J Stat Phys (2008) 131: 1103–1120
DOI 10.1007/s10955-008-9533-z

Thermodynamical Approach to the Longest Common Subsequence Problem

Saba Amsalu · Heinrich Matzinger · Marina Vachkovskaia

Received: 14 October 2007 / Accepted: 20 March 2008 / Published online: 11 April 2008

Abstract We introduce an interacting particle model in a random media and show that this particle process is equivalent to the Longest Common Subsequence (LCS) problem of two binary sequences. We derive a differential equation which links the mean LCS-curve to the average speed of the particles given their density and prove that the average speed of the particles and density converges uniformly on every scale which is somewhat larger than $\sqrt{n}$.

Keywords Longest common subsequence · Interacting particle systems · Optimal sequence alignment

1 Introduction

A common subsequence of two strings S_1 and S_2 is a sequence which is a subsequence of S_1 as well as of S_2. A Longest Common Subsequence (LCS) of S_1 and S_2 is a common subsequence of S_1 and S_2 of maximal length.

Let us give a numerical example. Take the two strings $S_1 = alaman$ and $S_2 = allemand$. Both of these string derive from the name of a Germanic tribe of the Allamans. (S_1 is a small town close to Lausanne. The second string means *German* in French.) The LCS of S_1 and S_2

All three authors are grateful to SFB 701. M.V. is grateful to CNPq (304561/2006–1 and 471925/2006–3) and FAPESP (thematic grant 04/07276–2) for partial support.

S. Amsalu
University of Bielefeld, Postfach 10 01 31, 33501 Bielefeld, Germany
e-mail: saba@uni-bielefeld.de

H. Matzinger
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332–0160, USA
e-mail: matzi@math.gatech.edu

M. Vachkovskaia (✉)
Instituto de Matemática, Estatística e Computação Científica, Universidade de Campinas, Caixa Postal 6065, CEP 13083–970, Campinas SP, Brasil
e-mail: marinav@ime.unicamp.br

is given by *alman*. The length of the LCS is relatively long which indicates a high degree of similarity between S_1 and S_2. We can represent the LCS *alman* by an alignment with gaps. We align all the letters which appear in the LCS whilst the other letters get aligned with gaps:

a	l			a	m	a	n	
a	l	l	e		m	a	n	d

LCS are thus used to recognize similar strings. The method can also be modified a little. Instead of looking for the same subsequence in S_1 and S_2 one could search for similar subsequences. This is done by a scoring function which associates with each pair of letters a score where a high score indicates a high similarity between letters. One looks then for an alignment which maximizes the total score. The total score is obtained as the sum of the scores of the aligned pairs of letters minus a penalty for the total number of gaps. The alignment we find in this way is called *optimal alignment*.

LCS and optimal alignments are some of the main methods used in modern computation biology as well as computational linguistics. For an overview of the former topic see [21, 23]. Despite the practical importance of the field, many mathematical question have not yet been answered rigorously.

Let S_1 and S_2 to be two binary i.i.d. strings of length n independent of each other. Let L_n designate the length of the LCS of S_1 and S_2. A simple subadditivity argument was used by Chvatal and Sankoff [6] to prove that $\mathbf{E}[L_n]/n$ converges as n goes to infinity. However, the exact value of the limit remains unknown up to this day. We can also consider strings of different lengths. Fix $p \in [0.5, 2]$. If S_1 has length np and S_2 has length n, we denote by $L_n(p)$ the length of LCS of S_1 and S_2. Again assuming that they are i.i.d. binary strings independent of each other, the same subadditivity argument implies that $\mathbf{E}[L_n(p)]/n$ converges. Let us designate the limit by $\varphi(p)$:

$$\varphi(p) := \lim_{n\to\infty} \frac{\mathbf{E}[L_n(p)]}{n}.$$

Again for non-trivial values of p the exact value of $\varphi(p)$ is not known.

The LCS problem can be viewed as a last passage percolation problem with correlated weights. For this let $f : \{0, 1\}^2 \to \{0, 1\}$ designate the map such that

$$f(0,0) = f(1,1) = 1, \qquad f(0,1) = f(1,0) = 0.$$

Now our last passage percolation goes as follows: we look for a path π on $\mathbb{N} \times \mathbb{N}$ which goes from the origin to the point $(n, n) \in \mathbb{N}$. Assume that at each step, π moves one unit to the right or one unit up or diagonally to the right and up by one unit at the same time. The vertical and horizontal edges have weight 0 whilst the diagonal edges $((i, j), (i + 1, j + 1))$ have weight

$$w(((i, j), (i + 1, j + 1))) := f(X_{i+1}, Y_{j+1}).$$

Then the path with maximum weight corresponds to the LCS of X and Y. The maximum weight gives the length of the LCS. In the last passage percolation language, the curve $p \to \varphi(p)$ corresponds to the rescaled shape of the wet region.

In this article we introduce an interacting particle system in random media. In Sect. 2, we explain heuristically that the function which gives the average speed of the particles given the density uniquely determines the curve $\varphi(p)$. In Sect. 3 we present the main result, relating the average speed of the particles and their density to the LCS problem. In Sect. 4.1

we give an heuristic argument leading to this result, and in Sect. 4.2 we provide a rigorous proof.

For the limit we let the text length go to infinity. The speed converges if we always look at the same proportion in the text. Having established the equivalence between the LCS-problem and our interacting particle problem could prove useful for investigating the LCS-problem more thoroughly: any bound on the speed of particles can now be translated in an inequality for the mean LCS length. Connections between similar problems and interacting particles systems have proven useful in other settings, see [1, 18, 19].

Let us mention a little bit more about the history of this problem:

As already mentioned, Chvatal and Sankoff [6] prove that the limit

$$\gamma := \lim_{n\to\infty} \frac{\mathbf{E}L_n}{n}$$

exists. The exact value of γ remains however unknown. Chvatal and Sankoff [6] derive upper and lower bounds for γ, and similar upper bounds were found by Baeza-Yates, Gavalda, Navarro and Scheihing [5] using an entropy argument. These bounds have been improved by Deken [8], and subsequently by Dancik and Paterson [7, 17]. For sequences with many equiprobable letters, Kiwi, Loebl and Matousek [12] determined the asymptotic value of γ.

Arratia and Waterman [4] derive a law of large deviation for L_n for fluctuations on scales larger than $\sqrt{n}$. This is very useful and shows that in no case the fluctuation of the optimal score can be larger than order $\sqrt{n}$.

Using the first passage percolation methods, Alexander [2] proves that $\mathbf{E}L_n/n$ converges at a rate of order $\sqrt{\log n/n}$. As mentioned, a long open standing problem in the LCS context is to determine the exact order of the fluctuation of the LCS length. Steele [20] proved that $\operatorname{Var} L_n \leq n$. Monte Carlo simulations led Chvatal and Sankoff [6] to conjecture that $\operatorname{Var} L_n = O(n^{2/3})$. Waterman [22] conjectured that for i.i.d. sequences the variance of L_n grows linearly.

