Journal of Statistical Physics, Vol. 58, Nos. 1/2, 1990

Monte Carlo Generation of Self-Avoiding Walks with Fixed Endpoints and Fixed Length

N. Madras,[1] A. Orlitsky,[2] and L. A. Shepp[2]

Received July 10, 1989

We propose a new class of dynamic Monte Carlo algorithms for generating self-avoiding walks uniformly from the ensemble with fixed endpoints and fixed length in any dimension, and prove that these algorithms are ergodic in all cases. We also prove the ergodicity of a variant of the pivot algorithm.

KEY WORDS: Self-avoiding walk; self-avoiding polygon; Monte Carlo; ergodicity; pivot algorithm.

1. INTRODUCTION

The self-avoiding walk (SAW) is a widely studied lattice model of a polymer molecule with excluded volume.[1,2] It is also important in the study of critical phenomena, since it is equivalent to the $N = 0$ limit of the N-vector model.[3]

Through the years, many dynamic Monte Carlo algorithms have been used to study the SAW. Their strategy is as follows: begin with an arbitrary SAW; make some random change to it to get another SAW; repeat. This produces a random (correlated) sequence of SAWs, from which statistics may be taken. A "local" algorithm is one in which the random change is always restricted to a bounded number of adjacent bonds (typically three or less). Local methods have been widely used (for a survey, see ref. 4), but they suffer from two difficulties: (1) The resulting sequence of SAWs is highly correlated, and (2) if the changes are length-conserving as well as local, then the algorithm will not be ergodic[5] (i.e., given any initial N-step SAW, there exist many other N-step SAWs which can never be produced by the algorithm).

[1] Department of Mathematics, York University, North York, Ontario, Canada M3J 1P3.
[2] AT & T Bell Laboratories, Murray Hill, New Jersey.

0022-4715/90/0100-0159$06.00/0

In contrast, the pivot algorithm invented by Lal[6] makes nonlocal changes to SAWs of fixed length N: a pivot point is chosen at random on the SAW, and a random reflection/rotation is applied to the part of the SAW subsequent to the pivot point, using the pivot point as the origin. Several attempts may be necessary, but when a change is successful, the resulting SAW will likely be "very different" (with respect to global observables such as end-to-end distance). In practice, this Monte Carlo procedure seems to produce an "effectively independent" observation in computer time of order N (or perhaps $N \log N$), in the sense that the variance of a global observable is proportional to N (or perhaps $N \log N$) divided by the total amount of computer time (see ref. 7, Sections 3.2–3.4 for more details). In addition, it is ergodic (any N-step SAW can be transformed into any other by a sequence of allowed changes), and satisfies detailed balance (or "reversibility") for the usual equal-weight probability distribution on the set of N-step SAWs.

We remark that nonlocal changes are usually rather "nonphysical"; thus, if one wishes to investigate the dynamics of polymer molecules, then local changes may be more appropriate. However, if one is interested only in equilibrium expectation values of a particular model, then there can be no objection to nonlocal moves if they are more efficient. Indeed, in view of (2) above, any algorithm for the fixed-length ensemble of SAWs in principle MUST use nonlocal changes.

It is also of interest to study the subclass of SAWs with fixed endpoints. In particular, if the endpoints are nearest neighbors, then this is the subclass of "self-avoiding polygons" (which one can think of as simple closed curves). Berg and Foerster[8] and Aragão de Carvalho *et al.*[9,10] ("BFACF") proposed a "local" Monte Carlo algorithm (see Fig. 1) that generates SAWs with fixed endpoints, allowing N to change. This algorithm can be very slow; in fact, its exponential autocorrelation time is infinite.[11] Moreover, there are ergodicity problems in three dimensions, since the knot type of a self-avoiding polygon cannot be changed by the BFACF moves.

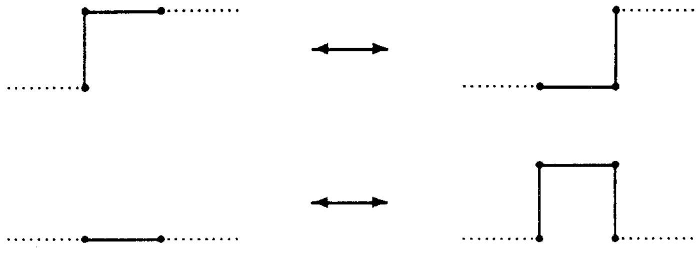

Fig. 1. Moves in the BFACF algorithm. The second move changes the length by 2.

To investigate the fixed-endpoints ensemble, one would like a set of nonlocal moves which could speed up the BFACF algorithm and make it ergodic. Dubins *et al.*[12] devised a length-conserving algorithm for self-avoiding polygons in two dimensions and proved its ergodicity.

In this paper, we propose a new class of algorithms for fixed-length, fixed-endpoints ensembles of SAWs on the simple (hyper) cubic lattice in any dimension. We also prove that they are ergodic in all cases; in particular, knots appear and disappear routinely, without requiring any special attention. The proof proceeds by induction on the dimension. It is interesting to observe that our proof of ergodicity in d dimensions, even for self-avoiding polygons, relies on the knowledge of ergodicity for all fixed-endpoints ensembles in $d-1$ dimensions.

We have not yet conducted an empirical estimation of the autocorrelation times of these new algorithms, but by analogy to Lal's pivot algorithm, it seems reasonable to hope that they produce an "effectively independent" observation in computer time of order N^2 (or better), depending on the observables in question. One can also consider algorithms which combine our nonlocal moves with the local BFACF moves to sample SAWs with fixed endpoints of varying length. These algorithms should be improvements over pure BFACF with respect to autocorrelation times as well as ergodicity. One such algorithm has recently been proposed and studied by Caracciolo *et al.*[14]; we also prove the ergodicity of this algorithm.

The paper is organized as follows. Section 2 defines the algorithms and sets the notation. Section 3 proves ergodicity in two dimensions, and Section 4 proves ergodicity in all other dimensions. Finally, Section 5 resolves a related question that was left open in ref. 7 regarding ergodicity of the variant of the pivot algorithm which only uses diagonal reflections.

2. STATEMENT OF RESULTS

We begin with notation and definitions. For a point x in d-dimensional space $\mathscr{R}^d$, let $(x^{(1)},..., x^{(d)})$ be its coordinate vector. Its l_1 norm is $|x| = |x^{(1)}| + \cdots + |x^{(d)}|$. The d-dimensional integer lattice is

$$\mathscr{Z}^d \stackrel{\text{def}}{=} \{(x^{(1)},..., x^{(d)}): x^{(i)} \text{ is an integer for } i = 1,..., d\}$$

An *N-step self-avoiding walk (SAW)* w in $\mathscr{Z}^d$ is a sequence $w_0, w_1,..., w_N$ of $N+1$ distinct points in $\mathscr{Z}^d$ such that each point is a nearest neighbor of its predecessor: $|w_i - w_{i-1}| = 1$ for $i = 1,..., N$. The points w_0 and w_N are the *endpoints* of w. For points A and B in $\mathscr{Z}^d$, let $S^N(A, B)$ be the set of all N-step SAWs having $w_0 = A$ and $w_N = B$. Obsrve that $S^N(A, B)$ is empty unless $N - |A - B|$ is a nonnegative even integer.

Given two points x_1 and x_2 in the d-dimensional real space $\mathscr{R}^d$, let $[x_1, x_2]$ be the line segment $\{tx_2 + (1-t)\,x_1 : 0 \leqslant t \leqslant 1\}$ oriented in the direction of increasing t. Given a sequence of points $x_0,..., x_k$ in $\mathscr{R}^d$, let $[x_0,..., x_k]$ be the piecewise-linear curve resulting from the concatenation of $[x_0, x_1]$, $[x_1, x_2]$,..., $[x_{k-1}, x_k]$ with the associated orientations. Observe that if $N \geqslant 2$ and $w_0, w_1,..., w_N$ is a SAW with $|w_0 - w_N| = 1$, then $[w_0, w_1,..., w_N, w_0]$ is a simple (i.e., non-self-intersecting) closed curve: a *self-avoiding polygon.*

Our dynamic Monte Carlo algorithm on the fixed-length, fixed-endpoints ensemble $S^N(A, B)$ works as follows. There is a finite set $\mathscr{F} \stackrel{\text{def}}{=} \{F_1,..., F_r\}$ of transformations of $S^N(A, B)$ into itself (here, $\mathscr{F}$ and r depend on N). Begin at time $t=0$ with any SAW $w^{[0]}$ in $S^N(A, B)$. At each successive integer time t, knowing $w^{[t-1]}$, choose a number $n(t)$ at random from $\{1,..., r\}$ according to a fixed probability distribution[3] (e.g., uniform), and put $w^{[t]} = F_{n(t)}(w^{[t-1]})$. The resulting sequence $w^{[0]}$, $w^{[1]}$, $w^{[2]}$,... of SAWs is a Markov chain on $S^N(A, B)$. The distribution of $w^{[t]}$ as $t \to \infty$ will converge to the uniform distribution on $S^N(A, B)$ if (e.g., ref. 15):

A1. The transition probabilities are *symmetric*: for every $w', w'' \in S^N(A, B)$,

$$\text{Prob}\{w^{[t]} = w' \mid w^{[t-1]} = w''\} = \text{Prob}\{w^{[t]} = w'' \mid w^{[t-1]} = w'\}$$

A2. For some $w' \in S^N(A, B)$:

$$\text{Prob}\{w^{[t]} = w' \mid w^{[t-1]} = w'\} > 0$$

(therefore, the chain is a *aperiodic*).

A3. The chain is *ergodic*: for every w' and w'' in $S^N(A, B)$, there is a $t > 0$ such that

$$\text{Prob}\{w^{[t]} = w' \mid w^{[0]} = w''\} > 0$$

In our algorithm, (A1) will follow directly from the fact that each F_i is its own inverse (it would suffice, in fact, just to have each F_i invertible and choose F_i and F_i^{-1} with equal probabilities); (A2) will follow as each w' is a fixed point of at least one $F_i \in \mathscr{F}$. The bulk of this paper is devoted to the proof of (A3).

