CAMBRIDGE JOURNALS

# Proof Delivery Form

## Combinatorics, Probability and Computing

**Date of delivery:**

**Journal and vol/article ref:** CPC 990

**Number of pages (not including this page):** 16

This proof is sent to you on behalf of Cambridge University Press. Please print out the file and check the proofs carefully. Make any corrections necessary on a hardcopy and answer queries on each page of the proofs.

Please return the **marked proof** within **4** days of receipt to:


**Carol Miller - Journals**
**Cambridge University Press**
**The Edinburgh Building,**
**Shaftesbury Road**
**Cambridge CB2 8RU, UK**


To avoid delay from overseas, please send the proof by airmail or courier.

If you have no corrections to make, please email **cmiller@cambridge.org** to save having to return your paper proof. If corrections are light, you can also send them by email, quoting both page and line number.

- You are responsible for correcting your proofs. Errors not found may appear in the published journal.
- The proof is sent to you for correction of typographical errors only. Revision of the substance of the text is not permitted, unless discussed with the editor of the journal.
- Please answer carefully any queries listed overleaf.
- A new copy of a figure must be provided if correction of anything other than a typographical error introduced by the typesetter is required.
- If you have problems with the file please contact cmiller@cambridge.org

Please note that this pdf is for proof checking purposes only. It should not be distributed to third parties and may not represent the final published version.

**Important:** you must return any forms included with your proof.

**Please do not reply to this email**

Please refer to our FAQs at
http://journals.cambridge.org/production_faqs

**Author queries:**

**Typesetter queries:**

**Non-printed material:**

transfer of copyright

CAMBRIDGE
UNIVERSITY PRESS

Please read the notes overleaf and then complete, sign, and return this form to **Journals Publishing, Cambridge University Press, The Edinburgh Building, Shaftesbury Road, Cambridge, CB2 8RU, UK** as soon as possible.

# COMBINATORICS, PROBABILITY AND COMPUTING

In consideration of the publication in **COMBINATORICS, PROBABILITY AND COMPUTING**

of the contribution entitled: ......................................................................................................................

......................................................................................................................................................

by (all authors' names): ..........................................................................................................................

......................................................................................................................................................

## 1 To be filled in if copyright belongs to you

**Transfer of copyright**

I/we hereby assign to Cambridge University Press, full copyright in all formats and media in the said contribution.

I/we warrant that I am/we are the sole owner or co-owners of the material and have full power to make this agreement, and that the material does not contain any libellous matter or infringe any existing copyright.

I/we further warrant that permission has been obtained from the copyright holder for any material not in my/our copyright including any audio and video material, that the appropriate acknowledgement has been made to the original source, and that in the case of audio or video material appropriate releases have been obtained from persons whose voices or likenesses are represented therein. I/we attach copies of all permission and release correspondence.

I/we hereby assert my/our moral rights in accordance with the UK Copyrights Designs and Patents Act (1988).

Signed (tick one) ☐ the sole author(s)

☐ one author authorised to execute this transfer on behalf of all the authors of the above article

Name (block letters) ..............................................................................................................................

Institution/Company ..............................................................................................................................

Signature: ......................................................... Date: ..........................................................................

*(Additional authors should provide this information on a separate sheet.)*

## 2 To be filled in if copyright does not belong to you

a Name and address of copyright holder ..................................................................................................

......................................................................................................................................................

......................................................................................................................................................

......................................................................................................................................................

b The copyright holder hereby grants to Cambridge University Press the non-exclusive right to publish the contribution in the journal and to deal with requests from third parties in the manner specified in paragraphs 4 and 5 overleaf.

(Signature of copyright holder or authorised agent) ..............................................................................

## 3 US Government exemption

I/we certify that the paper above was written in the course of employment by the United States Government so that no copyright exists.

Signature: ......................................................... Name (Block letters): .............................................................

## 4 Requests received by Cambridge University Press for permission to reprint this article should be sent to

**(see para. 4 overleaf)**

Name and address (block letters) ..........................................................................................................

......................................................................................................................................................

# Notes for contributors

1 The Journal's policy is to acquire copyright in all contributions. There are two reasons for this: *(a)* ownership of copyright by one central organisation tends to ensure maximum international protection against unauthorised use; *(b)* it also ensures that requests by third parties to reprint or reproduce a contribution, or part of it, are handled efficiently and in accordance with a general policy that is sensitive both to any relevant changes in international copyright legislation and to the general desirability of encouraging the dissemination of knowledge.

2 Two 'moral rights' were conferred on authors by the UK Copyright Act in 1988. In the UK an author's 'right of paternity', the right to be properly credited whenever the work is published (or performed or broadcast), requires that this right is asserted in writing.

3 Notwithstanding the assignment of copyright in their contribution, all contributors retain the following **non-transferable** rights:

- The right to post *either* their own version of their contribution as submitted to the journal (prior to revision arising from peer review and prior to editorial input by Cambridge University Press) *or* their own final version of their contribution as accepted for publication (subsequent to revision arising from peer review but still prior to editorial input by Cambridge University Press) on their **personal or departmental web page**, or in the **Institutional Repository** of the institution in which they worked at the time the paper was first submitted, or (for appropriate journals) in PubMedCentral, provided the posting is accompanied by a prominent statement that the paper has been accepted for publication and will appear in a revised form, subsequent to peer review and/or editorial input by Cambridge University Press, in **Combinatorics, Probability and Computing** published by Cambridge University Press, together with a copyright notice in the name of the copyright holder (Cambridge University Press or the sponsoring Society, as appropriate). On publication the full bibliographical details of the paper (volume: issue number (date), page numbers) must be inserted after the journal title, along with a link to the Cambridge website address for the journal. Inclusion of this version of the paper in Institutional Repositories outside of the institution in which the contributor worked at the time the paper was first submitted will be subject to the additional permission of Cambridge University Press (not to be unreasonably withheld).
- The right to post the definitive version of the contribution as published at Cambridge Journals Online (in PDF or HTML form) on their **personal or departmental web page**, no sooner than upon its appearance at Cambridge Journals Online, subject to file availability and provided the posting includes a prominent statement of the full bibliographical details, a copyright notice in the name of the copyright holder (Cambridge University Press or the sponsoring Society, as appropriate), and a link to the online edition of the journal at Cambridge Journals Online.
- The right to post the definitive version of the contribution as published at Cambridge Journals Online (in PDF or HTML form) in the **Institutional Repository** of the institution in which they worked at the time the paper was first submitted, or (for appropriate journals) in PubMedCentral, no sooner than **one year** after first publication of the paper in the journal, subject to file availability and provided the posting includes a prominent statement of the full bibliographical details, a copyright notice in the name of the copyright holder (Cambridge University Press or the sponsoring Society, as appropriate), and a link to the online edition of the journal at Cambridge Journals Online. Inclusion of this definitive version after one year in Institutional Repositories outside of the institution in which the contributor worked at the time the paper was first submitted will be subject to the additional permission of Cambridge University Press (not to be unreasonably withheld).
- The right to make hard copies of the contribution or an adapted version for their own purposes, including the right to make multiple copies for course use by their students, provided no sale is involved.
- The right to reproduce the paper or an adapted version of it in any volume of which they are editor or author. Permission will automatically be given to the publisher of such a volume, subject to normal acknowledgement.

4 We shall use our best endeavours to ensure that any direct request we receive to reproduce your contribution, or a substantial part of it, in another publication (which may be an electronic publication) is approved by you before permission is given.

5 Cambridge University Press co-operates in various licensing schemes that allow material to be photocopied within agreed restraints (e.g. the CCC in the USA and the CLA in the UK). Any proceeds received from such licenses, together with any proceeds from sales of subsidiary rights in the Journal, directly support its continuing publication.

6 It is understood that in some cases copyright will be held by the contributor's employer. If so, Cambridge University Press requires non-exclusive permission to deal with requests from third parties, on the understanding that any requests it receives from third parties will be handled in accordance with paragraphs 4 and 5 above (note that your approval and not that of your employer will be sought for the proposed use).

7 Permission to include material not in your copyright
If your contribution includes textual or illustrative material not in your copyright and not covered by fair use / fair dealing, permission must be obtained from the relevant copyright owner (usually the publisher or via the publisher) for the non-exclusive right to reproduce the material worldwide in all forms and media, including electronic publication. The relevant permission correspondence should be attached to this form.

If you are in doubt about whether or not permission is required, please consult the Permissions Controller, Cambridge University Press, The Edinburgh Building, Shaftesbury Road, Cambridge CB2 8RU, UK. Fax: +44 (0)1223 315052.
Email: lnicol@cambridge.org.

