COMBINATORIAL OPTIMIZATION PROBLEMS FOR WHICH ALMOST EVERY ALGORITHM IS ASYMPTOTICALLY OPTIMAL!

September 28, 1994

Wojciech Szpankowski*
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

Consider a class of optimization problems for which the cardinality of the set of feasible solutions is m and the size of every feasible solution is N. We prove in a general probabilistic framework that the value of the optimal solution and the value of the worst solution are asymptotically almost surely (a.s.) the same *provided* $\log m = o(N)$ as N and m become large. This result implies that for such a class of combinatorial optimization problems almost *every* algorithm finds asymptotically optimal solution! The quadratic assignment problem, the location problem on graphs, and a pattern matching problem fall into this class.

*This research was primary done while the author was visiting INRIA, Rocquencourt, France, and he wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support. In addition, support was provided by NSF Grants CCR-9201078, NCR-9206315 and INT-8912631, by Grant AFOSR-90-0107, and in part by NATO Collaborative Grant 0057/89.

1. INTRODUCTION

We consider a class of optimization problems that can be formulated as follows: Let n be an integer (e.g., number of vertices in a graph, size of a matrix, number of keys in a digital tree, etc.), and $\mathcal{S}_n$ a set of objects (e.g., set of vertices, elements of a matrix, keys, etc). We shall investigate the asymptotic behaviour of the optimal values $Z_{\max}(\mathcal{S}_n)$ and $Z_{\min}(\mathcal{S}_n)$ defined as follows

$$Z_{\max}(\mathcal{S}_n) = \max_{\alpha\in\mathcal{B}_n}\left\{\sum_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha)\right\} \quad , \quad Z_{\min}(\mathcal{S}_n) = \min_{\alpha\in\mathcal{B}_n}\left\{\sum_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha)\right\} , \tag{1}$$

and

$$Z_{\max}(\mathcal{S}_n) = \max_{\alpha\in\mathcal{B}_n}\min_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha) \quad , \quad Z_{\min}(\mathcal{S}_n) = \min_{\alpha\in\mathcal{B}_n}\max_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha) \tag{2}$$

where $\mathcal{B}_n$ is a set of all feasible solutions, $\mathcal{S}_n(\alpha)$ is a set of objects from $\mathcal{S}_n$ belonging to the α-th feasible solution, and $w_i(\alpha)$ is the weight assigned to the i-th object in the α-th feasible solution. We often write $Z_{\max}$ and $Z_{\min}$ instead of $Z_{\max}(\mathcal{S}_n)$ and $Z_{\min}(\mathcal{S}_n)$, respectively.

For example, in the traveling salesman problem, $\mathcal{B}_n$ represents the set of all Hamiltonian paths in a graph built over n vertices, $\mathcal{S}_n(\alpha)$ is the set of edges belonging to the α-th Hamiltonian path, and $w_i(\alpha)$ is the length (weight) of the i-th edge. Traditionally, formulation (1) is called the optimization problem with sum-objective function, while (2) is either known as the capacity optimization problem ($Z_{\max}$) or the bottleneck optimization problem ($Z_{\min}$).

Combinatorial optimization problems arise in many areas of science and engineering. Among others we mention here: the (capacity and bottleneck) assignment problem [14], the (bottleneck and capacity) quadratic assignment problem [8, 12], the minimum spanning tree, the minimum weighted k-clique problem [10], geometric location problems [11], and some others not directly related to optimization such as the height and depth of digital trees, the maximum queue length and hashing with lazy deletion [1], pattern matching [3], and so forth. We analyze this class of problems in a probabilistic framework which assumes that the weights $w_i(\alpha)$ are random variables drawn from a common distribution function $F(\cdot)$. We also assume that the cardinality of the feasible set is m (i.e., $|\mathcal{B}_n| = m$) and the cardinality of $\mathcal{S}_n(\alpha)$ is N for every $\alpha \in \mathcal{B}_n$.

Our interest lies in identifying a class of combinatorial problems for which $Z_{\min} \sim Z_*$ and $Z_{\max} \sim Z^*$ (a.s.) for $N, m \to \infty$ where Z_* and Z^* are the worst solutions of the above optimization problems. This will imply that almost every solution of such an optimization problem is asymptotically optimal in the sense that the relative error $(Z_{\max} - Z_*)/Z_{\max}$ (resp. $(Z_{\min} - Z^*)/Z_{\min}$) converges to zero in a probabilistic sense. As a simple consequence, one can pick any algorithm to solve these problems, and with high probability it will be asymptotically optimal!