Matzinger and Lember [14] determined the order of magnitude of the fluctuation of L_n when one sequence is not random but periodic. They were also able to determine [13] that the order of $\operatorname{Var}[L_n]$ is equal to n when the binary i.i.d. sequences are such that one and zero have very different probabilities. Houdre, Lember and Matzinger [11] determined the asymptotic distribution of the Longest Increasing Common Subsequence of two i.i.d. sequences of length n. Amsalu, Matzinger and Popov [3] discovered a fundamental relationship between transversal fluctuation and macroscopic uniqueness. Several of the previous mentioned papers are based on a combinatorial approach using *m-matches* and large deviation developed by Martinez, Matzinger and Hauser [10] and by Matzinger, Hauser and Durringer [9]. In these papers the authors present a Monte Carlo based method to bound the constant γ and the curve $p \to \varphi(p)$.

2 Particle Process and LCS

2.1 The Particle Process

We now describe the interacting particle process on a random media. Time is discrete. The media we consider is denoted by

$$X : \mathbb{N} \to \{0, 1\}; \qquad s \mapsto X_s.$$

The particles are located on $\mathbb{N} = \{1, 2, \ldots\}$. There can never be two particles on the same location. Particles can only move to the left (but not necessarily to the nearest neighbour) and can not jump over one another. Also, new particles can appear. At every step, to decide how the particles move, we flip a fair coin. One side of the coin is denoted by 0, the other by 1. The same coin flip is used for all the particles at the same time. The particles move to a position in the media which has the same value as the outcome of our flip in the following way. Suppose that at time t we have some particles and the outcome of the $t+1$ coin flip is Y_{t+1}. Then each particle looks at the current configuration and chooses for itself the leftmost available position which has the value Y_{t+1} and then all the particles jump simultaneously to the chosen positions. A new particle appears at the leftmost position which has the value Y_{t+1} and no particles to the right.

Let us look at a numerical example. Take $X = 10000011\ldots$. Hence

X_i	1	0	0	0	0	0	1	1	...
i	1	2	3	4	5	6	7	8	...

Assume now that at time t, we have particles distributed as follows on the media X:

•			•		•		•	...
1	0	0	0	0	0	1	1	...

where an "•" designates a position occupied by a particle. Hence we have a particle in 1, 4, 6, 8 at time t. Assume now that $Y_{t+1} = 0$. This means that all the particles which are going to move must do so to a point with color 0, thus at time $t+1$ the configuration of our particles will be

•	•			•			•	...
1	0	0	0	0	0	1	1	...

Let us now define formally the particle process. Let $\eta_i(t) \in \mathbb{N} \cup \{+\infty\}$ be the position of i-th particle at time t. For convenience we suppose that there is an infinite reservoir with particles at $+\infty$. The initial configuration is $\eta_i(0) = +\infty$ for all i (i.e., there are no particles in $\mathbb{N}$). Assume that $\eta_i(t) = s_1$, $\eta_{i+1}(t) = s_2$, where $s_1 < s_2 \leq +\infty$. That is, at time t there is a particle at s_1 and one at s_2, but no particle in between; if $s_2 = +\infty$, then there are no particles to the right of s_1. Let $a \in \{0, 1\}$. Suppose that when we flip the coin for the $(t+1)$-th time we obtain $Y_{t+1} = a$. Then, we have:

– If for all s such that $s_1 < s < s_2$, we have $X_s \neq a$, then $\eta_{i+1}(t+1) = s_2$ (at time $t+1$, the particle configuration on $[s_1 + 1, s_2]$ is the same as at time t).
– If there exist s such that $s_1 < s < s_2$ and $X_s = a$, then $\eta_{i+1}(t+1) = s^{(a)}$, where

$$s^{(a)} = \min_{s_1 < s < s_2} \{s : X_s = a\}. \tag{2.1}$$

Note that for nontrivial $X_1 X_2 \ldots$ at time t we have t particles in $\mathbb{N}$.

2.2 Connection Between LCS and the Particle Process

Let $X_1 X_2 \ldots$ and $Y_1 Y_2 \ldots$ be i.i.d. binary sequences independent from each other. Denote by

$$\mathcal{L}(X_1 X_2 \ldots X_s; Y_1 Y_2 \ldots Y_t)$$

the length of the LCS of $X_1X_2\dots X_s$ and $Y_1Y_2\dots Y_t$ and let

$$F_t(s)=\mathcal{L}(X_1X_2\dots X_s;Y_1Y_2\dots Y_t).$$

For every $s\in\mathbb{N}$ such that

$$F_t(s)>F_t(s-1),\quad \text{that is,}\quad F_t(s)=F_t(s-1)+1,$$

we say that there is a particle located in point s at time t. Note that with this definition $F_t(s)$ is equal to the number of particles located in the interval $[0,s]$ at time t. In this section we show that these particles follow the same motion mechanism as the particles described in the previous section. Take $X=X_1X_2\dots$ as random media on which the particles move and Y_{t+1} is the coin we flip to decide what the configuration at time $t+1$ will be. Let us first show an example before we explain why the particle process defined by $F_t(s)$ has the same evolution mechanism than the particle process in the previous section.

Let $X_1=0$, $X_2=0$, $X_3=1$ and $X_4=1$. Hence, $X=0011\dots$. Assume that $Y_1=0$, $Y_2=1$, $Y_3=1$ and $Y_4=1$. Then:

$$\begin{aligned}\mathcal{L}(\emptyset,Y_1Y_2Y_3Y_4)&=F_4(0)=0,\\ \mathcal{L}(X_1,Y_1Y_2Y_3Y_4)&=F_4(1)=1,\\ \mathcal{L}(X_1X_2,Y_1Y_2Y_3Y_4)&=F_4(2)=1,\\ \mathcal{L}(X_1X_2X_3,Y_1Y_2Y_3Y_4)&=F_4(3)=2,\\ \mathcal{L}(X_1X_2X_3X_4,Y_1Y_2Y_3Y_4)&=F_4(4)=3.\end{aligned}$$

We see that the map $s\mapsto F_4(s)$ has a point of increase in $s=1,3,4$. Hence there is a particle in point $s=1$, $s=3$, $s=4$ whilst the point $s=2$ is empty at time $t=4$.

Hence at time $t=4$ the particle configuration we have is

•		•	•
0	0	1	1

Assume now that $Y_5=0$. First look at the particle configuration. The new configuration (according to the rules for the dynamics of the particles in the previous section) at time 5 is thus:

•	•		•
0	0	1	1

It is immediate to check that the map $s\mapsto F_5(s)$ has points of increase in $s=1,2,4$. Hence, at time $t=5$ there are particles in point 1, 2 and 4. Note that this is the particle configuration we found using the particle dynamics mechanism. See Fig. 1 for the dynamics for $t\leq 5$.

