Expected length of the longest common subsequence for large alphabets

Marcos Kiwi*

Depto. Ing. Matemática and
Ctr. Modelamiento Matemático UMR 2071,
University of Chile
Correo 3, Santiago 170–3, Chile
e-mail: mkiwi@dim.uchile.cl

Martin Loebl†

Dept. of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic
e-mail: loebl@kam.mff.cuni.cz

Jiří Matoušek‡

Dept. of Applied Mathematics and
Institute of Theoretical Computer Science (ITI)
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic
e-mail: matousek@kam.mff.cuni.cz

*Gratefully acknowledges the support of ICM P01–05 and Fondecyt 1010689.

†Gratefully acknowledges the support of ICM-P01-05. This work was done while visiting the Dept. Ing. Matemática, U. Chile.

‡This research was done while visiting the Ctr. de Modelamiento Matemático, UMR–UChile 2071, U. Chile, Santiago, supported by Fondap in Applied Mathematics 2000–05.

Abstract

We consider the length L of the longest common subsequence of two randomly uniformly and independently chosen n character words over a k-ary alphabet. Subadditivity arguments yield that $\mathbf{E}[L]/n$ converges to a constant γ_k. We prove a conjecture of Sankoff and Mainville from the early 80's claiming that $\gamma_k\sqrt{k} \to 2$ as $k \to \infty$.

1 Introduction

Consider two sequences of length n, with letters from a size k alphabet Σ, say μ and ν. The longest common subsequence (LCS) problem is that of finding the largest value L for which there are $1 \leq i_1 < i_2 < \ldots < i_L \leq n$ and $1 \leq j_1 < j_2 < \ldots < j_L \leq n$ such that $\mu_{i_t} = \nu_{j_t}$, for all $t = 1, 2, \ldots, L$.

The LCS problem has emerged more or less independently in several remarkably disparate areas, including the comparison of versions of computer programs, cryptographic snooping, and molecular biology. The biological motivation of the problem is that long molecules such as proteins and nucleic acids like DNA can be schematically represented as sequences from a finite alphabet. Taking an evolutionary point of view, it is natural to compare two DNA sequences by finding their closest common ancestors. If one assumes that these molecules evolve only through the process of inserting new symbols in the representing strings, then ancestors are substrings of the string that represent the molecule. Thus, the length of the longest common subsequence of two strings is a reasonable measure of how close both strings are. In the mid 1970's, Chvátal and Sankoff [5] proved that the expected length of the LCS of two random k-ary sequences of length n when normalized by n converges to a constant. The value of this constant γ_k is unknown although much effort has been spent in finding good upper an lower bounds for it (see, for example, [3] and references therein). The best known upper and lower bounds for γ_k do not have a closed form. There are obtained either as numeric approximation to the solutions of a nonlinear equation or as a numeric evaluation of some series expansion (see [6] for a survey of such results).

Although the problem of determining γ_k has a simple statement, it has turned out to be a challenging mathematical endeavor. Moreover, its quite naturally motivated. Indeed, a claim that two DNA sequences of length n

are far apart makes sense provided their LCS differs significantly from $\gamma_4 n$ (since DNA sequence have 4 basis elements).

We analyze the behavior of γ_k for k tending to infinity, and more generally, we consider the expected length of the LCS when k is an (arbitrarily slowly growing) function of n and $n \to \infty$. The focus on the case where k grows with n is partly inspired by the work of Kiwi and Loebl [13]. For a bipartite graph G over two size n totally ordered color classes A and B, they considered

$$L(G) = \max\{L : \exists a_1 < \ldots < a_L,\, b_1 < \ldots < b_L,\, a_i b_i \in E(G),\, 1 \leq i \leq L\},$$

and studied its behavior when G is uniformly chosen among all possible d-regular bipartite graphs on A and B. They established that $L_n(G)/\sqrt{dn} \to 2$ as $n \to \infty$ provided $d = o(n^{1/4})$. Under this latter condition, any node of the d-regular bipartite graph can potentially be matched to a $d/n \to 0$ fraction of the other color class nodes. In the case of interest here, that is the LCS problem with $k \to \infty$, it also happens that any sequences' character can be matched to an expected $1/k \to 0$ fraction of the other sequence's characters. Both for this work and in [13], the vanishing fraction of (expected) potential matches is a key issue.

In this paper we confirm a conjecture of Sankoff and Mainville from the early 80's [17] stating that

$$\lim_{k\to\infty} \gamma_k \sqrt{k} = 2\,. \tag{1}$$

(See [16, § 6.8] for a discussion of work on lower and upper bounds on γ_k as well as a stronger version, due to Arratia and Steele, of the above stated conjecture.)

The constant 2 in (1) arises from a connection with another celebrated problem known as the longest increasing sequence (LIS) problem. The problem is also referred to as "Ulam's problem." (e.g., in [12, 4, 15]). Some (e.g., [16]) incorrectly credit Ulam for raising it in [20] where he mentions (without reference) a "well–known theorem" asserting that given n^2+1 integers in any order, it is always possible to find among them a monotone subsequence of $n+1$. The theorem is due to Erdős and Szekeres [7]. The discussion in [20] concerns only the behavior of the monotonic subsequence of a randomly and uniformly chosen permutation of n^2+1 elements. Monte Carlo simulations are reported in [2], where it is observed that over the range $n \leq 100$, the limit of the LIS of n^2+1 randomly chosen elements, when normalized by n, approaches 2. Hammersley [9] gave a rigorous proof of the existence of the limit and conjectured it was equal to 2. Later, Logan and Shepp [14], based on a result by Schensted [18], proved that $\gamma \geq 2$; finally, Vershik and Kerov [21] obtained that $\gamma \leq 2$. In a major recent breakthrough due to Baik, Deift, Johansson [4] the asymptotic distribution of the longest increasing sequence random variable has been determined. For a detailed account of these results, history and related work see the surveys of Aldous and Diaconis [1] and Stanley [19].