More precisely, we prove that for the sum-objective function $Z_{\max} = N\mu + o(N)$ (a.s.) and $Z_{\min} = N\mu - o(N)$ (a.s.), and for the bottleneck and capacity optimization problems respectively $Z_{\min} \sim F^{-1}(1 - o(1))$ and $Z_{\max} \sim F^{-1}(o(1))$ (a.s.) *provided* $\log m = o(N)$ where μ and $F^{-1}(\cdot)$ are the average value and the inverse of the distribution for weights $w_i(\alpha)$. Observe that for the sum-objective function $Z_{\min}$ is also the worst solution for the optimization problem $Z_{\max}$ and *vice versa.*

There are many combinatorial problems that fall under our model. We mention here the quadratic assignment problem (cf. [8, 12]), a class of location problems, the pattern matching problem (cf. [2, 3]), and so forth (cf. [4]). We shall discuss some details of these problems in the last section.

The formulation of the problem and its solution seemed to be new, even if the analysis present in this paper is quite simple. There are some scattered results in this direction (cf. [10, 13, 14]), but none of them addresses this issue in its generality. There is, of course, a huge volume of literature on combinatorial optimization problems but usually one assumes $\log m = O(N)$ and every problem is treated case by case.

During the revision of this paper, we have learned that in 1985 Burkard and Fincke [4] studied exactly the same problem. However, the authors of [4] proved their result only for *bounded* distribution on $[0, 1]$ and *only* for convergence in probability. These restrictions are crucial for the proof presented in [4]. Actually, our almost sure convergence solves the problem posed by Burkard and Fincke [4]. Needless to say, our technique of the proof is completely different and this allows to extend the results of Burkard and Fincke to a very general probabilistic framework.

2. RESULTS

We consider separately optimization problems with the sum-objective function, and the capacity and bottleneck optimization problems.

2.1 Optimization Problems with Sum-Objective Function

Throughout this paper, we adopt the following assumptions:

(A) The cardinality $|\mathcal{B}_n|$ of $\mathcal{B}_n$ is fixed and equal to m. The cardinality $|\mathcal{S}_n(\alpha)|$ of the set $\mathcal{S}_n(\alpha)$ does *not* depend on $\alpha \in \mathcal{B}_n$ and for all α it is equal to N, i.e., $|\mathcal{S}_n(\alpha)| = N$.

(B) For all $\alpha \in \mathcal{B}_n$ and $i \in \mathcal{S}_n(\alpha)$ the weights $w_i(\alpha)$ are identically and independently distributed (i.i.d.) random variables with common distribution function $F(\cdot)$, and the mean value μ, the variance σ^2, and the third moment μ_3 are finite.

Assumption (B) defines a *probabilistic model* of our problem (1). Assumption (B) can be boldly relaxed by imposing only stationarity and some mixing conditions on the weights (which do not necessary have to be identically distributed, too). Also, extensions of assumption (A) are possible (e.g., the size of $S_n(\alpha)$ may depend on a feasible solution α). We shall not explore these possibilities in the paper.

For our strongest result (i.e., the almost sure convergence) we need an additional assumption that basically says that our combinatorial structure has a *monotonicity property*:

(C) The objective function $Z_{\max}(\mathcal{S}_n)$ (resp. $Z_{\min}(\mathcal{S}_n)$) is a nondecreasing one (resp. nonincreasing) with respect to n, and also $|\mathcal{B}_{n+1}| \geq |\mathcal{B}_n|$.

Most of combinatorial problems satisfy (C). For example, all problems discussed in Section 3 fall under (C).

Now we are ready to present our first main result.

Theorem 1. *Under assumptions (A)-(C), as $N, m \to \infty$ with $n \to \infty$*

$$Z_{\min} = N\mu - o(N) \qquad \text{(a.s.)} \qquad Z_{\max} = N\mu + o(N) \tag{3}$$

provided

$$\log m = o(N) \ . \tag{4}$$

If assumption (C) is dropped, then (3) holds in a weaker sense, namely $Z_{\max} \sim Z_{\min} \sim N\mu$ in probability (pr.).