Let us now give a rigorous proof of the fact that the particle process defined via the function $F_t(s)$ follows the dynamics described in the previous section.

Lemma 2.1 *Suppose that we have constructed the particle configuration $\eta_i(t)$, $i=1,2\dots$ for $X=X_1\dots X_s$ and $Y=Y_1\dots Y_t$ (configuration at time t). Then, the new particle configuration for $X=X_1\dots X_s$ and $Y'=Y_1\dots Y_tY_{t+1}$, where $Y_{t+1}=a\in\{0,1\}$, is obtained from the previous one in the following way*:

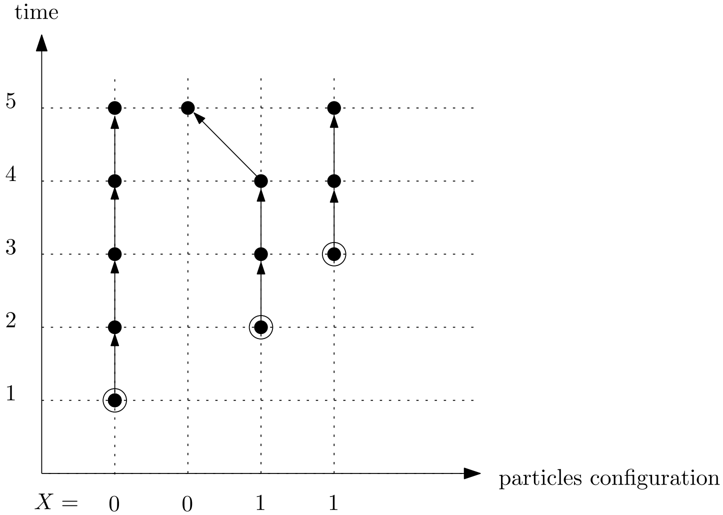

Fig. 1 Particles trajectories for $X = 0011\ldots$ and $Y = 01110\ldots$. The double circles correspond to the new particles

- *If there exists* s, $0 < s < \eta_1(t)$, *such that* $X_s = a$, *then*

$$\eta_1(t+1) = s_1^{(a)},$$

where

$$s_1^{(a)} = \min_{0<s<\eta_1(t)} \{s : X_s = a\}.$$

Otherwise, $\eta_1(t+1) = \eta_1(t)$.
- *If there exists* s, $\eta_i(t) < s < \eta_{i+1}(t)$, *such that* $X_s = a$, *then*

$$\eta_{i+1}(t+1) = s_{i+1}^{(a)},$$

where

$$s_{i+1}^{(a)} = \min_{\eta_i(t)<s<\eta_{i+1}(t)} \{s : X_s = a\}.$$

Otherwise, $\eta_{i+1}(t+1) = \eta_{i+1}(t)$.

Proof Suppose that $\eta_i = l$, $\eta_{i+1} = l'$, $X_k = a$ for some k, $l < k < l'$, and $X_m \neq a$ for all m, $l < m < k$. We want to show that in this case the particle which was at l' will jump to k, when we add a at the end of sequence Y. We look at the LCS as an optimal alignment of two sequences, where letters corresponding to the LCS are aligned.

As there are no particles between l and k, there exists an optimal alignment (LCS) of $X_1 \ldots X_k$ and $Y_1 \ldots Y_t$ which uses only $X_1 \ldots X_l$ part of $X_1 \ldots X_k$,

$$\mathcal{L}(X_1 \ldots X_k, Y_1 \ldots Y_t) = \mathcal{L}(X_1 \ldots X_l, Y_1 \ldots Y_t). \tag{2.2}$$

That is, this alignment does not use X_k. So, we can construct an optimal alignment of $X_1 \ldots X_k$ and $Y_1 \ldots Y_t a$ by taking the previous alignment of $X_1 \ldots X_l$ and $Y_1 \ldots Y_t$ and aligning $X_k = a$ with $Y_{t+1} = a$. This way, we get

$$\mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t a) = \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t) + 1. \tag{2.3}$$

Take some m, $l < m < k$. As there were no particles between l and l', and $X(m) \neq a$ for all m, $l < m < k$, we have

$$\begin{aligned}\mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t a) &= \mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t)\\ &= \mathcal{L}(X_1 \dots X_l, Y_1 \dots Y_t)\\ &= \mathcal{L}(X_1 \dots X_{m-1}, Y_1 \dots Y_t a),\end{aligned} \tag{2.4}$$

by the same argument that leads to (2.2). So, there will be no particles at positions $l + 1, \dots, k - 1$. Also, by (2.2), (2.3) and (2.4),

$$\begin{aligned}\mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t a) &= \mathcal{L}(X_1 \dots X_l, Y_1 \dots Y_t) + 1\\ &= \mathcal{L}(X_1 \dots X_{k-1}, Y_1 \dots Y_t a) + 1,\end{aligned}$$

and thus there will be a particle at position k. (Note that the same argument applies to the new particle at the leftmost position corresponding to letter a in the sequence X such that there are no particles to the right of this position.)

It remains to prove that there will be no particles at $k + 1, \dots, l'$. Suppose that we have a particle at m, $m < l'$. Then we have

$$\mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t a) > \mathcal{L}(X_1 \dots X_{m-1}, Y_1 \dots Y_t a).$$

Suppose first that there exists an optimal alignment of $X_1 \dots X_m$ and $Y_1 \dots Y_t Y_{t+1}$ in which X_m is not aligned to $Y_{t+1} = a$. Then we have

$$\mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t a) = \mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t),$$

thus,

$$\begin{aligned}\mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t) &> \mathcal{L}(X_1 \dots X_{m-1}, Y_1 \dots Y_t a)\\ &\geq \mathcal{L}(X_1 \dots X_{m-1}, Y_1 \dots Y_t),\end{aligned}$$

which contradicts the assumption that at time t there was no particle at m.