The information provided on this form will be held in perpetuity for record purposes. The name(s) and address(es) of the author(s) of the contribution may be reproduced in the journal and provided to print and online indexing and abstracting services and bibliographic databases

*Please make a duplicate of this form for your own records*

*Combinatorics, Probability and Computing* (2009) 00, 1–16. 
doi:10.1017/S0963548309009900 Printed in the United Kingdom

# On a Speculated Relation Between Chvátal–Sankoff Constants of Several Sequences

M. KIWI[1†] and J. SOTO[2‡]

[1]Departamento de Ingeniería Matemática, Centro de Modelamiento Matemático
(UMI 2807, CNRS), University of Chile
(e-mail: mkiwi@dim.uchile.cl)

[2]Department of Mathematics, MIT, Cambridge, MA 02139, USA
(e-mail: jsoto@math.mit.edu)

*Received 29 February 2008; revised 1 September 2008*

It is well known that, when normalized by $n$, the expected length of a longest common subsequence of $d$ sequences of length $n$ over an alphabet of size $\sigma$ converges to a constant $\gamma_{\sigma,d}$. We disprove a speculation by Steele regarding a possible relation between $\gamma_{2,d}$ and $\gamma_{2,2}$. In order to do that we also obtain some new lower bounds for $\gamma_{\sigma,d}$, when both $\sigma$ and $d$ are small integers.

## 1. Introduction

String matching is one of the most intensively analysed problems in computer science. Among string matching problems the longest common subsequence problem (LCS) stands out. This problem consists of finding the longest subsequence common to all strings in a set of sequences (often just two). The LCS problem is the basis of Unix's `diff` command, has applications in bioinformatics, and also arises naturally in remarkably distinct domains such as cryptographic snooping, the mathematical analysis of bird songs, and comparative genomics. In addition, the LCS problem offers a concrete basis for the illustration and benchmarking of mathematical methods and tools such as subadditive methods and martingale inequalities; see, for example, Steele's monograph [15].

Although the LCS problem has been studied under many different contexts there are several issues concerning it that are still unresolved. The most prominent of the outstanding questions relating to the LCS problem concerns the length $L_{n,\sigma,d}$ of an LCS of $d$ sequences of $n$ characters chosen uniformly and independently over some alphabet of size $\sigma$. Subadditivity arguments yield

† Gratefully acknowledges the support of CONICYT via FONDAP in Applied Mathematics and Anillo en Redes ACT08.
‡ Gratefully acknowledges the support of CONICYT via Anillo en Redes ACT08.

that for fixed $d$ and $n$ going to infinity, the expected value of $L_{n,\sigma,d}$ normalized by $n$ converges to a constant $\gamma_{\sigma,d}$. For $d, \sigma \geqslant 2$, the precise value of $\gamma_{\sigma,d}$ is unknown. The constant $\gamma_{2,2}$ is referred to as the Chvátal–Sankoff constant. The calculation of its exact value is an over three-decade-old open problem. The determination of its value has received a fair amount of attention, starting with the work of Chvátal and Sankoff [4], encompassing among others [1, 2, 6, 7, 8, 12], and is explicitly stated in several well-known texts such as those by Waterman [19, § 11.1.3], Steele [16, p. 3], Pevzner [13, p. 107], and Szpankowski [17, p. 109]. To the best of our knowledge the current sharpest bounds on $\gamma_{2,2}$ are due to Lueker [12], who established that $0.788071 \leqslant \gamma_{2,2} \leqslant 0.826280$.

The starting point for this investigation is the following comment by Steele [15]:

> It would be of interest to relate $c_3$ to $c^2$, and one is tempted to speculate that $c_3 = c^2$ (and more generally that $c_k = c^{k-1}$). Computational evidence does not yet rule this out.

Here, Steele uses $c$ to denote the limiting value of the longest common subsequence of two random sequences of length $n$ normalized by $n$ as $n$ goes to infinity, and in general, he uses $c_k$ to denote the analogous constant for $k$ sequences. However, it is unclear if in this comment he uses $c$ and $c_k$ to denote the constants $\gamma_{2,2}$ and $\gamma_{k,2}$ (*i.e.*, specifically for the case of alphabet size 2) or if he is generically denoting the constants for arbitrary alphabet size. Dančík [6] cites the previous statement as a conjecture by Steele using the second interpretation, *i.e.*, as the claim that, for all $d \geqslant 3$ and $\sigma \geqslant 2$,

$$\gamma_{\sigma,d} = \gamma_{\sigma,2}^{d-1}. \tag{1.1}$$

Dančík [6, Theorem 2.1, Corollary 2.1] shows that for $d \geqslant 2$

$$1 \leqslant \liminf_{\sigma\to\infty} \sigma^{1-1/d}\gamma_{\sigma,d} \leqslant \limsup_{\sigma\to\infty} \sigma^{1-1/d}\gamma_{\sigma,d} \leqslant e.$$

Hence, if (1.1) were true, then for $\epsilon > 0$ and $\sigma$ sufficiently large,

$$1-\epsilon \leqslant \sigma^{1-1/d}\gamma_{\sigma,d} = \sigma^{1-1/d}\gamma_{\sigma,2}^{d-1} \leqslant \sigma^{1-1/d}\left(\frac{e(1+\epsilon)}{\sqrt{\sigma}}\right)^{d-1}.$$

Dančík's results disprove (1.1) by observing that for $d > 2$ one may choose $\sigma$ large enough so as to make the rightmost term of the last displayed equation arbitrarily close to 0.

If we use the first interpretation of Steele's speculation quoted above, *i.e.*, considering only the case of binary alphabets as we believe it was intended, then (1.1) is not invalidated by Dančík's work.

In [15], Steele does not justify his speculation. The following non-rigorous argument gives some indication that one should expect that $\gamma_{2,3}$ is strictly bigger than $\gamma_{2,2}^2$. Indeed, let $A_1$, $A_2$ and $A_3$ be three independently and uniformly chosen binary sequences of length $n$. For $i \neq j$ and very large values of $n$ one knows that a longest common subsequence $\ell_{i,j}$ of sequences $A_i$ and $A_j$ would be of length approximately $\gamma_{2,2}n$. One would expect (although we can not prove it) that $\ell_{i,j}$ would behave like a uniformly chosen binary string of length $\gamma_{2,2}n$. Sequences $\ell_{1,2}$ and $\ell_{2,3}$ are clearly correlated. However, one might guess that the correlation is weak (again, we can certainly neither formalize nor prove such a statement). The previously stated discussion suggests that a longest common subsequence $\ell_{1,2,3}$ of $\ell_{1,2}$ and $\ell_{2,3}$ should be of length approximately $\gamma_{2,2}^2 n$. Since $\ell_{1,2,3}$ is clearly a longest common subsequence of $A_1$, $A_2$ and $A_3$, one is led to conclude that

$$\gamma_{2,3} \geqslant \gamma_{2,2}^2. \tag{1.2}$$

However, there are two good reasons to suspect that this last inequality should be strict.

- Since $\ell_{2,3}$ has only a fraction of $A_3$'s length, one expects that a longest common subsequence of $\ell_{1,2}$ and $A_3$ is significantly larger than a longest common subsequence of $\ell_{1,2}$ and $\ell_{2,3}$.
- The longest common subsequence of $A_1$, $A_2$ and $A_3$ might arise by taking a longest common subsequence on sub-optimal common subsequences $\ell'_{1,2}$ and $\ell'_{2,3}$ of $A_1$ and $A_2$, and $A_2$ and $A_3$, respectively.

This work's main contribution is to show that the inequality in (1.2) is indeed strict.

In Section 2 we give a simple argument that proves that when $\sigma$ is fixed and $d$ is large the identity $\gamma_{\sigma,d} = \gamma_{\sigma,2}^{d-1}$ does not hold. The underlying argument is essentially an application of the probabilistic method. However, it might still be possible for the relation to hold for some specific values of $\sigma$ and $d$. Of particular interest is the case of binary sequences, *i.e.*, $\sigma = 2$. In Section 3 we show that even this weaker identity does not hold, *i.e.*, that $\gamma_{2,3} \neq \gamma_{2,2}^2$. To achieve this goal, we rely on Lueker's [12] $U = 0.826280$ upper bound on $\gamma_{2,2}$ and determine a lower bound on $\gamma_{2,3}$ which is strictly larger than $U^2 \geqslant \gamma_{2,2}^2$. The lower bound on $\gamma_{2,3}$ is obtained by an approach similar to that used by Lueker [12] to lower-bound $\gamma_{2,2}$, although in our case we have to consider a non-binary alphabet. Aside from the extra notation needed to handle the cases $\sigma, d > 2$, our treatment is a straightforward generalization of the approach used by Lueker. (In fact, in order to keep the exposition as clear as possible we do not even use the optimization tweaks implemented by Lueker in order to take advantage of the symmetries inherent to the problem and objects that arise in its analysis.) We conclude with some final comments in Section 4.

