Matching with shift for one dimensional Gibbs measures

P.Collet [*]
C. Giardinà [†]
F. Redig [‡]

March 8, 2008

Abstract: We consider matching with shifts for Gibbsian sequences. We prove that the maximal overlap behaves as $c \log n$, where c is explicitly identified in terms of the thermodynamic quantities (pressure) of the underlying potential. Our approach is based on the analysis of the first and second moment of the number of overlaps of a given size. We treat both the case of equal sequences (and non-zero shifts) and independent sequences.

[*] Centre de Physique Théorique, CNRS UMR 7644, 91128 Palaiseau Cedex, France, *collet@cpht.polytechnique.fr*

[†] Department of Mathematics and Computer Science, Eindhoven University, P.O. Box 513 - 5600 MB Eindhoven, The Netherlands, *c.giardina@tue.nl*

[‡] Mathematisch Instituut Universiteit Leiden, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands, *redig@math.leidenuniv.nl*

1 Introduction

In sequence alignment, one wants to detect significant similarities between two (e.g. genetic or protein) sequences. In order to distinguish "significant" similarities, one has to compute the probability that a similarity of a certain size occurs for two independent sequences. The symbols in the sequences are however not necessarily occurring independently. From the point of view of statistical mechanics, it is quite natural to assume that the symbols in the sequence are generated according to a stationary Gibbs measure: this is the equilibrium measure which maximizes the entropy under physical constraints such as energy conservation. A priori there is no reason to assume that the symbols (bases) in e.g. a DNA sequence are i.i.d. or even Markov. It can however be plausible to assume that there is an underlying Markov chain of which the symbol sequence is a reduction: in that case we arrive at a so-called hidden Markov chain, and it is well-known that hidden Markov chains have generically infinite memory (though the symbol at a particular location only exponentially weakly depends on symbols far away). Therefore, proposing a Gibbs measure with exponentially decaying interaction as a model for the sequence seems quite natural. Besides motivation coming from sequence alignment, also in dynamical systems, one can ask for the probability of having a large "overlap" in a trajectory of length n, but without specifying the location of the piece of trajectory that is repeated. It is clear that this probability is related to the entropy, but not in such a straightforward way as the return time. In (hyperbolic) dynamical systems, by coding and partitioning one again naturally arrives at Gibbs measures with exponentially decaying interactions.

The first non-trivial problem associated with sequence alignment is the comparison of two sequences where it is allowed to shift one sequence w.r.t. the other. Remark that this problem is not easy even in the case of independent symbols in the sequence, because one allows for shifting one sequence w.r.t. the other. The comparison consists in the simplest case in finding the maximal number of consecutive equal symbols. Given two (independent) i.i.d. sequences, in [5], [6] it is proved that the maximal overlap, allowing shifts, behaves for large sequence length as $c \log n + X$ where n is the length of both sequences, c is a constant depending on the distribution of the sequence, and where X is a random variable with a Gumbel distribution. The analysis of these papers is based on large deviations, together with an analysis of random walk excursions. As the proofs use a form of permutation invariance, they cannot be extended to non-i.i.d. cases. In [9] the maximal alignment with shift is shown for Markov sequences, which requires a theory of excursions of random walk with Markovian increments.

In this paper, we focus on the more elementary question of showing that the maximal overlap, allowing shifts behaves as $c \log n$, but now in the context of general Gibbsian sequences. We also allow to match a sequence with *itself* (where of course we have to restrict to non-zero shifts). The constant c is explicitly identified and related to thermodynamic quantities associated to the potential of the underlying Gibbs measure.

Our approach is based on a first and second moment analysis of the random variable $N(\sigma, n, k)$ that counts the number of shift-matches of size k in a sequence σ of length n. One easily identifies the scale $k = k_n$ which discriminates the region where the first moment $\mathbb{E}N(\sigma, n, k_n)$ goes to zero (as $n \to \infty$) from the region where $\mathbb{E}N(\sigma, n, k)$ diverges. Via a second moment estimate, we then prove that this scale also separates the $N(\sigma, n, k) \to 0$ versus $N(\sigma, n, k) \to \infty$ (convergence in probability) region.

Our paper is organized as follows: in section 2 we introduce the basic preliminaries about

Gibbs measures, in section 3 we analyse the first moment of N in the case of matching a sequence with itself, and in section 4 we study the second moment. In section 5 we treat the case of two independent (Gibbsian) sequences with the same and with different marginal distributions.

