SCENERY RECONSTRUCTION: AN OVERVIEW

Heinrich Matzinger [1] and Jüri Lember [2]

GEORGIA TECH
School of Mathematics, Georgia Tech, Atlanta, Georgia 30332-0160, USA

UNIVERSITY OF TARTU
J. Liivi 2-513, Tartu, Estonia

Abstract

Scenery reconstruction is the problem of recovering a text which has been mixed up by a random walk. To recreate the text, one disposes exclusively of the observations made by a random walker. This topic stems from questions by Benjamini, Keane, Kesten, den Hollander, and others. The field of scenery reconstruction has been highly active quite recently. There exists a large variety of different techniques. Some of them are easy, some can be rather inaccessible. We describe where the different techniques are employed and explain several of the methods contained in the inaccessible papers. A reading guide to scenery reconstruction is provided. Most scenery papers will soon be available on the web: www.math.gatech.edu̇matzi

1 Introduction

The scenery reconstruction problem investigates whether one can identify a coloring of the integers, using only the color record seen along a random walk path. The problem originates from questions by Benjamini, Keane, Kesten, den Hollander, and others.

Specification of the problem. A (one dimensional) scenery is a coloring ξ of the integers $\mathbb{Z}$ with C_0 colors $\{1, \ldots, C_0\}$. Two sceneries are called equivalent if one of them is obtained from the other by a translation or reflection. Let $(S(t))_{t\geq 0}$ be a recurrent random walk on the integers. Observing the scenery ξ along the path of this random walk, one sees the color $\chi(t) := \xi(S(t))$ at time t. The *scenery reconstruction problem* is to retrieve the scenery ξ, given only the sequence of observations χ.
This problem can also be formulated as follows:
Does one path realization of the process $\{\chi(t)\}_{t\geq 0}$ uniquely determine ξ? The answer in those general terms is "no". However, under appropriate restrictions, the answer will become "yes". Let us explain these restrictions: First, if ξ and $\tilde{\xi}$ are equivalent, we can in general not distinguish whether the observations come from ξ or from $\tilde{\xi}$. Thus, we

[1]E-mail: matzi@math.gatech.edu Supported by SFB701 A3
[2]Supported by the Estonian Science Foundation Grant nr.5694 and SFB701 A3

can only reconstruct ξ up to equivalence. Second, the reconstruction can obviously work at best almost surely. Moreover, Lindenstrauss in [18] exhibits sceneries which can not be reconstructed. However, "typical" sceneries can be reconstructed up to equivalence (a.s.). We take the scenery ξ to be a realization of a random process (random scenery), and prove that almost every realization can be reconstructed a.s. up to equivalence. Most scenery reconstruction results assume that random scenery ξ and random walk S are independent of each other and distributed according to given laws μ and ν. Let us write $\xi \approx \psi$ when the sceneries ξ and ψ are equivalent. Scenery reconstruction results are formally formulated as follows:
Given that ξ and S are independent and follow the laws μ, respectively ν, there exists a measurable function

$$\mathcal{A} : C_0^{\mathbb{N}} \longrightarrow C_0^{\mathbb{Z}}$$

such that

$$P(\mathcal{A}(\chi) \approx \xi) = 1.$$

The methods used for scenery reconstruction change entirely when one modifies the number of colors in the scenery. (Except for [20], all scenery reconstruction papers so far, assume the scenery to be i.i.d.). Furthermore, taking another distribution for the random walk can completely modify the relation between χ and ξ. The scenery reconstruction methods are thus manifold. This is one of the reasons, why this subfield has become very active recently.

2 History

The first positive result about scenery reconstruction is Matzinger's Ph.D. Thesis [22]. This thesis was written under the supervision of Harry Kesten. Later Kesten noticed that the result [22] heavily relies on the skip-free property and the one-dimensionality of the random walk. This remark triggered intense research on the topic of scenery reconstruction. During the next three years at the Eurandom institute in Eindhoven, Jüri Lember, Mathias Löwe, Heinrich Matzinger, Franz Merkl and Silke Rolles devoted a large part of their time to scenery reconstruction. Later they all the worked in the group of Friedrich Götze in Bielefeld and continued devotin time to the subject of scenery reconstruction.
Recently, scenery reconstruction was also a topic present in Latin America: Andrew Hart, from Servet Martinez's Nucleo Millenio, worked in the area whilst Popov and Pichon from the [5].

Motivations coming from ergodic theory. Scenery reconstruction is part of the research area which investigates the ergodic properties of the color record χ. One of the motivations comes from ergodic theory, for example via the T, T^{-1} problem. The origin of this problem is a famous conjecture by Kolmogorov. He demonstrated that every Bernoulli shift T has a trivial tail-field (let us call the class of all transformations having

a trivial tail-field $\mathcal{K}$) and conjectured that also the converse is true. This was proved to be wrong by Ornstein, who presented an example of a transformation which is $\mathcal{K}$ but not Bernoulli. Evidently his transformation was constructed for the particular purpose to resolve Kolmogorov's conjecture. In 1971, Ornstein, Adler, and Weiss came up with a very natural example which is $\mathcal{K}$ but appeared not to be Bernoulli. This was the T, T^{-1}-transformation, and the T, T^{-1}-problem was to show that it was not Bernoulli. In a celebrated paper [10], Kalikow showed that the T, T^{-1}-transformation is not even loosely Bernoulli and therefore solved the T, T^{-1}-problem. A generalization of this result was recently proved by den Hollander and Steif [2].
The T, T^{-1}-transformation gives rise to a random process of pairs. The first coordinate of these pairs can be regarded as the position of a realization of simple random walk on the integers at time i. The second coordinate tells which color the walker would read at time i, if the integers were colored by an i.i.d. process with black and white in advance.

Observations of a random media by a random walk constitute a natural and important class of distributions. It is very differently behaved from most standard ergodic processes. The ergodic properties of the observations χ were investigated by Heicklen, Hoffman, Rudolph in [6], Kesten and Spitzer in [14], Keane and den Hollander in [11], den Hollander in [3], and den Hollander and Steif in [2].

A related topic: distinguishing sceneries. A related important problem is to distinguish sceneries: Let η_1 and η_2 be two given sceneries. Assume that either η_1 or η_2 is observed along a random walk path, but we do not know which one. Can we figure out which of the two sceneries was taken? The problem of distinguishing two sceneries which differ only in one point is called "detecting a single defect in a scenery". Benjamini, den Hollander, and Keane independently asked whether all non-equivalent sceneries could be distinguished. Kesten and Benjamini [1] proved that one can distinguish almost every pair of realizations of two independent random sceneries even in two dimensions and with only two colors. Before that, Howard had proven in [7, 8, 9] that any two periodic one dimensional non-equivalent sceneries are distinguishable, and that one can almost surely distinguish single defects in periodic sceneries. Kesten in [12] proved that one can a.s. recognize a single defect in a random scenery with at least five colors. He asked whether one can distinguish a single defect even if there are only two colors in the scenery.
Kesten's question was answered by the following result, proved in Matzinger's Ph.D. Thesis [22]: Almost every 2-color scenery can be reconstructed, a.s.. In [23], Matzinger proves that almost every 3-color scenery can be reconstructed, a.s..

3 Overview

3.1 3-color scenery seen along a simple random walk

In this subsection, we discuss the case presented in [23]. The assumptions for [23] as well as for this subsection are:

- The scenery ξ has three colors, so that $\xi : \mathbb{Z} \to \{0, 1, 2\}$. The scenery is i.i.d. and the three colors are equiprobable.
- The random walk S is a simple random walk, starting at the origin.

First, we need some notations:
For $x < y$, let ξ_x^y denote the piece of scenery ξ between the points x and y:

$$\xi_x^y := \xi(x)\xi(x+1)\dots\xi(y).$$

For $s, t \in \mathbb{N}$ with $s < t$, let χ_s^t denote the observations between time s and time t:

$$\chi_s^t := \chi(s)\chi(s+1)\dots\chi(t).$$

Assume that the random walk *crosses* between time s and t from point x to y. By this we mean that

$$S(s) = x, S(t) = y \tag{3.1}$$

and for all time $r \in (s, t)$, it holds

$$x < S(r) < y. \tag{3.2}$$

We can imagine that χ_s^t is obtained after copying the original information ξ_x^y incorrectly. The question is how much of the information of ξ_x^y is still contained in χ_s^t? It turns out that there is a large amount of information contained in the original sequence ξ_x^y, which can always be recovered if we are only given the observations χ_s^t.
In what follows, we describe a word, which depends only on ξ_x^y and can be recovered from χ_s^t. We call this word the *fingerprint* of ξ_x^y. The fingerprint is characteristic for the original sequence ξ_x^y and does not depend on the random walk path. The fingerprint is obtained by replacing in the words ξ_x^y any substring of the form aba by a, (where $a, b \in \{0, 1, 2\}$). We proceed until there is no more substring aba to be replaced. The final word obtained is defined to be the fingerprint of ξ_x^y. The order in which we process the different parts of the original word ξ_x^y does not matter, the result is always the same.
If we process the observations χ_s^t using the same method, we obtain the same fingerprint as for ξ_x^y. Hence, the fingerprint of ξ_x^y can always be reconstructed if we are only given χ_s^t.
Let us formalize these fundamental properties:

- The observations χ_s^t and the original sequence ξ_s^t are in the same equivalence class modulo $a = aba$, when (3.1) and (3.2) both hold.

- Every equivalence class mudolo $a = aba$ has exactly one minimal element.

The first property guarantees that given only χ_s^t, we can recover the fingerprint of ξ_s^t. The second property above states that the fingerprint is well defined. So, the fingerprint can be viewed as the smallest element of the modulo $a = aba$ equivalence class in the free semi-group over the three symbols $0, 1, 2$.

Let us look at a numerical example: Take the scenery ξ between 0 and 6 to be equal to:

$\xi(z)$	...	0	2	0	1	0	0	1	...
z	...	0	1	2	3	4	5	6	...

Assume that the random walk goes in ten steps from 0 to point 5 choosing the following path:

$$(S(0), S(1), S(2), \ldots, S(10)) = (0, 1, 2, 1, 2, 3, 4, 3, 4, 5, 6).$$

This means that the color record observed by the random walk between time 0 and time 10 is:

$$\chi(0)\chi(1)\ldots\chi(10) = 02020101001.$$

Let us process the piece of scenery ξ_0^6. By replacing stings aba by a, we obtain successively:

$$\xi_0^6 = 02\overbrace{010}^{0}01 \to \overbrace{020}^{0}01 \to 001.$$

The string 001 can not be further processed. Hence 001 is our fingerprint. Let us process ξ_0^6 in a different order:

$$\xi_0^6 = \overbrace{020}^{0}1001 \to \overbrace{010}^{0}01 \to 001.$$

We note that whatever processing order we chose, we always obtain the same final result 001.
Next, we process the observations χ_0^{10}:

$$\chi_0^{10} = \overbrace{020}^{0}20101001 \to \overbrace{020}^{0}101001 \to \overbrace{010}^{0}1001 \to \overbrace{010}^{0}01 \to 001.$$

the end result is the word 001. This is the same as the fingerprint of ξ_0^6. This shows how we can reconstruct the fingerprint of ξ_0^6 given only the observations χ_0^{10}. When the scenery contains many colors the fingerprint is not too different from the scenery. We first reconstruct the "fingeprint" of the whole scenery. Then we use statistical methods to fill in the gaps and reconstruct the scenery from the fingerprint.

3.2 2-color scenery seen along a simple random walk with holding

In this subsection, we discuss the case presented in Matzinger's thesis [22], as well as in the articles: [25, **?**, 26]. The assumptions for [25, 22, **?**, 26] as well as for this subsection are the following:

- The scenery ξ has two colors, so that $\xi : \mathbb{Z} \to \{0, 1\}$. Furthermore the random scenery is assumed to be i.i.d. with $P(\xi(z) = 0) = P(\xi(z) = 1) = 1/2$ for all $z \in \mathbb{Z}$.