[3] In other Monte Carlo algorithms the set of transformations and the probability distribution can depend on the current state, but the algorithms introduced in this paper are of the type discussed above.

We will now define the transformations F_i. To do this, we will first define some transformations T_i which perturb SAWs into objects that may or may not be SAWs. Then, for w in $S^N(A, B)$, we will put

$$F_i(w) = \begin{cases} T_i(w) & \text{if } T_i(w) \in S^N(A, B) \\ w & \text{if } T_i(w) \notin S^N(A, B) \end{cases}$$

In words: F_i attempts to deform w; the deformation is accepted if the result is a SAW and rejected otherwise.

Given a SAW $w = (w_0, ..., w_N)$ and integers k and l such that $0 \leqslant k < l \leqslant N$, define the *inversion* $T^{\text{inv}}_{k,l}(w)$ to be the sequence $w' = (w'_0, ..., w'_N)$ given by

$$w'_i = \begin{cases} w_k + w_l - w_{k+l-i} & \text{if } k \leqslant i \leqslant l \\ w_i & \text{otherwise} \end{cases}$$

Thus, $[w'_k, ..., w'_l]$ is the curve $[w_l, ..., w_k]$ inverted through the point $(w_k + w_l)/2$. For $d = 2$, this is a 180° rotation around $(w_k + w_l)/2$ as shown in Fig. 2.

Another way to view this transformation is via the sequence of *steps* $s_1, s_2, ..., s_N$ in w, where $s_i \stackrel{\text{def}}{=} w_i - w_{i-1}$. The steps of $T^{\text{inv}}_{k,l}(w)$ are $s_1, s_2, ..., s_k, s_l, s_{l-1}, ..., s_{k+2}, s_{k+1}, s_{l+1}, ..., s_N$. Observe that the inversions $T^{\text{inv}}_{k,k+2}(w)$ are precisely the length-preserving BFACF moves.

Next we define transformations which reflect a piece of a SAW through a hyperplane which makes angles of 45° with exactly two of the coordinate hyperplanes. Consider the case $d = 2$ first. For $w \in S^N(A, B)$, $0 \leqslant k < l \leqslant N$, and $m \in \{-1, +1\}$ define $T^{\text{ref},m}_{k,l}(w)$ as follows. If $w^{(1)}_l - w^{(1)}_k \neq$

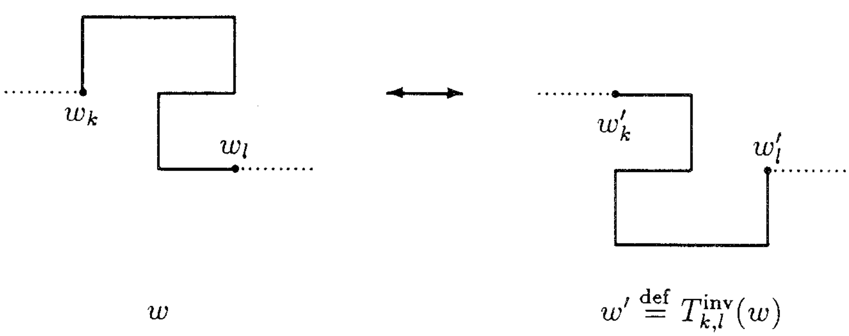

Fig. 2. Inversion transformation in $\mathscr{Z}^2$.

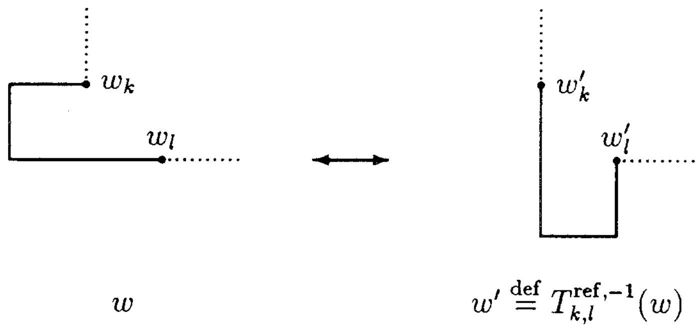

Fig. 3. Reflection in the perpendicular bisector of $[w_k, w_l]$.

$m(w_l^{(2)} - w_k^{(2)})$, then put $T^{\text{ref},m}_{k,l}(w) = w$; otherwise, put $T^{\text{ref},m}_{k,l}(w) = (w'_0, ..., w'_N)$, where

$$w'_i = \begin{cases} (w_k^{(1)} - m(w_{k+l-i}^{(2)} - w_l^{(2)}),\, w_k^{(2)} - m(w_{k+l-i}^{(1)} - w_l^{(1)})) & \text{if} \quad k \leqslant i \leqslant l \\ w_i & \text{otherwise} \end{cases}$$

Thus $[w'_k, ..., w'_l]$ is a *reflection* of $[w_k, ..., w_l]$ through the perpendicular bisector of $[w_k, w_l]$ if the line $[w_k, w_l]$ has slope m. This is illustrated in Fig. 3.

For $d \geqslant 3$, we proceed similarly. For $w \in S^N(A, B)$, $0 \leqslant k < l \leqslant N$, $m \in \{-1, +1\}$, and $1 \leqslant \alpha < \beta \leqslant d$, define $T^{\text{ref},m}_{k,l;\alpha,\beta}(w)$ as follows. If $w_l^{(\alpha)} - w_k^{(\alpha)} \neq m(w_l^{(\beta)} - w_k^{(\beta)})$ or if $w_l^{(\gamma)} \neq w_k^{(\gamma)}$ for some $\gamma \neq \alpha, \beta$, then put $T^{\text{ref},m}_{k,l;\alpha,\beta}(w) = w$; otherwise, put $T^{\text{ref},m}_{k,l;\alpha,\beta}(w) = (w'_0, ..., w'_N)$, where

$$(w'_i)^{(\gamma)} = \begin{cases} w_k^{(\gamma)} - m(w_{k+l-i}^{(\alpha+\beta-\gamma)} - w_l^{(\alpha+\beta-\gamma)}) & \text{if} \quad k \leqslant i \leqslant l \quad \text{and} \quad \gamma \text{ is } \alpha \text{ or } \beta \\ w_i^{(\gamma)} & \text{otherwise} \end{cases}$$

In three or more dimensions we require one more class of transformations. For $w \in S^N(A, B)$, $0 \leqslant k < l \leqslant N$, $m \in \{-1, +1\}$, and $1 \leqslant \alpha < \beta \leqslant d$, define $T^{\text{int},m}_{k,l;\alpha,\beta}(w)$ as follows. If $w_l^{(\alpha)} - w_k^{(\alpha)} \neq m(w_l^{(\beta)} - w_k^{(\beta)})$, then put $T^{\text{int},m}_{k,l;\alpha,\beta}(w) = w$; otherwise, let $w' \stackrel{\text{def}}{=} T^{\text{int},m}_{k,l;\alpha,\beta}(w)$ be the N-step walk whose steps $s'_i \stackrel{\text{def}}{=} w'_i - w'_{i-1}$ are

$$s_i'^{(\gamma)} \stackrel{\text{def}}{=} \begin{cases} m \cdot s_i^{(\beta)} & \text{if} \quad k < i \leqslant l \quad \text{and} \quad \gamma = \alpha \\ m \cdot s_i^{(\alpha)} & \text{if} \quad k < i \leqslant l \quad \text{and} \quad \gamma = \beta \\ s_i^{(\gamma)} & \text{otherwise} \end{cases}$$

The *interchange* transformation $T^{\text{int},m}_{k,l;\alpha,\beta}(w)$ interchanges the α and β coordinates of the steps $s_{k+1}, ..., s_l$; $T^{\text{int},+1}_{k,l;\alpha,\beta}(w)$ keeps the orientation of the interchanged coordinates, while $T^{\text{int},-1}_{k,l;\alpha,\beta}(w)$ reverses them. Notice that if

$d \geqslant 3$ and w is contained in the hyperplane $\{x: x^{(1)}=0\}$, say, then $T^{\text{ref},m}_{k,l;\alpha,\beta}(w)$ and $T^{\text{inv}}_{k,l}(w)$ will also be contained in that hyperplane; the only way to break out of the hyperplane is using $T^{\text{int},m}_{k,l;1,\alpha}(w)$, where $s_i^{(\alpha)} \neq 0$ for some $i \in \{k+1,..., l\}$. Thus, the T^{int} transformations are necessary in three or more dimensions; as shown in Section 3, they are not necessary for ergodicity in two dimensions.

The transformations T_i are now defined in d dimensions: there are

$$\binom{N+1}{2}\left(1+4\binom{d}{2}\right)$$

of them. A dynamic Monte Carlo algorithm may now be constructed as described above, using the family of transformations

$$\mathscr{F}^N_d = \{F^{\text{inv}}_{k,l}, F^{\text{ref},m}_{k,l;\alpha,\beta}, F^{\text{int},m}_{k,l;\alpha,\beta}: 0 \leqslant k < l \leqslant N, 1 \leqslant \alpha < \beta \leqslant d, \text{ and } m \in \{-1, 1\}\}$$

obtained from the corresponding T's. We say that two SAWs w and w' are *directly connected* if $w' = F(w)$ for some F in $\mathscr{F}^N_d$ [equivalently, if $w = F(w')$ for some F in $\mathscr{F}^N_d$; observe that each F is its own inverse]. We say that w and w' are *connected* if there is a sequence of SAWs $w[0],..., w[p]$ in $S^N(A, B)$ such that $w[0] = w$, $w[p] = w'$, and $w[i-1]$ is directly connected to $w[i]$ for each $i = 1,..., p$. Thus, ergodicity is equivalent to the property that any two SAWs in $S^N(A, B)$ are connected. This property will be proven in the next section for $d = 2$ and in Section 4 for higher dimensions.

Remarks.