Proof. We first prove (3) for the convergence in probability assuming only (A) and (B), and then by adding (C) we extend it to the almost sure convergence. Below, we consider only $Z_{\max}$. The lower bound trivially follows from the *Ergodic Theorem* (cf. [5]) and the fact that $\max_{\alpha \in \mathcal{B}_n} \{\sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha)\} \geq E\{\sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha)\} = N\mu$. We focus now on the upper bound.

Observe that we can rewrite (1) as

$$Z_{\max} = N\mu + \sigma\sqrt{N} \max_{\alpha \in \mathcal{B}_n} \left\{ \frac{\sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha) - N\mu}{\sigma\sqrt{N}} \right\} \ . \tag{5}$$

Let $X_\alpha = (\sum_{i \in \mathcal{S}_n(\alpha)} w_i(\alpha) - N\mu)/\sigma\sqrt{N}$. Then, the optimization problem is equivalent to finding the maximum over $\{X_\alpha\}_{\alpha \in \mathcal{B}_n}$.

Let $F_N(x) = \Pr\{X_\alpha \le x\}$. From Feller [7] (Chap. XVI.6) we know that for $x = o(\sqrt{N})$

$$\frac{1 - F_N(x)}{1 - \Phi(x)} = (1 + O(x/\sqrt{N}))\exp(\lambda_1 x^3/\sqrt{N}) \qquad \text{where} \qquad \lambda_1 = \frac{\mu_3}{6\sigma^3} \; , \tag{6}$$

and $\Phi(x)$ is the distribution function of the standard normal distribution. Now , by (6) and Boole's inequality for $x = o(\sqrt{N})$

$$\begin{aligned} \Pr\{\max_{\alpha \in \mathcal{B}_n} X_\alpha > x\} &= \Pr\{X_1 > x \text{ or } X_2 > x \text{ or } , \ldots , \text{ or } X_m > x\} \\ &\le m(1 - F_N(x)) = (1 + o(1))m(1 - \Phi(x))\exp(\lambda_1 x^3/\sqrt{N}) \; . \end{aligned} \tag{7}$$

Define a_m as the smallest solution to the following equation

$$m(1 - \Phi(a_m)) = 1 \; , \tag{8}$$

and observe that asymptotically $a_m \sim \sqrt{2\log m}$ (cf. [7]). Then, the inequality in (7) becomes for any $\varepsilon > 0$ as long as $a_m = o(\sqrt{N})$

$$\Pr\{\max_{\alpha \in \mathcal{B}_n} X_\alpha > a_m(1+\varepsilon)\} \le (1 + o(1))m(1 - \Phi(a_m(1+\varepsilon)))\exp\left(\lambda_1 a_m^3 (1+\varepsilon)^3/\sqrt{N}\right) \; .$$

But, asymptotically $1 - \Phi(a_m(1+\varepsilon)) \le (1 - \Phi(a_m))e^{-2\varepsilon a_m^2}$, and together with (8), this implies

$$\Pr\{\max_{\alpha \in \mathcal{B}_n} X_\alpha > a_m(1+\varepsilon)\} \le (1 + o(1))\exp\left(-a_m^2\left(2\varepsilon - \lambda_1 a_m (1+\varepsilon)^3/\sqrt{N}\right)\right) \; .$$

Finally, as long as $a_m/\sqrt{N} = o(1)$ (cf. (4)) one can find such $\delta > 0$ that

$$\Pr\{\max_{\alpha \in \mathcal{B}_n} X_\alpha > a_m(1+\varepsilon)\} \le \frac{1}{m^\delta} \tag{9}$$

which completes the proof of (3) for the convergence in probability.

