On compound Poisson approximation for sequence matching

Marianne Månsson

June 24, 1999

Abstract. Consider the sequences $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{j=1}^n$ of independent random variables, which take values in a finite alphabet, and assume that the variables $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ follow the distributions μ and ν, respectively. Two variables X_i and Y_j are said to match if $X_i = Y_j$. Let the number of matching subsequences of length k between the two sequences, when r, $0 \leq r < k$, mismatches are allowed, be denoted by W.

In this paper we use Stein's method to bound the total variation distance between the distribution of W and a suitably chosen compound Poisson distribution. To derive rates of convergence, the case where $E[W]$ stays bounded away from infinity, and the case where $E[W] \to \infty$ as $m, n \to \infty$, have to be treated separately. Under the assumption that $\ln n / \ln(mn) \to \rho \in (0,1)$, we give conditions on the rate at which $k \to \infty$, and on the distributions μ and ν, for which the variation distance tends to zero.

AMS 1991 Classification: 62E17, 60C05.
Key words and phrases: Compound Poisson approximation, sequence matching, Stein's method.

1 Introduction

Assume that $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{j=1}^n$ are sequences of independent random variables taking values in a finite alphabet A. Let $\{X_i\}$ and $\{Y_j\}$ follow the distributions μ and ν, respectively, and let

$$p = P(X_i = Y_j) = \sum_{a \in A} P(X_i = Y_j = a)$$

be the probability of a 'match' between X_i and Y_j, $i, j \geq 1$. In this paper we will study the distribution of the number of 'nearly matching' consecutive subsequences of a certain length between $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{j=1}^n$, where 'nearly matching' means that a number of mismatches is allowed.

Fix k and r, where $1 \leq k \leq \min\{m, n\}$, and $0 \leq r < k$, and let

$$I_{ij} \quad = \quad 1\{\sum_{t=0}^{k-1} 1\{X_{i+t} = Y_{j+t}\} = k - r\}. \tag{1.1}$$

Then $I_{ij} = 1$ when there are exactly $k - r$ matches and r mismatches between the subsequences $X_i, \ldots, X_{i+k-1}$ and $Y_j, \ldots, Y_{j+k-1}$. To avoid edge effects, we treat the sequences as 'circles', and let X_{m+i} and X_{-m+i} be identified with X_i, and similarly let Y_{n+j} and Y_{-n+j} be identified with Y_j, where $1 \leq i \leq m$, $1 \leq j \leq n$. The number of matching subsequences of length k with r mismatches can then be written as

$$W = \sum_{(i,j)\in\Gamma} I_{ij}, \tag{1.2}$$

where $\Gamma = \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$.

There is an extensive literature on various sequence matching problems. This is motivated for instance by the interest in molecular biology to compare DNA, RNA or protein sequences with the purpose of finding similarities which cannot be explained by chance. A variable which has received particular attention is the length of the longest matching subsequence with at most r mismatches, $r \geq 0$, between $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{j=1}^n$: $M = \max\{k : \sum_{t=0}^{k-1} 1\{X_{i+t} = Y_{j+t}\} \geq k - r, \text{ for some } i, j\}$. Analogues of the strong law of large numbers for M are given in Arratia and Waterman (1985, 1989).

Since W is a sum of indicators it might seem reasonable to approximate W by a Poisson variable, if p is small. However, since there is a strong dependence between indicators for overlapping subsequences, the matching subsequences tend to occur in clumps. In particular, if $I_{ij} = 1$ then the probability is high that also $I_{i-l,j-l} = 1$ or $I_{i+l,j+l} = 1$, $l < k$, especially if $r > 0$. Therefore 'usual' Poisson approximation to the total number of matching subsequences, W defined in (1.2), fails. However, under suitable conditions Poisson approximation of the number of clumps, W^* say, appropriately defined, has proved successful.

In e.g. Arratia, Gordon and Waterman (1986) and Karlin and Ost (1988), Bonferroni inequalities are used to show that W^* is asymptotically Poisson distributed. In more recent papers, the Stein-Chen method is used to derive rates of convergence for the approximations. Some examples are Arratia, Goldstein and Gordon (1989), Arratia, Gordon and Waterman (1990), Neuhauser (1996), Novak (1994, 1995), and Dembo, Karlin and Zeitouni (1994). The conditions in these references are somewhat different (for instance, whether $\mu \neq \nu$ and $r > 0$ are allowed), and some of them will be further discussed below Theorem 3.3.

The length of the longest matching subsequence, M, is related to the number of clumps, $W_k^* = W^*$, by $P(M < k) = P(W_k^* = 0)$, and hence distributional results for W^* carry over to M. In many of the above references, the main purpose is to find the distribution of M, and then Poisson approximation to the number of clumps is suitable to use. However, if the interest is in the total number of matching subsequences, W, then the clumping has to be taken into account in the approximating distribution, and a compound Poisson distribution is then a natural candidate. Stein's method is developed also for compound Poisson approximation, as is described in Section 2. In the present paper this method is used to approximate W by a compound Poisson variable.

2 Compound Poisson approximation by Stein's method

Let $W = \sum_{\alpha\in\Gamma} X_\alpha$, where Γ is a finite family of indices, and the X_α take values in $\mathbf{N}$; in our case, the X_α's are indicator variables. Here, we consider the compound Poisson approximation of W by means of Stein's method. Let CP(Λ), where $\Lambda = \sum_{i\geq 1} \delta_i \lambda_i$, with δ_i the unit mass at the point i, denote the compound Poisson distribution defined by

$$\mathrm{CP}\,(\Lambda) = \mathcal{L}\left(\sum_{i\geq 1} iN_i\right) = \mathcal{L}\left(\sum_{i=1}^{N} V_i\right), \tag{2.3}$$

where $(N_i,\ i \geq 1)$ are independent with $\mathcal{L}(N_i) = \mathrm{Po}\,(\lambda_i)$, and $\mathcal{L}(N) = \mathrm{Po}\,(\sum_{i\geq 1}\lambda_i)$ and V_i, $i = 1, 2, \ldots$, are independent of each other and of N, with $P(V_i = j) = \lambda_j / \sum_{i\geq 1}\lambda_i$.

Letting g_A be the solution of the Stein equation

$$\sum_{i\geq 1} i\lambda_i g(j+i) - jg(j) = 1\{j \in A\} - \mathrm{CP}\,(\Lambda)\{A\},$$

$j \geq 0$, it follows that

$$\begin{aligned} d_{TV}(\mathcal{L}(W), \mathrm{CP}(\Lambda)) &= \sup_{A\subset \mathbf{Z}^+} |P(W \in A) - \mathrm{CP}\,(\Lambda)\{A\}| \\ &= \sup_{A\subset \mathbf{Z}^+} \left| E\left[\sum_{i\geq 1} i\lambda_i g_A(W+i) - Wg_A(W)\right]\right|. \end{aligned} \tag{2.4}$$

Roos (1994a, 1994b) showed that it is often possible to find ε_0 and ε_1 such that

$$\left| E\left[\sum_{i\geq 1} i\lambda_i g(W+i) - Wg(W)\right]\right| \leq \varepsilon_0|g| + \varepsilon_1|\Delta g|, \tag{2.5}$$

for all bounded $g : \mathbf{N} \to \mathbf{R}$, where $|g| = \sup_{j\geq 1}|g(j)|$ and $|\Delta g| = \sup_{j\geq 1}|g(j+1) - g(j)|$, and the quantities λ_i are appropriately defined; and if (2.5) is satisfied, then

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\Lambda)) \leq \varepsilon_0 H_0(\Lambda) + \varepsilon_1 H_1(\Lambda), \tag{2.6}$$

from (2.4), where $H_0(\Lambda) = \sup_{A\subset\mathbf{Z}^+}|g_A|$ and $H_1(\Lambda) = \sup_{A\subset\mathbf{Z}^+}|\Delta g_A|$. In the 'local version' of Stein's method for compound Poisson approximation, which is stated in Theorem 2.1 below, the index set Γ must be partitioned for each $\alpha \in \Gamma$: $\Gamma = \{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw} \cup \Gamma_\alpha^b$, where

$$\Gamma_\alpha^{vs} = \{\beta \in \Gamma \setminus \{\alpha\} : I_\beta \text{ is 'very strongly' dependent on } I_\alpha\},$$

$$\Gamma_\alpha^{vw} = \{\beta \in \Gamma \setminus \{\alpha\} : I_\beta \text{ is 'very weakly' dependent on } I_\alpha\},$$

and

$$\Gamma_\alpha^b = \Gamma \setminus \{\{\alpha\} \cup \Gamma_\alpha^{vs} \cup \Gamma_\alpha^{vw}\}.$$

With $U_\alpha = \sum_{\beta\in\Gamma_\alpha^{vs}} I_\beta$, $Z_\alpha = I_\alpha + U_\alpha$, $X_\alpha = \sum_{\beta\in\Gamma_\alpha^b} I_\beta$, $Y_\alpha = \sum_{\beta\in\Gamma_\alpha^{vw}} I_\beta$, the following theorem holds.

