On Approximate Pattern Matching for a Class of Gibbs Random Fields

J.-R. Chazottes*
F. Redig†
E. Verbitskiy‡

March 3, 2005

Abstract: We prove an exponential approximation for the law of approximate occurrence of typical patterns for a class of Gibbsian sources on the lattice $\mathbb{Z}^d$, $d \geq 2$. From this result, we deduce a law of large numbers and a large deviation result for the the waiting time of distorted patterns.

Key-words: Gibbs measures, approximate matching, exponential law, lossy data compression, law of large numbers, large deviations.

1 Introduction

In recent years there has been growing interest in a detailed probabilistic analysis of pattern matching and approximate pattern matching. For example, in information theory, motivation comes from studying performance of idealized Lempel-Ziv coding schemes. In mathematical biology, one likes to have accurate estimates for the probability that two (e.g. DNA) sequences agree in a large interval with some error-percentage.

There is also considerable interest in the analysis of occurrence of patterns in the multi-dimensional setting, e.g., in the context of video-image compression [2], and more generally, lossy data compression [5, 6].

In this paper we study the following problem. Fix a pattern A_n in a cubic box of size n. Given a configuration σ of a Gibbs random field, what is the size of the "observation window" in which we do not necessary see exactly this pattern for the

*CPhT, CNRS-Ecole polytechnique, 91128 Palaiseau Cedex, France, jean-rene@cpht.polytechnique.fr

†Faculteit Wiskunde en Informatica, Technische Universiteit Eindhoven, Postbus 513, 5600 MB Eindhoven, The Netherlands, f.h.j.redig@tue.nl

‡Philips Research, Prof. Holstlaan 4, 5656 AA Eindhoven, The Netherlands, evgeny.verbitskiy@philips.com

first time, but any pattern obtained by distortion of the fixed pattern A_n ? By this we mean a pattern which contains a fixed fraction ϵ of spins different from those of A_n. We are interested in the behavior of the volume of this observation window, that we call "approximate hitting-time", when n grows.

Our main result (Theorem 1) can be phrased as follows. The distribution of the approximate hitting-time, when properly normalized, gets closer and closer to an exponential law. The normalization is the product of a certain parameter Λ_n and the probability of the set of distorted patterns $[A_n]^\epsilon$. In fact, we get a precise control of the error term which allows us to derive two corollaries for the approximate waiting-time: given a configuration η randomly chosen from an ergodic Gibbs random field, we increase the observation window in the random configuration σ drawn from the given Gibbs random field until we see approximately the pattern η_{C_n}. The first corollary implies a law of large numbers allowing to get the rate distortion function almost-surely from this approximate waiting-time. The second corollary is related to large deviations bounds. While the law of large numbers for approximate waiting-times appears in [6], the large deviation result is new. Moreover, the convergence in distribution to the exponential law of the rescaled approximate hitting-time is also new.

We briefly indicate the key ingredients needed to prove this exponential approximation. First, we assume that the Gibbs random field satisfies a certain strong mixing condition (non-uniformly φ-mixing condition). For instance, this property holds for all Markov random fields which satisfy the Dobrushin uniqueness condition. The second key ingredient is a result by Chi [4] allowing one to obtain the rate distortion function "à la Shannon-McMillan-Breiman". We take advantage of our previous work [1] in which we deal with "exact" hitting-times. The proof of the main result of the present work readily follows a large part of the proof in [1] but there is a crucial step which is different ("second moment estimate "). Moreover, one has to restrict to "good" patterns: if a pattern has "too much overlap" with its translates by vectors of size of order n, then one cannot hope to obtain an exponential distribution. These good patterns are shown to be typical in the sense that the measure of good patterns approaches one exponentially fast as n diverges (Proposition 2). When we have a random field distributed according to a Bernoulli measure, the goodness assumption on patterns can be removed. In this case, we prove (Theorem 2) that for any pattern Theorem 1 applies. Surprisingly, our proof involves the strong invariance principle for simple random walks. We have no idea to provide a simpler proof.

Outline of the paper. In the next section, we set notations and definitions, and state our main theorems. In section 3, we apply the exponential approximation of the previous section to approximate waiting times for which we obtain a.s. strong approximation and large deviations results. In section 4, we give all proofs.

Acknowledgment. We thank Z. Chi for providing us his preprint [4].

2 Set-up and main results

For the sake of simplicity, we consider a $\{0,1\}$-valued random field on the lattice $\mathbb{Z}^d$, $d \geq 2$. The results hold for arbitrary finite alphabets as well. Configurations are denoted η, σ, ω and collected in the set $\Omega = \{0,1\}^{\mathbb{Z}^d}$. Ω is provided with the Borel σ-field, and for $V \subseteq \mathbb{Z}^d$, $\mathcal{F}_V$ denotes the σ-algebra generated by $\{\sigma_x : x \in V\}$.

For a finite subset $V \subseteq \mathbb{Z}^d$ and configurations $\sigma, \eta \in \Omega$ we denote by

$$\Delta(V, \eta, \sigma) = \sum_{x \in V} |\eta_x - \sigma_x| \tag{2.1}$$

the number of mismatches between σ and η in the volume V, i.e., the Hamming distance between η_V and σ_V.

We denote by C_n the cube $[0,n]^d \cap \mathbb{Z}^d$. A n-pattern is a map $A_n : C_n \to \{0,1\}$. It is naturally associated to its cylinder $[A_n] = \{\sigma \in \Omega : \sigma_{C_n} = A_n\}$. For a pattern A_n and $x \in \mathbb{Z}^d$ we denote by $\theta_{-x}A_n$ the pattern supported on $C_n + x$ defined by $A_n(y+x) = A_n(y)$ $(y \in C_n)$.