To prove the stronger almost sure convergence result, we need some additional considerations. Note that (9) does not yet warrant an application of the Borel-Cantelli Lemma, hence we apply the idea presented in Kingman [9]. Let $Z_m = \max_{\alpha \in \mathcal{B}_n}\{X_\alpha\}$, and observe that under assumption (C) the quantity Z_m is a nondecreasing sequence with respect to n (hence also with respect to m due to (C) and (A)) such that $Z_m \sim \sqrt{2\log m}$ (pr.) with the rate of convergence as in (9). Fix now s, and find such r that $s2^r \le m \le (s+1)2^r$. The subsequence Z_{s2^r} almost surely converges to $\sqrt{2\log s2^r}$ by the Borel-Cantelli Lemma. Due to monotonicity of Z_m we also have for any m

$$\limsup_{n\to\infty} \frac{Z_m}{\sqrt{2\log m}} \le \limsup_{r\to\infty} \frac{Z_{(s+1)2^r}}{\sqrt{2\log(s+1)2^r}} \cdot \frac{\sqrt{2\log(s+1)2^r}}{\sqrt{2\log s2^r}} = 1 \qquad \text{(a.s.)} \; ,$$

and this completes the proof of the Theorem 1. ∎

Remark. In fact, from the proof one may conclude the following refinement of the upper bound: $Z_{\max} - N\mu = O(\sqrt{2\sigma^2 N \log m})$. It should be noted that the second term is of order $O(N)$ when $\log m = O(N)$, and our results brakes down. Nevertheless, even in the case $\log m = o(N)$ the second term may contribute significantly to the asymptotics, and in practice it cannot be completely ignored (cf. Section 3.3).

A direct consequence of Theorem 1 is the following corollary.

Corollary. *Let condition (4) hold. Then,*

$$\lim_{m\to\infty} \Pr\{Z_{\max} - Z_{\min} \le o(1) Z_{\min}\} = 1 \tag{10}$$

provided $N, m \to \infty$. ■

The above corollary says that *any* algorithm of the optimization problem almost always finds a good (i.e., asymptotically optimal) solution, *provided* condition (4) holds. Below, we discuss three well known combinatorial problems that fall under our assumptions.

In passing, we note that assumption (B) can be substantially relaxed. Indeed, the lower bound holds for all weights that form a stationary ergodic sequence. For the upper bound, we need an extension of (6) which holds for some stationary sequences with appropriate mixing conditions (cf. [5]). Also, the identically distributed weights can be replaced by a more general assumption as long as (6) can be established.

2.2 Bottleneck and Capacity Optimization Problems

In this subsection we consider the capacity and optimization problems defined in (2). In addition, we consider the worst solutions defined as

$$Z_*(\mathcal{S}_n) = \min_{\alpha\in\mathcal{B}_n} \min_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha) \quad , \quad Z^*(\mathcal{S}_n) = \max_{\alpha\in\mathcal{B}_n} \max_{i\in\mathcal{S}_n(\alpha)} w_i(\alpha) \ . \tag{11}$$

In the sequel, we write $Z_{\min}$ and Z^* instead of $Z_{\min}(\mathcal{S}_n)$ and $Z^*(\mathcal{S}_n)$, and we concentrate on the bottleneck optimization problems.

As above we adopt assumptions (A)-(C), however, we slightly modify assumption (B). Namely,

(B') The weights are i.i.d. random variables with distribution function $F(\cdot)$ that is a *strictly* increasing (continuous) function.

The important property of the bottleneck (and capacity) optimization problems – that allows us to obtain our results under general probabilistic framework (e.g., assumption (B') above) – is the so called *ranking-dependence* (cf. [13]). By this we mean that the optimal solution depends only on the rank of the weights $w_i(\alpha)$ but not on specific values of $w_i(\alpha)$. More formally, if $\mathcal{I}$ is the set of *strictly increasing* functions, then for every $f \in \mathcal{I}$ the following is true

$$f(Z_{\min}) = \min_{\alpha \in \mathcal{B}_n} \{ \max_{i \in \mathcal{S}_n(\alpha)} f(w_i(\alpha)) \} \ . \tag{12}$$

Since by assumption (B') the distribution function $F(\cdot)$ and its inverse $F^{-1}(\cdot)$ are strictly increasing, we can prove our result for a particular distribution (e.g., exponential or uniform), and then transform by $F^{-1}(\cdot)$ to any distribution.

Using the above ranking-dependence property of the bottleneck optimization problem, we prove the following result.