Theorem 2.1 *(Theorem 2 in Roos (1994a)). Using the notation introduced above, (2.5) holds if*

$$\begin{aligned} \lambda_i &= \frac{1}{i}\sum_{\alpha\in\Gamma} E[I_\alpha I[Z_\alpha = i]], \qquad i \geq 1, \qquad (2.7) \\ \varepsilon_0 &= \sum_{\alpha\in\Gamma}\sum_{i=1}^{|\Gamma_\alpha^{vs}|+1} E|E[I_\alpha I[Z_\alpha = i] \mid (I_\beta : \beta \in \Gamma_\alpha^{vw})] - E[I_\alpha I[Z_\alpha = i]]|, \\ \textit{and} \qquad \varepsilon_1 &= \sum_{\alpha\in\Gamma} E[I_\alpha]E[I_\alpha + U_\alpha + X_\alpha] + E[I_\alpha X_\alpha]. \end{aligned}$$

A 'coupling version' of the above theorem can be found in Roos (1994b).

Note that letting $\Gamma_\alpha^{vs} = \emptyset$, $\lambda_1 = \lambda$ and $\lambda_i = 0$, $i \geq 2$, the above reduces to the Poisson case:

$$d_{TV}(\mathcal{L}(W), \mathrm{Po}\,(\lambda)) \;\; \leq \;\; \varepsilon_0 H_0(\lambda) + \varepsilon_1 H_1(\lambda),$$

in which case there exist good bounds on $H_0(\lambda)$ and $H_1(\lambda)$, given respectively by $\min(1, \lambda^{-1/2})$ and $\lambda^{-1}(1-e^{-\lambda})$. Bounds on $H_0(\Lambda)$ and $H_1(\Lambda)$ as sharp as these cannot be found in the compound Poisson case for general Λ. The difficulties in deriving bounds in the compound Poisson approximation case are thoroughly discussed in Barbour (1997). In Barbour, Chen and Loh (1992) the following bound is given:

$$H_0(\Lambda), H_1(\Lambda) \leq \min\left(1, \frac{1}{\lambda_1}\right) \exp\left\{\sum_{i=1}^{\infty} \lambda_i\right\}, \tag{2.8}$$

valid for any $\Lambda = \sum_{i=1}^{\infty} \lambda_i \delta_i$. Because of the exponential term, (2.8) is unsatisfactory when $\sum_{i\geq 1} \lambda_i$ is large. If $i\lambda_i \searrow 0$ or $(\sum_{i\geq 1} i\lambda_i)^{-1} \sum_{i\geq 1} i(i-1)\lambda_i < 1/2$, considerably better bounds are given in Barbour, Chen and Loh (1992) and Barbour and Xia (1999), respectively. In the current problem, unfortunately neither of these conditions are satisfied in general, so the only available bound on $H_0(\Lambda)$ and $H_1(\Lambda)$ is that given in (2.8). However, Barbour and Utev (1998, 1999) developed an alternative way to handle the case when $\sum_{i\geq 1} \lambda_i$ is large. First we introduce the notation $\Omega = \sum_{i\geq 1} \lambda_i$, $\mu_i = \lambda_i/\Omega$, $\boldsymbol{\mu} = \sum_{i\geq 1} \mu_i \delta_i$ and the first moment $m_1 = \sum_{i\geq 1} i\mu_i$. The following result follows from Barbour and Utev (1999), where explicit expressions are given for all involved constants (in Theorems 1.10 and TV and Equations (1.24)–(1.28)).

Theorem 2.2 *Assume that (2.5) is satisfied for some nonnegative integer valued random variable W. Then there exist $S_l(\boldsymbol{\mu}) < \infty$, $l = 0, 1, 2$, and $3/4 < T(\boldsymbol{\mu}) < 1$, such that*

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}\,(\Lambda)) \;\; \leq \;\; \frac{\varepsilon_0 S_0(\boldsymbol{\mu})}{\Omega^{1/2}} + \frac{\varepsilon_1 S_1(\boldsymbol{\mu})}{\Omega} + P(W \leq T(\boldsymbol{\mu})\Omega m_1) S_2(\boldsymbol{\mu}),$$

if $\Lambda = \sum_{i\geq 1} \lambda_i \delta_i$ satisfies the conditions

(i) $\sum_{i\geq 1} \mu_i r_0^i < \infty$ *for some* $r_0 > 1$,

(ii) $\boldsymbol{\mu}$ *is aperiodic* ($\boldsymbol{\mu}\{l\mathbf{Z}_+\} < 1$ *for all* $l \geq 2$),

(iii) $\Omega > \frac{1}{2(1-T(\boldsymbol{\mu}))}$.

The expressions for $S_l(\boldsymbol{\mu})$ and $T(\boldsymbol{\mu})$ are rather involved; hence Theorem 2.2 is most applicable for asymptotic results in the case where $E[W]$ tends to infinity. In that case the following proposition can be useful, where the dependence on n is explicit in the indices.

Proposition 2.3 *Assume that, for each n, (2.5) is satisfied for some nonnegative integer valued random variable W_n. Then there exist $S_l < \infty$, $l = 0, 1, 2$, and $3/4 < T < 1$, which are independent of n (but depend on $\{\boldsymbol{\mu}_n\}_{n\geq 1}$), such that*

$$d_{TV}(\mathcal{L}(W_n), \mathrm{CP}\,(\Lambda_n)) \;\; \leq \;\; \frac{\varepsilon_{0n} S_0}{\Omega_n^{1/2}} + \frac{\varepsilon_{1n} S_1}{\Omega_n} + P(W_n \leq T\Omega_n m_{1n}]) S_2,$$

if the sequence $(\Lambda_n,\ n \geq 1)$ satisfies the conditions (*i*) $\sup_{n\geq 1}\sum_{i\geq 1}\mu_{in}r_0^i<\infty$ *for some* $r_0>1$,

(*ii*) *there exists a non-negative measure* $\boldsymbol{\nu}$ *on* $\mathbf{N}$, *such that the greatest common divisor of* $\{j:\nu_j>0\}$ *equals 1, and such that* $\mu_{in}\geq\nu_i$ *for all* i *and* n,

(*iii*) $\inf_{n\geq 1}\Omega_n>(2(1-T))^{-1}$.

Proof. Using the notation of Barbour and Utev (1999), let

$$m_j=\sum_{i\geq 1}i^j\mu_i,\;\; m_{[j]}=\sum_{i\geq 1}i(i-1)\cdots(i-j+1)\mu_i,\;\; m_*=\sup_{j\geq 1}\left\{\frac{m_{[j]}}{(j-4)!}\right\}^{1/j},$$

$$m_j^{(r)}=\sum_{i\geq 1}i^j\mu_i r^i \quad\text{and}\quad \rho^{*(r)}(\zeta)=\min\{\inf_{\zeta\leq\theta\leq\pi}\rho_1^{(r)}(\theta),\,\inf_{\zeta\leq\theta\leq\pi}\rho_2^{(r)}(\theta)/2,\,1\},$$

where

$$\rho_1^{(r)}(\theta)=1-\frac{1}{\sum_{k\geq 1}\mu_k r^k}\sum_{i\geq 1}r^i\mu_i\cos(i\theta)\text{ and }\rho_2^{(r)}(\theta)=1-\frac{1}{m_1^{(r)}}\sum_{i\geq 1}ir^i\mu_i\cos(i\theta),$$

and $n!=1$ if $n\leq 0$. $T(\boldsymbol{\mu})$ and $S_l(\boldsymbol{\mu})$, $l=0,1,2$, in Theorem 2.2 involve specific values of ζ and r, such that

$$0<\zeta\leq\pi\text{ and }\zeta<\sqrt{2m_2/m_4}, \tag{2.9}$$

and $1<r<r_0$ is small enough that

$$\frac{m_2^{(r)}}{m_4^{(r)}}>\frac{2m_2}{3m_4}\quad\text{and}\quad m_1^{(r)}(1-\rho^{*(r)}(\zeta))<m_1(1-\rho^{*(1)}(\zeta)/2), \tag{2.10}$$

and so that $s=\sqrt{r^2-1}$ satisfies

$$s<\frac{1}{9}\min\left\{1,\frac{1}{m_*},\frac{m_{[2]}+m_1/2}{m_*^3}\right\}. \tag{2.11}$$

Furthermore, some of the moments m_j and $m_{[j]}$ are involved in $S_l(\boldsymbol{\mu})$, $l=0,1,2$, in such a way that to prove this proposition, we must show the following:

(a) m_{jn}, $m_{[j]n}$ and m_{*n} are uniformly bounded above in n, and $m_{jn},m_{*n}\geq 1$,

(b) $\boldsymbol{\mu}_n$ is aperiodic for all n,

(c) there exist fixed choices of ζ and r such that (2.9)–(2.11) are valid for all n.

