Information Loss in Coarse Graining of Polymer Configurations via Contact Matrices

Patrik L. Ferrari* and Joel L. Lebowitz†

December 3, 2002

Abstract

Contact matrices provide a coarse grained description of the configuration ω of a linear chain (polymer or random walk) on $\mathbb{Z}^n$: $\mathcal{C}_{ij}(\omega) = 1$ when the distance between the position of the i-th and j-th step are less then or equal to some distance a, $\mathcal{C}_{ij}(\omega) = 0$ otherwise. We consider models in which polymers of length N have weights corresponding to simple and self-avoiding random walks, SRW and SAW, with a the minimal permissible distance. We prove that to leading order in N, the number of matrices equals the number of walks for SRW but not SAW. The coarse grained Shanon entropies for SRW agree with the fine grained ones for $n \leq 2$ but disagree for $n \geq 3$.

*Zentrum Mathematik, Technische Universität München, D-85747 Garching, Germany e-mail: ferrari@ma.tum.de

†Departments of Mathematics and Physics, Rutgers University, Piscataway, New Jersey e-mail: lebowitz@math.rutgers.edu

1 Introduction

The use of coarse grained descriptions is essential for systems with many degrees of freedom. The choice of the coarse grained variables is dictated by the nature of the system and by the questions of interest. One is then interesting in the amount of information lost in the coarse graining, at least in some statistical way [2].

In this paper we shall study this question for simple models of polymers, large molecules consisting of a linear sequence of N monomer units. A reduced description of this system can be based [7, 10, 13, 14] on associating to each polymer configuration a connectivity or contact matrix $\mathcal{C}$, such that $\mathcal{C}_{ij} = 1$ or 0 depending on whether the distance between the position of the i-th and j-th monomer, $\omega(i)$ and $\omega(j)$, is smaller or bigger than a certain specified value, a,

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{if } |\omega(i) = \omega(j)| \leq a, \quad i \neq j, \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

This is clearly a many to one coarse grained description of the polymer configuration. Still, knowledge of the contact matrix can predict many features of the vibrational spectra of certain proteins [1, 5].

To answer the question of how much information about a polymer is retained by its contact matrix, we consider an idealized version of the geometrical structure of a polymer in which the monomers occupy sites on the n-dimensional lattice, $\mathbb{Z}^n$. The configuration of a polymer containing $N + 1$ monomers is then specified by $\omega_N = (\omega(0), \omega(1), ..., \omega(N)) \in \Omega_N$ with $\omega(0) \equiv 0$ with $|\omega(i+1) - \omega(i)| = \pm e_\alpha$, where e_α is one of the unit directions on $\mathbb{Z}^n$, $\alpha = 1, ..., n$.

Let there be some given probability distribution $\mathbb{P}(\omega)$ on Ω. (We shall drop the subscript N whenever possible.) The set $\mathbf{C} = \{\mathcal{C}(\omega)\}_{\omega \in \Omega}$ partitions

Ω into sets $\Omega_C = \{\omega : \mathcal{C}(\omega) = C\}$, with

$$\deg \mathcal{C} = |\Omega_C| \tag{1.2}$$

the number of configurations $\omega \in \Omega_C$. The probability of ω being in Ω_C is then

$$\mathbb{P}(C) = \sum_{\omega \in \Omega_C} \mathbb{P}(\omega). \tag{1.3}$$

To measure the information lost in the coarse-graining we may compare the Shanon entropy of the coarse grained measure $\mathbb{P}(C)$, S_C, with the fine grained entropy S,

$$S = -\sum_{\omega \in \Omega} \mathbb{P}(\omega) \ln \mathbb{P}(\omega). \tag{1.4}$$

We then have

$$\begin{aligned} S_C = -\sum_{C \in \mathbf{C}} \mathbb{P}(C) \ln \mathbb{P}(C) &= -\sum_{C \in \mathbf{C}} \sum_{\omega \in \Omega_C} \mathbb{P}(\omega) \ln \left(\frac{\mathbb{P}(C)}{\mathbb{P}(\omega)} \mathbb{P}(\omega) \right) \\ &= S - \hat{S}_C \end{aligned} \tag{1.5}$$

where

$$\hat{S}_C = -\sum_{C \in \mathbf{C}} \mathbb{P}(C) \sum_{\omega \in \Omega_C} \mathbb{P}(\omega|C) \ln \mathbb{P}(\omega|C) \tag{1.6}$$

and

$$\mathbb{P}(\omega|C) = \mathbb{P}(\omega)/\mathbb{P}(C)|_{C=\mathcal{C}(\omega)} \tag{1.7}$$

is the conditional probability of ω given that it is in Ω_C. We can thus think of $\hat{S}_C$ as the average "conditional entropy" relative to $\mathcal{C}$. Since $\hat{S}_C \geq 0$ we clearly have $S_C \leq S$, and $S - S_C$ is then a measure of information lost in the coarse graining [2]. The question is how much. In particular we may ask how does

$$\delta_N \equiv S_C/S \tag{1.8}$$

behave as $N \to \infty$.