- The random walk S is a *simple random walk with holding*, starting at the origin, i.e.:

$$P(S(t+1)-S(t)=1)=P(S(t+1)-S(t)=-1)=P(S(t+1)-S(t)=0)=\frac{1}{3}$$

for all $t \in \mathbb{N}$.

Let $x \in \mathbb{Z}$ be such that $\xi(z) \neq \xi(z+1)$. In the case presented in the present subsection, the random walk can generate any observations by just hoping back and forth between the points z and $z+1$. Since the scenery is i.i.d., about half of the points z contained in a large interval satisfy $\xi(z) \neq \xi(z+1)$. This implies that in many places in the scenery ξ, the random walk can generate any pattern. Hence, in general, the observations do not contain "absolute sure information" about the underlying scenery.
This situation is similar to the typical situation encountered in statistics: We test a hypothesis, but we can not be sure whether the hypothesis holds or not. Rather, we decide if a hypothesis is likely or not given the observed data. For our scenery problem, this means the following:
Given χ_s^t we can infer certain features of the underlying piece of scenery ξ_x^y. Conditional on the observations, ξ_x^y might have very high likelihood to present certain features.
This situation is thus fundamentally different from the case presented in the last subsection, where we could reconstruct the fingerprint with certainty.

Let us illustrate the above remark by a numerical example. Take the scenery ξ between 0 and 6 to be equal to:

$\xi(z)$	...	0	1	1	1	0	0	1	...
z	...	0	1	2	3	4	5	6	...

Note that $\xi(0) \neq \xi(1)$. Hence, the random walk can generate any possible observation by just hoping back and forth between 0 and 1. Take the sequence 010001. The random walk can generate this sequence by following the path:

$$(S(0), S(1), S(2), S(3), S(4), S(5), \ldots) = (0, 1, 0, 0, 0, 1, \ldots).$$

Similarly, we have that $\xi(3) \neq \xi(4)$. Hence the random walk can generate any observations by moving only between the points 3 and 4. Assume that at time t the random walk is located on point 4: $S(t) = 4$. To generate the finite string 010001 directly after time t, the random walk can follow the path:

$$(S(t), S(t+1), S(t+2), S(t+3), S(t+4), S(t+5)) = (4, 3, 4, 4, 4, 3).$$

At first sight, the combinatorial method presented in the previous subsection seems completely useless for the present case. However, in his Ph.D. Thesis [22], Matzinger discovered that the algebraic-combinatorial structure explained in the previous subsection plays also an important role for the present case. More precisely, he discovered that the combinatorial structure of last section is contained in a hidden form in the conditional distribution of the observations χ given the scenery ξ. All the papers cited at the beginning of this subsection proceed in two steps:

- Given the observations χ, try to determine (approximately) the conditional distribution:

$$\mathcal{L}(\chi|\xi). \tag{3.3}$$

- Given the conditional distribution (3.3), determine the scenery ξ, (up to equivalence).

None of the papers cited in the beginning of this subsection are easily accessible. So, we decided to give the main idea of how to reconstruct ξ if we are given the conditional distribution (3.3) in Section 5. In this section, we also show how the conditional distribution (3.3) can be estimated with the help of only one realization of χ.

3.3 The development of the subfield of scenery reconstruction

The development of the subfield of scenery reconstruction took mainly place in three phases. In each phase, it become possible to reconstruct sceneries in a more complicated setting.

1. Combinatorial case: Multicolor scenery, simple random walk. This is the easiest situation. It occurs for example, when we observe a scenery with two or more colors along a simple random walk path. (The random scenery being i.i.d.). Subsection 3.1 has presented an example of the behavior typical in this setting: FROM FINITE MANY OBSERVATIONS, IT IS POSSIBLE TO RETRIEVE SOME INFORMATION ABOUT THE UNDERLYING SCENERY. THIS INFORMATION HOLDS WITH CERTAINTY, AS OPPOSED TO BE JUST "LIKELY". The situation in the 2-color and 3-color scenery reconstruction papers [24] and [23] are typical for this "combinatorial case". Although the methods in [24] and [23] are very different from each other, they are both based on some " algebraic approach". In [21] Matzinger, Merkl and Loewe, consider the case where the number of colors is strictly larger than the number of possibilities the random walk has at each step. This situation also belongs to the "combinatorial case". However, the reconstruction algorithm in [21] is not based on an algebraic approach.
Some very general ideas on scenery reconstruction like the zero-one-law for scenery reconstruction (see Subsection 4.6) and the stopping time replacement approach (see Subsections 4.4 and 4.5) valid for any type of scenery reconstruction, are presented in [21]. Difficult problems like the 2-dimensional reconstruction [19], heavily rely on these general ideas.
It is still an open problem to characterize the (joint) distributions of ξ and S which make the "combinatorial case" occur. (By this we mean: Which create a situation, where a finite number of observations contain sure information about the underlying scenery). We conjecture that the entropy in the scenery needs to exceeds the entropy of that of the random walk.

2. Semi-combinatorial case: 2-color scenery, simple random walk with holding. This is the situation described in Subsection 3.2. In the semi-combinatorial case, the method presented in Subsection 3.1 still plays an important role: the conditional distribution $\mathcal{L}(\chi|\xi)$ contains a hidden combinatorial structure similar to the one in Subsection 3.1.

The first reconstruction result for this situation was presented in Matzinger's thesis [22]. Later, Rolles and Matzinger [**?**, 26, 25] showed that it is possible to reconstruct a finite piece of scenery in polynomial time. They prove their result in the case of a simple random walk with holding and a 2-color scenery. The question about polynomial time reconstruction originated from Bejamini.
There is no easy accessible paper for the semi-combinatorial case. We present a rough sketch of some of the main ideas in Section 5.

3. Purely statistical case: Random walk with jumps and 2-color scenery. We say that the random walk S jumps if

$$P\left(S(t+1) - S(t) \in \{-1, 0, 1\} \right) \neq 1. \tag{3.4}$$

The purely statistical case, occurs when the random walk S jumps and when ξ is a 2-color i.i.d. scenery. In this situation, the combinatorial methods developed for the simple random walk are useless. Furthermore, the techniques developed for the random walk with holding do not work either: the conditional distribution $\mathcal{L}(\chi|\xi)$ is completely intractable when the random walk is allowed to jump (we explane this in Subsection 6.1). It is a very uneasy task to reconstruct sceneries in this setting. In [15, **?**], Matzinger and Lember solve the reconstruction problem in the 2-color, jump case. They use an information theoretical approach:
Instead of trying to reconstruct the scenery right away, they first ask what amount of information the observations contain about the underlying scenery. More precisely, they give a lower bound for the mutual information of the observations $\chi_0^{n^2}$ and the underlying piece of scenery ξ_0^n. They prove [15] that $I(\chi_0^{n^2}, \xi_0^n)$ is larger than order $\ln n$. This is a very small bound considering that $H(\xi_0^n) = n + 1$.
In [**?**], Lember and Matzinger prove that the lower bound for the mutual information $I(\chi_0^{n^2}, \xi_{-n}^n)$, implies that ξ can be a.s. reconstructed up to equivalence.

Let us mention other cases, which are strongly different from the three above.

Scenery reconstruction given disturbed input data. In [27], Matzinger and Rolles adapted the method proposed by Löwe, Merkl and Matzinger to the case where random errors occur in the observed color record. They show that the scenery can still be reconstructed provided the probability of the errors is small enough. When the observations are seen with random errors, the reconstruction of sceneries is closely related to some coin tossing problems. These have been investigated by Harris and Keane [4] and Levin, Pemantle and Peres [16]. The paper [27] on reconstruction with errors was motivated by their work and by a question of Peres: He asked for generalizations of the existing results on random coin tossing for the case of many biased coins. Hart and Matzinger [5] solve part of the reconstruction problem for a two color scenery seen along a random walk with bounded jumps when the observations are error corrupted.

Periodic scenery reconstruction Lember and Matzinger considered the problem of a periodic scenery seen along a random walk with jumps. This problem originated in a question by Benjamini and Kesten. The techniques used for this situation is very different from all other cases. It is related to the work of Levin and Peres [17]. They consider a scenery with only a finite numbers of one's. Furthermore, they take the observations to be error corrupted.

Scenery reconstruction in two dimensions. In [19], Löwe and Matzinger proved that sceneries in two dimensions can be reconstructed provided there are sufficiently many colors. Most researchers working on related problems, were surprised that scenery reconstruction is possible in two dimensions. The reason for this is the recurrence behavior of the random walk.

The scenery-reconstruction reading guide First we recommend the overview article by Kesten [13]. Then, we highly recommend all the articles on related topics which we cite, (as well as those which we might have forgotten).
The reader interested specifically in scenery reconstruction should probably start with the present review article. The two articles [23] and [24] for reconstrution with a simple random walk (the combinatorial case) are relatively easy and self contained. For the simple random walk with holding and a 2-color scenery (the semi-combinatorial case), there is no easy paper. For the purely statistical case, we advice to start with the simplified example at the beginning of [15]. For the purely statistical case, it might also be interesting to read [**?**]. Eventually we recommend the first section of [21]. This article is very rigorous, and the general structure of the paper can be used in many other contexts. Some aspects of the 2-dimensional reconstruction are nicely explained in [19] and should not be too difficult.

4 Basics

In this section, we explain some basic ideas and steps behind any scenery reconstruction approach: Constructing a finite piece of scenery, assembling the words, working with the stopping times, and a zero-one law.

4.1 What means "to reconstruct a finite piece"?

Every scenery reconstruction is based on some algorithms that reconstruct finite pieces of the scenery. In this subsection, we give an easy numerical example to explain what is meant by "reconstructing a finite piece of scenery". This simple example does not convey an idea on how the methods in more difficult situations works.

Let S designate a simple random walk starting at the origin. I want to test the scenery reconstruction ability of my friend Mr. Scenery Reconstruction (Mr. S.R.). For this I flip a fair coin several times in

order to create a 2-color scenery ξ (or at least a finite portion of it). Here is what I obtain:

$\xi(z)$	$\dots$	0	1	0	0	1	1	0	0	$\dots$
z	$\cdots$	-2	-1	0	1	2	3	4	5	$\dots$

Then, I flip a coin several times to determine the path of the simple random walk S. I obtain:

$S(t)$	$\dots$	0	-1	0	1	0	1	2	3	4	5	4	$\dots$
t	$\cdots$	0	1	2	3	4	5	6	7	8	9	10	$\dots$

The observations made by the random walker are:

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

I decide to give Mr. S.R. only the first 11 observations: $(0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0)$. What can he do with this ? Of course, I know ξ and the path of S but he does not. He knows however that S is a simple random walk starting at the origin. After several hour of thinking, Mr. S.R. comes back and tells me: the finite piece of scenery (i.e. binary word) 001100 is contained in the scenery ξ. My answer to him is: your statement is trivially correct: since the scenery ξ is i.i.d. every finite pattern will appear infinitely often in ξ, thus also 001100. Mr. S.R. sits back over his problem and after another hour of thinking makes the following statement: the finite piece 001100 is contained in the scenery ξ in the interval $[-9, 9]$. (This means that 001100 is a sub-word of the word $\xi(-9)\xi(-8)\ldots\xi(8)\xi(9)$.)
The way he could reach his conclusion is the following: a simple random walk can only generate the pattern 001100 if it walks in a straight way over the pattern 001100 in the scenery. ("Straight way" means: taking only steps in one direction.) But in the observations between time $t = 4$ and $t = 9$, we see the pattern 001100. Hence during that time, the random walk goes in a straight way over a place in the scenery ξ where that pattern appears. Now, up to time $t = 9$, the random walk remains in the interval $[-9, 9]$. It follows that the pattern 001100 appears in $[-9, 9]$.

The goal of the finite piece reconstruction is to construct a piece which is located in some given interval around the origin. The exact location of a finite piece cannot, in general, be determined from χ.