(a) By the definitions, $m_{jn},m_{*n}\geq 1$ for all n. To get upper bounds on m_{jn}, $m_{[j]n}$ and m_{*n} we start by noting that by condition (*i*), there exists $S<\infty$ such that $\sup_{n\geq 1}\sum_{i\geq 1}\mu_{in}r_0^i<S$ for some $r_0>1$. This implies that for each j, m_{jn} and $m_{[j]n}$ are uniformly bounded in n from above, since there exists c_j such that $i^j<r_0^i$ for $i>c_j$, and hence

$$m_{[j]n}\leq m_{jn}=\sum_{i\geq 1}i^j\mu_{in}\leq\sum_{i=1}^{c_j}i^j\mu_{in}+\sum_{i=c_j+1}^{\infty}r_0^i\mu_{in}\leq\sum_{i=1}^{c_j}i^j+S<\infty.$$

Furthermore, $\mu_{in} \leq S/r_0^i$, and hence

$$\begin{aligned}\left(\frac{m_{[j]n}}{(j-4)!}\right)^{1/j} &\leq S^{1/j}\left(\sum_{i\geq 1}\frac{i^j j^4}{r_0^i j!}\right)^{1/j} \\ &\sim \frac{S^{1/j} j^{7/(2j)} e}{j(2\pi)^{1/(2j)}}\left(\sum_{i\geq 1}\frac{i^j}{r_0^i}\right)^{1/j}, \qquad (2.12)\end{aligned}$$

by Stirling's formula. Now, for some $r_0^{-1} < b < 1$, $0 < c < \infty$ and $i \geq cj$,

$$\left(\frac{i+1}{i}\right)^j \leq e^{j/i} \leq b r_0,$$

and hence

$$\sum_{i\geq 1}\frac{i^j}{r_0^i} \leq \sum_{i<cj}\frac{i^j}{r_0^i} + \frac{(cj)^j}{r_0^{cj}}\sum_{i\geq 0} b^i \leq \left(\frac{r_0}{r_0-1} + (1-b)^{-1}\right)(cj)^j,$$

since $r_0 > 1$. Thus (2.12) is bounded by $4Se(r_0(r_0-1)^{-1} + (1-b)^{-1})c$, for all n and j, so that $\sup_{n\geq 1} m_{*n} < \infty$.

(b) The aperiodic assumption (ii) of Theorem 2.2 is expressed as follows in Assumption A in Barbour and Utev (1999): For any $0 < \zeta \leq \pi$, $\rho^{*(1)}(\zeta) > 0$.

To show that Assumption A holds for every n, we note that

$$\rho_{1n}^{(1)}(\theta) = 1 - \sum_{i\geq 1}\mu_{in}\cos(i\theta) = \sum_{i\geq 1}\mu_{in}(1-\cos(i\theta)) \geq \sum_{i\geq 1}\nu_i(1-\cos(i\theta)), \qquad (2.13)$$

for all $n \geq 1$. Now, $1-\cos(i\theta) \geq 0$ for all θ, and $1-\cos(i\theta)$ can be 0 only if $\theta = 2\pi p/q$, for some $p, q \in \mathbf{Z}^+$, and $1 \leq p \leq q/2$. Then those i for which $\cos(i\theta) = 1$ are $i = q, 2q, 3q, \ldots$, which have gcd $q \geq 2$. By condition (ii), the gcd for those i with $\nu_i > 0$ is 1, and so follows that there exists $j \neq q, 2q, 3q, \ldots$, for which $\nu_j(1-\cos(j\theta)) > 0$. Hence $\sum_{i\geq 1}\nu_i(1-\cos(i\theta)) > 0$ for all $0 < \theta \leq \pi$, and then by continuity $\inf_{\zeta\leq\theta\leq\pi}\sum_{i\geq 1}\nu_i(1-\cos(i\theta)) > 0$, which together with (2.13), show that $\inf_{\zeta\leq\theta\leq\pi}\rho_{1n}^{(1)}(\theta) > 0$. Similarly $\inf_{\zeta\leq\theta\leq\pi}\rho_{2n}^{(1)}(\theta)$ is bounded away from 0 uniformly in n, and hence Assumption A is satisfied for all n.

(c) By (a) it follows that there exists a single choice of ζ which satisfies (2.9) for all n. If $m_{jn}^{(r)} \to m_{jn}$ and $\rho_n^{*(r)}(\zeta) \to \rho_n^{*(1)}(\zeta)$ uniformly in n as $r \searrow 1$, then there exists also a single choice of r, such that (2.10) and (2.11) are satisfied for all n, since there exists $\delta > 0$ such that $\rho_n^{*(1)}(\zeta) > \delta$ for all n, by the proof of (b). What remains of the proof of (c) is thus to verify the uniform convergence.

The function $f(r) = r^i$, $i \geq 1$ is convex, thus $(r^i - 1)/(r-1) \leq dr^i/dr$, so that

$$\begin{aligned}|m_{jn}^{(r)} - m_{jn}| &= \sum_{i\geq 1}(r^i - 1)i^j\mu_{in} \\ &\leq (r-1)\sum_{i\geq 1} r^{i-1} i^{j+1}\mu_{in},\end{aligned}$$

and it follows from condition (i) that $m_{jn}^{(r)} \to m_{jn}$ as $r \searrow 1$, uniformly in n. Furthermore,

$$\begin{aligned}
|\rho_{1n}^{(r)}(\theta) - \rho_{1n}^{(1)}(\theta)| &= |\sum_{i\geq 1}(\frac{r^i}{\sum_{k\geq 1}\mu_{kn}r^k} - 1)\mu_{in}\cos(i\theta)| \\
&\leq |\sum_{i\geq 1}\sum_{k\geq 1}\mu_{kn}(r^i - r^k)\mu_{in}\cos(i\theta)| \\
&\leq 2\sum_{i\geq 1}\sum_{k<i} r^k(r-1)r^{i-k-1}(i-k)\mu_{kn}\mu_{in} \\
&\leq 2\sum_{i\geq 1}(r-1)r^{i-1}(i-1)\mu_{in},
\end{aligned}$$

and again uniform convergence follows from condition (i). Using similar arguments for $\rho_{2n}^{(r)}(\theta)$, it follows that $\rho_n^{*(r)}(\zeta) \to \rho_n^{*(1)}(\zeta)$ as $r \searrow 1$, uniformly in n. ■

3 Main results

Before we state the results, some notation need to be introduced. In the case where $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{i=1}^n$ have unequal distributions over the alphabet A, let

$$\mu_a = P(X_i = a), \qquad \nu_a = P(Y_j = a), \qquad (\mu\nu)^* = \max_{a\in A}(\mu_a\nu_a),$$

$i = 1, \ldots, m$, $j = 1, \ldots, n$,

$$q_{\mu k} = \sum_{a\in A}\mu_a^{k/2+1}\nu_a^{k/2}, \qquad q_{\nu k} = \sum_{a\in A}\mu_a^{k/2}\nu_a^{k/2+1}, \qquad \text{and} \qquad q_{\mu\nu k} = \sum_{a\in A}(\mu_a\nu_a)^{(k+1)/2}.$$

Recall that $p = P(X_i = Y_j)$, for any i, j, so that $p = \sum_{a\in A}\mu_a\nu_a$. Furthermore, if $k \geq 2$ is even, then $q_{\mu k}$ can be interpreted as the probability for a chain of k equalities starting and ending in sequence $\{X_i\}$, for instance

$$q_{\mu k} = P(X_1 = Y_1 = X_2 = Y_2 = \cdots = X_{k/2} = Y_{k/2} = X_{k/2+1}).$$

Similarly $q_{\nu k}$ is the probability for a chain of k equalities starting and ending in sequence $\{Y_j\}$, and if $k \geq 1$ is odd then $q_{\mu\nu k}$ is the probability for a chain of k equalities, if we start in $\{X_i\}$ and end in $\{Y_j\}$, or vice versa.

When the sequences $\{X_i\}_{i=1}^m$ and $\{Y_i\}_{i=1}^n$ have the same distribution, we use the notation in Novak (1995), and let

$$p_a = P(X_i = a) = P(Y_j = a) \quad \text{and} \quad q_k = \sum_{a\in A}p_a^{k+1},$$

where $q_k = q_{\mu k} = q_{\nu k}$ if k is even, and $q_k = q_{\mu\nu k}$ if k is odd.

