Electronic Journal of Probability

Vol. 13 (2008), Paper no. 15, pages 396–466.

Journal URL
http://www.math.washington.edu/~ejpecp/

Information Recovery From A Randomly Mixed Up Message-Text

Jüri Lember*
University Of Tartu
J. Liivi 2-513, Tartu, Estonia
e-mail: jyril@ut.ee

Heinrich Matzinger
University Of Bielefeld
Postfach 10 01 31, 33501 Bielefeld, Germany
e-mail: matzing@mathematik-uni-bielefeld.de

Abstract

This paper is concerned with finding a fingerprint of a sequence. As input data one uses the sequence which has been randomly mixed up by observing it along a random walk path. A sequence containing order $\exp(n)$ bits receives a fingerprint with roughly n bits information. The fingerprint is characteristic for the original sequence. With high probability the fingerprint depends only on the initial sequence, but not on the random walk path.

Key words: Scenery reconstruction, random walk in random environment.

AMS 2000 Subject Classification: Primary 60G50, 60K37.

Submitted to EJP on May 6, 2004, final version accepted May 22, 2007.

*Supported by Estoninan Science foundation Grant 5694

1 Introduction and Result

1.1 The information recovery problem

Let $\xi : \mathbb{Z} \to \{0, 1\}$ designate a double-infinite message-text with 2 letters. Such a coloring of the integers is also called a (2-color) scenery. Let $S = \{S(t)\}_{t\in\mathbb{N}}$ be a recurrent random walk on $\mathbb{Z}$ starting at the origin. In this paper we allow the random walk S to jump, i.e. $P(|S(t+1) - S(t)| > 1) > 0$. We use S to mix up the message-text ξ. For this we assume that ξ is observed along the path of S: At each point in time t, one observes $\chi(t) := \xi(S(t))$. Thus, χ designates the mixed up message-text, which is also the scenery ξ seen along the path of S. The *information recovery problem* can be described as follows: observing only one path realization of the process χ, can one retrieve a certain amount of information contained in ξ? A special case of the information recovery problem is when one tries to reconstruct the whole ξ. This problem is called *the scenery reconstruction problem* . In many cases being able to reconstruct a finite quantity of the information contained in ξ, already implies that one can reconstruct all of ξ. This paper is concerned with the information recovery problem in the context of a 2-color scenery seen along a random walk with jumps. The methods which exist so far seem useless for this case: Matzinger's reconstruction methods [Mat99a; Mat05] do not work when the random walk may jump. Furthermore, it seems impossible to recycle the method of Matzinger, Merkl and Löwe [LMM04] for the 2-color case with jumps. The reason is that their method, requires more than 2-colors. Hence, the fundamentally new approach is needed. That is presented in this paper.

1.2 Main assumptions

Let us explain the assumptions which remain valid throughout this paper:

- $\xi = \{\xi(z)\}_{z\in\mathbb{Z}}$ is a collection of i.i.d. Bernoulli variables with parameter 1/2. The path realization $\xi : z \mapsto \xi(z)$ is the scenery from which we want to recover some information. Often the realization of the process $\{\xi(z)\}_{z\in\mathbb{Z}}$ is also denoted by ψ.

- $S = \{S(t)\}_{t\in\mathbb{N}}$ is a symmetric recurrent random walk starting at the origin, i.e. $P(S(0) = 0) = 1$. We assume that S has bounded jump length $L < \infty$, where

  $$L := \max\{z|P(S(1) - S(0) = z) > 0\}.$$

  We also assume that S has positive probability to visit any point in $\mathbb{Z}$, i.e. for any $z \in \mathbb{Z}$ there exists $t \in \mathbb{N}$, such that $P(S(t) = z) > 0$.

- ξ and S are independent.

- $m = m(n)$ designates a natural number depending on n, so that

  $$\frac{1}{4}\exp\left(\frac{\alpha n}{\ln n}\right) \leq m(n) < \exp(2n)$$

  where $\alpha := \ln 1.5$

- For all $t \in \mathbb{N}$, let $\chi(t) := \xi(S(t))$. Let

$$\chi := (\chi(0), \chi(1), \ldots)$$

designate the observations made by the random walk S of the random scenery ξ. Hence χ corresponds to the scenery ξ seen along the path of the random walk S.

We need also a few notations:

- For every $k \in \mathbb{N}$, let $\xi_0^k := (\xi(0), \xi(1), \ldots, \xi(k))$ and let $\xi_0^{-k} := (\xi(0), \xi(-1), \ldots, \xi(-k))$.
- Let $f : D \to I$ be a map. For a subset $E \subset D$ we shall write $f|E$ for the restriction of f to E.

  Thus, when $[a, b] \in \mathbb{Z}$ is an integer interval and ξ is a scenery, then $\xi|[a, b]$ stands for the vector $(\xi(a), \ldots, \xi(b))$. We also write ξ_a^b for $\xi|[a, b]$ and ψ_a^b for $\psi|[a, b]$. The notation

$$\chi_0^{m^2} := (\chi(0), \chi(1), \chi(2), \ldots, \chi(m^2))$$

  is often used.
- Let $a = (a_1, \ldots, a_N)$, $b = (b_1, \ldots, b_{N+1})$ be two vectors with length N and $N + 1$, respectively. We write $a \sqsubseteq b$, if

$$a \in \{(b_1, \ldots, b_N), (b_2, \ldots b_{N+1})\}.$$

  Thus, $a \sqsubseteq b$ holds if a can be obtained from b by "removing the first or the last element".

1.3 Main result

The 2-color scenery reconstruction problem for a random walk with jumps is solved in two phases:

1. Given a finite portion of the observations χ only, one proves that it is possible to reconstruct a certain amount of information contained in the underlying scenery ξ.
2. If one can reconstruct a certain amount of information, then the whole scenery ξ can a.s. be reconstructed. This is proven in the second phase.

This paper solves the first of the two problems above. The second problem is essentially solved in the follow-up paper [LM02a]. In order to understand the meaning of the present paper, imagine that we want to transmit the word ξ_0^m. During transmission the lector head gets crazy and starts moving around on ξ following the path of a random walk. At time m^2, the lector head has reached the point m. Can we now, given only the mixed up information $\chi_0^{m^2}$, retrieve any information about the underlying code ξ_0^m? The main result of this paper theorem 1.1, shows that with high probability a certain amount of the information contained in ξ_0^m can be retrieved from the mixed up information $\chi_0^{m^2}$. This is the fingerprint of ξ_0^m, referred to in the abstract. Here is the main result of this paper.

Theorem 1.1. *For every $n > 0$ big enough, there exist two maps*

$$g : \{0,1\}^{m+1} \to \{0,1\}^{n^2+1}$$
$$\hat{g} : \{0,1\}^{m^2+1} \to \{0,1\}^{n^2}$$

and an event

$$E^n_{\text{cell\_OK}} \in \sigma(\xi(z)|z \in [-cm, cm])$$

with $c > 0$ not depending on n such that all the following holds:

1) *$P(E^n_{\text{cell\_OK}}) \to 1$ when $n \to \infty$.*

2) *For any scenery $\xi \in E^n_{\text{cell\_OK}}$,*

$$P\Big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big|S(m^2) = m, \xi\Big) > 3/4.$$

3) *$g(\xi_0^m)$ is a random vector with $(n^2 + 1)$ components which are i.i.d. Bernoulli variables with parameter $1/2$.*

The mapping g can be interpreted as a coding that compresses the information contained in ξ_0^m; the mapping $\hat{g}$ can be interpreted as a decoder that reads the information $g(\xi_0^m)$ from the mixed-up observations $\chi_0^{m^2+1}$. The vector $g(\xi_0^m)$ is the desired fingerprint of ξ_0^m. We call it the *g-information*. The function $\hat{g}$ will be referred to as the *g-information reconstruction algorithm*. Let us explain the content of the above theorem more in detail. The event $\{\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\}$ is the event that $\hat{g}$ reconstructs the information $g(\xi_0^m)$ correctly (up to the first or last bit), based on the observations $\chi_0^{m^2}$. The probability that $\hat{g}$ reconstructs $g(\xi_0^m)$ correctly is large given the event $\{S(m^2) = m\}$ holds. The event $\{S(m^2) = m\}$ is needed to make sure the random walk S visits the entire ξ_0^m up to time m^2. Obviously, if S does not visit ξ_0^m, we can not reconstruct $g(\xi_0^m)$.
The reconstruction of the g-information works with high probability, but conditional on the event that the scenery is nicely behaved. The scenery ξ behaves nicely, if $\xi \in E^n_{\text{cell\_OK}}$. In a sense, $E^n_{\text{cell\_OK}}$ contains " typical" (pieces of) sceneries. These are sceneries for which the g-information reconstruction algorithm works with high probability.
Condition **3**) ensures that the content of the reconstructed information is large enough. Indeed, if the piece of observations $\chi_0^{m^2}$ were generated far from ξ_0^m, i.e. the random walk S would start far from 0, then $g(\xi_0^m)$ were independent of $\chi_0^{m^2}$, and $P\big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big)$ would be about 2^{-n^2}. On the other hand, if S starts from 0 and n is big enough, then from **1)** and **2)**, it follows that

$$P\big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big|S(m^2) = m\big) > 3/4 \tag{1.1}$$

and

$$P\big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big) \geq \frac{3}{4} P(S(m^2) = m).$$

Since, by local central limit theorem, $P(S(m^2) = m)$ is of order $\frac{1}{m} \geq e^{-2n}$, we get that $P\big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big)$ is at least $O(e^{-2n})$. Although, for big n, the difference between 2^{-n^2} and e^{-2n} is negligible, it can be still used to make the scenery reconstruction possible.

1.4 History and related problems

A coloring of the integers $\xi : \mathbb{Z} \to \{0, 1, \dots, C-1\}$ is called a C-color scenery. In a sense, the scenery reconstruction started with the so-called *scenery distinguishing problem* that can be described as follows: Let ψ^a and ψ^b be two non-equivalent sceneries which are known to us. Assume that we are only given one realization of the observations $\chi := \psi \circ S$, where $\psi \in \{\psi^a, \psi^b\}$. Can we a.s. find out whether ψ is equal to ψ^a or ψ^b? If yes, we say that the sceneries ψ^a and ψ^b are distinguishable. Kesten and Benjamini [BK96] considered the case where the sceneries ψ^a and ψ^b are drawn randomly. They take ψ^a to be an i.i.d. scenery which is independent of ψ^b. In this setting, they prove that almost every couple of sceneries is distinguishable even in the two dimensional case and with only 2 colors. Before that Howard [How97] had shown that any two periodical non-equivalent sceneries are distinguishable. The problem of distinguishing two sceneries which differ only in one element is called the *single defect detection problem*. In [How96], Howard showed that single defects can always be detected in periodic sceneries observed along a simple random walk. Kesten [Kes96] showed that one can a.s. detect single defects in the case of 5-color i.i.d. sceneries.
The question of Kesten whether one can detect a single defect in 2-color sceneries lead Matzinger to investigate the *scenery reconstruction problem*: Given only one path realization of $\{\chi(t)\}_{t\in\mathbb{N}}$, can we a.s. reconstruct ξ? In other words, does one path realization of χ a.s. uniquely determine ξ? In general, it does not: in many cases it is not possible to distinguish a scenery from a shifted one. Furthermore, Lindenstrauss proved [Lin99] the existence of sceneries which can not be reconstructed. However, one can reconstruct "typical" sceneries: Matzinger takes ξ randomly, independent of S and shows that one can reconstruct a.s. the scenery up to shift and reflection. In [Mat05] and [Mat99a], he proves this for 2-color sceneries observed along the path of simple random walk or a simple random walk with holding. In [Mat99b], he reconstructs 3-color i.i.d. sceneries observed along a simple random walk path. The two cases require very different methods (for an overview of different techniques, see [ML06]). Later Kesten [Kes98] asked, whether one can also reconstruct two dimensional random sceneries. Loewe and Matzinger [LM02b] give a positive answer provided the scenery contains many colors. Another question was formulated first by Den Hollander: To which extent can sceneries be reconstructed when they are not i.i.d. in distribution. Loewe and Matzinger [LM03] characterize those distributions for which Matzinger's 3-color reconstruction works. Yet another problem comes from Benjamini: Is it possible to reconstruct a finite piece of a scenery close to the origin in polynomial time? We take for this polynomially many observations in the length of the piece we try to reconstruct. Matzinger and Rolles [MR03a; MR06] provide a positive answer.
The scenery reconstruction problem varies greatly in difficulty depending on the number of colors and the properties of the random walk. In general, when there are less colors and the random walk is allowed to jump, the problem gets more difficult. Kesten [Kes98] noticed, that Matzinger's reconstruction methods [Mat99a] and [Mat05] do not work when the random walk is allowed to jump. Matzinger, Merkl and Loewe [LMM04] showed that it is possible to reconstruct a.s. a scenery seen along the path of a random walk with jumps, provided the scenery contains enough colors. However, with more colors the system is completely differently behaved. This implies that the method of Matzinger, Merkl and Loewe is not useful for the 2-color case with jumps. The present paper is the first step towards reconstructing the 2-color scenery.
Let us mention some more recent developments and related works. A generalization of the

scenery reconstruction problem is the *scenery reconstruction problem for error-corrupted observations*. In that problem, there exists an error process $\nu_t, t \in \mathbb{N}$ and error-corrupted observations $\hat{\chi}$ so that $\hat{\chi}_t$ equals to the usual observation χ_t if and only if $\nu_t = 0$. The error process is i.i.d. Bernoulli and independent of everything else. The problem now is: is possible to reconstruct the scenery based on one realization of the process $\hat{\chi}$ only? Matzinger and Rolles [MR03b] showed that almost every random scenery seen with random errors can be reconstructed a.s. when it contains a lot of colors. However, their method cannot be used for the case of error corrupted 2-color sceneries. The error-corrupted observations were also studied by Matzinger and Hart in [HM06]. A closely related problem is the so-called *Harris-Keane coin tossing problem* introduced and studied by Harris and Keane in [HK97] and further investigated by Levin, Pemantle and Peres in [LPP01].
In the scenery reconstruction results above, the reconstructable scenery is a "typical" realization of a random (i.i.d. Bernoulli) scenery. A periodic scenery is not such kind of "typical" realization, so the abovementioned results do not apply for the case of periodic scenery. Howard [How97] proved that all periodic sceneries observed along a simple random walk path, can be reconstructed. This lead Kesten to ask what happens when the random walk is not simple. In [LM06], Matzinger and Lember give sufficient conditions for a periodic scenery being reconstructable when observed along a random walk with jumps.
A problem closely related to the reconstruction of periodic sceneries, is the reconstruction of sceneries with a finite number of ones. This problem was solved by Levin and Peres in [LP04], where they prove that every scenery which has only finite many one's can a.s. be reconstructed up to shift or reflection when seen along the path of a symmetric random walk. The used a more general framework of stochastic scenery. A *stochastic scenery* is a map $\xi : \mathbb{Z} \to I$, where I denotes a set of distributions. At time t, one observes the random variable $\chi(t)$, drawn according to the distribution $\xi(S_t) \in I$. Given S and ξ, the observations $\chi(t)$ for different t's are independent of each other. The observations are generated as follows : if at time t the random walk is at z, then a random variable with distribution $\eta(z)$ is observed. Hence, at time t, we observe $\chi(t)$, where $\mathcal{L}(\chi(t)|S(t) = z) = \eta(z)$.
Recently, Matzinger and Popov have been studied continuous sceneries [MP07]. They define a *continuous scenery* as a location of countably many bells placed on $\mathbb{R}$. In continuous case, instead of random walk, a Brownian motion is considered. Whenever the Brownian motion hits a bell, it rings. So, unlike the discrete scenery reconstruction, there are no colors: all the bells ring in the same way. The observations consists of time lengths between successive rings.
For a well-written overview of the scenery distinguishing and scenery reconstruction areas, we recommend Kesten's review paper [Kes98]. An overview of different techniques as well as the recent developments in scenery reconstruction can be found in [ML06].
Scenery reconstruction belongs to the field which investigates the properties of a color record obtained by observing a random media along the path of a stochastic process. The TT^{-1}-problem as studied by Kalikow [Kal82] is one motivation. The ergodic properties of observations have been investigated by Keane and den Hollander [KdH86], den Hollander [dH88], den Hollander and Steiff [dHS97] and Heicklen, Hoffman and Rudolph [HHR00]. An overview of mentioned results as well as many others can be found in [dHS06].

1.5 Organization of the paper

In order to explain the main ideas behind the g-information reconstruction algorithm, we first consider a simplified example in Subsection 1.6. In this example, ξ is a 3-color i.i.d. scenery instead of a 2-color scenery. The 2's are pretty rare in the scenery ξ: $P(\xi(z)=2)$ is of negative exponential order in n. The one's and zero's have equal probability: $P(\xi(z)=0)=P(\xi(z)=1)$. The (random) locations $\bar{z}_i$ of the 2's in ξ are called signal carriers. For each signal carrier $\bar{z}_i$, we define the *frequency of ones* at $\bar{z}_i$. The frequency of one's at $\bar{z}_i$ is a weighted average of ξ in the neighborhood of $\bar{z}_i$. The g-information $g(\xi_0^m)$ if a function of the different frequencies of ones of the signal carriers which are located in the interval $[0,m]$. The vector of frequencies works as a fingerprint for ξ_0^m. The reading of this fingerprint works as follows: Typically, the signal carriers are apart from each other by a distance of order e^n. Suppose that S visits a signal carrier. Before moving to the next signal carrier, it returns to the same signal carrier many times with high probability. By doing this, S generates many 2's in the observations at short distance from each other. This implies: when in the observations we see a cluster of 2's, there is a good reason to believe that they all correspond to the same 2 in the underlying scenery. In this manner we can determine many return times of S to the same signal carrier. This enables us to make inference about ξ in the neighborhood of that signal carrier. In particular, we can precisely estimate the frequencies of ones of the different signal carriers visited. This allows us to estimate $g(\xi_0^n)$. The estimator $\hat{g}$ is the desired decoder. The details are explained in Subsection 1.6. However, it is important to note, that between this simplified example and our general case there is only one difference: The signal carriers. In the general case we can no longer rely on the 2's and the signal carriers need to be constructed in a different manner. Everything else – from the definition of g and $\hat{g}$ up to the proof that the g-information reconstruction algorithm works with high probability – is exactly the same. (Note that the solution to our information recovery problem in the simplified 3-color case requires only five pages!)

For the general case with a 2-color scenery and a jumping random walk, the main difficulty consists in the elaboration of the signal carriers. In Section 2, we define many concepts which are subsequently used for the definition of the signal carriers. Also there, some technical results connected to the signal carriers are proved. The signal carriers are defined in Section 3.

The main goal of the paper is to prove that the g-reconstruction algorithm works with high probability (i.e. that the estimator $\hat{g}$ is precise). For this, we define two sets of events: the random walk dependent events and the scenery dependent event. All these events describe typical behavior of S or ξ. In Section 3, we define the scenery dependent events and prove that they have high probability. In Section 4 the same is done for the events that depend on S.

In section 5, we prove that if all these events hold, then the g-information reconstruction algorithm works, i.e. the event

$$E^n_{\text{g\_works}} := \{\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\}$$

holds. The results of Section 3 and Section 4 then guarantee that the g-information reconstruction algorithm works with high probability. This finishes the proof of Theorem 1.1.

1.6 3-color example

In this subsection, we solve the scenery reconstruction problem in a simplified 3-color case. We do not change the assumptions on S.

1.6.1 Setup

Recall that ξ_0^m and $\chi_0^{m^2}$ denote the piece of scenery $\xi|[0,m]$ and the first m observations $\chi|[0,m]$, respectively. We aim to construct two functions

$$g : \{0,1\}^{m+1} \to \{0,1\}^{n^2+1} \quad \text{and} \quad \hat{g} : \{0,1\}^{m^2+1} \to \{0,1\}^{n^2}$$

and a ξ-dependent event $E^n_{\text{cell\_OK}}$ such that

1) $P(E^n_{\text{cell\_OK}}) \to 1$

2) For every $\xi \in E^n_{\text{cell\_OK}}$, it holds

$$P\left(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m) \middle| S(m^2) = m, \xi\right) > \frac{3}{4}$$

3) $g(\xi_0^m)$ is i.i.d. binary vector where the components are Bernoulli random variables with parameter $\frac{1}{2}$.

Then (1.1) holds, implying that with high probability we can reconstruct $g(\xi_0^m)$ from the observations, provided that random walk S goes in m^2 steps from 0 to m.

Since this is not yet the real case, during the present subsection we will not be very formal. For this subsection only, let us assume that the scenery ξ has three colors instead of two. Moreover, we assume that $\{\xi(z)\}$ satisfies all of the following three conditions:

a) $\{\xi(z) : z \in \mathbb{Z}\}$ are i.i.d. variables with state space $\{0,1,2\}$,

b) $\exp(n/\ln n) \leq 1/P\left(\xi(0) = 2\right) \leq \exp(n)$,

c) $P(\xi(0) = 0) = P(\xi(0) = 1)$.

We define $m = n^{2.5}(1/P(\xi(0) = 2))$. Because of **b)** this means

$$n^{2.5} \exp(n/\ln n) \leq m(n) \leq n^{2.5} \exp(n).$$

The so defined scenery distribution is very similar to our usual scenery except that sometimes (quite rarely) there appear also 2's in this scenery. We now introduce some necessary definitions.

Let $\bar{z}_i$ denote the i-th place in $[0,\infty)$ where we have a 2 in ξ. Thus

$$\bar{z}_1 := \min\{z \geq 0 | \xi(z) = 2\}, \quad \bar{z}_{i+1} := \min\{z > \bar{z}_i | \xi(z) = 2\}.$$

We make the convention that $\bar{z}_0$ is the last location before zero where we have a 2 in ξ. For a negative integer $i < 0$, $\bar{z}_i$ designates the $i+1$-th point before 0 where we have a 2 in ξ. The random variables $\bar{z}_i$-s are called *signal carriers*. For each signal carrier, $\bar{z}_i$, we define the *frequency of ones* at $\bar{z}_i$. By this we mean the (conditional on ξ) probability to see 1 exactly after $e^{n^{0.1}}$ observations having been at $\bar{z}_i$. We denote that conditional probability by $h(\bar{z}_i)$ and will also write $h(i)$ for it. Formally:

$$h(i) := h(\bar{z}_i) := P\left(\xi(S(e^{n^{0.1}}) + \bar{z}_i) = 1 \middle| \xi\right).$$

It is easy to see that the frequency of ones is equal to a weighted average of the scenery in a neighborhood of radius $Le^{n^{0.1}}$ of the point $\bar{z}_i$. That is $h(i)$ is equal to

$$h(i) := \sum_{\substack{z\in[-Le^{n^{0.1}},Le^{n^{0.1}}]\\ z\neq\bar{z}_i}} \xi(z)P\big(S(e^{n^{0.1}})+\bar{z}_i=z\big) \tag{1.2}$$

(Of course this formula to hold assumes that there are no other two's in

$$[\bar{z}_i - Le^{n^{0.1}}, \bar{z}_i + Le^{n^{0.1}}]$$

except the two at $\bar{z}_i$. This is very likely to hold, see event E_6^n below).
Let

$$g_i(\xi_0^m) := I_{[0,0.5)}(h(i)).$$

We now define some events that describe the typical behavior of ξ.

- Let E_6^n denote the event that in $[0,m]$ all the signal carriers are further apart than $\exp(n/(2\ln n))$ from each other as well as from the points 0 and m. By the definition of $P(\xi(i)=2)$, the event $P(E_6^n)\to 1$ as $n\to\infty$.
- Let E_1^n be the event that in $[0,m]$ there are more than n^2+1 signal carrier points. Because of the definition of m, $P(E_1^n)\to 1$ as $n\to\infty$.

When E_1^n and E_6^n both hold, we define $g(\xi_0^m)$ in the following way:

$$g\left(\xi_0^m\right) := \left(g_1\left(\xi_0^m\right), g_2\left(\xi_0^m\right), g_3\left(\xi_0^m\right), \ldots, g_{n^2+1}\left(\xi_0^m\right)\right).$$

Conditional on $E_1^n \cap E_6^n$ we get that $g\left(\xi^m\right)$ is an i.i.d. random vector with the components being Bernoulli variables with parameter 1/2. Here the parameter 1/2 follows simply by symmetry of our definition (to be precise, $P(g_i\left(\xi_i^m\right)=1)=1/2-P(h(i)=1/2)$, but we disregard this small error term in this example) and the independence follows from the fact that the scenery is i.i.d. and $g_i(\xi_0^m)$ depends only on the scenery in a radius $Le^{n^{0.1}}$ of the point $\bar{z}_i$ and, due to E_6^n, the points $\bar{z}_i$ are further apart than $\exp(n/2\ln n) > L\exp(n^{0.1})$.
Hence, with almost no effort we get that when E_1^n and E_6^n both hold, then condition **3)** is satisfied. To be complete, we have to define the function g such that **3)** holds also outside $E_1^n \cap E_6^n$. We actually are not interested in g outside $E_1^n \cap E_6^n$ - it would be enough that we reconstruct g on $E_1^n \cap E_6^n$. Therefore, extend g in any possible way, so that $g\left(\xi_0^m\right)$ depends only on ξ_0^m and its component are i.i.d.

1.6.2 $\hat{g}$-algorithm

We show, how to construct a map $\hat{g}:\{0,1\}^{n^2}\mapsto\{0,1\}^n$ and an event $E^n_{\text{cell\_OK}}\in\sigma(\xi)$ such that $P(E^n_{\text{cell\_OK}})$ is close to 1 and, for each scenery belonging to $E^n_{\text{cell\_OK}}$, the probability

$$P\Big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\Big| S(m^2)=m,\xi\Big) \tag{1.3}$$

is also high. Note, when the scenery ξ is fixed, then the probability (1.3) depends on S, only.

The construction of $\hat{g}$ consists of several steps. The first step is the estimation of the frequency of one's $h(i)$. Note: due to E_6^n we have that in the region of our interest we can assume that all the signal carriers are further apart form each other than $\exp(n/(2\ln n))$. In this case we have that all the 2's observed in a time interval of length $e^{n^{0.3}}$ must come from the same signal carrier. We will thus take time intervals T of length $e^{n^{0.3}}$ to estimate the frequency of one's.

Let $T=[t_1,t_2]$ be a (non-random) time interval such that $t_2-t_1=e^{n^{0.3}}$. Assume that during time T the random walk is close to the signal carrier $\bar{z}_i$. Then every time we see a 2 during T this gives us a stopping time which stops the random walk at $\bar{z}_i$. We can now use these stopping times to get a very precise estimate of $h(i)$. In order to obtain the independence (which makes proofs easier), we do not take all the 2's which we observe during T. Instead we take the 2's apart by at least $e^{n^{0.1}}$ from each other.

To be more formal, let us now give a few definitions.
Let $\nu_{t_1}(1)$ denote the first time $t>t_1$ that we observe a 2 in the observations χ after time t_1. Let $\nu_{t_1}(k+1)$ be the first time after time $\nu_{t_1}(k)+e^{n^{0.1}}$ that we observe a 2 in the observations χ. Thus $\nu_{t_1}(k+1)$ is equal to $\min\{t|\chi(t)=2, t\geq \nu_{t_1}(k)+e^{n^{0.1}}\}$. We say that T is such that we can significantly estimate the frequency of one's for T, if there are more than $e^{n^{0.2}}$ stopping times $\nu_{t_1}(k)$ during T. In other words, we say that we can significantly estimate the frequency of one's for T, if and only if $\nu_{t_1}(e^{n^{0.2}})\leq t_2-e^{n^{0.1}}$.
Let $\hat{X}_{t_1}(k)$ designate the Bernoulli variable which is equal to one if and only if

$$\chi(\nu_{t_1}(k)+e^{n^{0.1}})=1.$$

When $\nu_{t_1}(e^{n^{0.2}})\leq t_2-e^{n^{0.1}}$ we define $\hat{h}_T$ the estimated frequency of one's during T in the following obvious way:

$$\hat{h}_T := \frac{1}{e^{n^{0.2}}}\sum_{k=1}^{e^{n^{0.2}}}\hat{X}_{t_1}(k).$$

Suppose we can significantly estimate the frequency of one's for T. Assume $E_6^n\cap E_1^n$ hold. Then all the stopping times $\nu_{t_1}(e^{n^{0.2}})$ stop the random walk S at one signal carrier, say $\bar{z}_i$. Because of the strong Markov property of S we get then that, conditional on ξ, the variables $X_{t_1}(k)$ are i.i.d. with expectations h_i. Now, by Höffding inequality,

$$P(|\hat{h}_T-h(i)|>e^{-n^{0.2}/4})\leq \exp(-(2e^{n^{0.2}/2}))$$

So that, with high probability, $\hat{h}_T$ is a precise estimate for $h(i)$:

$$|\hat{h}_T-h(i)|\leq e^{-n^{0.2}/4}. \tag{1.4}$$

The obtained preciseness of $\hat{h}_T$ is of the great importance. Namely, it is of smaller order than the typical variation of $h(i)$. In other words, with high probability $|h(i)-h(j)|$ is of much bigger order than $\exp(-n^{0.2}/4)$, $i\neq j$. To see this, consider (1.2). Note that, for each z,

$$\mu_i(z):=P(S(e^{n^{0.1}})+\bar{z}_i=z)$$

is constant, and, conditional under the event that in the radius of $L\exp(n^{0.1})$ are no more 2's in the scenery than $\bar{z}_i$, we have that $\xi(\bar{z}_i+z)$ are i.i.d. Bernoulli variables with parameter $\frac{1}{2}$. Hence

$$Var[h(i)]\leq \sum_{z\in[-Le^{n^{0.1}},Le^{n^{0.1}}]}\frac{1}{4}\Big(\mu_i(z)\Big)^2.$$

Since our random walk is symmetric we get that

$$\sum_{z\in[-Le^{n^{0.1}},Le^{n^{0.1}}]}\frac{1}{4}\left(\mu_i(z)\right)^2$$

is equal to 1/4 times the probability that the random walk is back at the origin after $2e^{n^{0.1}}$ time. By the local central limit theorem that probability is of order $e^{-n^{0.1}/2}$. This is much bigger than the order of the precision of the estimation of the frequencies of one's, $e^{-n^{0.2}/4}$. Since $h(i)$ is approximately normal, it is possible to show that with high probability all frequencies $h(0),h(1),\dots,h(n^2+1)$ are more than $\exp(-n^{0.11})$ apart from $\frac{1}{2}$. By the similar argument holds: If $\{\bar{z}_i\}_{i\in I}$ is the set of signal carriers that S encounters during the time $[0,m^2]$, then for each pair $i,j\in I$, the frequencies of ones satisfy

$$|h(i)-h(j)|>\exp(-n^{0.11}).$$

Let E_3^n be the set on which both statements ($h(i)$'s are more than $\exp(-n^{0.11})$ apart from 1/2 and from each other) hold.