Before answering this question we note that

$$S_C \leq \bar{S}_C = \ln |\mathbf{C}|, \tag{1.9}$$

where $\bar{S}_C$ is the entropy of the distribution $\bar{S}_C$ which assigns equal weight to all $C \in \Omega_C$, i.e. $\bar{\mathbb{P}}(C) = |\mathbf{C}|^{-1}$, with $|\mathbf{C}| = W(N)$ the total number of different coarse grained components, i.e. contact matrices. We may then also define

$$\gamma_N = \bar{S}_C/S \tag{1.10}$$

and by (1.9)

$$\delta_N \leq \gamma_N \leq 1. \tag{1.11}$$

So far everything is completely general. We shall now specialize to the case where all permissible configurations ω, i.e. all those for which $\mathbb{P}(\omega) \neq 0$, have the same probability. Then

$$W(N) = |\mathbf{C}| = \sum_{\omega \in \Omega} 1/\deg \mathcal{C}(\omega) = |\Omega| \mathbb{E}(\deg \mathcal{C})^{-1}, \tag{1.12}$$

where $\mathbb{E}$ is the expectation value with respect to the relevant uniform distribution.

The examples we shall consider here are:

1) The weights are those of simple symmetric random walks (SRW) on $\mathbb{Z}^n$, i.e. $|\Omega| = (2n)^N$ and $\mathbb{P}(\omega) = (2n)^{-N}$ for all ω.

2) The polymers behave like self avoiding walks (SAW) on $\mathbb{Z}^n$, i.e. the configuration space Ω consists of all ω s.t. $\omega(i) \neq \omega(j)$ for $i \neq j$, and $\mathbb{P}(\omega) = |\Omega|^{-1}$, where $|\Omega| \sim \mu_{SAW}^N$ is the number of SAW on $\mathbb{Z}^n$ of length N. SAW model the steric exclusion effects of the monomers and are frequently used as a model for polymers [4, 9].

3) The chains behave like bond self avoiding walks (BAW) on $\mathbb{Z}^n$ in which case Ω consists of all ω such that the pair $[\omega(i), \omega(i+1)] \neq [\omega(j), \omega(j \pm 1)]$ for $i \neq j$, and $\mathbb{P}(\omega) = |\Omega|^{-1}$, $|\Omega| \sim \mu_{BAW}^N$, the number of BAW [11].

Note that for uniform distributions S is just the log of the total number of configurations, i.e. $S = \ln|\Omega|$, so γ_N is just the ratio of the logarithms of the numbers of contact matrices and random walks.

The behavior of γ_N was studied in [13] for the case of SAW, with

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{if } |\omega(i) - \omega(j)| = 1, |i-j| = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.13}$$

(The inequality can in fact be made an equality here since 1 is the minimal distance between $\omega(i)$ and $\omega(j)$ for $i \neq j$.) Numerical studies [13] for $n = 2$ indicated that γ_N remained strictly less than 1 in the limit $N \to \infty$. It is then natural to ask whether the same is true for the SRW when we again define $\mathcal{C}_{ij}(\omega) = 1$ when $\omega(i)$ and $\omega(j)$ are as close as they can be, i.e. when a in (1.1) is set equal to zero

$$\mathcal{C}_{ij}(\omega) = \begin{cases} 1 & \text{for } |\omega(i) - \omega(j)| = 0, i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{1.14}$$

Note that for this case $W(N)$ satisfies

$$W(N_1 + N_2) \geq W(N_1)W(N_2) \tag{1.15}$$

so γ_N is monotone nondecreasing and the $\lim_{N\to\infty} \gamma_N$ exists.

In the present work we prove some results about δ_N and γ_N for all the above examples. (Some of these generalize ready to other uniform distributions.)

2 Main Results

Theorem 2.1. *For SRW on $\mathbb{Z}^2$, there exists constants $\kappa, \kappa_1, \kappa_2 > 0$ such that for large N,*

$$\gamma_N \geq 1 - \frac{\kappa \ln N}{N^{1/3}}, \tag{2.1a}$$

$$1 - \frac{\kappa_1}{\ln N} \leq \delta_N \leq 1 - \frac{\kappa_2}{(\ln N)^2}. \tag{2.1b}$$

Consequently, γ_N and $\delta_N \to 1$, as $N \to \infty$.

Theorem 2.2. *For SRW on $\mathbb{Z}^n$, $n \geq 3$, there exist constants κ_n and $\kappa'_n > 0$ such that for N large enough,*

$$\gamma_N \geq 1 - \frac{\kappa_n}{N^{2/(n+2)}}, \tag{2.2a}$$

$$\delta_N \leq 1 - \kappa'_n. \tag{2.2b}$$

Consequently, $\gamma_N \to 1$, as $N \to \infty$ while $\limsup_{N\to\infty} \delta_N < 1$.

Theorem 2.3. *For SAW and BAW, on $\mathbb{Z}^n$, $n \geq 2$,* $\limsup_{N\to\infty} \gamma_N < 1$ *and ipso-facto* $\limsup_{N\to\infty} \delta_N < 1$.

We are indebted to Harry Kesten for the key idea in the proof of Theorem 2.3 for SAW. The extension to BAW is straightforward.

Intuitively we expect that the larger the number of intersections the more information is contained in the contact matrices. It is therefore not surprising that for recurrent RW, such as SRW in $n = 2$, both γ_N and δ_N would go to 1. (For the degenerate case $n = 1$ there are for all the cases just twice as many random walks as C's corresponding to whether the first step is to the right or the left.) One expects however that for RW which have a strong tendency to spread out such as SAW in $n \geq 2$, and SRW in $n \geq 3$ the contact matrices lose too much information. This is indeed reflected in $\delta_N < 1$ for SAW in $n \geq 2$ and SRW in $n \geq 3$. The same is true for γ_N for SAW. Surprisingly however $\gamma_N \to 1$ for SRW in all dimensions. The reason for this, as we shall see, is that the probability that an SRW of length N in $\mathbb{Z}^n$ visits $R_N \leq \epsilon N$ distinct sites goes to zero slower than exponentially when $N \to \infty$ for any fixed $\epsilon > 0$.