B1. In practice, the following Monte Carlo algorithm (or a variation of it) may work more efficiently than the straightforward procedure described above. First, pick a pair k, l from among all pairs $1 \leqslant k < l \leqslant N$ with some fixed probability distribution (e.g., uniform). Then, knowing $w^{[t-1]} = w$, make a "short list" of all the F's that are "candidates": $F^{\text{inv}}_{k,l}$ is always on the list; $F^{\text{ref},m}_{k,l;\alpha,\beta}$ is on the list if $w_l^{(\alpha)} - w_k^{(\alpha)} = m(w_l^{(\beta)} - w_k^{(\beta)})$ and $w_k^{(\gamma)} = w_l^{(\gamma)}$ for all $\gamma \neq \alpha, \beta$; $F^{\text{int},m}_{k,l;\alpha,\beta}$ is on the list if $w_l^{(\alpha)} - w_k^{(\alpha)} = m(w_l^{(\beta)} - w_k^{(\beta)})$. Choose an F uniformly from this short list and put $w^{[t]} = F(\omega)$. It is not hard to see that the transition probabilities are symmetric for this algorithm.

B2. When implementing this algorithm, one could generalize the definition of $T^{\text{ref},m}_{k,l;\alpha,\beta}(w)$ for $d \geqslant 3$ as follows. If $w_l^{(\alpha)} - w_k^{(\alpha)} \neq m(w_l^{(\beta)} - w_k^{(\beta)})$, then put $T^{\text{ref},m}_{k,l;\alpha,\beta}(w) = w$; otherwise, put $T^{\text{ref},m}_{k,l;\alpha,\beta}(w) = (w'_0,..., w'_N)$, where

$$(w'_i)^{(\gamma)} = \begin{cases} w_k^{(\gamma)} - m(w_{k+l-i}^{(\alpha+\beta-\gamma)} - w_l^{(\alpha+\beta-\gamma)}) & \text{if } \; k \leqslant i \leqslant l \; \text{ and } \; \gamma \text{ is } \alpha \text{ or } \beta \\ w_i^{(\gamma)} & \text{otherwise} \end{cases}$$

The definition given earlier is only the minimum necessary to ensure ergodicity.

B3. Caracciolo *et al.*[14] have studied an algorithm (henceforth referred to as the CPS algorithm) which uses the BFACF moves and inversions. This generates a Markov chain on

$$S(A, B) \stackrel{\text{def}}{=} \bigcup_{N=0}^{\infty} S^N(A, B)$$

which is the set of all SAWs beginning at A and ending at B. The proofs given in Sections 3 and 4 imply that condition (A3) above holds for the CPS algorithm, as will be briefly shown at the ends of those sections.

3. PROOF OF ERGODICITY IN TWO DIMENSIONS

Ergodicity will be proven by induction on the dimension d. The first step is $d=2$.

Theorem 1. Fix endpoints A and B in $\mathscr{Z}^2$, and fix a length $N \geqslant |A-B|$ having the same parity as $|A-B|$. Then the Monte Carlo algorithm on $S^N(A, B)$ is ergodic.

Proof. Let $R^N(A, B)$ be the set of all SAWs in $S^N(A, B)$ which are subsets of the boundary of some rectangle; that is, $w \in R^N(A, B)$ if there exists $c \in \mathscr{Z}^2$ and integers $k_1, k_2 > 0$ such that

$$[w_0,..., w_N] \subseteq [c, c+(k_1, 0), c+(k_1, k_2), c+(0, k_2), c]$$

Notice that if $w \in R^N(A, B)$, then $[w_0,..., w_N]$ has at most one right-angle

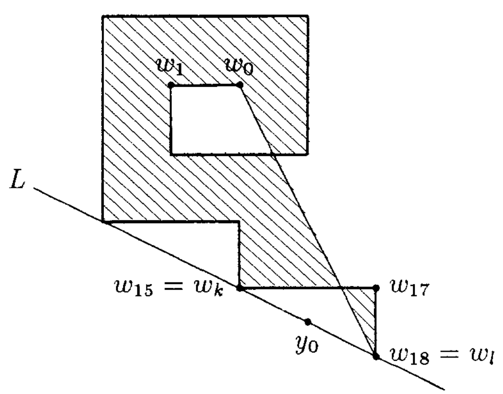

Fig. 4. $W(w)$ (thick lines and $[w_{18}, w_0]$) and $\text{Int}(W(w))$ (shaded region).

turn if $N=|A-B|$ and 2, 3, or 4 right-angle turns if $N>|A-B|$. The theorem will follow once the following two statements are proven:

C1. Any SAW in $S^N(A, B)$ is connected to some SAW in $R^N(A, B)$.

C2. Any two SAWs in $R^N(A, B)$ are connected.

We prove statement (C1) first; statement (C2) will be verified later in this section. For w in $S^N(A, B)$, let $H(w)$ be the boundary of the convex hull of w, considered as a subset of the plane, and let $W(w) \stackrel{\text{def}}{=} [w_0, w_1,..., w_N, w_0]$ be the "trace" of w, including the line connecting w_0 and w_N, as illustrated in Fig. 4. We require the following lemma.

Lemma 1. If $H(w) \subseteq W(w)$, then $w \in R^N(A, B)$.

Proof. If $[w_0, w_1,..., w_N]$ is the line segment $[A, B]$, the result is obvious. Otherwise, $H(w)$ is a simple closed curve. Since $[w_0, w_1,..., w_N]$ does not intersect itself, it cannot contain all of $H(w)$, so the hypothesis implies that some point of (A, B) $(\stackrel{\text{def}}{=} [A, B]\backslash\{A, B\})$ is in $H(w)$. Since $[A, B]$ is in the convex hull of w, it follows that $[A, B]$ is a subset of $H(w)$. Let $U=H(w)\backslash(A, B)$. U is a curve with no self-intersections, having endpoints A $(=w_0)$ and B $(=w_N)$, and U is contained in the curve $[w_0, w_1,..., w_N]$, which also has no self-intersections; hence $U=[w_0, w_1,..., w_N]$ and $H(w)=W(w)$. This implies that each line segment in U is parallel to a coordinate axis. Since $H(w)=U \cup (A, B)$ is the boundary of a convex set, it follows that U consists of at most four line segments and that U is contained in the boundary of some rectangle. ▮

Next, we define a function $f\colon S^N(A, B) \to \mathscr{R}$ such that every w in $S^N(A, B)\backslash R^N(A, B)$ is directly connected to a SAW w' in $S^N(A, B)$ satisfing $f(w')>f(w)$. As $S^N(A, B)$ is a finite set, this will prove statement (C1) above. Following the proof of the theorem, we show that the range of f contains at most $N^4/8+1$ values.

Let W be a closed curve. For a point y in W^c (the complement of W), the *winding number of W around y* is

$$I(y, W) \stackrel{\text{def}}{=} \frac{1}{2\pi i}\int_W \frac{dz}{z-y}$$

(here we view points of the plane as complex numbers). The properties of winding numbers are well known: $I(y, W)$ is an integer, 0 if y is in the unbounded component of W^c, and ± 1 if W is a simple closed curve with y in the bounded component of W^c. Define

$$\text{Ext}(W)=\{y \in W^c\colon I(y, W) \text{ is even}\}$$

$$\text{Int}(W)=\{y \in W^c\colon I(y, W) \text{ is odd}\}$$

Thus, $\text{Ext}(W) \cup \text{Int}(W) \cup W$ is the entire plane. Roughly speaking, y is in $\text{Int}(W)$ if and only if $y \in W^c$ and any curve from y to infinity crosses W an odd number of times. Finally, we define $f(w)$ to be the area of $\text{Int}(W(w))$, as depicted in Fig. 4.

Now suppose $w \in S^N(A, B) \setminus R^N(A, B)$. From Lemma 1 there is a point $y_0 \in H(w) \setminus W(w)$. This y_0 is an interior point of one of the line segments of $H(w)$. Let L be the (infinite) line containing this line segment and let k and l be the unique integers (see Fig. 4) such that:

1. $0 \leqslant k < l \leqslant N$.
2. $w_k,\ w_l \in L$.
3. The line segment $[w_k, w_l]$ contains y_0.
4. $w_j \notin L$ for all $j \in \{k+1,..., l-1\}$.

Let $w' = T_{k,l}^{\text{inv}}(w)$; we show that w' is a SAW. Since w_k and w_l are on the boundary of the convex hull of w, the closed half-plane having boundary L and containing w_{k+1} must contain all of the points of w. Since $w_j \notin L$ for $k < j < l$, this closed half-plane contains none of the points w_j' for $k < j < l$. Hence w' is self-avoiding.

To show that $f(w') > f(w)$, first abbreviate $W = W(w)$ and $W' = W(w')$. Let

$$Q = [w_k, w'_{k+1},..., w'_{l-1}, w_l, w_{l-1},..., w_{k+1}, w_k]$$

as shown in Fig. 5. Here $[w_k,..., w_l]$ and $[w'_k,..., w'_l]$ are simple curves, each residing on a different side of L. Hence Q is a simple closed curve. Aside from the points w_k and w_l, Q is the symmetric difference

$$Q = W \mathbin{\Delta} W' \stackrel{\text{def}}{=} (W \setminus W') \cup (W' \setminus W)$$

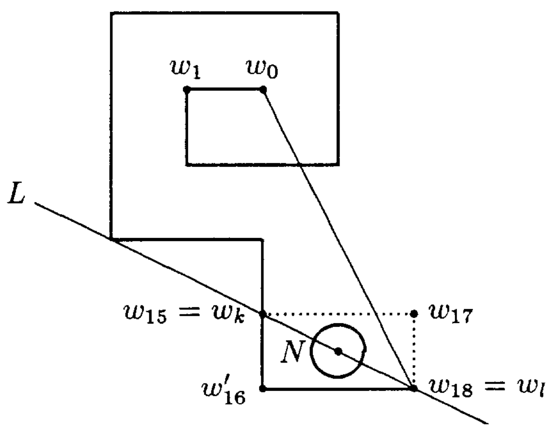

Fig. 5. $W(w')$ and $Q = [w_{15}, w'_{16}, w'_{17}, w_{18}, w_{17}, w_{16}, w_{15}]$.