For a pattern A_n we denote by $[A_n]^\epsilon$ the set of configurations which ϵ-match with A_n :

$$[A_n]^\epsilon = \{\omega \in \Omega : \Delta(C_n, \omega, A_n) \leq \epsilon |C_n|\}\,. \tag{2.2}$$

The set of configurations $[A_n]^\epsilon$ can also be viewed as a set of n-patterns, and we will (with a slight abuse of notation) use the same symbol for the set of configurations and the set of n-patterns which are restrictions of configurations in $[A_n]^\epsilon$ to C_n.

Definition 1. *The approximate hitting time of $[A_n]^\epsilon$ in a configuration σ is defined as*

$$\mathbf{T}_{[A_n]^\epsilon}(\sigma) = \min\{|C_k| : k > 0, \exists x \in \mathbb{Z}^d, C_n + x \subseteq C_k \text{ and } \theta_{-x}\sigma \in [A_n]^\epsilon\}\,. \tag{2.3}$$

For $\epsilon = 0$ (exact matching time or occurrence time of a pattern), we obtained in [1] an exponential approximation for the law of $\mathbf{T}_{[A_n]^\epsilon}$ under the hypotheses of non-uniform φ-mixing and Gibbsianess of the random field. We recall here what this mixing assumption is. For $m > 0$ define

$$\varphi(m) = \sup \frac{1}{|A_1|} |\, \mathbb{P}\left(E_{A_1} | E_{A_2}\right) - \mathbb{P}\left(E_{A_1}\right) |\, \tag{2.4}$$

where the supremum is taken over all finite subsets A_1, A_2 of $\mathbb{Z}^d$, with $d(A_1, A_2) \geq m$ ([1]) and $E_{A_i} \in \mathcal{F}_{A_i}$, with $\mathbb{P}(E_{A_2}) > 0$. Note that this $\varphi(m)$ differs from the usual φ-mixing function since we divide by the size of the dependence set of the event E_{A_1}.

Definition 2. *A random field is* **non-uniformly exponentially φ-mixing** *if there exist constants $C_1, C_2 > 0$ such that*

$$\varphi(m) \leq C_1 e^{-C_2 m} \quad \text{for all} \quad m > 0. \tag{2.5}$$

[1] As usual, $d(A_1, A_2) := \inf\{d(x,y) : x \in A_1, y \in A_2\}$ and $d(x,y) := ||x-y||_\infty = \max_{1 \leq i \leq d} |x_i - y_i|$.

A typical example of a Gibbs field satisfying this assumption is the 2d-Ising model above critical temperature. In general, it is satisfied in the so-called high-temperature regime of Dobrushin uniqueness. We refer the reader to [8, 9] for more details on this and on Gibbs measures in general.

An important property of Gibbs measures is the so-called "finite energy" property. This means that there is a continuous version of the conditional probability $\mathbb{P}(\sigma_0 = 0|\sigma_{\mathbb{Z}^d\setminus\{0\}})$ such that

$$\delta < \mathbb{P}(\sigma_0 = 0|\sigma_{\mathbb{Z}^d\setminus\{0\}}) < (1-\delta) \tag{2.6}$$

where $\delta \in (0, \frac{1}{2})$ is independent of σ. This immediately implies the existence of $\kappa > 0$ such that for all $V \subseteq \mathbb{Z}^d$, and all $\eta \in \Omega$

$$\mathbb{P}(\{\sigma : \sigma_V = \eta_V\}) \leq e^{-\kappa|V|}\,. \tag{2.7}$$

We will use the following estimate:

Lemma 1. *Under the assumption that $\mathbb{P}$ is a Gibbs measure, there exist $\epsilon_c > 0$ and $K = K(\epsilon_c) > 0$ such that for any pattern A_n and any $\epsilon < \epsilon_c$*

$$\mathbb{P}([A_n]^\epsilon) \leq e^{-Kn^d}\,.$$

Proof. This is an immediate consequence of the estimate (2.7) and the estimate

$$|[A_n]^\epsilon| \leq \sum_{k=0}^{\epsilon n^d} \binom{n^d}{k} \leq e^{n^d I(\epsilon)}$$

with $I(\epsilon) \downarrow 0$ if $\epsilon \downarrow 0$. □

Contrarily to the situation for exact matching, we will need an assumption on the patterns in order to obtain an exponential law. This can be compared with the condition of not being "badly self-repeating" needed to obtain the exponential law for the return times in [1]. As we shall see, being a "good" pattern is a typical property.

Definition 3. *Given $0 < \alpha < 1, 0 \leq \epsilon < 1$, we say that a n-pattern A_n is called (ϵ, α)-good if the set $[A_n]^\epsilon \cap \theta_x[A_n]^\epsilon$ is empty for all $x \in \mathbb{Z}^d$ such that $|x| \leq \alpha n$. The set of all (ϵ, α)-good patterns is denoted by $\mathcal{G}_n(\epsilon, \alpha)$. By abuse of notation, we use the same symbol for the set of configurations ω such that ω_{C_n} is (ϵ, α)-good.*

For $\epsilon = 0$ and $\alpha < 1/2$, $\mathcal{G}_n(\epsilon, \alpha)$ coincides with the set of non badly self-repeating patterns in [1], definition 5.1.