It has been speculated that the behavior of the longest strictly/weakly increasing subsequence of a uniform random word of length n, with letters from Σ may have "connections with the subject of sequence comparison statistics, motivated by DNA sequence matching ..." [1]. Our work re-enforces this speculation and in fact does more. It partly elicits the nature of the connection and the conditions under which sequence matching statistics relate to the behavior of longest increasing sequences.

2 Statement of Results

Let A and B henceforth denote two disjoint totally ordered sets. We assume that the elements of A are numbered $1, 2, \ldots, |A|$ and those of V are numbered $1, 2, \ldots, |B|$. We denote by r and s the size of $|A|$ and $|B|$, respectively. Typically, we have $r = s = n$.

Now, let G be a bipartite graph with color classes A and B. Two distinct edges ab and $a'b'$ of G are said to be *noncrossing* if a and a' are in the same order as b and b'; in other words, if $a < a'$ and $b < b'$ or $a' < a$ and $b' < b$. A matching of G is called *planar* if every distinct pair of its edges is noncrossing. We let $L(G)$ denote the number of edges of a maximum size planar matching in G (note that $L(G)$ depends on the graph G *and* on the ordering of its color classes).

We will focus on the following two models of random graphs:

- The *random words model* $\Sigma(K_{n,n}; k)$: the distribution over the set of subgraphs of $K_{n,n}$ obtained by uniformly and independently assigning each node of $K_{n,n}$ one of k characters and keeping those edges whose endpoints are associated to equal characters. Note that only disjoint unions of complete bipartite graphs may appear in this model.

- The *binomial random graph model* $G(K_{n,n}; p)$: the distribution over the set of subgraphs of $K_{n,n}$ where each edge of $K_{n,n}$ is included with probability p, and these events are mutually independent. (This is an obvious modification of the usual $G(n,p)$ model for bipartite graphs with ordered color classes.)

In order to keep the presentation simple, we first formulate and prove the results for the random words model. Then, in Section 7, we state analogous results for the binomial random graph model. These results' proofs are almost identical to the case of the random words model, and we only briefly comment on them.

Our results essentially say that $L(\Sigma(K_{n,n};k)) \cdot \sqrt{k}/n$ converges to 2 as $k \to \infty$, provided that n is sufficiently large in terms of k.

Theorem 1 *For every $\varepsilon > 0$ there exist k_0 and C such that for all $k > k_0$ and all n with $n/\sqrt{k} > C$ we have*

$$(1-\varepsilon) \cdot \frac{2n}{\sqrt{k}} \;\leq\; \mathbf{E}\left[L(\Sigma(K_{n,n};k))\right] \;\leq\; (1+\varepsilon) \cdot \frac{2n}{\sqrt{k}}\,.$$

Moreover, there is an exponentially small tail bound; namely, for every $\varepsilon > 0$ there exists $c > 0$ such that for k and n as above,

$$\mathbf{P}\left[\left|L(\Sigma(K_{n,n};k)) - \frac{2n}{\sqrt{k}}\right| \geq \varepsilon \frac{2n}{\sqrt{k}}\right] \leq e^{-cn/\sqrt{k}}.$$

Corollary 2 *The limit $\gamma_k = \lim_{n\to\infty} \mathbf{E}\left[L(\Sigma(K_{n,n};k))/n\right]$ exists, and*

$$\lim_{k\to\infty} \gamma_k \sqrt{k} = 2.$$

3 Tools

The crucial ingredient in our proofs is a sufficiently precise result on the distribution of the length of the longest increasing subsequence in a random permutation. We state a remarkable strong result of Baik, Deift and Johansson [4, eqn. (1.7) and (1.8)] (our formulation slightly weaker than theirs, in order to make the statement simpler). A much weaker tail bound than provided by them would actually suffice for our proof.

Theorem 3 *Let LIS_N be the random variable corresponding to the length of the longest increasing subsequence of a randomly chosen permutation of*

$\{1,\ldots,N\}$. *There are positive constants* B_0, B_1, *and* c *such that for every* λ *with* $B_0/N^{1/3} \le \lambda \le \sqrt{N}-2$,

$$\mathbf{P}\Big[\mathrm{LIS}_N \ge 2\sqrt{N} + \lambda\sqrt{N}\Big] \le B_1 \exp\Big(-c\lambda^{3/5}N^{1/5}\Big),$$

and for every λ *with* $B_0/N^{1/3} \le \lambda \le 2$,

$$\mathbf{P}\Big[\mathrm{LIS}_N \le 2\sqrt{N} - \lambda\sqrt{N}\Big] \le B_1 \exp\Big(-c\lambda^{3}N\Big).$$

We will also need a suitable version of Talagrand's inequality; see, e.g., [10, Theorem 2.29].

Theorem 4 (Talagrand's inequality) *Suppose that* $Z_1,\ldots,Z_N$ *are independent random variables taking their values in some set* Λ. *Let* $X = f(Z_1,\ldots,Z_N)$, *where* $f : \Lambda^N \to \mathbf{R}$ *is a function such that the following two conditions hold for some number* c *and a function* ψ:

- *(L) If* $z, z' \in \Lambda^N$ *differ only in the* k*th coordinate, then* $|f(z) - f(z')| \le c$.
- *(W) If* $z \in \Lambda^N$ *and* $r \in \mathbf{R}$ *with* $f(z) \ge r$, *then there exists a witness* $(\omega_j : j \in J)$, $J \subseteq \{1,\ldots,N\}$, $|J| \le \psi(r)/c^2$, *such that for all* $y \in \Lambda^N$ *with* $y_i = \omega_i$ *when* $i \in J$, *we have* $f(y) \ge r$.