It follows that for y in $(W \cup W')^c$,

$$I(y, Q) \equiv I(y, W) + I(y, W') \pmod 2$$

So

$$\mathrm{Int}(Q) = \mathrm{Int}(W) \Delta \mathrm{Int}(W') = \mathrm{Int}(W) \backslash \mathrm{Int}(W') \cup \mathrm{Int}(W') \backslash \mathrm{Int}(W)$$

To show that $f(w') > f(w)$, we need to show that

$$\text{area of } [\mathrm{Int}(W') \backslash \mathrm{Int}(W)] > \text{area of } [\mathrm{Int}(W) \backslash \mathrm{Int}(W')]$$

Now, there is a circular neighborhood N centered at y_0 which does not intersect $W \cup W'$ (see Fig. 5). Also, $y_0 \in \mathrm{Int}(W') \backslash \mathrm{Int}(W)$. Therefore:

1. $\mathrm{Int}(W') \backslash \mathrm{Int}(W)$ contains $\mathrm{Int}([w'_k, w'_{k+1}, \ldots, w'_l, w'_k]) \cup N$, which has area

$$\tfrac{1}{2}[\text{area of } \mathrm{Int}(Q)] + \tfrac{1}{2}[\text{area of } (N)]$$

2. $\mathrm{Int}(W) \backslash \mathrm{Int}(W')$ is contained in $\mathrm{Int}([w'_k, w'_{k+1}, \ldots, w'_l, w'_k]) \backslash N$, which has area

$$\tfrac{1}{2}[\text{area of } \mathrm{Int}(Q)] - \tfrac{1}{2}(\text{area of } N)$$

Therefore $f(w') - f(w) \geqslant (\text{area of } N)$; in particular, $f(w') > f(w)$ and statement (C1) is proven.

It remains to prove statement (C2). We distinguish between two cases: $N = |A - B|$ and $N > |A - B|$.

1. $N = |A - B|$: If $[A, B]$ is parallel to a coordinate axis, then $S^N(A, B)$ consists of only one SAW. Otherwise, there are exactly two SAWs in $R^N(A, B)$, say w and w', and $w' = T^{\mathrm{inv}}_{0,N}(w)$.

2. $N > |A - B|$: By rotating the coordinate system if necessary, we can assume $A^{(1)} < B^{(1)}$ and $A^{(2)} \geqslant B^{(2)}$. Let w^* be the unique SAW in $R^N(A, B)$ having $w^*_1 = A - (0, 1)$ and $w^*_{N-1} = B - (0, 1)$. Let w be an arbitrary SAW in $R^N(A, B)$; we will show that w is connected to w^*. We may assume that either $\min\{w_j^{(2)}: 0 \leqslant j \leqslant N\} < \min\{A^{(2)}, B^{(2)}\}$ or there exists an integer $J < A^{(1)}$ such that $[w_0, w_1, \ldots, w_N] = [A, (J, A^{(2)}), (J, B^{(2)}), B]$ [if not, replace w by $T^{\mathrm{inv}}_{0,N}(w)$]. In particular, w_1 is either $A - (0, 1)$ or $A - (1, 0)$. If $w_1 = A - (0, 1)$, put $w' = w$; otherwise, we define w' as follows. Let

$$k \overset{\text{def}}{=} \max\{i: w_i = A - (i, 0)\}$$

$$l \overset{\text{def}}{=} \max\{i: w_i = w_k - (0, i - k)\}$$

$$m \overset{\text{def}}{=} \max\{i: w_i = w_l + (0, i - l)\}$$

Thus w_k, w_l, and w_m are three of the corners of the rectangle determined by w. Now define (see Fig. 6) the SAWs

$$w^{[1]} = T^{\text{inv}}_{0,l-1}(w)$$
$$w^{[2]} = T^{\text{ref},-1}_{l-k-1,l+k+1}(w^{[1]})$$
$$w' = T^{\text{inv}}_{l+1,m}(w^{[2]})$$

Now w' is in $R^N(A, B)$, $w'_1 = A - (0, 1)$ and w'_{N-1} equals either $B - (0, 1)$ or $B + (1, 0)$. If $w'_{N-1} = B - (0, 1)$, then $w' = w^*$ and we are done. Otherwise, $w'_{N-1} = B + (1, 0)$, and we can repeat the algorithm of the preceding paragraph (with the necessary trivial modifications) to show that w' is connected to a SAW w'' in $R^N(A, B)$ with $w''_{N-1} = B - (0, 1)$ and $w''_1 = A - (0, 1)$; but then $w'' = w^*$ and Theorem 1 is proven. ∎

We now show that the function $f(w) \stackrel{\text{def}}{=}$ area of Int($W(w)$), defined in the above proof, attains at most $N^4/8 + 1$ values. This implies that every SAW is connected to a SAW in $R^N(A, B)$ via a sequence of at most $N^4/8$ inversions. The last part of the proof above implies that any two SAWs in $R^N(A, B)$ are connected via a sequence of at most 16 transformations. It follows that every two SAWs in $S^N(A, B)$ are connected via a sequence of at most $N^4/4 + 16$ transformations.

If $A^{(1)} = B^{(1)}$ or $A^{(2)} = B^{(2)}$, then $W(w)$ follows grid lines and the area $f(w)$ is an integer between 0 and $N^2/4$. Otherwise, assume, without loss of generality, that $A^{(1)} < B^{(1)}$ and $A^{(2)} < B^{(2)}$. Redraw w in the plane, mapping $w_i = (w_i^{(1)}, w_i^{(2)})$ to the planar point $((B^{(2)} - A^{(2)})\, w_i^{(1)},\ (B^{(1)} - A^{(1)})\, w_i^{(2)})$. This ensures that the line $[A, B]$ is at 45° with the coordinate axes and goes through at least one grid point. Hence, the area of the redrawn Int($W(w)$) is an integer multiple of 1/2. Yet the redrawn area is $(B^{(1)} - A^{(1)})(B^{(2)} - A^{(2)})$ times larger than the (original) area of Int($W(w)$). Thus, the area of Int($W(w)$) is an integer multiple of $1/[2(B^{(1)} - A^{(1)})(B^{(2)} - A^{(2)})]$. The area of Int($W(w)$) is at most $N^2/4$; hence $f(w)$ attains at most $(N^2/4)\, 2(B^{(1)} - A^{(1)})(B^{(2)} - A^{(2)}) + 1 \leqslant N^4/8 + 1$ values.

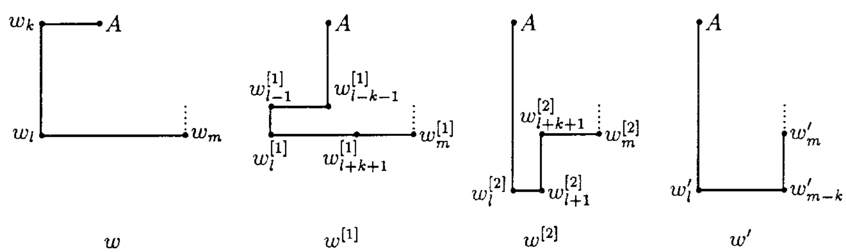

Fig. 6. Transforming a SAW in $R^N(A, B)$ into another.

Corollary 1. The CPS algorithm is ergodic in $\mathscr{Z}^2$.

Proof. As shown above, any SAW in $S^N(A, B)$ is connected to some SAW in $R^N(A, B)$ by a sequence of inversions. It is easy to see that any two SAWs in $\bigcup_{N=0}^{\infty} R^N(A, B)$ are connected by BFACF moves. [We remark that the subset of length-conserving BFACF moves alone does not connect all SAWs in $S^N(A, B)$; see ref. 5.] ∎

4. PROOF OF ERGODICITY IN HIGHER DIMENSIONS

Throughout this section, we assume that the dimension d is at least 3. Let $w = w_0, \ldots, w_N$ be a SAW in $\mathscr{Z}^d$. For $i \in \{1, \ldots, N\}$, exactly one of the d components of the ith step $s_i \stackrel{\text{def}}{=} w_i - w_{i-1}$ is ± 1 and the rest are zero. If $s_i^{(\delta)} = \pm 1$, we say that s_i is a step in the δth coordinate and write $\text{coor}(s_i) = \delta$. A sequence $w_j, \ldots, w_k$ where $0 \leqslant j \leqslant k \leqslant N$ is a *segment* of w. Its *length* is $k - j$; it steps are $s_{j+1}, \ldots, s_k$. A segment $w_j, \ldots, w_k$ is *straight* if all its steps are identical: $s_{j+1} = \cdots = s_k$ (a segment of length 0 or 1 is always straight). A straight segment is in the δth coordinate if one (hence all) of its steps are in the δth coordinate. A straight segment of w is *maximal* if it is not contained in any other straight segment of w. A maximal straight segment has length of at least 1. Every SAW "decomposes" uniquely into maximal straight segments that intersect only at their endpoints.

Let w be a SAW that decomposes into m maximal straight segments, and suppose that the ith segment is in the δ_ith coordinate. w is *canonical* if (1) $\delta_i \neq \delta_j$ for all $1 \leqslant i < j \leqslant m$, or (2) $\delta_i \neq \delta_j$ for all $1 \leqslant i < j < m$ and $s_1 = -s_N$.

In the first case, the length of w is $|w_N - w_0|$ and w is a *minimal length* canonical walk. The two types of canonical SAWs in $\mathscr{Z}^3$ are illustrated in Fig. 7.

Let $A, B \in \mathscr{Z}^d$. We show that

D1. Every SAW in $S^N(A, B)$ can be transformed into a canonical SAW in $S^N(A, B)$ via at most $(\frac{1}{4}N^4 + 18)(N^2 + 1)^{d-2} - 2$ transformations in $\mathscr{F}_d^N$.

D2. Any canonical SAW in $S^N(A, B)$ can be transformed into any other canonical SAW in $S^N(A, B)$ via at most $d^2/2$ transformations in $\mathscr{F}_d^N$.