The outline of the rest of the paper is as follows. In section 3 we first present some general inequalities and then prove the results on γ_N for SRW. In section 4 we give some bounds on the degeneracy of SRW and then prove the results on δ_N for SRW. In section 5 we prove Theorem 2.3 for SAW and BAW.

3 Proof of Results for γ_N

Some Inequalities

Using Jensen's inequality on the average over Ω_C, gives

$$\begin{aligned}\sum_{\omega\in\Omega_C} \frac{1}{\deg \mathcal{C}}(\mathbb{P}(\omega)\deg \mathcal{C})\ln(\mathbb{P}(\omega)\deg \mathcal{C}) \quad &\geq \quad \sum_{\omega\in\Omega_C} \mathbb{P}(\omega)\ln\left(\sum_{\omega\in\Omega_C}\mathbb{P}(\omega)\right) \\ &= \quad \mathbb{P}(C)\ln\mathbb{P}(C). \qquad (3.1)\end{aligned}$$

Writing now

$$S = -\sum_{C\in\mathbf{C}}\sum_{\omega\in\Omega_C}\mathbb{P}(\omega)(\ln(\mathbb{P}(\omega)\deg \mathcal{C}) - \ln\deg \mathcal{C}) \qquad (3.2)$$

we obtain[1]

$$S \le S_C + \mathbb{E}(\ln \deg \mathcal{C}), \tag{3.3}$$

which yields

$$1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{S} \le \delta_N \le 1. \tag{3.4}$$

We next give an upper bound for the degeneracy of the contact matrices defined in (1.14). This is the key to our results for SRW. Define the range R_N of an SRW with length N to be the number of distinct sites visited by the walk.

Lemma 3.1. *Let ω have a range $R_N = M$ and let $\mathcal{C}(\omega)$ be its contact matrix. Then*

$$\deg \mathcal{C}(\omega) \le (2n)^M. \tag{3.5}$$

PROOF. The contact matrix $\mathcal{C}(\omega)$ has $N+1-M$ columns with "1"'s in the upper triangular part, because we have $N+1-M$ intersections. Let us now construct all random walks ω' such that $\mathcal{C}(\omega') = \mathcal{C}(\omega)$. Consider now $\omega'(k)$ with $k \neq 0$, then there are two possible cases:

1. there exists a $i < k$ such that $\mathcal{C}_{ik}(\omega) = 1$,
2. for all $i < k$, $\mathcal{C}_{ik}(\omega) = 0$.

In the first case, $\omega'(k) = \omega'(i)$ and therefore we have only one choice for it. In the second case, the k^{th} step will occupy a place which was never occupied before. Therefore we have *at most* $2n$ possibles choices.

For a ω' with a contact matrix $\mathcal{C}(\omega') = \mathcal{C}(\omega)$ there are $M-1$ steps for which we are in the second case because the starting point is fixed at the

[1] The right side of (3.3) is the maximum of the entropy over all measures, $\mu(\omega)$ such that $\mu(\mathcal{C}) = \mathbb{P}(\mathcal{C})$.

origin and $N+1-M$ steps for which we are in the first one. Therefore there are *at most* $(2n)^{M-1} \le (2n)^M$ different ω' satisfying $\mathcal{C}(\omega') = \mathcal{C}(\omega)$, i.e. $\deg \mathcal{C}(\omega) \le (2n)^M$. $\square$

Proof of (2.1a). We first note that $\delta_N, \gamma_N \to 1$ for general recurrent RW, which includes SRW on $\mathbb{Z}^2$. For such walks $\mathbb{E}(R_N)/N \to 0$ as $N \to \infty$, therefore using (3.4) and Lemma 3.1,

$$\delta_N \ge 1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{N \ln 2n} \ge 1 - \frac{\mathbb{E}(R_N)}{N} \to 1 \text{ as } N \to \infty. \tag{3.6}$$

To prove (2.1a) consider the subset Ω^α of SRW on $\mathbb{Z}^2$ defined as

$$\Omega^\alpha = \{\omega \in \Omega_N \text{ s.t. } \omega(k \cdot 4[N^\alpha]) = 0, k = 0, 1, \ldots, k_{max}\}. \tag{3.7}$$

where k_{max} is the largest integer k such that $k4[N^\alpha] \le N$.Let us take $0 < \alpha < 1/2$. Each $\omega \in \Omega^\alpha$ returns to the origin after $4[N^\alpha]$ steps, it is therefore contained in a cube of edge length $4[N^\alpha]$ (except eventually for the last $2[N^\alpha]$ steps). Using Stirling formula we have, for some $C > 0$,

$$\mathbb{P}(\omega(4M) = 0) \ge \frac{(4M)!}{(M!)^4 4^{4M}} \ge CM^{-3/2} \tag{3.8}$$

because $\{\omega(4M) = 0\} \supset \{\omega(4M) = 0 \text{ with } M \text{ steps in each direction}\}$. Then

$$\mathbb{P}(R_N \le N^\beta \equiv 4^2[N^\alpha]^2 + 2[N^\alpha]) \ge \mathbb{P}(\omega \in \Omega^\alpha) \ge (C/[N^\alpha]^{3/2})^{N/4[N^\alpha]}. \tag{3.9}$$

Therefore combining (1.12), (3.5) and (3.9) for $n = 2$ obtain $W(N) \ge 4^N \mathbb{P}(R_N \le N^\beta)/4^{(N^\beta)}$, which implies

$$1 \ge \gamma_N \ge 1 - \kappa \frac{\ln N}{[N^\alpha]} - \frac{4^2[N^\alpha]^2 + 2[N^\alpha]}{N}, \tag{3.10}$$

for a $\kappa > 0$. For $\alpha = 1/3$ the RHS of (3.10) is optimized and the term with the logarithm dominates the last one. This proves the bound for $n = 2$. $\square$