Let m *be a median of* X. *Then, for all* $t \ge 0$,

$$\mathbf{P}[X \ge m + t] \le 2e^{-t^2/4\psi(m+t)}.$$

and

$$\mathbf{P}[X \le m - t] \le 2e^{-t^2/4\psi(m)}.$$

We will also need the following version of Chebyshev's inequality:

Lemma 5 *Let* $X_1,\ldots,X_N$ *be random variables attaining values* 0 *and* 1, *and let* $X = \sum_{i=1}^N X_i$. *Let* $\Delta = \sum_{i\ne j} \mathbf{E}[X_iX_j]$. *Then, for all* $t > 0$,

$$\mathbf{P}[\,|X - \mathbf{E}[X]| \ge t] \le \frac{1}{t^2}\Big(\mathbf{E}[X](1-\mathbf{E}[X]) + \Delta\Big).$$

Proof: Since $\mathbf{P}[|X - \mathbf{E}[X]| \ge t] \le \mathrm{Var}[X]/t^2$ and

$$\begin{aligned}\mathrm{Var}[X] &= \sum_{i,j}(\mathbf{E}[X_iX_j] - \mathbf{E}[X_i]\,\mathbf{E}[X_j]) \\ &= \sum_i \mathbf{E}\left[X_i^2\right] - \sum_{i,j}\mathbf{E}[X_i]\,\mathbf{E}[X_j] + \sum_{i\ne j}\mathbf{E}[X_iX_j]\,,\end{aligned}$$

the desired conclusion follows by additivity of expectation and the fact that since X_i is an indicator variable, $X_i^2 = X_i$. ■

4 Small graphs

In this section we derive a result essentially saying that Theorem 1 holds if k is sufficiently large in terms of n. For technical reasons, we also need to consider bipartite graphs with color classes of unequal sizes.

Proposition 6 *For every $\delta > 0$, there exists a (large) positive constant C such that:*

(i) *If $rs \geq Ck$ and $(r+s)\sqrt{rs} \leq \delta k^{3/2}/6$, then with $m_u = m_u(r,s) = 2(1+\delta)\sqrt{rs/k}$, we have*

$$\mathbf{P}[L(\Sigma(K_{r,s};k)) \geq m_u + t] \leq 2e^{-t^2/8(m_u+t)}$$

for all $t \geq 0$.

(ii) *If $rs \geq Ck$ and and $r+s \leq \delta k/6$, then with m_u as above and $m_l = m_l(r,s) = 2(1-\delta)\sqrt{rs/k}$, we have*

$$\mathbf{P}[L(\Sigma(K_{r,s};k)) \leq m_l - t] \leq 2e^{-t^2/8m_u}$$

for all $t \geq 0$.

Let G be a random bipartite graph generated according to the random words model $\Sigma(K_{r,s};k)$. The idea of the proof is simple: we show that (ignoring degree 0 nodes) G is "almost" a matching, and the size of the largest planar matching in a random matching corresponds precisely to the length of the longest increasing sequence in a random permutation of the appropriate size.

We have to deal with the (usually few) vertices of degree larger than one. To this end, we define a graph G' obtained from G by removing all edges incident to nodes of degree at least 2. Throughout, E and E' denote $E(G)$ and $E(G')$, respectively.

We clearly have $\mathbf{E}[|E|] = rs/k$. We will need a tail bound for large deviation from the expectation; a simple second-moment argument (Chebyshev's inequality) suffices.

Lemma 7 *For every $\eta > 0$,*

$$\mathbf{P}\left[\left||E| - \frac{rs}{k}\right| \geq \eta \cdot \frac{rs}{k}\right] \leq \frac{1}{\eta^2(rs/k)}\,.$$

Proof: For $e \in E(K_{r,s})$ let X_e be the indicator of the event $e \in E$. Furthermore, let $X = |E| = \sum_{e \in E} X_e$. The X_e's are indicator random variables with expectation $1/k$. Moreover, since $\mathbf{E}[X_e X_f] = 1/k^2$ for $e \neq f$, we have $\sum_{e \neq f} \mathbf{E}[X_e X_f] = rs(rs-1)/k^2 = (\mathbf{E}[X])^2 - \mathbf{E}[X]/k$. Thus, Lemma 5 yields

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq \eta \mathbf{E}[X]] \leq \frac{1}{\eta^2 \mathbf{E}[X]} \left(1 - \frac{1}{k}\right).$$

The desired conclusion follows immediately. ■

Now we bound above the expectation of $|E \setminus E'|$.

Lemma 8

$$\mathbf{E}[|E \setminus E'|] \leq (r+s)\frac{rs}{k^2}.$$

Proof: Let Y_w equal the degree of w if it is at least 2 and 0 otherwise. Define $Y = \sum_{w \in V(G)} Y_w$. Note that $|E \setminus E'| \leq Y$ (equality does not necessarily hold since both endpoints of an edge might be incident on nodes of degree at least 2). Let P_d be the probability that a vertex in color class A has exactly d incident edges. For any node a in color class A,

$$\mathbf{E}[Y_a] = \sum_{d=2}^{s} d P_d = \mathbf{E}[\deg_G(a)] - P_1 = \frac{s}{k} - \frac{s}{k}\left(1 - \frac{1}{k}\right)^{s-1} \leq \left(\frac{s}{k}\right)^2$$

(using $(1-x)^h \geq 1 - hx$). Similarly $\mathbf{E}[Y_b] \leq (r/k)^2$ for all nodes b in color class B, and so

$$\mathbf{E}[|E \setminus E'|] \leq \mathbf{E}[Y] \leq (r+s)\frac{rs}{k^2}.$$

■

Proof of Proposition 6. Changing one of the characters associated to a vertex of a bipartite graph G changes the value of $L(G)$ by at most 1. Hence $L(G)$ is 1-Lipschitz. Furthermore, the characters associated to 2ω nodes of G suffice to certify the existence of ω noncrossing edges (and thus $L(G) \geq \omega$). So Talagrand's inequality applies and, with m denoting a median of $L(G)$, yields

$$\mathbf{P}[L(G) \geq m + t] \leq 2e^{-t^2/8(m+t)} \quad \text{and} \quad \mathbf{P}[L(G) \leq m - t] \leq 2e^{-t^2/8m}.$$