Fig. 7. Canonical SAWs.

Therefore, any SAW in $S^N(A, B)$ can be transformed into any other SAW in $S^N(A, B)$ via at most $2[(\frac{1}{4}N^4+18)(N^2+1)^{d-2}-2]+d^2/2$ transformations in $\mathscr{F}_d^N$. For large N, this is less than N^{2d}. We begin by proving statement (D1); statement (D2) will be verified in Lemma 3. We first need the following simple lemma.

Lemma 2. Let $w \in S^N(A, B)$ be a minimal-length canonical SAW that is not straight and let $x \in \{w_1, ..., w_{N-1}\}$. Then $x \notin T_{0,N}^{\text{inv}}(w)$.

Proof. $x = w_I$ for some $0 < I < N$. Since the walk is not straight, there are two distinct coordinates δ_1 and δ_2 such that $\delta_1 = \text{coor}(s_i)$ for some $i \leqslant I$ and $\delta_2 = \text{coor}(s_j)$ for some $j > I$. Since w is a minimal-length canonical SAW, (1) $x^{(\delta_1)} \neq A^{(\delta_1)}$ and (2) $x^{(\delta_2)} \neq B^{(\delta_2)}$.

Suppose that $x \in \omega'$. Then $x = w'_J$ for some $0 < J < N$. Now $w_J'^{(\delta_2)} = x^{(\delta_2)} \neq B^{(\delta_2)}$ implies that w'_J precedes (or is in the middle of) the segment of w' that is in the δ_2th coordinate. But $w' \stackrel{\text{def}}{=} T_{0,N}^{\text{inv}}(w)$ is a minimal-length canonical SAW in which the (unique) segment in the δ_2th coordinate precedes the (unique) segment in the δ_1th coordinate, hence $x^{(\delta_1)} = w_J'^{(\delta_1)} = A^{(\delta_1)}$, contradicting (1). ∎

We now prove by induction on d that if $A, B \in \mathscr{Z}^d$, then every SAW in $S^N(A, B)$ can be transformed into a canonical SAW in $S^N(A, B)$ via at most

$$f_d \stackrel{\text{def}}{=} (\tfrac{1}{4}N^4+18)(N^2+1)^{d-2}-2$$

transformations in $\mathscr{F}_d^N$. In the last section we saw that any SAW in $\mathscr{Z}^2$ can be transformed into a canonical SAW via at most $\frac{1}{4}N^4+16$ transformations in $\mathscr{F}_2^N$. This provides the induction basis. For $\delta \in \{1, ..., d\}$, a subset H of $\mathscr{Z}^d$ is a *δ-hyperplane* if $x^{(\delta)} = y^{(\delta)}$ for all $x, y \in H$. We prove the induction step by creating successively larger SAWs in d-hyperplanes, then using the induction hypothesis to transform them into canonical SAWs. Appealing to geometric intuition, our terminology refers to the dth coordinate axis as vertical. We therefore say that a point $x \in \mathscr{Z}^d$ is *higher* (*lower*) than a point $y \in \mathscr{Z}^d$ if $x^{(d)} > y^{(d)}$ ($x^{(d)} < y^{(d)}$). We say that x is *directly above* (*directly below*) y if, in addition, $x^{(\delta)} = y^{(\delta)}$ for all $\delta \in \{1, ..., d-1\}$. We also say that a step of w is *vertical* if it is a step in the dth coordinate and that it is *horizontal* otherwise. A segment is vertical if all its steps are vertical and horizontal if all its steps are horizontal. Define

$$\text{top}(w) \stackrel{\text{def}}{=} \max\{w_i^{(d)}: 0 \leqslant i \leqslant N\}$$

to be the dth component of the "highest" vertices in w and let the *top hyperplane*

$$\text{Top}(w) \stackrel{\text{def}}{=} \{x: x^{(d)} = \text{top}(w)\}$$

be the "highest" d-hyperplane containing a vertex of w. All the figures from here on represent our three-dimensional intuition; they are given merely to help visualize the definitions. To prove the induction step, we establish the following claim.

Claim. For all A, $B \in \mathscr{Z}^d$, every SAW $w \in S^N(A, B)$ can either be transformed to a canonical SAW in $S^N(A, B)$ via at most f_{d-1} transformations in $\mathscr{F}_d^N$; or it can be transformed via at most $f_{d-1}+2$ transformations in $\mathscr{F}_d^N$ to a SAW $w' \in S^N(A, B)$ such that (1) $\text{top}(w') > \text{top}(w)$ ("w' has a higher top than w"), or (2) $\text{top}(w') = \text{top}(w)$ and $|w' \cap \text{Top}(w')| > |w \cap \text{Top}(w)|$ ("w' has the same top as w with more vertices in the top hyperplane"). ∎

Since $A^{(d)} \leqslant \text{top}(w) \leqslant A^{(d)} + N$ and $|w \cap \text{Top}(w)| \leqslant N$ for every walk in $S^N(A, B)$, after at most N^2 applications of this process, we obtain a SAW that is either canonical or can be transformed to a canonical SAW using at most f_{d-1} transformations. Therefore, every SAW in $S^N(A, B)$ can be transformed to a canonical SAW. The number of transformations required is at most

$$\begin{aligned}(f_{d-1}+2)\,N^2 + f_{d-1} &= [(\tfrac{1}{4}N^4+18)(N^2+1)^{d-3}]\,N^2 + (\tfrac{1}{4}N^4+18)(N^2+1)^{d-3} - 2\\ &= (\tfrac{1}{4}N^4+18)(N^2+1)^{d-2} - 2\\ &= f_d\end{aligned}$$

Proof of Claim. The proof covers many individual cases and is therefore somewhat involved. A schematic flow chart describing these cases is given in Fig. 16, at the end of the paper.

Consider the intersection of w and $\text{Top}(w)$. There are two possibilities.

1. $w \cap \text{Top}(w)$ consists of several, necessarily disjoint, SAWs. Then there are two integers t_1 and $t_2 > t_1 + 1$ such that w_{t_1}, $w_{t_2} \in \text{Top}(w)$ and $w_i \notin \text{Top}(w)$ for all $t_1 < i < t_2$. In this case, $w' \stackrel{\text{def}}{=} T^{\text{inv}}_{t_1,t_2}(w)$ is a SAW with $\text{top}(w') > \text{top}(w)$ as illustrated in Fig. 8.

2. $w \cap \text{Top}(w)$ consists of a single SAW. Then there are two (possibly identical) integers t_1 and $t_2 \geqslant t_1$ such that $w_i \in \text{Top}(w)$ for all $t_1 \leqslant i \leqslant t_2$, and $w_i \notin \text{Top}(w)$ for all $0 \leqslant i < t_1$ and all $t_2 < i \leqslant N$. Since $w_{t_1}, \ldots, w_{t_2}$ is contained in $\text{Top}(w)$, it is a $(d-1)$-dimensional SAW. Also, w has no other vertices in $\text{Top}(w)$. By the induction hypothesis, $w_{t_1}, \ldots, w_{t_2}$ can be transformed via at most f_{d-1} transformations in $\mathscr{F}_{d-1}^N \subseteq \mathscr{F}_d^N$ into a canonical SAW in $\text{Top}(w)$. We further distinguish between two possibilities.

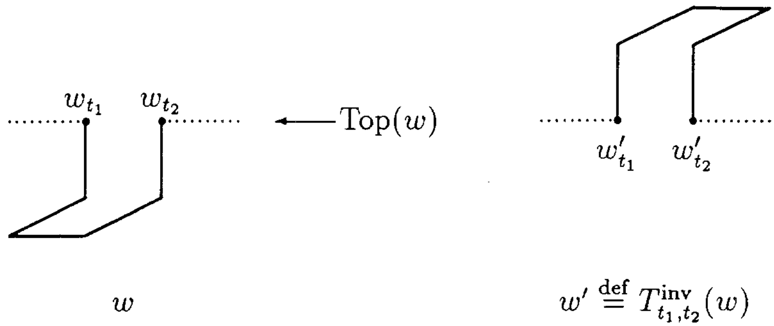

Fig. 8. $w \cap \text{Top}(w)$ consists of several SAWs.

(a) $w_{t_1},..., w_{t_2}$ is not minimal length. Then, the first and last straight segments of $w_{t_1},..., w_{t_2}$ are in the same coordinate, say δ. Necessarily, $\delta \neq d$ and $s_{t_1+1} = -s_{t_2}$; let $m = s^{(\delta)}_{t_1+1}$. Let t' be the penultimate vertex in the first segment of $w_{t_1},..., w_{t_2}$ and let t'' be the second vertex in the last segment. Then $w' \stackrel{\text{def}}{=} T^{\text{int},m}_{t',t'';\delta,d}(w)$ is a SAW with $\text{top}(w') = \text{top}(w) + 1$ as illustrated in Fig. 9.

(b) $w_{t_1},..., w_{t_2}$ is a minimal-length canonical SAW. If the segments $w_0,..., w_{t_1}$ and $w_{t_2},..., w_N$ are both straight, then w is canonical (this also takes care of the case $t_1 = 0$ and $t_2 = N$). If not, then there is an integer $I \in \{1,..., N\} - \{t_1,..., t_2 + 1\}$ such that s_I is horizontal: $s^{(d)}_I = 0$. We distinguish between two cases:

(i) $w_{t_1},..., w_{t_2}$ is straight: $s_{t_1+1} = \cdots = s_{t_2}$ (this case includes $t_1 = t_2$). If the walk is contained in a hyperplane, then by the induction hypothesis it can be transformed into a canonical SAW.[4] Otherwise, let $\delta = \text{coor}(s_{t_2})$; choose a coordinate α different from δ and d. Since w is not contained in a hyperplane, there exists a nonzero real number b such that the hyperplane $\{x \in \mathscr{R}^d\colon\ b(x^{(\alpha)} - w^{(\alpha)}_{t_2}) + (x^{(d)} - w^{(d)}_{t_2}) = 0\}$ contains $w_{t_1},..., w_{t_2}$ and

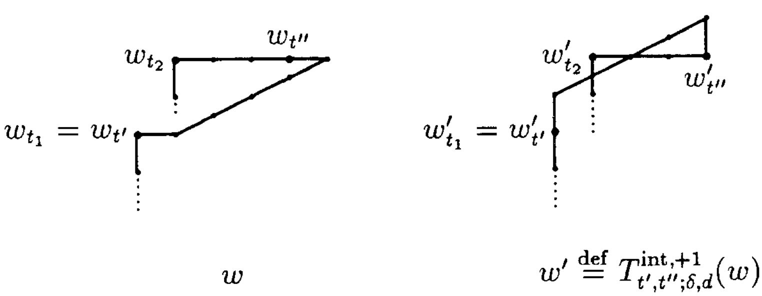

Fig. 9. $w_{t_1},..., w_{t_2}$ is not minimal length.