2 Definitions and Preliminaries

We consider random stationary sequences $\sigma = \{\sigma(i) : i \in \mathbb{Z}\}$ on the lattice $\mathbb{Z}$, where $\sigma(i)$ takes values in a finite set $\mathcal{A}$. The joint distribution of $\{\sigma(i) : i \in \mathbb{Z}\}$ is denoted by $\mathbb{P}$. We treat the case where $\mathbb{P}$ is a Gibbs measure with exponentially decaying interaction, see section 2.3 below for details. The configuration space $\Omega = \mathcal{A}^{\mathbb{Z}}$ is endowed with the product topology (making it into a compact metric space). The set of finite subsets of $\mathbb{Z}$ is denoted by $\mathcal{S}$. For $V, W \in \mathcal{S}$ we put $d(V, W) = \min\{|i-j| : i \in V, j \in W\}$. For $V \in \mathcal{S}$, the diameter is defined via $diam(V) = \max\{|i-j|, i, j \in V\}$. For $V \in \mathcal{S}$, $\mathcal{F}_V$ is the sigma-field generated by $\{\sigma(i) : i \in A\}$. For $V \in \mathcal{S}$ we put $\Omega_V = \mathcal{A}^V$. For $\sigma \in \Omega$ and $V \in \mathcal{S}$, $\sigma_V \in \Omega_V$ denotes the restriction of σ to V. For $i \in \mathbb{Z}$ and $\sigma \in \Omega$, $\tau_i\sigma$ denotes the translation of σ by i: $\tau_i\sigma(j) = \sigma(i+j)$. For a local event $E \subseteq \Omega$ the dependence set of E is defined the minimal $V \in \mathcal{S}$ such that E is $\mathcal{F}_V$ measurable. We denote $1\!\!1$ for the indicator function.

2.1 Patterns and cylinders

For $n \in \mathbb{N}, n \geq 1$ let $C_n = [1, n] \cap \mathbb{Z}$. An element $A_n \in \Omega_{C_n}$ is called a n-pattern or a pattern of size n. For a pattern $A_n \in \Omega_{C_n}$ we define the corresponding cylinder $\mathscr{C}(A_n) = \{\sigma \in \Omega : \sigma_{C_n} = A_n\}$. The collection of all n-cylinders is denoted by $\mathscr{C}_n = \cup_{A_n \in \Omega_{C_n}} \mathscr{C}(A_n)$. Sometimes, to denote the probability of the cylinder associated to the pattern A_n, we will use the abbreviation

$$\mathbb{P}(A_k) := \mathbb{P}(\mathscr{C}(A_k)) = \mathbb{P}(\sigma_{C_k} = A_k) \tag{2.1}$$

For $A_k = (\sigma(1), \sigma(2), \ldots, \sigma(k))$ a k-pattern and $1 \leq i \leq j \leq n$ we define the pattern $A_k(i, j)$ to be the pattern of length $j - i + 1$ consisting of the symbols $(\sigma(i), \sigma(i+1), \ldots, \sigma(j))$. For two patterns A_k, B_l we define their concatenation $A_k B_l$ to be the pattern of length $k + l$ consisting of the k symbols of A_k followed by the l symbols of B_l. Concatenation of three or more patterns follows obviously from this.

2.2 Shift-Matches

We will study properties of the following basic quantities.

Definition 2.2 (Number of shift-matches). *For every configuration $\sigma \in \Omega$ and for every $n \in \mathbb{N}$, $k \in \mathbb{N}$, with $k \leq n$, we define the number of matches with shift of length k up to n as*

$$\begin{aligned} N(\sigma, n, k) &= \frac{1}{2} \sum_{i=0}^{n-k} \sum_{j=0, j\neq i}^{n-k} 1\!\!1\{(\tau_i\sigma)_{C_k} = (\tau_j\sigma)_{C_k}\} \qquad (2.3) \\ &= \sum_{i\neq j=0}^{n-k} 1\!\!1\left(\sigma(i+1) = \sigma(j+1), \sigma(i+2) = \sigma(j+2), \ldots, \sigma(i+k) = \sigma(j+k)\right) . \end{aligned}$$