The proposition will follow once we show that $m_l \le m \le m_u$. To prove that $m \le m_u$, it suffices to verify that

$$\mathbf{P}[L(G) \ge m_u] \le \frac{1}{2}\,. \tag{2}$$

Let $\eta > 0$ be a suitable real parameter which we will specify later. We observe that since $|E'| \le |E|$ and $L(G) - L(G') \le |E \setminus E'|$,

$$\begin{aligned}\mathbf{P}[L(G) \ge m_u] \;\;\le\;\; & \mathbf{P}\left[|E| \ge (1+\eta)\frac{rs}{k}\right] \\ & + \mathbf{P}\left[|E \setminus E'| \ge \delta\sqrt{\frac{rs}{k}}\right] \\ & + \mathbf{P}\left[L(G') > (2+\delta)\sqrt{\frac{rs}{k}}, |E'| < (1+\eta)\frac{rs}{k}\right].\end{aligned}$$

We bound the terms one by one. By Lemma 8 and Markov's inequality,

$$\mathbf{P}\left[|E \setminus E'| \ge \delta\sqrt{\frac{rs}{k}}\right] \le \frac{r+s}{\delta k}\sqrt{\frac{rs}{k}} \le \frac{1}{6}\,. \tag{3}$$

Taking $N = (1+\eta)rs/k$ and $\lambda = [(2+\delta)/\sqrt{1+\eta}] - 2 > 0$ in Theorem 3, we get that

$$\begin{aligned}&\mathbf{P}\left[L(G') \ge (2+\delta)\sqrt{\frac{rs}{k}}, |E'| < (1+\eta)\frac{rs}{k}\right] \\ &\quad \le B_1 \exp\left(-c\lambda^{3/5}N^{1/5}\right) \le B_1 \exp\left(-c\lambda^{3/5}\left(\frac{rs}{k}\right)^{1/5}\right).\end{aligned} \tag{4}$$

From Lemma 7, (3), and (4), it follows that

$$\mathbf{P}[L(G) \ge m_u] \le \frac{1}{\eta^2(rs/k)} + \frac{1}{6} + B_1 \exp\left(-c\lambda^{3/5}\left(\frac{rs}{k}\right)^{1/5}\right).$$

So, (2) follows by taking, say, $\eta = \sqrt{6/C}$ and using $rs \ge Ck$.

To establish that $m_l \le m$, we proceed as before, i.e., we show that

$$\mathbf{P}[L(G) \le m_l] \le \frac{1}{2}\,. \tag{5}$$

Indeed, observe that since $|E'| = |E| - |E \setminus E'|$ and $L(G') \le L(G)$,

$$\begin{aligned}\mathbf{P}[L(G) \le m_l] \;\;\le\;\; & \mathbf{P}\left[|E| \le (1-\eta)\,\frac{rs}{k}\right] \\ & + \mathbf{P}\left[|E \setminus E'| \ge \delta \cdot \frac{rs}{k}\right] \\ & + \mathbf{P}\left[L(G') \le 2(1-\delta)\sqrt{\frac{rs}{k}}, |E'| > (1-\eta-\delta)\frac{rs}{k}\right].\end{aligned}$$

We again bound the terms one by one, applying as done above Lemma 8, Markov's inequality and Theorem 3, respectively. Indeed, for a suitable real value $\eta > 0$ and $\lambda = 2 - [2(1-\delta)/\sqrt{1-2\eta}] > 0$ we get

$$\mathbf{P}[L(G) \leq m_l] \leq \frac{1}{\eta^2(rs/k)} + \frac{1}{6} + B_1 \exp\left(-c\lambda^3 \frac{rs}{k}\right) .$$

So, (5) follows by taking again $\eta = \sqrt{6/C}$ and using $rs \geq Ck$. Proposition 6 is proved. ■

5 The lower bound in Theorem 1

In this section we establish the lower bound on the expectation of $L(\Sigma(K_{n,n};k))$ and the lower tail bound for its distribution.

Given ε, let $\delta > 0$ be such that $(1-2\delta)^2 = 1-\varepsilon$, and let $C = C(\delta)$ be as in Proposition 6. Fix $\widetilde{C} \geq \sqrt{C}$ large enough so that

$$\exp\left(-\frac{\delta^2}{4(1+\delta)} \cdot \widetilde{C}\right) \leq \delta .$$

Let $\tilde{n}(k) = \tilde{n} = \lfloor \delta k/12 \rfloor$. Proposition 6 applies for $k \geq k_0$ where k_0 is such that $\tilde{n}(k_0) \geq \widetilde{C}\sqrt{k_0}$. It follows that

$$\begin{aligned}
\mathbf{E}\left[L(\Sigma(K_{\tilde{n},\tilde{n}};k)\right] &\geq (1-2\delta) \cdot \frac{2\tilde{n}}{\sqrt{k}} \cdot \mathbf{P}\left[L(G) \geq 2(1-2\delta)\frac{\tilde{n}}{\sqrt{k}}\right] \\
&\geq (1-2\delta) \cdot \frac{2\tilde{n}}{\sqrt{k}} \left(1 - 2\exp\left(-\frac{\delta^2}{4(1+\delta)} \cdot \frac{\tilde{n}}{\sqrt{k}}\right)\right) \\
&\geq (1-\varepsilon) \cdot \frac{2\tilde{n}}{\sqrt{k}} .
\end{aligned}$$

The desired lower bound on the expectation follows since by subadditivity, $(1/n) \cdot \mathbf{E}\left[L(\Sigma(K_{n,n};k)\right]$ is nondecreasing.