Define

$$E^n_{\text{cell\_OK}}:=E_1^n\cap E_3^n\cap E_6^n.$$

Since E_1^n, E_3^n and E_6^n all depend on ξ, only, so does $E^n_{\text{cell\_OK}}$. From now on we assume that $E^n_{\text{cell\_OK}}$ hold and we describe the $\hat{g}$-construction algorithm in this case:

Phase I Determine the intervals $T\subseteq[0,m^2]$ containing more than $e^{n^{0.2}}$ two's (in the observations.) Let T_j designate the j-th such interval. Recall that these are the intervals where we can significantly estimate the frequency of one's. Let K designate the total number of such time-intervals in $[0,m^2]$.

Let $\pi(j)$ designate the index of the signal carrier $\bar{z}_i$ the random walk visits during time T_j (due to E_6^n, the visited signal carriers are further apart than $Le^{n^{0.2}}$ from each other and, hence, there is only one signal carrier that can get visited during time T_j. Thus the definition of $\pi(j)$ is correct.)

Phase II Estimate the frequency of one's for each interval T_j, $j=1,\dots,K$. Based on the observations $\chi_0^{m^2}$ only, obtain the vector

$$(\hat{h}_{T_1},\dots,\hat{h}_{T_K})=\Big(\hat{h}(\pi(1)),\hat{h}(\pi(2)),\dots,\hat{h}(\pi(K))\Big).$$

Here $\hat{h}(i)$ denotes the estimate of $h(i)$, obtained by time interval T_j, with $\pi(j)=i$.

The further construction of the $\hat{g}$-reconstruction algorithm bases on an important property of the mapping

$$\pi:\{1,\dots,K\}\to\mathbb{Z}.$$

Namely, with high probability π is a skip free walk, i.e. $|\pi(j)-\pi(j+1)|\leq 1$. Clearly, after being near to the point $\hat{z}_i$, S moves to the neighborhood of $\hat{z}_{i+1}$ or $\hat{z}_{i-1}$ (recall that on $E^n_{\text{cell\_OK}}$, 2's are rather far from each other). Say, it goes to the neighborhood of $\hat{z}_{i+1}$. The important property of S is that, with high probability, before moving to the vicinity of the next 2 (that is

located in $\hat{z}_{i+2}$ or $\hat{z}_i$) it visits $\hat{z}_{i+1}$ sufficiently many times. This means that there exists a time interval $[t_1, t_2]$ of length $e^{n^{0.3}}$ such that $\nu_{t_1}(e^{n^{0.2}}) \leq t_2 - e^{n^{0.1}}$. For big n, this clearly holds holds, if, after visiting $\hat{z}_{i+1}$ once, before $e^{n^{0.3}} - e^{n^{0.1}}$ steps, S visits $\hat{z}_{i+1}$ again at least $e^{n^{0.21}}$ times. This can be proven to hold with high probability.
Hence, the random walk during time $[0, m^2]$ is unlikely to go from one signal carrier to another without signaling all those in-between. By signaling those in-between, we mean producing in the observations for each signal carrier $\bar{z}_i$ a time intervals of for which one can significantly estimate the frequency of one's $h(i)$. In other words, with high probability, the mapping π is a skip-free random walk. In particular, $\pi(1) \in \{0.1\}$, i.e. $\pi_* \leq 1$, where

$$\pi_* := \min\{\pi(j) : j = 1, \ldots, K\}, \quad \pi^* := \max\{\pi(j) : j = 1, \ldots, K\}.$$

If $S(m^2) = m$, then by the event E_1^n, it holds $\pi^* > n^2$.

Phase III Apply clustering to the vector $(\hat{h}_{T_1}, \hat{h}_{T_2}, \ldots, \hat{h}_{T_K})$, i.e. define

$$C_i := \{\hat{h}_{T_j} : |\hat{h}_{T_j} - \hat{h}_{T_i}| \leq \exp(-n^{0.12})\},\ \hat{f}_i := \frac{1}{|C_i|} \sum_{j \in C_i} \hat{h}_{T_j}, \quad i = 1, \ldots, K.$$

Formally there are K clusters. However, if E_3^n holds and for every T, the estimate $\hat{h}_T$ satisfies (1.4), then for each different i, j either $C_i = C_j$ or $C_i \cap C_j = \emptyset$. To see this note that $|\hat{h}_{T_j} - \hat{h}_{T_i}| \leq \exp[-n^{0.12}]$ if and only if they estimate the same signal carrier. Indeed, if $\hat{h}_{T_j}$ and $\hat{h}_{T_i}$ estimate the same signal carrier, then by (1.4), their difference is at most $2\exp[-n^{0.2}/4] < \exp[-n^{0.12}]$. On the other hand, if $\hat{h}_{T_i}$ and $\hat{h}_{T_j}$ estimate $h(i) \neq h(j)$, respectively, then

$$|\hat{h}_{T_i} - \hat{h}_{T_j}| \geq |h(i) - h(j)| - 2\exp[-n^{0.2}/4] \geq \exp[-n^{0.11}] - 2\exp[-n^{0.2}/4] > \exp[-n^{0.12}],$$

since on E_3^n, $\exp[-n^{0.11}] < |h(i) - h(j)|$. Hence the clusters C_i and C_j coincide if and only if $\pi(i) = \pi(j)$, otherwise they are disjoint. Thus, $\hat{f}_j$ is the average of all estimates of $h(\pi(j))$ and, therefore, $\hat{f}_j$ is a good estimate of $h(\pi(j))$. Since C_i and C_j coincide if and only if $\pi(i) = \pi(j)$, it obviously holds that

$$\hat{f}_i = \hat{f}_j \quad \text{if and only if} \quad \pi(i) = \pi(j). \tag{1.5}$$

We denote $\hat{f}(\bar{z}_i) := \hat{f}_j$, if $\pi(j) = i$; (1.5) implies $\hat{f}(\bar{z}_i)) \neq \hat{f}(\bar{z}_j)$, if $i \neq j$.
After phrase III we, therefore, (with high probability) end up with a sequence of estimators

$$\hat{f}(\bar{z}_{\pi(1)}), \ldots, \hat{f}(\bar{z}_{\pi(K)})$$

that correspond to the sequence of frequencies $h(\pi(1)), \ldots, h(\pi(1))$. Or, equivalently, $j \mapsto \hat{f}_j$ is a path of a skip-free random walk π on the set of different reals $\{\hat{f}(\bar{z}_{\pi_*}), \ldots, \hat{f}(\bar{z}_{\pi^*})\}$.
The problem is that the estimates, $\hat{f}(\bar{z}_{\pi(1)}), \ldots, \hat{f}(\bar{z}_{\pi(K)})$ are in the wrong order, i.e. we are not aware of the values $\pi(j)$, $j = 1, \ldots, K$. But having some information about the values $\pi(j)$ is necessary for estimating the frequencies $h(1), \ldots, h(n^2+1)$. So the question is: How can get from the sequence $\hat{f}(\bar{z}(_{\pi(1)}), \ldots, \hat{f}(\bar{z}_{\pi(K)})$ the elements $\hat{f}(\bar{z}_1), \ldots, \hat{f}(\bar{z}_{n^2+1})$? Or, equivalently: after observing the path of π on $\{\hat{f}(\bar{z}_{\pi_*}), \ldots, \hat{f}(\bar{z}_{\pi^*})\}$, how can we deduce $\hat{f}(\bar{z}_1), \ldots, \hat{f}(\bar{z}_{n^2+1})$?

1.6.3 Real scenery reconstruction algorithm

We now present the so-called *real scenery reconstruction algorithm* - $\mathcal{A}_n^{\mathbb{R}}$. This algorithm is able to answer to the stated questions up to the (swift by) one element.

The algorithm works due to the particular properties of π and $\{\hat{f}(\bar{z}_{\pi_*}),\ldots,\hat{f}(\bar{z}_{\pi^*})\}$. These properties are:

A1) $\pi(1)\in\{0,1\}$, i.e. the first estimated frequency of one's, $\hat{f}_1$ must be either an estimate of $h(1)$ or of $h(0)$. Unfortunately there is no way to find out which one of the two signal carriers $\bar{z}_0$ or $\bar{z}_1$ was visited first. This is why our algorithm can reconstruct the real scenery up to the first or last bit, only;

A2) $\pi(K)>n^2$. This is true, because we condition on $S(m^2)=m$ and we assume that there are at least n^2+1 2-s in $[0,m]$ (event E_1^n);

A3) π is skip-free (it does not jump);

A4) $\hat{f}(\bar{z}_i)\neq\hat{f}(\bar{z}_j)\ \forall j\neq i,\quad i,j\in\{\pi_*,\ldots,\pi^*\}$.

Algorithm 1.2. *Let $\hat{f}=(\hat{f}_1,\hat{f}_2,\ldots,\hat{f}_K)$ be the vector of real numbers such that the number of different reals in $\hat{f}$ is at least n^2+1. The vector $\hat{f}$ is the input for $\mathcal{A}_n^{\mathbb{R}}$.*

- *Define $\mathcal{R}_1:=\hat{f}_1$.*
- *From here on we proceed by induction on j : once $\mathcal{R}_j$ is defined, we define*

$$\mathcal{R}_{j+1}:=\hat{f}_s \text{ where } s:=1+\max\{j:\hat{f}_j=\mathcal{R}_j\}.$$

- *Proceed until $j=n^2+1$ and put*

$$\mathcal{A}_n^{\mathbb{R}}(\hat{f}):=\big(\mathcal{R}_2,\mathcal{R}_3,\ldots,\mathcal{R}_{n^2+1}\big).$$

The idea of the algorithm is very simple: Take the first element $\hat{f}_1$ of $\hat{f}$ and consider all elements of the input vector $\hat{f}$ that are equal to $\hat{f}_1$ and find the one with the biggest index (the last $\hat{f}_1$). Let j_1 be this index. Then take $\hat{f}_{j_1+1}$ as the first output. By **A1)**, $\hat{f}_1$ is either $\hat{f}(\bar{z}_0)$ or $\hat{f}(\bar{z}_1)$; by **A2)** and **A3)**, $\hat{f}_{j_1+1}$ ie either $\hat{f}(\bar{z}_1)$ or $\hat{f}(\bar{z}_2)$. Now look for the last $\hat{f}_{j_1+1}$. Let the corresponding index be j_2 and take $\hat{f}_{j_2+1}$ as the second output. By **A2)** and **A3)**, $\hat{f}_{j_1+1}$ is either $\hat{f}(\bar{z}_2)$ or $\hat{f}(\bar{z}_3)$ (depending whether the first output were $\hat{f}(\bar{z}_1)$ or $\hat{f}(\bar{z}_2)$). Proceed so n^2 times. As a result, on obtains one of the following vectors

$$(\hat{f}(\bar{z}_1),\ldots,\hat{f}(\bar{z}_{n^2}))\quad\text{or}\quad(\hat{f}(\bar{z}_2),\ldots,\hat{f}(\bar{z}_{n^2+1})).$$

This means $\mathcal{A}_n^{\mathbb{R}}(\hat{f})\in\{(\hat{f}(\bar{z}_1),\ldots,\hat{f}(\bar{z}_{n^2})),(\hat{f}(\bar{z}_2),\ldots,\hat{f}(\bar{z}_{n^2+1}))\}$, i.e.

$$\mathcal{A}_n^{\mathbb{R}}(\hat{f})\sqsubseteq(\hat{f}(\bar{z}_1),\ldots,\hat{f}(\bar{z}_{n^2+1})). \tag{1.6}$$

Phase IV Apply $\mathcal{A}_n^{\mathbb{R}}$ to $\hat{f}$. Denote the output $\mathcal{A}_n^{\mathbb{R}}(\hat{f})$ by $(f_1,\ldots,f_{n^2})$. By (1.6) ,

$$(f_1,\ldots,f_{n^2})\sqsubseteq\big(\hat{f}(\bar{z}_1),\ldots,\hat{f}(\bar{z}_{n^2+1})\big). \tag{1.7}$$

Now recall that we are interested in reconstructing the $g_i(\xi_0^m) := I_{[0,5)}(h(i))$ rather than $\hat{h}(i)$. Thus, having estimates for $h(\bar{z}_i)$, namely $\hat{f}(\bar{z}_i)$, we use the obvious estimator for g_i: $I_{[0,0.5)}(f_i)$.

Phase V Define the final output of $\hat{g}$

$$\hat{g}(\chi_0^{m^2}) := \Big(I_{[0.5,1]}(f_1), \ldots I_{[0.5,1]}(f_{n^2})\Big).$$

Recall that because of E_3^n, with high probability all random variables $h(1), \ldots, h(n^2+1)$ are more than $\exp(-n^{0.11})$ apart from $\frac{1}{2}$. Since $\exp(-n^{0.11})$ is much bigger than the preciseness of our estimate, with high probability we have $\hat{f}(\bar{z}_i) < 0.5$ if and only if $h(\bar{z}_i) < 0.5$. By (1.7) this means

$$\hat{g}(\chi_0^{m^2}) = \Big(I_{[0.5,1]}(f_1), \ldots I_{[0.5,1]}(f_{n^2})\Big) \sqsubseteq \Big(I_{[0.5,1]}\big(h(\bar{z}_1)\big), \ldots I_{[0.5,1]}\big(h(z_{n^2+1})\big)\Big) = g(\xi_0^m).$$

Hence, when $E^n_{\text{cell\_OK}}$ holds, then $\hat{g}$ is properly defined and the probability (1.3) is high. In particular, by choosing n big enough, it can be proven to be greater that $\frac{3}{4}$. Since we are not interested in $\hat{g}$ beyond $E^n_{\text{cell\_OK}}$, we extend the definition of $\hat{g}$ arbitrary.

2 Whole truth about signal probabilities

In the previous section we considered the case where the scenery has three colors: $\{0,1,2\}$. The locations of the 2's where called signal carriers. The i-th such place was denoted by $\bar{z}_i$. In reality we have only two colors 0 and 1. Thus, we need to show that with 2 colors we also manage to define signal carriers $\bar{z}_i$ in such a way that all of the following holds:

a) Whenever the random walk passes by a signal carrier, we can recognize that the random walk is close to a signal carrier by looking at the observations (with high probability).

b) The probability to be induced in error by the observations, so that one infers that at a certain time one is close to a signal carrier when one is not, is small. This type of mistake never happens up to time m^2.

c) When we pass a signal carrier we are able to estimate its frequency of one's with high precision (with high probability).

In the present section, we define and investigate an important concept that leads to the signal carriers: Markov signal probability.

2.1 Definitions

In this subsection, we define the main notions of the section: delayed signal probability, strong signal probability and Markov signal probability. We also give a few equivalent characterizations of these concepts, and we try to explain their meaning. In the end of the subsection we give a formal definition of the frequency of ones.

- Let $D \subset \mathbb{Z}$ and let $\zeta : D \rightarrow \{0,1\}$. For example, ζ can be the scenery, ξ or the observations, χ. Let $T = [t_1, t_2] \subset D$ be an integer interval of length at least 3. Then we say that T is a *block* of ζ if and only if we have that

$$\zeta(t_1) = \zeta(t_2) \neq \zeta(t), \forall t \in]t_1, t_2[$$

We call $t_2 - t_1$ the length of the block T. The point t_1 is called the beginning of the block. For example, T is a block of ζ with length 4, if $\zeta|T = 01110$.

- Let $T = T(\chi) \subset \mathbb{N}$ be a time interval, possibly depending on the observations. For example, T can be a block of χ or $T = [t, t+n^2]$ can be a time interval of length n^2+1 such that $\chi(t) = \chi(t+1) = \cdots = \chi(t+n^2)$. Let $I \subset \mathbb{Z}$ be an integer interval (a location set). We say that T *was generated* (by S) *on* I, if and only if $\forall t \in T, S(t) \in I$.

- We now define the delayed signal probability. To simplify the notations afterwards, define

$$M = M(n) := n^{1000} - n^2, \quad \tilde{M} := n^{1000} - 2n^2.$$

Fix $z \in \mathbb{Z}$ and let S_z denote the random walk translated by z, i.e. for all $t \in \mathbb{N}$, $S_z(t) := S(t) + z$. We define the random variable δ_z^d in the following way:

$$\delta_z^d := P\left(\xi(S_z(M)) = \cdots = \xi\left(S_z\left(n^{1000}-1\right)\right) = \xi\left(S_z\left(n^{1000}\right)\right) \Big| \xi\right). \tag{2.1}$$

In other words, δ_z^d is the conditional probability (conditional on ξ) to observe only one color in the time interval $[n^{1000} - n^2, n^2]$ if the random walk starts at z. We shall call δ_z^d *delayed signal probability at* z.
During time n^{1000} the random walk can not move more than Ln^{1000}. Thus, δ_z^d depends only on the scenery ξ in the interval $\left[z - Ln^{1000}, z + Ln^{1000}\right]$.

- Let, for each $z \in \mathbb{Z}$

$$I_z := [z - Ln^{1000}, z + Ln^{1000}]. \tag{2.2}$$

We have that δ_z^d is a random variable which is measurable with respect to $\sigma(\xi(s)|s \in I_z)$. Since the distribution of ξ is translation invariant, the distribution of δ_z^d does not depend on z.

- For some technical reason, we need a stronger version of the delayed signal probability. Again, let $z \in \mathbb{Z}$. We define the *strong signal probability* at z, $\tilde{\delta}_z^d$, as follows

$$\tilde{\delta}_z^d := P\Big(\xi(S_z(M)) = \cdots = \xi(S_z(n^{1000})), \quad S_z(M+1), S_z(M+2), \ldots, S_z(n^{1000}) \in [z - L\tilde{M}, z + L\tilde{M}] \Big| \xi\Big).$$

Note that $\tilde{\delta}_z^d$ is measurable with respect to the sigma algebra $\sigma(\xi(s)|s \in [z - L\tilde{M}, z + L\tilde{M}])$. Also note that, obviously, $\delta_z^d \geq \tilde{\delta}_z^d$. However, the difference is not too big. Indeed, Höffding inequality states that for some constant $d > 0$

$$\begin{aligned}\delta_z^d - \tilde{\delta}_z^d &= P\Big(\xi(S_z(M)) = \cdots = \xi(S_z(n^{1000})), \quad \exists s \in \{M, \ldots, n^{1000}\} : |z - S_z(s)| > L\tilde{M} \Big| \xi\Big) \\ &\leq P\Big(|S(M)| > L(\tilde{M} - n^2)\Big) \leq \exp(-dn^{999}).\end{aligned} \tag{2.3}$$

Next we define the Markov signal probability at z. Roughly speaking, the Markov signal probability at z, denoted by δ_z^M, is the conditional (on ξ) probability to have (at least) n^2+1 times the same color generated on I_z exactly $n^{1000}-n^2$ after we observe n^2+1 times the same color generated on I_z. In this formulation the part "after we observe a string of n^2+1 times the same color generated on I_z" needs to be clarified. The explanation is the following: every time there is in the observations n^2+1 times the same color generated on I_z, we introduce a stopping time $\nu_z(i)$. The position of the random walk at these stopping times defines a Markov chain with state space I_z. As we will prove later, this Markov chain $\{S(\nu_z(k))\}_{k\geq 1}$ converges very quickly to a stationary measure, say μ_z. So, by "M after we observe n^2+1 times the same color generated on I_z" we actually mean: "M time after starting the random walk from an initial position distributed according to μ_z". Since the distribution of $S(\nu_z(i))$ converges quickly to μ_z, δ_z^M is close to the probability of observing n^2+1 times the same color generated on I_z exactly M time after time $\nu_z(i)$. In other words, δ_z^M is close to the conditional (on ξ) probability of the event that we observe only one color in the time interval $[\nu_z(i)+n^{1000}-n^2, \nu_z(i)+n^{1000}]$ and that during that time interval the random walk S is in I_z. Thus (for k big enough) δ_z^M is close to:

$$P\Big(\chi(\nu_z(i)+M)=\cdots=\chi(\nu_z(i)+n^{1000}) \quad \text{and} \quad S(\nu_z(i)+M),\ldots,S(\nu_z(i)+n^{1000})\in I_z|\xi\Big). \tag{2.4}$$

The ergodic theorem then implies that on the long run the proportion of stopping times $\nu_z(i)$ which are followed after M by n^2+1 observations of the same color generated on I_z converges a.s. to δ_z^M. Actually, to make some subsequent proofs easier, we do not take a stopping time $\nu_z(i)$ after each n^2+1 observations of the same color generated on I_z. Rather we ask that the stopping times be apart by at least $e^{n^{0.1}}$.

In order to prove how quickly we converge to the stationary measure, we also view the explained notions in terms of a regenerative process. The renewal times will be defined as the stopping times, denoted by $\vartheta_z(k)$, which stop the random walk at the point $z-2Le^{n^{0.1}}$. To simplify some proofs, we also require that there is at least one stopping $\nu_z(i)$ between $\vartheta_z(k)$ and $\vartheta_z(k+1)$. Thus $\vartheta_z(0)$ denotes the first visit by the random walk S to the point $z-2Le^{n^{0.1}}$. We define $\nu_z(1)$ to be the first time after $\vartheta_z(0)$ where there happens to be n^2+1 times the same color generated on I_z. Then, $\vartheta_z(1)$ is the first return of S to $z-2Le^{n^{0.1}}$ after $\nu_z(1)$ and so on. Let us give the formal definitions of all introduced notions.

- Let $\vartheta_z(0)$ denote the first visit of S to the point $z-2Le^{n^{0.1}}$. Thus

$$\vartheta_z(0)=\min\{t\geq 0|\, S(t)=z-2Le^{n^{0.1}}\}.$$

- Let $\nu_z(1)$ designate the first time after $\vartheta_z(0)$ where we observe n^2+1 zero's or one's in a row, generated on I_z. More precisely:

$$\nu_z(1):=\min\left\{t>\vartheta_z(0)\,\middle|\, \begin{array}{c}\chi(t)=\chi(t-1)=...=\chi\left(t-n^2\right)\\ \text{and } S(t-n^2),S(t-n^2+1),\ldots,S(t)\in I_z\end{array}\right\}.$$

Once $\nu_z(i)$ is well defined, define $\nu_z(i+1)$ in the following manner:

$$\nu_z(i+1):=\min\left\{t>\nu_z(i)+e^{n^{0.1}}\,\middle|\, \begin{array}{c}\chi(t)=\chi(t-1)=...=\chi\left(t-n^2\right)\\ \text{and } S(t-n^2),S(t-n^2+1),\ldots,S(t)\in I_z\end{array}\right\}.$$

- Let $\vartheta_z(k)$ denote the consecutive visits of S to the point $z-2Le^{n^{0.1}}$ provided that between two visits random walk S generates (at least once) $n+1$ consecutive 0-s or 1-s on I_z. More precisely,

$$\vartheta_z(k+1) := \min\{t > \vartheta_z(k) | S(t) = z - 2Le^{n^{0.1}}, \exists j : \vartheta_z(k) < \nu_z(j) < t\}, \quad k = 1, 2 \dots.$$

  Basically, the definition above says: if $\vartheta_z(k)$ is defined, we wait until we observe n^2+1 same colors generated on I_z. Since $S(\vartheta_z(k)) = z - 2Le^{n^{0.1}}$, then the first n^2+1 same colors generated on I_z can not happen earlier than $e^{n^{0.1}}$ times after $\vartheta_z(k)$. This means, the first n^2+1 same colors generated on I_z can not happen earlier than $e^{n^{0.1}}$ times after last stopping time ν_z, say $\nu_z(i)$ (this happens before $\vartheta_z(k)$). Thus, the first n^2+1 same colors generated on I_z is actually $\nu_z(i+1)$. Observing $\nu_z(i+1)$, we just wait for the next visit of S to the $z-2Le^{n^{0.1}}$. This defines $\vartheta_z(k+1)$.

- Let $X_{z,i}$, $i = 1, 2, \dots$ designate the Bernoulli variable which is equal to one if exactly after time M the stopping time $\nu_z(i)$ is followed by a sequence of n^2+1 one's or zero's generated on I_z. More precisely, $X_{z,i} = 1$ if and only if

$$\chi(\nu_z(i) + M) = \chi(\nu_z(i) + M + 1) = \dots = \chi(\nu_z(i) + n^{1000})$$

  and

$$S(j) \in I_z \quad \forall j = \nu_z(i) + M, \dots, \nu_z(i) + n^{1000}$$

- Define $\kappa_z(0) := 0$. Let $\kappa_z(k)$ designate the number of stopping times $\nu_z(k)$ occurring during the time from $\vartheta_z(0)$ to $\vartheta_z(k)$. Thus $\kappa_z(k)$ is defined by the inequalities:

$$\nu_z(\kappa_z(k)) \leq \vartheta_z(k) < \nu_z(\kappa_z(k) + 1).$$

  For all k, $S(\vartheta_z(k)) = z - 2Ln^{1000}$. Hence, for all i, $\vartheta_z(k) \neq \nu_z(i)$ and the inequalities above are strict.

- Define the following variables:

$$\mathcal{X}_z(k) = \sum_{i=\kappa(k-1)+1}^{\kappa(k)} X_{z,i}, \quad \mathcal{Z}_z(k) = \kappa(k) - \kappa(k-1), \quad k = 1, 2, \dots$$

  Thus, $\mathcal{Z}_z(k)$ is the number of stopping times occurring during the time interval from time $\vartheta_z(k-1)$ to time $\vartheta_z(k)$. Note that $\mathcal{Z}_z(k) \geq 1$, $\forall k$. The random variable $\mathcal{X}_z(k)$ designates the number of such stopping times which, during the same time interval, have been followed exactly after time M by a sequence of n^2+1 0's or 1's generated on I_z. Note that conditional on ξ the variables $\mathcal{X}_z(1)$, $\mathcal{X}_z(2), \dots$ are i.i.d. and the same holds for $\mathcal{Z}_z(1)$, $\mathcal{Z}_z(2), \dots$.

- Fix ξ and z. Let $Y_i := S(\nu_z(i))$, $i = 1, 2, \dots$ denote the Markov chain obtained by stopping the random walk S by $\nu_z(i)$. The state space of Y_i is I_z. Because of the nature of S, Y_i is finite, irreducible aperiodic and, therefore, an ergodic Markov chain.

- Let μ_z denote the stationary distribution of $\{Y_k\}$. In the present section z is fixed, so we write μ. The measure μ is a discrete probability measure on I_z, so $\mu = (\mu(j))_{j\in I_z}$. For each state, $j \in I_z$ define the hitting times $\tau_j(l)$, $l = 1, 2, 3, \ldots$. Formally,

$$\tau_j(1) := \min\{i \geq 1 : Y_i = j\}, \quad \tau_j(l) := \min\{i > \tau_j(l-1) : Y_i = j\}, \quad l = 2, 3 \ldots.$$

- We define:

$$\delta_z^M := \frac{E\left[\mathcal{X}_z(1)|\,\xi\right]}{E\left[\mathcal{Z}_z(1)|\,\xi\right]}. \tag{2.5}$$

We call δ_z^M *Markov signal probability* at z.

In the following we give some equivalent forms of (2.5). Note that conditional on ξ, $X_{z,i}$ is a regenerative process with respect to the renewal $\kappa_z(k)$. Hence, conditioning on ξ, we have

$$\lim_{r\to\infty} \sum_{i=1}^{r} \frac{X_{z,i}}{r} = \lim_{k\to\infty} \sum_{i=1}^{\kappa(k)} \frac{X_{z,i}}{\kappa(k)} = \lim_{k\to\infty} \frac{\sum_{i=1}^{k} \mathcal{X}_z(i)}{\sum_{i=1}^{k} \mathcal{Z}_z(i)} = \frac{E\left[\mathcal{X}_{z,1}|\,\xi\right]}{E\left[\mathcal{Z}_{z,1}|\,\xi\right]}. \quad \text{a.s.} \tag{2.6}$$

We count (up to time r) all sequences of length n^2+1 of one's or zero's, generated on the interval I_z according to the stopping times $\nu_z(i)$, $k = 1, 2, \ldots$. Among such sequences, the proportion of those sequences which are followed after exactly time M by another sequence of n^2+1 zero's or one's generated on the interval I_z converges a.s. to δ_z^M, as r goes to infinity.

On the other hand,

$$\frac{1}{r}\sum_{i=1}^{r} X_{z,i} = \sum_j \frac{N_j(r)}{r} \frac{1}{N_j(r)} \sum_{l=1}^{N_j(r)} X_{z,\tau_j(l)},$$

where $N_j(r) := \max\{l : \tau_j(l) \leq r\}$, $r = 1, 2, 3, \ldots$. Since $\tau_j(l)$, $l = 1, 2, 3, \ldots$ is a (delayed) renewal process with the corresponding renewal numbers $N_j(r)$ and with the expected renewal time $\frac{1}{\mu(j)}$ we get

$$\frac{N_j(r)}{r} \to \mu(j) \quad \text{a.s.}.$$

On the other hand, $X_{z,i}$ is a regenerative process with respect to each $\tau_j(l)$, $l = 1, 2, 3, \ldots$. Hence

$$\frac{1}{N_j(r)} \sum_{l=1}^{N_j(r)} X_{z,\tau_j(l)} \to E[X_{z,\tau_j(2)}], \quad \text{as} \quad r \to \infty \quad \text{a.s.}.$$

Since $E[X_{z,\tau_j(2)}] = P(X_{z,\tau_j(2)} = 1)$. The latter equals

$$P\Big(S_j(M), S_j(M+1), \cdots, S_j(n^{1000}) \in I_z \quad \text{and} \quad \xi(S_j(M)) = \xi(S_j(M+1)) = \cdots = \xi(S_j(n^{1000}))\Big).$$

This can be rewritten as

$$\sum_{l\in I_z} P(j,l)\delta_z(l),$$

where $P(j, l) := P(S(M) = j - l)$ and

$$\delta_z(l) := P\Bigg(S_l(0), S_l(1), \ldots, S_l(n^2) \in I_z \quad \text{and} \quad \xi(S_l(0)) = \xi(S_l(1)) = \ldots = \xi(S_l(n^2))\Bigg) \tag{2.7}$$

Hence

$$\delta_z^M = \sum_{j\in I_z} \mu(j) P\Big(S_j(M), S_j(M+1), \cdots, S_j(n^{1000}) \in I_z, \quad \xi(S_j(M)) = \cdots = \xi(S_j(n^{1000})\Big) \tag{2.8}$$

or

$$\delta_z^M = \sum_{j,l\in I_z} \mu(j) P(j,l) \delta_z(l). \tag{2.9}$$

Using the same notation, we have an equivalent form of writing the delayed signal probability

$$\delta_z^d = \sum_{l=I_z} P(z,l)\delta_z(l). \tag{2.10}$$

Formula (2.9) can be interpreted as follows: let U be a random variable with distribution μ_z and let S be a random walk, independent of U. Let S_U denote the translation of S by U, i.e., for each t, $S_U(t) = U + S(t)$. Then (2.9) states

$$\delta_z^M = P\Big(\xi(S_U(M)) = \cdots = \xi(S_U(n^{1000})) \quad \text{and} \quad S_U(M), \cdots, S_U(n^{1000}) \in I_z|\xi\Big). \tag{2.11}$$

Thus, δ_z^M is the limit-version of (2.4) when $i \to \infty$.