4.2 How to reconstruct a finite piece of ξ?

Again, we start with a simplified example. The main idea here however is important and appears often in more complicated settings.
Assume for a moment that instead of being a two color scenery, ξ would be a four color scenery, i.e. $\xi : \mathbb{Z} \longrightarrow \{0, 1, 2, 3\}$. Let us imagine furthermore, that there are two integers x, y such that $\xi(x) = 2$ and $\xi(y) = 3$, but outside x and y the scenery has everywhere color 0 or 1, (i.e. for all $z \in \mathbb{Z}$ with $z \neq x, y$ we have that $\xi(z) \in \{0, 1\}$.) The simple random walk $\{S(k)\}_{k\geq 0}$ can go with each step one unit to the right or one unit to the left. This implies that the shortest possible time for the random walk $\{S(k)\}_{k\geq 0}$ to go from the point x to the point y is $|x - y|$. When the random walk $\{S(k)\}_{k\geq 0}$ goes in shortest possible time from x to y, it goes in a straight way, which means that between the time it is at x and until it reaches y, it only moves in one direction. During that time, the random walk $\{S(k)\}_{k\geq 0}$ reveals the portion of ξ lying between x and y. So, if

between time t_1 and t_2 the random walk goes in a straight way from x to y, (that is if $|t_1 - t_2| = |x - y|$ and $S(t_1) = x, S(t_2) = y$), then the word $\chi(t_1)\chi(t_1+1)\dots\chi(t_2)$ is equal to the word $\xi(x)\xi(x+u)\xi(x+2u)\dots\xi(y)$, where $u := (y-x)/|y-x|$. Since the random walk $\{S(k)\}_{k\geq 0}$ is recurrent, it goes at least once in the shortest possible way from the point x to the point y, a.s.. Because we are given infinitely many observations, we can then (a.s.) figure out the distance between x and y. Indeed, the distance between x and y is the shortest time laps that a "3" will ever appear in the observations χ after a "2". When, on the other hand, a "3" appears in the observations χ in shortest possible time after a "2", then between the time we see that "2" and until we see the next "3", we observe a copy of $\xi(x)\xi(x+u)\xi(x+2u)\dots\xi(y)$ in the observations χ. This fact allows us to reconstruct the finite piece $\xi(x)\xi(x+u)\xi(x+2u)\dots\xi(y)$ of the scenery: Choose any couple of integers t_1, t_2 with $t_2 > t_1$, minimizing $|t_2 - t_1|$ under the condition that $\chi(t_1) = 2$ and $\chi(t_2) = 3$. Then $\chi(t_1)\chi(t_1+1)\dots\chi(t_2)$ is equal to $\xi(x)\xi(x+u)\xi(x+2u)\dots\xi(y)$, a.s..

Let the scenery ξ be such that: $\xi(-2) = 0$, $\xi(-1) = 2$, $\xi(0) = 0$, $\xi(1) = 1$, $\xi(2) = 1$, $\xi(3) = 3$, $\xi(4) = 0$. Assume furthermore that the scenery ξ has a 2 and a 3 nowhere else then in the points -1 and 3. Imagine that χ the observation given to us would start as follows:

$$\chi = (0, 2, 0, 1, 0, 1, 3, 0, 3, 1, 1, 1, 1, 0, 2, 0, 1, 1, 3, \dots)$$

By looking at all of χ we would see that the shortest time a 3 occurs after a 2 in the observations is 4. In the first observations given above there is however already a 3 only four time units after a 2. The binary word appearing in that place , between the 2 and the 3 is 011. We deduce from this that between the place of the 2 and the 3 the scenery must look like: 011.

In reality, the scenery we want to reconstruct is i.i.d. and does not have a 2 and a 3 occurring in only one place . So, instead of the 2 and the 3 in the example above, we will use a special pattern in the observations which will tell us when the random walk is back at the same spot. One possibility (although not yet the one we will eventually use) would be to use binary words of the form: 001100 and 110011. As mentioned in the previous subsection, the only possibility for the word 001100, resp. 110011 to appear in the observations, is when the same word 001100, resp. 110011 occurs in the scenery and the random walk reads it. So, imagine (to give another example of a simplified case) the scenery would be such that in a place x there occurs the word 001100, and in the place y there occurs the word 110011 , but these two words occur in no other place in the scenery. These words can then be used as markers: Consider the place in the observations, where the word 110011 occurs in shortest time after the word 001100. In that place in the observations we see a copy of the piece of the scenery ξ comprised between 110011 and 001100. The very last simplified example is unrealistic in at least two reasons. At first, the scenery is an outcome of an i.i.d. random scenery. Thus, any word will occur infinitely often in the scenery. Secondly, if the random walk S is with holding, it can generate any pattern in very many places (for every z such that $\xi(z) \neq \xi(z+1)$).
So, the simple markers described above are not suitable for practice, and we use more sophisticated markers instead. Moreover, in most cases, these markers have to be subtle "localization tests". The techniques use to build efficient markers depend heavily on the

number of colors of ξ and the distribution of S, and they differ very much. The nature of the marker technique basically determines the nature and approach of the scenery reconstruction as explained in Subsection 3.3.

4.3 Assembling pieces of scenery

Most scenery reconstruction algorithms work as follows: They first reconstruct an increasing sequence of finite pieces of ξ. Then, these finite pieces are assembled. The limit when the size of the pieces goes to infinity is a scenery which one proves to be equivalent to ξ. In the previous two subsections, we briefly explained the problem of reconstructing of a finite piece of scenery. In the present subsection, we try to explain the basics of the assembling.
For the assembling to work, each piece needs to be contained in only one place in the next piece. (Or at least this should hold a.s. for all but a finite number of pieces). Let us explain what is meant by "contained in only one place":
Let v and w be two finite words (sequences). We say that v occurs in a unique place in w and write $v \preccurlyeq_1 w$ if there is exactly one subword of w equal to v or v^t. (The transpose of v is designated by v^t.)

I give Mr S.R. more observations. Using some advanced reconstruction skills of his, he reconstructs two additional finite pieces of ξ. He proudly shows them to me:

$$v^2 = 101001100$$

and

$$v^3 = 010100110010.$$

(Let v^1 the first piece of scenery Mr. S.R. has reconstructed. Hence $v^1 := 001100$). Mr. S.R. notices that in each of these pieces the previous one occurs only in one place. He uses this to assemble his pieces. He first puts down v^1, then v^2, then v^3. Every time he places the next word over the previous one so that they coincide. Here is what Mr. S.R. gets after placing down the word v^1:

							0	0	1	1	0	0					
...	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	...

After placing the word v_2 "over v_1" Mr. S.R. obtains:

				1	0	1	0	0	1	1	0	0					
...	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	...

After putting v^3 over v^2:

			0	1	0	1	0	0	1	1	0	0	1	0			
...	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	...

To reconstruct the whole scenery ξ up to equivalence, Mr. S.R. has to keep reconstructing bigger and bigger words and assemble them. The end result after infinite time should be a scenery equivalent to ξ. For this he needs (among others) the pieces to be contained in one another in a unique place.

Let v^m designate the m-th finite piece reconstructed. Our assembling procedure yields a.s. as limit a scenery equivalent to ξ, if there exists $i_m, m \in \mathbb{N}$ a positive increasing sequence such that

$$\lim_{m\to\infty} i_m = +\infty$$

and such that all the three following conditions are satisfied for all but (possibly) a finite number of m's:

- The piece v^m is contained in a unique place in v^{m+1}:

$$v^m \preccurlyeq_1 v^{m+1}. \tag{4.1}$$

- The piece v^m is a piece of ξ located close to the origin. More precisely:

$$v^m \quad \preccurlyeq_1 \quad \xi(-i_m)\xi(-i_m+1)\dots\xi(i_m). \tag{4.2}$$

- The pieces v^m have to become "larger and larger", so as to cover the whole scenery ξ in the end. This condition can be expressed as follows:

$$\xi(-i_m)\xi(-i_m+1)\dots\xi(i_m) \preccurlyeq_1 \quad v^{m+1} \quad . \tag{4.3}$$

Hence, the problem of reconstructing ξ up to equivalence is reduced to constructing a sequence of finite pieces satisfying (4.1), (4.2) and (4.3):

> **Problem of reconstructing a sequence of finite piece of ξ:** Find a sequence of algorithms $\mathcal{A}^1, \mathcal{A}^2, \dots, \mathcal{A}^m, \dots$ such that: if v^m designates the finite piece of scenery (word) reconstructed by $\mathcal{A}^m$, then conditions (4.1), (4.2) and (4.3) hold a.s. for all but a finite number of m's.

4.4 Stopping times

Another important problem for scenery reconstruction is to develop statistical tests to find out when the random walk is close to the origin. If we know when the random walk is in the vicinity of the origin, we can use this information to reconstruct a finite piece of ξ close to the origin. If we are not able to determine when the random walk is close to the origin, then the finite pieces which we reconstruct might be located far away form the origin. This would imply that (4.2) is violated, and this might lead to a failure in reconstructing ξ.

I decide to play on with Mr. S.R. I use the same scenery ξ, but determine a new path of S. Also this time I decide to give him 100 observations, instead of just 11. The random walk goes between time $t = 95$ and $t = 100$ from point 0 to point 5. Hence the observations during the time interval $[95, 100]$ are

$$\chi(95)\chi(96)\chi(97)\chi(98)\chi(99)\chi(100) = 001100.$$

Assume that the pattern 001100 does not occur in the observations before time $t = 95$. Mr. S.R. deduces that the word 001100 appears in the scenery ξ within the interval $[-100, 100]$. This is a rather small amount of information: The pattern 001100 has six digits. With our i.i.d. scenery made of Bernoulli variables with parameter 0.5, this pattern has a probability of $(1/2)^6 = 1/64$. Therefore, the probability that 001100 appears somewhere in a given interval of length 200 is thus rather large.
I decide to help Mr. S.R. by giving him extra information. I tell him that at time $t = 40$, 66, and 95, the random walk is at the origin. (Hence $S(40) = S(66) = S(95) = 0$.) After receiving this information Mr. S.R. deduces that the word 001100 appears in ξ in the interval $[-5, +5]$. Assume that I would have given him, less information. I could say to him that at time $\tau_1 = 40$, $\tau_2 = 66$, and $\tau_3 = 95$ the random walk is in the interval $[-8, 8]$. (Hence $S(40), S(66), S(95) \in [-8, 8]$.) From this, he could have deduced that the word 001100 appears somewhere in ξ in the interval $[-13, 13]$.

We see how useful it is to have some information about the times when the random walk stays close to the origin. In fact, it is even necessary to determine times when the random walk is close to the origin, in order to reconstruct a finite piece of scenery. Let us explain this with the help of an example. Assume that one tries to reconstruct a piece of ξ located in the interval $[-5, 5]$ using the first hundred observations of χ only. It is very likely that the random walk spends most time before $t = 100$, outside the interval $[-5, 5]$. Since the scenery ξ is i.i.d., the observations made outside $[-5, 5]$ do not contain any information about ξ inside $[-5, 5]$. Hence to reconstruct some information about the piece of scenery $\xi|[-5, 5]$, we need to be able to determine when the random walk stays in the interval $[-5, 5]$.
In many scenery papers, the problem of reconstructing of a finite piece of scenery is decomposed into two sub-problems:

1. The problem of determining from observations the times τ_i indicating when the random walk is close to the origin. We require that the decision that at time t, the random walk is close to the origin, depends only on the observations $\chi(1), \ldots, \chi(t)$, only. Hence, the times τ_i are σ_χ-adapted stopping times, where σ_χ stands for the filtration $\sigma_\chi := \cup_{i=0}^{\infty} \sigma(\chi(0), \chi(1), \ldots, \chi(i))$.

2. The problem of reconstructing a finite piece of ξ located close to the origin with the help of χ and the stopping times τ_i (the additional information about the times when the random walk is close to the origin).