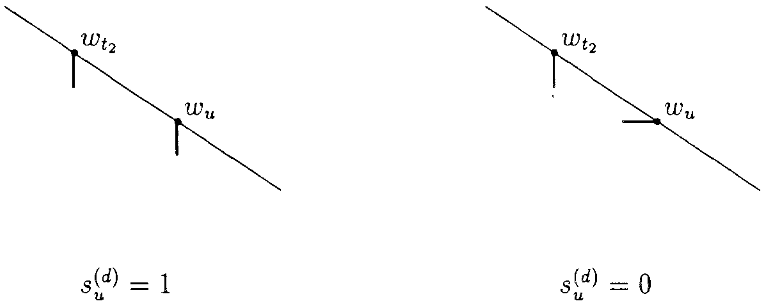

Fig. 10. Two-dimensional view when $w_{t_1},..., w_{t_2}$ is straight.

at least one other vertex w_u of w, and that the half-space $\{x \in \mathscr{R}^d$: $b(x^{(\alpha)} - w_{t_2}^{(\alpha)}) + (x^{(d)} - w_{t_2}^{(d)}) \leqslant 0\}$ contains all of w. Without loss of generality, $u > t_2$ and w_j is not in this hyperplane for every $j \in \{t_2 + 1,..., u - 1\}$. Then $w' \stackrel{\text{def}}{=} T_{t_2,u}^{\text{inv}}(w)$ is self-avoiding. Now, t_2 and u were chosen so that $w_i^{(d)} < w_{t_2}^{(d)} = \text{top}(w)$ for all $t_2 < i \leqslant u$ and $w_{u-1}^{(d)} \leqslant w_u^{(d)}$. Therefore, if $w_i^{(d)} < w_u^{(d)}$ for some $i \in \{t_2 + 1,..., u - 1\}$, then w' has a higher top than w; otherwise, $w'_{t_2+1} \in \text{Top}(w)$ as $s_u^{(d)} = 0$, hence w' has the same top with more vertices. These cases are illustrated in Figs. 10a and 10b.

(ii) The minimal-length canonical SAW $w_{t_1},..., w_{t_2}$ is not straight. This case is more involved. Let

$$\sec(w) \stackrel{\text{def}}{=} \max\{k < \text{top}(w): w_i^{(d)} = w_{i+1}^{(d)} = k \text{ for some } i \in \{0,..., N-1\}\}$$

and let

$$\text{Sec}(w) \stackrel{\text{def}}{=} \{x \in \mathscr{Z}^d: x^{(d)} = \sec(w)\}$$

Visually, if we trace the SAW w, then $\text{Sec}(w)$ is the second highest d-hyperplane containing a horizontal line. Since the segments $w_0,..., w_{t_1}$ and $w_{t_2},..., w_N$ are not both straight, $\sec(w)$ and $\text{Sec}(w)$ are well defined. Without loss of generality, assume that $\text{Sec}(w)$ contains a segment w_i, w_{i+1} for $i > t_2$. Let

$$u_1 \stackrel{\text{def}}{=} \min\{i \in \{t_2,..., N\}: w_i \in \text{Sec}(w)\}$$

and let

$$u_2 \stackrel{\text{def}}{=} \min\{i \in \{u_1 + 1,..., N\}: w_i \in \text{Sec}(w)\}$$

Defined this way, u_2 is the first integer larger than t_2 such that w_{u_2} is in $\text{Sec}(w)$ but is not directly below w_{t_2}. Define τ to be the unique point in

[4] We use a slight generalization of the induction hypothesis that allows the SAWs to be in any $(d-1)$-dimensional coordinate hyperplane of $\mathscr{Z}^d$, not just $\mathscr{Z}^{d-1}$.

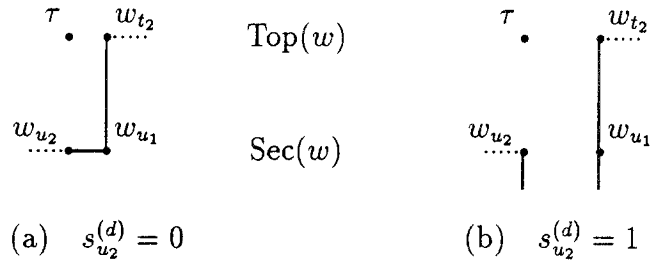

Fig. 11. Two-dimensional view when $w_{t_1},..., w_{t_2}$ is not straight.

Top(w) directly above w_{u_2}. Note that τ is not necessarily in w. There are two possibilities for s_{u_2}. Either $s^{(d)}_{u_2} = 0$, and then $u_2 = u_1 + 1$, or $s^{(d)}_{u_2} = 1$, and then $u_1 + 1 < u_2$ and $w^{(d)}_i < \sec(w) = w^{(d)}_{u_1} = w^{(d)}_{u_2}$ for all $u_1 < i < u_2$. These possibilities are illustrated in Figs. 11a and 11b.

The case where $s^{(d)}_{u_2} = 0$ is easier and we consider it first. At w_{u_2}, the walk can end ($w_{u_2} = B$), continue horizontally ($s^{(d)}_{u_2+1} = 0$), go down ($s^{(d)}_{u_2+1} = -1$), or up ($s^{(d)}_{u_2+1} = 1$). In the first three cases, let $v_2 \stackrel{\text{def}}{=} u_2$. If the walk goes up, then by choice of sec(w), it goes straight up, then ends at B; in that case, let $v_2 \stackrel{\text{def}}{=} N$.

If τ is not in w, then $w' \stackrel{\text{def}}{=} T^{\text{inv}}_{t_2, v_2}(w)$ is self-avoiding with a higher top (if $v_2 \neq u_2$) or with the same top and one more vertex in the top (if $v_2 = u_2$). If $\tau \in w$, then, since $w_{t_1},..., w_{t_2}$ is a minimal-length canonical SAW that is not straight, w_{t_1} cannot be τ (which is adjacent to w_{t_2}). Therefore, by Lemma 2, $w^* \stackrel{\text{def}}{=} T^{\text{inv}}_{t_1, t_2}(w)$ does not contain τ and $|w^* \cap \text{Top}(w^*)| = |w \cap \text{Top}(w)|$. So we can let $w' \stackrel{\text{def}}{=} T^{\text{inv}}_{t_2, v_2}(w^*)$. From here on, we assume that $s^{(d)}_{u_2} = 1$. We distinguish among three possibilities.

A. $\tau \notin w$ (see Fig. 12a). Then $w' \stackrel{\text{def}}{=} T^{\text{inv}}_{t_2, u_2}(w)$ has a higher top. To see that it is self-avoiding, note that the segment $w'_0,..., w'_{t_2}$ is the original self-avoiding segment $w_0,..., w_{t_2}$ and the segment $w'_{u_2},..., w'_N$ is the original segment $w_{u_2},..., w_N$. The segment $w'_{t_2+1},..., w'_{t_2+u_2-u_1-1}$ (originating from $w_{u_1+1},..., w_{u_2-1}$) lies strictly above Top(w) where there are no other vertices of w'. The segment $w'_{t_2+u_2-u_1},..., w'_{u_2}$ (originating from $w_{t_2},..., w_{u_1}$) lies in the straight line connecting τ and w_{u_2} where there are no other vertices of w'.

B. $\tau \in w$ and $\tau \neq w_{t_1}$ (see Fig. 12b). As we did in Case 2b(ii), first let $w^* \stackrel{\text{def}}{=} T^{\text{inv}}_{t_1, t_2}(w)$. Since $w_{t_1},..., w_{t_2}$ is a minimal-length canonical walk, Lemma 2 implies that τ is not in w^*, the new walk has the same top as w and the same number of vertices on top. Hence we are back in case A.

C. $\tau \in w$ and $\tau = w_{t_1}$ (see Fig. 12c). Let w_{u_3} be the last point of w that is in Sec(w). Since Sec(w) contains at least one horizontal segment, $u_3 \neq u_2$. At w_{u_3}, the walk can end ($w_{u_3} = B$), go down ($s^{(d)}_{u_3+1} = -1$), or up

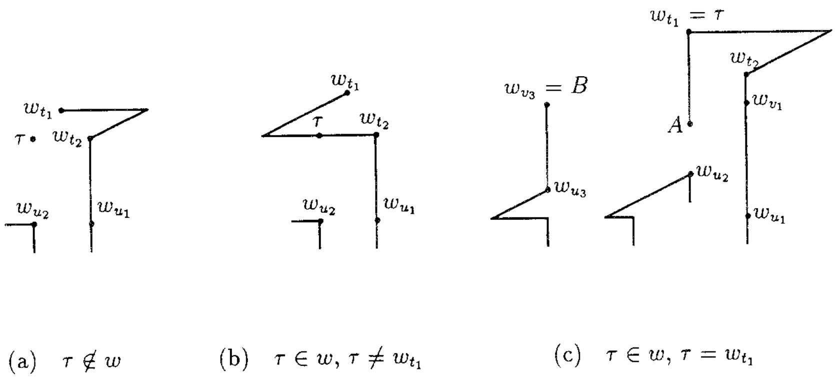

Fig. 12. Three-dimensional view when $s_{u_2}^{(d)} = 1$.