So, we should have $X(m) = a$ and $X(m)$ should be aligned to Y_{t+1} for any optimal alignment of $X_1 \dots X_m$ and $Y_1 \dots Y_t Y_{t+1}$. In this case, we have

$$\begin{aligned}\mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t a) &= \mathcal{L}(X_1 \dots X_m, Y_1 \dots Y_t) + 1\\ &= \mathcal{L}(X_1 \dots X_l, Y_1 \dots Y_t) + 1\\ &= \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t a)\end{aligned}$$

analogously to (2.2) and (2.3). Thus, there is in fact no particle at m. Consider now $m = l'$ and suppose that there is a particle at l'. Suppose again that there exists an optimal alignment of $X_1 \dots X_{l'}$ and $Y_1 \dots Y_t Y_{t+1}$ in which $X_{l'}$ is not aligned to $Y_{t+1} = a$. Then,

$$\mathcal{L}(X_1 \dots X_{l'}, Y_1 \dots Y_t a) = \mathcal{L}(X_1 \dots X_{l'}, Y_1 \dots Y_t)$$

$$\begin{aligned}&= \mathcal{L}(X_1 \dots X_{l'-1}, Y_1 \dots Y_t) + 1\\&= \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t) + 1\\&= \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t a)\end{aligned}$$

which means that there is no particle at l'. If $X_{l'}$ is aligned to Y_{t+1} for any optimal alignment of $X_1 \dots X_{l'}$ and $Y_1 \dots Y_t Y_{t+1}$, we get again

$$\begin{aligned}\mathcal{L}(X_1 \dots X_{l'}, Y_1 \dots Y_t a) &= \mathcal{L}(X_1 \dots X_{l'-1}, Y_1 \dots Y_t) + 1\\&= \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t) + 1\\&= \mathcal{L}(X_1 \dots X_k, Y_1 \dots Y_t a).\end{aligned}$$

Thus, there is no particle at l' and Lemma 2.1 is proved. □

3 Main Result

Let us first define the *speed* (or rather *average speed*) V. Let

$$s = \max\{\eta_i(n) : \eta_i(n + n\Delta) \le np\},$$

that is, s is the position of the rightmost particle at time $t = n$ which reaches $[0, np]$ within the time interval $[n, n + n\Delta]$. Let

$$V = V(p, \Delta, n) := \frac{s - np}{n\Delta}.$$

Similarly we define the *average density* ϱ_a to be the number of particles in the interval $[np, s]$ at time $t = n$ divided by $s - np$. Noting that by definition $s = np + Vn\Delta$, we obtain

$$\varrho_a = \varrho_a(p, \Delta, n) := \frac{F_n(np + Vn\Delta) - F_n(np)}{Vn\Delta}. \tag{3.1}$$

It is easy to see that for $p \in [0, 1/2]$ we have $\varphi(p) = p$, and for $p \in [2, \infty[$ we have that $\varphi(p) = 1$. (Indeed, for such values of p, the length of one sequence is at most half of the length of the other, and thus, in the limit, the proportion of the symbols of smaller sequence used in the LCS will be 1. This gives $\varphi(p) = p$ for $p \in [0, 1/2]$, and $\varphi(p) = 1$ for $p \in [2, \infty[$.) This means that in the interval $[0, n/2]$ at time $t = n$ the particles are dense and there are no holes. Hence the speed of the particles in that region is zero. Above the point $2n$ the particles are extremely rare, so if a particle would be there it would have a lot of free space to move and its speed might not be bounded above. It seems however reasonable to assume that for any p_1, p_2 such that

$$0.5 < p_1 < p_2 < 2 \tag{3.2}$$

and every β_1, β_2 such that

$$0 < \beta_1 < \beta_2 < 0.5, \tag{3.3}$$

there exists a compact interval $[c_1, c_2]$, $0 < c_1 < c_2 < \infty$, where c_1 and c_2 are constants not depending on p or β such that for all n large enough, for all $p \in [p_1, p_2]$ and for all $\beta \in [\beta_1, \beta_2]$, taking $\Delta = n^{\beta - 0.5}$, we have a.s. $V(p, \Delta, n) \in [c_1, c_2]$.

Also, it is generally believed (see e.g. the discussion in [3]) that the function $\varphi(\cdot)$ should be "well-behaved". On the other hand, it seems to be extremely difficult task to prove rigorously something concrete about the differentiability properties of this function; this situation is quite usual in the problems related to percolation, where one usually easily obtains the existence of critical curves (surfaces) in the parameter space, but the fine analysis of these objects can be almost unreachable (see e.g. [16]).

In view of the above discussion, it is reasonable to assume the following. Let $[p_1, p_2] \subset]0.5, 2[$, $[\beta_1, \beta_2] \subset]0, 0.5[$ and $\Delta = n^{\beta-0.5}$.

(A1) Suppose that φ is twice continuously differentiable on an open interval containing $[p_1, p_2]$ and φ as well as φ' are both bounded away from zero on that interval.
(A2) Suppose that there exists c_1, c_2, $0 < c_1 < c_2 < \infty$, such that for all n larger than some (random) n' we have

$$V(p, \Delta, n) \in [c_1, c_2]. \tag{3.4}$$

Theorem 3.1 *Assume* (A1) *and* (A2). *Then, there exists* (*random*) n_0 *such that for all* $n \geq n_0$, *for all* $p \in [p_1, p_2]$ *and all* $\beta \in [\beta_1, \beta_2]$ *we have a.s.*

$$|\varrho_a(p, \Delta, n) - \varphi'(p)| \leq Cn^{\alpha}\sqrt{\log n} \tag{3.5}$$

and

$$\left| V(p, \Delta, n) - \left(\frac{\varphi(p)}{\varphi'(p)} - p \right) \right| \leq Cn^{\alpha}\sqrt{\log n}, \tag{3.6}$$

where $\alpha := \max\{\beta - 0.5, -\beta\}$ (*note that* $\alpha < 0$), $\Delta = n^{\beta-0.5}$, *and* $C > 0$ *is a constant not depending on* n.

4 Proof of the Main Result

4.1 Heuristics

4.1.1 Fundamental Equality

In this section, we heuristically derive the fundamental differential equation which links $\varphi(p)$ to the speed of the particles. We assume here that the speed of the particles in a certain region converges as n goes to infinity. This makes the argument more intuitive. In the Sect. 4.2, we prove the result rigorously without making this assumption.

Consider the following intuitive reasoning. When particles flow through a conductor with speed V, how many have passed a given point after time Δt? Assuming that the particle density is ϱ and that the speed V is constant, the answer is $\varrho V \Delta t$.

We are going to use the same reasoning applied to our situation.