In a previous Mr. S.R. example, we used the fact that the random walk had performed after the time τ_3 a "straight walk" of 6 steps. This idea will be used in actual scenery reconstruction: We make sure that having enough stopping times τ_i, with high probability, some of them will be followed by a little piece of straight walk.

4.5 Solving the stopping-time-problem

In the previous subsection, we saw that the reconstruction of a finite piece of scenery can be decomposed into two parts:

1. Construct an increasing sequence of σ_χ-adapted stopping times τ_i which all "stop" the random walk close to the origin.

2. Using χ and the τ_i-s, in order to reconstruct a finite pieces of ξ close to the origin.

In the early papers, the stopping time problem is often the more difficult one. An important progress was made in [21], where it was showed that the stopping time problem can actually be solved with the solution to the second problem. This seems very strange at first sight, since the second problem is to reconstruct a piece of ξ with the help of stopping times. However, we can discover when the random walk is close to the origin, using a reconstruction algorithm for a finite piece of ξ. And paradoxically, the reconstruction algorithm used to construct the stopping times, requires itself to be feed with stopping times.

Let us go back to the Mr. S.R. example to see how a reconstruction algorithm can be used to tell us when the random walk is close to the origin. Recall that Mr. S.R. using only the hundred first observations $\chi(0)\dots\chi(99)$ reconstructed the finite piece v_3:

$$v^3 = 010100110010.$$

Let us thus assume that the reconstruction algorithms $\mathcal{A}^m$ work with finite input instead of using the whole of χ. Here for example, Mr. S.R. using his algorithm $\mathcal{A}^3$ with the input $\chi(0)\dots\chi(99)$ obtains the piece of scenery v^3, hence:

$$\mathcal{A}^3\Big(\chi(0)\dots\chi(99)\Big) = v^3.$$

Let $s > 100$ be a relatively large (non-random) number. We ask Mr. S.R. if he thinks that at time s the random walk was "close" to the origin. He has only the observations χ to base his guess upon. Mr. S.R. applies the reconstruction algorithm $\mathcal{A}^3$ to the first hundred observations after time s. Imagine that he finds:

$$\mathcal{A}^3\Big(\chi(s)\chi(s+1)\dots\chi(s+100-1)\Big) = v^3.$$

He is surprised to see that he obtains the same result as when he applied the algorithm $\mathcal{A}^3$ to $\chi(0)\dots\chi(99)$. From this he deduces that with high probability

$$S(s) \in [-200, 200].$$

His reasoning goes as follows: If we would have that $S(s) \notin [-200, 200]$, then during the time interval $[s, s+100-1]$ the simple random walk S would remain outside the interval $[-100, 100]$. The observations $\chi(s)\chi(s+1)\dots\chi(s+100)$ would then only depend on $\big(\xi(z)\big)_{z\notin[-200,200]}$ and the path of S. Hence, these observations would be independent of $\big(\xi(z)\big)_{z\in[-100,100]}$. In this case, since the scenery ξ is i.i.d., $\mathcal{A}^3\Big(\chi(s)\dots\chi(s+100)\Big)$ is independent of $\big(\xi(z)\big)_{z\in[-100,100]}$. But v_3 is a piece of $\big(\xi(z)\big)_{z\in[-100,100]}$. It follows that if $S(s) \notin [-200, 200]$, then v^3 is independent of $\mathcal{A}^3\Big(\chi(s)\dots\chi(s+100-1)\Big)$ and it would be an unlikely coincidence if v^3 would be exactly equal to $\mathcal{A}^3\Big(\chi(s)\dots\chi(s+100)\Big)$. (By "if $S_s \notin [-200, 200]$", we mean conditional on $S(s) \notin [-200, 200]$.)

4.6 Zero-one law for scenery reconstruction

In [21], Matzinger introduces a zero-one law for scenery reconstruction. The exact formulation goes as follows: Assume that there is an event A depending only on the observations

χ and such that the probability to reconstruct ξ correctly given A is strictely larger than $1/2$ then the scenery can be almost surely reconstructed. In many cases this is useful because it implies that we can assume for the reconstruction that we know a finite portion of the scenery to start with.

5 The semi-combinatorial case.

In this section, we consider the case where S is a simple random walk with holding (see Subsection 3.2) and the scenery ξ has two colors. We explained in Subsection 3.2, how the combinatorial methods in this case fail. In this section, we show how to reconstruct ξ given the conditional distribution $\mathcal{L}(\chi|\xi)$. For this we use the combinatorial methods for the simple random walk and apply them to $\mathcal{L}(\chi|\xi)$.

5.1 Conditional distribution of the observed blocks

A *block* is a substring of maximal length containing only zeros or ones. For a random walk with no jumps, each observed block of χ is generated on exactly one block of ξ. Let B_i denote the i-th block of χ and let $|B_i|$ denote its length. Roughly speaking, the following holds:
If B_i is generated on a block of ξ of length m, then $|B_i|$ is distributed like the first hitting time of $\{-1, m\}$ by the random walk S. (Recall that S starts at the origin).

Let us look at a numerical example. Let χ be

$$\chi(0)\chi(1)\chi(2)\chi(3)\chi(4)\chi(5)\chi(6)\chi(7)\ldots = 00111001\ldots.$$

We adopt the following convention: the first bits of χ, which are equal to each other, are not counted as a block. (This convention is made to simplify notations later). In our present example, this means that $\chi(0)\chi(1)$ does not count as a block. Hence, the first block B_1 of χ consists of the first three ones which come directly after $\chi(0)\chi(1)$:

$$00\overbrace{111}^{B_1}001\ldots.$$

The block B_1 consists of three digits and is hence of length 3, so that $|B_1| = 3$. The block B_1 starts after $\chi(1)$ and ends just before $\chi(5)$. We sometimes identify a block with its start point and end point. In this example, this gives that B_1 would get identified with the pair $(1, 5)$. Since the block B_1 consists of ones, we say that is has color 1. The second block B_2 corresponds to the two zeros after B_1:

$$00111\overbrace{00}^{B_2}1\ldots.$$

The block B_2 consists of two digits and has therefore length 2, so that $|B_2| = 2$. Start point and end point are 4 and 7, so that we can identify the block B_2, with the pair: $(4, 7)$.

We define the multicolor scenery $\psi : \mathbb{Z} \to \mathbb{N}$ to be the double-infinite sequence consisting of the lengths of the blocks of ξ. (These lengths are taken in the order as they appear in ξ).

Let us take the following numerical example for ξ:

$\xi(z)$	$\dots$	0	1	0	0	1	1	0	0	0	1	$\dots$
z	$\cdots$	-2	-1	0	1	2	3	4	5	6	7	$\dots$

We numerate the blocks of ξ from left to right. The block at the origin is defined to be the 0-th block. This gives for our numerical example, that $\xi(0)\xi(1)$ is the 0-th block of ξ. (This block can also be represented as the pair $(-1, 2)$). The block $(-1, 2)$ has length 2, so that $\psi(0) = 2$. The first block to the right of the 0-th block, is the 1-st block of ξ. In this case, this is the block consisting of the two ones: $\xi(2)\xi(3)$. This block has length two, and thus $\psi(1) = 2$. The block immediately to the right of the first block of ξ, is the 2-nd block of ξ. In this example, it consist of three zeroes. Hence $\psi(2) = 3$. The block immediately to the left of the zero-th block is block number -1. Here, it consists of one 1 and has therefore length 1. It follows, that $\psi(-1) = 1$. The multicolor scenery ψ in this case is equal to:

$\psi(z)$	$\dots$	1	2	2	3	$\dots$
z	$\dots$	-1	0	1	2	$\dots$

Let D be an integer interval. A path $r : D \to \mathbb{Z}$ is called a *nearest neighbor walk* if r takes only steps of one unit. More precisely, $r : D \to \mathbb{Z}$ is a nearest neighbor walk, if and only if for all $t, t + 1 \in D$

$$r(t + 1) - r(t) \in \{-1, 1\}.$$

Let R be the nearest neighbor walk describing in which order the random walk S visits the blocks of ξ: if the t-th block visited by S is block number z of ξ, then $R(t) = z$. Since $S(0) = 0$ and we discard the first identical bits of χ, it holds Therefore $R(1) \in \{-1, 1\}$.

Assume that the beginning of the path of S is given by:

$$(S(0), S(1), S(2), S(3), S(4), S(5), S(6), S(7), S(8)) = (0, 1, 2, 2, 2, 1, 1, 2).$$

With the scenery we consider in this example of this subsection, this gives the observations:

$$\chi(0) \dots \chi(7) = 00111001.$$

Note that the first block B_1 in the observations is $\chi(2)\chi(3)\chi(4)$. This block is generated on the block number 1 of the scenery. (This means that for time $t = 2, 3, 4$ the random walk stays in the block number 1 of ξ.) Hence, $R(1) = 1$. The second block B_2 of the observations is $\chi(5)\chi(6)$. This block is generated when S is in block number 0 of ξ. Hence $R(2) = 0$.

Let $\Pi(t)$ denote the length of the block of ξ on which B_t was generated. We can describe $\Pi(t)$ to be the observation made by R of ψ at time t:

$$\Pi(t) = \psi(R(t)).$$

Hence the color record $\Pi(1), \Pi(2), \dots$ corresponds to the observations of the scenery ψ made by the nearest neighbor walk R.
Next, we determine the joint distribution of the $|B_i|$'s given ξ. For this we need a few definitions. Let T_m denote the first hitting time of $\{-1, m\}$ by the random walk S:

$T_m := \min\{k \geq 0 : S_k \in \{-1, m\}\}$. (Recall that S starts at the origin.) Let λ^m denote the infinite dimensional vector which defines the distribution of T^m:

$$\lambda^m := (P(T_m = 1), P(T_m = 2), P(T_m = 3), P(T_m = 4), \ldots).$$

Let λ_l^m and λ_r^m be the two defective distributions which decompose λ^m into two parts according to whether the random walk first hits on -1 or on m. We have:

$$\lambda_l^m := (P(T_m = 1, S_{T_m} = -1), P(T_m = 2, S_{T_m} = -1), P(T_m = 3, S_{T_m} = -1), \ldots)$$

and

$$\lambda_r^m := (P(T_m = 1, S_{T_m} = m), P(T_m = 2, S_{T_m} = m), P(T_m = 3, S_{T_m} = m), \ldots).$$

We get: $\lambda^m = \lambda_l^m + \lambda_r^m$.
The length of an observed block given that it is generated on a block of ξ of length m has distribution λ^m. When, additionally, we ask that the random walk crosses the block of ξ, the conditional defective distribution of the length of the observed block equals λ_r^m. When we ask that the random walk S enters and exists the block on the same side, the conditional defective distribution equals λ_l^m.
Roughly speaking, we have the following situation:
If the block B_t is generated on a block of ξ of length m, then $|B_t|$ has conditional distribution $\lambda_?^m$. But when ξ and R are given, then B_t is generated on a block of length $\psi(R(t))$. Hence, given ξ and R, we have that $|B_t|$ has distribution $\lambda_?^{\pi_t}$, where $\pi_t = \psi(R(t))$. Similarly, the joint conditional distributions of the $|B_t|$'s, given ξ and R, is the direct product of $\lambda_?^{\pi(t)}$, where the sequence $\pi(1), \pi(2), \ldots$ is equal to $\psi(R(1)), \psi(R(2)), \ldots$. This is the content of the next lemma.

Lemma 5.1 *Let $r : [1, k] \to \mathbb{Z}$ denote a nearest neighbor walk starting at $+1$ or -1. Then, we have that the conditional joint distribution of the lengths of the observed blocks:*

$$\mathcal{L}\big(|B_1|, |B_2|, \ldots, |B_k| \big| \xi, R = r\big)$$

is equal, up to a positive constant to

$$\lambda_{l_1}^{\pi(1)} \otimes \lambda_{l_2}^{\pi(2)} \otimes \ldots \lambda_{l_k}^{\pi(k)}$$

where for all $t \in [1, k]$, we have

$$\pi(t) = \psi(r(t))$$

and

$$l_t = \begin{cases} r & \text{if } r(t-1) \neq r(t+1), \\ l & \text{if } r(t-1) = r(t+1). \end{cases}$$

(In the lemma above, by $R = r$, we mean that $R(0)R(1)\dots R(k) = r(0)r(1)\dots r(k)$.)