Theorem 2. (i) *For a class of bottleneck optimization problems satisfying assumptions (A), (B') and (C) the following holds as $N, m \to \infty$ with $n \to \infty$*

$$Z_{\min} = F^{-1}(1 - o(1)) \qquad \text{(a.s.)} \qquad Z^* = F^{-1}(1 - o(1)) \tag{13}$$

provided (4) holds, that is, $\log m = o(N)$. *Actually,* $F^{-1}(1 - \log m/N) \le Z_{\min} \le F^{-1}(1 - 1/N)$ *(a.s.). Thus,* $\lim_{m\to\infty} \Pr\{Z_{\min} - Z^* \le o(1) Z_{\min}\} = 1$.

(ii) *For the capacity optimization problem under the same assumptions as in (i) we have* $Z_{\max} \sim Z_* \sim F^{-1}(o(1))$. *Also* $\lim_{m\to\infty} \Pr\{Z_{\max} - Z_* \le o(1) Z_{\max}\} = 1$.

Proof. We only prove part (i). Let $X_\alpha = \max_{i \in \mathcal{S}_n} w_i(\alpha)$. Then,

$$\Pr\{Z_{\min} \le x\} \le m \Pr\{X_\alpha \le x\} = m\,(F(x))^N \ .$$

Let b_n be a solution of the following equation $mF^N(b_n) = 1$. Then, for any $\varepsilon > 0$ and uniform distribution (we select here our distribution that fits best to our purpose), the above becomes

$$\Pr\{Z_{\min} \le (1-\varepsilon) b_n\} \le m b_n^N (1-\varepsilon)^N = (1-\varepsilon)^N = o(1)$$

where the first equality of the above follows from the fact that $1 = mF^N(b_n) = m b_n^N$. Solving this equation we obtain $b_n = m^{-1/N} = e^{-\log m/N} = 1 - O(\log m/N) = 1 - o(1)$ since by (4) $\log m = o(N)$. This proves the lower bound (a.s.).

To obtain an upper bound, we consider Z^*, and as before we derive the following bound

$$\Pr\{Z^* > x\} \le Nm \Pr\{w_i(\alpha) > x\} = Nm(1 - F(x)) \ . \tag{14}$$

(Observe that we could also bound $Z_{\min}$ by $Z_{\min} \le \max_{i \in \mathcal{S}_n} w_i(\alpha) = X_\alpha$, and then in (14) Nm should be replaced by N.) Now, we consider the *exponential distribution*, and define a_n

as a solution of $Nme^{-a_n} = 1$, that is, $a_n = \log mN$. Note that the above becomes for any $\varepsilon > 0$

$$\Pr\{Z^* > (1+\varepsilon)a_n\} \leq \frac{1}{(nN)^\varepsilon} = o(1)$$

which proves the upper bound for the convergence in probability. To extend this result to almost sure convergence, we follow the footsteps of the proof of Theorem 1. ■

3. APPLICATIONS

In this section we discuss in some details three optimization problems, namely, the quadratic assignment problem, the location problem, and the pattern matching problem. We restrict the discussion to optimization problems with the sum-objective function. An extension to bottleneck and capacity optimization problems is easy (cf. [4]).

3.1 The Quadratic Assignment Problem

Let $A = (a_{ij})$ and $B = (b_{ij})$ be two real $n \times n$ matrices, and let $\pi(\cdot)$ be a permutation of $\{1, 2, \ldots, n\}$. Then, the *quadratic assignment problem* (QAP) is defined as (cf. [8])

$$Z_{\min} = \min_{\pi \in \mathcal{B}_n} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{\pi(i)\pi(j)} \right\} \tag{15}$$

where $\mathcal{B}_n$ is the set of all permutations of $\{1, \ldots, n\}$. Clearly, the QAP falls into the general formulation (1) with $N = n^2$ and $m = n!$. Note that $\log m \sim n \log n = o(n^2)$, so our condition (4) holds. Therefore, if assumption (B) is satisfied (e.g., this will hold if the matrices are generated independently from a common distribution), then Theorem 1 holds and $Z_{\min} \sim Z_{\max} \sim n^2\mu$ (a.s.) where $\mu = Ea_{ij}Eb_{ij}$. In fact, from the remark after the Corollary, we know that $Z_{\min} - n^2\mu = O(n^{3/2}\sqrt{\log n})$, as also proved by Rhee [12] in a more sophisticated probabilistic model.