## 2. Disproving $\gamma_{\sigma,d} = \gamma_{\sigma,2}^{d-1}$ for large $d$

We start this section by introducing some notation. Given strings $A_1, \ldots, A_d$ of length $n$, we denote by $L(A_1, \ldots, A_d)$ the length of the longest common subsequence of all $A_i$s. Let $\mathcal{U}_{n,\sigma}$ be the distribution of sequences of length $n$ whose characters are chosen uniformly and independently from $\Sigma = \{1, \ldots, \sigma\}$. We denote by $L_{n,\sigma,d}$ the random variable $L(A_1, \ldots, A_d)$ when all the $A_i$ are chosen according to $\mathcal{U}_{n,\sigma}$. Finally, we let $\gamma_{\sigma,d}$ denote the limit of $\mathbf{E}L_{n,\sigma,d}/n$ when $n \to \infty$ (the existence of this limit follows from standard subadditivity arguments [4]).

In what follows, we give a lower bound for $\gamma_{\sigma,d}$ that is independent of $d$. This bound is based on the following simple fact. If $X$ is chosen according to $\mathcal{U}_{n,\sigma}$ and $n$ is large, then the number of occurrences of a fixed character in $\Sigma$ is roughly $n/\sigma$. Intuitively, this means that for a set of $d$ random strings of (very large) length $n$, with very high probability a sequence formed by roughly $\lfloor n/\sigma \rfloor$ equal characters will be a common subsequence of all the $d$ random strings.

**Lemma 2.1.** *For all $d$ and $\sigma$, we have $\gamma_{\sigma,d} \geqslant 1/\sigma$.*

**Proof.** Let $A_1, \ldots, A_d$ be $d$ independent random strings chosen according to $\mathcal{U}_{n,\sigma}$. Let $X_i$ denote the number of times the character $c \in \Sigma$ appears in $A_i$, and $X = \min\{X_1, \ldots, X_d\}$. The string $c^X$ formed by $X$ copies of the character $c$ is a common subsequence of all $X_i$s. It follows that $L(A_1, \ldots, A_d) \geqslant X$.

Each $X_i$ is a binomial variable with parameter $p = 1/\sigma$. By a standard Chernoff bound [9, Remark 2.5] we have that for any $0 < \varepsilon < 1$,

$$\Pr[X_i \leqslant (1-\varepsilon)np] \leqslant \exp(-2n(p\varepsilon)^2).$$

Applying Markov's inequality, and recalling that the $X_i$s are independent, it follows that

$$\mathbf{E}X \geqslant (1-\varepsilon)np \Pr[X \geqslant (1-\varepsilon)np] \geqslant (1-\varepsilon)np[1-\exp(-2n(p\varepsilon)^2)]^d.$$

Letting $n$ be sufficiently large that $[1-\exp(-2n(p\varepsilon)^2)]^d \geqslant (1-2\varepsilon)/(1-\varepsilon)$, we obtain $\mathbf{E}X \geqslant np(1-2\varepsilon)$. Therefore,

$$\frac{\mathbf{E}L_{n,\sigma,d}}{n} = \frac{\mathbf{E}L(A_1,\ldots,A_d)}{n} \geqslant \frac{\mathbf{E}X}{n} \geqslant (1-2\varepsilon)p = \frac{1-2\varepsilon}{\sigma}.$$

It follows that $\gamma_{\sigma,d} \geqslant (1-2\varepsilon)/\sigma$. Since this is true for any $\varepsilon > 0$, we conclude that $\gamma_{\sigma,d} \geqslant 1/\sigma$. □

It is now easy to disprove that $\gamma_{\sigma,d} = \gamma_{\sigma,2}^{d-1}$ for large $d$. Indeed, note that since $\gamma_{\sigma,2} < 1$ [4], then $\lim_{d\to\infty} \gamma_{\sigma,2}^{d-1} = 0$. On the other hand, the previous lemma asserts that $\gamma_{\sigma,d} \geqslant 1/\sigma$ for all $d$, and hence for $d$ large enough, $\gamma_{\sigma,2}^{d-1} < \gamma_{\sigma,d}$.

In particular, for the case $\sigma = 2$, Lueker [12] proved that $\gamma_{2,2} \leqslant U$ for $U = 0.826280$. Thus, for all $d \geqslant 5$, we have the strict inequality

$$\gamma_{2,2}^{d-1} \leqslant (0.826280)^{d-1} < 1/2 \leqslant \gamma_{2,d}.$$

## 3. Disproving $\gamma_{2,3} = \gamma_{2,2}^2$

### 3.1. Diagonal common subsequence

As already mentioned, the best-known provable lower bound for $\gamma_{2,2}$ found so far is due to Lueker [12]. The starting point of Lueker's lower bound technique is a result by Alexander [1], who related the expected length of the LCS of two random strings of the same length $n$, to the expected length of the LCS of two random strings whose lengths sum up to $2n$. Below, we establish an analogue of Alexander's result but for the case of $d$ randomly chosen sequences.

Let $C[j..k]$ denote the substring $C[j]C[j+1]\cdots C[k]$ formed by all the characters between the $j$th and $k$th positions of $C$. Given strings $A_1,\ldots,A_d$ of length at least $n$, we say that $B$ is an $n$-diagonal common subsequence of $A_1,\ldots,A_d$ if $B$ is a common subsequence of a set of prefixes of $A_1,\ldots,A_d$ whose lengths sum to $n$, *i.e.*, if for some indices $i_1,\ldots,i_d$ such that $i_1+\cdots+i_d = n$, the string $B$ is a common subsequence of $A_1[1..i_1], A_2[1..i_2],\ldots,A_d[1..i_d]$.

Let $D_n(A_1,\ldots,A_d)$ denote the length of a longest $n$-diagonal common subsequence of the strings $A_1,\ldots,A_d$. We denote by $D_{n,\sigma,d}$ the random variable $D_n(A_1,\ldots,A_d)$ where the strings $A_1,\ldots,A_d$ are chosen according to $\mathcal{U}_{n,\sigma}$.

The main objective of this section is to prove the following extension of a result of Alexander [1, Proposition 2.4] for the $d = 2$ case.

**Theorem 3.1.** *For all $n \geqslant d$,*

$$d\cdot\mathbf{E}D_{n,\sigma,d} - d^{3/2}\sqrt{2n\ln n} \leqslant \mathbf{E}L_{n,\sigma,d} \leqslant \mathbf{E}D_{nd,\sigma,d}.$$

*In particular, for all $\sigma$ there exists $\delta_{\sigma,d}$ such that*

$$\delta_{\sigma,d} = \lim_{n\to\infty} \frac{\mathbf{E}D_{n,\sigma,d}}{n} = \frac{\gamma_{\sigma,d}}{d}.$$

For the sake of clarity of exposition, before proving Theorem 3.1 we establish some intermediate results.

**Lemma 3.2.** *For all $n$ and $d$, $\mathbf{E}L_{n,\sigma,d} \leqslant \mathbf{E}D_{nd,\sigma,d}$.*

**Proof.** Let $A_1,\dots,A_d$ be random strings independently chosen according to $\mathcal{U}_{nd,\sigma}$. Since a longest common subsequence of $A_1[1..n],\dots,A_d[1..n]$ is also an $nd$-diagonal common subsequence of $A_1,\dots,A_d$,

$$L(A_1[1..n],\dots,A_d[1..n]) \leqslant D_{nd}(A_1,\dots,A_d).$$

Taking expectation on both sides of the previous inequality yields the desired conclusion. □

**Lemma 3.3.** *For all $n \geqslant d$,*

$$d\cdot \mathbf{E}D_{n,\sigma,d} - d^{3/2}\sqrt{2n\ln n} \leqslant \mathbf{E}L_{n,\sigma,d}.$$

**Proof.** Let $A_1,\dots,A_d$ be a list of words of length $n$. Note that if we change one character of any word in the list, then the values $L(A_1,\dots,A_d)$ and $D_n(A_1,\dots,A_d)$ will change by at most one unit. It follows that the random variables $L_{n,\sigma,d}$ and $D_{n,\sigma,d}$ (seen as functions from $(\Sigma^n)^d$ to $\mathbb{R}$) are both 1-Lipschitz. Applying Azuma's inequality (as treated in, for example, [9, § 2.4]), we get

$$\Pr\left[D_{n,\sigma,d} \leqslant \mathbf{E}D_{n,\sigma,d} - \sqrt{n/2}\right] \leqslant \exp\left(-\frac{2(n/2)}{nd}\right) = e^{-1/d} < \frac{d}{d+1},$$

where the last inequality holds since $e^{-x} < 1/(x+1)$ for all $x > 0$.