To prove the results below, Stein's method as described in the previous section will be used. In particular, Theorem 2.1 will be used both to derive bounds and to define the parameters λ_l, $l \geq 1$, in the approximating compound Poisson distributions. Recall that λ_l, $l \geq 1$, defined as in (2.7), depends on how the index set Γ is partitioned. We choose

$$\Gamma_{ij}^{vs} = \{(i', j') \in \Gamma\backslash\{(i,j)\} : -k < i - i' = j - j' < k\};$$

the choice of Γ^b_{ij} and Γ^{vw}_{ij} can be found in Section 4, to which the proofs of the main theorems are deferred. Note that in our case $Z_{ij} = I_{ij} + \sum_{(i'j')\in\Gamma^{vs}_{(ij)}} I_{i'j'}$ have the same distribution for all $(i,j) \in \Gamma$, so that

$$\lambda_l \;=\; \frac{E[W]}{l} P(Z_{11} = l | I_{11} = 1), \tag{3.14}$$

where W is defined in (1.2), and

$$E[W] = mn\binom{k}{r} p^{k-r}(1-p)^r.$$

For $r = 0$, the parameters are

$$\lambda_l \;=\; \begin{cases} mnp^{k+l-1}(1-p)^2, & l = 1,\dots,k-1, \\ mn((2k-l-2)p^{k+l-1}(1-p)^2 + 2p^{k+l-1}(1-p))/l, & l = k,\dots,2k-2, \\ mnp^{3k-2}/(2k-1), & l = 2k-1. \end{cases}$$

When $r > 0$ and k is at all large, there are many combinations of matches and mismatches for which $Z_{11} = l$, and a formula for the general case seems hard to find. However, given r and k, the parameters can be calculated by computor.

The following lemma is the first step towards bounds on the total variation distance between the distribution of W and a compound Poisson distribution with the above parameters, and is proved in Subsection 4.1 by means of Theorem 2.1.

Lemma 3.1 *Let W and λ_l, $l \geq 1$, be defined by (1.2) and (3.14), respectively. For any bounded function $g : \mathbf{N} \to \mathbf{R}$,*

$$\left| E[Wg(W) - \sum_{l\geq 1} l\lambda_l g(W+l)] \right| \;\leq\; |\Delta g|\varepsilon,$$

where

$$\varepsilon \;=\; mn\binom{k}{r}^2 \left\{ p^{2(k-r)}6k(m+n) + 4k^2 Q_{2(k-r)} + 2mkq_{\mu 2}^{k-r} + 2nkq_{\nu 2}^{k-r} \right\}, \tag{3.15}$$

with $q_{\mu 2} = q_{\nu 2} = q_2$ if $\mu = \nu$, and

$$Q_{2(k-r)} \;=\; \begin{cases} q_{2(k-r)}, & \text{if } \mu = \nu \\ (\mu\nu)^{*(k-r)}, & \text{otherwise.} \end{cases} \tag{3.16}$$

In the case where $\mu = \nu$ is the uniform distribution,

$$\varepsilon \;=\; mn\binom{k}{r}^2 |A|^{-2(k-r)} 8k(m+n). \tag{3.17}$$

A bound with somewhat better constants is given in (4.34), in the proof of this lemma.

The following result is valid for any m, n, k and r, and follows immediately from (2.6), (2.8), and Lemma 3.1. However, because of the exponential term it is useful only when the total number of clumps $\sum \lambda_l$ is small.

Theorem 3.2 *Let W, $\Lambda = \sum_l \lambda_l \delta_l$ and ε be defined by (1.2), (3.14) and (3.15), respectively. Then*

$$d_{TV}(\mathcal{L}(W), \mathrm{CP}(\Lambda)) \quad \leq \quad \left(1 \wedge \frac{1}{\lambda_1}\right) \exp\left\{\sum_{l\geq 1} \lambda_l\right\} \varepsilon.$$

If k and r are kept fixed while m and n tend to infinity, then the total variation distance obviously tends to infinity. The asymptotics considered here is when k, m and n tend to infinity at suitable rates, while r is fixed. To emphasise the dependence on m and n, we will henceforth use the subscript $_{mn}$ when it seems necessary. The first part of the theorem below follows directly from Theorem 3.2, while the second is proved by means of Lemma 3.1 and Proposition 2.3 in Section 4.2.

Theorem 3.3 *Let W_{mn} and $\Lambda_{mn} = \sum_l \lambda_{lmn}\delta_l$ be defined by (1.2) and (3.14), respectively. If k_{mn} is chosen such that:*

(*i*) *$E[W_{mn}]$ stays bounded away from ∞ as $m, n \to \infty$, then*

$$\begin{aligned} d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) &= O\left(mnk_{mn}^{2r+1}(mq_{\mu2}^{k_{mn}-r} + nq_{\nu2}^{k_{mn}-r} + k_{mn}Q_{2(k_{mn}-r)})\right) \\ &\leq O\left(mnk_{mn}^{2r+1}(m+n)(\mu\nu)^{*(k_{mn}-r)}\right), \end{aligned}$$

(*ii*) *$E[W_{mn}] \to \infty$ as $m, n \to \infty$, then*

$$\begin{aligned} d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) &= O\left(\frac{k_{mn}^{r+1}}{p^{k_{mn}-r}}(mq_{\mu2}^{k_{mn}-r} + nq_{\nu2}^{k_{mn}-r} + k_{mn}Q_{2(k_{mn}-r)})\right) \\ &\leq O\left(\frac{k_{mn}^{r+1}}{p^{k_{mn}-r}}(m+n)(\mu\nu)^{*(k_{mn}-r)}\right), \end{aligned}$$

where $Q_{2(k_{mn}-r)}$ is defined in (3.16), $q_{\mu2} = q_{\nu2} = q_2$ if the sequences have equal distributions. In case of uniform distribution,

$$d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) = \begin{cases} O\left(mnk_{mn}^{2r+1}(m+n)|A|^{-2(k-r)}\right), & \textit{in case (i)}, \\ O\left(k_{mn}^{r+1}(m+n)|A|^{-(k-r)}\right), & \textit{in case (ii)}. \end{cases}$$

We will now consider the asymptotic behaviour of the bounds above when

$$k_{mn} \quad = \quad \lfloor b \log_{1/p} mn + cr \log_{1/p}(\log_{1/p} mn) + d \rfloor, \tag{3.18}$$

where $\lfloor \cdot \rfloor$ denotes the greatest integer function, and $b, c, d \in \mathbf{R}$. The choices of b and c determine the limit of $E[W_{mn}]$:

$$E[W_{mn}] \to \begin{cases} 0, & \text{if } b > 1, \text{ or } b = 1, c > 1, \\ p^d(1-p)^r/(p^r r!), & \text{if } b = c = 1, \\ \infty, & \text{if } b = 1 \text{ and } c < 1, \text{ or } 0 < b < 1. \end{cases}$$

Thus, by choosing $b = c = 1$ and $d = \log_{1/p}\{(1-p)^r/(p^r\lambda r!)\}$ we get $E[W_{mn}] \to \lambda$.

With k_{mn} as in (3.18), it follows that for any $0 < x < 1$

$$x^{k_{mn}-r} = O((mn)^{b \log_{1/p} x}(\ln mn)^{cr \log_{1/p} x}). \tag{3.19}$$

Using this expression for $x = q_{\mu 2}, q_{\nu 2}, q_2, p$ and $(\mu\nu)^*$ yields the order of the bounds on $d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn}))$ given in Theorem 3.3. In particular, if $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are uniformly distributed, then

$$q_{\mu 2}^{k-r} = q_{\nu 2}^{k-r} = Q_{2(k-r)} = p^{2(k-r)} \quad \text{and} \quad \log_{1/p} p^{2(k-r)} = -2(k-r),$$

so that

$$d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) =$$

$$\begin{cases} O((m+n)(\ln mn)^{1+2r(1-c)}(mn)^{1-2b}), & \text{if } E[W_{mn}] \to C,\ C \geq 0, \\ O((m+n)(\ln mn)^{r+1-cr}(mn)^{-b}), & \text{if } E[W_{mn}] \to \infty. \end{cases}$$

To make the bounds in Theorem 3.3, with k_{mn} as in (3.18), approach zero as $m, n \to \infty$, we need some conditions on q_2, $q_{\mu 2}$, $q_{\nu 2}$, and on the relative growth rates of m and n. Conditions for convergence in the case where $\ln n/\ln(mn) \to \rho \in (0,1)$ are summarised in the following corollary. Note that $\rho = 1/2$ if $m = n$.