We shall need a result by Chi [4] on the rate distortion function. We recall briefly the definition of the rate distortion function and refer the reader to [3] for more information and background and to [6] for a discussion on lossy data compression. Given a stationary and ergodic measure $\mathbb{Q}$ and a stationary and ergodic Gibbs measure $\mathbb{P}$, the rate distortion function $\mathcal{R}(\mathbb{Q}, \mathbb{P}, \epsilon)$ is defined as follows:

$$\mathcal{R}(\mathbb{Q}, \mathbb{P}, \epsilon) = \lim_{n\to\infty} \mathcal{R}_n(\mathbb{Q}, \mathbb{P}, \epsilon) \tag{2.8}$$

$$\mathcal{R}_n(\mathbb{Q}, \mathbb{P}, \epsilon) = \inf_{J_n} \frac{1}{|C_n|} H(\mathbb{J}_n \parallel \mathbb{Q}_n \times \mathbb{P}_n) \tag{2.9}$$

where the infimum taken over all joint distributions $\mathbb{J}_n$ on $\{0,1\}^{n^d} \times \{0,1\}^{n^d}$ such that the $\{0,1\}^{n^d}$-marginal of $\mathbb{J}_n$ is $\mathbb{Q}_n$ and

$$\int \frac{\Delta(C_n, \omega, \sigma)}{|C_n|} \, d\mathbb{J}_n(\omega, \sigma) \leq \epsilon \, .$$

$H(\mathbb{J}_n \parallel \mathbb{Q}_n \times \mathbb{P}_n)$ is the relative entropy between $\mathbb{J}_n$ and $\mathbb{Q}_n \times \mathbb{P}_n$.

We have the following key result which follows from [4] and [6, Theorem 25].

Proposition 1. *Let $\mathbb{Q}$ be a stationary and ergodic measure and $\mathbb{P}$ be a stationary and ergodic Gibbs measure. Then*

$$\mathcal{R}(\mathbb{Q}, \mathbb{P}, \epsilon) = -\lim_{n\to\infty} \frac{1}{|C_n|} \log \mathbb{P}([\omega_{C_n}]^\epsilon) \quad \mathbb{Q} - \text{almost-surely} \, . \tag{2.10}$$

Moreover, $\mathcal{R}$ is a convex (and hence continuous) function of ϵ and is non-zero in some interval $[0, \epsilon_0)$.

The property (2.10) is called the generalized asymptotic equipartition property in [6]. Throughout we will simply write $\mathcal{R}(\epsilon)$ instead of $\mathcal{R}(\mathbb{Q}, \mathbb{P}, \epsilon)$.

We can now state our main result.

Theorem 1. *Suppose that $\mathbb{P}$ is a non-uniformly exponentially φ-mixing Gibbs measure and $\mathbb{Q}$ is a stationary and ergodic Gibbs measure. Assume that the rate distortion function (2.8) is strictly positive in $[0, \epsilon_0)$. Then for all $\alpha \in (0,1)$ and $\epsilon > 0$ small enough: namely,*

$$\frac{\epsilon}{\alpha} < \epsilon_0,$$

there exist $\Lambda_1, \Lambda_2, C, c \in (0, \infty)$, such that and for every $t > 0$, $n \geq 1$, and $\mathbb{Q}$-almost all ω with $\omega_{C_n} \in \mathcal{G}_n(\epsilon, \alpha)$, the following estimate holds:

$$\left| \mathbb{P}\left(\mathbf{T}_{[\omega_{C_n}]^\epsilon} > \frac{t}{\Lambda_n \mathbb{P}([\omega_{C_n}]^\epsilon)} \right) - e^{-t} \right| \leq C e^{-ct} \, e^{-Kn^d} \tag{2.11}$$

where $\Lambda_n = \Lambda(\omega_{C_n})$ is such that

$$\Lambda_1 \leq \Lambda_n \leq \Lambda_2 \, . \tag{2.12}$$

Dependence of the parameters in Theorem 1 on ϵ and α will be discussed after the proof, see Remark 1.

Let us briefly comment on the difference between Theorem 1 and the one obtained in [1] for exact matching, that is the case corresponding to $\epsilon = 0$. First of all, we need to restrict ourselves to special patterns, i.e., (ϵ, α)-good patterns, whereas in [1] result applies to all patterns. Secondly, the error term that we obtain in [1] is of the form $Ce^{-ct}\mathbb{P}([\omega_{C_n}])^\rho$, where $\rho > 0$. Of course, the factor $\mathbb{P}([\omega_{C_n}])^\rho$ is uniformly exponentially small for Gibbs measures. This is no longer true for $\mathbb{P}([\omega_{C_n}]^\epsilon)$ if ϵ is too large. This is precisely why we need Lemma 1. Thirdly, a crucial step in the

proof of Theorem 1, which differs slightly from that in [1] for the case $\epsilon = 0$, involves Proposition 1. This explains why we need to restrict to typical configurations in the sense of this result.

Let us close this set of remarks by noticing that $\mathbb{Q}$ has to be a stationary and ergodic measure, but not necessarily Gibbsian. But for later use of Theorem 1 we shall also need the latter assumption, so we already impose it to state the theorem.

The following proposition shows that "$\omega_{C_n} \in \mathcal{G}(\epsilon, \alpha)$", i.e. that a pattern being (ϵ, α)-good, is a typical property.

Proposition 2. *Let $\mathbb{Q}$ be a stationary Gibbs measure. Then, if $\alpha < 1/2$ and $\epsilon > 0$ is small enough, there exists $\nu > 0$ such that for all $n \geq 1$*

$$\mathbb{Q}(\mathcal{G}_n(\epsilon, \alpha)) > 1 - e^{-\nu n^d} . \tag{2.13}$$

It turns out that if the random field has a non-trivial dependence structure, then the restriction to (ϵ, α)-good patterns is unavoidable. However, in the case of a random field distributed according to a Bernoulli measure, the exponential law (2.11) holds for *all* approximate patterns. This is expressed by the following theorem.

Theorem 2. *If $\mathbb{P}$ is the Bernoulli measure with $\mathbb{P}(\sigma_0 = 1) = 1/2$, then (2.11) holds without the restriction that ω_{C_n} is (ϵ, α)-good.*

3 Approximate waiting-time fluctuations

The purpose of this section is to derive two consequences of Theorem 1 and Proposition 2. The first one implies a strong law of large numbers for the approximate waiting-time. It was previously derived in [6] directly using the mixing property (2.5). The second one concerns large deviations of the approximate waiting-time and it is a new result. Given two configurations ω, σ, the **approximate waiting time** is $\mathbf{W}_n^\epsilon(\omega, \sigma) := \mathbf{T}_{[\omega_{C_n}]^\epsilon}(\sigma)$.