$(s_{u_3+1}^{(d)} = 1)$. If it ends or goes down, let $v_3 \stackrel{\text{def}}{=} u_3$. Otherwise, the walk goes straight up and then ends at $w_N = B$; let $v_3 \stackrel{\text{def}}{=} N$. It is easy to verify that $w' \stackrel{\text{def}}{=} T_{t_1,v_3}^{\text{inv}}(w)$ has a higher top than w; we show that it is self-avoiding. Let w_{v_1} be the vertex of w that lies in the d-hyperplane containing w_{v_3} directly above w_{u_1}. The segment $w'_0,\ldots, w'_{t_1-1}$ is the original segment $w_0,\ldots, w_{t_1-1}$; no vertex in it coincides with an earlier one. The segment $w'_{t_1},\ldots, w'_{t_1+v_3-v_1}$ is a rigid transformation of $w_{v_1},\ldots, w_{v_3}$ and therefore it does not intersect itself. It lies on the d-hyperplane $\text{Top}(w)$ or above it where there are no other vertices of w' [we are in case 2, which assumes that $w \cap \text{Top}(w)$ is a single SAW]. The straight segment $w'_{t_1+v_3-v_1+1},\ldots, w'_{t_1+v_3-t_2-1}$ (originating from $w_{t_2+1},\ldots, w_{v_1-1}$) lies directly below $w'_{t_1+v_3-v_1}$ and is strictly higher than w_{v_3}. The only other vertices of w' in that region are directly below w'_{t_1} and therefore do not coincide. The segment $w'_{t_1+v_3-t_2},\ldots, w'_{v_3}$ originating from $w_{t_1},\ldots, w_{t_2}$ lies in the d-hyperplane $\{x\colon x^{(d)} = w_{v_3}^{(d)}\}$. The only other vertex of w' in this hyperplane belongs to the vertical segment $w'_0,\ldots, w'_{t_1}$. Denote the vertex, if it exists, by $w'_s = w_s$. If w'_s does not intersect the segment, then w' is a SAW, as $w'_{v_3+1},\ldots, w'_N$ is the original $w_{v_3+1},\ldots, w_N$, lying lower than $\text{Sec}(w)$. If w'_s intersects the segment, then, as in case 2b(ii), we first let $w^* \stackrel{\text{def}}{=} T_{t_1,t_2}^{\text{inv}}(w)$ and then $w' \stackrel{\text{def}}{=} T_{t_1,v_3}^{\text{inv}}(w^*)$. To show that w' is self-avoiding, we need only show that w'_s does not intersect the segment $w'_{t_1+v_3-t_2},\ldots, w'_{v_3}$ at w'_{v_3} (originating from w_{t_1}) or at $w'_{t_1+v_3-t_2}$ (originating from w_{t_2}). w'_s does not coincide with w'_{v_3} as w'_{v_3} is indirectly above w_{u_3}, whereas w'_s is directly above w_{u_2}. w'_s does not coincide with $w'_{t_1+v_3-t_2}$, as $w'_{t_1+v_3-t_2}$ is directly below $w'_{t_1+v_3-u_1}$ (originating from w_{u_1}), whereas w'_s is directly below $w'_{t_1+v_3-u_3}$ (originating from w_{u_3}). ∎

It remains to prove statement (D2).

Lemma 3. Any canonical SAW in $S^N(A, B)$ can be transformed into any other canonical SAW in $S^N(A, B)$ using at most $d^2/2$ transformations in $\mathscr{F}_d^N$.

Proof. Let w and w' be canonical SAWs in $S^N(A, B)$. We distinguish between canonical SAWs that are minimal length and those that are not.

1. $N = |A - B|$. Then both w and w' consist of the same number of maximal straight segments. The segments corresponding to each coordinate are of the same length. The only difference is the order of the segments. The order of two consecutive segments can be reversed using one inversion transformation. Hence any order can be achieved using at most

$$(d-1) + (d-2) + \cdots + 1 = d(d-1)/2$$

inversion transformations.

2. $N > |A - B|$. If w and w' are contained in the same two-dimensional plane, then the three transformations illustrated in Fig. 6 and one additional inversion suffice to transform w into w'. [A canonical SAW is a special case of a SAW in $R^N(A, B)$.] Otherwise, let $\delta = \mathrm{coor}(s_1)$ and $\varepsilon = \mathrm{coor}(s'_1)$.

(a) If $\delta = \varepsilon$, then $s_1 = s'_1$ or $s_1 = -s'_1$. The SAWs w and w' [or, in the latter case, w and $T^{\mathrm{inv}}_{0,N}(w)$] decompose into the same maximal straight segments; the δ-coordinate segments are first and last, while the other segments appear in possibly different order. Using at most $(d-2)(d-3)/2$ inversion transformations as above, the order of the segments can be modified.

(b) If $\delta \neq \varepsilon$, assume $\delta < \varepsilon$. Any $\alpha \in \{1,..., d\} \backslash \{\delta, \varepsilon\}$ is the coordinate of a maximal straight segment in w if and only if it is the coordinate of a maximal straight segment of w' and these two segments are of equal lengths. Using at most $d-2$ inversion transformations, "move" the ε-coordinate segment of w, if it exists, so that it is first. Call the new SAW $w^{[1]}$. Let $w^{[2]} \stackrel{\mathrm{def}}{=} T^{\mathrm{int},m}_{k,N;\delta,\varepsilon}(w^{[1]})$, where $k = |A^{(\delta)} - B^{(\delta)}| + |A^{(\varepsilon)} - B^{(\varepsilon)}|$ and where $m = 1$ if $s_k^{[1](\varepsilon)} = s_{k+1}^{[1](\delta)}$ and $m = -1$ otherwise (this ensures that $w^{[2]}$ is self-avoiding). Now, $w^{[2]}$ is canonical and $\mathrm{coor}(s_1^{[2]}) = \mathrm{coor}(s'_1) = \varepsilon$, so we are in case (a) and at most $(d-2)(d-3)/2 + 1$ inversion transformations are needed to transform $w^{[2]}$ into w'. The full details are left to the reader.

The total number of transformations used is at most $(d-2) + 1 + (d-2)(d-3)/2 + 1 \leqslant d^2/2$. ∎

Corollary 2. The CPS algorithm is ergodic in $\mathscr{Z}^d$ for $d \geqslant 3$.

Proof. Reflection and interchange transformations were used in the above proofs only in Lemma 3 and case 2a in the proof of the claim to show that when $w \cap \mathrm{Top}(w)$ is a canonical walk that is not minimal, w can be transformed to a walk with a higher top.

It is easy to see that in both cases, BFACF moves can be substituted for reflection and interchange transformations. ▌

5. A SINGLE TRANSFORMATION FOR TWO-DIMENSIONAL SAWS WITH FIXED LENGTH AND FREE ENDPOINTS

In this section we return to self-avoiding walks in two dimensions, but we focus on the free-endpoint ensemble: $S^N(A)$, the set of all N-step SAWs in $\mathscr{L}^2$ with $w_0 = A$. The pivot algorithm[6,7,16] is a highly efficient Monte Carlo algorithm on $S^N(A)$, in which the set of transformations $\mathscr{F}_2^N$ consists of: (i) reflections through vertical and horizontal lines, (ii) reflections through lines of slope ± 1, (iii) 180° rotations, (iv) 90° rotations.

In all cases, the transformation is applied to a segment of the form $w_k, w_{k+1},..., w_N$, with $0 \leqslant k \leqslant N$. It was shown in ref. 7 that this algorithm is ergodic, even if $\mathscr{F}_2^N$ includes only classes (i) and (ii), or (i) and (iv), or (ii) and (iii), or (ii) and (iv) of transformations; however, the algorithm is not ergodic if $\mathscr{F}_2^N$ only includes the class (i), or (iii), or (iv). In this section, we prove that class (ii) alone suffices for ergodicity. In fact, we show that any SAW in $S^N(A)$ can be transformed into any other SAW in $S^N(A)$ by at most $2N$ transformations fromf class (ii); the best bound proven before now is $4N$, even when $\mathscr{F}_2^N$ contains all four classes. Our result is also useful if one wants to write the simplest possible computer program for the pivot algorithm.

For an SAW $w = w_0,..., w_N$, an integer $k \in \{0,..., N\}$, and $m \in \{-1, +1\}$, define the reflection transformation $T_k^m(w) \stackrel{\text{def}}{=} (w'_0,..., w'_N$ by

$$w'_i \stackrel{\text{def}}{=} \begin{cases} w_i & \text{for } 0 \leqslant i \leqslant k \\ (w_k^{(1)} + m(w_i^{(2)} - w_k^{(2)}), w_k^{(2)} + m(w_i^{(1)} - w_k^{(1)})) & \text{for } k < i \leqslant N \end{cases}$$

As shown in Fig. 13, T_k^m reflects the tail of w, from w_k to w_N, in the line of slope m going through w_k. Therefore, $T_k^m(w)$ is a walk of length N starting at w_0. However, it is not necessarily self-avoiding. Note also that T_k^m is its own inverse: $T_k^m(T_k^m(w)) = w$.