In passing, we should note that the *linear assignment problem* (LAP) does not fall into our category. In this case, a single matrix A is given, and

$$Z_{\min} = \min_{\pi \in \mathcal{B}_n} \sum_{i=1}^{n} a_{i\pi(i)} \ .$$

Then, $N = n$ and $m = n!$, and hence $\log m \neq o(N)$. Theorem 1 does not apply to this situation. In fact, for the uniform distribution of weights we know that $1.43 \leq EZ_{\min} \leq 2$ (cf. [6]). It is conjectured that $EZ_{\min} \sim \pi^2/6 \approx 1.67\ldots$. On the other hand, it is easy to prove that for the exponential distribution of weights $Z_{\max} \sim n \log n$ (pr.) while for the normally distributed weights $Z_{\max} \sim n\sqrt{2 \log n}$ (pr.) (cf. [10, 13]).

3.2 Location Problem on Graphs

A general location problem can be formulated as follows (see also [11]). Let $x_1, x_2, \ldots, x_n$ be a given set of points. The median problem selects L points $c_1, c_2, \ldots, c_L$ so as to minimize (maximize) the distance between these points and the points $x_1, x_2, \ldots, x_n$. To formulate the problem in terms of our general optimization problem (1), we introduce a distance function (random variable) $d(x_i, x_j)$ which represents weight for a pair (x_i, x_j). As a feasible solution $\alpha = (c_1, \ldots, c_L)$, we accept any choice of L points out of n, so that cardinality of $|\mathcal{B}_n| = \binom{n}{L}$. Then, we have

$$Z_{\min} = \min_{\alpha \in \mathcal{B}_n} \sum_{i=1}^{n-L} \min_{1 \le j \le L} \{d(x_i, c_j)\} \ .$$

Some simplification of the problem can be achieved if one considers the location problem on a (complete directed) graph. Indeed, let w_{ij} be a weight assigned to the (i,j)-edge with the distribution function $F(\cdot)$. By a feasible solution, we understand a subset $\alpha = \{c_1, \ldots, c_L\} \subset \mathcal{M} = \{1, 2, \ldots, n\}$ of cardinality L of vertices in a complete graph K_n. Then, the L median problem becomes (for the maximum)

$$Z_{\max} = \max_{\alpha \in \mathcal{B}_n} \sum_{i \in \mathcal{M} - \alpha} \max_{j \in \alpha} \{w_{ij}\} \ .$$

Note that $|\mathcal{B}_n| = \binom{n}{L} \sim n^L/L!$ for bounded L. Let us define $W_i(\alpha) = \max_{j \in \alpha} w_{ij}$. Note that under assumption (B) the distribution $F_W(x)$ of $W_i(\alpha)$ is $F^L(x)$. The average value EW of $W_i(\alpha)$ is rather easy to evaluate in most interesting cases. For example, if the weights are exponentially distributed, then $EW = H_L$ where H_L is the L-th harmonic number; if the weights are uniformly distributed on $[0,1]$, then $EW = L/(L+1)$, and so forth. Since, $m = |\mathcal{B}_n| = n^L/L!$, and $N = n - L$, then for bounded L our condition (4) of Theorem 1 holds, and therefore

$$Z_{\min} \sim Z_{\max} = (n-L)EW + O(\sigma_W \sqrt{2nL \log n}) \sim (n-L)EW \qquad \text{(a.s.)} \ .$$

In particular, $Z_{\min} \sim Z_{\max} \sim (n-L)H_L$ (a.s.) for the exponential distribution of weights, and $Z_{\max} \sim Z_{\min} = (n-L)L/(L+1)$ (a.s.) for the uniformly distributed weights.

3.3 Pattern Matching Problem

We consider the following string matching problem: Given are two strings, a *text* string $\mathbf{a} = a_1 a_2 \ldots a_n$ and a *pattern* string $\mathbf{b} = b_1 b_2 \ldots b_K$ of lengths n and K respectively, such that symbols a_i and b_j belong to a V-ary alphabet $\Sigma = \{1, 2, \ldots, V\}$. Let C_i be the number of positions at which the substring $a_i a_{i+1} \ldots a_{i+K-1}$ agrees with the pattern $\mathbf{b}$. That is,

$C_i = \sum_{j=1}^{K} equal(a_{i+j-1}, b_j)$ where $equal(x, y)$ is one if $x = y$, and zero otherwise (the index j that is out of range is understood to stand for $1 + (j \bmod n)$). We are interested in the quantity

$$M_{m,K} = \max_{1\le i\le n} \{C_i\} = \max_{1\le i\le n} \left\{ \sum_{j=1}^{K} equal(a_{i+j-1}, b_j) \right\}$$

which represents the best matching between $\mathbf{b}$ and any K-substring of $\mathbf{a}$, and could be viewed as a measure of similarity between these strings. Clearly, the above problem falls into our general formulation with $m = n$ and $N = K$.