Now we establish the lower tail bound. Let $\tilde{n} = \lceil C\sqrt{k}\, \rceil$ and $q = \lfloor n/\tilde{n} \rfloor$. Moreover, let G be chosen according to $\Sigma(K_{n,n};k)$ and let G_i be the subgraph induced in G by the vertices $(i-1) \cdot \tilde{n} + 1, \ldots, i \cdot \tilde{n}$ in each color class, $i = 1, \ldots, q$. We observe that $L(G_1), \ldots, L(G_q)$ are independent identically distributed with distribution $\Sigma(K_{\tilde{n},\tilde{n}};k)$ and $L(G) \geq L(G_1) + \cdots + L(G_q)$.

Let $\mu = \mathbf{E}[L(G_i)]$ and $t = \varepsilon(2n/\sqrt{k})$. Since $n \leq (q+1)\tilde{n}$, the lower bound on μ proved above yields that

$$\mathbf{P}\left[L(G) \leq (1-3\varepsilon) \cdot \frac{2n}{\sqrt{k}}\right] \leq \mathbf{P}\left[\sum_{i=1}^{q} L(G_i) \leq q\mu - t + (\mu - t)\right].$$

An argument similar to the one used above to derive the bound $\mu \geq (1-\varepsilon)2\tilde{n}/\sqrt{k}$ can be used to obtain $\mu \leq (1+\varepsilon)2\tilde{n}/\sqrt{k}$ from Propostion 6. Let n be large enough so that $n \geq \tilde{n}(1+2\varepsilon)/\varepsilon$. Thus, $q \geq (1+\varepsilon)/\varepsilon$ and $t \geq \varepsilon q\mu/(1+\varepsilon) \geq \mu$. Hence, a standard Chernoff bound [10, Theorem 2.1] implies that

$$\mathbf{P}\left[L(G) \leq (1-3\varepsilon) \cdot \frac{2n}{\sqrt{k}}\right] \leq \mathbf{P}\left[\sum_{i=1}^{q} L(G_i) \leq q\mu - t\right] \leq \exp\left(-\frac{t^2}{2q\mu}\right) \leq \exp\left(-\frac{\varepsilon^2}{2(1+\varepsilon)} \cdot \frac{2n}{\sqrt{k}}\right).$$

6 The upper bound in Theorem 1

We will only discuss the tail bound since $L(\Sigma(K_{n,n}; k)) \leq n$ always, and so the claimed estimate for the expectation follows from the tail bound.

Let $\varepsilon > 0$ be fixed. We choose a sufficiently small $\delta = \delta(\varepsilon) > 0$, much smaller than ε. Requirements on δ will be apparent from the subsequent proof.

Henceforth, we fix constants $1/2 < \alpha < \beta < 3/4$ (any choice of α and β in the specified range would suffice for our purposes). In this section, we will always assume that $k \geq k_0$ for a sufficiently large integer $k_0 = k_0(\varepsilon)$, and that n is sufficiently large compared to k: $n \geq k^\beta$, say. Note that for $n \leq k^\beta$ (and k sufficiently large), the tail bound of Theorem 1 follows from Proposition 6.

Block partitions. Let us write

$$m_{\max} = (1+\varepsilon) \cdot \frac{2n}{\sqrt{k}}$$

for the upper bound on the expected size of a planar matching as in Theorem 1. We also define an auxiliary parameter

$$\ell = k^\alpha.$$

This is a somewhat arbitrary choice (but given by a simple formula). The essential requirements on ℓ are that ℓ be much larger than $\sqrt{k}$ and much smaller than $k^{3/4}$. We note that n/ℓ is large by our assumption $n \geq k^\beta$.

Let M be a planar matching with $m_{\max}$ edges on the sets A and B, $|A| = |B| = n$. We define a partition of M into blocks of consecutive edges. There will be roughly n/ℓ blocks, each of them containing at most

$$e_{\max} = \left\lfloor \frac{1}{\delta} \cdot \frac{\ell}{n} \cdot m_{\max} \right\rfloor$$

edges of M. So $e_{\max}$ is of order $\ell/\sqrt{k}$, which by our assumptions can be assumed to be larger than any prescribed constant. Moreover, we require that no block is "spread" over more than ℓ consecutive nodes in A or in B.

Formally, the ith block of the partition will be specified by nodes $a_i, a'_i \in A$ and $b_i, b'_i \in B$; $a_i b_i \in M$ is the first edge in the block and $a'_i b'_i \in M$ is the last edge (the block may contain only one edge, and so $a_i b_i = a'_i b'_i$ is possible). The edge $a_1 b_1$ is the first edge of M, and $a_{i+1} b_{i+1}$ is the edge of M immediately following $a'_i b'_i$. Finally, given $a_i b_i$, the edge a'_i, b'_i is taken as the rightmost edge of M such that

- the ith block has at most $e_{\max}$ edges of M, and
- $a'_i - a_i \leq \ell$ and $b'_i - b_i \leq \ell$ (here and in the sequel, with a little abuse of notation, we regard the nodes in A and those in B as natural numbers $1, 2, \ldots, n$, although of course, the nodes in A are distinct from those of B).

Let q denote the number of blocks obtained in this way. It is easily seen that $q = O(n/\ell)$.

A block partition is schematically illustrated in Fig. 1.

Counting the types. Let e_i be the number of edges of M in the ith block. Let us call the $5q$-tuple $T = (a_1, a'_1, b_1, b'_1, e_1, \ldots, a_q, a'_q, b_q, b'_q, e_q)$ the *type* of the block partition of M, and let us write $T = T(M)$. Let $\mathcal{T}$ denote the set of all possible types of block partitions of planar matchings as above.

Lemma 9 *We have*

$$|\mathcal{T}| \leq \exp\left(C_1 \frac{n}{\ell} \log \ell\right)$$

with a suitable absolute constant C_1.