Let $\lambda = \mathbf{E}D_{n,\sigma,d} - \sqrt{n/2}$. Since $D_{n,\sigma,d} > \lambda$ implies that there are positive indices $i_1,\dots,i_d$ such that $i_1+\cdots+i_d = n$ and $L(A_1[1..i_1],\dots,A_d[1..i_d]) \geqslant \lambda$,

$$\Pr[D_{n,\sigma,d} > \lambda] \leqslant \sum_{\substack{0<i_1,\dots,i_d<n,\\ i_1+\cdots+i_d=n}} \Pr[L(A_1[1..i_1],\dots,A_d[1..i_d]) > \lambda].$$

Let $I$ be the number of summands on the right-hand side. Note that $I = \binom{n-1}{d-1}$ since it counts the number of ways of partitioning $n$ into $d$ positive summands. It follows that there exist positive $j_1,\dots,j_d$ summing to $n$ such that

$$\Pr\left[L(A_1[1..j_1],\dots,A_d[1..j_d]) > \lambda\right] > \frac{1}{I}\left(1-\frac{d}{d+1}\right) = \frac{1}{I(d+1)}.$$

Note that the distribution of the random variable $L(A_1[1..j_1],\dots,A_d[1..j_d])$ is the same as the distribution of $L(A_1[1..j_{\tau(1)}],\dots,A_d[1..j_{\tau(d)}])$ for any permutation $\tau : [d] \to [d]$. It is also easy to see that the distribution of $L(A_1[a_1..b_1],\dots,A_d[a_d..b_d])$ and $L(A_1[a'_1..b'_1],\dots,A_d[a'_d..b'_d])$ is the same when $b_m - a_m = b'_m - a'_m$ for all $1 \leqslant m \leqslant d$.

Now, let $\tau$ be the cyclic permutation $(12\cdots d)$, and for $0 \leqslant m \leqslant d-1$ let $\mathcal{E}_m$ denote the event

$$L\left(A_1\left[\sum_{l=0}^{m-1} j_{\tau^l(1)}+1\,..\sum_{l=0}^{m} j_{\tau^l(1)}\right],\ldots,A_d\left[\sum_{l=0}^{m-1} j_{\tau^l(d)}+1\,..\sum_{l=0}^{m} j_{\tau^l(d)}\right]\right) > \lambda.$$

In particular, $\mathcal{E}_0$ is the event $\{L(A_1[1..j_1],\ldots,A_d[1..j_d]) > \lambda\}$ whose probability was bounded above. Note that the events $\mathcal{E}_0,\ldots,\mathcal{E}_{d-1}$ are equiprobable. Since each of the $\mathcal{E}_m$s depends on a different set of characters, they are independent. Moreover, if $\mathcal{E}_0,\ldots,\mathcal{E}_{d-1}$ simultaneously occur, then by concatenating the common subsequences of each block of characters we get that $L(A_1,\ldots,A_d) > d\lambda$. Hence,

$$\left(\frac{1}{I(d+1)}\right)^d < \prod_{m=0}^{d-1}\Pr[\mathcal{E}_m] = \Pr[\mathcal{E}_0,\mathcal{E}_1,\ldots,\mathcal{E}_{d-1}] \leqslant \Pr[L_{n,\sigma,d} > d\lambda]. \tag{3.1}$$

Applying Azuma's inequality again, we have

$$\Pr\left[L_{n,\sigma,d} \geqslant \mathbf{E}L_{n,\sigma,d}+\sqrt{\frac{nd^2\ln(I(d+1))}{2}}\right] \leqslant \left(\frac{1}{I(d+1)}\right)^d. \tag{3.2}$$

Combining (3.1) and (3.2) and recalling that $\lambda = \mathbf{E}D_{n,\sigma,d}-\sqrt{n/2}$, we obtain

$$\Pr\left[L_{n,\sigma,d} \geqslant \mathbf{E}L_{n,\sigma,d}+\sqrt{\frac{nd^2\ln(I(d+1))}{2}}\right] < \Pr\left[L_{n,\sigma,d} > d\mathbf{E}D_{n,\sigma,d}-d\sqrt{\frac{n}{2}}\right].$$

Hence,

$$\mathbf{E}L_{n,\sigma,d}+\sqrt{\frac{nd^2\ln(I(d+1))}{2}} \geqslant d\mathbf{E}D_{n,\sigma,d}-d\sqrt{\frac{n}{2}}.$$

Since $2 \leqslant d \leqslant n$, $(d+1)I = (d+1)\binom{n-1}{d-1} \leqslant n^d$, and so

$$d\mathbf{E}D_{n,\sigma,d} \leqslant \mathbf{E}L_{n,\sigma,d}+d\sqrt{\frac{n}{2}}+\sqrt{\frac{nd^2\ln(I(d+1))}{2}} \leqslant \mathbf{E}L_{n,\sigma,d}+d^{3/2}\sqrt{2n\ln(n)}. \qquad \square$$

**Proof of Theorem 3.1.** Lemmas 3.2 and 3.3 already give the bounds on $\mathbf{E}L_{n,\sigma,d}$. To complete the proof we need to show that $\lim_{n\to\infty}\mathbf{E}D_{n,\sigma,d}/n$ exists and that its value is $\gamma_{\sigma,d}/d$. By Lemmas 3.2 and 3.3 we have

$$\mathbf{E}L_{n,\sigma,d} \leqslant \mathbf{E}D_{nd,\sigma,d} \leqslant \frac{1}{d}\mathbf{E}L_{nd,\sigma,d}+d^{1/2}\sqrt{2nd\ln(nd)}.$$

Dividing by $n$, it follows that $\lim_{n\to\infty}\mathbf{E}D_{nd,\sigma,d}/n = \gamma_{\sigma,d}$. Furthermore, $\mathbf{E}D_{n,\sigma,d}$ is non decreasing in $n$, so

$$\frac{\lfloor n/d\rfloor}{n/d}\cdot\frac{\mathbf{E}D_{d\lfloor n/d\rfloor,\sigma,d}}{\lfloor n/d\rfloor} \leqslant \frac{\mathbf{E}D_{n,\sigma,d}}{n/d} \leqslant \frac{\lceil n/d\rceil}{n/d}\cdot\frac{\mathbf{E}D_{d\lceil n/d\rceil,\sigma,d}}{\lceil n/d\rceil}.$$

Since both the left-hand side and right-hand side terms above converge to $\gamma_{\sigma,d}$ when $n\to\infty$, the middle term also converges to that value, and so $\lim_{n\to\infty}\mathbf{E}D_{n,\sigma,d}/n = \gamma_{\sigma,d}/d$ as claimed. $\square$

### 3.2. Longest common subsequence of two words over a binary alphabet

In this section we describe Lueker's [12] approach for finding a lower bound on $\gamma_{d,\sigma}$ when $d = \sigma = 2$. Later on, we will generalize Lueker's technique to the cases of arbitrary $d$ and $\sigma$.

Let $X_1$ and $X_2$ be two random sequences chosen from $\mathcal{U}_{n,2}$, *i.e.*, strings of length $n$ such that all their characters are chosen uniformly and independently from the binary alphabet $\{0,1\}$. Lueker defines, for any two strings $A$ and $B$ over the binary alphabet, the quantity

$$W_n(A,B) = \mathbf{E}\Big[\max_{i+j=n} L(AX_1[1..i], BX_2[1..j])\Big].$$

Informally, $W_n(A,B)$ represents the expected length of an LCS of two strings with prefixes $A$ and $B$, respectively, and suffixes formed by uniformly and independently choosing $n$ characters in $\{0,1\}$. It is easy to see that $W_n(A,B)$ behaves as $D_{n,2,2}$ as $n \to \infty$. Hence, applying Alexander's $d = 2$ version of Theorem 3.1, Lueker observes that, for all $A, B \in \{0,1\}^*$,

$$\gamma_{2,2} = \lim_{n\to\infty} \frac{W_{2n}(A,B)}{n}.$$

A natural idea is to approximate $\gamma_{2,2}$ by $W_{2n}(A,B)/n$. Fix the length $l \in \mathbb{N}$ of the strings $A$ and $B$ and denote by $w_n$ the $2^{2l}$-dimensional vector whose coordinates correspond to the values $W_n(A,B)$ when $A$ and $B$ vary over all binary sequences of length $l$. For example, when $l = 2$, the vector $w_n$ has the following form:

$$w_n = \begin{pmatrix} w_n[00,00] \\ w_n[00,01] \\ \vdots \\ w_n[11,10] \\ w_n[11,11] \end{pmatrix} = \begin{pmatrix} W_n(00,00) \\ W_n(00,01) \\ \vdots \\ W_n(11,10) \\ W_n(11,11) \end{pmatrix}.$$

Lueker established a lower bound for each component of $w_n$ as a function of the components of $w_{n-1}$ and $w_{n-2}$. To reproduce that lower bound, we need to introduce some more notation. If $A = A[1]A[2]\cdots A[l]$ is a sequence of length $l \geqslant 2$, let $h(A)$ denote the *head* of $A$, *i.e.*, its first character, and let $T(A)$ denote its *tail*, *i.e.*, the substring obtained from $A$ by removing its head. In other words, $h(A) = A[1]$ and $T(A) = A[2..l]$. It is easy to see that the following relations among $w_n$,$w_{n-1}$ and $w_{n-2}$ hold.