Definition 2.4 (Maximal shift-matching). *For every configuration $\sigma \in \Omega$ and for every $n \in \mathbb{N}$ we define $M(\sigma, n)$ to be the maximal length of a shift-matching up to n, that is the maximal $k \in \mathbb{N}$ (with $k \leq n$) such that there exist $i \in \mathbb{N}$ and $j \in \mathbb{N}$ (with $0 \leq i < j \leq n - k$) satisfying*

$$(\tau_i \sigma)_{C_k} = (\tau_j \sigma)_{C_k} \tag{2.5}$$

where we adopt the convention $\max(\varnothing) = 0$*.*

Definition 2.6 (First occurrence of a shift-matching). *For every configuration $\sigma \in \Omega$ and for every $k \in \mathbb{N}$ we define $T(\sigma, k)$ to be the first occurrence of a shift-match, that is the minimal $n \in \mathbb{N}$ (with $k \leq n$) such that there exist $i \in \mathbb{N}$ and $j \in \mathbb{N}$ (with $0 \leq i < j \leq n - k$) satisfying*

$$(\tau_i \sigma)_{C_k} = (\tau_j \sigma)_{C_k}. \tag{2.7}$$

where we adopt the convention $\min(\varnothing) = \infty$*.*

The following proposition follows immediately from these definitions.

Proposition 2.8. *The probability distributions of the previous quantities are related by the following "duality" relations:*

$$\mathbb{P}(N(\sigma, n, k) = 0) = \mathbb{P}(M(\sigma, n) < k) = \mathbb{P}(T(\sigma, k) > n) \; . \tag{2.9}$$

2.3 Gibbs measures

We now state our assumptions on $\mathbb{P}$, and recall some basic facts about Gibbs measures. The reader familiar with this can skip this section.

We choose for $\mathbb{P}$ the unique Gibbs measure corresponding to an exponentially decaying translation-invariant interaction. In dynamical systems language this corresponds to the unique equilibrium measure of a Hölder continuous potential.

Definition 2.10. *A translation-invariant interaction is a map*

$$U : \mathcal{S} \times \Omega \to \mathbb{R}, \tag{2.11}$$

such that the following conditions are satisfied:

1. *$U(A, \sigma)$ depends on $\sigma(i)$, with $i \in A$ only.*

2. *Translation invariance:*

$$U(A + i, \tau_{-i}\sigma) = U(A, \sigma) \qquad \forall A \in \mathcal{S}, i \in \mathbb{Z}, \sigma \in \Omega. \tag{2.12}$$

3. *Exponential decay: there exist $\gamma > 0$ such that*

$$||U||_\gamma := \sum_{A \ni 0} e^{\gamma diam(A)} \sup_{\sigma \in \Omega} |U(A, \sigma)| < \infty \; . \tag{2.13}$$

The set of all such interactions is denoted by $\mathcal{U}$. Here are some standard examples of elements of $\mathcal{U}$.

1. Ising model with magnetic field h: $\mathcal{A} = \{-1, 1\}$, $U(\{i, i+1\}, \sigma) = J\sigma_i\sigma_{i+1}$, $U(\{i\}, \sigma) = h\sigma_i$ and all other $U(A, \sigma) = 0$.

2. General finite range interactions. An interaction U is called *finite-range* if there exists an $R > 0$ such that $U(A, \sigma) = 0$ for all $A \in \mathcal{S}$ with $\text{diam}(A) > R$.

3. Long range Ising models $U(\{i, j\}, \sigma) = J_{j-i}\sigma_i\sigma_j$ with $|J_k| \leq e^{-\gamma k}$ for some $\gamma > 0$ and all other $U(A, \sigma) = 0$

For $U \in \mathcal{U}$, $\zeta \in \Omega$, $\Lambda \in \mathcal{S}$, we define the finite-volume Hamiltonian with boundary condition ζ as

$$H_\Lambda^\zeta(\sigma) = \sum_{A \cap \Lambda \neq \varnothing} U(A, \sigma_\Lambda \zeta_{\Lambda^c}) \tag{2.14}$$

and the Hamiltonian with free boundary condition as

$$H_\Lambda(\sigma) = \sum_{A \subseteq \Lambda} U(A, \sigma), \tag{2.15}$$

which depends only on the spins inside Λ.