We analyze $M_{m,K}$ under the following probabilistic assumption: *symbols from the alphabet* Σ *are generated independently, and symbol* $i \in \Sigma$ *occurs with probability* p_i. This probabilistic model is known as the *Bernoulli model*. It is equivalent to our assumption (B). From Theorem 1 we conclude that $M_{n,K} \sim KP$ (a.s.) provided $\log n = o(K)$, where $P = \sum_{i=1}^{V} p_i^2$ is the average value of a match in a *given* position. The case $\log n = O(K)$ was treated in Arratia *et al.* [2].

From the proof of Theorem 1 we also conclude that for the case $\log n = o(K)$ we have $M_{n,K} \sim KP + O(\sqrt{2(P - P^2)K \log n})$ (pr.). However, a precise estimate of the second term in the above is quite involved. Recently, Atallah *et al.* [3] proved that for a wide range of input probabilities p_i the following is true: $M_{n,K} \sim KP + \sqrt{2(P - T)K \log n}$ (pr.) where $T = \sum_{i=1}^{V} p_i^3$.

ACKNOWLEDGMENT

The author would like to thank P. Jacquet (INRIA, Rocquencourt) for interesting discussions concerning this paper, and M. Golin (Hong Kong University) for pointing out the necessity of assumption (C) for our (a.s.) result. We also thank A. Frieze (Carnegie-Mellon) for an encouraging to publish this note.

References

[1] D. Aldous, M. Hofri and W. Szpankowski, Maximum Size of a Dynamic Data Structure: Hashing With Lazy Deletion Revisited, *SIAM J. Computing*, 21, pp. 713-732, 1992.

[2] Arratia, R., Gordon, L., and Waterman, M., The Erdös-Rényi Law in Distribution, for Coin Tossing and Sequence Matching, *Annals of Statistics*, 18, 539-570, 1990.

[3] M. Atallah, P. Jacquet and W. Szpankowski, Pattern Matching with Mismatches: A Randomized Algorithm and Its Analysis, *Random Structures & Algorithms*, 4, 191-213, 1993.

[4] R. Burkard and U. Fincke, Probabilistic Asymptotic properties of some Combinatorial Optimization Problems, *Discrete Applied Mathematics*, 12, 21-29, 1985.

[5] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York 1968.

[6] M. Dyer, A. Frieze, and C.J.H. McDiarmid, On Linear Programs with Random Costs, *Mathematical Programming*, 35, 3-16, 1986.

[7] Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley & Sons, New York, 1971.

[8] Frenk, J., van Houweninge, M. and Rinnooy Kan, A., Asymptotic Properties of the Quadratic Assignment Problem, *Mathematics of Operations Research*, 10, 100– 116, 1985.

[9] J.F.C. Kingman, *Subadditive Processes*, in Ecole d'Eté de Probabilités de Saint-Flour V-1975, Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin 1976.

[10] Lueker, G., Optimization Problems on Graphs with Independent Random Edge Weights, *SIAM J. Computing*, 10, 338–351, 1981.

[11] Papadimitriou, C., Worst-Case and Probabilistic Analysis of a Geometric Location Problem, *SIAM J. Computing*, 10, 542–557, 1981.

[12] W.T. Rhee, A Note on Asymptotic Properties of the Quadratic Assignment Problem, *Operations Research Letters*, 7, 197-200, 1988.

[13] W. Szpankowski, Combinatorial optimization through order statistics, *Proc. Second International Symposium on Algorithms*, Taipei, Taiwan, Lecture Notes in Computer Science, 557 (eds. W.L. Hsu and R.C.T. Lee), 208-217, Springer-Verlag 1991.

[14] Weide, B., Random Graphs and Graph Optimization Problems, *SIAM J. Computing*, 9, 552–557, 1980.