- If $h(A) = h(B)$, then

$$w_n[A,B] \geqslant 1 + \frac{1}{4} \sum_{(c,c')\in\{0,1\}^2} w_{n-2}[T(A)c, T(B)c'].$$

- If $h(A) \neq h(B)$, then

$$w_n[A,B] \geqslant \frac{1}{2}\max\Big\{\sum_{c\in\{0,1\}} w_{n-1}[T(A)c, B], \sum_{c\in\{0,1\}} w_{n-1}[A, T(B)c]\Big\}.$$

Using the previous inequalities one can define a function $F : \mathbb{R}^{2^{2l}} \times \mathbb{R}^{2^{2l}} \to \mathbb{R}^{2^{2l}}$ such that for all $n \geqslant 2$, $w_n \geqslant F(w_{n-1}, w_{n-2})$. Furthermore, the function $F$ can be decomposed in two simpler functions $F_=$ and $F_{\neq}$ such that if $\Pi_=$ and $\Pi_{\neq}$ are the projections of the vectors onto the coordinates corresponding to the pairs of words with the same and different heads respectively, then

$$\Pi_=(w_n) \geqslant F_=(w_{n-2}) \quad \text{and} \quad \Pi_{\neq}(w_n) \geqslant F_{\neq}(w_{n-1}).$$

It might be useful to see some examples of these transformations. For instance, to obtain a lower bound of $w_n[001,011]$, one considers

$$\begin{aligned} w_n[001,011] &\geqslant F_=(w_{n-2})[001,011] \\ &= 1 + \frac{1}{4}(w_{n-2}[010,110] + w_{n-2}[010,111] + w_{n-2}[011,110] + w_{n-2}[011,111]). \end{aligned}$$

And to bound $w_n[001,111]$,

$$\begin{aligned} w_n[001,111] &\geqslant F_{\neq}(w_{n-1})[001,111] \\ &= \frac{1}{2}\max\{w_{n-1}[010,111] + w_{n-1}[011,111], w_{n-1}[001,110] + w_{n-1}[001,111]\}. \end{aligned}$$

### 3.3. Longest common subsequence of $d$ words over general alphabets

In this section we extend Lueker's lower bound arguments as described in the previous section to the general case of $d$ strings whose characters are uniformly and independently chosen over an alphabet of size $\sigma$.

Let $X_1,\ldots,X_d$ be a collection of $d$ independent random strings chosen according to $\mathcal{U}_{n,\sigma}$ and let $A_1,\ldots,A_d$ be a collection of $d$ finite sequences over the same alphabet. We now consider

$$W_n(A_1,\ldots,A_d) = \mathbf{E}\Big[\max_{i_1+\cdots+i_d=n} L(A_1X_1[1..i_1],\ldots,A_dX_d[1..i_d])\Big].$$

This quantity represents the expected length of an LCS of $d$ words with prefixes $A_1,\ldots,A_d$, respectively, and $d$ suffixes whose lengths sum up to $n$ and whose characters are uniformly and independently chosen in $\Sigma = \{1,\ldots,\sigma\}$. Since $W_n(A_1,\ldots,A_d)$ and $D_{n,\sigma,d}$ behave similarly as $n \to \infty$, Theorem 3.1 implies that, for all $A_1,\ldots,A_d$,

$$\gamma_{\sigma,d} = \lim_{n\to\infty} \frac{W_{nd}(A_1,\ldots,A_d)}{n}. \tag{3.3}$$

Just as in the $d=2$ case, fix $l \in \mathbb{N}$ and let $w_n$ denote the $\sigma^{ld}$-dimensional vector whose coordinates are all the values of $W_{nd}(A_1,\ldots,A_d)$ when $A_1,\ldots,A_d$ vary over all sequences in $\Sigma^l$. We again seek a lower bound for $w_n$ as a function of vectors $w_m$, with $m < n$.

It is easy to see that if all the strings $A_1,\ldots,A_d$ start with the same character, then

$$w_n[A_1,\ldots,A_d] \geqslant 1 + \frac{1}{|\Sigma^d|}\sum_{\vec{c}\in\Sigma^d} w_{n-d}[T(A_1)c(1), T(A_2)c(2),\ldots,T(A_d)c(d)].$$

Informally, the previous inequality asserts that if all the words start with the same character then the expected length of the LCS of all of them, allowing $n$ random extra characters, is at least 1 (the first character) plus the average of the expected length of the LCS of the words obtained by eliminating the first character and 'borrowing' $d$ of the $n$ random characters.

If not all the words start with the same character, we can still find a lower bound, but to write it down we need to introduce some additional notation. For any two sets $X$ and $Y$, we follow the standard convention of denoting by $Y^X$ the set of all mappings from $X$ to $Y$. Also, for a $d$-tuple of strings $A = (A_1, \ldots, A_d)$ and $z \in \Sigma$, we let $N_z(A)$ denote the set of indices $j \in \{1, \ldots, d\}$ such that $A_j$'s head is not equal to $z$, *i.e.*, to the set of string indices *not* starting with $z$. For a mapping $c : N_z(A) \to \Sigma$, we define $\tau_z(A, c)$ as the the $d$-tuple of strings obtained from $A$ by replacing each string $A_i$ that does not start with $z$ by the sequence obtained by eliminating its first character and adding the character $c(i)$ at its tail. Formally, $\tau_z(A, c) = (A'_1, \ldots, A'_d)$, where

$$A'_i = \begin{cases} A_i, & \text{if } h(A_i) = z, \\ T(A_i)c(i), & \text{if } h(A_i) \neq z. \end{cases}$$

A crucial fact is that for a $d$-tuple of strings $A$, if its coordinates do not all start with the same character, then

$$w_n[A] \geqslant \max_{z \in \Sigma} \frac{1}{|\Sigma^{N_z(A)}|} \sum_{c \in \Sigma^{N_z(A)}} w_{n-|N_z(A)|}[\tau_z(A, c)].$$

Informally, each term over which the maximum is taken corresponds to the expected length of the LCS of the strings one would obtain by disregarding all first characters of sequences not starting with $z$, and concatenating to the tail of these strings an element randomly chosen over the alphabet $\Sigma$.

For the sake of illustration, consider the following example of the derived inequalities when $\sigma = 2$ and $d = 4$:

$$\begin{aligned} w_n[001, 011, 101, 001] \geqslant \max \Bigg\{ & \frac{1}{2} \sum_{c \in \{0,1\}^{\{3\}}} w_{n-1}[001, 011, 01c(3), 001], \\ & \frac{1}{2^3} \sum_{c \in \{0,1\}^{\{1,2,4\}}} w_{n-3}[01c(1), 11c(2), 101, 01c(4)] \Bigg\}. \end{aligned}$$

In the previous example only the third string over which $w_n$ is evaluated does not start with 0. Hence, the first term over which the maximum is taken is the average of the values of $w_{n-1}$ evaluated at the two possible 4-tuples of strings obtained from $A$ by removing the initial 1 from the third string and adding a 0 or 1 final character. On the other hand, $w_n$ is evaluated at three strings that do not start with a 1. Hence, the second term over which the maximum is taken is the average of the values of $w_{n-3}$ over all the 4-tuples of strings obtained from $A$ by removing all the initial 0s and adding a 0 or 1 final character to those same strings.