We consider by how much that number of particles in the interval $[0, np]$ increases between time t and $t + \Delta t$. (Here p is a positive constant which does not depend on n.) Assume that in the neighborhood of point np at time t the density of particles is ϱ and the average speed is V, we find that during the time interval $[t, t + \Delta t]$ there are $\varrho V \Delta t$ particles entering the interval $[0, np]$. Hence

$$F_t(np) + \varrho V \Delta t = F_{t+\Delta t}(np). \tag{4.1}$$

We replace t by n and Δt by $\Delta \cdot n$, before dividing (4.1) by n. This yields:

$$\frac{F_n(np)}{n} + \varrho V \Delta = (1+\Delta) \frac{F_{n+\Delta n}((n+\Delta n)\frac{p}{1+\Delta})}{n+\Delta n}. \tag{4.2}$$

As mentioned it is known that $F_t(tp)/t$ converges as t goes to infinity and we denote the limit by $\varphi(p)$:

$$\varphi(p) = \lim_{t\to\infty} \frac{F_t(tp)}{t}.$$

Thus, letting t go to infinity whilst leaving Δ fixed, we have:

$$\lim_{n\to\infty} \frac{F_n(np)}{n} = \varphi(p) \tag{4.3}$$

and

$$\lim_{n\to\infty} \frac{F_{n+\Delta n}((n+\Delta n)\frac{p}{1+\Delta})}{n+\Delta n} = \varphi\left(\frac{p}{1+\Delta}\right).$$

Using the above limits and letting n go to infinity in (4.2), we find

$$\varphi(p) + \lim_{n\to\infty} \varrho V \Delta = (1+\Delta)\varphi\left(\frac{p}{1+\Delta}\right)$$

from which we obtain

$$\frac{\varphi(p) - \varphi(\frac{p}{1+\Delta})}{\Delta} + \lim_{n\to\infty} \varrho V = \varphi\left(\frac{p}{1+\Delta}\right). \tag{4.4}$$

Assuming that φ is differentiable and letting Δ go to zero, equality (4.4) becomes

$$p\varphi'(p) + \lim_{\Delta\to 0}\lim_{n\to\infty} \varrho V = \varphi(p). \tag{4.5}$$

It only remains to see what happens to the average particle density ϱ at the limit. Note for this that the average particle density is simply the number of particles divided by the length of the interval. The interval we consider is $[np, np + V\Delta n]$. The length of this interval is $V\Delta n$ whilst the number of particles in this interval at time $t = n$ is equal to $F_n(np+V\Delta n) - F_n(np)$. This implies that the average particle density ϱ of this interval at time n is equal to

$$\varrho(n, \Delta, p) = \frac{F_n(np + V\Delta n) - F_n(np)}{V\Delta n}. \tag{4.6}$$

Letting n go to infinity, we find that

$$\lim_{n\to\infty} \frac{F_n(np + V\Delta n)}{n} = \varphi(p + V\Delta). \tag{4.7}$$

Using (4.7) and (4.3) in (4.6), we find:

$$\lim_{n\to\infty} \varrho(n, \Delta, p) = \frac{\varphi(p + V\Delta) - \varphi(p)}{V\Delta}.$$

Assuming that φ has a derivative, the last expression above goes to $\varphi'(p)$ as Δ goes to zero. Hence:

$$\lim_{\Delta\to\infty}\lim_{n\to\infty}\varrho(n,\Delta,p)=\varphi'(p).$$

We find thus an interpretation for the derivative of $\varphi(p)$: the particle density at the limit. Going back to (4.5) and plugging in our expression for the limit of ϱ we get

$$p\varphi'(p)+\varphi'(p)V=\varphi(p) \tag{4.8}$$

which we can also write as

$$p\varphi'(p)+\varphi'(p)V(p)=\varphi(p). \tag{4.9}$$

In this derivation of the fundamental equality (4.8), the only thing not rigorous is that we treated the speed of the particles $V(p)$ as a constant not depending on n and ϱ.

4.1.2 Velocity/Density Map Determines Expected LCS-Curve

In this section we describe heuristically the ideas behind the proof of the fact that if we know the velocity V as a function of the density ϱ, then this uniquely determines the map $\varphi(p)$.

We assume for the moment that the density at the limit is a function of p and that this function is a bijection. We can then express the velocity of the particles near np in terms of the density. Let W denote the speed as function of the density ϱ. Hence, $W(\varrho):=V(p(\varrho))$ and hence $W=V\circ\varrho^{-1}$, where ϱ^{-1} denotes the inverse function of $p\to\varrho(p)$. Equation (4.8) becomes

$$p\varphi'(p)+\varphi'(p)W(\varrho(p))=\varphi(p).$$

Taking the derivative with respect to p in the last equation above yields

$$p\varphi''+\varphi''W(\varrho)+\varphi'\varrho'W'(\varrho)=0.$$

Recall that $\varphi'=\varrho$ and hence the last equation above can be written as

$$p\varphi''+\varphi''W(\varrho)+\varrho\varphi''W'(\varrho)=0.$$

Assuming φ'' bounded from 0, we have that

$$p+W(\varrho(p))+\varrho(p)W'(\varrho(p))=0. \tag{4.10}$$

If we define the map G as follows:

$$G(s):=W(s)+sW'(s)$$

and assuming that W is invertible, (4.10) yields

$$\varrho(p)=G^{-1}(-p) \tag{4.11}$$

which finally gives (assuming that $G^{-1}(-s)$ is integrable)

$$\varphi(p)=\int_0^p G^{-1}(-s)ds. \tag{4.12}$$

The last equation allows us to obtain φ from the map W.

4.2 Proof of Theorem 3.1

In the previous section, we derived heuristically the fundamental equality (4.8). It was not a rigorous proof because we implicitly assumed the speed to converge when we hold Δ fixed and let n go to infinity. In this section we present a formal proof. By our assumption, φ is differentiable, let designate by ϱ the derivative of φ:

$$\varrho(p) = \varphi'(p).$$

(At this stage we don't make any assumption that $\varrho(p)$ represents the particle density at the limit.) In order to simplify notation, we write V for $V(p, \Delta, n)$, ϱ_a for $\varrho_a(p, \Delta, n)$ and ϱ for $\varrho(p)$.

In what follows we will need to define a certain number of approximation errors (in what follows, all error terms depend on n, Δ, p). Let us define them at this stage:

Let ε_1, ε_2, ε_3, ε_4, ε_5, ε_6 be defined by:

$$\varepsilon_1 := \frac{F_n(np)}{n} - \varphi(p), \tag{4.13}$$

$$\varepsilon_2 := \frac{F_{n+\Delta n}((n+\Delta n)\frac{p}{1+\Delta})}{n+\Delta n} - \varphi\left(\frac{p}{1+\Delta}\right), \tag{4.14}$$

$$\varepsilon_3 := \frac{F_n(np + V\Delta n)}{n} - \varphi(p + V\Delta), \tag{4.15}$$

$$\varepsilon_4 := \frac{\varphi(p) - \varphi\left(\frac{p}{1+\Delta}\right)}{\Delta} - p\varphi'(p), \tag{4.16}$$

$$\varepsilon_5 := \varphi\left(\frac{p}{1+\Delta}\right) - \varphi(p), \tag{4.17}$$

$$\varepsilon_6 := \frac{\varphi(p + \Delta V) - \varphi(p)}{\Delta V} - \varphi'(p). \tag{4.18}$$

The same argument which leads to equality (4.2) applied to the current definition of ϱ_a and V yields:

$$\frac{F_n(np)}{n} + \varrho_a V\Delta = (1+\Delta)\frac{F_{n+\Delta n}((n+\Delta n)\frac{p}{1+\Delta})}{n+\Delta n}.$$