Let ξ be

$\xi(z)$	…	0	1	1	0	1	0	1	0	0	1	1	1	1	0	0	1	1	0	0	…
z	…	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8	9	10	…

We have that $(0, 5)$ is the first block of ξ. This block has length 4 and hence $\psi(1) = 4$. (This block correspond to the four ones $\xi(1)\xi(2)\xi(3)\xi(4)$). The next block to the right is $(4, 7)$. This is the second block of ξ and has length 2, so that $\psi(2) = 2$. The third block of ξ consists of the two zeros $\xi(7)\xi(8)$. This block has length 2, so that $\psi(3) = 2$. The zero-th block of ξ consists of the two zeros $\xi(-1)\xi(0)$. Hence, $\psi(0) = 2$. Furthermore, we have that $\psi(-1) = 1$, since the first block to the left of block zero has length one.
Assume next that the random walk S makes the following first steps:

$S(t)$	0	−1	0	1	2	3	4	3	4	5	6	5	4	…
t	0	1	2	3	4	5	6	7	8	9	10	11	12	…

the observations in this case are:

$\chi(t)$	0	0	0	1	1	1	1	1	1	0	0	0	1	…
t	0	1	2	3	4	5	6	7	8	9	10	11	12	…

In this case the first block in the observations χ is $(2, 9)$. (Note that the first three 1's of χ do not count as a block. The block $(2, 9)$ of χ is of length 6 and of color 1. The second block of χ is of length 3 and color 0. It is the block $(8, 12)$. The first block of χ is generated by the random walk S on the block $(0, 5)$ of ξ. It is generated when the random walk S crosses $(0, 5)$. By this we mean that S enters the block $(0, 5)$ on one side and exists on the other side. The second block of χ is generated by S on the block $(4, 7)$ of ξ. But this time the block is generated in a different way: S enters on one end of $(4, 7)$ and exists on the same side.
Let B_k denote the k-th block of χ and let $|B_k|$ denote the length of that block. In our numerical example we find $B_1 = (2, 9)$ and $|B_1| = 6$. Furthermore, $B_2 = (8, 12)$ and $|B_2| = 3$.
Next we want to try to understand what the conditional distribution of $|B_1|$ is, (conditional on ξ). Let E denote the event that S hits on 1 before it hits on -2. When E holds the block B_1 is generated on $(0, 5)$. When E does not hold then B_1 is generated on $(-3, -1)$. Let T designate the first hitting time of $\{1, -2\}$ by S. Let b_k^-, resp. b_k^+ denote the left end, resp. the right end of the block B_k. Thus, B_k is equal to the block (b_k^-, b_k^+) of χ. When the scenery ξ is like in our example, we find $T = b_1^- + 1$. When furthermore E holds we get :

- $S(b_1^-) = 0$ and $S(b_1^- + 1) = 1$
- b_1^+ is the first hitting time of $\{0, 5\}$ by S after time $b_1^- + 1$. Thus, in this case $b_1^+ = \min\{t \geq b_1^- + 1 | S(t) \in \{0, 5\}\}$.

This implies that the conditional distribution of $|B_1| = b_1^+ - b_1^- - 1$ given the scenery ξ and conditional on E is like the distribution of the first exit time by S of an interval of length 5. This conditional distribution in the case of our example, equals:

$$\mathcal{L}\left(|B_1| \,||\, \xi, E\right) = \lambda^4.$$

We see that the conditional distribution of the length of an block of χ is λ^m. In this case, m designates the length of the block of ξ on which the block of χ was generated.
What is the conditional joint distribution of the $|B_k|$'s given ξ? Again we look at the case when the scenery ξ is like our numerical example. In this case the random walk S can for example cross the block $(0, 5)$ then the block $(4, 7)$ and finally the block $(6, 9)$. If the random walk crosses these blocks in the

above mentioned manner and order, the joint conditional distribution for $|B_1|, |B_2|, |B_3|$ is proportional to $\lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_r^2$. More precisely, if the scenery ξ is like in our example, we get:

$$\mathcal{L}\left(|B_1|, |B_2|, |B_3|| \, \xi, S(b_1^+) = 5, S(b_2^+) = 7, S(b_3^+) = 9\right) = \frac{1}{|\lambda_r^4||\lambda_r^2||\lambda_r^2|} \lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_r^2.$$

(Here $|\lambda_r^m| := \sum_{i=1}^{\infty} P(T_m = i, S_{T_m} = m)$.) Another possibility for the random walk S is to first cross the block $(0, 5)$ then the block $(4, 7)$ and finally enter the block $(6, 9)$ and exit on the same side. The conditional joint distribution of $|B_1|, |B_2|, |B_3|$ in this case is proportional to $\lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_l^2$. Then

$$\mathcal{L}\left(|B_1|, |B_2|, |B_3|| \, \xi, S(b_1^+) = 5, S(b_2^+) = 7, S(b_3^+) = 6\right) = \frac{1}{|\lambda_r^4||\lambda_r^2||\lambda_l^2|} \lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_l^2.$$

Yet another possibility for the random walk S would be to first cross the block $(0, 5)$ then enter the block $(4, 7)$ and exit on the same side and finally cross the block $(0, 5)$ from right to left. In this case, the conditional joint distribution of $|B_1|, |B_2|, |B_3|$ is proportional to $\lambda_r^4 \otimes \lambda_l^2 \otimes \lambda_r^4$. Then

$$\mathcal{L}\left(|B_1|, |B_2|, |B_3|| \, \xi, S(b_1^+) = 5, S(b_2^+) = 4, S(b_3^+) = 0\right) = \frac{1}{|\lambda_r^4||\lambda_l^2||\lambda_r^4|} \lambda_r^4 \otimes \lambda_l^2 \otimes \lambda_r^4.$$

The random walk can also choose to first visit the block to the left of zero. For example it could first cross from right to left the block $(-3, -1)$, then cross the block $(-4, -2)$ and finally cross also from right to left the block $(-5, -3)$. In this case the defective conditional joint distribution of $|B_1|, |B_2|, |B_3|$ is $\lambda_r^1 \otimes \lambda_r^1 \otimes \lambda_r^1$. There are many other possibilities. In total for the first three block crossed there are $2(2^3)$ different possibilities. The joint conditional distribution of $|B_1|, |B_2|, |B_3|$ is obtained by adding the defective distributions for all these different cases. The sum can be decomposed into two groups of cases: the cases which start with $S(b_1^-) = 0$ and those which start with $S(b_1^-) = -1$. For the numerical example considered here, this gives

$$\begin{aligned}\mathcal{L}\left(|B_1|, |B_2|, |B_3| \, \middle| \xi\right) = P(S(b_1^-) = 0) \cdot \left(\lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_r^2 + \lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_l^2 + \lambda_r^4 \otimes \lambda_l^2 \otimes \lambda_r^4 + \ldots\right) \\ + P(S(b_1^-) = -1) \cdot \left(\lambda_r^1 \otimes \lambda_r^1 \otimes \lambda_r^1 + \lambda_r^1 \otimes \lambda_r^1 \otimes \lambda_l^1 + \ldots\right)\end{aligned}$$

With the scenery ξ of this example, ψ is equal to:

$\psi(z)$	$\ldots$	1	2	4	2	2	$\ldots$
z	$\ldots$	-1	0	1	2	3	$\ldots$

Let us analyze the different terms in the last sum above. The first term is

$$\lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_r^2. \tag{5.1}$$

This correspond to when R walks in a straight way from point 1 to 3, and hence $R(1) = 1, R(2) = 2, R(3) = 3$. In other words, the random walk S visits first block one of ξ, then block 2 before block 3. The sequence of superscripts of (5.1) is $4, 2, 2$. This corresponds to the length of the blocks visited. In this case,

$$(4, 2, 2) = (\psi(1), \psi(2), \psi(3)).$$

Take now the second term of the sum of conditional joint distribution of the $|B_i|$'s. This term is

$$\lambda_r^4 \otimes \lambda_r^2 \otimes \lambda_l^2. \tag{5.2}$$

It corresponds to the random walk visiting first block 1 of ξ, then block 2 before going back to block 1. Hence, this corresponds to $R(1) = 1, R(2) = 2, R(3) = 1$. The sequence of superscripts of expression (5.2) is the sequence of the lengths of the blocks visited by S. In this case the sequence $4, 2, 2$, is equal

to $\psi(R(1)), \psi(R(2)), \psi(R(3)$ for $R(1)R(2)R(3) = 121$.
In the sum of the conditional joint distribution of the $|B_i|$'s, we observe that each term corresponds to one possibility for the nearest neighbor walk R. So, we have to consider all nearest neighbor walk paths starting at 1 or -1. Each one gives one term in our sum. The connection between the terms and the corresponding nearest neighbor walks is the following: The sequence of superscripts of the term is equal to the observations of ψ made by the nearest neighbor walk. (By observations of ψ made by r, we mean $\psi \circ r$).

Let us now formulate the essence of the last example. For this, some notations are needed.
Let $\mathcal{R}^k$ denote the set of all nearest neighbor walks $R : [1, k] \to \mathbb{Z}$ such that $R(1) \in \{1, -1\}$. Let $\mathcal{R}^k_+$, resp. $\mathcal{R}^k_-$ denote the set of all nearest neighbor walks in $\mathcal{R}^k$ starting at 1, resp. at -1.
Let $r \in \mathcal{R}^k$. We denote by π_r the observations made by r of ψ. Thus, $\forall t \in [1, k]$,

$$\pi_r(t) = \psi(r(t)).$$

Let $l_r(t)$ be the variable which describes if the nearest neighbor walk at time t moves back to the point where it was at time $t - 1$. More precisely, $l_r(t) = l$ if $r(t - 1) = r(t + 1)$ and $l_r(t) = r$ if $r(t - 1) \neq r(t + 1)$. (Note that we use the same symbol r, for two very different things: one is the nearest neighbor walk r. The other thing is just a symbol, which tells us when the random walk S exits a block on the opposite side than where it entered.) With this notation, we formalize the conditional distribution of the lengths of the first k blocks.