Proof of (2.2a). For $n > 2$, $\mathbb{P}(R_N = X) \sim \exp(-aN/X^{2/n})$ when $X \to \infty$, $\frac{X}{N} \to 0$ (see [12] and pp. 88-92 of [17]). Therefore for $\alpha \in (0,1)$, $\mathbb{P}(R_N = N^\alpha) \sim \exp(-aN^{1-2\alpha/n})$. But for an ω with $R_N = M$, $\deg \mathcal{C}(\omega) \leq (2n)^M$, see Lemma 3.1. Therefore restricting the sum in (1.12) to $\omega \in \Omega_N^\alpha$, we have $W(N) \geq (2n)^N \mathbb{P}(R_N = N^\alpha)/(2n)^{N^\alpha}$, since the numerator is just the number of terms in that sum. This implies, for large N,

$$1 \geq \gamma_N \geq 1 - \sigma(N, \alpha) \tag{3.11}$$

where $\sigma(N,\alpha) = N^{\alpha-1} + aN^{-2\alpha/n}/\ln 2n$. Choosing $\alpha \in (0,1)$ which minimizes $\sigma(N,\alpha)$ for large N, we obtain $\alpha - 1 = -2/(n+2)$. Taking $\kappa_n = (\ln 2n + a)/\ln 2n$ completes the proof. □

This theorem implies that for N large we have (up to smaller corrections),

$$W(N) \geq (2n)^N (2n)^{-\kappa_n N^{n/(n+2)}}. \tag{3.12}$$

There exists also an upper bound on γ_N which depends on the decrease of $\mathbb{P}(R_N/N < \varepsilon)$.

Proposition 3.2. *For all fixed $\varepsilon > 0$, there exists a constant $\kappa' > 0$ such that for N large enough,*

$$\gamma_N \leq 1 - \kappa' \frac{|\ln \mathbb{P}(R_N/N \leq \varepsilon)|}{N}. \tag{3.13}$$

The outline of the proof will be given in the Appendix.

4 Bounds on the degeneracy of SRW

4.1 SRW on $\mathbb{Z}^2$

For SRW on $\mathbb{Z}^2$, $R_N \sim \frac{\pi N}{\ln N}$, more precisely (see e.g. [15]),

$$\mathbb{E}(R_N) = \frac{\pi N}{\ln 8N} \left(1 + \mathcal{O}\left(1/\ln N\right)\right). \tag{4.1}$$

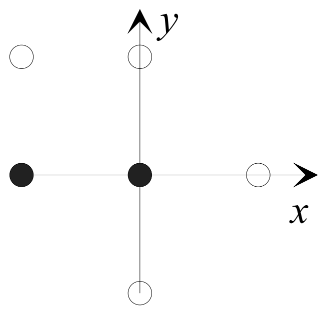

Figure 1: The pattern Q. The visited sites of Q are black.

Next we apply a result of van Wijland, Caser and Hilhorst [15]. Let the support of ω be defined to be the set of points visited by ω. Consider two finite disjoint sets of lattice points A_u and A_v. The "pattern" centered at $\mathbf{x}$ associated with the sets A_u and A_v is a configuration of $|A_v|$ visited sites $\mathbf{x}+\mathbf{z}$, $\mathbf{z} \in A_v$ and of $|A_u|$ unvisited sites $\mathbf{x}+\mathbf{z}'$, $\mathbf{z}' \in A_u$. We say that the pattern appears in the support of ω at $\mathbf{x}$ if the lattice points $\mathbf{x}+\mathbf{z}$, $\mathbf{z} \in A_v$, are in the support of ω and the lattice points $\mathbf{x}+\mathbf{z}'$, $\mathbf{z}' \in A_u$, are not in the support of ω. The numbers of times that a pattern appears in the support of ω is then the number of different $\mathbf{x} \in \mathbb{Z}^n$ such that it appears at $\mathbf{x}$.

Let us consider the "pattern Q" defined as the set composed of the following sets A_v and A_u: $A_v(Q) = \{(0,0),(-1,0)\}$ and $A_u(Q) = \{(1,0),(0,-1),(0,1),(-1,1)\}$ (see Figure 1).

Let $Q_N = Q_N(\omega)$ be the number of times that Q appears in the support of ω. Then using [15], $\mathbb{E}(Q_N) = \frac{\pi^2 N}{(\ln 8N)^2} m_1 + \mathcal{O}\left(\frac{N}{(\ln 8N)^3}\right)$ where $m_1 = m_1(Q)$ is a constant, and $Q_N - \mathbb{E}(Q_N) \simeq \frac{2\mathcal{A}}{\ln 8N} \mathbb{E}(Q_N)\gamma(N)$ where $\gamma(N)$ is a random variable (Varadhan's renormalized local time of self-intersections) with mean 0 and variance 1. $\mathcal{A}$ is a constant given in [15] whose value is ~ 1.3034. We computed m_1 finding $m_1 \cong 2.78 \cdot 10^{-3}$.

These results imply the following proposition.