Expressing all the derived inequalities in vector form we have that there is a function $F : (\mathbb{R}^{\sigma^{ld}})^d \to \mathbb{R}^{\sigma^{ld}}$ such that

$$w_n \geqslant F(w_{n-1}, w_{n-2}, \ldots, w_{n-d}). \tag{3.4}$$

For the ensuing discussion it will be convenient to rewrite $F$ in an alternative way. For each $z \in \Sigma$ we define the linear transformation $F_z : (\mathbb{R}^{\sigma^{ld}})^d \to \mathbb{R}^{\sigma^{ld}}$ such that

$$F_z(v_1, \ldots, v_d)[A] = \begin{cases} \dfrac{1}{|\Sigma^{N_z(A)}|} \displaystyle\sum_{c \in \Sigma^{N_z(A)}} v_{|N_z(A)|}[\tau_z(A, c)], & \text{if } |N_z(A)| \neq 0, \\ 0, & \text{if } |N_z(A)| = 0. \end{cases} \tag{3.5}$$

Then, if we let $b \in \mathbb{R}^{\sigma^{ld}}$ be the vector with value 1 in the coordinates associated to $d$-tuples of strings of length $l$ starting all with the same character and 0 in the rest of the coordinates, $F$ can be expressed as

$$F(v_1,\dots,v_d) = b + \max_{z\in\Sigma} F_z(v_1,\dots,v_d). \tag{3.6}$$

### 3.4. Finding a lower bound for $\gamma_{\sigma,d}$

In the preceding section we established that for any $d$-tuple of strings $A = (A_1,\dots,A_d)$, each of length $l$, we have $\gamma_{\sigma,d} = \lim_{n\to\infty} w_{nd}[A]/n$. To lower-bound this latter quantity one is tempted to try the following approach: (1) for a fixed word length $l$, compute explicitly $w_0,\dots,w_{d-1}$, and (2) define a new sequence of vectors $(v_n)_{n\in\mathbb{N}}$ as $v_i = w_i$ for $0 \leqslant i \leqslant d-1$, and then iteratively define $v_n = F(v_{n-1}, v_{n-2},\dots,v_{n-d})$, for all $n \geqslant d$. Since $F$ is monotone and by (3.4), we have that $v_n \leqslant w_n$ for every $n \in \mathbb{N}$. It is natural to fix an arbitrary $d$-tuple of strings $A = (A_1,\dots,A_d)$ and estimate a lower bound for $\gamma_{\sigma,d}$ by $\lim_{n\to\infty} v_{nd}[A]/n$ for large enough $n$.

Unfortunately, for the approach discussed in the previous paragraph to work one would need to determine for which values of $n$ the quantity $v_{nd}[A]/n$ is effectively a lower bound for $\gamma_{\sigma,d}$. Indeed, $v_{nd}[A]/n$ does not even need to be increasing and $w_{nd}[A]/n$ equals $\gamma_{\sigma,d}$ only in the limit when $n \to \infty$. We will pursue a different approach that relies on the next lemma which is a generalization of an observation by Lueker [12] for the $d = \sigma = 2$ case.

**Lemma 3.4.** *Let $\mathcal{F} : (\mathbb{R}^{\sigma^{ld}})^d \to \mathbb{R}^{\sigma^{ld}}$ be a transformation that satisfies the following properties.*

(1) **Monotonicity.** *If the inequality $(v_1, v_2,\dots,v_d) \leqslant (w_1, w_2,\dots,w_d)$ holds component-wise, then the inequality $\mathcal{F}(v_1, v_2,\dots,v_d) \leqslant \mathcal{F}(w_1, w_2,\dots,w_d)$ also holds component-wise.*

(2) **Translation invariance.** *Let $\mathbf{1}$ be the vector of ones in $\mathbb{R}^{\sigma^{ld}}$ and $\vec{\mathbf{1}} = (\mathbf{1},\dots,\mathbf{1})$ be the vector of ones in $(\mathbb{R}^{\sigma^{ld}})^d$. Then, for any $r \in \mathbb{R}$ and for all $(v_1, v_2,\dots,v_d) \in (\mathbb{R}^{\sigma^{ld}})^d$,*

$$\mathcal{F}((v_1, v_2,\dots,v_d) + r\vec{\mathbf{1}}) = \mathcal{F}(v_1,\dots,v_d) + r\mathbf{1}.$$

(3) **Feasibility.** *There exists a* feasible triplet for $\mathcal{F}$, *i.e., a $(u, r, \varepsilon)$ with $u \in \mathbb{R}^{\sigma^{ld}}$, $r \in \mathbb{R}$, and $0 \leqslant \varepsilon \leqslant r$ such that*

$$\mathcal{F}(u + (d-1)r\mathbf{1},\dots,u + 2r\mathbf{1}, u + r\mathbf{1}, u) \geqslant u + (dr - \varepsilon)\mathbf{1}.$$

*Then, for any sequence $(v_n)_{n\in\mathbb{N}}$ of vectors in $\mathbb{R}^{\sigma^{ld}}$ such that $v_n \geqslant \mathcal{F}(v_{n-1},\dots,v_{n-d})$ for all $n \geqslant d$, there exists a vector $u_0$ in $\mathbb{R}^{\sigma^{ld}}$ such that, for all $n \geqslant 0$,*

$$v_n \geqslant u_0 + n(r - \varepsilon)\mathbf{1}. \tag{3.7}$$

**Proof.** Let $\mathcal{F}$ be a transformation satisfying the hypothesis of the lemma and $(u, r, \varepsilon)$ a feasible triplet for $\mathcal{F}$. Let $(v_n)_{n\in\mathbb{N}}$ be a sequence of vectors as in the lemma's statement and let $\alpha \in \mathbb{R}$ be large enough so that, for all $j \leqslant d-1$,

$$v_j + \alpha\mathbf{1} \geqslant u + j(r - \varepsilon)\mathbf{1}.$$

For example, set $\alpha$ to be the largest component of the vector $\max_{0\leqslant j\leqslant d-1}(u + j(r-\varepsilon)\mathbf{1} - v_j)$.

Note that $u_0 = u - \alpha\mathbf{1}$ satisfies (3.7) for all $n \leqslant d-1$. We will prove by induction that this holds for all $n \in \mathbb{N}$. Suppose that (3.7) holds up to $n-1$. Using the inductive hypothesis we have

$$\begin{aligned}
&(v_{n-1}, \ldots, v_{n-d}) \\
&\quad \geqslant (u_0 + (n{-}1)(r{-}\varepsilon)\mathbf{1}, \ldots, u_0 + (n{-}j)(r{-}\varepsilon)\mathbf{1}, \ldots, u_0 + (n{-}d)(r{-}\varepsilon)\mathbf{1}) \\
&\quad = (u + (d{-}1)r\mathbf{1}, \ldots, u + (d{-}j)r\mathbf{1} + (j{-}1)\varepsilon\mathbf{1}, \ldots, u + (d{-}1)\varepsilon\mathbf{1}) \\
&\qquad + ((n{-}d)(r{-}\varepsilon) - (d{-}1)\varepsilon - \alpha)\vec{\mathbf{1}} \\
&\quad \geqslant (u + (d{-}1)r\mathbf{1}, \ldots, u + (d{-}j)r\mathbf{1}, \ldots, u) \ + ((n{-}d)(r{-}\varepsilon) - (d{-}1)\varepsilon - \alpha)\vec{\mathbf{1}}.
\end{aligned}$$

Evaluating $\mathcal{F}$ at the terms on both sides of the previous inequality we get, by monotonicity and translation invariance, that

$$\begin{aligned}
v_n &\geqslant \mathcal{F}(v_{n-1}, \ldots, v_{n-d}) \\
&\geqslant \mathcal{F}(u + (d{-}1)r\mathbf{1}, \ldots, u + (d{-}j)r\mathbf{1}, \ldots, u) + ((n{-}d)(r{-}\varepsilon) - (d{-}1)\varepsilon - \alpha)\mathbf{1}.
\end{aligned}$$

Since $(u, r, \varepsilon)$ is a feasible triplet, it follows that

$$\begin{aligned}
v_n &\geqslant u + (dr - \varepsilon)\mathbf{1} + ((n{-}d)(r{-}\varepsilon) - (d{-}1)\varepsilon - \alpha)\mathbf{1} \\
&= u - \alpha\mathbf{1} + n(r - \varepsilon)\mathbf{1} = u_0 + n(r - \varepsilon)\mathbf{1}.
\end{aligned}$$

This completes the proof. □

From $F$'s definition it easily follows that $F$ is monotone and invariant under translations. If we find a feasible triplet $(u, r, \varepsilon)$ for $F$ then, by Lemma 3.4, we can conclude that the sequence of vectors $(w_n)_{n\in\mathbb{N}}$ satisfy $w_n \geqslant u_0 + n(r - \varepsilon)\mathbf{1}$ for all $n$. It follows from (3.3) that

$$\gamma_{\sigma,d} \geqslant d(r - \varepsilon).$$

The key point we are trying to make is that in order to establish a good lower bound for $\gamma_{\sigma,d}$ one only needs to exhibit a good feasible triplet, namely one such that $(r - \varepsilon)$ is as large as possible.