We now define the frequency of ones. To obtain the consistency with the Markov signal probability, we formally define the frequency of ones in terms of regenerative processes. However, we also derive the analogue of (2.11), which explains the meaning of the notion.

- Let $U_{z,i} = \xi(S(\nu_z(i) + e^{n^{0.1}}))$ and define

$$\mathcal{U}_z(k) := \sum_{i=\kappa(k-1)+1}^{\kappa(k)} U_{z,i}.$$

- Let

$$h(z) := \frac{E(\mathcal{U}_z(1)|\xi)}{E(\mathcal{Z}_z(1)|\xi)}.$$

  The random variable $h(z)$ is $\sigma(\xi(i) : i \in [z - L(n^{1000} + e^{n^{0.1}}), z + L(n^{1000} + e^{n^{0.1}})])$-measurable; $h(z)$ is called as *frequency of ones* at z. As in (2.6), conditioning on ξ, we have

$$\lim_{r\to\infty} \sum_{i=1}^{r} \frac{\mathcal{U}_{z,i}}{r} = h(z) \quad \text{a.s..}$$

  With the same argument as above, we get

$$\lim_{r\to\infty} \frac{1}{r}\sum_{i=1}^{r} U_{z,i} = \lim_{r\to\infty} \sum_j \frac{N_j(r)}{r}\frac{1}{N_j(r)}\sum_{l=1}^{N_j(r)} U_{z,\tau_j(l)} = \sum_j \mu(j) E(U_{z,\tau_j(2)}).$$

  Now,

$$E(U_{z,\tau_j(2)}) = \sum_{i=j-Le^{n^{0.1}}}^{i=j+Le^{n^{0.1}}} \xi(i) P\big(S_j(e^{n^{0.1}}) = i\big)$$

and, therefore

$$h(z)=\sum_{j=I_z}\mu(j)\sum_{i=j-Le^{n^{0.1}}}^{j+Le^{n^{0.1}}}\xi(i)P\big(S_j(e^{n^{0.1}})=i\big)=\sum_{i=z-L(n^{1000}+e^{n^{0.1}})}^{z+L(n^{1000}+e^{n^{0.1}})}\xi(i)\sum_{j=I_z}\mu(j)P(S_j(e^{n^{0.1}})=i). \tag{2.12}$$

Now, it is easy to see that in terms of U and S as in (2.11), i.e. U and S are independent, U has law μ_z, we have

$$h(z)=P(\xi(U+S(e^{n^{0.1}}))=1|\xi)=E[\xi(U+S(e^{n^{0.1}}))|\xi], \tag{2.13}$$

2.2 Auxiliary results

In the present section we investigate the relations between δ_z^M and δ_z^d. Note that they only depend on the scenery ξ in the interval $[z-Ln^{1000},z+Ln^{1000}]$. In other words,

$$\delta_z^M,\delta_z^d\in\sigma\Big(\xi(j)|j\in[z-Ln^{1000},z+Ln^{1000}]\Big).$$

The distribution of both δ_z^M and δ_z^d does not depend on particular choice of z. Hence, without loss of generality, in the following we consider the point $z=0$, only.

Define $p_M:=\max\{P(S(M)=z)|z\in\mathbb{Z}\}$. We call a block *big*, if its length is bigger than $\frac{n}{\ln n}$.

Proposition 2.1. *For any $c_\delta\in[p_M,2p_M]$, the following statement hold:*

a $P(\delta_z^d\geq c_\delta)\leq\exp(-\alpha n/\ln n)$, where $\alpha:=\ln(1.5)$

b $P\left(\delta_z^d\geq c_\delta\right)\geq(0.5)^n>\exp\left(-n\right)$

c If all blocks of $\xi|[z-Ln^{1000},z+Ln^{1000}]$ are shorter than $n/\ln n+1$, then $\delta_z^d<c_\delta$. Formally:

$$\left\{\delta_z^d\geq c_\delta\right\}\subseteq\{\ [z-Ln^{1000},z+Ln^{1000}]\text{ contains a big block of }\xi\ \}$$

d Conditional on $\left\{\delta_z^d\geq c_\delta\right\}$ it is likely that $[z-Ln^{1000},z+Ln^{1000}]$ contains at most $0.5\ln n$ big blocks of ξ. More precisely:

$$P\left(E_{\delta,z}^c\middle|\,\delta_z^d\geq c_\delta\right)\leq\left(2Ln^{1000}\right)^{0.5\ln n}(0.5)^{-0.5n}$$

where

$$E_{\delta,z}:=\{\ [z-Ln^{1000},z+Ln^{1000}]\text{ has less than }0.5\ln n\text{ big blocks of }\xi\ \}$$

In order to prove Proposition 2.1, we use the following lemma. The proof of it can be found in [LMM04].

Lemma 2.1. *There exists a constant $a>0$ such that for each $t,r\in\mathbb{N}$, for each subset $I\subset\mathbb{Z}$, and for each $j\in I$ and for every mapping $\zeta:\mathbb{Z}\to\{0,1\}$, the following implication holds*

if all blocks of ζ in I are shorter or equal to r, then

$$P\left(\begin{array}{c}\zeta\left(S_j\left(0\right)\right)=\zeta\left(S_j\left(1\right)\right)=\cdots=\zeta\left(S_j\left(t\right)\right)\\ \textit{and } S_j\left(0\right),S_j\left(1\right),...,S_j\left(t\right)\in I\end{array}\right)\le\exp\left(-\frac{at}{r^2}\right).$$

Proof that c holds: Without loss of generality assume $z=0$. Suppose that the length of all blocks of $\xi|[-Ln^{1000},Ln^{1000}]$ is at most $n/\ln n$. Let $I:=[-Ln^{1000},Ln^{1000}]$. Denote $\delta(l)=\delta_0(l)$, where $\delta_0(l)$ is as in (2.7). If the all the blocks in I are not longer than $n/\ln n$ we get by Lemma 2.1 that for all $j\in I$

$$\delta(j)\le\exp\Big(-\frac{an^2}{(n/\ln n)^2}\Big)=n^{-a\ln n}.$$

By (2.10) we get that

$$\delta_0^d=\sum_{j=-Ln^{1000}}^{Ln^{1000}}P(0,j)\delta(j)\le\sum_{j=-Ln^{1000}}^{Ln^{1000}}P(0,j)n^{-a\ln n}\le n^{-a\ln n}\tag{2.14}$$

The expression on the right side of the last inequality is of smaller order than any negative polynomial order in n. By the local central limit theorem p_M is of order $n^{-\frac{M}{2}}$. Thus, for n big enough

$$\delta_0^d<p_M\le c_\delta.$$

Proof that a holds: Without loss of generality assume $z=0$. Define the event

$$E_z:=\{\xi(z)=\xi(z+1)=\cdots=\xi(z+\frac{n}{\ln n})\}$$

Part c states that

$$\{\delta_0^d\ge c_\delta\}\subseteq\bigcup_{z\in[-Ln^{1000},Ln^{1000}]}E_z.$$

Thus,

$$P\left(\delta_0^d\ge c_\delta\right)\le\sum_{z=-Ln^{1000}}^{Ln^{1000}}P(E_z).$$

Now, clearly

$$P(E_z)=\exp\left(-\frac{\ln(2)n}{\ln n}\right).$$

So,

$$P(\delta_0^d\ge c_\delta)\le 2Ln^{1000}\exp\left(-\frac{\ln\left(2\right)n}{\ln n}\right).\tag{2.15}$$

The dominating term in the product on the right side (2.15) is $\exp\left(-\ln\left(2\right)n/\ln n\right)$. Hence, for n big enough, the expression on the right side of (2.15) is smaller than $\exp(-\frac{\ln(1.5)n}{\ln n})$.

Proof that b holds: It suffices to prove that

$$P(\delta_z^d \geq 2p_M) \geq (0.5)^n.$$

Without loss of generality assume $z = 0$. Define $E := \{\xi(0) = \xi(1) = \cdots = \xi(n)\}$. We are going to show that

$$E \subseteq \{\delta_0^d \geq 2p_M\} \quad \text{and} \quad P(E) \geq \exp(-n).$$

Recall the definition of $\delta(j)$. If E holds, then for any $j \in [0, n]$ we have

$$\delta(j) \geq P\Big(S_j(t) \in [0, n], \forall t \in [0, n^2]\Big)$$

Now, because of the central limit theorem, there is a constant $b > 0$ not depending on n, such that for all $j \in [n/3, 2n/3]$ we have:

$$P\Big(S_j(t) \in [0, n], \forall t \in [0, n^2]\Big) > b.$$

By the local central limit theorem, again, for all $j \in [n/3, 2n/3]$ we have, for n big enough, that

$$P(0, j) \geq \frac{p_M}{2}. \tag{2.16}$$

Using (2.10) and (2.16) we find that when E holds, then

$$\delta_0^d \geq \sum_{j=\frac{n}{3}}^{\frac{2n}{3}} bP(0, j) \geq \frac{bnp_M}{6}. \tag{2.17}$$

For n big enough, obviously the right side of (2.17) is bigger than $2p_M$. This proves $E \subseteq \{\delta_0^d \geq 2p_M\}$. Furthermore, we have that $P(E) = 0.5^n$. The inequality $0.5^n > \exp(-n)$ finishes the proof.

Proof that d holds: Without loss of generality assume $z = 0$. For a block T, the point $\inf T$ is called the beginning of the block. Let $t_1, t_2, \ldots$ denote the beginnings of the consecutive big blocks in $[-Ln^{1000}, \infty)$. Define $t_0 := -Ln^{1000}$ and $g_i := t_i - t_{i-1}$, $i = 1, 2, \ldots$. So, g_i measures the distances between consecutive big blocks. Clearly, g_i-s are i.i.d. Note,

$$E_{\delta,0}^c \subset \Big\{ \sum_{i=1}^{0.5 \ln n} g_i \leq 2Ln^{1000} \Big\} \subset \cap_{i=1}^{0.5 \ln n} \Big\{ g_i < 2Ln^{1000} \Big\}.$$

Note

$$P(g_1 < 2Ln^{1000}) \leq \sum_{z=t_0}^{Ln^{1000}-1} P(\text{a big block begins at } z) \leq 2Ln^{1000}(0.5)^{\frac{n}{\ln n}}.$$

Hence,

$$P(E_{\delta,0}^c) \leq P(g_i \leq 2Ln^{1000})^{0.5 \ln n} = \big(2Ln^{1000}\big)^{0.5 \ln n} (0.5)^{0.5n}.$$

Combining this with b, we get

$$P(E_{\delta,0}^c | \delta_0^d > c_\delta) \leq \frac{P(E_{\delta,0}^c)}{P(\delta_0^d > c_\delta)} \leq \big(2Ln^{1000}\big)^{0.5 \ln n} (0.5)^{-0.5n} \to 0.$$

Lemma 2.2.

$$P\left(\delta_z^d \geq c_\delta\right) \left(2Ln^{1000}\right)^{-0.5 \ln n} \leq 2P\left(\delta_z^d \wedge \delta_z^M \geq c_\delta(1 - O(M^{-\frac{1}{2}}))\right).$$

2.3 Proof of Lemma 2.2

In the present subsection we prove Lemma 2.2. To the end of the section we assume $z = 0$. At first we define fences.

Fences

An interval $[t, t+4L-1] \subset D$ is called a *fence* of ζ, if

$$0 = \zeta(t) = \zeta(t+1) \cdots = \zeta(t+L-1) \neq \zeta(t+L) = \cdots = \zeta(t+2L-1) \neq \\ \zeta(t+2L) = \cdots = \zeta(t+3L-1) \neq \zeta(t+3L) = \cdots = \zeta(t+4L-1)$$

The point $t+2L$ is the *breakpoint* of the fence. So, T is a fence of ζ corresponding to the $L = 3$, if and only if $\zeta|T = 000111000111$.

Let $z_0 := -Ln^{1000}$ and let z_k, $k = 1, 2, \ldots$ be defined inductively: z_k denotes the breakpoint of the first fence of scenery ξ in $[z_k + 4L, \infty)$. We call the points z_k the breakpoints of consecutive fences (of scenery ξ). Define $l_i := z_i - z_{i-1}$, $i = 1, 2, \ldots$ and $N := \max\{k : z_{k-1} \leq Ln^{1000}\} < Ln^{1000}$. The random variables l_i measure the distances between the breakpoints of consecutive fences, they are i.i.d. Let $l := Ln^{1000} - z_N$, $l \leq l_{N+1}$. The moment generating function of l_1, say $M(\lambda)$, does not depend on n and it is finite, if $\lambda > 0$ is small enough. Let $M := \exp(\lambda l_1) < \infty$ and choose $C > 1$ such that $\lambda C > 1$. Now define the event

$$E_b := \{l, l_i \leq Cn, \quad i = 1, 2, \ldots, N\}$$

and apply the large deviation inequality to see $P(l_1 > Cn) = P(\lambda l_1 > \lambda Cn) < Me^{-\lambda Cn}$. Now,

$$P(E_b^c) \leq \sum_{i=1}^{Ln^{1000}} P(l_i > Cn) = Ln^{1000} P(l_1 > Cn) < Ln^{1000} M e^{-\lambda Cn}.$$

Applying b, we get

$$P(E_b^c | \delta_0^d \geq c_\delta) \leq \frac{P(E_b^c)}{P(\delta_0^d \geq c_\delta)} \leq Ln^{1000} M e^{(1-\lambda C)n} \to 0. \tag{2.18}$$

Mapping

Let $\mathcal{O}$ denote the set of all possible pieces of sceneries in $I := [-Ln^{1000}, Ln^{1000}]$, i.e. $\mathcal{O} := \{0,1\}^I$. The random variables δ_0^d, δ_0^M as well as the events $\{\delta_0^d > c_\delta\}$, $E_{\delta,0}$, E_b depend on the restriction of the scenery to I, only. Hence they can be defined on the probability space $(\mathcal{O}, 2^{\mathcal{O}}, P)$, where P stands for the normalized counting measure.

Define

$$\mathcal{C} := \{\delta_0^d > c_\delta\} \cap E_{\delta,0} \cap E_b \subset \mathcal{O}.$$

Hence $\mathcal{C}$ consists of all pieces of sceneries, η, with the following properties: $\delta_0^d(\eta)$ is bigger than c_δ, the number of big blocks is less than $0.5 \ln n$ and the gaps between the breakpoints of the consecutive fences in I is at most Cn.

Let $\eta \in \mathcal{C}$ and let $z_0, z_1, \ldots, z_N$ be the breakpoints of consecutive fences (restricted to I) of η. Since $\eta \subset E_b$, we have $N \geq 2Ln^{999}$. Now partition the interval I as follows:

$$I = I_1 \cup I_2 \cup \cdots \cup I_N \cup I_{N+1}, \tag{2.19}$$

where $I_k := [z_{k-1}, z_k - 1]$, $k = 1, \ldots, N, I_{N+1} := [I_N, Ln^{1000}]$. Let $l(I_k) := z_k - k_{k-1}$ denote the length of I_k. We shall call the partition (2.19) the fence-partition corresponding to η. The fences guarantee that any block of η, that is longer than L is a proper subset of one interval I_k. Since $\eta \in \{\delta_0^d > c_\delta\} \cap E_{\delta,0}$, there is at least one and at most $0.5 \ln n$ big blocks. Let I_k^*, $k = 1, \ldots N^*$, $N^* \leq 0.5 \ln n$ denote the $k-th$ interval containing at least one big block. Similarly, let I_k^o, $k = 1, \ldots, N + 1 - N^*$ denote the $k - th$ interval with no big blocks. Clearly, most of the intervals I_k are without big blocks, in particular $\sum_k l(I_k^o) > Ln^{1000}$. Define

$$j^o := \min\{j : \sum_{k=1}^{j} l(I_k^o) > Ln^{1000}\}.$$

To summarize - to each $\eta \in \mathcal{C}$ corresponds an unique fence-partition, an unique labelling of the interval according to the blocks, and, therefore, unique j^o. We now define a mapping $B : \mathcal{C} \to \mathcal{O}$ as follows:

$$B(\eta) := (\eta|I_1^o, \eta|I_2^o, \ldots, \eta|I_{j^o}^o, \eta|I_1^*, \ldots, \eta|I_{N^*}^*, \eta|I_{j^o+1}^o, \ldots, \eta|I_{N+1-N^*}^o).$$

We also define the corresponding permutation:

$$\Pi_\eta : I \to I, \quad \Pi_\eta(I) = (I_1^o, I_2^o, \ldots, I_{j^o}^o, I_1^*, \ldots, I_{N^*}^*, I_{j^o+1}^o, \ldots, I_{N+1-N^*}^o).$$

Thus, $B(\eta) = \eta \circ \Pi_\eta$.

Since all big blocks of η are contained in the intervals I_k, the mapping B keeps all big blocks unchanged, and just moves them closer to the origin.

The mapping B is clearly not injective. However, $B(\eta_1) = B(\eta_2)$ implies that the fence-partitions corresponding to η_1 and η_2 consists of the same intervals, with possibly different order. Also the intervals with big blocks (marked with star) are the same, but possibly differently located. Moreover, the ordering of the similarly marked blocks corresponding to η_1 and η_2 are the same (i.e. if the 8-th interval, I_8, of the partition corresponding to η_1 is the 20-th interval, I_{20}, of the partition corresponding to η_2, then their marks are the same. If I_8 in its partition is the seventh interval with o ($I_8 = I_7^o$ in the partition corresponding to the η_1), then the same block in the second partition must be also the seventh interval with o ($I_{20} = I_7^o$ in the partition corresponding to η_2). Therefore, the partition corresponding to η_1 and η_2 differ on the location of the star-intervals, only. Since the number of intervals is smaller than $2Ln^{1000}$ and the number of star-intervals is at most $0.5 \ln n$, the number of different partitions with the properties described above, is less than $(2Ln^{1000})^{0.5 \ln n}$. This means

$$|B(\mathcal{C})|(2Ln^{1000})^{0.5 \ln n} > |\mathcal{C}|. \tag{2.20}$$

Proof of Lemma 2.2: Because of the counting measure and (2.20) we get

$$\frac{P(B(\mathcal{C}))}{P(\mathcal{C})} = \frac{|B(\mathcal{C})|}{|\mathcal{C}|} > (2Ln^{1000})^{-0.5 \ln n}.$$

By Propositions 2.2 and 2.3,

$$P(B(\mathcal{C})) \le P\Big(\delta_0^d \wedge \delta_0^M \ge c_\delta(1 - O(M^{-\frac{1}{2}}))\Big).$$

By (2.18) and d) of Proposition 2.1, we get:

$$\frac{P(\mathcal{C})}{P(\delta_0^d > c_\delta)} = P(E_{\delta,0} \cap E_b | \delta_0^d \ge c_\delta) > 0.5,$$

provided n is big enough. These relations yield:

$$P\Big(\delta_0^d \wedge \delta_0^M \ge c_\delta(1 - O(M^{-\frac{1}{2}}))\Big) \ge (2Ln^{1000})^{-0.5 \ln n} \cdot 0.5 \cdot P(\delta_0^d > c_\delta).$$

The lemma is proved.

Proposition 2.2. *For any $\varsigma \in B(\mathcal{C})$ we have*

$$\delta_0^d(\varsigma) \ge c_\delta[1 - O(M^{-\frac{1}{2}})].$$

Proof. Let $\varsigma \in B(\mathcal{C})$. Choose $\eta \in B^{-1}(\varsigma)$. Let $\{I_k\}$ be the fence-partition corresponding to η. Let $\delta_z^\eta(l)$, $\delta_z^\varsigma(l)$ denote the probabilities defined in (2.7), with ξ replaced by η and ς, respectively. As already noted, because of the fencing-structure, any sequence of consecutive one's or zero's can be generated on the one interval I_k, only. More precisely, if $l \in I_k$, then

$$\delta_0^\eta(l) = P\big(S_l(0), \dots, S_l(n^2) \in I_k\,,\ \eta(S_l(0)) = \dots = \eta(S_l(n^2))\big). \tag{2.21}$$

By the argument of the proof of c of Proposition 2.1, we get that each interval without big blocks, I_k^o, has the property: the probability of generating $n^2 + 1$ consecutive zeros or ones is smaller than $n^{-a \ln n}$. In other words $\delta_0^\eta(l) \le n^{-a \ln n}$, $\quad \forall l \in I^o$, where $I^o := \cup_k I_k^o$. Denote $I^* := \cup_k I_k^*$. Now, by (2.10) and (2.21) we have

$$\begin{aligned}
\delta_0^d(\eta) &= \sum_{l \in I} P(0,l)\delta_0^\eta(l) = \Big(\sum_{l \in I^o} + \sum_{l \in I^*}\Big) P(0,l)\delta_0^\eta(l) \\
&\le \sum_{l \in I^o} P(0,l) n^{-a \ln n} + \sum_{l \in I^*} P(0,l)\delta_0^\eta(l) \\
&\le n^{-a \ln n} + \sum_{l \in I^*} P(0,l)\delta_0^\eta(l) \le n^{-a \ln n} + p_M \sum_{l \in I^*} \delta_0^\eta(l).
\end{aligned}$$

Since $\eta \in \mathcal{C}$, $\delta_0^d(\eta) \ge c_\delta \ge p_M$, we have

$$\sum_{l \in I^*} \delta_0^\eta(l) \ge \frac{c_\delta - n^{-a \ln n}}{p_M} \ge 1 - \frac{n^{-a \ln n}}{p_M} = 1 - O\Big(\frac{\sqrt{M}}{n^{a \ln n}}\Big), \tag{2.22}$$

Clearly $O(\frac{\sqrt{M}}{n^{a \ln n}}) = o(n^{-\alpha})$, for all $\alpha \ge 0$.

Now consider $\varsigma = M(\eta)$. Let $J_1, J_2, \dots J_{N+1}$ denote the new location of the intervals I_i after applying the mapping Π_η to I. Fix an $j \in I$ and let $j \in J_k$. The equation $\varsigma | J_k = \eta | I_k$ and (2.21) imply

$$\begin{aligned}\delta_0^{\varsigma}(j) &= P\big(S_j(0),\dots,S_j(n^2)\in I, \varsigma(S_j(0))=\cdots=\varsigma(S_j(n^2))\big)\\ &\geq P\big(S_j(0),\dots,S_j(n^2)\in J_k, \varsigma(S_j(0))=\cdots\varsigma(S_j(n^2))\big)\\ &= P\big(S_l(0),\dots,S_l(n^2)\in I_k, \eta(S_l(0))=\cdots=\eta(S_l(n^2))\big)=\delta_0^{\eta}(l),\end{aligned}$$

where $l=\Pi(j)\in I_k$. This means $\delta_0^{\varsigma}(j)\geq\delta_0^{\eta}(\Pi_\eta(j)), \quad \forall j\in I$. In particular,

$$\sum_{j\in J_k}\delta_0^{\varsigma}(j)\geq\sum_{l\in I_k}\delta_0^{\eta}(j) \tag{2.23}$$

If $I_1=J_1$ and $I_{N+1}=J_{N+1}$, i.e. the first and last intervals do not contain big blocks, then, obviously, (2.23) is an equation.

Let $J^*=\Pi_\eta(I^*)$, i.e. J^* is the union of all intervals with big blocks in the new location. The length of I^* (and, therefore, that of J^*) is at most $0.5Cn\ln n$. Thus, J^* is at most $Cn+0.5Cn\ln n$ from the origin. Let n be so big, that $Cn+0.5Cn\ln n\leq n^2$. Then, $j\leq n^2$ for each for each $j\in J^*$. Denote by:

$$p_o=\min\{P(S(M)=i):|i|\leq n^2\}.$$

Now from (2.22) and (2.23) we get

$$\begin{aligned}\delta_0^d(\varsigma) &= \sum_j P(0,j)\delta_0^{\varsigma}(l)\geq\sum_{j\in J^*}P(0,j)\delta_0^{\varsigma}(j)\geq\sum_{l\in I^*}P(0,j)\delta_0^{\eta}(l)\\ &\geq p_o\sum_{l\in I^*}\delta_0^{\eta}(l)\geq(c_\delta-n^{-a\ln n})\frac{p_o}{p_M}=c_\delta\Big(1-\frac{p_M-p_o}{p_M}-\frac{n^{-a\ln n}p_o}{c_\delta p_M}\Big)\\ &= c_\delta[1-O(M^{-\frac{1}{2}})]-O\Big(\frac{\sqrt{M}}{n^{a\ln n}}\Big)=c_\delta[1-O(M^{-\frac{1}{2}})].\end{aligned}$$

Proposition 2.3. *For any* $\varsigma\in B(\mathcal{C})$ *we have*

$$\delta_0^M(\varsigma)\geq c_\delta[1-O(M^{-\frac{1}{2}})].$$

Proof. We use the notation and the results of the previous proof. By the representation (2.8) we have

$$\delta_0^M(\varsigma)=\sum_{i,j\in I}\mu(i)P(i,j)\delta_0^{\varsigma}(j)\geq\sum_{i,j\in J^*}\mu(i)P(i,j)\delta_0^{\varsigma}(j) \tag{2.24}$$

where $\mu=\{\mu(i)\}_{i\in I}$ is the stationary measure of $Y_k=S(\nu_0(k))$, $k=1,2,\dots$.

Use local central limit theorem (CLT in the sequel) to estimate

$$\begin{aligned}\min_{i,j\in J^*}P(j,i) &\geq \min\{P(i,j):|i-j|\leq n^2\}\geq\frac{c}{\sqrt{M}}\exp\Big(-\frac{dn^2}{M}\Big)-O(M^{-1})\\ &= \frac{c}{\sqrt{M}}\Big(1-O\Big(\frac{n^2}{M}\Big)\Big)-O(M^{-1})=p_M\Big(1-O\Big(\frac{1}{\sqrt{M}}\Big)\Big).\end{aligned} \tag{2.25}$$

with d,c being constants not depending on n.

Hence, because of (2.24), (2.22) and (2.25)

$$\begin{aligned}\delta_0^M(\varsigma) \geq &\mu(J^*)[p_M\big(1-O(\frac{1}{\sqrt{M}})\big)]\frac{c_\delta - n^{-a\ln n}}{p_M}\\ =&\mu(J^*)\big(1-O(\frac{1}{\sqrt{M}})\big)(c_\delta - n^{-a\ln n}) = \mu(J^*)\big(1-O(\frac{1}{\sqrt{M}})\big)c_\delta.\end{aligned} \tag{2.26}$$

We now estimate $\mu(J^*)$. We shall show that

$$P(Y_{k+1}\in J^*|Y_k=j) \geq 1-o(M^{-1}) \quad \forall j\in I.$$

Then $\mu(J^*)=\sum_j P(Y_{k+1}\in J^*|Y_k=j)\mu(j)\geq 1-o(M^{-1})$ and, by (2.26)

$$\delta_0^M(\varsigma)\geq \mu(J^*)\geq (1-o(M^{-1}))c_\delta[1-O(M^{-\frac{1}{2}})] = c_\delta[1-O(M^{-\frac{1}{2}})].$$

Estimation of $\mu(J^*)$

Fix an $j\in I$ and define ν as the first time after $e^{n^{0.1}}$ when n^2+1 consecutive 0-s or 1-s are generated on I. Formally,

$$\nu := \min\left\{t\geq e^{n^{0.1}}\,\middle|\, \begin{array}{c}\chi\left(t\right)=\chi\left(t-1\right)=...=\chi\left(t-n^2\right)\\ \text{and } S_j(i)\in I, \forall i=t-n^2,\dots,t\end{array}\right\}$$

where $\chi=\varsigma\circ S_j$. Clearly

$$P(S_j(\nu)\in J^*) = P(Y_{k+1}\in J^*|Y_k=j).$$

Thus, it suffices to estimate $P(S_j(\nu)\in J^*)$.

At first note that by (2.22) and (2.23,) we get $\sum_{j\in J^*}\delta_0^\eta(j)\to 1$. Since $|J^*|\leq n^2$ (and n is big enough), we deduce the existence of $j^*\in J^*$ such that

$$\delta_0^\eta(j^*) > \frac{1}{n^3}. \tag{2.27}$$

Then, because of the fences we have:

$$\{S_j(\nu)\notin J^*\} = \{S_j(\nu-n^2),\dots,S_j(\nu)\in I\backslash J^*,\ \ \chi(\nu-n^2)=\dots=\chi(\nu)\}.$$

Now, let τ_k be the k-th visit after time $e^{n^{0.1}}-n^2$ to the interval I. Let τ_k^* be the k-th visit after time $e^{n^{0.1}}-n^2$ to the point j^*. Define the events

$$F_k := \{S_j(\tau_k-n^2),\dots,S_j(\tau_k)\in I\backslash J^*,\ \ \chi(\tau_k-n^2)=\dots=\chi(\tau_k)\},\ \ k=1,2\dots$$

$$F_k' = \cup_{i=0}^{n^{2000}-1}\{S_j(\tau_k+i)=j^*\},\quad k=1,2,\dots$$

$$F_k^* = \{\chi(\tau_k^*)=\dots=\chi(\tau_k^*+n^2)\},\ \ k=1,2,\dots$$

We consider the events

$$E_1 := \{\nu > \tau_{n^{2020}}\} \cup \{S_j(\nu) \in J^*\},\ \ E_2 := \{\tau^*_{n^{10}} \leq \tau_{n^{2020}} - n^2\},\ \ E_3 := \cup_{k=1}^{n^{10}} F^*_k$$

The event E_1 ensures that within the first n^{2020} visits of S_j to I no consecutive 0's or 1's were generated on $I\backslash J^*$. The event E_2 ensures that before time $\tau_{n^{2020}} - n^2$ the random walk visits at least n^{10} times the point j^*. Finally, the event E_3 ensures that during these n^{10} visits of j^*, at least one of them is a beginning of n^2 consecutive 0's or 1's. If these events hold, then $\nu \leq \tau_{n^{2020}}$ and $S_j(\nu) \in J^*$. Thus

$$E_1 \cap E_2 \cap E_3 \subset \{S_j(\nu) \in J^*\}.$$

Next, we give upper bounds for the probabilities $P(E_1), P(E_2), P(E_3)$.