Proposition 4.1. *For simple random walks on $\mathbb{Z}^2$, there exists a $\nu > 0$ such that*

$$\lim_{N\to\infty} \mathbb{P}\left(\deg \mathcal{C}(\omega) \geq e^{\nu \frac{N}{(\ln N)^2}}\right) = 1. \tag{4.2}$$

Proof. Suppose that the pattern Q, centered at $\zeta \in \mathbb{Z}^2$, exists in the support of a random walk ω. Let us consider the following transformation:

$$\begin{aligned} T_\zeta : \Omega_N &\longmapsto \Omega_N \\ \omega &\longrightarrow T_\zeta(\omega) = \begin{cases} \omega(i) & \text{if } \omega(i) \neq \zeta, \\ \zeta + (-1,1) & \text{if } \omega(i) = \zeta. \end{cases} \end{aligned} \tag{4.3}$$

In other words we exchange the points ζ and $\zeta + (-1,1)$. This application does not change the contact matrix of the random walk, because $\zeta + (-1,1)$ is connected only with $\zeta + (-1,0)$. We have to prove that the probability of having the pattern Q in the support of a random walk at least $M = \nu N/(\ln N)^2$ times goes to 1 as $N \to \infty$. A RW with M times the pattern Q appearing in its support is at least 2^M times degenerate: we can apply or not apply T_ζ independently for each ζ such that Q appears in the support of ω (centered in ζ).

We want an upper bound of $\mathbb{P}\left(Q_N < \alpha \frac{\mu N}{(\ln N)^2}\right)$ for $\alpha \in (0,1)$ and $\mu = m_1 \pi^2$. For each $k > 0$ and N large enough,

$$\mathbb{P}\left(Q_N < \alpha \frac{\mu N}{(\ln N)^2}\right) \leq \mathbb{P}\left(Q_N - \mathbb{E}(Q_N) \leq -k a_Q \frac{N}{(\ln 8N)^3}\right) \tag{4.4}$$

with $a_Q = 2\mu\mathcal{A}$. In fact, for N large enough, $\mathbb{E}(Q_N) = \frac{\mu N}{(\ln 8N)^2} + \mathcal{O}(N/(\ln 8N)^3)$ and therefore for each $\alpha < 1$, $\frac{\mu N}{(\ln 8N)^2} + \mathcal{O}(N/(\ln 8N)^3) - k a_Q \frac{N}{(\ln 8N)^3} \geq \alpha \frac{\mu N}{(\ln N)^2}$. Thus

$$\begin{aligned} \mathbb{P}\left(Q_N < \alpha \frac{\mu N}{(\ln N)^2}\right) &\leq \mathbb{P}\left(Q_N - \mathbb{E}(Q_N) \leq -k a_Q \frac{N}{(\ln 8N)^3}\right) \\ &\leq \frac{\mathbb{E}\left(Q_N - \mathbb{E}(Q_N)\right)^2}{k^2 a_Q^2 \frac{N^2}{(\ln 8N)^6}} \overset{N\to\infty}{\longrightarrow} \frac{1}{k^2}. \end{aligned} \tag{4.5}$$

Therefore for each $\alpha \in (0,1)$ we have $\forall\, k > 0$,

$$\lim_{N\to\infty} \mathbb{P}\left(Q_N < \alpha \frac{\mu N}{(\ln N)^2}\right) \leq \frac{1}{k^2}. \tag{4.6}$$

This implies that for all $\alpha \in (0,1)$, $\lim_{N\to\infty} \mathbb{P}\left(Q_N \geq \alpha \frac{\mu N}{(\ln N)^2}\right) = 1$. All the random walks with a number of pattern Q in their support more than $\alpha \frac{\mu N}{(\ln N)^2}$ are more degenerate than $e^{\alpha\mu N \ln 2/(\ln N)^2}$. Then, since $\alpha \in (0,1)$, for all choice of $\nu < \mu \ln 2$ we have (4.2). $\square$

4.2 SRW on $\mathbb{Z}^n$, $n \geq 3$

Let us define, for $n = 3$, the pattern P as consisting of a set A_v of visited sites and a set A_u of unvisited sites as follows: $A_v(P) = \{(0,0,0), (-1,0,0)\}$ and $A_u(P) = \{(1,0,0), (0,1,0), (0,-1,0), (0,0,1), (0,0,-1), (-1,1,0)\}$. Let $P_N = P_N(\omega)$ be the number of times that P appears in the support of ω. Then using [16] we have $\mathbb{E}(P_N) = m_1 N + \mathcal{O}(\sqrt{N})$ and $P_N - \mathbb{E}(P_N) \simeq a_P \sqrt{N \ln N}\,\eta(N)$ where $m_1 = 2.5 \cdot 10^{-3}$, $a_P = 1.2 \cdot 10^{-2}$ and $\eta(N)$ is a random variable with normal distribution $\mathcal{N}(0,1)$.

Proposition 4.2. *For simple random walks on $\mathbb{Z}^3$, there exists a $\nu > 0$ such that*

$$\lim_{N\to\infty} \mathbb{P}\left(\deg \mathcal{C}(\omega) \geq e^{\nu N}\right) = 1. \tag{4.7}$$

Proof. The proof is very close to the one of Proposition 4.1. This time we exchange the sites ζ and $\zeta + (-1,1,0)$ (if P appears centered in ζ) and we prove that $\forall k > 0$ and N large enough $\mathbb{P}(P_N < \alpha m_1 N) \leq \frac{1}{k^2}$ if $\alpha < 1$. Then for all choice of $\nu < m_1 \ln 2$ (4.7) holds. $\square$

In dimension $n \geq 4$ the same result (with a different value of m_1) is expected to hold. In fact a similar pattern in $n \geq 4$ has $\mathbb{E}(P_N) = m_1 N +$

$\mathcal{O}(\ln N)$ in $n = 4$, $\mathbb{E}(P_N) = m_1 N + \mathcal{O}(1)$ in $n \geq 5$ and $\mathbb{E}(P_N - \mathbb{E}(P_N))^2 = C_P N + o(N)$, see [16]. The only point that one should prove for $n \geq 4$ is that $m_1 \neq 0$.