Proof: The number of choices for $a_1, \ldots, a_q$ is at most the number of ways of choosing q elements out of n, i.e., $\binom{n}{q}$. Since $m_{\max} \leq n$, the number

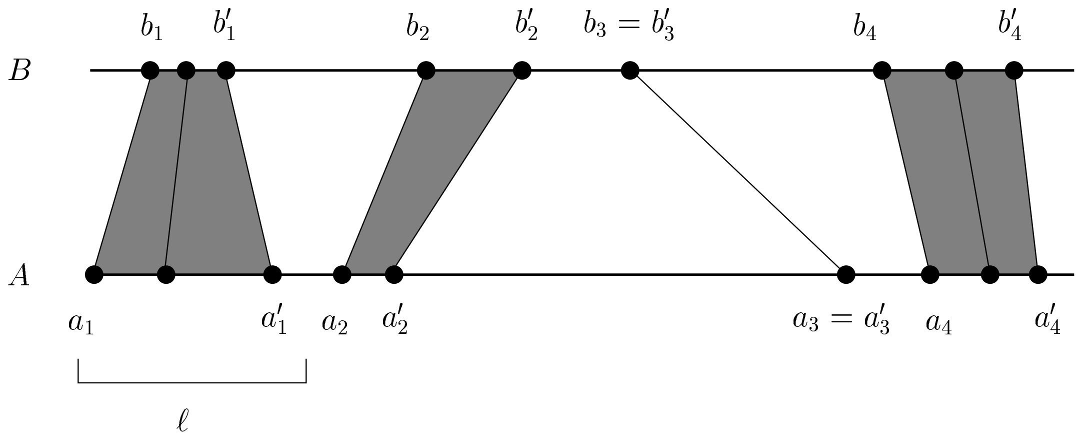

Figure 1: A block partition.

of choices for the e_i is no larger than the number of partitions of n into q positive summands, which is $\binom{n}{q}$. Grossly overestimating, for a fixed q we can thus bound the number of types by $\binom{n}{q}^5$. Using the standard estimate $\binom{n}{q} \le (en/q)^q$ and $q = O(n/\ell)$, we get $\log |\mathcal{T}| = O((n/\ell)\log \ell)$ as claimed. ■

The probability of a matching with a given type of block partition. Next we show that for every fixed type T, the probability that our random graph contains a planar matching of size $m_{\max}$ with that type of block partition is very small.

Lemma 10 *Let n and k be as above. For any given type $T \in \mathcal{T}$, the probability p_T that the random graph $\Sigma(K_{n,n}; k)$ contains a planar matching M with $m_{\max}$ edges and with $T(M) = T$ satisfies*

$$p_T \le \exp\left(-c\varepsilon^2\delta \cdot \frac{n}{\sqrt{k}}\right)$$

with a suitable absolute constant $c > 0$.

Proof: Let G_i denote the subgraph of the considered random graph $\Sigma(K_{n,n}; k)$ induced by the nodes $a_i, a_i + 1, \ldots, a'_i$ and $b_i, b_i + 1, \ldots, b'_i$. We note that the distribution of G_i is the same as that of $\Sigma(K_{r_i,s_i}; k)$, where $r_i = a'_i - a_i + 1$ and $s_i = b'_i - b_i + 1$.

A necessary condition for the existence of a planar matching M with $T(M) = T$ is $L(G_i) \geq e_i$ for all $i = 1, 2, \ldots, q$. Crucially for the proof, the events $L(G_i) \geq e_i$ are independent for distinct i, and so we have

$$p_T \leq \prod_{i=1}^{q} \mathbf{P}[L(\Sigma(K_{r_i,s_i}; k)) \geq e_i] .$$

The plan is to apply Proposition 6(i) for each i. The construction of the block partition guarantees that $r_i, s_i \leq \ell$, and so the condition $(r_i + s_i)\sqrt{r_i s_i} \leq \delta K^{3/2}/6$ in Proposition 6 is satisfied. However, the condition $r_i s_i \geq Ck$ may fail. To remedy this, we artificially enlarge the blocks; clearly, this can only increase the probability that a planar matching of size e_i is present.

Let us call the ith block *short* if it is the last block, i.e., $i = q$, or if $e_i = e_{\max}$. Let $S \subseteq [q]$ denote the set of all indices of short blocks. We have $(|S|-1)e_{\max} \leq m_{\max}$, and since $e_{\max} \geq \frac{1}{\delta} \cdot \frac{\ell}{n} \cdot m_{\max} - 1$, we obtain $|S| \leq 2\delta n/\ell$.

The blocks that are not short are called *regular*, and we write $R = [q] \setminus S$. For a regular block i, we have $\max(a_{i+1} - a_i, b_{i+1} - b_i) \geq \ell$ by the construction of the block partition.

Now we define the sizes of the artificially enlarged graphs, which will replace the G_i in the subsequent calculation. Namely, for a short block $(i \in S)$, we set

$$\bar{r}_i = \bar{s}_i = \ell.$$

For a regular block $(i \in R)$, we distinguish two cases. If $a_{i+1} - a_i \geq \ell$, we set $\bar{r}_i = \ell$ and $\bar{s}_i = \max(\delta\ell, s_i)$. Otherwise, we set $\bar{r}_i = \max(\delta\ell, r_i)$ and $\bar{s}_i = \ell$.

In the first case above, we have $\bar{r}_i \leq a_{i+1} - a_i$ and $\bar{s}_i - s_i \leq \delta\ell$, and similarly for the second case. Therefore, $\sum_{i\in R} \bar{r}_i \leq n + \delta\ell \cdot |R| = (1 + O(\delta))n$, with an absolute constant in the $O(\cdot)$ notation, and similarly $\sum_{i\in R} \bar{s}_i = (1 + O(\delta))n$. For $i \in S$ we find $\sum_{i\in S} \bar{r}_i, \sum_{i\in S} \bar{s}_i \leq |S| \cdot \ell \leq 2\delta n$. Altogether

$$\sum_{i=1}^{q} \bar{r}_i \leq (1 + O(\delta))n, \qquad \sum_{i=1}^{q} \bar{s}_i \leq (1 + O(\delta))n. \tag{6}$$