1) Note that: $E_1^c \subset \cup_{k=1}^{n^{2020}} F_k$, implies: $P(E_1^c) \leq \sum_{k=1}^{n^{2020}} P(F_k)$. For each k,

$$\begin{aligned}P(F_k) = \sum_{l\in I\backslash J^*} & P[S_l(0),\dots,S_l(n^2) \in I\backslash J^*, \varsigma(S_l(0)) = \cdots = \varsigma(S_l(n^2))]\times \\ & \times P(S_j(\tau_k - n^2) = l).\end{aligned}$$

There is no big blocks in $I\backslash J^*$, hence by the argument of c:

$$P[S_l(0),\dots,S_l(n^2) \in I\backslash J^*, \varsigma(S_l(0)) = \cdots = \varsigma(S_l(n^2))] \leq n^{-a\ln n},$$

implying that:

$$P(E_1^c) \leq n^{2020 - a\ln n}.$$

2) To estimate $P(E_2)$ we use the Höffding inequality. By central limit theorem there exists a constant $p > 0$ not depending on n such that $P(F'_k) \geq p$. Also note that F'_k and F'_l are independent if $|k-l| \geq n^{2000}$. Hence, the set $\{F'_k\}$, $k = 1,\dots,n^{2020}$ contains a subset $\{F'_{k_i}\}$ $i = 1,\dots n^{20}$ consisting of independent events. Let $X_i := I_{F'_{k_i}}$. Now, $\tau_{n^{2018}} + n^{2000} \leq \tau_{n^{2019}} \leq \tau_{n^{2020}} - n^2$, if n is big enough. This means

$$\Big\{\sum_{i=1}^{n^{18}} X_i \geq n^{10}\Big\} \subset E_2.$$

Now, when n is big enough, we have

$$\begin{aligned}P(E_2^c) &\leq P\Big(\sum_{i=1}^{n^{18}} X_i < n^{10}\Big) = P\Big(\sum_{i=1}^{n^{18}}(X_i - EX_i) < n^{10} - \sum_{i=1}^{n^{18}} EX_i\Big) \\ &\leq P\Big(\sum_{i=1}^{n^{18}}(X_i - EX_i) < -(n^{18}p - n^{10})\Big) \leq P\Big(\sum_{i=1}^{n^{18}}(X_i - EX_i) < -n^{17}\Big) \leq \\ &\leq \exp(-\frac{2n^{34}}{n^{18}}) = \exp(-2n^{16}).\end{aligned}$$

3) Note F^*_l, F^*_k are independent, if $|k-l| > n^2$ Let $\{F^*_{k_i}\}$, $i = 1,2,\dots,n^7$ be a subset of $\{F^*_k\}$ consisting on independent events, only. By (2.27), $P(F^*_k) > \frac{1}{n^3}$, $\forall k$. Now

$$P(E_3^c) \leq P(\cap_{i=1}^{n^7} F^*_{k_i}) = \prod_{i=1}^{n^7}(1 - P(F^*_{k_i})) \leq \Big(1 - \frac{1}{n^3}\Big)^{n^7}. \tag{2.28}$$

The right side of (2.28) is smaller than $(0.5)^{n^4}$ if n is big enough.

Thus,

$$P(S_j(\nu) \in J^*) \geq 1 - [n^{2020-a\ln n} + \exp(-2n^{16}) + (0.5)^{n^4}]$$
$$= 1 - O(n^{-2020+a\ln n}) = 1 - o(M^{-1}).$$

2.4 Corollaries

We determine the critical value c_r. Since we choose it within the interval $[p_M, 2p_M]$, it has all properties stated in Proposition 2.1 and Lemma 2.2. However, we also have to ensure that with high probability the signal probabilities δ_z^d and δ_z^M are significantly away from c_r. By "significantly" we mean that the difference between these probabilities and c_r is bigger than a polynomially small quantity in n. This polynomially small quantity will be denoted by Δ. Thus, c_r must be properly chosen and that will be done with the help of Corollary 2.2.

At first, some preliminary observations.

Proposition 2.4. *For any $j > 2$, there exists an interval $[a, b] \subset [p_M, 2p_M]$ of length $p_M / \left(n^{j+2}\right)$ such that*

$$P(\delta_0^d < b | \delta_0^d \geq a) \leq \frac{1}{n^j} \tag{2.29}$$

Proof. We do the proof by contradiction. Assume on the contrary that there exists no interval $[a, b] \subset [p_M, 2p_M]$ of length $l := p_M / n^{j+2}$ such that (2.29) is satisfied. Let $a_i := p_M + il$, $i = 0, \ldots, n^{j+2}$. Since $[a_i, a_{i+1}] \subset [p_M, 2p_M]$ is an interval of length l, by assumption:

$$P(\delta_0^d \geq a_{i+1} | \delta_0^d \geq p_M + a_i) \leq \left(1 - \frac{1}{n^j}\right), \quad i = 1, \ldots, n^j - 1.$$

Now, by b) of Proposition 2.1:

$$e^{-n} < P(\delta^d \geq 2p_M) = \prod_{i=0}^{n^{j+2}-1} P(\delta_0^d \geq a_{i+1} | \delta_0^d \geq a_i) \leq \left(1 - \frac{1}{n^j}\right)^{n^{j+2}}. \tag{2.30}$$

Since $(1 - \frac{1}{n^j})^{n^j} < e^{-1}$, we have $(1 - \frac{1}{n^j})^{n^{j+2}} < e^{-n^2}$. Thus, (2.30) implies $e^{-n} < e^{-n^2}$ - a contradiction. $\square$

Corollary 2.1. *Let $[x, y] \subset [p_M, 2P_M]$ be an interval of length l. Then there exists an subinterval $[u, v] \subset [x, y]$ of length $\frac{l}{e^{2n}}$ such that*

$$P(\delta_0^d < v | \delta_0^d > u) \leq \frac{1}{e^n}. \tag{2.31}$$

Proof. The proof of the corollary follows the same argument that the proof of Proposition 2.4: (2.31) together with the statement b) of Proposition 2.1 yield the contradiction: $\exp(-n) < P(\delta_0^d \geq 2p_M) \leq P(\delta_0^d \geq v) \leq \left[\left(1 - \frac{1}{e^n}\right)^{e^n}\right]^{e^n} < \exp(-e^n)$. $\square$

The next proposition proves the similar result for $\delta_0^M \wedge \delta_0^d$. Since we do not have the analogue of b) of Proposition 2.1, we use Lemma 2.2, instead.

Proposition 2.5. *Let $[a,b] \subset [p_M, 2p_M]$ be such that $2p_M - b > p_M O(M^{-\frac{1}{2}})$. For any $i > 2$ there exists an interval $[x,y] \subset [a,b]$ with length $(b-a)/n^{i+2}$ such that, for n big enough*

$$P(\delta_0^M < y | \delta_0^M \wedge \delta_0^d > x) \leq P(\delta_0^M \wedge \delta_0^d < y | \delta_0^M \wedge \delta_0^d > x) \leq \frac{1}{n^i}. \tag{2.32}$$

Proof. Suppose that such a (sub)interval does not exists. Then follow the argument of the previous proof to get

$$P\Big(\delta_0^M \wedge \delta_0^d \geq 2p_M(1 - O(M^{-\frac{1}{2}}))\Big) \leq P\big(\delta_0^M \wedge \delta_0^d \geq b\big) \leq \Big(1 - \frac{1}{n^i}\Big)^{n^{i+2}} < \exp(-n^2). \tag{2.33}$$

By Lemma 2.2 and b) of Proposition 2.1

$$P\Big(\delta_0^M \wedge \delta_0^d \geq 2p_M(1 - O(M^{-\frac{1}{2}}))\Big) \geq 0.5(2Ln^{1000})^{-0.5\ln n}\exp(-n). \tag{2.34}$$

For n big enough, the right side of (2.34) is bigger than e^{-2n}. This contradicts (2.33). □

The following corollary specifies c_r and Δ.

Corollary 2.2. *Let $\Delta := (p_M/8)n^{-10054}$, $\tilde{\Delta} = \Delta e^{-2n}$. Then there exists $c_r \in [p_M + \Delta, 2p_M - \Delta]$ such that, for n big enough, simultaneously,*

$$P\left(\delta_0^d \geq c_r - \Delta\right) \leq \exp((\ln n)^3) P\left(\delta_0^d \wedge \delta_0^M \geq c_r - \Delta\right); \tag{2.35}$$

$$P(\delta_0^M < c_r + \Delta | \delta_0^M \wedge \delta_0^d \geq c_r - \Delta) \leq n^{-10000} \tag{2.36}$$

and

$$P(\delta_0^d < c_r - \Delta + \tilde{\Delta} | \delta_0^d \geq c_r - \Delta) \leq \exp(-n). \tag{2.37}$$

Proof. By Proposition 2.4 there exists an interval $[a,b] \subset [p_M, 2p_M]$ of length p_M/n^{52} such that

$$\frac{P(\delta_0^d \geq b)}{P(\delta_0^d \geq a)} = P(\delta_0^d \geq b | \delta_0^d \geq a) > 1 - \frac{1}{n^{50}} > 0.5. \tag{2.38}$$

We now consider the interval $[a, \frac{a+b}{2}]$. Note that:

$$2p_M - \frac{a+b}{2} \geq b - \frac{b+a}{2} = \frac{b-a}{2} = \frac{p_M}{2n^{52}} > p_M O(M^{-\frac{1}{2}}).$$

Now use Proposition 2.5 with $i = 10000$ to find a subset $[x,y] \in [a, \frac{a+b}{2}]$ with length $l := \frac{b-a}{2}n^{-10002} = \frac{p_M}{2}n^{-10054}$ such that (2.32) holds.
Let us now take $z = x + \frac{l}{4}$. By Corollary 2.1, there exists an subinterval $[u, u + \tilde{\Delta}] \in [x, z]$ with length $\frac{l}{4e^{2n}}$ such that

$$P(\delta_0^d < u + \tilde{\Delta} | \delta^d > u) \leq \exp(-n). \tag{2.39}$$

Now take $\Delta := \frac{l}{4} = \big(p_M/8\big)n^{-10054}$, $c_r := u + \Delta$. Since $[c_r - \Delta, c_r + \Delta] \subset [x, y]$, we have that

$$
\begin{aligned}
&P(\delta_0^M < c_r + \Delta | \delta_0^M \wedge \delta_0^d > c_r - \Delta) \leq P(\delta_0^M \wedge \delta_0^d < c_r + \Delta | \delta_0^M \wedge \delta_0^d > c_r - \Delta) \leq \\
&P(\delta_0^M \wedge \delta_0^d < y | \delta_0^M \wedge \delta_0^d > c_r - \Delta) = \frac{P(\Delta - c_r < \delta_0^M \wedge \delta_0^d < y)}{P(\delta_0^M \wedge \delta_0^d > \Delta - c_r)} \leq \\
&\frac{P(y > \delta_0^M \wedge \delta_0^d > x) - P(x \leq \delta_0^M \wedge \delta_0^d \leq c_r - \Delta)}{P(\delta_0^M \wedge \delta_0^d > x) - P(x < \delta_0^M \wedge \delta_0^d \leq c_r - \Delta)} \leq \frac{P(y > \delta_0^M \wedge \delta_0^d > x)}{P(\delta_0^M \wedge \delta_0^d > x)} = \\
&P(\delta_0^M \wedge \delta_0^d < y | \delta_0^M \wedge \delta_0^d > x) \leq \frac{1}{n^{10000}}.
\end{aligned}
$$

Hence, (2.36) holds.
Since $u = c_r - \Delta$, we also have that (2.37) holds.
It only remains to show that the chosen c_r also satisfies (2.35).
Clearly $\Delta > 2p_M O(M^{-\frac{1}{2}}) > c_r O(M^{-\frac{1}{2}})$. That implies:

$$P\Big(\delta_0^d \wedge \delta_0^M \geq c_r(1 - O(M^{-\frac{1}{2}})\Big) \leq P(\delta_0^d \wedge \delta_0^M \geq c_r - \Delta).$$

Combine this with Lemma 2.2 to get

$$P(\delta_0^d \geq c_r) 0.5(2Ln^{1000})^{-0.5 \ln n} \leq P(\delta_0^d \wedge \delta_0^M \geq c_r - \Delta) \tag{2.40}$$

Since $[c_r - \Delta, c_r + \Delta] \subset [a, b]$ we have

$$P(\delta_0^d \geq a) \geq P(\delta_0^d \geq c_r - \Delta) \geq P(\delta_0^d \geq c_r) \geq P(\delta_0^d \geq b).$$

Now, by (2.38)

$$\frac{P(\delta_0^d \geq c_r)}{P(\delta_0^d \geq c_r - \Delta)} \geq \frac{P(\delta_0^d \geq b)}{P(\delta_0^d \geq a)} > 0.5.$$

The last inequality above, together with (2.40) implies

$$P(\delta_0^d \geq c_r - \Delta) \leq 0.25(2Ln^{1000})^{0.5 \ln n} P(\delta_0^d \wedge \delta_0^M \geq c_r - \Delta) \tag{2.41}$$

Now, the relation

$$0.25(2Ln^{1000})^{0.5 \ln n} \leq \exp((\ln n)^3)$$

together with (2.41) establishes (2.35). □

3 Scenery-dependent events

In the present section we define and investigate the signal points and Markov signal points. We show that with high probability the location of the signal points follows certain clustering structure. This structure gives us the desired signal carriers in the 2-color case.

3.1 Signal points

We are now going to define the Markov signal points, strong signal points and signal points – these are the location points, where the corresponding signal probabilities are above the critical value c_r. The Markov signal points form the core of the signal carriers, the (strong) signal points will be used in our proofs. In an oversimplified way, we could say that the Markov signal points are places in the scenery ξ where the conditional probability to see in the observations some rare unusual pattern is above c_r. The unusual pattern is basically a string of n^2, zero's or one's.

In the present subsection, with the help of the signal points, we define many other important notions, and we also investigate their properties.

In the following, Δ and c_r are as in Corollary 2.2. In particular,

$$\Delta = \frac{p_M}{8} n^{-10054}, \quad p_M = \max\{P(S(M) = z) | z \in \mathbb{Z}\}.$$

- A (location) point $z \in \mathbb{Z}$ is called *signal point*, if $\delta_z^d > c_r - \Delta$.
- A (location) point $z \in \mathbb{Z}$ is called *strong signal point*, if $\tilde{\delta}_z^d > c_r - \Delta$.
- A (location) point $z \in \mathbb{Z}$ is called *Markov signal point*, if

$$\delta_z^d > c_r - \Delta \quad \text{and} \quad \delta_z^M > c_r - \Delta.$$

- We call a Markov signal point z *regular*, if $\delta_z^M > c_r + \Delta$.
- Let $\bar{z}_1$ be the first Markov signal point in $[0, \infty)$. Let $\bar{z}_k$ be defined inductively: $\bar{z}_k$ is the first Markov signal point in $[\bar{z}_{k-1} + 2Ln^{1000}, \infty)$. Let $\bar{z}_0$ be the Markov signal point in $(-\infty, 0]$ which laies closest to the origin. Let $\bar{z}_{-k}$ be defined inductively: $\bar{z}_{-k}$ is the rightmost Markov signal point in $(-\infty, \bar{z}_{-(k-1)} - 2Ln^{1000}]$. Thus $\ldots, \bar{z}_{-2}, \bar{z}_{-1}, \bar{z}_0, \bar{z}_1, \bar{z}_2, \ldots$ is a sequence of ordered random variables which we call *signal carrier points*.
- For given z, the set

$$\mathcal{N}_z := [z - L(n^{1000} + e^{n^{0.3}}), z - L(n^{1000})] \cup [(z + Ln^{1000}, z + L(n^{1000} + e^{n^{0.3}})]$$

  is called the *neighborhood* of z. We say that the neighborhood of z is *empty*, if $\mathcal{N}_z$ does not contain any block of ξ longer than $n^{0.35}$. Thus, $\{\mathcal{N}_z \text{ is empty }\} \subset \sigma(\xi_i, i \in \mathcal{N}_z)$.
- We say that z has *empty border*, if the set $I_z - [z - \tilde{M}, z + \tilde{M}]$ does not contain any block of ξ longer than $n^{0.35}$. Thus, $\{\mathcal{N}_z \text{ has empty border }\} \subset \sigma(\xi_i, i \in I_z - [z - \tilde{M}, z + \tilde{M}])$.
- Let p, $\tilde{p}$ and p^d be the probability, that a fixed point is a Markov signal point, a strong signal point or a signal point, respectively. From (2.3), part a) of Proposition 2.1 and by (2.35) of Corollary 2.2 we know

$$p^d - \exp(-dn^{999}) < \tilde{p} \leq p^d; \tag{3.1}$$

$$p \leq p^d \leq \exp(-\frac{\alpha n}{\ln n}); \tag{3.2}$$

$$\frac{p^d}{p} \leq \exp((\ln n)^3). \tag{3.3}$$

- For each $j = 0, 1, 2, \dots, 2Ln^{1000}$ partition the set $\mathbb{Z} \cap [-Ln^{1000} + j, \infty)$ into adjacent integer intervals of length $2Ln^{1000}$. Let $I_{k,j}$ denote the k-th interval of the partition who's first interval starts at $-Ln^{1000} + j$. Thus,

$$I_{1,j} = [j - Ln^{1000}, j + Ln^{1000}], \quad I_{2,j} = [j + Ln^{1000} + 1, j + 3Ln^{1000} + 1],$$

$$I_{3,j} = [j + 3Ln^{1000} + 2, j + 5Ln^{1000} + 2],$$

$$\dots$$

$$I_{k,j} = [j + kLn^{1000} + k - 1, j + (k+2)Ln^{1000} + k - 1].$$

- Let $z_{j,k}$ denote the midpoints of $I_{k,j}$. Hence

$$z_{j,1} = j, \quad z_{j,2} = j + 2Ln^{1000} + 1, \quad \dots, \quad z_{j,k} = j + 2kLn^{1000} + (k-1).$$

For each j, the intervals $I_{k,j}$, $k = 1, 2, \dots$ are disjoint. Thus, the events

$$\{z_{k,j} \quad \text{is a Markov signal point}\}, \quad k = 1, 2, \dots$$

are independent with the same probability p.

- Let k' denote the integer valued random variable that shows the index of the first interval $I_{k,0}$ which has its midpoint being a Markov signal point. By such a counting we disregard the first interval. Thus, $k' > 1$ and, formally, k' is defined by the relations

$$\delta_{z_{2,0}} \wedge \delta^M_{z_{2,0}} \le c_r - \Delta, \ \dots \quad \delta^M_{z_{k'-1,0}} \wedge \delta^d_{z_{k'-1,0}} \le c_r - \Delta, \ \delta^M_{z_{k',0}} \wedge \delta^d_{z_{k',0}} > c_r - \Delta$$

Clearly, $k' - 1$ is a geometrical random variable with parameter p and, hence, $Ek' = \frac{1}{p} + 1$.

- Let Z be the location of the first Markov signal point after $2Ln^{1000}$. Recall $\bar{z}_1$ is the location of the first Markov signal point after 0. Note, that for each $i \ge 0$, we have

$$P(\bar{z}_1 \le i) < P(\cup_{j=0}^{i}\{i \text{ is a Markov signal point}\}) \le pi \tag{3.4}$$

and

$$P(Z \le i) \le p(i - 2Ln^{1000}), \quad i \ge 2Ln^{1000}. \tag{3.5}$$

From (3.4) and (3.2) we get

$$P(\bar{z}_1 \le 2Ln^{1000}) \le p2Ln^{1000} \le 2Ln^{1000}\exp(-\frac{\alpha n}{\ln n}) \to 0. \tag{3.6}$$

- We now estimate EZ. For this note: $Z \le z_{k',0} = 2k'Ln^{1000} + k' - 1$ and

$$EZ \le (\frac{1}{p} + 1)2Ln^{1000} + \frac{1}{p} \le \frac{3}{p}Ln^{1000}. \tag{3.7}$$

From (3.3) we get

$$EZp^d \le 3\frac{p^d}{p}Ln^{1000} \le 3Ln^{1000}\exp((\ln n)^3). \tag{3.8}$$

On the other hand by (3.5) we have, for each x, $EZ \geq xP(Z \geq x) \geq x(1-px)$. Now, take $x = (2p)^{-1}$ and use (3.2) to get

$$EZ \geq \frac{1}{4p} \geq \frac{1}{4}\exp(\frac{\alpha n}{\ln n}). \tag{3.9}$$

- Take $m(n) = \ulcorner n^{2.5}EZ \urcorner$.

  By (3.3) and b) of Proposition 2.1 we get

$$n^{2.5}EZ \leq \frac{3Ln^{1002.5}}{p^d}\exp((\ln n)^3) \leq 3Ln^{1002.5}\exp((\ln n)^3 + n) < \exp(2n),$$

  implying

$$\frac{1}{4}\exp(\frac{\alpha n}{\ln n}) \leq m < \exp(2n), \tag{3.10}$$

  provided n is big enough.

- Next, we define the random variables which we are using later:

$$X_z := I_{\{\delta_z^d > c_r - \Delta, \quad \delta_i^M > c_r - \Delta\}}, \quad z = 0, 1, 2, \ldots.$$

  Thus, X_z indicates, whether z is a Markov signal point or not. The random variables X_z are identically distributed with mean p.

- We estimate the number of Markov signal points in $[0, cm]$, where $c > 1$ is a fixed integer, not depending on n . For this define:

$$E_0 := \Big\{\sum_{z=0}^{cm} X_z \leq n^{10000}\Big\}.$$

  Thus, when E_0 holds, the interval $[0, cm]$ contains at most n^{10000} Markov signal points.

  To estimate $P(E_0)$ we use the Markov inequality and (3.7)

$$P(E_0^c) = P\Big(\sum_{i=0}^{cm} X_i > n^{10000}\Big) < \frac{(cm+1)p}{n^{10000}} \leq \frac{c(n^{2.5}EZ+1)p+1}{n^{10000}}$$
$$< c3Ln^{1002.5-10000} + (c+1)n^{-10000} = o(1).$$

- Finally, define $Z_0 < Z_1 < \cdots < Z_k < \cdots$ as follows:

  $Z_0 := 0$, $Z_1 := Z$, and, let Z_{k+1} be the first Markov signal point that is greater than $2Ln^{1000} + Z_k$. Note the differences: Z, $Z_2 - Z_1$, $Z_3 - Z_2$, …, $Z_{k+1} - Z_k, \ldots$ are i.i.d. Also:

$$\{\text{No Markov signal points in } [0, 2Ln^{1000}]\} = \{Z_i = \bar{z}_i \text{ for all } i\} := E_s^n. \tag{3.11}$$

  From (3.6) we know that

$$P(E_s^n) \to 1. \tag{3.12}$$

3.2 Scenery-dependent events

Next, we describe the typical behavior of the signal points in the interval $[0, cm]$. Here $c > 1$ is a fixed integer, not depending on n. Among others we show that, with high probability, for each signal carrier point $\bar{z}_i$ in $[0, cm]$, the corresponding frequency of ones, $h(\bar{z}_i)$, vary more than $e^{-n^{0.11}}$ (events $\bar{E}_3^n$ and $\bar{E}_4^n$ below). We also show that, with high probability, all signal points in $[0, cm]$ have empty neighborhood.

All the properties listed below depend on the scenery ξ only. Therefore we refer to them as the *scenery dependent events*.

We now define all scenery dependent events, $\bar{E}_1^n, \ldots, \bar{E}_9^n$ and prove the convergence of their probabilities. All the events will be defined on the interval $[0, cm]$, where $c > 1$ is a fixed integer. Thus, if a point z is such that $\mathcal{N}_z \notin [0, cm]$, by the neighborhood of z, we mean $\mathcal{N}_z \cap [0, cm]$. This means $\bar{E}_i^n \in \sigma(\xi_z : z \in [0, cm])$. The exact value of c will be defined in the next chapter (in connection with the event $E_{2,S}^n$). During this chapter, c is assumed to be any fixed integer bigger than 1.

At first, we list the events of interest:

$\bar{E}_1^n := \{\bar{z}_{n^2+1} \leq m\};$

$\bar{E}_2^n := \{\text{every signal point in } [0, cm] \text{ has an empty neighborhood}\};$

$\bar{E}_3^n := \{\text{every pair } \bar{z}_1, \bar{z}' \text{ of signal carrier points in } [0, cm] \text{ satisfies}: \ |h(\bar{z}) - h(\bar{z}')| \geq e^{-n^{0.11}} \text{ if } \bar{z} \neq \bar{z}'\};$

$\bar{E}_4^n := \{\text{every signal carrier point } \bar{z}, \text{ in } [0, cm] \text{ satisfies}: \ |h(\bar{z}) - \frac{1}{2}| \geq e^{-n^{0.11}}\};$

$\bar{E}_5^n := \{\text{every signal point } z \in [0, cm] \text{ satisfies } \delta_z^M \notin [c_r - \Delta, c_r + \Delta]\};$

$\bar{E}_6^n := \{\text{ for all signal carrier points } \bar{z}_i \text{ in } [0, cm] \text{ we have } EZn^{11001} \geq |\bar{z}_i - \bar{z}_{i+1}| \geq EZn^{-11001}\};$

$\bar{E}_7^n := \{\text{no signal carrier points in } [m - EZn^{-11001}, m + EZn^{-11001} \wedge cm] \cup [0, EZn^{-11001}]\};$

$\bar{E}_8^n := \{\text{every strong signal point in } [0, cm] \text{ has empty border}\};$

$\bar{E}_9^n := \{\text{every signal point in } [0, cm] \text{ is a strong signal point}\}.$

Proof that $P(\bar{E}_1^n) \to 1$

If $\bar{E}_1^n$ holds, then in $[0, m]$ we have more than n^2 signal carrier points .

Define the random variables $Z_0 < Z_1 < \cdots < Z_k < \cdots$ as in (3.11). Let $E_{1a}^n := \{Z_{n^2+1} \leq m\}$. Since $E_s \cap E_{1a}^n \subset \bar{E}_1^n$, it suffices to show that $P(E_{1a}^n) \to 1$. To see this, we use the Markov inequality:

$$P(E_{1a}^{nc}) = P(Z_{n^2+1} > m) \leq \frac{EZ_{n^2+1}}{m} \leq \frac{(n^2+1)}{n^{2.5}} \to 0.$$

Proof that $P(\bar{E}_2^n) \to 1$

$$\bar{E}_2^{nc} = \{\,\text{there exists a signal point in } [0, cm] \text{ with non} - \text{empty neighborhood}\}.$$

Clearly,

$$\bar{E}_2^{nc} = \cup_{z=0}^{cm} E_2(z), \quad \text{where} \quad E_2(z) := \{z \text{ is a signal point and } \mathcal{N}_z \text{ is not empty}\}.$$

For each z, the events $\{\mathcal{N}_z \text{ is empty}\}$ and $\{\delta_z > c_r - \Delta\}$ are independent. Thus, for each z,

$$P(E_2(z)) = P(\delta_z > c_r - \Delta)P(\mathcal{N}_z \text{ is empty}) = p^d P(\mathcal{N}_z \text{ is not empty}).$$

We obviously have $P(\mathcal{N}_z \text{ is empty}) = P(\mathcal{N}_o \text{ is empty})$ and

$$\begin{aligned}&P(\mathcal{N}_o \text{ is not empty}) = \\ &P(\mathcal{N}_o \text{ contains at least one block longer than } n^{0.3}) < 2L\exp(n^{0.3})2^{-n^{0.35}}.\end{aligned}$$

Hence, from (3.8):

$$\begin{aligned}P(\bar{E}_2^{nc}) &\le cmp^d 2L\exp(n^{0.3})(\frac{1}{2})^{n^{0.35}} \le 6cn^{2.5}L^2 n^{1000}\exp((\ln n)^3 + n^{0.3})2^{-n^{0.35}} \\ &= 6cL^2 n^{1002.5}\exp(n^{0.3} + (\ln n)^3)2^{-n^{0.35}} \to 0,\end{aligned}$$

if $n \to \infty$.

Proof that $P(\bar{E}_8^n) \to 1$

For each z, the events $\{\delta_z^d > c_r - \Delta\}$ and $\{z$ has empty border $\}$ are independent. Now use the same argument as in the previous proof.

Proof that $P(\bar{E}_5^n) \to 1$

Note

$$\bar{E}_5^{nc} = \{\text{there exists a non} - \text{regular Markov signal point } z \in [0, cm]\}.$$

As in the previous proof, write:

$$\bar{E}_5^n = \cup_{z=0}^{cm} E_5(z), \quad \text{where} \quad E_5(z) := \{z \text{ is a non} - \text{regular Markov signal point}\}.$$

For each z,

$$\begin{aligned}P(E_5^c(z)) &= P(\delta_z^M \wedge \delta_z^d > c_r - \Delta)P(\delta_z^M \le c_r + \Delta | \delta_z^M \wedge \delta_z^d > c_r - \Delta) \\ &= pP(\delta_z^M \le c_r + \Delta | \delta_z^M \wedge \delta_z^d > c_r - \Delta).\end{aligned}$$

From (2.36) of Corollary 2.2 we have:

$$P(\delta_z^M \le c_r + \Delta | \delta_z^M \wedge \delta_z^d > c_r - \Delta) \le n^{-10^5}.$$

Thus, from (3.7) $P(\bar{E}_5^{nc}) \le cmpn^{-10^5} \le c(n^{2.5}EZ + 1)pn^{-10^5} = c3Ln^{1002.5-100000} + cpn^{-10^5} \to 0$, as $n \to \infty$.