Empirically, one observes that for any set of initial vectors $v_0, \ldots, v_{d-1}$, if one makes $v_{n+d} = F(v_{n+d-1}, \ldots, v_n)$ for all $n \in \mathbb{N}$, then the sequence $(v_n)_{n\in\mathbb{N}}$ is such that $v_n/n$ seems to converge to a vector with all its components taking the same value. In fact, one observes that for large values of $n$ the vectors $v_n$ and $v_{n+1}$ differ essentially by a constant (independent of $n$) times the all ones vector. Roughly, there exists a real value $r$ such that $v_{n+1} - v_n$ is approximately $r\mathbf{1}$ for all large enough $n$. Since, by definition $v_{n+d} = F(v_{n+d-1}, \ldots, v_{n+1}, v_n)$, this implies that

$$F(v_n + (d{-}1)r\mathbf{1}, v_n + (d{-}2)r\mathbf{1}, \ldots, v_n + r\mathbf{1}, v_n) \sim v_n + dr\mathbf{1}.$$

It follows that one possible approach to find a feasible triplet is to consider an $n$ large enough so that the difference between $v_n$ and $v_{n-1}$ is essentially a constant times the all ones vector. Then, set $u = v_n$, and define $r$ as the maximum value such that $v_n - v_{n-1} \geqslant r\mathbf{1}$ and $\varepsilon$ as the minimum possible value such that the triplet $(u, r, \varepsilon)$ is feasible for $F$. The following result validates the approach just described.

**Lemma 3.5.** *Let $\mathcal{F} : (\mathbb{R}^{\sigma^{ld}})^d \to \mathbb{R}^{\sigma^{ld}}$ be a monotone and translation-invariant transformation. Let $v_0, \ldots, v_{d-1} \in \mathbb{R}^{\sigma^{ld}}$ and $v_{n+d} = \mathcal{F}(v_{n+d-1}, \ldots, v_{n+1}, v_n)$ for all $n \in \mathbb{N}$. If for some $r \in \mathbb{R}$, $n_0 \geqslant 1$*

*and* $\varepsilon > 0$ *we have* $\|v_{n+1} - v_n - r\mathbf{1}\|_\infty \leqslant \varepsilon/2d$ *for all* $n \in \{n_0, \dots, n_0{+}d{-}1\}$, *then* $(v_{n_0}, r, \varepsilon)$ *is a feasible triplet for* $\mathcal{F}$.

**Proof.** First, observe that the monotonicity and translation invariance property of $\mathcal{F}$ implies that

$$\|\mathcal{F}(x_0, \dots, x_{d-1}) - \mathcal{F}(y_0, \dots, y_{d-1})\|_\infty \leqslant \max_{i=0,\dots,d-1} \|x_i - y_i\|_\infty.$$

Let $u = v_{n_0}$ and note that $\|v_{n_0+i} - (u + ir\mathbf{1})\|_\infty \leqslant i\varepsilon/2d < \varepsilon/2$ for $0 \leqslant i \leqslant d$. Hence, by definition of $v_{n_0+d}$,

$$\|v_{n_0+d} - \mathcal{F}(u + (d{-}1)r\mathbf{1}, u + (d{-}2)r\mathbf{1}, \dots, u + r\mathbf{1}, u)\|_\infty \leqslant \varepsilon/2.$$

Since $\|v_{n_0+d} - (u + dr\mathbf{1})\|_\infty \leqslant \varepsilon/2$ it follows that

$$\|(u + dr\mathbf{1}) - \mathcal{F}(u + (d{-}1)r\mathbf{1}, u + (d{-}2)r\mathbf{1}, \dots, u + r\mathbf{1}, u)\|_\infty \leqslant \varepsilon.$$

In other words, $(u, r, \varepsilon)$ is a feasible triplet for $\mathcal{F}$. □

It is easy to check that $F$ satisfies the hypothesis of Lemma 3.5. This justifies, together with the empirical observation that $v_{n+1} - v_n$ is approximately $r\mathbf{1}$ for large values of $n$, the general approach described in this section for finding a feasible triplet for $F$, and thus a lower bound for $\gamma_{\sigma,d}$. It is important to stress here that there is no need to prove the convergence of $v_n/n$ to $r\mathbf{1}$ in order to establish the lower bound $\gamma_{\sigma,d} \geqslant d(r - \varepsilon)$. We only need to find a feasible triplet $(u, r, \varepsilon)$ for $F$. The characteristics of $F$, empirical observations and Lemma 3.5, efficiently lead to such feasible triplets.

### 3.5. Implementation and results; new bounds

In this section we describe the procedure we implemented in order to find a feasible triplet $(u, r, \varepsilon)$ for $F$ and, as a corollary, a lower bound for $\gamma_{\sigma,d}$. The procedure is called FEASIBLETRIPLET; it is parametrized in terms of the number of sequences $d$ and the alphabet $\Sigma$, and its pseudocode is given in Algorithm 1. In order to implement $F$ we rely on the characterization given by (3.5) and (3.6). Since the $F_z$s are linear transformations, they can be represented as matrices. This allows for fast evaluation of the $F_z$s, but requires a prohibitively large amount of main memory for all but small values of $\sigma$, $l$ and $d$. In order to optimize memory usage, we use the fact that by distinguishing (3.5) according to the cardinality of $N_z(A)$ where $A \in (\Sigma^l)^d$, $F_z$ can be written as

$$F_z(v_1, \dots, v_d) = \frac{1}{\sigma^1} F_{z,1}(v_1) + \cdots + \frac{1}{\sigma^d} F_{z,d}(v_d),$$

where

$$F_{z,i}(v_i)[A] = \begin{cases} \displaystyle\sum_{c \in \Sigma^{N_z(A)}} v_i[\tau_z(A, c)], & \text{if } |N_z(A)| = i, \\ 0, & \text{otherwise.} \end{cases}$$

Note in particular that every $F_{z,i}$ can be represented as a 0–1 sparse matrix.

**Algorithm 1** Procedure for computing a feasible triple for $F$

**procedure** $\textsc{FeasibleTriplet}_{d,\Sigma}(l,n)$ $\quad\triangleright\ l \in \mathbb{N}$ parameter, $n \in \mathbb{N}$ iteration steps
  **for** $i = 0,\dots,d-1$ **do**
    $v_i \leftarrow \mathbf{0}$ $\quad\triangleright$ Where $\mathbf{0}$ denotes the vector of zeros in $\mathbb{R}^{\sigma^{ld}}$
  **end for**
  $(u,r,\varepsilon) \leftarrow (v_0,0,0)$
  **for** $i = d,\dots,n$ **do**
    $v_i \leftarrow F(v_{i-1},v_{i-2},\dots,v_{i-d})$
    $R \leftarrow \max_{A\in(\Sigma^l)^d}(v_i - v_{i-1})[A]$
    $W \leftarrow v_i + dR\mathbf{1} - F(v_i+(d-1)R\mathbf{1},\dots,v_i+R\mathbf{1},v_i)$
    $E \leftarrow \max\{0,\max_{A\in(\Sigma^l)^d} W[A]\}$
    **if** $R - E \geqslant r - \varepsilon$ **then**
      $(u,r,\varepsilon) \leftarrow (v_i,R,E)$
    **end if**
  **end for**
  **return** $(u,r,\varepsilon)$
**end procedure**

*Table 1.* Best-known lower bounds for $\gamma_{\sigma,2}$ (in boldface).

| $\sigma$ | $\gamma_{\sigma,2}$ | | |
|---|---|---|---|
| | This work | Lower bound from [2] | Lower bound from [5, 8] |
| 3 | **0.671697** | 0.63376 | 0.61538 |
| 4 | **0.599248** | 0.55282 | 0.54545 |
| 5 | **0.539129** | 0.50952 | 0.50615 |
| 6 | **0.479452** | 0.46695 | 0.47169 |
| 7 | 0.444577 | – | **0.44502** |
| 8 | 0.356545 | – | **0.42237** |
| 9 | 0.327935 | – | **0.40321** |
| 10 | 0.303490 | – | **0.38656** |

In our experiments we ran Algorithm 1 for different values of $l$ and alphabet sizes $\sigma$. As one would expect, the derived lower bounds improve as $l$ grows. However, the memory resources required to perform the computation also increases. Indeed, throughout the second loop of Algorithm 1 we need to store $d$ vectors of dimension $\sigma^{ld}$. Also, a simple analysis of the definition of the sparse matrix $F_{z,i}$ shows that it has $\binom{d}{i}\sigma^{(l-1)d}(\sigma-1)^i\sigma^i$ non-zero entries. It follows that a sparse matrix representation of $F_z$ has roughly $\sigma^{ld}(\sigma-1)^d$ non-zero entries. Hence, the necessary computations are feasible only for small values of $\sigma$, $l$ and $d$, unless additional features of the matrices involved are taken advantage of in order to optimize memory usage.

Table 1 summarizes the lower bounds we obtain for $\gamma_{\sigma,2}$ and contrasts them with previously derived ones. To the best of our knowledge, for the $d = 2$ case and alphabet sizes 3, 4, 5, and 6, this work provides the currently best-known lower bounds for $\gamma_{\sigma,2}$. It might be worth mentioning that, as can be seen in that table, the bound of [5, 8] is better than the bound of the more recent

*Table 2.* Lower bounds for $\gamma_{\sigma,d}$.