Now $\bar{r}_i$ and $\bar{s}_i$ already satisfy the requirements of Proposition 6(i), since we have $\bar{r}_i\bar{s}_i \geq \delta\ell^2 = \delta k^{2\alpha} > Ck$ and $(\bar{r}_i + \bar{s}_i)\sqrt{\bar{r}_i\bar{s}_i} \leq 2\ell^2 = 2k^{2\alpha} < \delta k^{3/2}/6$. We thus have, by Proposition 6,

$$\mathbf{P}[L(\Sigma(K\bar{r}_i, \bar{s}_i; k)) \geq e_i] \leq 2e^{-(e_i - m_u(\bar{r}_i, \bar{s}_i))^2/8e_i}$$

for all i such that $e_i \geq m_u(\bar{r}_i, \bar{s}_i)$, where $m_u(r,s) = (1+\delta)2\sqrt{rs/k}$. In the denominator of the exponent, we estimate $e_i \leq e_{\max}$. We thus have

$$p_T \leq \prod_{i=1}^{q} 2e^{-\max(0, e_i - m_u(\bar{r}_i, \bar{s}_i))^2/8e_{\max}}$$

(note that the factors for i with $e_i < m_u(\bar{r}_i, \bar{s}_i)$ equal 1). We consider the logarithm of p_T, we use the Cauchy–Schwarz inequality, and the inequality $\max(0,x) + \max(0,y) \geq \max(0, x+y)$:

$$\begin{aligned}
-\ln p_T &\geq \frac{1}{8e_{\max}} \sum_{i=1}^{q} \max(0, e_i - m_u(\bar{r}_i, \bar{s}_i))^2 - q\ln 2 \\
&\geq \frac{1}{8e_{\max}} \cdot \frac{1}{q} \cdot \left(\sum_{i=1}^{q} \max(0, e_i - m_u(\bar{r}_i, \bar{s}_i)) \right)^2 - q\ln 2 \\
&\geq \Omega(1) \cdot \frac{1}{e_{\max}} \cdot \frac{\ell}{n} \left(\sum_{i=1}^{q} e_i - \sum_{i=1}^{q} m_u(\bar{r}_i, \bar{s}_i) \right)^2 - q\ln 2 \\
&\geq \Omega\left(\frac{\delta\sqrt{k}}{n}\right) \left((1+\varepsilon)\frac{2n}{\sqrt{k}} - \frac{2(1+\delta)}{\sqrt{k}} \sum_{i=1}^{q} \sqrt{\bar{r}_i \bar{s}_i} \right)^2 - q\ln 2.
\end{aligned}$$

The function $(x,y) \mapsto \sqrt{xy}$ is subadditive: $\sqrt{xy} + \sqrt{x'y'} \leq \sqrt{(x+x')(y+y')}$. Thus, using (6), we have

$$\sum_{i=1}^{q} \sqrt{\bar{r}_i \bar{s}_i} \leq (1 + O(\delta))n\,,$$

and so, since $q = O(n/l)$ and $l \geq \sqrt{k}$,

$$-\ln p_T \geq \Omega\left(\frac{\delta\sqrt{k}}{n}\right) \left((1+\varepsilon)\frac{2n}{\sqrt{k}} - (1+O(\delta))\frac{2n}{\sqrt{k}} \right)^2 - q\ln 2 = \Omega\left(\varepsilon^2\delta \cdot \frac{n}{\sqrt{k}}\right).$$

Lemma 10 is proved. ■

Proof of Theorem 1. We have

$$\mathbf{P}[L(\Sigma(K_{n,n};k)) \geq m_{\max}] \leq \sum_{T \in \mathcal{T}} p_T \leq |\mathcal{T}| \cdot \max_T p_T\,.$$

The sought after estimate

$$\mathbf{P}[L(\Sigma(K_{n,n};k)) \geq m_{\max}] \leq \exp\left(-\Omega(\varepsilon^2\delta n/\sqrt{k}\,)\right),$$

follows from Lemmas 9 and 10. ■

7 Extensions

Similarly one can prove results for the Erdős model analogous to those obtained in previous sections (essentially, k is now replaced by $1/p$):

Theorem 11 *For every $\varepsilon > 0$ there exist constants $p_0 \in (0,1)$ and C such that for all $p < p_0$ and all n with $n\sqrt{p} > C$ we have*

$$(1-\varepsilon)\cdot 2n \cdot \sqrt{p} \leq \mathbf{E}\left[L(G(K_{n,n};p))\right] \leq (1+\varepsilon)\cdot 2n\cdot \sqrt{p}\,.$$

Moreover, there is an exponentially small tail bound; namely, for every $\varepsilon > 0$ there exists $c > 0$ such that for p and n as above,

$$\mathbf{P}[|L(G(K_{n,n};p)) - 2n\sqrt{p}| \geq \varepsilon 2n\sqrt{p}] \leq e^{-cn\sqrt{p}}\,.$$

Subadditivity arguments yield that $\mathbf{E}\left[L(G(K_{n,n};p))\right]/n$ converges to a constant Δ_p as $n \to \infty$. The previous theorem thus implies that $\Delta_p/\sqrt{p} \to 2$ as $p \to 0$.

Also, similar results hold for the $G(K_{r,s};p)$ model as those derived for $\Sigma(K_{r,s};k)$. Specifically,

Proposition 12 *For every $\delta > 0$, there exists a (large) positive constant C such that:*

(i) *If $rs \geq C/p$ and $(r+s)\sqrt{rs} \leq \delta/6p^{3/2}$, then with $m_u = m_u(r,s) = 2(1+\delta)\sqrt{rsp}$, we have*

$$\mathbf{P}[L(G(K_{r,s};p)) \geq m_u + t] \leq 2e^{-t^2/8(m_u+t)}$$

for all $t \geq 0$.

(ii) *If $rs \geq C/p$ and and $r+s \leq \delta/6p$, then with m_u as above and $m_l = m_l(r,s) = 2(1-\delta)\sqrt{rsp}$, we have*

$$\mathbf{P}[L(G(K_{r,s};p)) \leq m_l - t] \leq 2e^{-t^2/8m_u}$$

for all $t \geq 0$.