Proof that $P(\bar{E}_9^n) \to 1$

We use the same argument as in the previous proof. Note

$$\bar{E}_9^{nc} = \{\text{there exists a signal point } z \in [0, cm] \text{ that is not a strong signal point}\}.$$

As in the previous proof, write

$$\bar{E}_9^{nc} = \cup_{z=0}^{cm} E_9(z), \quad \text{where} \quad E_9(z) := \{z \text{ is a non} - \text{strong signal point}\}.$$

Recall (2.3): $\tilde{\delta}_z^d > \delta_z^d - \exp(-dn^{999})$. Since, for n big enough, $\exp(-dn^{999}) < \tilde{\Delta} = \Delta \exp(-2n)$, we get

$$\tilde{\delta}_z^d > \delta_z^d - \tilde{\Delta}.$$

Now, for each z,

$$\begin{aligned} P(E_9(z)) &= P(\delta_z^d > c_r - \Delta) P(\tilde{\delta}_z^d \leq c_r - \Delta | \delta_z^d > c_r - \Delta) \\ &= p^d P(\tilde{\delta}_z^d \leq c_r - \Delta | \delta_z^d > c_r - \Delta) \leq p^d P(\delta_z^d - \tilde{\Delta} \leq c_r - \Delta | \delta_z^d > c_r - \Delta) \\ &\leq p^d P(\delta_z^d \leq c_r - \Delta + \tilde{\Delta} | \delta_z^d > c_r - \Delta). \end{aligned}$$

By (2.37) of Corollary 2.2 we now have:

$$P(E_9(z)) \leq p^d \exp(-n).$$

Hence, by (3.8):

$$P(\bar{E}_9^{nc}) \leq cmp^d \exp(-n) \leq p^d c(EZn^{2.5}+1) \exp(-n) \leq c3Ln^{1000} \exp{(\ln n)^3} \exp(-n) + o(1) = o(1).$$

Proof that $P(\bar{E}_6^n) \to 1$

Consider random variables $Z_0 < Z_1 < \cdots < Z_k < \cdots$ as in (3.11). Let $N = \max\{i : Z_i \leq cm\}$. Define

$$E_{6b}^n := \{Z_i - Z_{i-1} \leq EZn^{10001}, \quad i = 1, 2, \ldots, n^{1000}\} \tag{3.13}$$

$$\bar{E}_{6c}^n := \{Z_i - Z_{i-1} \geq EZn^{-11001}, \quad i = 1, 2, \ldots, n^{1000}\} \tag{3.14}$$

and note that:

$$E_s \cap E_{6b}^n \cap E_{6a}^n \cap \{N \leq n^{10000}\} \subset \bar{E}_6^n.$$

Since $E \subset \{N \leq n^{10000}\}$, we get $P(N \leq n^{10000}) \to 1$. We also know that $P(E_s) \to 1$. Thus, it suffices to show that $P(E_{6b}^{nc}), P(E_{6c}^{nc}) \to 0$ as $n \to \infty$. Now, by the Markov inequality, (3.5) and

(3.7):

$P(E_{6b}^{nc}) = P(\exists 1 \leq i \leq n^{10000} \text{ such that : } Z_i - Z_{i-1} > EZn^{10001})$

$\leq \sum_{i=1}^{n^{10000}} P(Z_i - Z_{i-1} > EZn^{10001}) = n^{10000} P(Z > EZn^{10001}) \leq$

$n^{10000} \frac{EZ}{EZn^{10001}} = \frac{1}{n};$

$P(E_{6c}^{nc}) = P(\exists\ 1 \leq i \leq n^{10000} \text{ such that : } Z_i - Z_{i-1} < EZn^{-11001})$

$\leq \sum_{i=1}^{n^{10000}} P(Z_i - Z_{i-1} < EZn^{-11001}) \leq n^{10000} P(Z < EZn^{-11001}) <$

$pEZn^{-1001} \leq 3Ln^{1000-1001} = \frac{3L}{n}.$

Proof that $P(\bar{E}_7^n) \to 1$

Consider the event

$$\{\text{there is no signal carrier points in } [0, EZn^{11001}]\}.$$

Every signal carrier point is a Markov signal point. Hence, for the proof, it suffices to show, that with high probability there is no Markov signal points in the interval $[0, EZn^{11001}]$.

Now, by (3.4) and (3.7)

$$P(\text{No Markov signal points in } [0, EZn^{11001}]) =$$

$$P(Z^o > EZn^{-11001}) \leq pEZn^{-11001} \leq 3Ln^{-11001+1000} = o(1).$$

Thus $P(\text{No Markov signal points in } [0, EZn^{-11001}]) \to 1$.

Now repeat the same argument for the intervals $[m, m - EZn^{-11001}]$ and $[m, m + EZn^{-11001}]$.

3.3 Proof of $P(\bar{E}_3^n) \to 1$ and $P(\bar{E}_4^n) \to 1$

The proof relies on the rate of convergence in the local central limit theorem (LCLT in sequel). In the next subsection we present some technical preliminaries related to the proof.

3.3.1 Some preliminaries

Let S be the symmetric random walk with span 1. Define: $p_N(k) = P(S(N) = k)$. The random walk S has lattice $+\backslash - z$,$z \in Z$; its variance is σ^2.

Use local CLT ([Pet95], page 197):

$$\sup_k \left| \sigma\sqrt{N} p_N(k) - \frac{1}{\sqrt{2\pi}} \exp\{-\frac{k^2}{2\sigma^2 N}\} \right| = O(\frac{1}{\sqrt{N}}) \tag{3.15}$$

or

$$\sup_k \left| p_N(k) - \frac{1}{\sigma\sqrt{N}\sqrt{2\pi}} \exp\{-\frac{k^2}{2\sigma^2 N}\} \right| = O(\frac{1}{N}).$$

Denote

$$q_N(k) := \frac{1}{\sigma\sqrt{N}\sqrt{2\pi}} \exp\{-\frac{k^2}{2\sigma^2 N}\} \quad |k| \leq LN.$$

Let $t_N := (\ln N)^b$, $b > 1$.

We estimate:

$$\begin{aligned} |p_N^2(k) - q_N^2(k)| &\leq (p_N(k) + q_N(k)) \sup_k |p_N(k) - q_N(k)| \\ &\leq [2q_N(k) + O(\frac{1}{\sqrt{N}})] O(\frac{1}{N}) = O(\frac{1}{\sqrt{N}N}) \end{aligned}$$

and

$$\sum_{k>t_N+j}^{L\sqrt{N}} [p_N^2(k) - q_N^2(k)] \leq (L\sqrt{N}) O(\frac{1}{\sqrt{N}N}) = O(\frac{1}{N}), \quad j = -t_N, \cdots, t_N.$$

Estimate:

$$\begin{aligned} &\frac{p_N^2(k)}{\sum_{k>t_N+j} p_N^2(k)} \leq \frac{p_N^2(k)}{\sum_{k>t_N+j}^{L\sqrt{N}} p_N^2(k)} \leq \frac{q_N^2(k) + O(\frac{1}{N})}{\sum_{k>t_N+j}^{L\sqrt{N}} [p_N^2(k) - q_N^2(k)] + \sum_{k>t_N+j}^{L\sqrt{N}} q_N^2(k)} \\ &\leq \frac{O(\frac{1}{N})}{\sum_{k>t_N+j}^{L\sqrt{N}} q_N^2(k) - O(\frac{1}{N})}, \end{aligned}$$

for all k and $j = -t_N \ldots, t_N$.

Now,

$$\sum_{k>t_N+j}^{L\sqrt{N}} q_N^2(k) = \frac{1}{2\sigma^2 \pi N} \sum_{k>t_N+j}^{L\sqrt{N}} \exp(-\frac{k^2}{\sigma^2 N})$$

and

$$\sum_{k>t_N+j}^{L\sqrt{N}} \exp(-\frac{k^2}{\sigma^2 N}) \geq \sum_{k>2t_N}^{L\sqrt{N}} \exp(-\frac{k^2}{\sigma^2 N}) > \sum_{k>2t_N}^{L\sqrt{N}} \exp(-\frac{L^2}{\sigma^2}) = M(L\sqrt{N} - 2t_N).$$

Thus, for each $j = -t_N, \ldots, t_N$,

$$\sup_k \frac{p_N^2(k)}{\sum_{k>t_N+j} p_N^2(k)} \leq \frac{O(\frac{1}{N})}{\frac{K}{N}(L\sqrt{N} - 2t_N) - O(\frac{1}{N})} = l \frac{K_4}{K_1\sqrt{N} - K_2 t_N - K_3} = O(\frac{1}{\sqrt{N}}) \quad (3.16)$$

where K, K_1, K_2, K_3, K_4 are constants.

Let μ be a probability distribution on $\{-t_N, -t_N + 1, \ldots, 0, \ldots, t_N - 1, t_N\}$. Consider the convolutions

$$u_N(k) = \sum_{j=-t_N}^{t_N} p_N(k-j)\mu_j, \quad k = -(LN - t_N), \ldots, LN + t_N. \quad (3.17)$$

If $p_N(k) \geq p_N(k+1)$ for all $k \geq 0$, then for each $k > t_N$, we have the bounds

$$p_N(k+t_N) \leq u_N(k) \leq p_N(k-t_N). \tag{3.18}$$

In this case,

$$\sum_{k>t_N}^{t_N+LN} u_N(k) \geq \sum_{l>2t_N}^{N} p_N(l).$$

And from (3.16), taking $j = t_N$ we may deduce that:

$$\sup_{t_N<k} \frac{u_N^2(k)}{\sum_{k>t_N} u_N^2(k)} \leq \sup_{0<k} \frac{p_N^2(k)}{\sum_{k>2t_N} p_N^2(k)} \leq O(\frac{1}{\sqrt{N}}). \tag{3.19}$$

Generally, choose an atom $\lambda := \mu_j > 0$. Then

$$u_N(k) \geq \lambda p_N(k+j), \quad u_N^2(k) \geq \lambda^2 p_N^2(k+j)$$

and

$$\sum_{k>t_N}^{t_N+LN} u_N^2(k) \geq \lambda^2 \sum_{k>t_N+j}^{N} p_N^2(k). \tag{3.20}$$

Since $\sup_{k>t_N} u_N^2(k) \leq \sup_{k>0} p_N^2(k)$, we get from (3.16):

$$\sup_{t_N\leq k} \frac{u_N^2(k)}{\sum_{k>t_N} u_N^2(k)} \leq \sup_k \frac{p_N^2(k)}{\lambda^2 \sum_{k>t_N+j} p_N^2(k)} = O(\frac{1}{N^{\frac{1}{4}}}). \tag{3.21}$$

In particular, from (3.21) follows:

$$\frac{\sum u_N^3(k)}{\sum u_N^2(k)\sqrt{\sum u_N^2(k)}} \leq \max_k u_N(k) \frac{\sum u_N^2(k)}{\sum u_N^2(k)\sqrt{\sum u_N^2(k)}} \leq \max_k \frac{u_N(k)}{\sqrt{\sum u_N^2(k)}} \leq O\Big(\frac{1}{N^{\frac{1}{4}}}\Big). \tag{3.22}$$

Suppose that arrays $u_k := u_N(k)$ and $v_k := v_N(k)$, $t_N < k \leq LN + t_N$ both satisfy (3.22). Then

$$\frac{\sum(u_k^3+v_k^3)}{\sum(u_k^2+v_k^2)\sqrt{\sum(u_k^2+v_k^2)}} \leq \max\{u_k, v_k\} \frac{\sum(u_k^2+v_k^2)}{\sum(u_k^2+v_k^2)\sqrt{\sum(u_k^2+v_k^2)}} \tag{3.23}$$

$$\leq \max\{\max_k \frac{u_k}{\sqrt{\sum u_k^2}}, \max_k \frac{v_k}{\sqrt{\sum v_k^2}}\} = O(N^{-\frac{1}{4}}) \tag{3.24}$$

Let us make one more observation. Since $\exp(\frac{-9t_N^2}{2\sigma^2 N}) \to 1$, there exists a $c' > 0$ such that

$$\exp(\frac{-9t_N^2}{2\sigma^2 N}) > c'$$

for each N big enough. Thus, there exists a constant $c > 0$ such that

$$p_N(k) > \frac{c}{\sqrt{N}}, \quad \forall |k| \leq 3t_N.$$

Take λ as previously. Then

$$u_N(k) \geq p(k+j)\lambda \geq \frac{c\lambda}{\sqrt{N}}.$$

Hence there exists $C > 0$: $u(l) \geq \frac{C}{\sqrt{N}}$ $\forall l$ such that $|l+j| \leq 3t_N$.

In particular

$$u_N(k) \geq \frac{C}{\sqrt{N}}, \quad -2t_N \geq k \leq 2t_N. \tag{3.25}$$

3.3.2 Proof that $P(\bar{E}_3^n) \to 1$

Define the random variables z_1, z_2, ... as follows: z_1 is the first Markov signal point in $[0, \infty)$, z_k is the first Markov signal point in $[z_{k-1} + e^{n^{0.3}}, \infty)$. Note that a.s. there are infinitely many such points.

From the signal carrier part we know that, if each Markov signal point in $[0, cm]$ has empty neighborhood, i.e. $\bar{E}_2^n$ holds, then they form clusters which have radius at most $2Ln^{1000}$ and lie at least $e^{n^{0.3}}$ apart from each other. In this case all signal carrier points in $[0, cm]$ coincide with the z_i's defined above. We define the event:

$$E_{3a}^n := \Big\{\text{for each} \quad i, j \leq n^{10000}, i \neq j \text{ we have } |h(z_i) - h(z_j)| \geq \exp(-n^{0.11})\Big\}.$$

Then:

$$E_{3a}^n \cap \bar{E}_2^n \cap E_0 \subset \bar{E}_3^n.$$

Since $P(E_{3a}^n \cap E_0) \to 1$, it suffices to show that $P(E_{3a}^n) \to 1$ as $n \to \infty$.

Let z_i, z_j, $i \neq j$. For simplicity denote them as z and z' Let

$$\epsilon_n := \exp(-n^{0.11}).$$

Consider the event:

$$E_n(i,j) := \{|h(z) - h(z')| \geq \epsilon_n\}.$$

For each $y \in Z$, define the random vector:

$$\xi_n(y) := \Big(\xi(y - Ln^{1000} - e^{n^{0.1}}),\, \xi(y - Ln^{1000} - e^{n^{0.1}} + 1), \ldots, \xi(y + Ln^{1000})\Big).$$

Now, let $\xi_n := \xi_n(z)$ and $\xi'_n := \xi_n(z')$. They are independent.

$$f_n := \sum_{k=z+Ln^{1000}+1}^{z+L(n^{1000}+e^{n^{0.1}})} u_n(k)\xi(k), \quad f'_n := \sum_{k=z'+Ln^{1000}+1}^{z'+L(n^{1000}+e^{n^{0.1}})} u'_n(k)\xi(k),$$

where

$$u_n(k) := \sum_{i=z-Ln^{1000}}^{z+Ln^{1000}} P(S_i(e^{n^{0.1}}) = k)\mu_i, \;\; u'_n(k) := \sum_{i=z'-Ln^{1000}}^{z'+Ln^{1000}} P(S_i(e^{n^{0.1}}) = k)\mu'_i$$

and μ_i, $i = z - Ln^{1000}, \cdots, z + Ln^{1000}$ and μ'_i, $i = z' - Ln^{1000}, \cdots, z' + Ln^{1000}$ denote the atoms of the stationary measure corresponding to z and z', respectively.

Recall that by (2.13)

$$h(z) := \sum_{k=z-L(n^{1000}+e^{n^{0.1}})}^{z+L(n^{1000}+e^{n^{0.1}})} u_n(k)\xi(k), \quad f'_n := \sum_{k=z'-L(n^{1000}+e^{n^{0.1}})}^{z'+L(n^{1000}+e^{n^{0.1}})} u'_n(k)\xi(k).$$

Note that conditioning on ξ_n, the coefficients $u_n(k)$ become constants.
(More precisely, f_n has the same distribution as

$$\tilde{f}_n := \sum_{k>Ln^{1000}}^{L(n^{1000}+e^{n^{0.1}})} \tilde{u}_n(k)\xi(k),$$

with

$$\tilde{u}_n(k) := \sum_{j=-Ln^{1000}}^{Ln^{1000}} P(S_j(e^{n^{0.1}})=k)\tilde{\mu}_j = \sum_{j=-Ln^{1000}}^{Ln^{1000}} P(S(e^{n^{0.1}})=k-j)\tilde{\mu}_j,$$

with $\tilde{\mu} := \{\tilde{\mu}_j\} := \{\mu_{z+j}\}$, $-Ln^{1000} \le j \le Ln^{1000}$ being a random probability measure independent of $\xi_{Ln^{1000}+1}, \dots \xi_{e^{n^{0.1}}}$. In this setup, conditioning on ξ_n means conditioning on $\tilde{\mu}$.)
Hence

$$P\Big(\frac{f_n - Ef_n}{\sqrt{Df_n}} \le x|\xi_n\Big) = P\Big(\frac{\sum_{k>Ln^{1000}}^{L(e^{n^{0.1}}+N^{1000})} u_n(k)(\xi_(k)-\frac{1}{2})}{\frac{1}{2}\sqrt{\sum_{k>Ln^{1000}}^{L(e^{n^{0.1}}+N^{1000})} u_n^2(k)}} \le x|\xi_n\Big),$$

where $(u_n(k))$ are the fixed coefficients of type (3.17) (with $N = e^{n^{0.1}}$, $b = 10000$). Now the Berry-Esseen inequality for independent random variables (see, [Pet95], Thm 3, p.111) states:

$$\sup_x \Big|P\Big(\frac{\sum u_n(k)(\xi(k)-\frac{1}{2})}{\frac{1}{2}\sqrt{\sum u_n^2(k)}} \le x|\xi_n\Big) - \Phi(x)\Big| \le A\frac{\sum u_n^3(k)}{\sum u_n^2(k)\sqrt{\sum u_n^2(k)}}, \tag{3.26}$$

with some constant A not depending on n and $u_n(k)$-s. By (3.22) (with $N = e^{n^{0.1}}$, $b = 10000$), the right side of (3.26) is bounded by $O(e^{\frac{-n^{0.1}}{4}})$. Here Φ stands for the standard normal distribution function.

By similar argument, conditioning on (ξ_n, ξ'_n) and using (3.23) instead of (3.22) yields:

$$\sup_x \Big|P\big(\frac{f_n - f'_n - \mu_n}{\sigma_n} \le x|\xi_n, \xi'_n\big) - \Phi(x)\Big| = O(e^{\frac{-n^{0.1}}{4}}), \tag{3.27}$$

with $\mu_n := E(f_n - f'_n)$, $\sigma_n := \sqrt{Df_n + Df'_n}$ where (f_n and f'_n are independent.)
Let $g_n := h_n - f_n$, $g'_n := h'_n - f'_n$. The event $E_n(i,j)$ can be written as:

$$E_n^c(i,j) := \{f_n - f'_n \in g_n - g'_n + [-\epsilon_n, \epsilon_n]\}.$$

Given ξ_n and ξ'_n, the random variable $g_n - g'_n$ is a constant. By (3.27) we have

$$P(E_n^c(i,j)|\xi_n, \xi'_b) = P\Big(\tfrac{f'_n - f_n - \mu_n}{\sigma_n} \in \tfrac{g_n - g'_n + [-\epsilon_n, \epsilon_n] - \mu_n}{\sigma_n}|\xi_n, \xi'_n\Big) \le$$

$$2\sup_x \Big|P\big(\tfrac{f'_n - f_n - \mu_n}{\sigma_n} \le x|\xi_n, \xi'_n\big) - \Phi(x)\Big| + \sup\Big\{\Phi(a) - \Phi(b)\big|a - b = \tfrac{2\epsilon_n}{\sqrt{2\pi}\sigma_n}\Big\} \le$$

$$O(e^{\frac{-n^{0.1}}{4}}) + \sqrt{\tfrac{2}{\pi}}\tfrac{\epsilon_n}{\sigma_n}.$$

Next, we estimate the standard deviation σ_n. For that note: because of (3.25) $u_n^2(z+Ln^{1000}+1)\geq C^2e^{-n^{0.1}}$, $u'^2_n(z'+Ln^{1000}+1)\geq C^2e^{-n^{0.1}}$ if n is big enough. Thus,

$$\sigma_n=\sqrt{Df_n+Df'_n}=\frac{1}{2}\sqrt{\sum u_N^2(k)+\sum u'^2_N(k)}>\frac{1}{2}\sqrt{2C^2e^{n^{0.1}}}=\sqrt{2}Ce^{\frac{-n^{0.1}}{2}}.$$

Hence, for n big enough there exists a constant $C_2<\infty$ such that

$$\sqrt{\frac{2}{\pi}}\frac{\epsilon_n}{\sigma_n}\leq\frac{1}{\sqrt{\pi}}\exp(-n^{0.11}+\frac{n^{0.1}}{2})\leq C_2\exp(-n^{0.05}). \tag{3.28}$$

Thus, (3.28), (3.27) give:

$$P(E^n(i,j))\leq O(e^{\frac{-n^{0.11}}{4}})+O(e^{-n^{0.05}})=O(e^{-n^{0.05}}).$$

By definition

$$E_{3a}^n=\cap_{i,j,i\neq j}^{n^{10000}}E^n(i,j)$$

and

$$P(E_{3a}^{nc})\leq\sum_{i,j,i\neq j}^{n^{10000}}P(E^{nc}(i,j))<n^{20000}O(e^{-n^{0.05}})=o(1).$$

Outline of the proof that $P(\bar{E}_4^n)$ is close to one

Denote the Use (3.26) to get:

$$\begin{aligned}
P(\bar{E}_4^{nc}|\xi_n)&=P(|f_n+g_n-0.5|\leq\epsilon_n|\xi_n)=P(f_n+g_n\in[0.5-\epsilon_n,0.5+\epsilon_n]|\xi_n)\\
&=P(f_n\in[(0.5-g_n)-\epsilon_n,0.5-g_n+\epsilon_n]|\xi_n)\\
&=P\Big(\frac{f_n-Ef_n}{\sqrt{Df_n}}\in\Big[\frac{0.5-Ef_n-g_n-\epsilon_n}{\sqrt{Df_n}},\frac{0.5-Ef_n-g_n+\epsilon_n}{\sqrt{Df_n}}\Big]|\xi_n\Big)\\
&\leq 2\sup_x P\Big(\frac{f_n-Ef_n}{\sqrt{Df_n}}\leq x|\xi_n\Big)+\sup\Big\{\Phi(a)-\Phi(b)\Big|a-b=\sqrt{\frac{2}{\pi}}\frac{\epsilon_n}{\sqrt{Df_n}}\Big\}\\
&\leq O(e^{\frac{-n^{0.1}}{4}})+\sqrt{\frac{2}{\pi}}\frac{\epsilon_n}{\sqrt{Df_n}}=O(e^{-n^{0.05}}),
\end{aligned}$$

because $\sqrt{Df_n}>C\exp(-\frac{n^{0.1}}{2})$. The rest of the proof goes as previously.

- In the following we consider the scenery dependent events defined on $[-cm,cm]$. For this, we define the events $\tilde{E}_i^n$, $i=1,\ldots,9$, where $\tilde{E}_i^n$ is defined exactly as $\bar{E}_i^n$, with $[-cm,0]$ instead of $[0,cm]$.

- Finally, we define the events:

  $$E_i^n:=\tilde{E}_i^n\cap\bar{E}_i^n.$$

  The results of the present section show that $\forall\ i=1,\ldots,9$,

  $$P(E_i^n)\to 0,\quad n\to\infty.$$

3.4 What is a signal carrier?

Let us briefly summarize the main ideas of the previous sections.

A signal carrier is a place in the scenery, where the probability to generate a block of n^2+1 times the same color is high. However, it is clear that such a place can not be too small. In the 3-color example the signal carrier depends on only one bit of the scenery. In the 2-color case, it takes many more bits to make the scenery (locally) atypical. We saw in Proposition 2.1, that for z to be a signal point, it is necessary that the interval I_z contains at least one big (longer than $n/\ln n$) block of ξ. Thus, if a point z is a (Markov, strong) signal point or not, depends on $\xi|I_z$.

If z is a signal point, then the scenery ξ is atypical in the interval I_z: δ_z^d is high. Thus, signal points would be our candidates for the signal carriers, if, for each z, we could estimate δ_z^d. The latter would be easy, if we knew when the random walk visits z. Then just take all such visits and consider the proportion of those visits that were followed by n^2+1 same colors after M steps. Unfortunately, we do not know when the random walk S visits z. But we do know (we observe) when S generates blocks with length at least n^2. Thus we can take these observations (times) as the visits of (the neighborhood of) z and estimate the probability of generating n^2+1 times the same color, M steps after previously observing n^2+1 times the same color. This idea yields the Markov signal probability. The problem now is to localize the area where the random walk (during a given time period) can generate n^2+1 times the same colors in the observations. If this area was too big, we could neither estimate the Markov signal probability nor understand where we are. To localize the described area, we showed (event E_2^n) that signal points have empty neighborhood. In the next section we shall see that the probability to generate a block of n^2+1 times the same color on the empty neighborhood is very small. This means, if S is close to a signal point z, then, with high probability, (and during a certain time period) all strings of n^2+1 times the same colors in the observations are be generated on I_z. The fact that all signal points have also empty borders (events E_8^n and E_9^n) makes the latter statement precise. Thus, a Markov signal point seems to be a reasonable signal carrier. But which one? Note, if z is a Markov signal point, i.e. I_z contains at least one big block, then, very likely the point $z+1$ is a Markov signal point, too. In other words, Markov signal points come in clusters. However, when E_2^n holds, then each point in such a cluster has empty neighborhood. On the other hand, for z to be a Markov signal point, it is necessary to have at least one big block of ξ in I_z. This means that the diameter of every cluster of Markov signal points is at most $2Ln^{1000}$. The distances between the clusters are at least $L(e^{n^{0.3}}-n^{1000})$. Hence, in 2-color case one can think of signal carriers as clusters of Markov signal points (provided E_2^n holds, but this holds with high probability). However, to make some statements more formal, for each cluster we have one representator, namely the signal carrier point. Since the diameters of the clusters are at most $2Ln^{1000}$, our definition of signal carrier points ensures that different signal carrier points belong to different clusters. If the cluster is located in $[0,\infty)$, then the signal carrier point is the left most Markov signal point in the cluster; if the cluster is located in $(-\infty,0)$, then the signal carrier point is the right most Markov signal point in the cluster. The event E_7^n ensures that there are no Markov signal points in the $2Ln^{1000}$-neighborhood of 0, so $\bar{z}_1$ and $\bar{z}_0$ belong to different clusters, too. The exact choice of a signal carrier point is irrelevant. However, it is important to note that given a cluster, everything that makes this cluster a signal carrier cluster (namely, the big blocks of scenery) is inside the interval $I_{\bar{z}}$, where $\bar{z}$ is the signal carrier point corresponding to the cluster. In particular, all blocks in the observations that are longer than

n^2 will be generated on $I_{\bar{z}}$. This means that the signal carrier points, $\bar{z}_i$ (or the corresponding intervals $I_{\bar{z}_i}$) serve as signal carriers as well. At least, if we are able to estimate $\delta^M_{\bar{z}_i}$ with great precision. This is the subject of the next section.

4 Events depending on random walk

In the previous section we saw: if all scenery dependent events hold, then the signal carrier points are good candidates for the signal carriers. In this case the signal is an untypically high Markov signal probability. Hence, to observe this signal, we must be able to estimate the Markov signal probability. In the present section we define these estimators and in the next section we will see that they perform well, if the random walk S behaves typically. We describe the typical behavior of S in terms of several events depending on S. The main objective of the present section is to show that the (conditional) probability of such events tends to 1 as n tends to infinity.

4.1 Some preliminaries

As argued in Subsection 3.4, the main idea of the estimation of the Markov signal probability is very simple - given a time interval T, consider all blocks in the observations $\chi|_T$ that are bigger than n^2. Among these observations calculate the proportions of such blocks, that after exactly M steps, are followed by another such block. The time interval used for this estimation must be big enough to get a precise estimate but, on the other hand, it must be in correspondence with the size of an (empty) neighborhood. Recall that the neighborhood $\mathcal{N}_z$ consisted of two intervals of length $Le^{n^{0.3}}$. Hence, the optimal size of the interval T is $e^{n^{0.3}}$.

We now define the necessary concepts related to the described estimate - stopping times (that stop when a string of at least n^2+1 times the same color is observed) and the Bernoulli variables that show whether the stopping times are followed (after M step) by another such string or not. For technical reasons after stopping the process, we wait at least $e^{n^{0.1}}$ steps until we look for the next block.

- Let $t > 0$ and let $\hat{\nu}_t(1)$ be the smallest $s \geq t$ such that:

$$\chi(t) = \chi(t-1) = \cdots = \chi(t-n^2). \tag{4.1}$$

  We define the stopping times $\hat{\nu}_t(i)$, $i = 2, 3, \dots$ inductively: $\hat{\nu}_t(i)$ is the smallest $t \geq \hat{\nu}_t(i-1) + e^{n^{0.1}}$ such that (4.1) holds.

- Let $X_{t,i}$ be the Bernoulli random variable that is one if and only if:

$$\chi(\hat{\nu}_t(i) + M) = \chi(\hat{\nu}_t(i) + M + 1) = ... = \chi(\hat{\nu}_t(i) + M + n^2).$$

  Let $T = T(t) := [t, t + e^{n^{0.3}}]$. Define:

$$\hat{\delta}^M_T = \begin{cases} \frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} X_{t,i} & \text{if } \hat{\nu}_t(e^{n^{0.2}}) < t + e^{n^{0.3}} - e^{n^{0.1}} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

- We define some analogues of $\hat{\nu}_t$ and X_t. Let $z \in \mathbb{Z}$ and $t \in \mathbb{N}$. Let $\nu_{z,t}(1)$ designate the first time after t where we observe n^2 zero's or one's in a row, generated on the interval I_z. More precisely:

$$\nu_{z,t}(1) := \min \left\{ s > 0 \;\middle|\; \begin{array}{c} \chi(s) = \chi(s-1) = \cdots = \chi(s-n^2) \\ S(j) \in I_z, \ \forall j = s-n^2, \ldots, s \end{array} \right\}.$$

Once $\nu_{z,t}(i)$ is well defined, define $\nu_{z,t}(i+1)$ in the following manner:

$$\nu_{z,t}(i+1) := \min \left\{ t \geq \nu_{z,t}(i) + e^{n^{0.1}} \;\middle|\; \begin{array}{c} \chi(s) = \chi(s-1) = \ldots = \chi(s-n^2) \\ S(j) \in I_z, \ \forall j = s-n^2, \ldots, s \end{array} \right\}.$$

- Let $X_{z,t,i}$, $i = 1, 2, \ldots$ designate the Bernoulli variable which is equal to one if exactly after time M the stopping time $\nu_{z,t}(i)$ is followed by a sequence of n^2+1 one's or zero's generated on I_z. More precisely, $X_{z,t,i} = 1$ if and only if

$$\chi(\nu_{z,t}(i) + M) = \chi(\nu_{z,t}(i) + M + 1) = \cdots = \chi(\nu_{z,t}(i) + n^2) \quad \text{and}$$
$$S(\nu_{z,t}(i) + M), \ldots, S(\nu_{z,t}(i) + n^{1000}) \in I_z.$$

Define

$$\hat{\delta}^M_{z,t} := \frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} X_{z,t,i}.$$

As argued in Subsection 2.1, $\{S(\nu_{z,t,i})\}$ is an ergodic Markov process with state space I_z and with the stationary measure I_z. Hence,

$$\frac{1}{j} \sum_{i=1}^{j} X_{z,t,i} \to \delta^M_z, \quad \text{a.s.}$$

Now we can apply some large deviation inequality to see that if $j \geq \exp(n^{0.2})$, then $\hat{\delta}^M_{z,t}$ gives a very precise estimate of δ^M_z.