4.3 Proof of (2.1b) and (2.2b)

The previous results on the degeneracy lead to the following results.

Proof of (2.1b). Using Propositions 4.1 and (4.1) we obtain bounds on δ_N, for $n = 2$.

$$\delta_N \leq 1 - \mathbb{P}\left(\deg \mathcal{C} \geq e^{\nu N/(\ln N)^2}\right) \frac{\nu}{\ln 4 (\ln N)^2}. \tag{4.8}$$

By Proposition 4.1, there exists a $\nu > 0$ such that $\mathbb{P}\left(\deg \mathcal{C} \geq e^{\nu N/(\ln N)^2}\right) \to 1$ as $N \to \infty$. This gives

$$\lim_{N\to\infty} (1 - \delta_N)(\ln N)^2 \geq \nu / \ln 4. \tag{4.9}$$

Consequently for $\kappa_2 < \nu/\ln 4$ and N large enough, $\delta_N \leq 1 - \frac{\kappa_2}{(\ln N)^2}$.

On the other hand,

$$\delta_N \geq 1 - \frac{\mathbb{E}(\ln \deg \mathcal{C})}{N \ln 2n} \geq 1 - \frac{\mathbb{E}(R_N)}{N}. \tag{4.10}$$

Then

$$\lim_{N\to\infty} (1 - \delta_N) \ln N \leq \lim_{N\to\infty} \mathbb{E}(R_N) \frac{\ln N}{N} = \pi, \tag{4.11}$$

consequently for $\kappa_1 > \pi$ and N large enough, $\delta_N \geq 1 - \kappa_1/\ln N$. $\square$

Proof of (2.2b). For all $\nu > 0$,

$$\delta_N \leq 1 - \frac{\mathbb{P}(\deg \mathcal{C} \geq e^{\nu N})}{\ln (2n)} \nu. \tag{4.12}$$

By Proposition 4.2, there exists, for $n \geq 3$, a $\nu > 0$ such that $\mathbb{P}(\deg \mathcal{C} \geq e^{\nu N}) \to 1$ as $N \to \infty$. Therefore for N large enough

$$\delta_N \leq 1 - \frac{\nu/2}{\ln(2n)} < 1, \tag{4.13}$$

and $\limsup_{N\to\infty} \delta_N < 1$. $\square$

5 Proof of Theorem 2.3

The contact matrix of ω is now defined by (1.13).

Let us consider the case of SAW. We introduce some notation: we consider a cube $D = \{x \in \mathbb{Z}^n \text{ s.t. } c_i \leq x^{(i)} \leq c_i + b, 1 \leq i \leq n\}$ for some $c = (c_1, \ldots, c_n) \in \mathbb{Z}^n$ with its boundary $\partial D = \{x \in \mathbb{Z}^n \text{ s.t. } x^{(i)} = c_i + b \text{ or } x^{(i)} = c_i, 1 \leq i \leq n\}$. A path P is a SAW of finite length, say k, starting at the origin, i.e. $P = \{X_i(P), 0 \leq i \leq k\}$ with $X_0(P) = 0$.

We consider only paths such that there exists a cube D with $X_0(P) = 0$ and $X_k(P)$ two of its vertices and $X_j(P) \in D$ for all $0 \leq j \leq k$. We say that (P, D) occurs at the r^{th} step ω if

1. $X_{r+j}(\omega) - X_r(\omega) = X_j(P)$ for all $j = 0, \ldots, k$ and
2. ω does not occupy any other points of D.

$\chi_N(j, (P, D))$ is the number of $\omega \in \Omega_N$ such that (P, D) occurs *at most* at j steps.

Theorem 5.1 (Kesten Pattern Theorem [8]). *Let P be a SAW and D a cube such that D has 0 and $X_k(P)$ as two of its vertices and contains P. Then*

$$\limsup_{N\to\infty} \left(\frac{\chi_N(aN, (P, D))}{|\Omega_N|} \right)^{1/N} < 1 \text{ for some } a > 0 \tag{5.1}$$

where $|\Omega_N|$ is the total number of SAW of length N.

It is known that $|\Omega_N| \simeq \mu_{SAW}^N$ with $\mu_{SAW} > 0$.

Proof of Theorem 2.3. Let us take $b > 2$ and consider a SAW path of length $k + 2n$ constructed as follows. The firsts n steps of P connect the points $(0, \dots, 0)$ and $(1, \dots, 1)$. The following k steps connect the points $(1, \dots, 1)$ and $(b-1, \dots, b-1)$ with a SAW remaining always in $D \setminus \partial D$. The last n steps of P connect the points $(b-1, \dots, b-1)$ and $(b, \dots, b)$. Let us divide the set Ω_N into a sum of two disjoint parts: $\Omega_N = \Omega_N^a \cup (\Omega_N^a)^c$ where $\Omega_N^a = \{\omega \in \Omega_N \text{ s.t. } (P, D) \text{ occurs at most } aN \text{ times}\}$ and $(\Omega_N^a)^c$ its complementary set. It follows from Theorem 5.1 that

$$\exists \, \zeta > 0 \text{ s.t. } \mathbb{P}(\omega \in \Omega_N^a) \leq e^{-\zeta N}.$$

Let us take a $\omega \in (\Omega_N^a)^c$. Then (P, D) occurs at least aN times in ω. Consider an occurrence of (P, D) in the piece of P between its t^{th} and its $(t+k+2n)^{th}$ steps. We apply an axis rotation of $2\pi/n$ degrees to the cube $D \setminus \partial D$, where the axis is its diagonal of direction $(1, \dots, 1)$. This transformation does not change the contact matrix, and we can apply it n times obtaining each time a different SAW. For the chosen ω it can be done independently in at least aN different places, therefore the corresponding contact matrix is at least n^{aN} times degenerate.