Proposition 3. *Under the assumptions of Theorem 1 and Proposition 2, there exists $\gamma_0 > 0$ such that for all $\gamma > \gamma_0$:*

$$-\gamma \log n \leq \log(\mathbf{W}_n^\epsilon(\omega, \sigma)\mathbb{P}([\omega_{C_n}]^\epsilon)) \leq \log(\log n^\gamma) \tag{3.1}$$

$\mathbb{Q} \times \mathbb{P}$-eventually almost surely. In particular

$$\lim_{n\to\infty} \frac{1}{|C_n|} \log \mathbf{W}_n^\epsilon(\omega, \sigma) = \mathcal{R}(\mathbb{Q}, \mathbb{P}, \epsilon) \quad \mathbb{Q} \times \mathbb{P} - \text{almost surely} .$$

With Proposition 3 we recover the results of Theorems 26 and 27 in [6]. However, there is a substantial difference in conditions on random fields. We have to restrict ourselves to measures $\mathbb{Q}$ which are stationary and ergodic Gibbs measures, while in [6] $\mathbb{Q}$ is only assumed to be stationary and ergodic. On the other hand, we permit

Gibbs $\mathbb{P}$, while in [6] $\mathbb{P}$ must be Bernoulli. The reason for our assumptions on $\mathbb{Q}$ is that Proposition 2 is valid for Gibbs measures. We do not know if it can be extended to more general situations.

Let us also remark that by a basic result in Probability Theory, this strong approximation implies that if a central limit theorem holds for $-(1/|C_n|)\ \log \mathbb{P}([\omega_{C_n}]^\epsilon))$, then it holds also for $(1/|C_n|)\ \log \mathbf{W}_n^\epsilon(\omega,\sigma)$. Unfortunately, the former seems to be a difficult issue, except in the iid case. We refer the reader to [6] for some results in that direction.

We have the following (partial) large deviation results. We first need the following lemma showing that we can define the generalized conditional q-order Rényi entropy for Gibbs random fields. This was first done in [11] for (α-mixing) stochastic processes ($d = 1$) with the difference that here we need to condition on (ϵ,α)-good patterns and use the Gibbs property instead of mixing.

Lemma 2. *Let $\mathbb{Q},\mathbb{P}$ be stationary Gibbs measures and assume that $\alpha < 1/2$ and $0 \le \epsilon < 1$. Then, for all $q \in \mathbb{R}$, the following function is well-defined:*

$$\mathcal{E}_\epsilon(q) := \mathcal{E}_\epsilon(q;\mathbb{Q},\mathbb{P}) = \lim_{n\to\infty} \frac{1}{|C_n|}\ \log \int \mathbb{P}([\omega_{C_n}]^\epsilon)^q\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega)\,. \tag{3.2}$$

($\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}$ denotes the measure $\mathbb{Q}$ conditioned on the set of good patterns.)

The generalized q-order Rényi entropy should be defined as $-\mathcal{E}_\epsilon(-q)/q$.

We now have the following theorem. By $a_n \approx b_n$ we mean that $\max\{a_n/b_n, b_n/a_n\}$ is bounded from above.

Theorem 3. *Let $\mathbb{P}$ be a non-uniformly exponentially φ-mixing Gibbs measure and $\mathbb{Q}$ a stationary and ergodic Gibbs measure. If $\epsilon > 0$ is small enough, then for any $\alpha_0 \le \alpha < 1/2$, we have*

$$\int\int (\mathbf{W}_n^\epsilon(\omega,\sigma))^q\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega)\ d\mathbb{P}(\sigma) \approx \int \mathbb{P}([\omega_{C_n}]^\epsilon)^{-q}\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega) \quad \text{if } q \ge -1 \tag{3.3}$$

and

$$\int\int (\mathbf{W}_n^\epsilon(\omega,\sigma))^q\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega)\ d\mathbb{P}(\sigma) \approx \int \mathbb{P}([\omega_{C_n}]^\epsilon)\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega) \quad \text{if } q < -1 \tag{3.4}$$

In particular,

$$\lim_{n\to\infty} \frac{1}{|C_n|} \log \int\int (\mathbf{W}_n^\epsilon(\omega,\sigma))^q\ d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega)\ d\mathbb{P}(\sigma) = \begin{cases} \mathcal{E}_\epsilon(-q) & \text{if } q \ge -1 \\ \mathcal{E}_\epsilon(1) & \text{if } q < -1\,. \end{cases} \tag{3.5}$$

It follows from this theorem that Theorem 4.5.20 in [7] applies to large deviations of $((1/|C_n|)\log \mathbf{W}_n^\epsilon(\eta,\sigma))_n$. We do not know under which conditions the function $q \mapsto \mathcal{E}_\epsilon(q)$ is, for example, continuously differentiable for ϵ small enough. It it were the case, we would get a large deviation principle for $((1/|C_n|)\log \mathbf{W}_n^\epsilon(\eta,\sigma))_n$ with a rate function given by the Legendre transform of $\mathcal{E}_\epsilon(-q)$ for $q > -1$.

4 Proofs

4.1 Proof of Theorem 1

The proof of Theorem 1 is quite similar to the proof of exponential law in [1]. We describe briefly the common approach and indicate the differences. We also provide the necessary modifications of the proof.