Theorem 5.1 *The joint conditional distribution*

$$\mathcal{L}\left(|B_1|, |B_2|, \ldots, |B_k| \mid \xi\right)$$

is equal to the sum

$$p_1 \sum_{r \in \mathcal{R}^{k+1}_+} \lambda^{\pi_r(1)}_{l_r(1)} \otimes \lambda^{\pi_r(2)}_{l_r(2)} \otimes \ldots \otimes \lambda^{\pi_r(k)}_{l_r(k)} \quad + \quad p_{-1} \sum_{r \in \mathcal{R}^{k+1}_-} \lambda^{\pi_r(1)}_{l_r(1)} \otimes \lambda^{\pi_r(2)}_{l_r(2)} \otimes \ldots \otimes \lambda^{\pi_r(k)}_{l_r(k)} \tag{5.3}$$

where

$$p_1 := P(R(1) = 1), \quad p_{-1} := P(R(1) = -1).$$

5.2 Reconstruction of ξ from $\mathcal{L}(\chi|\xi)$

We want to reconstruct ξ given $\mathcal{L}(\chi|\xi)$, only.
Let W^k be the set of all possible observations of ψ made by a nearest neighbor walk belonging to $\mathcal{R}^k$:

$$W^k := \left\{\pi_r | r \in \mathcal{R}^k\right\}.$$

Let $W := \cup_{k=1}^{\infty} W^k$.
Assume first that all the distribution-vectors λ^m_l and λ^m_r for $m \in \mathbb{N}$, are linearly independent of each other. (This is not exactly the case, let us first imagine it.) We could linearly

decompose $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \, |\xi)$ and find each term in the sum given in Theorem 5.1. Each term in the sum is a direct product of distributions λ_l^m and λ_r^m. For a given term, taking the sequence of superscripts yields π_r. Hence, we would be able to determine all the possible observations π_r of ψ made by a nearest neighbor walk $r \in \mathcal{R}^k$. In other words, we would obtain the set W^k.
In reality, not all distributions λ_l^m and λ_r^m for $m \in \mathbb{N}$, are independent of each other. However many of them are. So, it is still possible determine all the terms in the sum of Theorem 5.1 using the distributions $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \, |\xi)$ alone. For this, one uses linear decomposition along some subspaces and heavy combinatorics. To explain the details, ugly and complicated notations are needed. This one of the reasons, why there are no easily accessible papers about the reconstruction for a random walk with holding.
Next, note that the set W is the set of all possible observations of ψ made by a nearest neighbor walk belonging to $\mathcal{R}^k$. The nearest neighbor walks in $\mathcal{R}^k$ are without holding. Therefore, we can apply the techniques for the simple random walk, which were presented in Subsection 3.1. Of course, in Subsection 3.1, we considered the case, where we have only one realization of the observations. Here, W consists of all possible observations of ψ by a $r \in \mathcal{R}^k$. However, if instead of one observation, the set of all possible observations is given, the reconstruction becomes much easier. Hence, with the techniques described in Subsection 3.1, it is easy to reconstruct ψ up to equivalence from the set W.
Since S starts at the origin, ψ and $\chi(0)$ determine ξ up to equivalence.

An example to show how to reconstruct ξ up to equivalence from ψ and $\chi(0)$. Assume

$$\psi(-1) = 3, \psi(0) = 2, \psi(1) = 1, \psi(2) = 3$$

and $\chi(0) = 0$. This then implies, that the scenery ξ around the origin looks like this:

$$\ldots 01110010001 \ldots$$

Next, we describe the algorithm to reconstruct ξ if we are given $\mathcal{L}(\chi|\xi)$.

Algorithm 5.1

1. *Decompose*
$$\mathcal{L}(|B_1|, \ldots, |B_k||\xi)$$
linearly and use combinatorics to obtain the set W^k. Do this for every $k \in \mathbb{N}$.

2. *Use the combinatorial methods for the simple random walk, to reconstruct ψ from the set of observations W.*

3. *From ψ and $\chi(0)$ reconstruct ξ up to equivalence.*

5.3 Approximation of $\mathcal{L}(\chi|\xi)$

In the previous subsection, we explained how to reconstruct ξ provided, we are given $\mathcal{L}(\chi|\xi)$. It remains to explain, how to obtain $\mathcal{L}(\chi|\xi)$. To begin with, assume that on top

of the observations χ, we are given an increasing sequence $(\tau_i)_{i\in\mathbb{N}}$ of σ_χ-adapted stopping times. Let each of these stopping times stop the random walk at the origin: $S(\tau_i) = 0$ for all $i \in \mathbb{N}$. Then, because of the strong Markov property of S, we have that the empirical distribution of the first k observations after τ_i, converges to

$$\mathcal{L}(\chi(0)\chi(1)\dots\chi(k-1)|\xi),$$

as the number of stopping times goes to infinity. In other words, the empirical distribution of the color records

$$\chi(\tau_i)\chi(\tau_i+1)\dots\chi(\tau_i+k-1)$$

where $i \in [1, n]$, converges to $\mathcal{L}(\chi(0)\chi(1)\dots\chi(k-1)|\xi)$ as $n \to \infty$.
In general, it is not possible to construct many σ_χ-adapted stopping times which all tell when S is exactly at the origin. It is only possible to construct stopping times which all stop S in a given interval I. If we then take the empirical distribution of the observations after the stopping time τ_i, this will not be an approximation of $\mathcal{L}(\chi(0)\chi(1)\dots\chi(k-1)|\xi)$. Rather, it will be an approximation of the mixture

$$\sum_{z\in I} a_z \mathcal{L}(\chi_z|\xi), \tag{5.4}$$

where a_z denotes the proportion of stopping times for which $S(\tau_i) = z$ and χ_z denote the observations made by a random walk starting at z:

$$\chi_z := (\xi(z), \xi(z+S(1)), \xi(z+S(2)), \dots, \xi(z+S(k)).$$

Given the mixture (5.4), the reconstruction of ξ is very similar to the reconstruction with the help of $\mathcal{L}(\chi|\xi)$.
For the construction of the stopping times we refer the reader to [22, 21] or [25].

6 The purely statistical case

6.1 Why the method for the semi-combinatorial case fails in the purely statistical case

Here we explain why the method described in the previous section is impossible when the random walk is allowed to jump. Recall the definition of the random walk with jumps in (3.4). In this section, ξ is a 2-color scenery and S is a symmetric recurrent random walk with jumps (the symmetry is assumed to simplify the notation). We assume that S has bounded jump length $L < \infty$, where

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

Let x and y be two points of $\mathbb{N}$. We say that x and y are *equivalent* with respect to ξ if there exist a possible path for the random walk S going from x to y and such that we

observe only the same color during the whole trip from x to y. Formally: x and y are equivalent with respect to ξ if there exists $0 < s < t$ such that

$$P(\chi(s) = \chi(s+1) = \ldots = \chi(t), S(s) = x, S(t) = y|\xi) > 0.$$

An *island* is a maximal set of points in $\mathbb{Z}$ which are all equivalent with respect to ξ. If the random walk has no jumps, then the island is a block.

Let ξ be

$\ldots$	1	1	0	1	0	0	1	1	0	0	0	1	0	1	0	1	1	1	1	0	1	0	0	0	$\ldots$
$\cdots$	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	$\ldots$

.

Let S be such that

$$P(S(t+1) - S(t) = k) > 0, \quad \text{if and only if } k = -2, -1, 0, 1, 2.$$

Then $L = 2$. The following points are equivalent:

$$\{-4, -3, -1\}, \{-2, 0, 1\}, \{2, 3\}, \{4, 5, 6, 8, 10\}, \{7, 9, 11, 12, 13, 14, 16\}, \{15, 17, 18, 19\}$$

and the islands of this example are

$$\{-2, 0, 1\}, \{2, 3\}, \{4, 5, 6, 8, 10\}, \{7, 9, 11, 12, 13, 14, 16\}.$$

Let S be such that

$$P(S(t+1) - S(t) = k) > 0, \quad \text{if and only if } k = -3, -2, -1, 0, 1, 2, 3.$$

So, the maximum step of S is at the length of 3, i.e. $L = 3$. Then the following points are equivalent:

$$\{-4, -3, -1, 2, 3\}, \{-2, 0, 1, 4, 5, 6, 8, 10\}, \{7, 9, 11, 12, 13, 14, 16\}, \{15, 17, 18, 19\}$$

and the only island in this example is

$$\{7, 9, 11, 12, 13, 14, 16\}.$$

Let S be such that

$$P(S(t+1) - S(t) = k) > 0, \quad \text{if and only if } k = -3, -1, 0, 1, 3.$$

So, $L = 3$, but the moves with the length 2 are not allowed. Then the following points are equivalent:

$$\{-4, -3, -1, 2, 3\}, \{-2, 0, 1, 4, 5, 6, 8\}, \{10\}, \{7\}, \{9, 11, 12, 13, 14\}, \{16\}, \{15, 17, 18, 19\}$$

and the islands of this example are:

$$\{10\}, \{7\}, \{9, 11, 12, 13, 14\}, \{16\}.$$

We see, how the islands depend on the nature of S.

If the random walk can jump, then a block B of χ is generated on an island of ξ, but not necessarily on a block of ξ. It turns out that this difference is crucial. Similarly to the case for a random walk with no jumps, the conditional distribution

$$\mathcal{L}\big(|B_1|, |B_2|, \ldots, |B_k| \ \big|\xi\big)$$

can be written as a positive linear combination of direct product of distributions. These distributions are now the conditional distributions of the length of an observed block given the underlying island. However, there are important differences to the case of a random walk with no jumps. The main differences are the following.

1. There is no explicit formula for the conditional distribution of $|B_i|$ given the island it was generated on. This conditional distribution depends on some eigenvalues for which, in general, there is no explicit formula. Recall that in the case of no jumps, there is a simple explicit expression for the distribution of $|B_i|$ given that it was generated on a block of length m.

2. A block of length m has a fixed shape. If the random walk can jump, then, for an island with m elements, there are exponentially many (in m) possible shapes. Hence, there are exponentially many possible conditional distributions for $|B_i|$.

These differences cause the reasons why the approach from the previous section does not work at all, if the random walk with can jump. The reasons are the following.

1. Scenery reconstruction is not just about reconstruction. It is also about proving that the reconstruction works. Since we use an estimation of $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \ |\xi)$ instead of $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \ |\xi)$ itself, we need to be able to bound the approximation error. In the linear decomposition, the approximation error depends on how "linearly independent" the distributions λ^m are of each other. Unfortunately, without the explicit formulas it is not possible to show how "linearly independent"the distributions λ^m are of each other. And then it is not possible to evaluate the effect of the approximation error in our decomposition.

2. In the previous section, the linear decomposition of the approximation of

$$\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \ |\xi)$$

was possible, because the number of components was relatively small. With an exponential number of distribution, it is not more possible to decompose our approximation of $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \ |\xi)$. Indeed, since there are so many possible distributions, many will be very close to each other and hence the approximation error will make it impossible to recognize which one really appear in $\mathcal{L}(|B_1|, |B_2|, \ldots, |B_k| \ |\xi)$.

The reasons above make the method form last section completely unsuitable for the random walk with jump and a fundamental new approach was needed.

6.2 How to reconstruct a small amount of information

In Section 3.1, we introduced the concept of fingerprint. A fingerprint is a transformation of a piece of observation χ_s^t that gives us certain information about the underlying piece of scenery ξ_a^b on which χ_s^t was generated. In the setup considered in Section 3.1, a fingerprint was a relatively easy defined and well understood transformation. In the setup of the present section (2-color scenery observed along a random walk with jumps), such fingerprints do not work. However, it is still possible to construct a transformation of a piece of observation that can be used as a fingerprint. The construction of such fingerprints is more complicated and, as typical to the statistical approach of the

scenery reconstruction, they reveal the desired information with certain probability, only. However, these fingerprints constitute the basis of the scenery reconstruction in this setup.

In the following, we present the fingerprint existence theorem. This is the main result of the paper [15].
Let us introduce and recall some notation. Recall that

$$\chi_0^{m^2} = \chi(0)\dots\chi(m^2), \quad \xi_0^m = \xi(0)\dots\xi(m).$$

Let $a = a_1 \dots a_N$, $b = b_1 \dots b_{N+1}$ be two words with length N and $N+1$, respectively. We write $a \sqsubseteq b$, if

$$a \in \{b_1 \dots b_N, b_2 \dots b_{N+1}\}.$$

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

Theorem 6.1 *There exists constants $c, \alpha > 0$ not depending on n such that:*
For every $n > 0$ big enough, there exist an integer $m(n)$ satisfying

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

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 $E^n_{\text{cell\_OK}} \in \sigma(\xi(z) | z \in [-cm, cm])$ such that all the following holds:

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

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

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

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

$$\left\{\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\right\}$$

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, 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, given that $\xi \in E^n_{\text{cell\_OK}}$, the probability $P\big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\big)$ is about $P(S(m^2) = m)$ which is of order $\frac{1}{m} \geq e^{-2n}$. Although, for big n, this difference is negligible, it can be still used to make the scenery reconstruction possible.