Using (4.13) and (4.14) in the previous equation, we find:

$$\varphi(p) + \varepsilon_1 + \varrho_a V\Delta = (1+\Delta)\left(\varphi\left(\frac{p}{1+\Delta}\right) + \varepsilon_2\right),$$

which yields

$$\frac{\varphi(p) - \varphi(\frac{p}{1+\Delta})}{\Delta} + \varrho_a V = \varphi\left(\frac{p}{1+\Delta}\right) + \varepsilon_2 + \frac{\varepsilon_2}{\Delta} - \frac{\varepsilon_1}{\Delta}.$$

With the help of (4.16) and (4.17) the last equality above becomes:

$$p\varphi'(p) + \varrho_a V = \varphi(p) + \varepsilon_2 + \frac{\varepsilon_2}{\Delta} - \frac{\varepsilon_1}{\Delta} - \varepsilon_4 + \varepsilon_5.$$

Let ε be equal to

$$\varepsilon = \varepsilon_2 + \frac{\varepsilon_2}{\Delta} - \frac{\varepsilon_1}{\Delta} - \varepsilon_4 + \varepsilon_5, \tag{4.19}$$

so that we get:

$$p\varphi'(p) + \varrho_a V = \varphi(p) + \varepsilon. \tag{4.20}$$

Applying (4.13) and (4.15) to the definition (3.1) of ϱ_a, we find

$$\varrho_a = \frac{\varphi(p + V\Delta) - \varphi(p)}{V\Delta} + \frac{\varepsilon_3}{V\Delta} - \frac{\varepsilon_1}{V\Delta}$$

and hence with the help of (4.18) we get

$$\varrho_a = \varphi'(p) + \varepsilon_6 + \frac{\varepsilon_3}{V\Delta} - \frac{\varepsilon_1}{V\Delta}. \tag{4.21}$$

Let ε_a denote the error-term:

$$\varepsilon_a = \varepsilon_6 + \frac{\varepsilon_3}{V\Delta} - \frac{\varepsilon_1}{V\Delta}$$

so that

$$\varrho_a = \varphi'(p) + \varepsilon_a. \tag{4.22}$$

Note that when ε_a goes to zero then ϱ_a goes to $\varphi'(p)$. When ε also converges and $\varphi'(p) \neq 0$, then it follows from (4.20) that $V(p, \Delta, n)$ converges. Hence, the limit $V(p)$ satisfies then (4.9). It remains thus to investigate when ε and ε_a converge.

We are going to show that, by a concentration inequality and a speed of convergence theorem, we have that ε_1, ε_2 and ε_3 are approximately of order $1/\sqrt{n}$. For ε_3 to be of that order we also need a constant upper bound on V.

Assuming that φ is twice continuously differentiable, we find that, as Δ goes to zero, the error terms ε_4, ε_5 and ε_6 are of order Δ. To prove that ε_6 is of order Δ we also need a constant upper bound on V.

Looking at the expressions for ε and ε_a,we see thus that for convergence we need ε_i/Δ for $i = 1, 2, 3$ to converge to zero and V to be bounded away from zero. Roughly speaking, this means that we have to take Δ that converges somewhat slower than $1/\sqrt{n}$ whilst making sure V is in a compact interval away from zero.

Let us return to (4.20) which can be rewritten as

$$V = \frac{1}{\varrho_a}(-p\varphi'(p) + \varphi(p) + \varepsilon). \tag{4.23}$$

From (4.22) it follows that

$$\frac{1}{\varrho_a} - \frac{1}{\varphi'(p)} = \frac{-\varepsilon_a}{\varrho_a \varphi'(p)}.$$

With the help of (4.22) equality (4.23) becomes

$$V(p, \Delta, n) = -p + \frac{\varphi(p)}{\varphi'(p)} + \varepsilon_v, \tag{4.24}$$

where

$$\varepsilon_v := \frac{\varepsilon}{\varrho_a} + \frac{-\varepsilon_a}{\varrho_a \varphi'(p)}(-p\varphi'(p) + \varphi(p)). \tag{4.25}$$

Clearly when ε_a and ε both converge to zero and assuming that $\varphi'(p)$ and ϱ_a are bounded away from zero, we get that ε_v also converges to zero. This then yields that V converges to $-p + \varphi(p)/\varphi'(p)$.

To prove the Theorem 3.1, we need, among others, to show that $F_n(nq)/n$ converges sufficiently fast. We decompose $F_n(nq)/n - \varphi(q)$ into two parts:

$$\frac{F_n(nq)}{n} - \varphi(q) = \left(\frac{F_n(nq)}{n} - \frac{\mathbf{E}[F_n(nq)]}{n}\right) + \left(\frac{\mathbf{E}[F_n(nq)]}{n} - \varphi(q)\right).$$

The next lemma deals with the first part of the sum above.

Lemma 4.1 *For all q_1, q_2, such that $0 < q_1 < q_2 < \infty$, there exists $K > 0$ such that a.s. for all n large enough and for all $q \in [q_1, q_2]$ we have*

$$\left|\frac{F_n(nq)}{n} - \frac{\mathbf{E}[F_n(nq)]}{n}\right| \leq K\frac{\sqrt{\log n}}{\sqrt{n}}. \tag{4.26}$$

This lemma is an immediate consequence of the Borel-Cantelli lemma and the following proposition (a version of which can be found, for example, in McDiarmid [15]).

Proposition 4.1 *Let $m \in \mathbb{N}$. Let $W_1, W_2, \ldots$ be a sequence of i.i.d. variables taking values in a set A. Suppose that $f(w_1, w_2, \ldots, w_m)$ is a function from A^m to $\mathbb{R}$ with the property that changing any of the arguments of f while holding the others fixed changes the value of f by a quantity whose absolute value is less or equal to $c > 0$. Then for all $\varepsilon > 0$ we have*

$$P(|f(W_1, \ldots, W_m) - \mathbf{E}[f(W_1, \ldots, W_m)]| \geq \varepsilon) \leq 2e^{-2\varepsilon^2/(mc^2)}.$$

The next lemma we need is

Lemma 4.2 *For all q_1, q_2, such that $0 < q_1 < q_2 < \infty$, there exists $K > 0$ such that for all n large enough and for all $q \in [q_1, q_2]$ we have*

$$\left|\varphi(q) - \frac{\mathbf{E}[F_n(nq)]}{n}\right| \leq K\frac{\sqrt{\log n}}{\sqrt{n}}. \tag{4.27}$$

The similar result can be found in [2]. His proof is applicable to our case (see remark in the end of Sect. 2 of [2]), with just few small changes.