It is well known that a random variable Z has an exponential distribution if and only if

$$\mathbb{P}(Z > s + t | Z > t) = \mathbb{P}(Z > s)$$

or, equivalently,

$$\mathbb{P}(Z > s + t) = \mathbb{P}(Z > s)\mathbb{P}(Z > t)\,.$$

The basic ingredient of the proof in [1] was Lemma 4.4 ("Iteration Lemma"). This result establishes that for a pattern A_n and any finite number of cubes $C_i \subseteq \mathbb{Z}^d$, $i = 1, \ldots, k$, with equal volumes

$$|C_i| = \Big(\frac{1}{\mathbb{P}(A_n)}\Big)^\gamma$$

we have

$$\mathbb{P}\Big(A_n \text{ does not occur in } \bigcup_{i=1}^{k} C_i\Big) \approx \mathbb{P}\Big(A_n \text{ does not occur in } C_1\Big)^k. \tag{4.1}$$

In [1] we also observed that the Iteration Lemma remains valid if a pattern A_n is replaced by the event $[A_n]^\epsilon$, with $[A_n]^\epsilon$ does not occur in volume V if any pattern $B_n \in [A_n]^\epsilon$ does not occur in volume V.

Another important ingredient of the proof is the control of the parameter of the exponential distribution. Lemma 4.3 ("The parameter") in [1] concerns non-triviality of the parameter Λ_n, that is, the fact that it is neither null nor infinite. To prove Lemma 4.3, we established a *uniform* second moment estimate for the number of occurrences of a pattern A_n in a configuration σ restricted to a box that has later to be taken of size $1/\mathbb{P}([A_n])$. It is the proof of this second moment estimate that we have to modify completely. In remark 4.1 in [1], we noticed that if $E_n \in \mathcal{F}_{C_n}$ are events such that $\mathbb{P}(E_n) < e^{-cn^d}$ for some $c > 0$, and such that

$$\limsup_{n\to\infty} \sum_{0<|x|<n} \frac{\mathbb{P}(E_n \cap \theta_x E_n)}{\mathbb{P}(E_n)} < \infty \tag{4.2}$$

then this implies, together with the mixing property (2.5) and the Gibbs property (2.7), that the desired uniform second moment estimate holds. In turn, this implies the non-triviality of the parameter (2.12) (Lemma 4.3 in [1]).

Thus, we turn to prove (4.2) when the event E_n is $[A_n]^\epsilon$, where A_n is a good and typical pattern. We assume that patterns A_n are such that $\cap_n A_n = \{\sigma\}$, with σ

chosen from the $\mathbb{Q}$-measure one set of Proposition 1 and such that A_n is good in the sense of Definition 3.

We have to show for patterns $A_n \in \mathcal{G}_n(\epsilon, \alpha)$ with $\epsilon/\alpha < \epsilon_0$, there exists a finite $C = C(\epsilon, \alpha)$ such that for all n

$$\sum_{0<|x|\le n} \frac{\mathbb{P}([A_n]^\epsilon \cap \theta_x[A_n]^\epsilon)}{\mathbb{P}([A_n]^\epsilon)} \le C(\epsilon, \alpha)\,. \tag{4.3}$$

First of all, since $A_n \in \mathcal{G}_n(\epsilon, \alpha)$ (see Definition 3), the terms corresponding to x with $|x| < \alpha n$ are equal to 0. Therefore we have to estimate the sum

$$\sum_{\alpha n\le|x|\le n} \frac{\mathbb{P}([A_n]^\epsilon \cap \theta_x[A_n]^\epsilon)}{\mathbb{P}([A_n]^\epsilon)} \tag{4.4}$$

Note that for x with $|x| \ge \alpha n$, the intersection $(C_n + x) \cap C_n$ is not very large:

$$|(C_n + x) \cap C_n| \le (1 - \alpha)n^d.$$

Note also that $\Delta(V, \omega, A_n)$ denotes the number of differences between ω and A_n in the volume V, see (2.1) . Then we can write

$$\mathbb{P}([A_n]^\epsilon \cap \theta_x[A_n]^\epsilon) = \mathbb{P}\left(\omega : \Delta(C_n, \omega, A_n) \le \epsilon n^d \cap \Delta(C_n + x, \omega, \theta_{-x}A_n) \le \epsilon n^d\right) \tag{4.5}$$

where by $\theta_{-x}A_n$ we mean $\theta_{-x}A_n(y+x) = A_n(y)$, $y \in C_n$. For the sake of convenience, we simply write C for C_n and C_x for $C_n + x$ in the course of this proof. We also introduce the short-hand notations

$$\begin{aligned} S_1 &= \Delta(C \setminus C_x, \omega, A_n) \\ S_2 &= \Delta(C \cap C_x, \omega, A_n) \\ S_3 &= \Delta(C \cap C_x, \omega, \theta_{-x}A_n) \\ S_4 &= \Delta(C \setminus C_x, \omega, \theta_{-x}A_n)\,. \end{aligned} \tag{4.6}$$

With these notations what we have to estimate

$$\sum_{\alpha n\le|x|\le n} \frac{\mathbb{P}([A_n]^\epsilon \cap \theta_x[A_n]^\epsilon)}{\mathbb{P}([A_n]^\epsilon)} = \sum_{\alpha n\le|x|\le n} \mathbb{P}(S_3 + S_4 \le \epsilon n^d | S_1 + S_2 \le \epsilon n^d). \tag{4.7}$$

The following estimate is a corollary of [4] and a basic property of a Gibbs measures: for any configuration ξ:

$$\mathbb{P}(\{\omega : \Delta(V_n, \omega, \sigma) \le \epsilon|V_n|\}|\ \xi_{V_n^c}) \le \exp(-|V_n|\ \mathcal{R}(\epsilon) + c|\partial V_n|) \tag{4.8}$$

Indeed, the unconditioned statement is proved in [4], and conditioning can at most introduce a term of order $\exp(c|\partial V_n|)$.