Theorem 6.1 gives us a lower bound to the mutual information $I\big(\xi_0^m); \chi_0^{m^2}\big)$. Indeed, from Fano's inequality follows that, for n big enough, there exists an $\beta > 0$ such that

$$I\big(\xi_0^m); \chi_0^{m^2}\big) \geq I\big(\xi_0^m); \hat{g}(\chi_0^{m^2})\big) \geq H(\xi_0^m) - 1 - (1 - P_{\sqsubseteq}) \log(2^m - 1) \geq \beta \ln m,$$

where

$$P_{\sqsubseteq} = P\Big(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m)\Big).$$

6.3 3-color example

In this subsection, we solve the fingerprint problem in a simplified 3-color case. This example gives the main ideas behind Theorem 6.1.

6.3.1 Setup

Recall that we want to construct two functions

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

such that

1) with high probability

$$P\left(\hat{g}(\chi_0^{m^2}) \sqsubseteq g(\xi_0^m) \,\middle|\, S(m^2) = m\right).$$

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

In other words, **1)** states 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) $\exp(n/\ln n) \le 1/P(\xi(0) = 2) \le \exp(n)$,

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

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

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

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

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:

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

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:

$$h(i) := \sum_{\substack{z \in [-Le^{n^{0.1}}, Le^{n^{0.1}}] \\ z \ne \bar{z}_i}} \xi(z) P\big(S(e^{n^{0.1}}) + \bar{z}_i = z\big) \tag{6.1}$$

(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 $E^n_{6\_2}$ 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 $E^n_{6\_2}$ 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 $P(E^n_{6\_2}) \to 1$ as $n \to \infty$.

- Let $E^n_{1\_2}$ be the event that in $[0,m]$ there are more than n^2+1 signal carrier points. Because of the definition of m, $P(E^n_{1\_2}) \to 1$ as $n \to \infty$.

When $E^n_{1\_2}$ and $E^n_{6\_2}$ both hold, we define $g(\xi^m_0)$ in the following way:

$$g\left(\xi^m_0\right) := \left(g_1\left(\xi^m_0\right), g_2\left(\xi^m_0\right), g_3\left(\xi^m_0\right), \ldots, g_{n^2+1}\left(\xi^m_0\right)\right)$$

Conditional on $E^n_{1\_2} \cap E^n_{6\_2}$ 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, $P(g_i\left(\xi^m_i\right)=1) = 1/2 - P(h(i)=1/2)$, but we disregard this small error term in this example] and the independence follows from the fact that the scenery is i.i.d. [indeed, $g_i(\xi^m_0)$ depends only on the scenery in a radius $Le^{n^{0.1}}$ of the point $\bar{z}_i$ and, due to E_{6_2}, the points $\bar{z}_i$ are further apart than $\exp(\frac{n}{2\ln n}) > L\exp(n^{0.1})$]. Hence, with almost no effort we get that when $E^n_{1\_2}$ and $E^n_{6\_2}$ both hold, then condition **2)** is satisfied. To be complete, we have to define the function g such that **2)** holds also outside $E^n_{1\_2} \cap E^n_{6\_2}$. We actually are not interested in g outside $E^n_{1\_2} \cap E^n_{6\_2}$ - it would be enough that we reconstruct g on $E^n_{1\_2} \cap E^n_{6\_2}$. Therefore, extend g in any possible way, so that $g\left(\xi^m_0\right)$ depends only on ξ^m_0 and its component are i.i.d.

6.3.2 $\hat{g}$-algorithm

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

$$P\Big(\hat{g}(\chi^{m^2}_0) \sqsubseteq g(\xi^m_0) | S(m^2) = m\Big) \tag{6.2}$$

is also high. Note, when the scenery ξ is fixed, then the probability (6.2) depends on S. 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_2} 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 $E^n_{6\_2} \cap E_{1\_2}$ 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,

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

Hence, with high probability, $\hat{h}_T$ is a precise estimate for $h(i)$. 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 (6.1). 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

$$Var[h(i)] \leq \sum_{[-Le^{n^{0.1}}, Le^{n^{0.1}}]} \frac{1}{4}\Big(\mu_{0.2}(z)\Big)^2.$$

Since our random walk is symmetric we get that

$$\sum_{z \in [-Le^{n^{0.1}}, Le^{n^{0.1}}]} \frac{1}{4}\left(\mu_{0.2}(z)\right)^2$$

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 $h(0), h(1), \ldots, h(n^2+1)$ 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 $E^n_{3\_2}$ be the set on which both statements holds.

Define

$$E_{OK} := E^n_{1\_2} \cap E^n_{3\_2} \cap E^n_{6\_2}.$$

From now on we assume that E_{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 $E^n_{6\_2}$, 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, $j = 1, \ldots, K$. Based on the observations $\chi_0^{m^2}$ only, obtain the vector

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

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 $\pi : \{1, \ldots, K\} \to \mathbb{Z}$ - with high probability π is a skip free walk, i.e. $|\pi(j) - \pi(j+1)| \leq 1$. 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 length $e^{n^{0.3}}$ for which one can significantly estimate the frequency of one's $h(i)$. In particular, the skip-freeness implies that $\pi(1) \in \{0, 1\}$. The skip-freeness of π is proved in Theorem 5.2.
Let $\pi_* := \min\{\pi(j) : j = 1, \ldots, K\}$. Now $\pi_* \leq 1$. Let $\pi^* := \max\{\pi(j) : j = 1, \ldots, K\}$. If $S(m^2) = m$, then, by $E^n_{1\_2}$, $\pi^* > n^2$.

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

$$C_i := \{\hat{h}_{T_j} : |\hat{h}_{T_j} - \hat{h}_{T_i}| \leq 2\exp(-n^{0.12})\},\ \hat{f}_i := \frac{1}{|C_i|}\sum_{j\in C_i} \hat{h}_{T_j}, \quad i = 1, \ldots, K.$$

By $E^n_{3\_2}$, we have $5\exp(-n^{0.12}) < \exp(-n^{0.11}) < |h(i) - h(j)|$, if n is big enough. Hence, $\hat{h}_{T_j} \in C_i$ if and only if $\pi(i) = \pi(j)$. Thus, for each different i, j either $C_i = C_j$ or $C_i \cap C_j = \emptyset$. Hence, $\hat{f}_j$ is the average of all estimates of $h(\pi(j))$ and, therefore, $\hat{f}_j$ is a good estimate of $h(\pi(j))$. Obviously,

$$\hat{f}_i = \hat{f}_j \quad \text{if and only if} \quad \pi(i) = \pi(j). \tag{6.3}$$

Thus, we can denote $\hat{f}(\bar{z}_i) := \hat{f}_j$, if $\pi(j) = i$ and (6.3) implies $\hat{f}(\bar{z}_i)) \neq \hat{f}(\bar{z}_j)$, if $i \neq j$.
After phrase III we, therefore, 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 $\hat{f}(\bar{z}_1), \ldots, \hat{f}(\bar{z}_{n^2+1})$?

6.3.3 Real scenery reconstruction algorithm

We now present the so-called *real scenery reconstruction algorithm* - $\mathcal{A}^{\mathbb{R}}_n$. 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 $E^n_{1\_2}$);

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 6.1 *Let $\varkappa = (\varkappa_1, \varkappa_2, \ldots, \varkappa_K)$ be the vector of real numbers such that the number of different reals in $\varkappa$ is at least $n^2 + 1$. The vector $\varkappa$ constitutes the input for $\mathcal{A}^{\mathbb{R}}_n$.*
Define $\mathcal{R}_1 := \varkappa_1$. From here on we proceed by induction on j : once $\mathcal{R}_j$ is defined, we define $\mathcal{R}_{j+1} : \varkappa_s$, with $s := 1 + \max\{j : \varkappa_j = \mathcal{R}_j\}$. Proceed until $j = n^2 + 1$ and put

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

The idea of the algorithm is very simple: take the first element $\varkappa_1$ of $\varkappa$ and consider all elements of the input vector $\varkappa$ that are equal to $\varkappa_1$ and find the one with the biggest index (the last $\varkappa_1$). Let j_1 be this index. Then take $\varkappa_{j_1+1}$ as the first output and look for the last $\varkappa_{j_1+1}$. Let the corresponding index be j_2 and take $\varkappa_{j_2+1}$ as the second output. Proceed so n^2+1 times.
Let us proof that the algorithm $\mathcal{A}_n^{\mathbb{R}}$ works. In our case the input vector is $\hat{f} := (\hat{f}_1, \dots, \hat{f}_K)$.

Proposition 6.1 *Let $\{\hat{f}(\bar{z}_{\pi_*}), \dots, \hat{f}(\bar{z}_{\pi^*})\}$ and π satisfy A1), A2), A3), A4). Then*

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

Phase IV Apply $\mathcal{A}_n^{\mathbb{R}}$ to $\hat{f}$. Denote the output $\mathcal{A}_n^{\mathbb{R}}(\hat{f})$ by $(f_1, \dots, f_{n^2})$. By Proposition 6.1,

$$(f_1, \dots, f_n) \sqsubseteq (\hat{f}(\bar{z}_1), \dots, \hat{f}(\bar{z}_{n^2+1}). \tag{6.4}$$

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 $\hat{g}$

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

Recall that because of $E^n_{3\_2}$, with high probability all random variables $h(1), \dots, h(n^2+1)$ 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 (6.4) this means

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

Hence, when E_{OK} holds, then $\hat{g}$ is properly defined and the probability (6.2) is high. Since we are not interested in $\hat{g}$ when E_{OK} does not hold, we extend the definition of $\hat{g}$ arbitrary to E^c_{OK}.

6.4 How to reconstruct a word

In this subsection, we try to explain the main ideas behind the procedure that uses the fingerprints (as in Theorem 6.1) to reconstruct a word. This procedure is the content of the paper [**?**].

6.4.1 Ladder word selection

Recall the basic word reconstruction procedure in Subsection 4.2: We assumed there exists two special locations x and y so that we can immediately see, when S is located at x or at y. Then the shortest word between x any y in observations is ξ_x^y, a.s.. We are now considering the case when the random walk can jump. Suppose that $x < y$ and $y - x = cL$, where c is a positive integer, and L is the length of the maximal jump of S. In this case, the shortest observation word between x any y is not ξ_x^y, but r the word

$$\xi(x)\xi(x+L)\xi(x+2L)\dots\xi(x+(c-1)L)\xi(y). \tag{6.5}$$

The word (6.5) is called a *ladder word*. So, if S can jump, then the procedure described above gives us a ladder word. However, if we have sufficiently many ladder words, then the scenery can still be reconstructed.
Let us now explain the reconstruction of a ladder word a bit more precisely. Suppose that for a pair (x, y) such that $y = x + cL$, there exist integers $u, v < \infty$ and functions

$$G(\xi_y^{y+u}) =: G_y, \quad G^*(\xi_{x-u}^x) =: G_x^*$$
$$\hat{G}(w), w \in \{0,1\}^{v+1}, \quad \hat{G}^*(w), w \in \{0,1\}^{v+1}$$

such that the following hold:

1 if $S(t) \le x$, then $\hat{G}^*(\chi_{t-v}^t) = G_x^*$, if $S(t) \ge y$, then $\hat{G}(\chi_t^v) = G_y$;

2 if $S(t) > x$, then $\hat{G}^*(\chi_{t-v}^t) \ne G_x^*$, if $S(t) < y$, then $\hat{G}(\chi_t^v) \ne G_y$.