Corollary 3.4 *Let k_{mn} be given by (3.18), and assume that $\ln n/\ln(mn) \to \rho \in (0,1)$ as $m, n \to \infty$. Then*

$$d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) \quad \to \quad 0 \ \textit{as}\ m, n \to \infty,$$

under the following conditions:

Case 1: $E[W_{mn}] \to C$, $C \geq 0$, ($b > 1$ or $b = 1, c \geq 1$):

$$\textit{conditions:} \qquad \log_{1/p} q_{\mu 2} < (\rho - 2)b^{-1} \ \textit{and}\ \log_{1/p} q_{\nu 2} < -(\rho+1)b^{-1}, \qquad (3.20)$$

Case 2: $E[W_{mn}] \to \infty$ ($b = 1$ and $c < 1$, or $0 < b < 1$):

$$\textit{conditions:} \qquad \log_{1/p} q_{\mu 2} < (\rho - 1)b^{-1} - 1 \ \textit{and}\ \log_{1/p} q_{\nu 2} < -\rho b^{-1} - 1. \quad (3.21)$$

In particular, for uniform distribution, $m = n$ and k_{mn} chosen so that $E[W_{mn}]$ stays bounded away from 0 and infinity,

$$d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) \quad = \quad O(n^{-1}\ln n).$$

Proof. Using Theorem 3.3, (3.19) and that $m \sim (mn)^{1-\rho}$ and $n \sim (mn)^{\rho}$ yields the following bounds.
Case 1:

$$\begin{aligned} d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) \quad = \quad & O((\ln n)^{1+r(2+c\log_{1/p} q_{\mu 2})} n^{(2-\rho+b\log_{1/p} q_{\mu 2})/\rho} \\ & +(\ln n)^{1+r(2+c\log_{1/p} q_{\nu 2})} n^{(\rho+1+b\log_{1/p} q_{\nu 2})/\rho}) \\ & +(\ln n)^{2(1+r)+rc\log_{1/p}(\mu\nu)^*} n^{(1+b\log_{1/p}(\mu\nu)^*)/\rho}). \end{aligned}$$

Since $\log_{1/p}(\mu\nu)^* < \log_{1/p} p = -1$, the last term in this bound tends to zero.
Case 2:

$$\begin{aligned} d_{TV}(\mathcal{L}(W_{mn}), \mathrm{CP}\,(\Lambda_{mn})) \quad = \quad & O(n^{(1-\rho+b(\log_{1/p} q_{\mu 2}+1))/\rho}(\ln n)^{r+1+cr(\log_{1/p} q_{\mu 2}+1)} \\ & +n^{(\rho+b(\log_{1/p} q_{\nu 2}+1))/\rho}(\ln n)^{r+1+cr(\log_{1/p} q_{\nu 2}+1)}) \\ & +(\ln n)^{r+2+cr(\log_{1/p}(\mu\nu)^*+1)} n^{b(\log_{1/p}(\mu\nu)^*+1)/\rho}). \end{aligned}$$

■

Remark 3.1.

- In case of equal distributions and $\rho = 1/2$, (3.20) is satisfied since then $q_{\mu 2} = q_{\nu 2} = q_2$, $\min\{\rho - 2, -(\rho + 1)\}b^{-1} = -3/(2b) \geq -3/2$, and by Lemma 4.1 (*iii*), $\log_{1/p} q_2 < \log_{1/p} p^{3/2} = -3/2$.

- In case of uniform distribution, $\log_{1/p} q_2 = \log_{1/p} p^2 = -2$, so the bounds tends to zero for any ρ in case 1.

- To achieve convergence to zero in the above corollary, a necessary condition is that $b > 1/2$. This condition is also sufficient for uniform distribution, if $\rho = 1/2$. □

With k chosen so that the expected number of clumps of matching subsequences, $E[W^*_{mn}]$, stays bounded away from 0 and infinity, and $\ln n / \ln(mn) \to \rho \in (0,1)$, conditions for convergence of the total variation distance between the distribution of the number of clumps of matches to a Poisson variable can be found in Arratia, Gordon and Waterman (1990) and in Neuhauser (1996). Using the notation $\theta_h = -\log_{1/p} P(Y_2 = X_1 | X_1 = Y_1)$ and $\theta_v = -\log_{1/p} P(X_2 = Y_1 | X_1 = Y_1)$, then if $r = 0$, the conditions for convergence in Case 2 of Corollary 3.4 equals $\rho \in (1 - \theta_v, \theta_h)$, which coincides with the condition for convergence in Arratia, Gordon and Waterman (1990). In Neuhauser (1996) it is shown that convergence to a Poisson limit can be achieved under the somewhat weaker condition

$$\frac{H(\gamma,\nu)}{\ln 1/p} < \rho < 1 - \frac{H(\gamma,\mu)}{\ln 1/p},$$

where $H(\gamma,\alpha) = \sum_{a \in A} \gamma_a \ln(\gamma_a / \alpha_a)$ is the relative entropy, and γ is defined by $\gamma_a = \mu_a \nu_a / p$. It seems plausible that the same technique could be used here to widen the possible ranges for $q_{\mu 2}$ and $q_{\nu 2}$.

Furthermore, in Neuhauser (1996) approximation by a compound Poisson distribution is considered under the condition that $0 < \rho < \frac{H(\gamma,\nu)}{\ln 1/p}$ or $1 - \frac{H(\gamma,\mu)}{\ln 1/p} < \rho < 1$.

In the case of no mismatches, $r = 0$, and $m = n$ ($\rho = 1/2$), and equal distributions, the order of the bounds for Poisson approximation for the number of clumps given in Arratia, Gordon and Waterman (1990), Neuhauser (1996) and Novak (1994) is $O(n^{-1} \ln n)$, which is the same as the order achieved here for compound Poisson approximation of the total number of matches. In the case of mismatches, Novak (1995) presents bounds of order $O(1/\ln n)$. Here we have proved that under the same conditions the order is $O(n^{-1} \ln n)$ for compound Poisson approximation, as for no mismatches.

3.1 At most r mismatches, and the longest matching subsequence

All the above results concern approximation of the number of matching subsequences of length k when *exactly* r mismatches are allowed. Maybe the case where *at most* r mismatches are allowed is more useful, and we conclude this section with some results in this case. Let

$$\tilde{W} = \sum_{(i,j) \in \Gamma} \tilde{I}_{ij},$$

where $\tilde{I}_{ij} = 1$ if $\sum_{t=0}^{k-1} 1\{X_{i+t} = Y_{j+t}\} \geq k - r$, and let $\tilde{\Lambda}$ be defined by means of the same partitioning of Γ as before.

Corollary 3.5 *Let $\tilde{W}$ and $\tilde{\Lambda}$ be defined as above.*

- *If* $\binom{k}{r}$ *is replaced by* $\binom{k+r}{r}$ *in (3.15) and (3.17), then* Lemma 3.1 *remains true.*
- *The bounds given in* Theorem 3.2 *and* Theorem 3.3*, and the conditions for convergence in* Corollary 3.4*, remain true for* $d_{TV}(\mathcal{L}(\tilde{W}), \mathrm{CP}\,(\tilde{\Lambda}))$.

Proof. To adapt the proof of Lemma 3.1 to the present situation, only minor changes are necessary. The equality corresponding to (4.26) here looks like

$$\sum_{(i',j')\in\Gamma_\alpha^{bl}} E[I_\alpha I_{i'j'}] = \sum_{(i',j')\in\Gamma_\alpha^{bl}} \sum_{t=0}^{r} \sum_{t'=0}^{r} \sum_{c_t=1}^{\binom{k}{t}} \sum_{c_{t'}=1}^{\binom{k}{t'}} P(I_{\alpha c_t} = 1, I_{i'j'c'_{t'}} = 1),$$

$l = 1, 2, 3$, where c_t and $c'_{t'}$ are configurations consisting of t and t' mismatches, respectively. If $\mu = \nu$, then $P(I_{\alpha c_t} = 1, I_{i'j'c'_{t'}} = 1) \leq q_{2(k-r)}$ for all $0 \leq t, t' \leq r$, since $q_a \leq q_b$ if $a \geq b$, and hence, for Γ_α^{b1},

$$\begin{aligned}\sum_{(i',j')\in\Gamma_\alpha^{b1}} E[I_\alpha I_{i'j'}] &\leq \sum_{(i',j')\in\Gamma_\alpha^{b1}} q_{2(k-r)} \sum_{t=0}^{r} \binom{k}{t} \sum_{t'=0}^{r} \binom{k}{t'} \\ &\leq \sum_{(i',j')\in\Gamma_\alpha^{b1}} q_{2(k-r)} \binom{k+r}{r}^2.\end{aligned}$$

The corresponding bounds involving $I_{i'j'}$, $(i', j') \in \Gamma_\alpha^{bl}$, $l = 2, 3$, and for $\mu \neq \nu$ are changed similarly, and hence Lemma 3.1 holds with $\binom{k+r}{r}$ instead of $\binom{k}{r}$.