The problem is that the random variables $X_{z,t,i}$ and, hence, the estimate $\hat{\delta}^M_{z,t}$ is *a priori* not observable. This is because we cannot observe whether a string of n^2+1 times the same color in the observations is generated on I_z or not. Thus, we can not observe neither $\nu_{t,z}(i)$ nor $X_{t,z,i}$. However, the event $E^n_{3,S}$, stated below, ensures that with high probability $\hat{\delta}^M_{z,t}$ is the same as $\hat{\delta}^M_T$, provided that during the time interval T, the random walk S is close to z (the sense of closeness will specified later).

- We define the estimates for the frequency of ones. Again, we define a general, observable, estimate: $\hat{h}_t$ and its theoretical, *a priori* not-observable counterpart: $\hat{h}_{z,t}$.

Define

$$\hat{h}_t := \begin{cases} \frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} \chi(\nu_t(i) + e^{n^{0.1}}) & \text{if}, \hat{\nu}_t(e^{n^{0.2}}) < t + e^{n^{0.3}} - e^{n^{0.1}} \\ 0 & \text{otherwise.} \end{cases},$$

$$\hat{h}_{z,t} := \frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} \chi(\nu_{z,t}(i) + e^{n^{0.1}}).$$

- Finally, we define the stopping time that stop the walk, when a new signal carrier is visited. Let $\ldots, \bar{z}_{-1}, \bar{z}_0, \bar{z}_1, \ldots$ denote the signal carrier-points in $\mathbb{R}$. Denote $I_i := I_{z_i}$ and let $\rho(k)$ denote the time of the k-th visit of S to one of the intervals I_i in the following manner: when an interval I_i is visited, then the next stop is on a different interval.

  More precisely, let $\rho(0)$ be the first time $t \geq 0$ such that $S(t) \in \cup_i I_i$. Denote $I(\rho(k))$ the interval I_i visited at time $\rho(k)$. Then define $\rho(k)$ inductively:

$$\rho(k+1) = \min\{t > \rho(k) | S(t) \in \cup_i I_i, \quad S(t) \notin I(\rho(k))\}.$$

4.2 Random walk-dependent events

In this section, we define the events that characterize the typical behavior of the random walk S on the typical scenery on the interval $[-cm, cm]$. The (piece of) scenery $\xi|[-cm, cm]$ is typical if it satisfies all the scenery-dependent events E_i^n, $i = 1, \ldots, 9$. Recall, that the events E_i^n are the same as the events $\bar{E}_i^n$ defined in Section 4.2 with $[0, cm]$ replaced by $[-cm, cm]$. Also recall that $c > 1$ is an arbitrary fixed constant not depending on n, and $m = \ulcorner n^{2.5} EZ \urcorner$. Hence, throughout the section we consider the sceneries belonging to the set:

$$E_{\text{cell\_OK}} := \cap_{i=1}^{9} E_i^n. \tag{4.3}$$

Clearly, $E_{\text{cell\_OK}}$ depends on n. We know that $P(E_{\text{cell\_OK}}) \to 1$ if $n \to \infty$.

Let $\psi : \mathbb{Z} \to \{0, 1\}$ be a (non random) scenery. Let $P_\psi(\cdot)$ designate the measure obtained by conditioning on $\{\xi = \psi\}$ and $\{S(m^2) = m\}$. Thus,

$$P_\psi(\cdot) := P(\cdot | \xi = \psi, S(m^2) = m). \tag{4.4}$$

Let $P(\cdot|\psi)$ denote $P(\cdot|\xi = \psi)$. We now list the events that describe the typical behavior of S. The objective of the section is to show: if n is big and $\psi_n :=: \psi \in E_{\text{cell\_OK}}$ then all listed events have big conditional probabilities P_ψ. The events depending on the random walk are:

$E^n_{1,\mathrm{S}} := \left\{S(m^2) = m\right\};$

$E^n_{2,S} := \left\{\forall t \in [0, m^2] \text{ we have that } S(t) \in [-cm, cm]\right\};$

$E^n_{3,S} := \{\forall t \in [0, m^2], \text{ it holds}: \ \hat{\delta}^M_T \leq c_r, \text{ if } \delta^d_{S(s)} \leq c_r - \Delta \ \forall s \in T(t)\};$

$E^n_{4,S} := \{\rho(n^{25000}) \geq m^2\};$

$E^n_{5,S} := \{\forall k \leq n^{25000} \ \text{ we have: if } \rho(k) \leq m^2 \text{ then } \hat{\nu}_{\rho(k)}(e^{n^{0.2}}) \leq \rho(k) + e^{n^{0.3}} - e^{n^{0.1}}\};$

$$E^n_{6,S} := \left\{\begin{array}{c} \text{for any } t \in [0, m^2] \text{ satisfying } \chi(t) = \cdots = \chi(t + n^2) \\ \text{there exists } s \in [t, t + n^2] \text{ such that } S(s) \\ \text{is contained in a block of } \xi \text{ bigger than } n^{0.35} \end{array}\right\};$$

$E^n_{7,S}(z, t) := \left\{\left|\hat{\delta}^M_{z,t} - \delta^M_z\right| < e^{-n^{0.12}}\right\}, \quad z \in \mathbb{Z}, \ t > 0;$

$E^n_{7,S} := \cap^{cm}_{z=-cm} \cap^{m^2}_{t=0} E^n_{13,S}(z, t);$

$E^n_{8,S}(z, t) := \left\{\left|\hat{h}_{z,t} - h(z)\right| < e^{-n^{0.12}}\right\}, \quad z \in \mathbb{Z}, \ t > 0;$

$E^n_{8,S} := \cap^{cm}_{z=-cm} \cap^{m^2}_{t=0} E^n_{8,S}(z, t);$

We now estimate the conditional probabilities of all listed events. In most cases we prove statements like $P_\psi(E^n_{j,S}) \to 1$. This means: for an arbitrary sequence $\psi_n \in E^n_{\text{cell\_OK}}$, we have:

$$\lim_{n\to\infty} P(E^n_{j,S}|S(m^2) = m, \xi = \psi_n) = 1.$$

4.3 Proofs

At first note that by LCLT, we have:

$$P(E_{1,S}) = \frac{1}{m} + O(\frac{1}{m^2}).$$

Clearly, $E_{1,S}$ does not depend on ξ, i.e. $P(E_{1,S}|\psi) = P(E_{1,S})$. Using (3.10) we get:

$$P(E_{1,S}|\psi) \geq \exp(-2n) - O(\exp(-4n)) \geq \exp(-3n). \tag{4.5}$$

From (4.5) follows that for any event E,

$$P_\psi(E) = \frac{P(E, S(m^2) = m|\psi)}{P(S(m^2) = m|\psi)} \leq \frac{P(E|\psi)}{\exp(-3n)}. \tag{4.6}$$

Proposition 4.1. *For each $\epsilon > 0$ there exists $c(\epsilon)$, independent of n, such that for each ψ, $P_\psi(E^n_{2,S}) \geq 1 - \epsilon$, provided n is big enough.*

Proof. At first note, that, for each n, the event $E^n_{2,S}$ is independent of the scenery ψ. Thus,

$$P_\psi(E^n_{2,S}) = P(E^n_{2,S}|S(m^2) = m).$$

Define

$$E^n_a(c) = \{\forall t \in [0, m^2] \text{ we have that } S(t) \leq cm\}$$

$$E^n_b(c) = \left\{\forall t \in [0, m^2] \text{ we have that } S(t) \geq -cm\right\}$$

Clearly,

$$E^n_{2,S} = E^n_a(c) \cap E^n_b(c).$$

We now find c, not depending on n such that $P_\psi(E^{nc}_a(c)), P_\psi(E^{nc}_b(c)) \leq \frac{\epsilon}{2}$.
Let us define the stopping time ϑ:

$$\vartheta := \min\{t|S(t) > cm\}.$$

Let for all $j \in 1, \ldots, L$:

$$p_j := P\Big(S(m^2) = m,\, \vartheta \leq m^2 \text{ and } S(\vartheta) = cm + j\Big)$$

We have that

$$P\left(E^{nc}_a(c) \cap E^n_{1,S}\right) = \sum_{j=1}^{L} p_j.$$

Our random walk S is symmetric. By the reflection principle, for all $j \in 1, \ldots, L$, we have:

$$p_j = P(S(m^2) = cm + j + (cm + j - m) = 2cm + 2j - m,\, \vartheta \leq m^2 \text{ and } S(\vartheta) = cm + j).$$

Thus $p_j \leq P\left(S\left(m^2\right) = 2cm - m + 2j\ \right)$ and

$$P\left(E^{nc}_a(c) \cap E^n_{1,S}\right) \leq \sum_{j=1}^{L} P\left(S(m^2) = m(2c - 1) + 2j\right). \tag{4.7}$$

By LCLT, for big m, the right side of (4.7) can be made arbitrary small in comparison to $P\left(S\left(m^2\right) = m\right)$ by taking c big enough. In other words, there exists c , not depending on n such that:

$$\frac{\sum_{j=1}^{L} P\left(S\left(m^2\right) = 2cm + m + 2j\ \right)}{P\left(S\left(m^2\right) = m\right)} \leq \frac{\epsilon}{2}.$$

This means

$$\frac{P\left(E^{nc}_a(c) \cap E^n_{1,S}\right)}{P\left(E^n_{1,S}\right)} = P_\psi\left(E^{nc}_a(c)\right) \leq \frac{\epsilon}{2}.$$

Similar argument gives $P_\psi\left(E^{nc}_b(c)\right) \leq \frac{\epsilon}{2}$. ◻

Note, that the choice of c does not depend on n. From now on, we fix c such that Proposition 4.1 holds with $\epsilon > \frac{1}{8}$. This particular c is used in the definition of all scenery-dependent events and, therefore, in the definition of $E_{\text{cell\_OK}}$ as well as in the definitions $E^n_{4,S}$, $E^n_{5,S}$.

In what follows, we often use these versions of the Hoeffding inequality:
Let $X_1, \ldots, X_N$ be independent random variables with range in $[a,b]$. Let S_N denote their sum. Then:

$$\begin{aligned} P(|S_N - ES_N| \geq \epsilon) &\leq 2\exp(-2\frac{\epsilon^2}{N(b-a)^2}) \leq \exp(-\frac{d'\epsilon^2}{N}); \\ P(\frac{1}{N}|S_N - ES_N| \geq \epsilon) &\leq 2\exp(-2\frac{\epsilon^2 N}{(b-a)^2}) \leq \exp(-d'\epsilon^2 N). \end{aligned} \tag{4.8}$$

For our random walk, this gives:

$$\begin{aligned} P(|S(N)| \geq \epsilon) &\leq 2\exp(-\frac{\epsilon^2}{4L^2 N}) \leq \exp(-\frac{d\epsilon^2}{N}) \\ P(|\frac{S(N)}{N}| \geq \epsilon) &\leq 2\exp(-\frac{\epsilon^2 N}{4L^2}) \leq \exp(-d\epsilon^2 N), \end{aligned} \tag{4.9}$$

for some $d', d > 0$.

We also use the following results: let $X_1, \ldots, X_N$ be i.i.d. random variables with mean 0 and finite variance σ^2. Let $M_n^+ = \max_{i=1,\ldots,N} S_i$, $M_n = \max_{i=1,\ldots,N} |S_i|$. Then

$$\frac{M_N^+}{\sigma\sqrt{N}} \Rightarrow \sup_{0\leq t\leq 1} W_t, \quad \text{and} \quad \Big(\frac{M_N}{\sigma\sqrt{N}}, \frac{S(N)}{\sigma\sqrt{N}}\Big) \Rightarrow (\sup_{0\leq t\leq 1} |W_t|, W(1)), \tag{4.10}$$

where W_t is standard Brownian motion. It is well-known that $\forall x > 0$, $P(\sup_{0\leq t\leq 1} W_t \leq x) = 2\Phi(x) - 1$.

Proof that $\liminf_n P_\psi(E^n_{4,S}) \geq 1 - \frac{1}{8}$.

For each n, fix an arbitrary $\psi_n \in E^n_{\text{cell\_OK}}$. Since $\psi_n \in E^n_{\text{cell\_OK}} \subset E^n_6$, we have that for every signal carrier point $\bar{z}_i \in [-cm, cm]$:

$$\bar{z}_{i+1} - \bar{z}_i, \bar{z}_i - \bar{z}_{i-1} \geq EZn^{-11001}. \tag{4.11}$$

For this proof, let $\mu := EZ$ and $N(n) := \mu^2 n^{-24000}$. Since $m \leq n^{2.5}\mu + 1$, we have $n^{25000} \times N = n^{25000} \times \mu^2 n^{-24000} = \mu^2 n^{1000} > m^2$. Hence, if $E^n_{4,S}$ fails, then there must be at least one $k \in \{0, \ldots, n^{25000} - 1\}$ such that $\rho(k+1) - \rho(k) < N$. Moreover, if $E^n_{4,S}$ fails, then for each $k \in \{0, \ldots, n^{25000} - 1\}$ it holds $\rho(k) \leq m^2$. We formalize this observation. Let:

$$\begin{aligned} E_{a,4}(k) &:= \{\rho(k+1) - \rho(k) \geq N, \quad \rho(k) \leq m^2\} \\ E_{a,4} &:= \cap_{k=0}^{n^{25000-1}} E_{a,4}(k). \end{aligned} \tag{4.12}$$

It holds

$$E^{nc}_{4,S} \subset E^c_{a,4}. \tag{4.13}$$

By Proposition 4.1, for n big enough, $P_\psi(E^n_{2,S}) \le \frac{1}{8}$. Thus,

$$P_\psi(E^{nc}_{4,S}) \le P_\psi(E^c_{4,a} \cap E^n_{2,S}) + P_\psi(E^{nc}_{2,S}) \le \frac{1}{8} + \sum_{k=0}^{n^{25000-1}} P_\psi(E^c_{a,4}(k) \cap E^n_{2,S}). \tag{4.14}$$

We now bound $P_\psi(E_{a,4}(k))$.

Suppose $E^n_{2,S}$ holds. Then $\rho(k) \le m^2$ implies that the signal carrier visited at time $\rho(k)$ is in $[-cm, cm]$. By (4.11) this means that the closest signal carrier point is at least at distance μn^{-11001}. Let I_i be $I(\rho(k))$. Then

$$\inf\{|t-s| : t \in I_i, s \in I_j\} \ge \mu n^{-11001} - 2Ln^{1000}, \tag{4.15}$$

where $j \in \{i-1, i+1\}$. By (3.9), $\mu^2 > n^{25000}$. Then $\mu > n^{12500} \ge 2Ln^{12002}$, implying

$$\mu n^{-11001} - 2Ln^{1000} \ge \mu n^{-11002}. \tag{4.16}$$

We consider the event

$$E_{a,4}(k)^c \cap E^n_{2,S} \subseteq \{\rho(k+1) - \rho(k) < N, \quad S(\rho(k)) \in [-cm, cm]\}.$$

From (4.15) and (4.16) it follows that:

$$P\Big(\rho(k+1) - \rho(k) < N, \quad S(\rho(k)) \in [-cm, cm]\Big|\psi_n\Big) \le P(\rho(k+1) - \rho(k) < N\Big|S(\rho(k)) \in [-cm, cm], \xi = \psi_n) \le$$
$$P\Big(\sup_{l \le N} |S(l)| > \mu n^{-11001} - 2Ln^{1000}\Big) \le P(\sup_{l \le N} |S(l)| > \mu n^{-11002}\}.$$

Use the following maximal inequality:

$$P(\max_{l \le N} |S(l)| > \mu n^{-11002}) \le 3 \max_{l \le N} P\Big(|S(l)| > \frac{\mu}{3} n^{-11002}\Big). \tag{4.17}$$

By the Hoeffding inequality, for each $l \le N$:

$$P\Big(|S(l)| > \frac{\mu}{3} n^{-11002}\Big) \le \exp(-\frac{d\mu^2 n^{-22004}}{9l}) \le \exp\big(-\frac{d\mu^2 n^{-22004}}{9N}\big)$$
$$\le \exp\big(-\frac{dn^{24000-22004}}{9}\big) = \exp\big(-\frac{dn^{1996}}{9}\big).$$

Hence,

$$P(E_{a,4}(k)|\psi) \le \exp\big(-\frac{dn^{1996}}{9}\big), \; P(E_{a,4}|\psi) \le n^{25000} \exp\big(-\frac{dn^{1996}}{9}\big).$$

By (4.6), we get

$$P_\psi(E^{nc}_{a,4}) \le n^{25000} \exp\big(3n - \frac{dn^{2996}}{9}\big).$$

The right side of the last inequality tends to 0 if $n \to \infty$. Relation (4.13) now finish the proof.

Proof that $P_\psi(E^n_{3,S}) \to 1$

Let $t \geq 0$ be an integer and define the stopping times $\hat{\nu}_t^o(1), \hat{\nu}_t^o(2), \ldots$ as follows: $\hat{\nu}_t^o(1)$ is the smallest time $s \geq t + e^{n^{0.1}}$ such that:

$$\chi(s-n^2) = \chi(s-n^2+1) = \cdots = \chi(s) \text{ and } \delta^d_{S(s)} \leq c_r - \Delta. \tag{4.18}$$

Once $\hat{\nu}_t^o(k)$ is well defined, define $\hat{\nu}_t^o(k+1)$ to be the smallest time $s \geq \hat{\nu}_t^o(k) + e^{n^{0.1}}$ such that (4.18) holds.

Let $X^o_{t,k}$ be the Bernoulli variable which is equal to one if and only if

$$\chi(\hat{\nu}_t^o(k) + M) = \chi(\hat{\nu}_t^o(k) + M + 1) = \cdots = \chi(\hat{\nu}_t^o(k) + M + n^2).$$

Finally define:

$$\hat{\delta}^M_{o,t} := \frac{1}{e^{n^{0.2}}} \sum_{k=1}^{e^{n^{0.2}}} X^o_{t,k}.$$

Let

$$E^n_{3,S}(t) := \left\{ \hat{\delta}^M_{o,t} < cr \right\}.$$

Clearly,

$$\bigcap_{t \in 0,\ldots,m^2} E^n_{3,S}(t) \subseteq E^n_{3,S}, \quad \text{imlpying} \quad P(E^{nc}_{3,S}|\psi) \leq \sum_{t=0}^{m^2} P(E^{nc}_{3,S}(t)|\psi), \tag{4.19}$$

where ψ is an arbitrary fixed scenery.

Note, for any fixed scenery ψ, the random variables $X^o_{t,1}, X^o_{t,2}, \ldots$ are clearly independent (but not necessarily identically distributed). However, for each i, $E(X^o_{t,i}|\psi) \leq c_r - \Delta$, implying that

$$c_r - \frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} E(X^o_{t,i}|\psi) \geq \Delta.$$

Recall $\Delta = \frac{p_M}{n^{10054}}$. We know that $\Delta \geq n^{-\beta}$, where β is an integer bigger than 11000. Thus, by (4.8)

$$\begin{aligned} P(E^{nc}_{3,S}(t)|\psi) = P(\hat{\delta}^M_{o,t} \geq c_r|\psi) &= P\Big(\frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} X^o_{t,i} \geq c_r|\psi\Big) \\ &\leq P\Big(\frac{1}{e^{n^{0.2}}} \sum_{i=1}^{e^{n^{0.2}}} (X^o_{t,i} - EX^o_{t,k}) \geq \Delta|\psi\Big) \leq \exp(-d'\Delta^2 e^{n^{0.2}}) \\ &\leq \exp\big(-(d' n^{-2\beta} e^{n^{0.2}})\big). \end{aligned}$$

Now, use (4.6), (4.19) and (3.10) to get

$$P_\psi(E^{nc}_{3,S}) \leq m^2 \exp(-d' n^{-2\beta} e^{n^{0.2}} + 3n) \leq \exp\big(7n - (d' n^{-2\beta} e^{n^{0.2}})\big) \to 0,$$

as $n \to \infty$.

Proof that $P_\psi(E^n_{6,S}) \to 1$

Let:

$$E^n_{6,S}(t) = \left\{ \begin{array}{c} \text{if } \chi(t) = \chi(t+1) = \cdots = \chi(t+n^2) \\ \text{then } \exists s \in [t, t+n^2] \text{ such that} \\ S(s) \text{ is contained in a block of } \xi \ \text{ longer than } n^{0.35} \end{array} \right\}.$$

We have that

$$E^n_{6,S} = \bigcap_{t \in [0,m^2]} E^n_{6,S}(t)$$

and thus

$$P_\psi(E^{nc}_{6,S}) \leq \sum_{t=0}^{m^2} P_\psi(E^{nc}_{6,S}(t)).$$

Note:

$$E^{nc}_{6,S}(t) = \left\{ \begin{array}{c} \forall s \in [t, t+n^2] \ \text{ the random walk } S(s) \\ \text{is contained in a block of } \xi \text{ with length at most } n^{0.35} \\ \text{and } \chi(t) = \chi(t+1) = \cdots = \chi(t+n^2) \end{array} \right\}.$$

Fix a scenery ψ. Let $I = \mathbb{Z}/\cup B(\psi)$, where $B(\psi_n)$ is a block of ψ bigger than $n^{0.35}$ and the union is taken over all such blocks. Note $I = \cup_k I_k$, where I_k are disjoint intervals, at least $n^{0.35}$ far from each other. Thus, if $S(t) \in I_k$, then $S(t+s) \notin I_l$ for each $l \neq k$ and for each $s = 1, \ldots, n^2$. Hence

$$P(E^{nc}_{6,S}(t)|\psi) = \sum_{j \in I} P\Big(S(t), \ldots, S(t+n^2) \in I \text{ and } \chi(t) = \cdots = \chi(t+n^2) | S(t) = j\Big) P(S(t) = j)$$

$$\sum_k \sum_{j \in I_k} P\Big(S_j(0), \ldots, S(n^2) \in I_k \text{ and } \chi(t) = \cdots = \chi(t+n^2)\Big) P(S(t) = j).$$

By Lemma 2.1, there exists a constant $a > 0$ not depending on n such that, for each j,

$$P\Big(S_j(0), \ldots, S(n^2) \in I_k \text{ and } \chi(t) = \cdots = \chi(t+n^2)\Big) \leq \exp\Big(-\frac{an^2}{n^{0.7}}\Big). \tag{4.20}$$

Then,

$$P(E^{nc}_{6,S}(t)|\psi) \leq \exp(-an^{1.3}).$$

Thus, by (4.6):

$$P_\psi(E^{nc}_{6,S}(t)) \leq \exp\left(-an^{1.3} + 3n\right) \to 0$$

and by (3.10)

$$m^2 \exp(-an^{1.2} + 3n) \leq e^{7n - an^{1.3}} \to 0.$$

Proof that $P_\psi(E^n_{7,S}) \to 1$

Preliminaries

Recall that the definitions of stopping times involved:

a) $\vartheta_z(k)$, $k=0,1,\ldots$ stands for consecutive visits of S to the point $z-2Le^{n^{0.1}}$, provided that between $\vartheta_z(k)$ and $\vartheta_z(k+1)$ at least once n^2+1 same colors have been generated on I_z;

b) $\nu_z(1)$ ($\nu_z(i)$, $i=2,3,\ldots$) is the first time after $\vartheta_z(0)$, (after $\nu_z(k-1)+e^{n^{0.1}}$) observing n^2+1 times the same color generated on I_z.

In Section 2.1 the stopping times $\vartheta_z(k)$, $\nu_z(i)$ as well as the random variables $X_{z,i}$ were used to define the random variables $\kappa_z(k)$, $\mathcal{X}_z(k)$ and $\mathcal{Z}_z(k)$. The latter were used to define δ_z^M. Then we fix an arbitrary time t and define the counterparts of all the above-mentioned stopping times and random variables starting from t. In Section 4.1 we already defined the t-counterpart of $\nu_z(i)$ and $X_{z,i}$, namely $\nu_{z,t}(i)$, and $X_{z,t,i}$, $i=1,2,\ldots$. Recall that in the definition of $\nu_{z,t}(1)$, the starting point $\vartheta_z(0)$ was replaced by t, the induction for $\nu_{z,t}(i)$ is the same as the one for $\nu_z(i)$, $i=2,3,\ldots$. The Bernoulli random variables $X_{z,t,i}$ were defined exactly as $X_{z,i}$ with the stopping times $\nu_{z,t}(i)$ instead of the $\nu_z(i)$'s.

We define the t-counterpart of $\vartheta_z(k)$, $k=0,1,\ldots$.

- Let $\vartheta_{z,t}(0)=t$ and let

  $$\vartheta_{z,t}(k):=\{\min s>\vartheta_{z,t}(k-1):S(s)=z-2Le^{n^{0.1}},\ \exists j:s>\nu_{z,t}(j)>\vartheta_{z,t}(k-1)\}.$$

  We use $\vartheta_{z,t}(k)$ to define the t-analogues of κ_z, $\mathcal{Z}_z$ and $\mathcal{X}_z$.

- More precisely, let $\kappa_{z,t}(0)=0$ and let $\kappa_{z,t}(k)$ be defined by the inequalities

  $$\nu_{z,t}(\kappa_{z,t}(k))<\vartheta_{z,t}(k)<\nu_{z,t}(\kappa_{z,t}(k)+1).$$

  The definition of $\mathcal{Z}_{z,t}$ and $\mathcal{X}_{z,t}$ is straightforward:

  $$\mathcal{X}_{z,t}(k)=\sum_{i=\kappa_{z,t}(k-1)+1}^{\kappa_{z,t}(k)}X_{z,t,i},\quad \mathcal{Z}_{z,t}(k)=\kappa_{z,t}(k)-\kappa_{z,t}(k-1),\quad k=1,2,\ldots$$

Note that, if ξ is fixed, then, for all $t>0$, the random variables $\mathcal{X}_{z,t}(1),\mathcal{X}_{z,t}(2),\ldots$ are independent and the random variables $\mathcal{X}_{z,t}(2),\mathcal{X}_{z,t}(3),\ldots$ are i.i.d. with the same distribution as $\mathcal{X}_z(k)$. The same holds for $\mathcal{Z}_{z,t}(1)$, $\mathcal{Z}_{z,t}(2),\ldots$. Also note, that $\mathcal{Z}_{z,t}(k)\geq 1$, $k=1,2,\ldots$.

Hence, for all $t>0$,

$$\delta_z^M=\delta_z^M(\xi)=\frac{E(\mathcal{X}_{z,t}(2)|\xi)}{E(\mathcal{Z}_{z,t}(2)|\xi)}=\lim_{k\to\infty}\frac{\sum_{i=1}^k\mathcal{X}_{z,t}(i)}{\sum_{i=1}^k\mathcal{Z}_{z,t}(i)}.$$

We are now going to show that for each ξ, t, z, the first $e^{n^{0.2}}$ observations of $X_{z,t,i}$ are enough to estimate $\delta_z^M(\xi)$ very precisely, i.e. $\hat\delta_{z,t}^M$ is close to δ_z^M.