We are now ready to finish the proof of Theorem 3.1. We first prove (3.5). For this let $0 < q_1 < q_2 < \infty$ be such that $\varphi(p) \in C^2[q_1, q_2]$ and

$$q_1 < p_1, p_2 < q_2. \tag{4.28}$$

Then, because of (4.26), (4.27), and (4.28) we find that for all n (random) large enough and for all $p \in [p_1, p_2]$ it holds

$$|\varepsilon_1| \leq 2K\frac{\sqrt{\log n}}{\sqrt{n}}. \tag{4.29}$$

Note that since $\beta \in [\beta_1, \beta_2]$ we find that

$$\Delta \leq n^{\beta_2 - 0.5}, \tag{4.30}$$

where $n^{\beta_2-0.5}$ goes to zero as n goes to infinity since we assumed that $\beta_2 < 0.5$.

Hence, since we also assumed V to be bounded uniformly (see (3.4)), for all n large enough and for all $p \in [p_1, p_2]$, we have that $p + V\Delta \in [q_1, q_2]$. Hence again inequalities (4.26) and (4.27) apply and we find

$$|\varepsilon_3| \leq 2K \frac{\sqrt{\log n}}{\sqrt{n}}. \tag{4.31}$$

As $\varphi(p) \in C^2[q_1, q_2]$, the second derivative of $\varphi(\cdot)$ is bounded on $[q_1, q_2]$. Denote by K_6 the maximum

$$K_6 := \max_{p \in [q_1, q_2]} |\varphi''(p)|.$$

As mentioned, for n (random) large enough we have that $p + V\Delta$ is in $[q_1, q_2]$. By the Mean Value Theorem, we have then that

$$|\varepsilon_6| \leq K_6 \Delta V$$

and hence with the help of (3.4), we get

$$|\varepsilon_6| \leq K_6 c_2 n^{\beta - 0.5}. \tag{4.32}$$

Using now the bound (4.29) and (3.4), we obtain the inequality

$$\left|\frac{\varepsilon_1}{V\Delta}\right| \leq 2K \frac{\sqrt{\log n}}{\sqrt{n}} \cdot \frac{1}{c_1 n^{\beta - 0.5}}$$

and hence

$$\left|\frac{\varepsilon_1}{V\Delta}\right| \leq \frac{2K}{c_1}\sqrt{\log n}\, n^{-\beta}. \tag{4.33}$$

In the same way we obtain

$$\left|\frac{\varepsilon_3}{V\Delta}\right| \leq \frac{2K}{c_1}\sqrt{\log n}\, n^{-\beta}. \tag{4.34}$$

Together, (4.32), (4.33), (4.34) and imply that

$$|\varepsilon_a| \leq C_1 n^{\alpha} \sqrt{\log n} \tag{4.35}$$

where we take

$$C_1 := K_6 c_2 + \frac{4K}{c_1}.$$

Then, (4.35) and (4.22) together imply inequality (3.5).

We are now going to prove the bound (3.6). We chose q_1 and q_2 in the same way as was done in the previous computations. Because Δ goes to zero uniformly when n goes to infinity (see inequality (4.30)), there exists n_0 such that for all $n \geq n_0$, for all $p \in [p_1, p_2]$ and all $\beta \in [\beta_1, \beta_2]$ we have

$$\frac{p}{1+\Delta} \in [q_1, q_2].$$

Hence, inequality (4.26) and (4.27) apply, and we obtain that for all n large enough

$$|\varepsilon_2| \leq 2K \frac{\sqrt{\log(n+\Delta n)}}{\sqrt{n+\Delta n}} \leq 2K \frac{\sqrt{\log(n+\Delta n)}}{\sqrt{n}}. \tag{4.36}$$

Also, we may assume that for n large enough $n + \Delta n \leq 2n$. This assumption together with (4.36) implies

$$|\varepsilon_2| \leq 4K \frac{\sqrt{\log n}}{\sqrt{n}} \tag{4.37}$$

which also leads to

$$\left|\frac{\varepsilon_2}{\Delta}\right| \leq 4K\sqrt{\log n} n^{-\beta}. \tag{4.38}$$

For the same reasons we obtain

$$\left|\frac{\varepsilon_1}{\Delta}\right| \leq 2K\sqrt{\log n}\, n^{-\beta}. \tag{4.39}$$

Next consider the map

$$x \mapsto \varphi\left(\frac{p}{1+x}\right).$$

Obviously there exists an interval $[a, b]$, such that $0 \in (a, b)$, $-1 \notin [a, b]$, and such that for all $x \in [a, b]$ we have $p/(1+x) \in [q_1, q_2]$. The above map is then well defined and it is twice continuously differentiable on $[a, b]$. Since we know that Δ goes to zero uniformly and $0 \in (a, b)$, we have that for all n large enough and all $\beta \in [\beta_1, \beta_2]$, we get $\Delta \in [a, b]$. Let $c_3 > 0$ designate the maximum of the absolute value of the second derivative of $\varphi(p/(1+x))$:

$$c_3 := \max_{x \in [a,b]} \left\{ \left| \frac{d^2\varphi(p/(1+x))}{d^2x} \right| \right\}.$$

By Taylor's formula, we obtain

$$\begin{aligned}
\varphi\left(\frac{p}{1+\Delta}\right) &= \varphi(p) + \frac{d}{dx}\varphi\left(\frac{p}{1+x}\right)\Big|_{x=0} \Delta + \frac{1}{2}\frac{d^2}{dx^2}\varphi\left(\frac{p}{1+x}\right)\Big|_{x=\xi} \Delta^2 \\
&= \varphi(p) - p\varphi'(p)\Delta + \frac{1}{2}\frac{d^2}{dx^2}\varphi\left(\frac{p}{1+x}\right)\Big|_{x=\xi} \Delta^2
\end{aligned}$$

where $0 < \xi < \Delta$. Thus,

$$\begin{aligned}
|\varepsilon_4| &= \left| \frac{\varphi(p) - \varphi(p/(1+\Delta))}{\Delta} - p\varphi'(p) \right| \\
&\leq \left| \frac{1}{2}\frac{d^2}{dx^2}\varphi\left(\frac{p}{1+x}\right)\Big|_{x=\xi} \Delta \right|
\end{aligned}$$

$$\begin{aligned} &\leq c_3 \Delta \\ &= c_3 n^{\beta-0.5} \end{aligned} \tag{4.40}$$

for all n large enough and for all $\beta \in [\beta_1, \beta_2]$.