In [11], Johansson implicitly considers a model somewhat related to the $G(K_{n,n};p)$ model. Specifically, a distribution $G^*(K_{n,n};p)$ over weighted instances of $K_{n,n}$. The weight of each edge is a geometrically distributed random variable taking the value $k \in \mathbf{N}$ with probability $(1-p)^k p$, and the edge weights are mutually independent. Denoting the maximum weight planar matching of an instance drawn according to $G^*(K_{n,n};p)$ by $L(G^*(K_{n,n};p))$, Johansson's result [11, Theorem 1.1] says that for all $pi \in (0,1)$,

$$\lim_{n\to\infty} \frac{1}{n} \cdot \mathbf{E}\left[L(G^*(K_{n,n};p))\right] = \frac{(1+\sqrt{1-p})^2}{p}\,.$$

Note that an instance G of $G(K_{n,n};p)$ can be obtained from one drawn according to $G^*(K_{n,n};p)$ by including in G only those edges of $K_{n,n}$ with nonzero weight. Hence,

$$\mathbf{E}\left[L(G(K_{n,n};p))\right] \quad\leq\quad \mathbf{E}\left[L(G^*(K_{n,n};p))\right]\,.$$

It follows that $\Delta_p \leq (1+\sqrt{1-p})^2/p$, for all $p \in (0,1)$. We shall see below that known results imply a much stronger bound on Δ_p for not too large values of p.

Gravner, Tracy and Widom [8] consider processes associated to random $(0,1)$–matrices where each entry takes the value 1 with probability p, independent of the values of other matrix entries. In particular they study a process called *oriented digital boiling* (ODB) and analyze the behavior of a so called *height function* which equals, in distribution, the longest sequence (i_l, j_l) of positions in a random $(0,1)$–matrix of size $n \times n$ which have entry 1 such that the i_l's are increasing and the j_l's are nondecreasing. In contrast, $L(G(K_{n,n};p))$ equals in distribution the longest such sequence with both i_l's and j_l's increasing. This latter model is referred to as *strict oriented digital boiling* in [8], but no results are claimed for it. Clearly, an ODB process dominates that of a strict ODB process. Hence, [8, §3, (1)] implies that for any $p < 1/2$,

$$\Delta_p \leq \kappa_p := \lim_{n\to\infty} \frac{1}{n} \cdot \mathbf{E}\left[L(G(K_{n,n};p))\right] = 2\sqrt{p(1-p)}\,,$$

which in turn implies that $\limsup_{p\to 0} \Delta_p/\sqrt{p} \leq 2$. Nevertheless, our derivation of this latter limit value is elementary in comparison with the highly technical nature of [8].

Acknowledgments

We thank Ricardo Baeza for calling to our attention reference [17].

References

[1] D. Aldous and P. Diaconis. Longest increasing subsequences: From patience sorting to the Baik–Deift–Johansson theorem. *Bulletin of the AMS*, 36(4):413–432, 1999.

[2] R.M. Baer and P. Brock. Natural soritng over permutation spaces. *Mathematics of Computation*, pages 385–410, 1967.

[3] R. Baeza-Yates, G. Navarro, R. Gavaldá, and R. Schehing. Bounding the expected length of the longest common subsequences and forests. *Theory of Computing Systems*, 32(4):435–452, 1999.

[4] J. Baik, P. A. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutaions. *J. Amer. Math. Soc.*, 12:1119–1178, 1999.

[5] V. Chvátal and D. Sankoff. Longest common subsequences of two random sequences. *J. Appl. Prob*, 12:306–315, 1975.

[6] V. Dančík. *Expected Length of Longest Common Subsequences*. PhD thesis, Department of Computer Science, University of Warwick, September 1994.

[7] P. Erdős and G. K. Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[8] J. Gravner, C. A. Tracy, and H. Widom. Limit theorems for height fluctuations in a class of discrete space time growth models. *J. Stat. Phys.*, 102:1085–1132, 2001.

[9] J. M. Hammersley. A few seedlings of research. In *Proc. Sixth Berkeley Sympos. Math. Stat. Prob.*, pages 345–394, Berkeley, Calif., 1972. Univ. of California Press.

[10] S. Janson, T. Łuczak, and A. Ruci nski. *Random Graphs*. Wiley, 2000.

[11] K. Johannson. Shape fluctuations and random matrices. *Commun. Math. Phys.*, 209:437–476, 2000.

[12] J. F. C. Kingman. Subadditive ergodic theory. *The Annals of Probability*, 1(6):883–909, 1973.

[13] M. Kiwi and M. Loebl. Largest planar matching in random bipartite graphs. *Random Structures and Algorithms*, 21(2):162–181, 2002.

[14] B. F. Logan and L. A. Shepp. A variational problem or random Young tableaux. *Adv. in Math.*, 26:206–222, 1977.

[15] A. Okounkov. Random matrices and ramdom permutations. *International Mathematics Research Notices*, pages 1043–1095, 2000.

[16] P.A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach.* MIT Press, 2000.

[17] D. Sankoff and J. Kruskal, editors. *Common subsequences and monotone subsequences*, chapter 17, pages 363–365. Addison–Wesley, Reading, Mass., 1983.

[18] C. Schensted. Longest increasing and decreasing subsequences. *Canad. J. Math.*, 13:179–191, 1961.

[19] R. P. Stanley. Recent progress in algebraic combinatorics. *Bulletin of the AMS*, 40(1):55–68, 2002.

[20] S. Ulam. Monte carlo calculations in problems of mathematical physics. In *Modern Mathematics for the Engineers*, pages 261–281. McGraw-Hill, 1961.

[21] A. M. Vershik and S. V. Kerov. Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tableaux. *Dokl. Akad. Nauk SSSR*, 233:1024–1028, 1977.