Fix z,t,ψ and define:

$$\mathcal{Z}_k:=\mathcal{Z}_{z,t}(k),\quad \mathcal{X}_k:=\mathcal{X}_{z,t}(k),\quad X_i:=X_{k,t,i},\quad E\mathcal{X}=E(\mathcal{X}_2|\psi),\quad E\mathcal{Z}=E(\mathcal{Z}_2|\psi),\quad P(\cdot)=P(\cdot|\psi).$$

Thus:

$$\delta_z^M = \delta_z^M(\psi) = \frac{E\mathcal{X}}{E\mathcal{Z}}.$$

Let $a = \ulcorner e^{3n^{0.1}} \urcorner$ and define:

$$\mathcal{Z}_k^a = \mathcal{Z}_k \wedge a, \quad \mathcal{X}_k^a = \mathcal{X}_k \wedge a, \quad E\mathcal{X}^a := E(\mathcal{X}_2^a|\psi), \quad E\mathcal{Z}^a := E(\mathcal{Z}_2^a|\psi).$$

Finally, define:

$$\Delta := e^{-\frac{n^{0.2}}{4}}.$$

We consider the events:

$$\begin{aligned}
E_{7,a} &= \Big\{ \mathcal{Z}_k \leq a, \quad k = 1, 2, \ldots, e^{n^{0.2}} \Big\}, \\
E_{7,b} &= \Big\{ \Big| \frac{\mathcal{X}_1^a + \cdots + \mathcal{X}_k^a}{k} - E\mathcal{X}^a \Big| \leq \frac{\Delta}{3}, \quad \forall k \in [\frac{e^{n^{0.2}}}{a}, e^{n^{0.2}}] \Big\} \text{ and} \\
E_{7,c} &= \Big\{ \Big| \frac{\mathcal{Z}_1^a + \cdots + \mathcal{Z}_k^a}{k} - E\mathcal{Z}^a \Big| \leq \frac{\Delta}{3}, \quad \forall k \in [\frac{e^{n^{0.2}}}{a}, e^{n^{0.2}}] \Big\}.
\end{aligned}$$

First step

First we show that:

$$E_{7,a} \cap E_{7,b} \cap E_{7,c} \subset E_{7S}^n(z,t). \tag{4.21}$$

Let $\bar{\imath}$ be (random) number defined by the inequalities:

$$\mathcal{Z}_1 + \cdots + \mathcal{Z}_{\bar{\imath}} \leq e^{n^{0.2}} < \mathcal{Z}_1 + \cdots + \mathcal{Z}_{\bar{\imath}+1}. \tag{4.22}$$

Since $\mathcal{Z}_k \geq 1$, we have $\bar{\imath} \leq e^{n^{0.1}}$. Let $\bar{k} := \mathcal{Z}_1 + \cdots + \mathcal{Z}_{\bar{\imath}}$. Now,

$$\hat{\delta}_{z,t}^M = \frac{\sum_{i=1}^{e^{n^{0.2}}} X_i}{e^{n^{0.2}}} = \frac{\sum_{k=1}^{\bar{\imath}} \mathcal{X}_k + \sum_{i=\bar{k}+1}^{e^{n^{0.2}}} X_i}{\bar{k} + e^{n^{0.2}} - \bar{k}} = \frac{\frac{1}{\bar{\imath}} \sum_{k=1}^{\bar{\imath}} \mathcal{X}_k + \frac{1}{\bar{\imath}} \sum_{i=\bar{k}+1}^{e^{n^{0.2}}} X_i}{\frac{\bar{k}}{\bar{\imath}} + \frac{e^{n^{0.2}} - \bar{k}}{\bar{\imath}}}.$$

Denote

$$\Delta_I := E(\mathcal{X}^a - \mathcal{X}) + \frac{1}{\bar{\imath}} \sum_{i=1}^{\bar{\imath}} (\mathcal{X}_i - E\mathcal{X}^a) + \frac{1}{\bar{\imath}} \sum_{i=\bar{k}+1}^{e^{n^{0.2}}} X_i,$$

$$\Delta_{II} := E(\mathcal{Z}^a - \mathcal{Z}) + \frac{1}{\bar{\imath}} \sum_{i=1}^{\bar{\imath}} (\mathcal{Z}_i - E\mathcal{Z}^a) + \frac{1}{\bar{\imath}} \sum_{i=\bar{k}+1}^{e^{n^{0.2}}} Z_i.$$

Thus,

$$\hat{\delta}_{z,t}^M = \frac{E\mathcal{X} + \Delta_I}{E\mathcal{Z} + \Delta_{II}}.$$

Suppose now, that E_{7a} holds. Then, for each $i = 1, \ldots, e^{n^{0.2}}$, we have $\mathcal{Z}_i = \mathcal{Z}_i^a$, $\mathcal{X}_i = \mathcal{X}_i^a$. From (4.22) then follows that $e^{n^{0.2}} \leq \bar{\imath} a$, i.e.

$$e^{n^{0.2}} \geq \bar{\imath} \geq \frac{e^{n^{0.2}}}{a}. \tag{4.23}$$

When $\bar{\imath} = e^{n^{0.2}}$, then $e^{n^{0.2}} - \bar{k} = 0$, otherwise $e^{n^{0.2}} - \bar{k} \leq \mathcal{Z}_{i+1} \leq a$. Since $\sum_{i=\bar{\imath}+1}^{e^{n^{0.2}}} X_i \leq e^{n^{0.2}} - \bar{k}$, we get

$$\frac{1}{\bar{\imath}} \sum_{i=\bar{k}+1}^{e^{n^{0.2}}} X_i \leq \frac{e^{n^{0.2}} - \bar{k}}{\bar{\imath}} \leq \frac{a}{\bar{\imath}} \leq a^2 e^{-n^{0.2}} = \exp(6n^{0.1} - n^{0.2}) < \frac{\Delta}{6}, \tag{4.24}$$

provided n is big enough.
Hence, by (4.23) we have (recall that we assumed $E_{7,a}$)

$$\Big\{\Big|\frac{1}{\bar{\imath}}\sum_{k=1}^{\bar{\imath}}(\mathcal{X}_k - E\mathcal{X}^a)\Big| \leq \frac{\Delta}{3}\Big\} = \Big\{\Big|\frac{1}{\bar{\imath}}\sum_{k=1}^{\bar{\imath}}(\mathcal{X}_k^a - E\mathcal{X}^a)\Big| \leq \frac{\Delta}{3}\Big\} = \bigcup_{l=\frac{e^{n^{0.2}}}{a}}^{e^{n^{0.2}}} \Big\{\Big|\frac{1}{l}\sum_{k=1}^{l}(\mathcal{X}_k^a - E\mathcal{X}^a)\Big| \leq \frac{\Delta}{3}, \bar{\imath} = l\Big\}$$

$$\supset \Big\{\Big|\frac{1}{l}\sum_{k=1}^{l}(\mathcal{X}_k^a - E\mathcal{X}^a)\Big| \leq \frac{\Delta}{3}, l = \frac{e^{n^{0.2}}}{a}, \ldots, e^{n^{0.2}}\Big\} = E_{7,b}.$$

Similarly,

$$\Big\{\Big|\frac{1}{\bar{\imath}}\sum_{k=1}^{\bar{\imath}}(\mathcal{X}_k - E\mathcal{X}^a)\Big| \leq \frac{\Delta}{3}\Big\} \supset E_{7,c}.$$

Thus, by (4.24) on $E_{7a} \cap E_{7b} \cap E_{7c}$ we have

$$|\Delta_I| \leq |E\mathcal{X}^a - E\mathcal{X}| + 2\frac{\Delta}{3} = E(\mathcal{X} - \mathcal{X}^a) + 2\frac{\Delta}{3}$$
$$|\Delta_{II}| \leq |E\mathcal{Z}^a - E\mathcal{Z}| - 2\frac{\Delta}{3} = E(\mathcal{Z} - \mathcal{Z}^a) + 2\frac{\Delta}{3}.$$

Fix $k = 1, 2, \ldots$. Denote by $n_0, n_1, n_2, \ldots$ integers that satisfy $n_0 = 0$, $e^{2n^{0.1}} + 1 \geq n_i - n_{i-1} \geq e^{2n^{0.1}}$, $\forall i$. Let Y_j, $j = 0, 1, \ldots$ denote a Bernoulli random variable which is equal to 1 if and only if between time $\nu(\vartheta(k) + 1 + n_j)$ and $\nu(\vartheta(k) + 1 + n_{j+1})$ the random walk does not visit the point $z^* := z - 2Le^{n^{0.1}}$. The random variables Y_j are independent.
By definition, $\nu(i+1) - \nu(i) \geq e^{n^{0.1}}$. Hence, $\nu(\vartheta(k) + 1 + n_{j+1}) - \nu(\vartheta(k) + 1 + n_j) \geq e^{3n^{0.1}}$. At time $\nu(\vartheta(k) + 1)$, the random walk is located on I_z and, therefore, no more than $3e^{n^{0.1}}$ from z^*. By (4.10), the probability to visit the point z^* within time $e^{3n^{0.1}}$ starting from the $3e^{n^{0.1}}$-neighborhood of z^* goes to 1 as $n \to \infty$. Hence, $\sup_j P(Y_j = 1) \to 0$. Let n be so big, that $P(Y_j = 1) \leq e^{-1}$, for all j. This means, for each

$$P(\mathcal{Z}_k \geq te^{2n^{0.1}}) \leq P(Y_j = 1, j = 0, \ldots, \ulcorner t \urcorner - 1) \leq \exp(-\ulcorner t \urcorner) \leq \exp(-t), \quad k = 1, 2, \ldots \tag{4.25}$$

Now,

$$E(\mathcal{Z} - \mathcal{Z}^a) = \int_{\{\mathcal{Z} \geq a\}} \mathcal{Z} dP - aP(\mathcal{Z} \geq a) = aP(\mathcal{Z} \geq a) + \int_{(a,\infty)} P(\mathcal{Z} > x)dx - aP(\mathcal{Z} \geq a) = \int_{(a,\infty)} P(\mathcal{Z} > x)dx.$$

By (4.25):

$$\int_{(a,\infty)} P(\mathcal{Z} > x)dx \leq \int_a^\infty \exp(-xe^{-2n^{0.1}})dx \leq e^{2n^{0.1}} \exp(-ae^{-2n^{0.1}})) \leq e^{2n^{0.1}} \exp(-e^{n^{0.1}}).$$

Thus, for n big enough:

$$E(\mathcal{Z}-\mathcal{Z}^a)\leq e^{2n^{0.1}}\exp(-e^{n^{0.1}})\leq\frac{\Delta}{3}.$$

Since, $\mathcal{X}\leq\mathcal{Z}$, we get:

$$E(\mathcal{X}-\mathcal{X}^a)=\int_{(a,\infty)}P(\mathcal{X}>x)dx\leq\int_{(a,\infty)}P(\mathcal{Z}>x)dx\leq\frac{\Delta}{3}.$$

Thus, on $E_{7a}\cap E_{7b}\cap E_{7c}$ we have:

$$|\Delta_I|,|\Delta_{II}|\leq\Delta. \tag{4.26}$$

Recall that we have

$$\hat{\delta}^M_{z,t}=\frac{E\mathcal{X}+\Delta_I}{E\mathcal{Z}+\Delta_{II}}.$$

Hence, by (4.26):

$$\frac{E\mathcal{X}-\Delta}{E\mathcal{Z}+\Delta}\leq\hat{\delta}^M_{z,t}\leq\frac{E\mathcal{X}+\Delta}{E\mathcal{Z}-\Delta}.$$

By Taylor's formula,

$$\frac{E\mathcal{X}-\Delta}{E\mathcal{Z}+\Delta}=\frac{E\mathcal{X}}{E\mathcal{Z}}-\Big(\frac{E\mathcal{X}+E\mathcal{Z}}{(E\mathcal{Z})^2}\Big)\Delta+o(\Delta).$$

Since $1\leq E\mathcal{X}\leq E\mathcal{Z}$, the latter means (for Δ small enough)

$$\Big|\frac{E\mathcal{X}-\Delta}{E\mathcal{Z}+\Delta}-\frac{E\mathcal{X}}{E\mathcal{Z}}\Big|\leq\Big(\frac{E\mathcal{X}+E\mathcal{Z}}{(E\mathcal{Z})^2}\Big)\Delta+o(\Delta)\leq 2\Delta+o(\Delta)<3\Delta.$$

Similarly

$$\Big|\frac{E\mathcal{X}+\Delta}{E\mathcal{Z}-\Delta}-\frac{E\mathcal{X}}{E\mathcal{Z}}\Big|<3\Delta.$$

Now, $\delta^M_{z,t}=\frac{E\mathcal{X}}{E\mathcal{Z}}$ implying that

$$|\delta^M_z-\hat{\delta}^M_{z,t}|<3\Delta<e^{-n^{0.12}}.$$

Thus, (4.21) holds.

Second step

We now show that $P(E^c_{7,a})$, $P(E^c_{7,b})$ and $P(E^c_{7,c})$ are of order $o(\exp(-n^{1000}))$.

Taking $t=e^{n^{0.1}}$ (4.25) yields:

$$P(\mathcal{Z}_k>a)\leq\exp(-e^{n^{0.1}}),\quad k=1,2,\ldots.$$

Thus:

$$P(E^c_{7,a})\leq\exp(n^{0.2})\exp(-e^{n^{0.1}})=\exp(n^{0.2}-e^{n^{0.1}})<\exp(-n^{1000}). \tag{4.27}$$

To estimate $P(E_{7,b})$ and $P(E_{7,c})$ we use the Hoeffding inequality. Fix $l\in[\frac{e^{n^{0.2}}}{a},e^{n^{0.2}}]$. By (4.8) we have:

$$P\Big(\Big|\frac{1}{l}\sum_{k=1}^{l}(\mathcal{X}^a_k-E\mathcal{X}^a_k)\Big|\geq\frac{\Delta}{6}\Big)\leq\exp\Big(-2l\big(\frac{\Delta}{a6}\big)^2\Big).$$

On the other hand, since $\mathcal{X}_k^a$, $k \geq 2$ are i.i.d., we have:

$$\Big|\frac{1}{l}\sum_{k=1}^{l} E\mathcal{X}_k^a - E\mathcal{X}^a\Big| = \frac{1}{l}|E\mathcal{X}^a - E\mathcal{X}_1^a| \leq \frac{2a}{l} \leq 2a^2 e^{-n^{0.2}} = 2\exp(6n^{0.1} - n^{0.2}) < \frac{\Delta}{6}.$$

Thus,

$$P\Big(\Big|\frac{1}{l}\sum_{k=1}^{l}\mathcal{X}_k^a - E\mathcal{X}^a\Big| \geq \frac{\Delta}{3}\Big) \leq P\Big(\Big|\frac{1}{l}\sum_{k=1}^{l}(\mathcal{X}_k^a - E\mathcal{X}^a)\Big| \geq \frac{\Delta}{6}\Big) \leq \exp\Big(-2l\big(\frac{\Delta}{a6}\big)^2\Big) \leq \exp(-Ke^{n^{0.2}}\frac{\Delta^2}{a^3}),$$

where $K = \frac{2}{36}$. Now,

$$e^{n^{0.2}}\frac{\Delta^2}{a^3} = \exp(n^{0.2} - \frac{1}{2}n^{0.2} - 9n^{0.1}) = \exp(\frac{1}{2}n^{0.2} - 9n^{0.1}) > \exp(\frac{n^{0.2}}{4})$$

and

$$P\Big(\Big|\frac{1}{l}\sum_{k=1}^{l}\mathcal{X}_k^a - E\mathcal{X}^a\Big| \geq \frac{\Delta}{3}\Big) \leq \exp(-Ke^{\frac{n^{0.2}}{4}}).$$

Finally

$$P(E_{7,b}^c) \leq \sum_{l=\frac{e^{n^{0.2}}}{a}}^{e^{n^{0.2}}} P\Big(\Big|\frac{1}{l}\sum_{k=1}^{l}\mathcal{X}_k^a - E\mathcal{X}^a\Big| \geq \frac{\Delta}{3}\Big) < e^{n^{0.2}}\exp(-Ke^{\frac{n^{0.2}}{4}}) < \exp(-e^{n^{0.1}}) < \exp(-n^{1000}). \tag{4.28}$$

The same bound holds for $P(E_{7,c}^c)$.

Because of (4.21), (4.27) and (4.28) we get:

$$P(E_{7S}^{nc}(c,t)) \leq 3\exp(-n^{1000}). \tag{4.29}$$

The bound in (4.29) do not depend on the choice of z, t and ψ. Note that on $[-cm, cm] \times [0, m^2]$, there are no more than $(cm)^3$ values of (z,t). Hence

$$P\left(E_{7,S}^{nc}\right) \leq \Sigma_{z\in[-cm,cm],t\in[0,m^2]} P\left(E_{7,S}^{nc}(z,t)\right).$$

From (4.29) it follows

$$P\left(E_{7,S}^{nc}\right) \leq (cm)^3 3\exp(-n^{1000}). \tag{4.30}$$

Recall that by (3.10): $(cm)^3 \leq c^3 e^{6n}$. Hence, the right side of (4.30) is less than $3c^3\exp(6n - n^{1000})$. This is of order $o(\exp(-3n))$. By (4.6), we therefore have:

$$P_\psi(E_{7S}^{nc}) \to 0.$$

Outline of the proof that $P_\psi(E_{8,S}^n) \to 1$

Note that in the previous proof the exact nature of $X_{z,i}$, $\mathcal{X}_z(k)$ as well as $X_{z,t,i}$, $\mathcal{X}_{z,t}(k)$ was not used. Hence, the proof holds, if they were replaced by $U_{z,i}$, $\mathcal{U}_z(k)$, $\chi(\nu_{z,t}(i)+e^{n^{0.1}})$ and

$$\sum_{\kappa(k-1)+1}^{\kappa(k)} \chi(\nu_{z,t}(i)+e^{n^{0.1}}),$$

respectively. By (2.12) this proves that $P_\psi(E^n_{8,S}) \to 1$.

Proof that $P_\psi(E^n_{5,S}) \to 1$.

Fix $\psi_n \in E^n_{\text{cell\_OK}}$.

For each $k=0,1,2,\ldots$, let $\tau_k(0) := \rho(k)$ and for each $j=1,2,\ldots$, let $\tau_k(j)$ be the smallest time $t > \tau_k(j-1) + 2e^{n^{0.1}}$ for which $S(t) \in I(\rho(k))$.

Let $X_k(j)$ be the Bernoulli random variable which is equal to one if and only if during time $[\tau_k(j), \tau_k(j) + (n^{3000}+n^2)]$ we observe n^2+1 consecutive 0's or 1's. That is $X_k(j) = 1$ if and only if $\exists t \in [\tau_k(j), \tau_k(j)+n^{3000}]$ such that $\chi(t) = \chi(t+1) = \cdots = \chi(t+n^2)$.
Clearly, for each k, the random variables $X_k(j)$, $j=0,1,2,\ldots$ are independent

At first we show that there exists a constant $a>0$, not depending on n, such that for each k and j,

$$P(X_k(j)=1) \geq n^{-a\ln n} = e^{-a\ln^2 n}. \tag{4.31}$$

Fix $k=0,1,\ldots$ and let $I := I(\rho(k))$. Let $\bar{z}$ be the signal carrier point such that $I_{\bar{z}} = I$. Since $\bar{z}$ is a signal carrier point, then, by Corollary 2.2 and c) of Proposition 2.1, I contains at least one big block of ψ_n. Let $T=[a,b] \subset I$ be that block. Now, let $a < a^* < b^* < b$ be such that $a^*-a, b^*-a^*, b-b^* \geq \frac{|T|}{3} \geq \frac{\ln n}{3n}$. Let $T^* = [a^*, b^*]$. Now,

$$P(X_k(j)=1) \geq P(S(\tau_k(j)+n^{3000}) \in T^*) P(\chi(t) = \chi(t+1) = \cdots = \chi(t+n^2) | S(t) \in T^*).$$

Now, by LCLT:

$$P(S(\tau_k(j)+n^{3000}) \in T^*) \geq \frac{1}{cn^{1500}} - O(\frac{1}{n^{3000}}) \geq n^{-1501},$$

provided that n is big enough.
Let $N = (\frac{n}{\ln n})^2$ (w.l.o.g we assume that this is an integer) and estimate:

$$\begin{gathered} P(\chi(t)=\chi(t+1)=\cdots=\chi(t+n^2)|S(t)=j\in T^*) \geq P(S_j(i)\in T, \quad \forall i=1,2,\ldots,n^2) \geq \\ P\Big(\max_{i=1,\ldots,N} |S_j(i)| \leq \frac{|T|}{3}, S_j(N)\in T^*\Big)^{\ln^2 n} = P\Big(\max_{i=1,\ldots,N} \frac{|S_j(i)|}{\sqrt{N}} \leq \frac{1}{3}, \frac{S_j(N)}{\sqrt{N}} \in \frac{T^*}{\sqrt{N}}\Big)^{\ln^2 n}. \end{gathered} \tag{4.32}$$

Note: $|T^*| \geq \sqrt{N}$. By (4.10):

$$P\Big(\max_{i=1,\ldots,N} \frac{|S_j(i)|}{\sqrt{N}} \leq \frac{1}{3}, \frac{S_j(N)}{\sqrt{N}} \in \frac{T^*}{\sqrt{N}}\Big) \to P(\sup_{0\geq t\leq 1} |W_t| \leq \frac{1}{3\sigma}, W_1 \in I) > \gamma > 0.$$

Thus, for n big enough there exists $a < \infty$ such that the right side of (4.32) is bigger than $(\frac{1}{a})^{\ln^2 n} = n^{-c \ln n}$, with $c > 0$. Hence, (4.31) holds with $a = c + 1$.

Define the following events:

$$E_a(k) = \left\{ \begin{array}{c} \text{if } \rho(k) \leq m^2 \\ \text{then during the time } [\rho(k), \rho(k) + e^{n^{0.3}} - e^{n^{0.1}}] \\ S \text{ visits } I(\rho(k)) \text{ more than } e^{n^{0.22}} \text{ times} \end{array} \right\} \quad k = 0, 1, \ldots$$

and

$$E_a := \cap_{k=1}^{25000} E_a(k).$$

Also define

$$E_b(k) := \Big\{ \sum_{j=0}^{e^{n^{0.21}}} X_k(j) \geq e^{n^{0.2}} \Big\}, \quad E_b := \cap_{k=0}^{n^{25000}} E_b(k).$$

Now, clearly, on $E_a(k)$ we have $\tau_k(e^{n^{0.21}}) \leq \rho(k) + e^{n^{0.3}} - 2e^{n^{0.1}}$. Thus $E^n_{5,S}$ holds, if

$$\sum_{j=0}^{e^{n^{0.21}}} X_k(j) \geq e^{n^{0.2}}.$$

Hence

$$E_{5,S} \supset E_a \cap E_b \quad \text{and} \quad P_\psi(E^c_{5,S}) \leq P_\psi(E^c_a) + P_\psi(E^c_b).$$

We are now proving that $P_\psi(E^c_a) \to 0$ and $P_\psi(E^c_b) \to 0$.

Proof that $P_\psi(E^c_b) \to 0$
By (4.6) it is enough to show that:

$$P(E^c_b | \psi_n) = o(e^{-3n}). \tag{4.33}$$

Note that for big n, $\exp(n^{0.2} - n^{0.21}) < EX_k(j)$, $\forall j$. Thus,

$$\exp(n^{0.2} - n^{0.21}) < \frac{1}{e^{n^{0.21}}} \exp(-n^{0.21}) \sum_{j=0}^{e^{n^{0.21}}} E(X_k(j)) =: \bar{m}.$$

By the Hoeffding inequality we obtain that for a constant $K > 0$:

$$P(E^c_b(k)|\psi_n) = P\Big(\frac{1}{e^{n^{0.21}}} \sum_{j=0}^{e^{n^{0.21}}} X_k(j) < \exp(n^{0.2} - n^{0.21})\Big) \leq P\Big(\frac{1}{e^{n^{0.21}}} \sum_{j=0}^{e^{n^{0.21}}} X_k(j) < \frac{\bar{m}}{2}\Big) =$$

$$P\Big(\frac{1}{e^{n^{0.21}}} \sum_{j=0}^{e^{n^{0.21}}} \big(X_k(j) - EX_k(j)\big) < -\frac{\bar{m}}{2}\Big) \leq \exp(-K\bar{m}^2 e^{n^{0.21}}) \leq \exp(-Ke^{n^{0.21} - 2a \ln^2 n}).$$

Hence,

$$P(E^c_b | \psi_n) \leq n^{25000} \exp(-Ke^{n^{0.21} - 2a \ln^2 n}) = o(e^{-3n}).$$

Proof that $P_\psi(E_a^c) \to 0$
This proof is a little tricky because unlike the other proofs we have that $P(E_a|\psi_n)$ is much bigger than $P(S(m^2) = m)$.

Let $L = n^{100000}$ and consider the event:

$$C = \Big\{S\left(m^2(1 - n^{-3L})\right) \in \left[m(1 - n^{-L}), m(1 + n^{-L})\right] = [m - \frac{m}{n^L}, m + \frac{m}{n^L}]\Big\}.$$

Here and in the rest of the proof we assume (without loss of generality) that all ratios and exponents are integers. Also define

$$E_c(k) = \left\{\rho(k) \notin [m^2(1 - n^{-3L}), m^2]\right\}, \quad k = 0, 1, \ldots, \quad E_c := \cup_{k=1}^{25000} E_c(k).$$

The event E_c means that no stopping time $\rho(k)$ occurs in the time-interval $[m^2(1 - n^{-3L}), m^2]$, the event $E_a \cap E_c$ satisfies

$$E_a \cap E_c = E_a^* := \cap_{k=1}^{25000} E_a^*(k),$$

where

$$E_a^*(k) = \left\{ \begin{array}{c} \text{if } \rho(k) \leq m^2(1 - n^{-L}) \\ \text{then during the time } [\rho(k), \rho(k) + e^{n^{0.3}} - e^{n^{0.1}}] \\ S \text{ visits } I(\rho(k)) \text{ more than } e^{n^{0.22}} \text{ times} \end{array} \right\}.$$

We show that the probability $P(E_a|E^n_{1,S}, \psi_n)$, can be very well approximated by the probability $P(E_a^*|C, \psi_n)$ and the latter goes to 0 when $n \to \infty$. We proceed in three steps.

1) At first note: since

$$C^c \cap E^n_{1,S} = \{S(m^2(1 - n^{-3L})) \notin [m(1 - n^L), m(1 + n^L)], S(m^2) = m\},$$

we get, by the Hoeffdig inequality

$$\begin{aligned} P(C^c \cap E^n_{1,S}|\psi_n) =& P(C^c \cap E^n_{1,S}) = P(E^n_{1,S}|C^c)P(C^c) \leq P(E^n_{1,S}|C^c) \\ =& P\Big(\Big|S\Big(\frac{m^2}{n^{3L}}\Big)\Big| \geq \frac{m}{n^L}\Big) \leq \exp(-dn^L) = o(n^{-3n}). \end{aligned}$$

The latter implies

$$P_\psi(C^c) = o(1). \tag{4.34}$$

2) Second, use the inequalities:

$$P(E_a^{*c} \cap E^n_{1,S} \cap C|\psi_n) \leq P(E_a^c \cap E^n_{1,S} \cap C|\psi_n) \leq P(E_a^{*c} \cap E^n_{1,S} \cap C|\psi_n) + P(E_c^c \cap E^n_{1,S}|\psi_n).$$

Since $\psi \in E^n_7$, it has no signal carrier points in $[m - EZn^{-11001}]$. Hence, $E_c^c \cap E^n_{1,S}$ can hold only, if during time interval $[m^2(1 - n^{-3L}), m^2]$ the random walk covers a distance of at least $EZn^{-11001} - Ln^{1000}$. Thus,

$$P(E_c^c \cap E^n_{1,S}|\psi_n) \leq P\Big(\max_{l=1,\ldots,\frac{m^2}{n^{3L}}} |S(l)| \geq EZn^{-11001} - Ln^{1000}\Big) \leq P\Big(\max_{l=1,\ldots,\frac{m^2}{n^{3L}}} |S(l)| \geq \frac{m}{n^{11003}} - Ln^{1000}\Big).$$

Now use the maximal inequality (4.17) together with the Hoeffding inequality to estimate

$$\begin{aligned} P\Big(\max_{l=1,\dots,\frac{m^2}{n^{3L}}} |S(l)| \geq \frac{m}{n^{11003}} - Ln^{1000}\Big) &\leq \max_{l=1,\dots,\frac{m^2}{n^{3L}}} 3P\Big(|S(l)| \geq \frac{1}{3}\frac{m}{n^{12000}}\Big) \\ &\leq 3\exp(-dn^{3L-12000}) = o(e^{-3n}). \end{aligned}$$

This implies:

$$\frac{P(E_a^c \cap C \cap E_{1,S}^n|\psi_n) - P(E_a^{*c} \cap C \cap E_{1,S}^n, |\psi_n)}{P(E_{1,S}^n|\psi_n)} = P_\psi(E_a^c \cap C) - P_\psi(E_a^{*c} \cap C) = o(1). \quad (4.35)$$

3) Finally, note that:

$$P(E_a^{*c} \cap E_{1,S}^n \cap C|\psi_n) = P(E_a^{*c} \cap C|\psi_n)P(E_{1,S}^n|E_a^{*c} \cap C, \psi_n) = P(E_a^{*c} \cap C|\psi_n)P(E_{1,S}^n|C, \psi_n).$$

On the other hand,

$$P(E_{1,S}^n|\psi_n) \geq P(E_{1,S}^n \cap C|\psi_n) = P(E_{1,S}^n|C, \psi_n)P(C|\psi_n).$$

Hence,

$$P_\psi(E_a^{*c} \cap C) = \frac{P(E_a^{*c} \cap E_{1,S}^n \cap C|\psi_n)}{P(E_{1,S}^n|\psi_n)} \leq \frac{P(E_a^{*c} \cap C|\psi_n)P(E_{1,S}^n|C, \psi_n)}{P(E_{1,S}^n|C, \psi)P(C|\psi_n)} = P(E_a^{*c}|C, \psi_n). \quad (4.36)$$

By CLT, $P(C|\psi_n) = P(S\left(m^2(1 - n^{-3L}\right) \in [m - \frac{m}{n^L}, m + \frac{m}{n^L}])$ is of order $\frac{1}{n^K}$ for some big $K > 0$. We estimate the probability $P(E_a^{*c}|\psi_n)$.
To do this, fix k and let $T_1, T_2, \dots$ denote the waiting times of S between visits of the point $S(\rho(k))$ (when we start at the time $\rho(k)$). Although $ET_i = \infty$, it is known that $ET_i^{\frac{1}{3}} =: K' < \infty$ (see, e.g. [LMM04]). The number K', does not depend on n. Thus, by the Markov inequality we have

$$\begin{aligned} P(E_a^{*c}) &\leq P\Big(\sum_{i=1}^{e^{n^{0.22}}} T_i > e^{n^{0.3}} - e^{n^{0.1}}\Big) = P\left(\Big(\sum_{i=1}^{e^{n^{0.22}}} T_i\Big)^{\frac{1}{3}} > \left(e^{n^{0.3}} - e^{n^{0.1}}\right)^{\frac{1}{3}}\right) \\ &\leq P\Big(\sum_{i=1}^{e^{n^{0.22}}} T_i^{\frac{1}{3}} > \left(e^{n^{0.3}} - e^{n^{0.1}}\right)^{\frac{1}{3}}\Big) \leq \frac{e^{n^{0.22}}K'}{\left(e^{n^{0.3}} - e^{n^{0.1}}\right)^{\frac{1}{3}}} \leq e^{-n^{0.25}}. \end{aligned}$$

Thus, $P(E_{a*}^c) \leq n^{25000}e^{-n^{0.25}} = o(n^{-K})$ implying that

$$P(E_{a*}^c|C, \psi) \leq \frac{P(E_{a*}^c|\psi_n)}{P(C|\psi_n)} = o(1). \quad (4.37)$$

To complete the proof, use (4.34), (4.35), (4.37), (4.37) to get

$$\begin{aligned} P_\psi(E_a^c) &\leq P_\psi(E_a^c \cap C) + P_\psi(C^c) = P_\psi(E_a^{*c} \cap C) + P_\psi(E_a^c \cap C) - P(E_a^{*c} \cap C) + o(1) \\ &\leq P(E_a^{*c}|C, \psi_n) + o(1) = o(1). \end{aligned}$$

5 Combinatorics of g and ĝ

In this section we show: if all scenery dependent events and random walk dependent events hold, then our estimates $\hat{\delta}_T^M$ and $\hat{h}_t$ are precise. This means, we can observe our signals and, just like in our 3-color example, we can estimate the g-function.

Let us first give the definition of the g-function in the 2-colors case.