Since r is constant, the change of the bound in Lemma 3.1 does not affect the magnitude of order of the bounds, so that the rest of the results for exactly r mismatches hold true. ■

Finally, fix $r \geq 0$ and let

$$M = \max\{t : \sum_{l=0}^{t-1} 1\{X_{i+t} = Y_{j+t}\} \geq t - r \text{ for some } 1 \leq i \leq m, 1 \leq j \leq n\},$$

be the length of the longest matching subsequence with at most r mismatches. Then $P(M < k) = P(\tilde{W}^{(k)} = 0)$, where $\tilde{W}^{(k)} = \tilde{W}$, and the above error bounds hold also for $|P(M < k) - \exp\{-\sum_l \tilde{\lambda}_l^{(k)}\}|$.

4 Proofs

4.1 Proof of Lemma 3.1

In order to use Theorem 2.1 to prove Lemma 3.1, we must partition $\Gamma = \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$ in a suitable way. For each $(i,j) \in \Gamma$, let

$$\Gamma_{ij}^{vs} = \{(i',j') \in \Gamma\backslash\{(i,j)\} : -k < i - i' = j - j' < k\},$$

$$\Gamma_{ij}^{b} = \{(i',j') \in \Gamma\backslash\{(i,j) \cup \Gamma_{ij}^{vs}\} : |i - i'| < 2k - 1 \text{ or } |j - j'| < 2k - 1\},$$

and

$$\Gamma_{ij}^{vw} = \{(i',j') \in \Gamma : |i-i'| \geq 2k-1 \text{ and } |j-j'| \geq 2k-1\}.$$

Then $\Gamma = \Gamma_{ij}^{vw} \cup \Gamma_{ij}^{b} \cup \Gamma_{ij}^{vs} \cup \{i,j\}$, and by this partition of Γ we have that the $\{I_{i'j'}, (i',j') \in \Gamma_{ij}^{vw}\}$ are independent of both I_{ij} and the $\{I_{i''j''}, (i'',j'') \in \Gamma_{ij}^{vs}\}$, so that ε_0 in Theorem 2.1 equals zero. We need a finer division of Γ_{ij}^{b}:

$$\Gamma_{ij}^{b1} = \{(i',j') \in \Gamma_{ij}^{b} : |i-i'| < k, |j-j'| < k, i-i' \neq j-j'\},$$

$$\Gamma_{ij}^{b2} = \{(i',j') \in \Gamma_{ij}^{b} : |i-i'| < k, |j-j'| \geq k\},$$

$$\Gamma_{ij}^{b3} = \{(i',j') \in \Gamma_{ij}^{b} : |i-i'| \geq k, |j-j'| < k\},$$

and

$$\Gamma_{ij}^{b4} = \Gamma_{ij}^{b} \backslash \{\cup_{l=1}^{3} \Gamma_{ij}^{bl}\} = \{(i',j') \in \Gamma_{ij}^{b} : |i-i'|, |j-j'| \geq k\}.$$

Let $|\cdot|$ denote the number of elements in a set. Then

$$\begin{aligned}
|\Gamma_{ij}^{vs}| &= 2(k-1), \\
|\Gamma_{ij}^{b}| &= (4k-3)(m+n) - (4k-3)^2 - |\Gamma_{ij}^{vs}| - 1, \\
|\Gamma_{ij}^{b1}| &= (2k-1)^2 - 2(k-1) - 1 = 4k^2 - 6k + 2, \\
|\Gamma_{ij}^{b2}| &= (n-(2k-1))(2k-1), \\
|\Gamma_{ij}^{b3}| &= (m-(2k-1))(2k-1), \\
|\Gamma_{ij}^{b4}| &= 2(k-1)(m+n-6k+4). \qquad (4.22)
\end{aligned}$$

Note that I_{ij} are equally distributed for all $(i,j) \in \Gamma$; since we treat the sequences as circles this is true also for U_{ij}, X_{ij} and $I_{ij}X_{ij}$. Hence, by Theorem 2.1,

$$\begin{aligned}
&\left| E[Wg(W) - \sum_{l \geq 1} l\lambda_l g(W+l)] \right| \\
&\leq \; |\Delta g| mn \left(E[I_\alpha] E[I_\alpha + U_\alpha + X_\alpha] + E[I_\alpha X_\alpha] \right), \qquad (4.23)
\end{aligned}$$

where α is an arbitrary index in Γ.

It follows immediately that

$$E[I_\alpha + U_\alpha + X_\alpha] \;=\; (1 + |\Gamma_\alpha^{vs}| + |\Gamma_\alpha^{b}|) E[I_\alpha], \qquad (4.24)$$

and in order to bound $E[I_\alpha X_\alpha]$, we write

$$E[I_\alpha X_\alpha] \;=\; \sum_{l=1}^{4} \sum_{(i',j') \in \Gamma_\alpha^{bl}} E[I_\alpha I_{i'j'}].$$

The indicators $I_{i'j'}$ with $(i',j') \in \Gamma_\alpha^{b4}$ are independent of I_α, and hence

$$\sum_{(i',j') \in \Gamma_\alpha^{b4}} E[I_\alpha I_{i'j'}] \;=\; E[I_\alpha]^2 |\Gamma_\alpha^{b4}|. \qquad (4.25)$$

Before turning to $E[I_\alpha I_{i'j'}]$, where $(i',j') \in \Gamma_\alpha^{bl}$, $l = 1,2,3$, we collect some simple inequalities in a lemma. Some of these inequalities, and parts of the argument below can be found elsewhere, see e.g. Arratia, Gordon and Waterman (1990) and Novak (1995).

Lemma 4.1

Equal distributions: *(i)* *If* $s \in \{1, 2, \ldots\}$, *then* $p^s \leq q_s$.
(ii) *If* $s, t \in \{1, 2, \ldots\}$, *then* $q_s q_t \leq q_{s+t}$.
(iii) *If* $s \in \{2, 3, \ldots\}$, *then* $q_s < p^{(s+1)/2}$.

Unequal distributions: (iv) $p^2 \leq q_{\mu 2}$ *and* $p^2 \leq q_{\nu 2}$.

Proof. (i) Let Z be a random variable which assumes the value p_a with probability p_a. Then it follows from Jensen's inequality that

$$p^s = (\sum_{a \in A} p_a^2)^s = E[Z]^s \leq E[Z^s] = \sum_{a \in A} p_a^{s+1} = q_s.$$

(ii) Now, let Z assume the value p_a^{s+t} with probability p_a, and let $f(x) = x^{s/(s+t)}$. Then, again by Jensen's inequality,

$$q_s = \sum_{a \in A} (p_a^{s+t})^{s/(s+t)} p_a = E[f(Z)] \leq f(E[Z]) = (\sum_{a \in A} p_a^{s+t+1})^{s/(s+t)} = q_{s+t}^{s/(s+t)},$$

and similarly $q_t \leq q_{s+t}^{t/(s+t)}$, from which the inequality follows.
(iii) Since $(s+1)/2 > 1$ the following inequality holds:

$$p^{(s+1)/2} = (\sum_{a \in A} p_a^2)^{(s+1)/2} > \sum_{a \in A} p_a^{s+1} = q_s.$$

(iv) Follows in a similar way as (i). ■

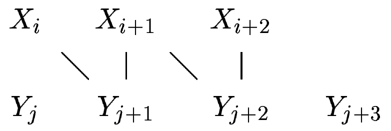
Note that if $I_\alpha = 1$, and $\alpha = (11)$ say, then $X_{i''} = Y_{j''}$ for exactly $k - r$ of the indices $(i'', j'') \in \{(s, s) : s = 1, \ldots, k\}$, while r are mismatches, $X_{i''} \neq Y_{j''}$. These mismatches can occur at $\binom{k}{r}$ different r-subsets of the index pairs, and we number the configurations of matches and mismatches from 1 to $\binom{k}{r}$. Let $I_{\alpha c}$ denote an indicator which equals one if the particular configuration c, $c = 1, \ldots, \binom{k}{r}$, of matches and mismatches take place. Let $I_{i'j'c'}$, $c' = 1, \ldots, \binom{k}{r}$, be the corresponding indicators pertaining to $I_{i'j'}$. Then

$$\sum_{(i',j') \in \Gamma_\alpha^{bl}} E[I_\alpha I_{i'j'}] \quad = \quad \sum_{(i',j') \in \Gamma_\alpha^{bl}} \sum_{c=1}^{\binom{k}{r}} \sum_{c'=1}^{\binom{k}{r}} P(I_{\alpha c} = 1, I_{i'j'c'} = 1). \tag{4.26}$$

First we will consider $P(I_{\alpha c} = 1, I_{i'j'c'} = 1)$ in the case where the sequences are equally distributed. The $2(k-r)$ equalities constitute $1, 2, \ldots,$ or $2(k-r)$ 'chains', depending on where the mismatches occur. This is illustrated by an example.