We proceed as follows

$$\begin{aligned}\mathbb{P}\big((S_1+S_2)\le \epsilon n^d \cap (S_3+S_4)\le \epsilon n^d\big) &\le \mathbb{P}\left((S_1+S_2)\le \epsilon n^d \cap S_4\le \epsilon n^d\right)\\ &\le \sup_{\xi}\mathbb{P}(S_4\le \epsilon n^d|\xi_{\mathbb{Z}^d\setminus(C_x\setminus C)})\mathbb{P}([A_n]^\epsilon)\\ &\le \exp\left(-\alpha n^d\mathcal{R}\left(\frac{\epsilon}{\alpha}\right)+cn^{d-1}\right)\mathbb{P}([A_n]^\epsilon)\ .\end{aligned}$$

Therefore

$$\sum_{\alpha n\le|x|\le n}\frac{\mathbb{P}([A_n]^\epsilon\cap\theta_x[A_n]^\epsilon)}{\mathbb{P}([A_n]^\epsilon)}\le n^d\exp\left(-\alpha n^d\mathcal{R}\left(\frac{\epsilon}{\alpha}\right)+cn^{d-1}\right)=:C_n(\epsilon,\alpha).$$

Taking into account that $\epsilon/\alpha<\epsilon_0$, and hence $\mathcal{R}(\epsilon/\alpha)>0$, we conclude that $C_n(\epsilon,\alpha)\to 0$ as $n\to\infty$, and hence

$$C(\epsilon,\alpha)=\sup_n C_n(\epsilon,\alpha)$$

is finite. This finishes the proof.

Remark 1. *The parameters of Theorem 1 depend on the choice of ϵ and α. The most interesting is the dependence of Λ_1 and Λ_2. The Parameter Lemma of [1] in fact shows that a uniform choice $\Lambda_2=2$ suffices. A more interesting question is whether we can give a uniform bound on Λ_1 for a large set of ϵ and α. The present modification of the second moment estimate together with the rest of Lemma 4.3 in [1], which remains unchanged, gives that for some c, dependent on ϵ alone, the following choice of $\Lambda_1=\Lambda_1(\epsilon,\alpha)$ will suffice*

$$\Lambda_1=\frac{1}{c+C(\epsilon,\alpha)}.$$

The rate distortion function $\mathcal{R}$ is a monotonically decreasing function. Hence, for a fixed $\epsilon>0$, $\alpha\mathcal{R}(\frac{\epsilon}{\alpha})$ is a monotonically increasing function of α, and finally, $C(\epsilon,\alpha)$ is monotonically decreasing in α. Therefore, if $\epsilon<\epsilon_0$, then for all $\alpha>\alpha_0:=0.99\frac{\epsilon}{\epsilon_0}$

$$\Lambda_1(\epsilon,\alpha)\ge\Lambda_1(\epsilon,\alpha_0)>0.$$

Therefore, for a fixed $\epsilon>0$ we obtain a uniform (in α) bound on the parameter Λ_1.

4.2 Proof of Proposition 2

For $\epsilon=0$ we know that most patterns are $(0,\alpha)$-good for any $\alpha<1$. Indeed, it is proved in [1] (lemma 5.3) that $\mathbb{Q}(\mathcal{G}_n(\epsilon,\alpha))\ge 1-e^{-\kappa' n^d}$, for some $\kappa'>0$.

Let us now argue that for small ϵ this is still the case. Suppose $\alpha<1/2$, that is, we are going to consider vectors $x\in\mathbb{Z}^d$ such that $|x|\le\frac{n}{2}$. An element A of $[A_n]^\epsilon\cap\theta_x[A_n]^\epsilon$ satisfies

$$\sum_{y\in C_x\cap C}|A(y)-A(y-x)|\le 2\epsilon n^d\ . \tag{4.9}$$

(Recall that $C = C_n$ and $C_x = C_n + x$. This implies that there exists a set $V_n \subseteq C$ and a disjoint translate $V_n + z \subseteq C$ such that $|V_n| > (1/2)^d n^d$ such that $\theta_{-z} A_{V_n+z}$ matches with error fraction $2^{d+1}\epsilon$ with A_{V_n}, this can be made as small as $e^{-\nu n^d}$, for some $\nu > 0$, for ϵ sufficiently small uniformly in A_{V_n} by Lemma 1. Therefore we obtain that

$$\mathbb{Q}(\mathcal{G}(\epsilon, \alpha_0)) > 1 - e^{-\nu n^d} \tag{4.10}$$

for all $\alpha < 1/2$ and ϵ small enough.

4.3 Proof of Theorem 2

We consider the case $d = 1$ only, because the case $d \geq 2$ is completely analogous. Start with the particular pattern $A_n = 0 \cdots 0$ that we simply denote by 0_n. The difficulty with this "bad pattern" comes from the fact that the second moment estimate does not apply, because (4.2) fails. Therefore, we have to prove by other means that there exists $\delta > 0$ such that for all $n \in \mathbb{N}$,

$$\delta < \mathbb{P}\left(\mathbf{T}_{[0_n]^\epsilon} > \frac{1}{\mathbb{P}([0_n]^\epsilon)} \right) < 1 - \delta \tag{4.11}$$

which would imply the non-triviality of the parameter Λ_n. We will first show that there exists a sequence $k_n \uparrow \infty$ such that

$$\delta < \mathbb{P}\left(\mathbf{T}_{[0_n]^\epsilon} > k_n \right) < 1 - \delta \, . \tag{4.12}$$

It will then follow easily from the Bernoulli character of $\mathbb{P}$ that k_n does not depend on the choice of the pattern, i.e., (4.12) holds with the same k_n for any pattern A_n. Then we can apply Theorem 1 for good patterns, and obtain $k_n = 1/\mathbb{P}([A_n]^\epsilon) = 1/\mathbb{P}([0_n]^\epsilon)$. We have the following identities:

$$\begin{aligned} \mathbb{P}(\mathbf{T}_{[0_n]^\epsilon} \leq k_n) &= \mathbb{P}\left(\min_{k=0}^{k_n} \sum_{i=k}^{k+n} \omega_i \leq n\epsilon \right) \\ &= \mathbb{P}\left(\max_{k=0}^{k_n} \sum_{i=k}^{k+n} (1 - 2\omega_i) \geq (1-2\epsilon)n \right) \\ &= \mathbb{P}\left(\max_{k=0}^{k_n} (S_{k+n} - S_k) \geq (1-2\epsilon)n \right) \end{aligned} \tag{4.13}$$

where S_n is the position of a simple random walk on $\mathbb{Z}$ (with $S_0 = 0$) after n steps. By Theorem 7.23 in [12], together with the strong invariance principle [12] p. 53, we have

$$\max_{k=0}^{k_n} (S_{k+n} - S_k) = a \log k_n + b \log\log k_n + c + o(1) + X \tag{4.14}$$

where X is a random variable with a Gumbel distribution. Therefore, if we choose k_n such that

$$(1-2\epsilon)n = a \log k_n + b \log\log k_n + c + o(1) \tag{4.15}$$

then (4.12) holds.