Now we have an upper bound for the total number of contact matrices:

$$\begin{aligned} W(N) &\leq \mathbb{P}(\omega \in \Omega_N^a)|\Omega_N| + \mathbb{P}(\omega \in (\Omega_N^a)^c)|\Omega_N| n^{-aN} \qquad (5.2) \\ &\leq \left(e^{-\zeta N} + e^{-Na \ln n}\right) |\Omega_N|. \end{aligned}$$

Defining $\alpha_M = \max\{\zeta, a \ln n\} > 0$ and $\alpha_m = \min\{\zeta, a \ln n\} > 0$, we obtain

$$\begin{aligned} \limsup_{N\to\infty} \gamma_N &\leq \lim_{N\to\infty} \frac{\ln\left(e^{-\alpha_m N}\left(1 + e^{-\frac{\alpha_M}{\alpha_m} N}\right)\right) + \ln|\Omega_N|}{\ln|\Omega_N|} \qquad (5.3) \\ &= 1 - \frac{\alpha_m}{\ln \mu_{SAW}} < 1. \end{aligned}$$

$\square$

Now we consider the case of bond-self-avoiding walks (BAW). We introduce some notation: we consider a cube D as for SAW and a cube $D^1 = \{x \in \mathbb{Z}^n \text{ s.t. } c_i - 1 \leq x^{(i)} \leq c_i + b + 1, 1 \leq i \leq n\}$ for some $c = (c_1, \ldots, c_n) \in \mathbb{Z}^n$. In this case a path P is a BAW instead of a SAW with the same conditions as for SAW. We consider only the paths such that there exists a cube D with $X_0(P) = 0$ and $X_k(P)$ two of its vertices and $X_j(P) \in D^1$ for all $0 \leq j \leq k$. We say that (P, D) occurs at the r^{th} step ω if

1. $X_{r+j}(\omega) - X_r(\omega) = X_j(P)$ for all $j = 0, \ldots, k$ and

2. ω does not occupy any other points of D.

$\chi_N(j, (P, D))$ is the number of $\omega \in \Omega_N$ such that (P, D) occurs *at most* at j steps. Theorem 5.1 holds also for BAW [3].

Proposition 5.2. *For BAW*

$$\limsup_{N \to \infty} \gamma_N < 1. \tag{5.4}$$

Proof. The proof is identical to the one of Theorem 2.3. $\square$

Remark 5.3. *As noticed by Kesten in [8], Theorem 5.1 could be proven also for other lattices in almost the same way, therefore Theorem 2.3 should hold for other lattices than $\mathbb{Z}^n$.*

Acknowledgments

We thank H. Kesten for supplying us with the argument needed to prove Theorem 2.3 for SAW, and S. Goldstein for useful discussions. Research supported by NSF Grant DMR 98-13268, AFOSR Grant AF 49620-01-1-0154, and DIMACS and its supporting agencies, the NSF under Contract

No. STC-91-19999 and the N.J. Commission on Science and Technology. Work of P.L. Ferrari partially supported by the Swiss fellowship Sunburst-Fonds.

A Outline of the proof of Proposition 3.2

Let $I_N = N + 1 - R_N$ be the number of intersections. Consider an interval $J = [k_0, k_1] \subset [0, 1]$ and the subset $\Lambda_N(J) = \{\omega \in \Omega_N \text{ s.t. } I_N(\omega)/N \in J\}$. We define the mean degeneracy on $\Lambda_N(J)$ by $\langle \deg \mathcal{C} \rangle_J = |\Lambda_N(J)|/W(N)_J$, where $W(N)_J$ is the number of contact matrices corresponding to RW with $I_N/N \in J$.We set $d(J) = \liminf_{N\to\infty} \frac{1}{N} \ln \langle \deg \mathcal{C} \rangle_J$.

Theorem A.1. *For SRW on $\mathbb{Z}^n$, $n \geq 2$, and $\pi' = \lim_{N\to\infty} \mathbb{E}(I_N)/N$,*

$$d(J = [k_0, k_1]) > 0 \textit{ for all } k_0 < \pi' \textit{ and } k_0 < k_1 < 1. \tag{A.1}$$

For an $\omega \in \Omega_N$, let us define $F(\omega)$ to be the number of loops of length 4 which do not intersect the remaining part of ω (called "free-4-loops").

Proposition A.2. *Let J be as in Theorem A.1. Then there exists an $\alpha_J > 0$ such that*

$$\beta_J = \liminf_{N\to\infty} -\frac{1}{N} \ln \mathbb{P}\{\omega \in \Lambda_N(J) \textit{ s.t. } F(\omega) \leq \alpha_J N\} > 0. \tag{A.2}$$

Proof of Theorem A.1. For $k_0 < \pi'$ and $k_1 \in (k_0, 1)$,

$$\begin{aligned} \frac{W(N)_J}{|\Lambda_N(J)|} &\leq \mathbb{P}\{F(\omega) \leq \alpha_J N \text{ for } \omega \in \Lambda_N(J)\} \qquad (A.3)\\ &+ \; 2^{-\alpha_J N} \mathbb{P}\{F(\omega) > \alpha_J N \text{ for } \omega \in \Lambda_N(J)\} \leq 2\exp(-\min\{\beta_J, \alpha_J \ln 2\} N) \end{aligned}$$

since a contact matrix with M free-4-loops is at least 2^M times degenerate. Then it follows by Proposition A.2 that $d(J) \geq \min\{\beta_J, \alpha_J \ln 2\} > 0$. $\square$