Example 4.1. Let $k = 3$, $r = 1$, $i' = i$ and $j' = j + 1$. In the figures below, which show two possibilities when $I_{ij} = 1$ and $I_{i'j'} = 1$, the bars correspond to equalities.

- First we let $X_i \neq Y_j$ and $X_{i+2} \neq Y_{j+3}$.



Here $X_i = Y_{j+1} = X_{i+1} = Y_{j+2} = X_{i+2}$, and the equalities constitute one chain.

- Next, we let $X_{i+2} \neq Y_{j+2}$ and $X_i \neq Y_{j+1}$.

$$\begin{tikzcd}
X_i \arrow[d, dash] & X_{i+1} \arrow[d, dash] \arrow[dr, dash] & X_{i+2} \arrow[dr, dash] & \\
Y_j & Y_{j+1} & Y_{j+2} & Y_{j+3}
\end{tikzcd}$$

  In this case $Y_j = X_i$, $Y_{j+1} = X_{i+1} = Y_{j+2}$ and $X_{i+2} = Y_{j+3}$, so that there are three chains of equalities.

□

So, all in all there are $2(k-r)$ equalities which are split into at most $2(k-r)$ chains. Assume that there are s chains, where $1 \leq s \leq 2(k-r)$, in the configurations c and c', and that the number of equalities in the chains are n_i, $i = 1, \ldots, s$, where $\sum_{i=1}^{s} n_i = 2(k-r)$. Then

$$\begin{aligned} P(I_{\alpha c} = 1, I_{i'j'c'} = 1) &\leq q_{n_1} q_{n_2} \cdots q_{n_s} \\ &\leq q_{n_1+n_2+\cdots+n_s} \\ &= q_{2(k-r)}, \end{aligned} \tag{4.27}$$

where the second inequality follows by Lemma 4.1 (ii).

We now turn to the case of unequal distributions. As before there are $2(k-r)$ equalities which split into chains. Assume s of these chains start and end in sequence $\{X_i\}$, and let $m_1, \ldots, m_s$ be the number of equalities these chains involve. Furthermore, assume t of the chains start and end in sequence $\{Y_j\}$, with $n_1, \ldots, n_t$ equalities each. Finally, assume that u of the chains start in $\{X_i\}$ and end in $\{Y_j\}$, or the other way around, with $o_1, \ldots, o_u$ equalities each. Then

$$P(I_{\alpha c} = 1, I_{i'j'c'} = 1) \leq q_{\mu m_1} \cdots q_{\mu m_s} q_{\nu n_1} \cdots q_{\nu n_t} q_{\mu\nu o_1} \cdots q_{\mu\nu o_u}. \tag{4.28}$$

To bound the right-hand side of this inequality, recall that $(\mu\nu)^* = \max_{a \in A}\{\mu_a \nu_a\}$ so that

$$q_{\mu m_i} = \sum_{a \in A} (\mu_a \nu_a)^{m_i/2} \mu_a \leq (\mu\nu)^{*m_i/2},$$

and analogously, $q_{\nu n_i} \leq (\mu\nu)^{*n_i/2}$. Furthermore,

$$q_{\mu\nu o_i} = \sum_{a \in A} (\mu_a \nu_a)^{(o_i+1)/2} \leq (\mu\nu)^{*o_i/2} \sum_{a \in A} (\mu_a \nu_a)^{1/2} \leq (\mu\nu)^{*o_i/2},$$

by using Cauchy-Schwarz inequality in the last step. Since $m_1 + \cdots + m_s + n_1 + \cdots + n_t + o_1 + \cdots + o_u = 2(k-r)$, we get

$$P(I_{\alpha c} = 1, I_{i'j'c'} = 1) \leq (\mu\nu)^{*(k-r)}, \tag{4.29}$$

by (4.28). Combining (4.26), (4.27), (4.29) and letting $Q_{2(k-r)} = q_{2(k-r)}$ in case of equal distributions, and $Q_{2(k-r)} = (\mu\nu)^{*(k-r)}$ otherwise, yields

$$\sum_{(i',j') \in \Gamma_\alpha^{b1}} E[I_\alpha I_{i'j'}] \leq |\Gamma_\alpha^{b1}| \binom{k}{r}^2 Q_{2(k-r)}. \tag{4.30}$$

Next, we show that

$$P(I_{\alpha c}=1, I_{i'j'c'}=1) \leq q_2^{k-r},$$

if $(i',j') \in \Gamma_\alpha^{b2}$, in the case where the sequences are equally distributed. If $I_\alpha = I_{i'j'} = 1$, then there are $2(k-r)$ equalities which split into chains, and (4.27) is valid also in this case. However, here we can do a little better by using the knowledge that the $2(k-r)$ equalities constitute chains with either one or two equalities in each chain, depending on where the mismatches occur. This is because $|j-j'| \geq k$, and is illustrated by the following example.

Example 4.2. Let $k=3$, $r=1$, $i'=i$ and $j'=j+k$, and let the mismatches be $X_i \neq Y_j$ and $X_{i+2} \neq Y_{j+k+2}$. Then we have the situation illustrated by the figure below.

$$\begin{array}{ccc} Y_j & Y_{j+1} & Y_{j+2} \\ & | & | \\ X_i & X_{i+1} & X_{i+2} \\ | & | & \\ Y_{j+3} & Y_{j+4} & Y_{j+5} \end{array}$$

□

Hence, the $2(k-r)$ equalities split into s chains, where $k-r \leq s \leq 2(k-r)$. Let the number of equalities in the chains be n_i, $i=1,\ldots,s$, where $n_i = 1$ or 2, and $\sum_{i=1}^{s} n_i = 2(k-r)$. A chain with n_i equalities has probability q_{n_i}. Since there must be an even number of n_i which equals 1, and $q_1^2 = p^2 \leq q_2$ by Lemma 4.1 (i), it follows that

$$P(I_{\alpha c}=1, I_{i'j'c'}=1) \leq q_{n_1} q_{n_2} \cdots q_{n_s} \quad \leq \quad q_2^{k-r}, \tag{4.31}$$

if $(i',j') \in \Gamma_\alpha^{b2}$. By symmetry, (4.31) holds also for $(i',j') \in \Gamma_\alpha^{b3}$.

In the case of unequal distributions, the bounds for $(i',j') \in \Gamma_\alpha^{b2}$ and $(i',j') \in \Gamma_\alpha^{b3}$ are somewhat different. As in the case of equal distributions, the $2(k-r)$ equalities split into chains with either one or two equalities in each chain. If $(i'j') \in \Gamma_\alpha^{b2}$ the chains with two equalities will consist of one variable from sequence $\{X_i\}$ and two from $\{Y_j\}$, and vice verse if $(i'j') \in \Gamma_\alpha^{b3}$. Using Lemma 4.1 (iv), we then get

$$P(I_{\alpha c}=1, I_{i'j'c'}=1) \quad \leq \quad \begin{cases} q_{\nu 2}^{k-r}, & \text{if } (i'j') \in \Gamma_\alpha^{b2}, \\ q_{\mu 2}^{k-r}, & \text{if } (i'j') \in \Gamma_\alpha^{b3}, \end{cases} \tag{4.32}$$

so that

$$\sum_{(i',j') \in \Gamma_\alpha^{b2} \cup \Gamma_\alpha^{b3}} E[I_\alpha I_{i'j'}] \quad \leq \quad \binom{k}{r}^2 \left\{ |\Gamma_\alpha^{b2}| q_{\nu 2}^{k-r} + |\Gamma_\alpha^{b3}| q_{\mu 2}^{k-r} \right\}, \tag{4.33}$$

where $q_{\mu 2} = q_{\nu 2} = q_2$ if $\mu = \nu$, by (4.26), (4.31), and (4.32).

From the definition of I_α, given in (1.1), it follows that

$$E[I_\alpha] \leq \binom{k}{r} p^{k-r},$$

which together with (4.23), (4.24), (4.25), (4.30) and (4.33) gives

$$\left|E[Wg(W) - \sum_{l\geq 1} l\lambda_l g(W+l)]\right| \leq |\Delta g| mn \binom{k}{r}^2$$
$$\left\{(1 + |\Gamma_\alpha^{vs}| + |\Gamma_\alpha^b| + |\Gamma_\alpha^{b4}|)p^{2(k-r)} + |\Gamma_\alpha^{b1}|Q_{2(k-r)} + |\Gamma_\alpha^{b2}|q_{\nu 2}^{k-r} + |\Gamma_\alpha^{b3}|q_{\mu 2}^{k-r}\right\}. \quad (4.34)$$

Using (4.22) proves Lemma 3.1.

4.2 Proof of Theorem 3.3

Theorem 3.3 concerns the asymptotic behaviour of the total variation bounds as $m, n \to \infty$. The dependence on m, n is not explicitly written out here, but note that $k, W, \lambda_l, \mu_l, \Omega, Z_\alpha$ all depend on m, n.