If we now choose any other pattern A_n, then under $\mathbb{P}$, $S_n = 2\sum_{i=0}^{n}(1/2-\sigma_i-A_n(i))$ is again distributed as a simple random walk, so we find the same k_n, which completes the proof of the theorem.

4.4 Proof of Proposition 3

By using Theorem 1 we immediately get

$$\mathbb{Q}\times\mathbb{P}\left\{(\omega,\sigma):\ \log\left(\mathbf{W}_n^{\epsilon}(\omega,\sigma)\mathbb{P}([\omega_{C_n}]^{\epsilon})\right)>\log t\right\}=$$

$$\int d\mathbb{Q}(\omega)\,\mathbb{P}\left\{\sigma:\ \log\left(\mathbf{T}_{[\omega_{C_n}]^{\epsilon}}(\sigma)\mathbb{P}([\omega_{C_n}]^{\epsilon})\right)>\log t\right\}\leq e^{-\Lambda_1 t}+Ce^{-Kn^d}+\sum_{A_n\in\mathcal{G}_n^c(\epsilon,\alpha)}\mathbb{Q}([A_n])\,.$$

Now we choose $t=t_n=\log(n^{\gamma})$ with $\gamma>0$ such that $\Lambda_1\gamma>1$. This makes the first term in the rhs summable in n. The last one equals the $\mathbb{Q}$-measure of the complement of $\mathcal{G}_n(\epsilon,\alpha)$, which is less than $e^{-\nu n^d}$ by Proposition 2. We thus get the upper bound in (3.1) by an application of the Borel-Cantelli Lemma.

Now we turn to prove the lower bound in (3.1). Proceeding as before, we get

$$\mathbb{Q}\times\mathbb{P}\left\{(\omega,\sigma):\ \log\left(\mathbf{W}_n^{\epsilon}(\omega,\sigma)\mathbb{P}([\omega_{C_n}]^{\epsilon})\right)\leq\log t\right\}\leq$$

$$1-e^{-\Lambda_2 t}+Ce^{-Kn^d}+\sum_{A_n\in\mathcal{G}_n^c(\epsilon,\alpha)}\mathbb{Q}([A_n])\leq$$

$$\Lambda_2 t+Ce^{-Kn^d}+e^{-\nu n^d}\,.$$

We have used Theorem 1 and Proposition 2. We now choose $t=t_n=n^{-\gamma}$, with $\gamma>1$, to get a summable upper bound in n for the above probability. An application of Borel-Cantelli Lemma gives the desired result and the proof of the proposition is complete.

4.5 Proof of Lemma 2

We only consider the case $q>0$ leaving the (very similar) proof for the case $q<0$ to the reader. Let $\mathcal{S}_{\square}$ be the system of all rectangular boxes of the form

$$V=\mathbb{Z}^d\cap\prod_{k=1}^{d}[m_k,n_k]\quad\text{with}\ \ m_k,n_k\in\mathbb{Z},m_k\leq n_k\,.$$

Before proceeding, we have to extend definition 3 somewhat. We will denote by $\mathcal{G}_V(\epsilon,\alpha)$ the set of good patterns supported on $V\in\mathcal{S}_{\square}$. We shall need Proposition 2 which remains valid if one replaces $\mathcal{G}_n(\epsilon,\alpha)$ with $\mathcal{G}_V(\epsilon,\alpha)$ and n by $|V|$ in (2.13).

We are going to prove that the function $a:\mathcal{S}_{\square}\to(-\infty,+\infty)$ defined as

$$a(V):=-\log\int\mathbb{P}^q([\sigma_V]^{\epsilon})\;d\mathbb{Q}_{\mathcal{G}_{V\cup V'}(\epsilon,\alpha)}(\sigma)$$

satisfies

$$a(V \cup V') \leq a(V) + a(V') + C\ |\partial(V \cup V')|$$

for all $V, V' \in \mathcal{S}_\square$ such that $V \cup V' \in \mathcal{S}_\square$ and $V \cap V' = \varnothing$, where C is a constant (depends on q), and where ∂V denotes the boundary of V. Of course, $|\partial(V \cup V')|$ is a surface order correction. If such a property holds (together with $a(V + x) = a(V)$, for all $x \in \mathbb{Z}^d$, $V \in \mathcal{S}_\square$ which is obvious by stationarity of the measure), then a generalized sub-additive lemma, obtained as a combination of a lemma found in [8] and another one given in [7], will guarantee that

$$\lim_{n\to\infty} \frac{a(C_n)}{|C_n|}$$

exists, as we wish. For all $q \in \mathbb{R}$, $V, V' \in \mathcal{S}_\square$ such that $V \cup V' \in \mathcal{S}_\square$ and $V \cap V' = \varnothing$, we have the following:

$$\mathbb{P}^q([\sigma_{V\cup V'}]^\epsilon) = \left(\sum_{\omega_{V\cup V'} \in [\sigma_{V\cup V'}]^\epsilon} \mathbb{P}([\omega_{V\cup V'}]) \right)^q \geq$$

$$e^{K_1\ |\partial(V\cup V')|} \left(\sum_{\omega_{V\cup V'} \in [\sigma_{V\cup V'}]^\epsilon} \mathbb{P}([\omega_V])\ \mathbb{P}(\omega_{V'}]) \right)^q \geq$$