Outline of the proof of Proposition A.2: Divide $\mathbb{Z}^n$ into disjoint n-cubes of edgelength 4. First we remark that at least aN cubes are visited by $\omega \in \Lambda_N(J)$, $a = (1-k_1)/4^n$, and at least $aN/2$ are visited at most by $2/a$ steps. Consider $\Lambda_N^{\alpha N}(J) = \{\omega \in \Lambda_N(J) \text{ s.t. } F(\omega) \leq \alpha N\}$, $\alpha \ll 1$. We do two successive operations on $\omega \in \Lambda_N^{\alpha N}(J)$.

1) We modify the free-4-loops so that the new RW $\widetilde{\omega}$ has $F(\widetilde{\omega}) = 0$. This is obtained by moving the 3rd step to the position of the 1st step of the free-4-loops.

2) We choose $2\alpha N$ cubes out of the first $aN/2$ visited less than $2/a$ steps. The choice can be made in $\binom{aN/2}{2\alpha N}$ different ways. $\widetilde{\omega}$ passes in a cube not more that $2/a$ times and at each time we replace the path inside the chosen cubes by another one of length increased by 2 which remains on the boundary of the cube and leaving the enter and exit points unchanged. Therefore the center of the cubes are now empty. Secondly we add a free-4-loop in the center of the cubes the first time that are visited by $\widetilde{\omega}$.

The final RW have length $n \in [N(1+20\alpha), N(1+c_2\alpha)]$ and $I_n/n \leq k_1 + c_2\alpha$, $c_2 = 17 + 8/a$. Then using some results of Hamana and Kesten on $\psi(k) = \lim_{N\to\infty} -\frac{1}{N} \ln \mathbb{P}(R_N/N \geq k)$ [6], we conclude that, if $\beta_J = 0$, for α small enough the number of constructed RW exceeds the total number of RW with $n \in [N(1+20\alpha), N(1+c_2\alpha)]$ and $I_n/n \leq k_1 + c_2\alpha$. Therefore $\beta_J > 0$.

Outline of the proof of proposition 3.2: $\mathbb{P}(R_N/N \leq \varepsilon)$ is not exponentially small in N (see e.g. [6] and proof of (2.1a)). Let $J_1 = [0, 1-\varepsilon)$ and $J_2 = [1-\varepsilon, 1]$. Since $d(J_1) > 0$, $W(N)_{J_1}$ is exponentially small compared with $W(N)_{J_2}$ for N large enough. Therefore for large N, $W(N) \simeq W(N)_{J_2} \leq |\Omega_N| \mathbb{P}(R_N/N \leq \varepsilon)$, from which follows (3.13).

The complete proof can be found at *http://www-m5.mathematik.tu-muenchen.de/pers/ferrari/homepage/download/appendix.ps.gz.*

References

[1] I. Bahar, A.R. Atilgan, and B. Erman, Folding Des. **2**, 173 (1997).

[2] P. Billingsley, *Ergodic Theory and Information*, New York, Wiley ed. (1965).

[3] P.L. Ferrari, *Contact matrices for random walks*, EPFL-Lausanne diploma thesis.

[4] K.F. Freed, *Polymers as self-avoiding walks*, Ann. Probab. **9**, 837–556 (1981).

[5] T. Haliloglu, I. Bahar, and B. Erman, *Gaussian Dynamics of Folded Proteins*, Phys. Rev. Lett. **79**, 3090–3093 (1997).

[6] Y. Hamana and H. Kesten, *A large deviation result for the range of random walk and for the Wiener sausage*, Probab. Theory Relat. Fields, **120** (2), 183–208 (2001).

[7] T.F. Havel, G.M. Crippen, and I.D. Kuntz, Biopolymers, **18**, 73 (1979).

[8] H. Kesten, *On the Number of Self-Avoiding Walks*, J. Math. Phys. **4**, 960 (1963).

[9] K.F. Lau and K.A. Dill, Macromolecules **22**, 3986 (1989).

[10] S. Lifson and C. Sander, Nature (London), **282**, 109 (1979).

[11] N. Madras and G. Slade, *The Self-Avoiding Walk*, Probability and Its Applications, Birkäuser Boston ed. (1996).

[12] Th. M. Nieuwenhuizen, *Trapping and Lifshitz Tails in Random Media, Self-Attracting Polymers and the Number of Distinct Sites Visited:*

A Renormalized Instanton Approach in Three Dimensions, Phys. Rev. Lett. **62** (4), 357–360 (1989).

[13] M. Vendruscolo, B. Subramanian, I. Kanter, E. Domany, and J.L. Lebowitz, *Statistical Properties of Contact Maps*, Phys. Rev. E, **59**, 977–984 (1999).

[14] M. Vendruscolo, E. Kussell, and E. Domany, Folding Des. **2**, 295 (1997).

[15] F. van Wijland, S. Caser, and H.J. Hilhorst, *Topology of the support of the two-dimensional random walk*, J. Phys. A **30**, 507 (1997).

[16] F. van Wijland and H.J. Hilhorst, *Universal Fluctuations in the Support of the Random Walk*, J. Stat. Phys. **89**, 119 (1997).

[17] F. van Wijland, *Thèse*, Université de Paris-Sud, U.F.R. Scientifique d'Orsay (1998).