A *diagonal support line* (DSL) of a walk w at w_k is a line of slope 1 or -1 containing w_k such that all vertices of w lie on one of its sides. As illustrated in Fig. 14, if w has a DSL of slope m at w_k, then $T_k^m(w)$ is self-avoiding, hence in $S^N(w_0)$. To see this analytically, note that w has a

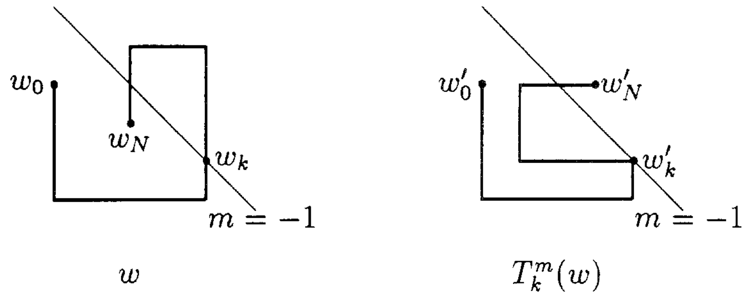

Fig. 13. Reflection at w_k with $m = -1$.

diagonal support line at w_k if $(w_i^{(1)} - mw_i^{(2)}) - (w_k^{(1)} - mw_k^{(2)})$ has the same sign for all $0 \leqslant i \leqslant N$. Hence, for $k < i \leqslant N$,

$$w_i'^{(1)} - mw_i'^{(2)} = w_k^{(1)} + m(w_i^{(2)} - w_k^{(2)}) - m(w_k^{(2)} + m(w_i^{(1)} - w_k^{(1)}))$$
$$= 2(w_k^{(1)} - mw_k^{(2)}) - (w_i^{(1)} - mw_i^{(2)})$$

Hence,

$$(w_i'^{(1)} - mw_i'^{(2)}) - (w_k^{(1)} - mw_k^{(2)}) = -[(w_i^{(1)} - mw_i^{(2)}) - (w_k^{(1)} - mw_k^{(2)})]$$

Therefore, the tail of w is mapped to a side of the DSL that previously contained no vertices of w.

For $1 \leqslant i \leqslant N$, the *ith step* of an N-step SAW $w = w_0, \ldots, w_N$ is the increment $s_i \stackrel{\text{def}}{=} w_i - w_{i-1}$. A vertex w_i of a walk w is a *turn vertex* of w if $0 < i < N$ and $s_i \neq s_{i+1}$. A SAW is *straight* if it has no turn vertices.

The next lemma shows that if a walk $w \in S^N(A)$ is not straight, then there is a turn vertex w_k and $m \in \{-1, +1\}$ such that $T_k^m(w)$ has one turn

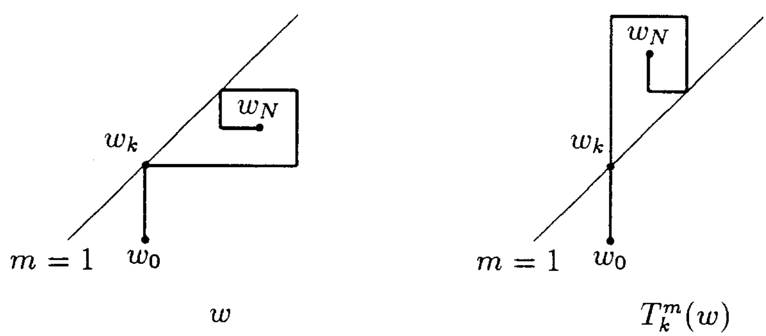

Fig. 14. Reflection in a diagonal support line at w_k.

vertex less than w. Therefore, any SAW can be transformed into a straight SAW by a sequence of at most $N-1$ reflections. Since two reflections suffice to transform any straight SAW into any other, and since any reflection transformation is its own inverse, it follows that any SAW in $S^N(A)$ can be transformed into any other using at most $2N$ successive transformation.

Lemma 4. For any SAW w that is not straight, there are $0 \leqslant k \leqslant N$ and $m \in \{-1, +1\}$ such that $T_k^m(w)$ has one turn less than w.

Proof. The following observations can be easily verified.

1. If w_k is a turn vertex of $w \in S^N(A)$ and $T_k^m(w)$ is self-avoiding, then $T_k^m(w)$ has one less turn than w.
2. All SAW consisting of more than a single vertex have four distinct DSLs.
3. All intersections of a SAW with a DSL are either end vertices or turn vertices.

If one of the DSLs whose slope is m intersects w at a turn vertex w_k, then, from Observation 1, the reflection $T_k^m(w)$ is a SAW with one turn vertex less than w. Otherwise, two DSLs, one with slope 1 and the other with slope -1, intersect at w_0 and the other two DSLs intersect at w_N. Necessarily, then, the first and last step of w are identical: $s_1 = s_N$ (see Fig. 15a). Consider the two DSLs that intersect at w_N. "Slide" them simultaneously in the $-s_N$ direction until at least one of them contains two or more vertices of w (see Fig. 15b). Let m be the slope of that DSL,[5] let w_l be the last vertex (maximal l) contained in the DSL, and let w_k be any other vertex contained in the DSL. The segment of w from w_l to w_N is straight. We show that w_k must be a turn vertex.

[5] If both DSLs contain two or more vertices of w, as in Fig. 15b, pick one of them arbitrarily.

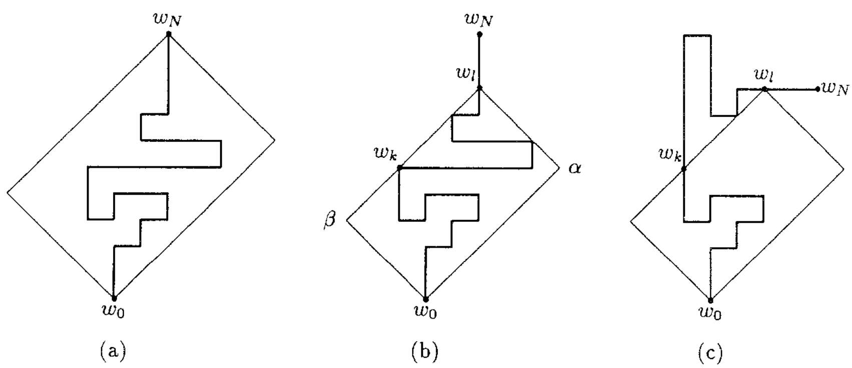

Fig. 15. Reducing the number of turns when DSLs hit endpoints.

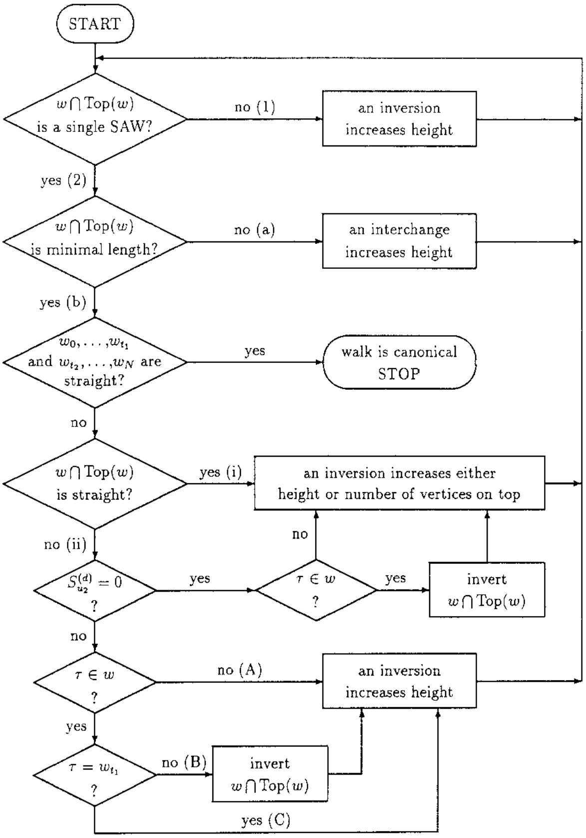

Fig. 16. Flow chart for the proof of statement (D1), Section 4.

By construction, if the DSL is perturbed in the direction of s_N, it intersects w only once, in the vicinity of w_l. Hence w_k is either a turn or an end vertex. If it is an end vertex, it must be w_0; but this is impossible because $s_1 = s_N$; hence any line perturbed from w_0 in the direction of $s_1 = s_N$ intersects the segment from w_0 to w_1. Therefore w_k is a turn vertex.

To see that $T_k^m(w)$ is self-avoiding, note that the two DSLs containing w_0 and the two modified DSLs form a (tilted) rectangle (w_0, α, w_l, β in Fig. 15b). The segment of w between w_0 and w_l is contained in the rectangle and the segment between w_l and w_N is straight. Hence, after the reflection, the segment between w'_k and w'_N will all be outside the rectangle and will not intersect the segment from w'_0 to w'_k.

Hence w_k is a turn vertex and $T_k^m(w)$ is self-avoiding. From Observation 1, $T_k^m(w)$ has one turn vertex less than w. ∎

ACKNOWLEDGMENTS

We thank Sergio Caracciolo and Alan Sokal for discussions of their work. One of the authors (N.M.) was supported by NSERC of Canada.

REFERENCES

1. C. Domb, *Adv. Chem. Phys.* **15**:229 (1969).
2. S. G. Whittington, *Adv. Chem. Phys.* **51**:1 (1982).
3. P. G. de Gennes, *Phys. Lett. A* **38**:339 (1972).
4. K. Kremer and K. Binder, *Comp. Phys. Rep.* **7**:259 (1988).
5. N. Madras and A. D. Sokal, *J. Stat. Phys.* **47**:573 (1987).
6. M. Lal, *Mol. Phys.* **17**:57 (1969).
7. N. Madras and A. D. Sokal, *J. Stat. Phys.* **50**:109 (1988).
8. B. Berg and D. Foerster, *Phys. Lett.* **106B**:323 (1981).
9. C. Aragão de Carvalho, S. Caracciolo, and J. Fröhlich, *Nucl. Phys. B* **215**[FS7]:209 (1983).
10. C. Aragão de Carvalho and S. Caracciolo, *J. Phys.* (Paris) **44**:323 (1983).
11. A. D. Sokal and L. E. Thomas, *J. Stat. Phys.* **51**:907 (1988).
12. L. E. Dubins, A. Orlitsky, J. A. Reeds, and L. A. Shepp, *IEEE Trans. Inform. Theory* **34**:1509 (1988).
13. A. Berretti and A. D. Sokal, *J. Stat. Phys.* **40**:483 (1985).
14. S. Caracciolo, A. Pelissetto, and A. D. Sokal, preprint.
15. S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes* (Academic Press, New York, 1975).
16. B. MacDonald, N. Jan, D. L. Hunter, and M. O. Steinitz, *J. Phys. A: Math. Gen.* **18**:2627 (1985).