$$e^{K_2\ |\partial(V\cup V')|} \left(\sum_{\omega_V \in [\sigma_V]^\epsilon} \mathbb{P}([\omega_V]) \right)^q \left(\sum_{\omega_{V'} \in [\omega_{V'}]^\epsilon} \mathbb{P}(\omega_{V'}]) \right)^q =$$

$$e^{K_2\ |\partial(V\cup V')|}\, \mathbb{P}^q([\sigma_{V\cup V'}]^\epsilon)\, \mathbb{P}^q([\sigma_{V\cup V'}]^\epsilon)$$

where K_1, K_2 are constants. The first inequality follows from the Gibbs property and the second one is a simple consequence of the Hamming distance property. To complete the proof we again use the Gibbs property to get

$$\int \mathbb{P}^q([\sigma_{V\cup V'}]^\epsilon)\ d\mathbb{Q}_{\mathcal{G}_{V\cup V'}(\epsilon,\alpha)}(\sigma) = \sum_{\omega_{V\cup V'} \in \{0,1\}^{V\cup V'}} \mathbb{P}^q([\omega_{V\cup V'}]^\epsilon)\ \mathbb{Q}_{\mathcal{G}_{V\cup V'}(\epsilon,\alpha)}([\omega_{V\cup V'}]) \geq$$

$$\mathbb{Q}(\mathcal{G}_{V\cup V'}(\epsilon,\alpha))\, e^{K_3\ |\partial(V\cup V')|} \times$$

$$\sum_{\omega_V \in \{0,1\}^V} \mathbb{P}^q([\omega_V]^\epsilon)\ \mathbb{Q}_{\mathcal{G}_{V\cup V'}(\epsilon,\alpha)}([\omega_V]) \times \sum_{\omega_{V'} \in \{0,1\}^{V'}} \mathbb{P}^q([\omega_{V'}]^\epsilon)\ \mathbb{Q}_{\mathcal{G}_{V\cup V'}(\epsilon,\alpha)}([\omega_{V'}]) \geq$$

$$\frac{1}{2}\ e^{K_3\ |\partial(V\cup V')|} \int \mathbb{P}^q([\sigma_V]^\epsilon)\ d\mathbb{Q}_{\mathcal{G}_V(\epsilon,\alpha)}(\sigma) \times \int \mathbb{P}^q([\sigma_{V'}]^\epsilon)\ d\mathbb{Q}_{\mathcal{G}_{V'}(\epsilon,\alpha)}(\sigma)$$

where K_3 is a constant. The second inequality is the consequence of Proposition 2 if $|V \cup V'|$ is large enough. The lemma is proved.

4.6 Proof of Theorem 3

Since the proof of this theorem is very similar to that of Theorem 2.7 in [1], we only sketch it to indicate the little differences between them.

The starting point is of course to write

$$\int\int (\mathbf{W}_n^\epsilon(\omega,\sigma))^q \; d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega) \; d\mathbb{P}(\sigma) = \int d\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}(\omega) \; \int \mathbf{T}^q_{[\omega_{C_n}]^\epsilon}(\sigma) \; d\mathbb{P}(\sigma) \; .$$

Then we can mimic the proof of Theorem 2.7 in [1] by using Theorem 1 and the analog of lemma 4.3 in [1], which holds true when $\mathbf{T}_{[\omega_{C_n}]}$ is replaced by $\mathbf{T}_{[\omega_{C_n}]^\epsilon}$, provided that ω_{C_n} be a (ϵ,α)-good pattern (see the beginning of the proof of Theorem 1), and ω be $\mathbb{Q}$-typical in the sense of Proposition 1. Notice that we integrate with respect to the conditional measure $\mathbb{Q}_{\mathcal{G}_n(\epsilon,\alpha)}$ which takes care of these two properties.

References

[1] M. Abadi, J.-R. Chazottes F. Redig and E. Verbitskiy, *Exponential distribution for the occurrence of rare patterns in Gibbsian random fields*, Commun. Math. Phys. **246** no. 2 (2004), 269–294.

[2] M. Alzina, W. Szpankowski and A. Grama, *2d-pattern matching, image and video-compression: theory, algorithms and experiments*, IEEE Transactions on Image Processing **11** no. 3, (2002), 318-331.

[3] T. Berger, *Rate Distorsion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[4] Z. Chi, *Conditional large deviation principle for finite state Gibbs random fields.* Preprint (2002).

[5] A. Dembo, I. Kontoyiannis, *The asymptotics of waiting times between stationary processes, allowing distortion*, Ann. Appl. Probab. **9** (1999), no. 2, 413–429.

[6] A. Dembo, I. Kontoyiannis, *Source coding, large deviations, and approximate matching*, IEEE Trans. Inform. Th. **48** no. 6 (2002), 1590–1615.

[7] A. Dembo, O. Zeitouni, *Large deviations techniques and applications.* Second edition. Applications of Mathematics (New York) **38**. Springer-Verlag, New York, 1998.

[8] H.-O. Georgii. *Gibbs Measures and Phase Transitions.* Walter de Gruyter & Co., Berlin, 1988.

[9] X. Guyon. *Random Fields on a Network. Modeling, Statistics and Applications*, Springer Verlag, New York, Berlin, 1995.

[10] I. Kontoyiannis, *Pattern matching and lossy data compression on random fields*, IEEE Trans. Inform. Th. **49** no. 4 (2003), 1047–1051.

[11] T. Luczak, W. Szpankowski, *A suboptimal lossy data compression based on approximate pattern matching*, IEEE Trans. Inform. Th. **43** no. 5 (1997), 1439–1451.

[12] P. Révész, *Random Walks in Random and Non-Random Environments*, World Scientific, Singapore, New Jersey, London, Hong Kong, (1990).