Note the difference: Instead of assuming that the element $\xi(y)$ is unique, we assume now (more realistically) that the piece of scenery ξ_y^{y+u} is somehow unique. The function G recognizes and captures the uniqueness of that piece. So, G_y can be considered as the name of the piece ξ_y^{y+u}. The function $\hat{G}$ reads the name from observations. The name G_y is assumed to be such that it cannot be read from a piece of observation χ_t^{t+v}, if by generating it, S did start left from y. The symmetric assumptions are made for x.
In this case, the reconstruction of the word (6.5) is straightforward. For each $t \ge 0$ define the observation-words

$$w^1(t) := \chi_{t-u}^t, \quad w^2(t) := \chi_t^{t+c}, \quad w^1(t) := \chi_{t+c}^{t+c+u} \tag{6.6}$$

and apply the functions $\hat{G}^*$ and $\hat{G}$ to $w^1(t)$ and $w^3(t)$, respectively. Because S recurrent, a.s. there exists a t such that $\hat{G}^*(w^1(t)) = G_x^*$ and $\hat{G}(w^3(t)) = G_y$. In particular, this implies that $S(t) \le x$ and $S(t+c) \ge y$. On the other hand, during c steps, the random walk cannot move more that cL. This is exactly the distance between x and y. Hence, the only possibility is that $S(t) = x$ and $S(t+c) = y$. In this case, $w^2(t)$ equals the ladder word (6.5).
So, under the assumptions **1** and **2**, there is a simple rule for selection a observation word w as the ladder word (6.5), and this rule works a.s. Unfortunately, the assumptions are

still very unrealistic. At first, in our setup, there is no way to construct the name reading functions $\hat{G}$ and $\hat{G}^*$ so that they read the names G_y and G^*_x for sure. It is more realistic to assume that they do it with certain probability, only. This means that the assumptions **1** hold with a positive probability. However, since the random walk is recurrent, the procedure above still holds.
Secondly, a necessary condition for the procedure above to work is that there is no $z < y$ such that $\xi_z^{z+u} = \xi_y^{y+u}$. Indeed, if there exists such a z, then the names $G_z = G(\xi_z^{z+u})$ and $G_y = G(\xi_y^{y+u})$ were equal, and $\hat{G}$ could read it from the left of y. Because any finite pattern exists infinitely often in ξ (a.s.), the condition obviously fails. Therefore, it is more realistic to assume that the word ξ_y^{y+u} is unique in a certain neighborhood, only. To apply the procedure, it is then necessary to know, when S is in the neighborhood of interest. This is done via the stopping times as explained in Subsections 4.4 and 4.5. To every stopping time τ_k corresponds a triple (w^1, w^2, w^3). If there are sufficiently many stopping times, then (with high probability), for some triples the name readers read the names.
Thirdly, the described procedure requires that we know the names $G^*_x = G(\xi^x_{x-u})$ and $G_y = G(\xi_y^{y+u})$. These names depend on unknown ξ. However, they can be read with positive probability. If we have sufficiently many word triples (w^1, w^2, w^3), then (with high probability), a certain portion of them satisfy $\hat{G}^*(w^1) = G^*_x$, $\hat{G}(w^3) = G_y$. So, there exists a pair of names, G^* and G such that the number of word triples (w^1, w^2, w^3) satisfying $\hat{G}^*(w^1) = G^*$, $\hat{G}(w^3) = G$ is above a pre-defined threshold. Unfortunately, there can be many pairs on names (G^*, G) having the same property. To choose the right pair, we ripe benefit from the condition **2**. Due to this condition, the right pair of names has an important characteristic – if $\hat{G}^*(w^1) = G^*_x$ and $\hat{G}(w^3) = G_y$, then the word w^2 is always (6.5) and hence the same. We can now formalize a more realistic but not yet definitive rule for selecting a word w^2 as the ladder word (6.5).

Simplified selection rule: The word is taken as (6.5), if there exists a pair of names G^*, G such that the following holds:

a) there exists a certain amount of triples (w^1, w^2, w^3) such that

$$\hat{G}^*(w^1) = G^*, \quad \hat{G}(w^3) = G; \tag{6.7}$$

b) for every triple (w^1, w^2, w^3) satisfying (6.7), it holds $w^2 = w$.

In [**?**], the simplified rule is modified so that the definite selection rule works with high probability, provided that we have suitable name- and name reading functions.

6.4.2 Reading the names

Let us now briefly explain the basic ideas behind the construction of the name and name reading functions. Unfortunately, we cannot find these functions so that the condition **2** were satisfied. Instead, we require that if $S(t) > x$ and $S(t) < y$, then the events

$\{\hat{G}^*(\chi^t_{t-v}) = G^*_x\}$ and $\{\hat{G}(\chi^v_t) = G_y\}$ have negligible probability comparing to the case of $S(t) \geq x$ and $S(t) \leq y$. To construct such functions, we use the fingerprints from Theorem 6.1. For several reasons, the fingerprints alone are not good enough. To get more powerful functions, we use the fingerprints iteratively. We take a l big enough, and we shall apply the functions g and $\hat{g}$ l times consecutively. Let $w = w(0)\dots w(lm) \in \{0,1\}^{lm+1}$, where m is as in Theorem 6.1. We define l sub-words, called *cells*

$$w_i = w((i-1)m)\cdots w(im), \quad i = 1,\dots,l.$$

Using the sub-words w_i, we naturally extend the definition of g to the words in $\{0,1\}^{lm+1}$

$$G : \{0,1\}^{lm+1} to \{0,1\}^{l(n^2+1)}, \quad G(w) = \big(g(w_1),\dots,g(w_l)\big).$$

Let $v = v(0)\dots v(lm^2) \in \{0,1\}^{lm^2+1}$. We define cells

$$v_i = v((i-1)m^2)\dots v(im^2), \quad i = 1,\dots,l.$$

Using the sub-words v_i, we extend the definition of $\hat{G}$ to the words in $\{0,1\}^{lm^2+1}$.

$$\hat{G} : \{0,1\}^{lm^2+1} \to \{0,1\}^{ln^2}, \quad \hat{G}(v) = (\hat{G}(v_1),\dots,\hat{G}(v_l)).$$

The functions $\hat{G}^*$ and G are defined similarly.
Finally, we have to relax the requirement $\hat{G}(v) = G(u)$. Since the name reading procedure is based on Theorem 6.1, it is natural to expect that $\hat{G}(\chi_t^{t+m^2l})$ "reads" $G(\xi_y^{y+ml})$, if the relation $\sqsubseteq$ holds cell-wise, i.e.

$$\hat{G}\big(\chi^{t+im^2}_{t+(i-1)m^2}\big) \sqsubseteq G\big(\xi^{y+im}_{y+(i-1)m}\big) \quad \text{for each} \quad i = 1,\dots,l.$$

To understand, why this definition would not work, note that Theorem 6.1 bounds the probability of the event $\hat{G}(\chi^{t+im^2}_{t+(i-1)m^2}) \sqsubseteq G(\xi^{y+im}_{y+(i-1)m})$ only if the piece of scenery $\xi^{y+(i-1)m+cm}_{y+(i-1)m-cm}$ belongs to the set $E^n_{\text{cell\_OK}}$. If this is the case, we say that the cell $\xi^{y+im}_{y+(i-1)m}$ is *OK*. Although $E^n_{\text{cell\_OK}}$ has the probability close to one, since l is big, we expect a proportion of cells not to be OK. For not OK cells, the statement **2)** of Theorem 6.1 needs not hold, and the cell-wise reproducing might fail. Hence, we relax the requirement of the full cell-wise reproducing to the requirement that the OK cells are reproduced. Whether a cell is OK or not, depends on unknown ξ. However, it can shown that for a suitable ϵ, the number of OK cells is bigger than $l(1-3\epsilon)$, provided l is big enough. Hence, the function $G(\chi_t^{t+m^2l})$ reads the name $G_y(\xi)$, if the relation $\sqsubseteq$ holds for at least $l(1-3\epsilon)$ cells. The name G^*_x is read similarly.

To get the better insight to the whole ladder word reconstruction procedure including the definite selection rule as well as the final definition of $\hat{G}$ and G functions, the reader is recommended to read the first section of [**?**].

6.4.3 Assembling the words

Having a working ladder words selection rule, with high probability, we collect many ladder words with fixed length. With high probability, again, these words can be uniquely assembled to get a ladder word of bigger length. Suppose now that we want to reconstruct a piece of scenery with the length M. For this we obviously need L different ladder words (modulo L), each of them having at least $\frac{M}{L}$ elements. Suppose we are able to construct these L ladder words. We are now faced another problem: How to assemble these different ladder words together? Since the locations of these ladder words in the original scenery ξ is obviously disjoint, to assemble them correctly, we need some more information. Like a (relatively) small piece of original scenery. In Subsection 4.3, we re-stated the scenery reconstruction problem as the problem of reconstructing an increasing sequence of finite pieces of ξ. So, for reconstructing the piece v^{m+1}, we can use the already reconstructed piece v^m as a piece of original scenery. This piece helps us to assemble the ladder words correctly. This part of the scenery reconstruction procedure is identical in different setups, and it is well explained in the first section of [21].

References

[1] Itai Benjamini and Harry Kesten. Distinguishing sceneries by observing the scenery along a random walk path. *J. Anal. Math.*, 69:97–135, 1996.

[2] Frank den Hollander and Jeff E. Steif. Mixing properties of the generalized T, T^{-1}-process. *J. Anal. Math.*, 72:165–202, 1997.

[3] W. Th. Frank den Hollander. Mixing properties for random walk in random scenery. *Ann. Probab.*, 16(4):1788–1802, 1988.

[4] Matthew Harris and Mike Keane. Random coin tossing. *Probab. Theory Related Fields*, 109(1):27–37, 1997.

[5] Andrew Hart and Heinrich Matzinger. Markers for error-corrupted observations. Preprint. Submitted., 2003.

[6] 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.

[7] C. Douglas Howard. Detecting defects in periodic scenery by random walks on $\mathbb{Z}$. *Random Structures Algorithms*, 8(1):59–74, 1996.

[8] C. Douglas Howard. Orthogonality of measures induced by random walks with scenery. *Combin. Probab. Comput.*, 5(3):247–256, 1996.

[9] C. Douglas Howard. Distinguishing certain random sceneries on $\mathbb{Z}$ via random walks. *Statist. Probab. Lett.*, 34(2):123–132, 1997.

[10] S. A. Kalikow. T, T^{-1} transformation is not loosely Bernoulli. *Ann. of Math. (2)*, 115(2):393–409, 1982.

[11] Mike Keane and W. Th. F. den Hollander. Ergodic properties of color records. *Phys. A*, 138(1-2):183–193, 1986.

[12] 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.

[13] 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.

[14] Harry Kesten and Frank Spitzer. A limit theorem related to a new class of self-similar processes. *Z. Wahrsch. Verw. Gebiete*, 50(1):5–25, 1979.

[15] Jyri Lember and Heinrich Matzinger. Information recovery from a randomly mixed up message-text, 2003. Submitted.

[16] David A. Levin, Robin Pemantle, and Yuval Peres. A phase transition in random coin tossing. *Ann. Probab.*, 29(4):1637–1669, 2001.

[17] David A. Levin and Yuval Peres. Random walks in stochastic scenery on $\mathbb{Z}$. Preprint, 2002.

[18] Elon Lindenstrauss. Indistinguishable sceneries. *Random Structures Algorithms*, 14(1):71–86, 1999.

[19] M atthias Löwe and Heinrich Matzinger. Scenery reconstruction in two dimensions with many colors. *Ann. Appl. Probab.*, 12(4):1322–1347, 2002.

[20] Matthias Löwe and Heinrich Matzinger. Reconstruction of sceneries with correlated colors. *Stochastic Processes and their Applications*, 105(2):175–210, 2003.

[21] 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.

[22] Heinrich Matzinger. *Reconstructing a 2-color scenery by observing it along a simple random walk path with holding.* PhD thesis, Cornell University, 1999.

[23] Heinrich Matzinger. Reconstructing a three-color scenery by observing it along a simple random walk path. *Random Structures Algorithms*, 15(2):196–207, 1999.

[24] Heinrich Matzinger. Reconstructing a 2-color scenery by observing it along a simple random walk path. *Random Structures Algorithms*. Accepted, 2004.

[25] Heinrich Matzinger and Silke W. W. Rolles. Finding blocks and other patterns in a random coloring of $\mathbb{Z}$. To appear in Random Structures Algorithms, 2006.

[26] Heinrich Matzinger and Silke W. W. Rolles. Retrieving random media. To appear in Probab. Theory Related Fields, 2006.

[27] 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.