)}}. \]

Also one has,
\[ \mathbb{E}(N^2) = \sum_{u,v} \mathbb{P}(g_u \geq \sqrt{2 \log(n!)}, g_v \geq \sqrt{2 \log(n!)}) \]
\[ \leq \sum_{u,v} \exp \left( -\frac{4 \log(n!)}{\text{Var}(g_u + g_v)} \right) \]
\[ \leq \sum_{u,v} (n!)^{-2/(1 + r_{u,v})}. \]

Using the Payley–Zygmund inequality, it follows that
\[ \mathbb{P}(-W_n \geq \sqrt{2 \log(n!)}) = \mathbb{P}(N > 0) \geq \frac{(\mathbb{E}N)^2}{\mathbb{E}(N^2)} \]
\[ \geq \frac{9}{200 \log(n!) \sum_{u,v} (n!)^{-2/(1 + r_{u,v})}}. \]

Using the classical concentration inequality (Chatterjee, 2014, Proposition A.7)
\[ (4) \quad \mathbb{P}(-W_n - \mathbb{E}(-W_n) > x) \leq e^{-x^2/2} \]
for \( x \geq 0 \) and combining the latter with the previous result yields
\[ \mathbb{E}(-W_n) \geq \sqrt{2 \log(n!)} - \sqrt{2} \left( \log \left( \sum_{u,v} (n!)^{-2/(1 + r_{u,v})} \right) + \log \log(n!) - \log(9/200) \right)^{1/2}. \]

The conclusion will follow if we prove that there exists \( \epsilon > 0 \) such that
\[ \log \left( \sum_{u,v} (n!)^{-2/(1 + r_{u,v})} \right) \leq (1 - \epsilon) \log(n!) + o(n!). \]
First, note that
\[ \sum_{u,v}(n!)^{-2/(1+r_{u,v})} = (n!)^2 \sum_{k=0}^{n} \frac{1}{k!} (n!)^{-2/(1+k/n)} \sum_{\ell=0}^{n-k} \frac{(-1)^\ell}{\ell!} \]
\[ \leq \sum_{k=0}^{n} \frac{1}{k!} (n!)^{2k/(n+k)} \]
\[ = \sum_{k=0}^{n} \frac{2k}{n+k} \frac{\log(k)}{\log(n)}. \]

Looking at the form of the exponent, one can see that the claim follows by choosing \( \epsilon \) small enough. For instance, taking \( k \leq n/3 \), we have
\[ (n!)^{\frac{2k}{n+k} \frac{\log(k)}{\log(n)}} \leq (n!)^{\frac{1}{3}}. \]

While, for \( k \geq n/3 \), it holds
\[ (n!)^{\frac{2k}{n+k} \frac{\log(k)}{\log(n)}} \leq (n!)^{\frac{(n/3) \log(n/3)}{n \log n}} \leq (n!)^{\frac{1}{3} + O\left(\frac{1}{\log n}\right)}. \]

Unfortunately, finding the right order of the exponent is not so straightforward. Further, it would not help much getting closer to the true value of the expectation as equation (4) is expected not to be sharp enough; the discussion following will point this out.

4. Further questions and discussion

For a generic Gaussian process \( \{X_t\}_{t \in T} \), the theory of concentration of measure tells us that the fluctuations of sup \( t \in T X_t \) are not larger than sup \( t \in T (E(X_t^2))^{1/2} \). In many cases, this is not sharp and equation (4) can be improved. When the fluctuations of the supremum are of a much lower size, the process is said to exhibit superconcentration. This phenomenon also appears in the setting considered here; finding the correct order of variance is thus an open, natural problem.

When the costs are Exp(1) or uniform over \([0,1]\), it is known that the variance goes to zero as \( n \) goes to \( \infty \) (see Wästlund (2005) or (Talagrand (1995)). It may be expected that this also occurs here.

Also, limit theorems for Gaussian random variables are well-covered in the literature, in particular in the weakly dependent case and for stationary sequences. In the latter case, assuming some regularity on the correlation function, Mittal
and Ylvisaker (1975) proved that Gumbel limits, mixture of Gumbel and Normal distributions and Normal distributions could arise depending on the fact that the correlation multiplied by $\log(n)$ goes (respectively) to 0, a constant or $\infty$. Can we adapt their proof to obtain a normal limit in the case considered here? Following their strategy seems to give one of the best potential ways to finally proving a limit theorem for one of the most difficult random optimization problems.

Finally, considering the largest eigenvalue of the correlation matrix of an equicorrelated vector with correlations $(\log n)^{-1+\epsilon}$, one easily gets that the largest eigenvalue is $\sim n(\log n)^{-1+\epsilon}$ and this eigenvalue is associated with the eigenvector $\mathbf{1}$. In the latter case, a normal limit distribution holds for the maximum. Getting back to the correlation matrix in the random assignment problem we study here, one obtains that the largest eigenvalue is $n!/n$ and is also associated with the eigenvector $\mathbf{1}$. Thus, the overall geometries of the two Gaussian arrays under consideration are relatively similar in terms of principal direction. This further suggests that a limiting Gaussian distribution should appear in the problem studied in this paper. Can this geometric fact serve to derive a central limit theorem?

We hope to see these questions solved.

References

Aldous, David (2001). “The $\zeta(2)$ limit in the random assignment problem”. In: Random Structures & Algorithms 18.4, pp. 381–418.
Aldous, David and Michael J. Steele (2004). “The objective method: probabilistic combinatorial optimization and local weak convergence”. In: Probability on discrete structures. Springer, pp. 1–72.
Chatterjee, Sourav (2014). Superconcentration and related topics. Vol. 15. Springer.
– (2019). “A general method for lower bounds on fluctuations of random variables”. In: The Annals of Probability 47.4, pp. 2140–2171.
de Montmort, P. R. (1708). “Essai d’Analyse sur les Jeux de Hazard, Paris”. In: Mézard, Marc and Giorgio Parisi (1987). “On the solution of the random link matching problems”. In: Journal de Physique 48.9, pp. 1451–1459.
Mittal, Yash and Donald Ylvisaker (1975). “Limit distributions for the maxima of stationary Gaussian processes”. In: Stochastic Processes and their Applications 3.1, pp. 1–18.
Talagrand, Michel (1995). “Concentration of measure and isoperimetric inequalities in product spaces”. In: Publications Mathématiques de l’Institut des Hautes Etudes Scientifiques 81.1, pp. 73–205.
Wästlund, Johan (2005). The variance and higher moments in the random assignment problem. Linköping University Electronic Press.