Finally, since Δ goes to zero uniformly, we find that for all n large enough and all $\beta \in [\beta_1, \beta_2]$ we have $p/(1+\Delta) \in [q_1, q_2]$. By the Mean Value Theorem, we obtain

$$|\varepsilon_5| \leq c_5 \Delta = c_5 n^{\beta-0.5}, \tag{4.41}$$

where

$$c_5 := \max_{p \in [q_1, q_2]} |\varphi'(p)|.$$

Looking at the definition (4.19) of the error term ε as well as the inequalities (4.37)–(4.41), we obtain

$$|\varepsilon| \leq C_2 \sqrt{\log n}\, n^{\alpha}, \tag{4.42}$$

where $\alpha = \max\{\beta - 0.5, -\beta\}$ and C_2 is the constant

$$C_2 := 10K + c_3 + c_5.$$

Let β_3 be equal to $\beta_3 := -\max\{\beta_2 - 0.5, -\beta_1\}$. Note that $\beta_3 > 0$ and, since $\beta \in [\beta_1, \beta_2]$, we have

$$n^{\alpha} \leq n^{-\beta_3}. \tag{4.43}$$

Hence, inequality (4.42) reads

$$|\varepsilon| \leq C_2 \sqrt{\log n}\, n^{-\beta_3}, \tag{4.44}$$

where the right-hand side converges to zero and does not depend on β. We assumed that for p in an open interval containing $[p_1, p_2]$ the derivative φ' is bounded below. Hence we get that there exists a constant $c_7 > 0$, such that for all $p \in [p_1, p_2]$ we have

$$\frac{1}{\varphi'(p)} \leq c_7. \tag{4.45}$$

We have proved in (4.35), that the difference between ϱ_a and $\varphi'(p)$ is less than a $C_1 n^{\alpha} \sqrt{\log n}$ for n large enough and all $\beta \in [\beta_1, \beta_2]$ and $p \in [p_1, p_2]$. Hence from our last comment above it follows that this difference is less than $\sqrt{\log n}\, n^{-\beta_3}$. Hence, $\varphi'(p)$ being bounded below, we obtain that for large enough n, ϱ_a is also bounded below. Hence, we get a uniform bound for all n large enough and all $\beta \in [\beta_1, \beta_2]$ and $p \in [p_1, p_2]$:

$$\frac{1}{\varrho_a} \leq c_8. \tag{4.46}$$

Since $[p_1, p_2]$ is a compact interval and since $\varphi(p)$ is continuously differentiable on it, we have that φ and φ' are both bounded from above and below on $[p_1, p_2]$. Hence, so is $-p\varphi'(p) + \varphi(p)$. Hence there exists $c_9 > 0$ so that for all $p \in [p_1, p_2]$ we have

$$|-p\varphi'(p) + \varphi(p)| \leq c_9.$$

Looking at the definition (4.25) of ε_v and using the last inequality above as well as (4.45) and (4.46), we find

$$|\varepsilon_v| \leq c_8|\varepsilon| + c_7c_8c_9|\varepsilon_a|.$$

Using (4.42) and (4.35), in the last inequality above, yields

$$|\varepsilon_v| \leq C_3\sqrt{\log n}\, n^{\alpha}, \tag{4.47}$$

where

$$C_3 := c_8C_2 + c_7c_8c_9C_1.$$

Inequality (4.47) together with (4.24), imply (3.6). This then finishes the proof of our main theorem.

References

1. Aldous, D., Diaconis, P.: Hammersley's interacting particle process and longest increasing subsequences. Probab. Theory Relat. Fields **103**, 199–213 (1995)
2. Alexander, K.S.: The rate of convergence of the mean length of the longest common subsequence. Ann. Appl. Probab. **4**(4), 1074–1082 (1994)
3. Amsalu, S., Matzinger, H., Popov, S.: Lower bounds for the probability of macroscopical non-unique alignments. ESAIM Probab. Stat. **11**, 281–300 (2007)
4. Arratia, R., Waterman, M.S.: A phase transition for the score in matching random sequences allowing deletions. Ann. Appl. Probab. **4**(1), 200–225 (1994)
5. Baeza-Yates, R.A., Gavaldà, R., Navarro, G., Scheihing, R.: Bounding the expected length of longest common subsequences and forests. Theory Comput. Syst. **32**(4), 435–452 (1999)
6. Chvatal, V., Sankoff, D.: Longest common subsequences of two random sequences. J. Appl. Probab. **12**, 306–315 (1975)
7. Dančík, V., Paterson, M.: Upper bounds for the expected length of a longest common subsequence of two binary sequences. Random Struct. Algorithms **6**(4), 449–458 (1995)
8. Deken, J.G.: Some limit results for longest common subsequences. Discrete Math. **26**(1), 17–31 (1979)
9. Durringer, C., Hauser, R., Matzinger, H.: Approximation to the mean curve in the LCS problem. Stoch. Process. Their Appl. **118**(4), 629–648 (2008)
10. Hauser, R., Matzinger, H., Martinez, S.: Large deviation based upper bounds for the lcs-problem. Adv. Appl. Probab. **38**(3), 827–852 (2006)
11. Houdre, C., Lember, J., Matzinger, H.: On the longest common increasing binary subsequence. C.R. Acad. Sci. Paris Ser. I **343**, 589–594 (2006)
12. Kiwi, M., Loebl, M., Matousek, J.: Expected length of the longest common subsequence for large alphabets. Adv. Math. **197**(2), 480–498 (2005)
13. Lember, J., Matzinger, H.: Standard deviation of the longest common subsequence (2006, submitted)
14. Lember, J., Matzinger, H., Durringer, C.: Deviation from the mean in sequence comparison when one sequence is periodic. Lat. Am. J. Probab. Math. **III** (2007)
15. McDiarmid, C.: On the method of bounded differences. In: Surveys in Combinatorics, pp. 148–188. Cambridge University Press, Cambridge (1989)
16. Menshikov, M.V., Popov, S.Yu., Sisko, V.: On the connection between oriented percolation and contact process. J. Theor. Probab. **15**(1), 207–221 (2002)
17. Paterson, M., Dančík, V.: Longest common subsequences. In: Mathematical Foundations of Computer Science 1994 (Košice, 1994). Lecture Notes in Comput. Sci., vol. 841, pp. 127–142. Springer, Berlin (1994)
18. Seppäläinen, T.: Exact limiting shape for a simplified model of first-passage percolation on the plane. Ann. Probab. **26**(3), 1232–1250 (1998)
19. Seppäläinen, T.: Large deviations for increasing sequences on the plane. Probab. Theory Relat. Fields **112**, 221–251 (1998)
20. Steele, M.J.: An Efron-Stein inequality for non-symmetric statistics. Ann. Stat. **14**, 753–758 (1986)
21. Waterman, M.S.: General methods of sequence comparison. Bull. Math. Biol. **46**(4), 473–500 (1984)
22. Waterman, M.S.: Estimating statistical significance of sequence alignments. Philos. Trans. R. Soc. Lond. B **344**, 383–390 (1994)
23. Waterman, M.S.: Introduction to Computational Biology. Chapman & Hall, London (1995)