Alphabet size $\sigma = 2$

| $d$ | $L$ such that $\gamma_{2,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.781281 | 10 |
| 3 | 0.704473 | 7 |
| 4 | 0.661274 | 5 |
| 5 | 0.636022 | 4 |
| 6 | 0.617761 | 3 |
| 7 | 0.602493 | 2 |
| 8 | 0.594016 | 2 |
| 9 | 0.587900 | 2 |
| 10 | 0.570155 | 1 |
| 11 | 0.570155 | 1 |
| 12 | 0.563566 | 1 |
| 13 | 0.563566 | 1 |
| 14 | 0.558494 | 1 |

Alphabet size $\sigma = 3$

| $d$ | $L$ such that $\gamma_{3,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.671697 | 6 |
| 3 | 0.556649 | 4 |
| 4 | 0.498525 | 3 |
| 5 | 0.461402 | 2 |
| 6 | 0.421436 | 1 |
| 7 | 0.413611 | 1 |
| 8 | 0.405539 | 1 |

Alphabet size $\sigma = 4$

| $d$ | $L$ such that $\gamma_{4,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.599248 | 5 |
| 3 | 0.457311 | 3 |
| 4 | 0.389008 | 2 |
| 5 | 0.335517 | 1 |
| 6 | 0.324014 | 1 |

Alphabet size $\sigma = 5$

| $d$ | $L$ such that $\gamma_{5,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.539129 | 4 |
| 3 | 0.356717 | 2 |
| 4 | 0.289398 | 1 |
| 5 | 0.273884 | 1 |

Alphabet size $\sigma = 6$

| $d$ | $L$ such that $\gamma_{6,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.479452 | 3 |
| 3 | 0.309424 | 2 |
| 4 | 0.245283 | 1 |

Alphabet size $\sigma = 7$

| $d$ | $L$ such that $\gamma_{7,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.444577 | 3 |
| 3 | 0.234567 | 1 |
| 4 | 0.212786 | 1 |

Alphabet size $\sigma = 8$

| $d$ | $L$ such that $\gamma_{8,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.356545 | 2 |
| 3 | 0.207547 | 1 |

Alphabet size $\sigma = 9$

| $d$ | $L$ such that $\gamma_{9,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.327935 | 2 |
| 3 | 0.186104 | 1 |

Alphabet size $\sigma = 10$

| $d$ | $L$ such that $\gamma_{10,d} \geqslant L$ | Parameter $l$ |
|---|---|---|
| 2 | 0.303490 | 2 |
| 3 | 0.168674 | 1 |

work of [2] for alphabet size 6, and that for bigger alphabet sizes, the bound of [5, 8] is still better than ours.

The best-known lower bound for $\gamma_{2,2}$ is still that established by Lueker [12]. Table 2 lists the distinct choices of $\sigma$ and $d$ for which we could execute Algorithm 1 and indicates the value of the parameter $l$ giving rise to the reported lower bound.

### 3.6. Disproving Steele's $\gamma_{2,2} = \gamma_{2,3}^2$ speculation

We showed in Section 2 that $\gamma_{2,d} > \gamma_{2,2}^{d-1}$ for all $d \geqslant 5$. We now establish that this is also the case when $d = 3$ and $d = 4$. Recall that Lueker [12] proved that $\gamma_{2,2} \leqslant U$ for $U = 0.826280$. From

Table 2 we see that for $d = 3$ and $d = 4$, the indicated lower bound for $\gamma_{2,d}$ is strictly greater than $U^{d-1}$, and is therefore also strictly greater than $\gamma_{2,2}^{d-1}$. This implies that $\gamma_{2,d} > \gamma_{2,2}^{d-1}$ for $d = 4$ and $d = 3$ as claimed. Together with the results of Section 2 this establishes that $\gamma_{2,d} > \gamma_{2,2}^{d-1}$ for all $d \geqslant 3$.

## 4. Final comments

As already mentioned at the start of this paper, Steele [15] pointed out that it would be of interest to find relations between the values of the $\gamma_{\sigma,d}$s, especially between $\gamma_{2,2}$ and $\gamma_{2,3}$. We think it would be very interesting if such a relation would exist. In fact, it might shed some light upon the longstanding open problem of determining the exact value of the Chvátal–Sankoff constant.

Lacking a relation among the $\gamma_{\sigma,d}$s it would still be interesting to relate these terms to some other constants that arise in connection with other combinatorial problems. A step in this direction was taken by Kiwi, Loebl and Matoušek [10], who showed that $\sqrt{\sigma}\gamma_{\sigma,2} \to c_2$ when $\sigma \to \infty$, where $c_2$ is a constant that turns up in the study of the Longest Increasing Sequence (LIS) problem (also known as Ulam's problem). Specifically, $c_2$ is the limit to which the expected length of a LIS of a randomly chosen permutation of $\{1, \ldots, n\}$ converges when normalized by $\sqrt{n}$. Logan and Shepp [11] and Vershik and Kerov [18] showed that $c_2 = 2$. Consider now the following experiment. Choose $n$ points in a unit $d$-dimensional cube $[0, 1]^d$ and let $H_d(n)$ be the random variable corresponding to the length of a longest chain (for the standard partial order in $\mathbb{R}^d$) of the $n$ chosen points. Bollobás and Winkler [3] proved that there are constants $c'_2, c'_3, \ldots$ such that $c'_d < e$, $\lim_{d\to\infty} c'_d = e$ and $\lim_{n\to\infty} H_d(n)/n^{1/d} = c'_d$. By labelling a set $S$ of points in $[0, 1]^2$ in increasing order of their $x$-coordinate and reading the labels in the order of their $y$-coordinates one can associate a permutation $\pi$ to the set $S$. It is easy to see that a chain of points in $S$ is in one-to-one correspondence to an increasing sequence of $\pi$. Hence, it follows that $c'_2 = c_2$. Soto [14] extended the results of [10] and showed that $\sigma^{1-1/d}\gamma_{\sigma,d} \to c'_d$ when $\sigma \to \infty$. We think that any similar type of result, or even a reasonable conjecture, that would hold for fixed $\sigma$ and $d$ would also be quite interesting.

## References


[1] Alexander, K. S. (1994) The rate of convergence of the mean of the longest common subsequence. *Ann. Appl. Probab.* **4** 1074–1083.

[2] Baeza-Yates, R., Navarro, G., Gavaldá, R. and Schehing, R. (1999) Bounding the expected length of the longest common subsequences and forests. *Theory Comput. Syst.* **32** 435–452.

[3] Bollobás, B. and Winkler, P. (1988) The longest chain among random points in Euclidean space. *Proc. Amer. Math. Soc.* **103** 347–353.

[4] Chvátal, V. and Sankoff, D. (1975) Longest common subsequences of two random sequences. *J. Appl. Probab.* **12** 306–315.

[5] Dančík, V. (1994) Expected length of longest common subsequences. PhD thesis, Department of Computer Science, University of Warwick.

[6] Dančík, V. (1998) Common subsequences and supersequences and their expected length. *Combin. Probab. Comput.* **7** 365–373.

[7] Dančík, V. and Paterson, M. (1995) Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Struct. Alg.* **6** 449–458.

[8] Deken, J. P. (1979) Some limit results for longest common subsequences. *Discrete Math.* **26** 17–31.

[9] Janson, S., Łuczak, T. and Rucinski, A. (2000) *Random Graphs*, Wiley.
[10] Kiwi, M., Loebl, M. and Matoušek, J. (2005) Expected length of the longest common subsequence for large alphabets. *Adv. Math.* **197** 480–498.
[11] Logan, B. and Shepp, L. (1977) A variational problem of random Young tableaux. *Adv. Math.* **26** 206–222.
[12] Lueker, G. (2003) Improved bounds on the average length of longest common subsequences. In *Proc. 14th Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 130–131.
[13] Pevzner, P. (2000) *Computational Molecular Biology: An Algorithmic Approach*, MIT Press.
[14] Soto, J. (2006) Variantes aleatorias de la subsecuencia común más grande. Departamento de Ingeniería Matemática, U. Chile. (In Spanish.)
[15] Steele, J. M. (1986) An Efron–Stein inequality for nonsymmetric statistics. *Ann. Statist.* **14** 753–758.
[16] Steele, J. M. (1996) *Probability Theory and Combinatorial Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.
[17] Szpankowski, W. (2000) *Average Case Analysis of Algorithms and Sequences*, Series in Discrete Mathematics and Optimization, Wiley InterScience.
[18] Vershik, A. and Kerov, S. (1977) Asymptotics of the Plancherel measure of the symmetric group and the limiting form of Young tableaux. *Doklady Akademii Nauk SSSR* **233** 1024–1028.
[19] Waterman, M. (1995) *Introduction to Computational Biology: Average Case Analysis of Algorithms and Sequences*. Series in Discrete Mathematics and Optimization, Chapman & Hall/CRC.