(i) If $E[W]$ stays bounded away from ∞, then this is true also for $\sum_l \lambda_l$, and the bound in Theorem 3.2 is of order $O(\varepsilon)$. Since $q_{\mu 2}, q_{\nu 2} \geq p^2$ by Lemma 4.1 (iv), Theorem 3.3 (i) follows immediately.

(ii) To prove Theorem 3.3 (ii), we will use Proposition 2.3, and first we will show that $E[W] \to \infty$ implies $\Omega = \sum_l \lambda_l \to \infty$.

Let the pair of subsequences of length k which start in α be referred to as the α-sequence. If there is a match in each end of the α-sequence, and $r+1$ mismatches directly to the left and to the right of the α-sequence, then every $(i'j')$-sequence, $(i'j') \in \Gamma_\alpha^{vs}$, will have at least $r+1$ mismatches. Thus $I_{i'j'} = 0$ for $(i'j') \in \Gamma_\alpha^{vs}$ in that case, and we get

$$\begin{aligned}
&P(Z_\alpha = 1 \mid I_\alpha = 1) \\
&\geq \; P(Z_\alpha = 1 \mid I_\alpha = 1, \text{a match in each end of the } \alpha\text{-sequence}) \frac{\binom{k-r}{2}}{\binom{k}{2}} \\
&\geq \; (1-p)^{2(r+1)} \frac{(k-r)(k-r-1)}{k(k-1)} > C, \qquad (4.35)
\end{aligned}$$

for some $C > 0$ which is independent of k (and hence of m and n). It follows that

$$\begin{aligned}
\Omega &= E[W] \sum_{l\geq 1} \frac{1}{l} P(Z_\alpha = l | I_\alpha = 1) \\
&\geq E[W]C, \qquad (4.36)
\end{aligned}$$

and hence Ω and $E[W]$ are of the same order. Then, if the conditions of Proposition 2.3 are satisfied,

$$\begin{aligned}
d_{TV}(\mathcal{L}(W), \mathrm{CP}(\Lambda)) &\leq O\left(\frac{\varepsilon}{\Omega} + P(W \leq TE[W])\right) \\
&\leq O\left(\frac{k^{r+1}}{p^{k-r}}(mq_{\mu 2}^{k-r} + nq_{\nu 2}^{k-r} + kQ_{2(k-r)}) + P(W \leq TE[W])\right),
\end{aligned}$$

by (3.15) and (4.36).

To show that condition (i) in Proposition 2.3 is satisfied, we need an upper bound on

$$\mu_l \;=\; \frac{E[W]}{l\,\Omega} P(Z_\alpha = l | I_\alpha = 1) \;\leq\; P(Z_\alpha = l | I_\alpha = 1)/C. \tag{4.37}$$

Assume that $I_\alpha = 1$ and $l > r$, and that the r mismatches in the α-sequence all lie in the beginning and/or in the end of the α-sequence. Then the probability of $Z_\alpha = l$ is less than p^{l-1-r} since it is enough with $l - 1 - r$ new matches for a clump of size l to occur. For all other configurations of matches and mismatches in the α-sequence, the probability of $Z_\alpha = l$ given that $I_\alpha = 1$ is smaller. Hence

$$P(Z_\alpha = l | I_\alpha = 1) \;\leq\; \min\{1, p^{l-1-r}\},$$

independently of k, and it follows that condition (i) of Proposition 2.3 is satisfied since for any z such that $1 < z < 1/p$,

$$\sum_{l \geq 1} z^l \mu_l \;\leq\; \sum_{l \geq 1} z^l \min\{1, p^{l-1-r}\} = O(1),$$

by (4.37).

By (4.35) and since $E[W] \geq \Omega$,

$$\mu_1 \;=\; \frac{E[W] P(Z_\alpha = 1 \mid I_\alpha = 1)}{\Omega} \geq C > 0.$$

Thus condition (ii) in Proposition 2.3 is satisfied, with $\nu_1 = C$ and $\nu_j = 0$, $j \geq 2$. Furthermore, condition (iii) is satisfied for large m and n since Ω tends to infinity.

All that remains of the proof of Theorem 3.3 is to bound $P(W \leq TE[W])$, which is easily done by means of Chebyshev's inequality:

$$\begin{aligned} P(W \leq TE[W]) &= P(E[W] - W \geq E[W](1-T)) \\ &\leq P((W - E[W]))^2 \geq E[W]^2(1-T)^2) \\ &\leq \frac{E[W^2] - E[W]^2}{E[W]^2(1-T)^2}. \end{aligned}$$

Recall the definitions of Y_{ij}, X_{ij}, and U_{ij} given above Theorem 2.1, and note that $\sum_{(i',j') \neq (i,j)} I_{i'j'} = W - I_{ij} = Y_{ij} + X_{ij} + U_{ij}$. Thus

$$\begin{aligned} E[W^2] &= E[\sum_{(i,j) \in \Gamma} I_{ij}^2] + E[\sum_{(i,j) \in \Gamma} \sum_{(i',j') \neq (i,j)} I_{ij} I_{i'j'}] \\ &= E[W] + \sum_{(i,j) \in \Gamma} E[I_{ij}(Y_{ij} + X_{ij} + U_{ij})] \\ &= E[W] + mn \left(|\Gamma_\alpha^{vw}| E[I_\alpha]^2 + \sum_{(i',j') \in \Gamma_\alpha^b \cup \Gamma_\alpha^{vs}} E[I_\alpha I_{i'j'}] \right) \\ &\leq E[W] + E[W]^2 + 4mn(m+n)kE[I_\alpha] \\ &= E[W](1 + 4(m+n)k) + E[W]^2, \end{aligned}$$

where the inequality follows since $|\Gamma_\alpha^{vw}| \leq mn$, $|\Gamma_\alpha^b \cup \Gamma_\alpha^{vs}| \leq 4(m+n)k$ by (4.22). Hence

$$\begin{aligned}\frac{E[W^2]-E[W]^2}{E[W]^2(1-T)^2} &\leq \frac{E[W](1+4(m+n)k)}{E[W]^2(1-T)^2} \\ &\leq O\left(\frac{(m+n)k}{E[W]}\right),\end{aligned}$$

which is of smaller order than ε/Ω, and the proof of Theorem 3.3 is concluded. ■

Acknowledgement

The author wishes to thank A. D. Barbour for suggesting the problem and for valuable discussions.

References

Arratia, R., Goldstein, L. and Gordon, L. (1989). Two moments suffice for Poisson approximations: The Chen-Stein method. *Ann. Probab.* **17**, 9–25.

Arratia, R., Gordon, L. and Waterman, M. S. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14**, 971–993.

Arratia, R., Gordon, L. and Waterman, M. S. (1990). The Erdős-Rényi law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18**, 539–570.

Arratia, R. and Waterman, M. S. (1985). Critical phenomena in sequence matching. *Ann. Probab.* **13**, 1236–1249.

Arratia, R. and Waterman, M. S. (1989). The Erdős-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17**, 1152–1169.

Barbour, A.D. (1997). Stein's method. *Encyclopaedia of Statistical Sciences*, Update, Vol. 1, (Eds. S. Kotz, C. B. Read and D. L. Banks), Wiley, New York, 513–521.

Barbour, A.D., Chen, L.H.Y. and Loh, W.-L. (1992). Compound Poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.* **20**, 1843–1866.

Barbour, A.D. and Utev, S. (1998). Solving the Stein Equation in compound Poisson approximation. *Adv. in Appl. Probab.* **30**, 449–475.

Barbour, A.D. and Utev, S. (1999). Compound Poisson approximation in total variation. *Stochastic Process. Appl.*, (to appear).

Barbour, A.D. and Xia A. (1999). Random perturbations. (Preprint)

Dembo, A., Karlin, S. and Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.* **22**, 2022–2039.

Karlin, S. Ost, F. (1988) Maximal length of common words among random letter sequences. *Ann. Probab.* **16**, 535–563.

NEUHAUSER, C. (1996). A phase transition for the distribution of matching blocks. *Combin. Probab. Comput.* **5**, 139–159.

NOVAK, S. Y. (1994). Poisson approximation for the number of long match patterns in random sequences. *Theory Probab. Appl.* **39**, 593–603.

NOVAK, S. Y. (1995). Long match patterns in random sequences. *Siberian Adv. Math.* **5** , 128–140.

ROOS, M. (1994a). Stein's method for compound Poisson approximation: the local approach. *Ann. Appl. Probab.*, **4**, 1177–1187.

ROOS, M. (1994b). Stein–Chen method for compound Poisson approximation: the coupling approach. *Probab. Theory and Math. Stat.*, Proceedings of the Sixth Vilnius Conference, 645–660.