5.1 Definition of g

In this subsection we give a formal definition of the function

$$g : \{0,1\}^{m+1} \mapsto \{0,1\}^{n^2+1}.$$

The function g depends on n. When n is fixed, we choose $m = \ulcorner n^{2.5} EZ \urcorner$, where the random variable Z is the location of the first Markov signal point after $2Ln^{1000}$ in ξ. We consider the signal carrier points $\bar{z}_1, \bar{z}_2, \ldots$, in $[0, m]$. Define the following subset of $\{0,1\}^{m+1}$:

$$E^* := \{\psi \in \{0,1\}^{m+1} : \bar{z}_1(\psi) \geq L(e^{n^{0.1}} + n^{1000}), \bar{z}_{n^2+1} \leq m - L(e^{n^{0.1}} + n^{1000})\}.$$

Here, $\bar{z}_i(\psi) = \infty$, if the piece of scenery ψ has less than i signal carrier points.

Clearly $E^n_{\text{cell\_OK}} \subset E^*$. If $\psi \in E^*$, then for each $\bar{z}_i(\psi)$ we define the vector of the frequency of ones $h(i)$, $i = 1, \ldots, n^2+1$. Recall from (2.13) that:

$$h(i) = h(\bar{z}_i(\psi)) = P(\psi(U + S(e^{n^{0.1}})) = 1),$$

where U is a random variable with distribution $\mu(\bar{z}_i)$.

Now, if $\psi \in E^*$, let:

$$g_i(\psi) = \begin{cases} 1 & \text{, if } h(i) > 0.5 \\ 0 & \text{, if } h(i) < 0.5 \\ \bar{z}_i(\psi) & \text{otherwise.} \end{cases} \tag{5.1}$$

When $\psi \notin E^*$, define

$$g_i(\psi) = \psi(i), \quad i = 2, 3, \ldots, n^2+2. \tag{5.2}$$

Definition 5.1. *$g(\psi) = (g_1(\psi), \ldots, g_{n^2+1}(\psi))$, where $g_i(\psi)$ is (5.1), if $\psi \in E^*$ and $g_i(\psi)$ is (5.2), if $\psi \notin E^*$.*

Definition 5.1 ensures that $g(\psi)$ depends only on ξ_0^m, and that $(g_1(\xi), \ldots, g_{n^2+1}(\xi))$ is an i.i.d. random vector, with the components being Bernoulli random variables with parameter $\frac{1}{2}$.

5.2 Definition of $\hat{g}$

Next, we formalize the construction of the $\hat{g}$-function. The function $\hat{g} : \{0,1\}^{m^2+1} \mapsto \{0,1\}^{n^2}$ aims to estimate the (non-observable) function g. The argument of $\hat{g}$ is the vector of observations $\chi_0^{m^2} := (\chi(0), \ldots, \chi(m^2))$, and the estimate is given up to the first or last bit. In other words, $\hat{g}$ aims to achieve $\hat{g}(\chi^{m^2}) \sqsubseteq g(\xi|[0, m])$.

The algorithm for computing $\hat{g}$ has 5 phases and it differs from the $\hat{g}$-reconstruction algorithm for the 3-color case (Subsection 1.6) by the first step, only. The rest of the construction is the same.

1. For all $T = [t, t + e^{n0.3}] \subset [0, m^2]$ compute the estimate of the Markov signal probability $\hat{\delta}^M_T$. Select all intervals $T_1 = [t_1, t_1 + e^{n^{0.3}}], T_2 = [t_2, t_2 + e^{n^{0.3}}], \ldots, T_K = [t_K, t_K + e^{n^{0.3}}]$, $t_1 < t_2 < \cdots < t_K$, where the estimated Markov signal probability are higher than c_r. Here K stands for the number of such intervals.

2. For all selected intervals, estimate the frequency of ones. Obtain the estimates $\hat{h}_{T_1}, \ldots, \hat{h}_{T_K}$, $i = 1, \ldots, K$.

3. Define clusters:

$$C_i := \{\hat{h}_{T_j} : |\hat{h}_{T_j} - \hat{h}_{T_i}| \leq 2\exp(-n^{0.12})\},\ \hat{f}_i := \frac{1}{|C_i|}\sum_{j \in C_i} \hat{h}_{T_j}, \quad i = 1, \ldots, K.$$

4. Apply the real scenery construction algorithm $\mathcal{A}^{\mathbb{R}}_n$ (see Subsection 1.6.3) to the vector $(\hat{f}_1, \ldots, \hat{f}_K)$. Denote the output, $\mathcal{A}^{\mathbb{R}}_n(\hat{f}_1, \ldots, \hat{f}_K)$, by

$$(f_1, \ldots, f_{n^2}). \tag{5.3}$$

If the number of different reals in $(\hat{f}_1, \ldots, \hat{f}_K)$ is less than n^2 (e.g. $K \leq n^2$), then complete the vector (5.3) arbitrarily.

5. Define the final output of $\hat{g}$ as follows

$$\hat{g}(\chi^{m^2}) := (I_{[0.5,1]}(f_1), \ldots, I_{[0.5,1]}(f_{n^2})).$$

5.3 Main proof

Next, we prove the main result: when all previously stated events hold, then the $\hat{g}$-algorithm *works*, i.e. $\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)$.

Recall $E^n_{\text{cell\_OK}} = \cap_{i=1}^9 E^n_i$. Similarly define the intersection of the random walk dependent events: $E^n_S := \cap_{i=1}^8 E^n_{i,S}$. Finally, let $E_{\text{g-works}}$ be the event that $\hat{g}$ works, i.e.:

$$E_{\text{g-works}} := \Big\{\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\Big\}. \tag{5.4}$$

At first we show that step 1 in the definition of $\hat{g}$ works properly, i.e. a time interval T is selected (i.e. $\hat{\delta}^M_T > c_r$) only if during the time T the random walk is close to a unique signal carrier point $\bar{z}$. The closeness is defined in the following sense: we say that during time period T, the random walk S is close to z, if there exists $s \in T$ such that $S(s) \in I_z$.

Proposition 5.1. *Suppose $E^n_{\text{cell\_OK}} \cap E^n_S$ holds. Let $T = [t, t + e^{n^{0.3}}] \subset [0, m]$. If during T, the random walk is close to a signal point z, and $\hat{\nu}_t(e^{n^{0.2}}) \leq t + e^{n^{0.3}} - e^{n^{0.1}}$, then $\hat{\delta}^M_T = \hat{\delta}^M_{z,t}$ and $\hat{h}_T = \hat{h}_{z,t}$.*

Proof. Since ξ and S are independent, we fix $\xi = \psi \in E^n_{\text{cell\_OK}}$ and show that the claim of the proposition holds.

Let S be close to the signal point z. By $E^n_2 \cap E^n_8 \cap E^n_9$, the point z has empty neighborhood and empty borders. Hence, in the area

$$([z - L(n^{1000} + e^{n^{0.3}}), z + L(n^{1000} + e^{n^{0.3}})] - [z - L\tilde{M}, z + L\tilde{M}]) \cap [-cm, cm]$$

there are no blocks that are bigger than $n^{0.35}$. Recall that $\tilde{M} = n^{1000} - 2n^2$. Since $2n^{0.35} < n^{0.4} < n^2$, this means: all blocks with length at least $n^{0.4}$ must lay inside the interval $[z - L(n^{1000} - n^2), z + L(n^{1000} - n^2)]$. In particular, this implies - if, during the time T the random walk S visits a block bigger than $n^{0.4}$, then during the n^2 step before and after that visit, it must stay in the interval I_z. Formally: if $\exists s \in T : S(s) \in B$, then

$$S(s - n^2), S(s - n^2 + 1), \ldots, S(s + n^2 - 1), S(s + n^2) \in I_z. \tag{5.5}$$

Here B stands for a block of ψ with length at least $n^{0.4}$.

We now take advantage of the event $E^n_{6,S}$: the random walk cannot generate $n^2 + 1$ times the same color, if it does not visit a block bigger than $n^{0.4}$. By (5.5) this means that all $n^2 + 1$ same colors must be generated on I_z. Hence, inside the time interval T, the stopping times $\hat{\nu}_t(i)$ are equal to the stopping times $\nu_{z,t}(i)$. Similarly, $X_{t,i} = X_{z,t,i}$, provided $\hat{\nu}_t(i) + n^{1000} \leq t + e^{n^{0.3}}$.

By assumption, there are at least $e^{n^{0.2}}$ stopping times $\hat{\nu}_t(i)$ in $[t, t + e^{n^{0.3}} - e^{n^{0.1}}]$ These stopping times are then equal to $\nu_{z,t}(i)$. Similarly, $X_{t,i} = X_{z,t,i}$, $i = 1, \ldots, e^{n^{0.2}}$. The latter means that the observable estimates $\hat{\delta}^M_T$ and $\hat{h}_T$ equals the non-observable estimates $\hat{\delta}^M_{z,t}$ and $\hat{h}_{z,t}$, respectively. □

Corollary 5.1. *Suppose $E^n_{\text{cell\_OK}} \cap E^n_S$ holds. Let $T = [t, t + e^{n^{0.3}}] \subset [0, m]$. If during T the random walk is close to a signal point z, then $\hat{\delta}^M_T > 0$ implies that $\hat{h}_T = \hat{h}_{z,t}$ and $\hat{\delta}^M_T = \hat{\delta}^M_{z,t}$.*

Proof. By definition, $\hat{\delta}^M_T > 0$ if in the time interval $[t, t + e^{n^{0.3}} - e^{n^{0.1}}]$ there are at least $e^{n^{0.2}}$ stopping times $\hat{\nu}_t(i)$. Now Proposition 5.1 applies. □

Lemma 5.1. *Suppose $E^n_{\text{cell\_OK}} \cap E^n_S$ holds. Let $T = [t, t + e^{n^{0.3}}] \subset [0, m]$ be such that $\hat{\delta}^M_T > c_r$. Then there exists an unique signal carrier point $\bar{z} \in [-cm, cm]$ such that S is close to $\bar{z}$ during T and $\hat{\delta}^M_T = \hat{\delta}^M_{\bar{z},t}$.*

Proof. Fix $\xi = \psi \in E^n_{\text{cell\_OK}}$. Note that, since E^n_2 holds, all signal points in $[-cm, cm]$ have empty neighborhood. Together with d) of Proposition 2.1 this means that all signal points in $[-cm, cm]$ are in clusters with diameter less than $2Ln^{1000}$. The distance between any two clusters, i.e. the distance between closest signal points in these clusters, is bigger than $e^{n^{0.3}}$. Moreover, by $E^n_8 \cap E^n_9$, all signal points have empty borders.

If $E^n_{2,S}$ holds, then during time $[0, m^2]$, our random walk stays in $[-cm, cm]$. Together with the clustering structure of the signal points, this means: if during the time interval $T \subset [0, m^2]$ of length $e^{n^{0.3}}$ the random walk S is close to some signal points, then they all belong to the same cluster. Hence, during T, S can be close to at most one signal carrier point (recall, every cluster has one representant, the signal carrier point). We have to show that if $\hat{\delta}^M_T > c_r$, then there exists at least one signal carrier point $\bar{z}$ such that, (during T) S is close to $\bar{z}$.

During T, the random walk S has 3 options :

- S is not close to any signal point
- S is close to the signal points that are not Markov signal points
- S is close to a Markov signal point.

If S is not close to any signal point, then by $E^n_{3,S}$, $\hat{\delta}^M_T \leq c_r$. This excludes the first possibility. Hence, $\hat{\delta}^M_T > c_r$ cannot happen, if during T, S is not close to any signal point.

Suppose now that there exists a signal point z such that (during T) S is close to z. By assumption we have $\hat{\delta}^M_T > c_r > 0$. By Corollary 5.1 we have that $\hat{\delta}^M_T = \hat{\delta}^M_{z,t}$. Now we reap benefit from the events E^n_5 and $E^n_{7,S}$. The event E^n_5 ensures that z is regular, i.e. $|\delta^M_z - c_r| \geq \Delta > e^{-n^{0.12}}$ (recall, Δ is polynomially small). On the other hand, the event $E^n_{7,S}$ ensures $|\hat{\delta}^M_T - \delta^M_z| = |\hat{\delta}^M_{z,t} - \delta^M_z| \leq \exp(-n^{0.12})$. Thus on $E^n_5 \cap E^n_{7,S}$ we have:

$$\hat{\delta}^M_T > c_r \quad \text{if and only if} \quad \delta^M_z > c_r - \Delta. \tag{5.6}$$

Suppose that we have the second possibility – S is close to some signal points, but not close to any Markov signal points. Then z is not a Markov signal point. Hence, (5.6) ensures that $\hat{\delta}^M_T \leq c_r$. This contradicts our assumption that $\hat{\delta}^M_T > c_r$. Hence, z must be a Markov signal point and our third option holds.

Thus $\hat{\delta}^M_T > c_r$ implies that during T, the random walk S is close to a Markov signal point. By clustering structure we know that S is close to a cluster of signal points with at least one Markov signal points. In Subsection 3.4 we argued that such a cluster serves as the signal carrier. However, to complete the proof we must show that, during T, S is also close to the corresponding signal carrier point, say $\bar{z}$.

The points $\bar{z}$ and z belong to the same cluster, i.e. $|\bar{z} - z| < 2Ln^{1000}$. Consider the interval

$$J_z := [z - L(\exp(n^{0.3}), z + L(\exp(n^{0.3})] \cap [-cm, cm].$$

This is the region, where the random walk S stays during time T. We know that the intervals I_z and $I_{\bar{z}}$ both have empty neighborhood and empty borders. Thus all blocks of $\psi|J_z$ that are longer than $n^{0.4}$ must lie in $I_z \cap I_{\bar{z}}$ (by c of Proposition 2.1, in $I_z \cap I_{\bar{z}}$ there is at least one big block of ψ). Argue as in the proof of Proposition 5.1: because of $E^n_{6,S}$, to generate n^2+1 consecutive 0's or 1's, S must visit a block with length at least $n^{0.4}$. To have $\hat{\delta}^M_T > 0$, during T, S must have at least $e^{n^{0.2}}$ such visits. All those blocks are in $I_z \cap I_{\bar{z}} \subset I_{\bar{z}}$. Thus, when $\hat{\delta}^M_T > 0$, then during T, S visits $\bar{z}$ at least $e^{n^{0.2}}$ times. This means that during T, S is close to $\bar{z}$. By Corollary 5.1, we get $\hat{\delta}^M_T = \hat{\delta}^M_{z,t}$. □

Theorem 5.2. *If $E^n_{\text{cell\_OK}}$ and E^n_S both hold, then, for n big enough, $\hat{g}$ works. In other words,*

$$E^n_{\text{cell\_OK}} \cap E_S \subset E_{\text{g-works}}. \tag{5.7}$$

Proof. Suppose $E^n_{\text{cell\_OK}} \cap E^n_S$ hold. Fix $\xi = \psi \in E^n_{\text{cell\_OK}}$ and let

$$g(\psi) = (g_1(\psi), \dots, g_{n^2+1}(\psi)).$$

We have to show: if E^n_S holds, then given the observations $\chi_0^{m^2}$, the function

$$\hat{g}(\chi_0^{m^2}) := (I_{[0.5,1]}(f_1), \ldots, I_{[0.5,1]}(f_{n^2}))$$

is equal to $\hat{g}(\psi)$ up to the first or last bit.

Let $\chi_0^{m^2}$ be the observations. Apply the $\hat{g}$-construction algorithm.

1) At the first step we pick the intervals $T_1 = [t_1, t_1 + e^{n^{0.1}}], \ldots, [t_K, t_K + e^{n^{0.1}}]$ such that for each j, $\hat{\delta}^M_T > c_r$, $j = 1, \ldots, K$. By Lemma 5.1 we know that each interval T_j corresponds to exactly one signal carrier point, say $\bar{z}_{\pi(j)}$.

Let us investigate the mapping $\pi : \{1, \ldots, K\} \mapsto \mathbb{Z}$, where $\pi(j)$ is the index of the signal carrier corresponding to the interval T_j. We now show that π posses the properties A1), A2), A3) that are familiar from the Subsection 1.6.3

A1) $\pi(1) \in \{0, 1\}$

A2) $\pi(K) \geq n^2 + 1$

A3) π is skip-free, i.e. $\forall j,\ |\pi(j) \pm \pi(j)| \leq 1$.

All these properties hold because of $E^n_{4,S} \cap E^n_{5,S}$. Indeed, during the time interval $[0, m^2]$ the random walk starts at 0 and, according to the event $E^n_{1,S}$, ends at m. Let $\bar{z}_1 \ldots, \bar{z}_u$ denote all signal carrier points of ψ in $[0, m]$. By E^n_1, $u > n^2$. The maximal length of a jump of S is L and, therefore, on its way, S visits all intervals $I_{\bar{z}_1}, \ldots I_{\bar{z}_u}$. Recall that the stopping times $\rho(k)$ denote the first visits of the new interval (the first visit of the next interval, not necessarily new for the past). By $E^n_{4,S} \cap E^n_{5,S}$, for each k such that $\rho(k) < m^2$ we have: there is at least $e^{n^{0.2}}$ stopping times $\hat{\nu}_{\rho(k)}(i)$ in $T := [\rho(k), \rho(k) + e^{n^{0.3}} - e^{n^{0.1}}]$. Let $\bar{z}$ be the signal carrier point such that $S(\rho(k)) \in I_{\bar{z}}$. Thus the assumptions of Proposition 5.1 hold and $\hat{\delta}^M_T = \hat{\delta}^M_{\bar{z},t}$. Moreover, by (5.6) we have that $\hat{\delta}^M_T > c_r$, i.e. the interval T will be selected in the first step of the $\hat{g}$-reconstruction.

To summarize: the random walk starts at 0, by convention the first signal carrier point in $[0, \infty)$ is $\bar{z}_1$, the biggest signal carrier point in $(-\infty, 0]$ is $\bar{z}_0$. From Lemma 5.1 we know - during T_1, S must be close to a signal carrier point. On the other hand $[\rho(0), \rho(0) + e^{n^{0.3}}]$ is the first time interval, during which S is close to a signal carrier point. We know that this interval will be selected. Hence $\pi(1) \in \{0, 1\}$.

On its way S visits all signal carrier interval $I_{\bar{z}_1}, \ldots I_{\bar{z}_u}$. Right after the first visit of a new signal carrier, $\rho(k)$, the random walk produces an interval $T = [\rho(k), \rho(k) + e^{n^{0.3}}]$ that will be selected. Together with Lemma 5.1 the latter yields that π is skip-free.

Recall that $\bar{z}_u$ is the last signal carrier point in $[0, m]$. Thus, the last signal carrier interval S visits during $[0, m^2]$ is $\bar{z}_u$ or $\bar{z}_{u+1}$. By E^n_7 we know that $\bar{z}_u$ lays in $[0, m - Le^{n^{0.3}}]$. Hence, if $S(\rho(k)) \in I_{\bar{z}_u}$, then $[\rho(k), \rho(k) + e^{n^{0.3}}]$ will be selected. We get that the last selected interval corresponds to the signal carrier that is at least $\bar{z}_{n^2+1}$. Thus $\pi(K) \geq n^2 + 1$.

Let $\pi_* := \min\{\pi(j) : j = 1, \ldots, K\}$, $\pi^* := \max\{\pi(j) : j = 1, \ldots, K\}$. We just saw that $\pi_* \leq 1$, $\pi^* \geq n^2 + 1$ and π is a skip-free random walk on $\{\pi_*, \pi_* + 1, \ldots, \pi^*\}$.

The rest of the algorithm was already explained in Subsection 1.6.3. However, in the following we give a bit more formal explanation.

2) At the second step we calculate $\hat{h}_{T_1}, \ldots, \hat{h}_{T_K}$. By Lemma 5.1, we know that, for each $j = 1, \ldots, K$

$$\hat{h}_{T_j} = \hat{h}_{\bar{z}_{\pi(j)}, t_j}.$$

3) Since $E^n_{8,S}$ holds, we know that, for each $j = 1, \ldots, K$,

$$|\hat{h}_{T_j} - h(\bar{z}_{\pi(j)})| = |\hat{h}_{\bar{z}_{\pi(j)}, t_j} - h(\bar{z}_{\pi(j)})| < \exp(-n^{0.12}).$$

This means: if $\pi(i) = \pi(j)$ then $|\hat{h}_{T_i} - \hat{h}_{T_j}| \leq 2\exp(-n^{0.12})$.
On the other hand, by E^n_3 we know that $\pi(i) \neq \pi(j)$ implies

$$|h(\bar{z}_{\pi(j)}) - h(\bar{z}_{\pi(i)})| \geq \exp(-n^{0.11}). \tag{5.8}$$

We assume n to be big enough to satisfy $\exp(-n^{0.12}) < 5\exp(-n^{0.11})$. Hence $\pi(i) \neq \pi(j)$ implies that $|\hat{h}_{T_i} - \hat{h}_{T_j}| > 2\exp(-n^{0.12})$. Thus, if $E^n_{8,S} \cap E^n_3$, then for each $i, j = 1, \ldots, k$ we have

$$\hat{h}_j \in C_i \quad \text{if and only if} \quad \pi(i) = \pi(j). \tag{5.9}$$

Hence the clusters C_i and C_j are either identical or disjoint; $C_i = C_j$ if and only if $\pi(j) = \pi(i)$. The same, obviously, holds for the averages:

$$\hat{f}_j = \hat{f}_i \quad \text{if and only if} \quad \pi(i) = \pi(j).$$

Let for each $i = \{\pi_*, \pi_* + 1, \ldots, \pi^*\}$, $\hat{f}(\bar{z}_i) = \hat{f}_j$, if $\pi(j) = i$. Hence, $\hat{f}(\bar{z}_i)$ is the estimate of $h(\bar{z}_i)$ and

$$\hat{f}_j = \hat{f}(\bar{z}_{\pi(j)}), \quad j = 1, \ldots, K.$$

Hence, $j \mapsto \hat{f}_j$ can be considered as the observations of the skip-free random walk π on the different reals $\{\hat{f}(\bar{z}_{\pi_*}), \hat{f}(\bar{z}_{\pi_*+1}), \ldots \hat{f}(\bar{z}_{\pi^*})\}$.

4) The real scenery construction algorithm $\mathcal{A}^{\mathbb{R}}_n$ is now able to reconstruct the numbers $\hat{f}(z_1), \ldots, \hat{f}(z_{n^2+1})$ up to the first or last number. Thus

$$(f_1, \ldots, f_{n^2}) = \mathcal{A}^{\mathbb{R}}(\hat{f}_1, \ldots, \hat{f}_K) \sqsubseteq (\hat{f}(\bar{z}_1), \ldots, \hat{f}(\bar{z}_{n^2+1})).$$

5) By E^n_4, we have that $|h(\bar{z}_i) - 0.5| \leq \exp(-n^{0.11})$. >From (5.8) and (5.9), it follows:

$$|\hat{f}_i - h(\bar{z}_{\pi(i)})| \leq \exp(-n^{0.12}).$$

The latter implies:

$$\hat{f}(\bar{z}_i) \geq 0.5 \quad \text{if and only if} \quad h(\bar{z}_i) \geq 0.5.$$

Hence , for each $i = 1, \ldots, n^2 + 1$, we have that $I_{[0.5,1]}(\hat{f}(\bar{z}_i)) = I_{[0.5,1]}(h(\bar{z}_i))$. Thus:

$$\hat{g}(\chi_0^{m^2}) = \Big(I_{[0.5,1]}(f_1), \ldots I_{[0.5,1]}(f_n^2)\Big) \sqsubseteq \Big(I_{[0.5,1]}(h(\bar{z}_1)), \ldots I_{[0.5,1]}(h(z_{n^2+1}))\Big) = g(\psi).$$

$\square$

Proof of Theorem 1.1 Fix $c > 0$ such that Proposition 4.1 holds for $\epsilon = \frac{1}{8}$. Use this particular c to define all scenery dependent events as well as all random walk-dependent vents.

The intersection of all scenery-dependent events is $E^n_{\text{cell\_OK}}$. In Section 3.2, we proved that $P(E^n_{\text{cell\_OK}}) \to 1$. Hence **1)** holds.

Now consider the event E^n_S. Use Theorem 5.2 to find the integer $N_1 < \infty$ such that for each $n > N_1$, (5.4) hold. Then, for each $n > N_1$, $\psi_n \in E^n_{\text{cell\_OK}}$ we have

$$P(g(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)|S(m^2) = m, \xi = \psi_n) \geq P(E^n_S|S(m^2) = m, \xi = \psi_n) = P_\psi(E^n_S).$$

In Section 4.3, we proved that $\liminf_n P_\psi(E^n_S) \geq 1 - \frac{1}{8}$. Let N_2 be so big that $P_\psi(E_n) > \frac{3}{4}$ $\forall n > N_1$. Take $N := N_1 \vee N_2$. With such N, **2**) holds.

Finally, the statement **3**) follows from the definition of g in Section 5.1.

References

[BK96] Itai Benjamini and Harry Kesten. Distinguishing sceneries by observing the scenery along a random walk path. *J. Anal. Math.*, 69:97–135, 1996. MR1428097

[dH88] W. Th. Frank den Hollander. Mixing properties for random walk in random scenery. *Ann. Probab.*, 16(4):1788–1802, 1988. MR0958216

[dHS97] Frank den Hollander and Jeff E. Steif. Mixing properties of the generalized T, T^{-1}-process. *J. Anal. Math.*, 72:165–202, 1997. MR1482994

[dHS06] Frank den Hollander and Jeffry Steif. Random walk in random scenery: a survey of some recent results. In *Dynamics&stochasics.*, volume 48, pages 53–65. IMS Lecture Lotes Monogr. Ser., 2006. MR2306188

[HHR00] D. Heicklen, C. Hoffman, and D. J. Rudolph. Entropy and dyadic equivalence of random walks on a random scenery. *Adv. Math.*, 156(2):157–179, 2000. MR1808244

[HK97] Matthew Harris and Mike Keane. Random coin tossing. *Probab. Theory Related Fields*, 109(1):27–37, 1997. MR1469918

[HM06] Andrew Hart and Heinrich Matzinger. Markers for error-corrupted observations. *Stochastic Processes and their Applications*, 116:807–829, 2006. MR2218337

[How96] C. Douglas Howard. Detecting defects in periodic scenery by random walks on $\mathbb{Z}$. *Random Structures Algorithms*, 8(1):59–74, 1996. MR1368850

[How97] C. Douglas Howard. Distinguishing certain random sceneries on $\mathbb{Z}$ via random walks. *Statist. Probab. Lett.*, 34(2):123–132, 1997. MR1457504

[Kal82] S. A. Kalikow. T, T^{-1} transformation is not loosely Bernoulli. *Ann. of Math. (2)*, 115(2):393–409, 1982. MR0647812

[KdH86] Mike Keane and W. Th. F. den Hollander. Ergodic properties of color records. *Phys. A*, 138(1-2):183–193, 1986. MR0865242

[Kes96] Harry Kesten. Detecting a single defect in a scenery by observing the scenery along a random walk path. In *Itô's stochastic calculus and probability theory*, pages 171–183. Springer, Tokyo, 1996. MR1439524

[Kes98] Harry Kesten. Distinguishing and reconstructing sceneries from observations along random walk paths. In *Microsurveys in discrete probability (Princeton, NJ, 1997)*, pages 75–83. Amer. Math. Soc., Providence, RI, 1998. MR1630410

[Lin99] Elon Lindenstrauss. Indistinguishable sceneries. *Random Structures Algorithms*, 14(1):71–86, 1999. MR1662199

[LM02a] Jüri Lember and Heinrich Matzinger. Reconstructing a piece of 2-color scenery. Eurandom http://www.eurandom.nl/EURANDOMreports.htm, 2002. Submitted.

[LM02b] M atthias Löwe and Heinrich Matzinger. Scenery reconstruction in two dimensions with many colors. *Ann. Appl. Probab.*, 12(4):1322–1347, 2002. MR1936595

[LM03] Matthias Löwe and Heinrich Matzinger. Reconstruction of sceneries with correlated colors. *Stochastic Processes and their Applications*, 105(2):175–210, 2003. MR1978654

[LM06] Jüri Lember and Heinrich Matzinger. Reconstruction of periodic scenery seen along a random walk with jumps. *Stochastic Processes and their Applications*, 116(11):1584–1599, 2006. MR2269217

[LMM04] Matthias Löwe, Heinrich Matzinger, and Franz Merkl. Reconstructing a multicolor random scenery seen along a random walk path with bounded jumps. *Electron. J. Probab.*, 9:no. 15, 436–507 (electronic), 2004. MR2080606

[LP04] D. A. Levin and Y. Peres. Identifying several biased coins encountered by a hidden random walk. *Random Structures Algorithms*, 25(1):91–114, 2004. MR2069666

[LPP01] David A. Levin, Robin Pemantle, and Yuval Peres. A phase transition in random coin tossing. *Ann. Probab.*, 29(4):1637–1669, 2001. MR1880236

[Mat99a] Heinrich Matzinger. *Reconstructing a 2-color scenery by observing it along a simple random walk path with holding.* PhD thesis, Cornell University, 1999.

[Mat99b] Heinrich Matzinger. Reconstructing a three-color scenery by observing it along a simple random walk path. *Random Structures Algorithms*, 15(2):196–207, 1999. MR1704344

[Mat05] Heinrich Matzinger. Reconstructing a 2-color scenery by observing it along a simple random walk path. *Ann. Appl. Prob*, 15(1):778–819, 2005. MR2114990

[ML06] Heinrich Matzinger and Jüri Lember. Scenery reconstruction: an overview. In *Information and Randomness (Maass, Martinez, San Martin (Eds))*, volume 66, pages 76–125. Hermann, Travaux en cours, 2006.

[MP07] Heinrich Matzinger and Sergue. Popov. Deteting local pertubation in a continuous scenery. *Electronic Journal of Probability*, 12:637–660, 2007. MR2318405

[MR03a] Heinrich Matzinger and Silke W.W. Rolles. Reconstructing a piece of scenery with polynomially many observations. *Stochastic Processes and their Applications*, 107(2):289–300, 2003. MR1999792

[MR03b] Heinrich Matzinger and Silke W.W. Rolles. Reconstructing a random scenery observed with random errors along a random walk path. *Probab. Theory Related Fields*, 125(4):539 – 577, 2003. MR1974414

[MR06] Heinrich Matzinger and Silke W. W. Rolles. Retrieving random media. *Probab. Theory Related Fields*, 136(3):469 – 507, 2006. MR2257132

[Pet95] V. V. Petrov. *Limit theorems of probability theory: sequences of independent random variables*. Clarendon Press, Oxford, 1995. MR1353441