Corresponding to the Hamiltonian in (2.14) we have the finite-volume Gibbs measures $\mathbb{P}_\Lambda^{U,\zeta}$, $\Lambda \in \mathcal{S}$, defined on Ω by

$$\int f(\xi)\, d\mathbb{P}_\Lambda^{U,\zeta}(\xi) = \sum_{\sigma_\Lambda \in \Omega_\Lambda} f(\sigma_\Lambda \zeta_{\Lambda^c}) \frac{e^{-H_\Lambda^\zeta(\sigma)}}{Z_\Lambda^\zeta}\,, \tag{2.16}$$

where f is any continuous function and Z_Λ^ζ denotes the partition function normalizing $\mathbb{P}_\Lambda^{U,\zeta}$ to a probability measure:

$$Z_\Lambda^\zeta = \sum_{\sigma_\Lambda \in \Omega_\Lambda} e^{-H_\Lambda^\zeta(\sigma)} \tag{2.17}$$

For a probability measure $\mathbb{P}$ on Ω, we denote by $\mathbb{P}_\Lambda^\zeta$ the conditional probability distribution of $\sigma(i), i \in \Lambda$, given $\sigma_{\Lambda^c} = \zeta_{\Lambda^c}$. Of course, this object is only defined on a set of $\mathbb{P}$-measure one. For $\Lambda \in \mathcal{S}, \Gamma \in \mathcal{S}$ and $\Lambda \subseteq \Gamma$, we denote by $\mathbb{P}_\Gamma(\sigma_\Lambda | \zeta)$ the conditional probability to find σ_Λ inside Λ, given that ζ occurs in $\Gamma \setminus \Lambda$.

Definition 2.18. *For $U \in \mathcal{U}$, we call $\mathbb{P}$ a Gibbs measure with interaction U if its conditional probabilities coincide with the ones prescribed in (2.16), i.e., if*

$$\mathbb{P}_\Lambda^\zeta = \mathbb{P}_\Lambda^{U,\zeta} \qquad \mathbb{P} - a.s. \qquad \Lambda \in \mathcal{S}, \zeta \in \Omega. \tag{2.19}$$

In our situation, with $U \in \mathcal{U}$, the Gibbs measure $\mathbb{P}$ corresponding to U is unique. Moreover it satisfies the following strong mixing condition: for all $V, W \in \mathcal{S}$ and all events $A \in \mathcal{F}_V$, $B \in \mathcal{F}_W$

$$\left| \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} - \mathbb{P}(A) \right| \leq e^{-cd(V,W)} \tag{2.20}$$

where $c > 0$ depends of course on the interaction U.

2.4 Thermodynamic quantities

We now recall some definitions of basic important statistical mechanics quantities.

Definition 2.21. *The pressure $p\,(U)$of the Gibbs measure $\mathbb{P}$ associated with the interaction U is defined as*

$$p\,(U) = \lim_{n\to\infty} \frac{1}{n} \log Z_n \tag{2.22}$$

where

$$Z_n = \sum_{\sigma_{C_n}\in\Omega_{C_n}} \exp(-\sum_{A\subseteq C_n} U(A,\sigma))$$

is the partition function with the free boundary conditions.

Definition 2.23. *The entropy $s(U)$ of the Gibbs measure $\mathbb{P}$ associated with the interaction U is defined as*

$$s(U) = \lim_{n\to\infty} -\frac{1}{n} \sum_{A_n\in\Omega_{C_n}} \mathbb{P}(\mathscr{C}(A_n)) \log \mathbb{P}(\mathscr{C}(A_n))\,. \tag{2.24}$$

In terms of the interaction U we have the following basic thermodynamic relation between pressure, entropy and the Gibbs measure $\mathbb{P}$ corresponding to U.

$$s(U) = p\,(U) + \int f_U \; d\mathbb{P}, \tag{2.25}$$

where

$$f_U(\sigma) = \sum_{A\ni 0} \frac{U(A,\sigma)}{|A|}$$

denotes the average internal energy per site.

We also have the following relation between f_U and the Hamiltonian:

$$H^{\xi}_{\Lambda}(\sigma) = \sum_{i\in\Lambda} \tau_i f_U(\sigma) + O(1) \tag{2.26}$$

where $O(1)$ is a quantity which is uniformly bounded in Λ, σ, ξ.

The function f_U is what is called the potential in the dynamical systems literature. An exponentially decaying interaction U then corresponds to a Hölder continuous potential f_U.

The following is a standard property of (one-dimensional) Gibbs measures with interaction $U \in \mathcal{U}$. For the proof see [3].

Proposition 2.27. *For the unique Gibbs measure $\mathbb{P}$ with interaction U there exists a constant $\gamma > 1$ such that for any configuration $\sigma \in \Omega$ and for any pattern $A_k \in \Omega_{C_k}$, we have*

$$\gamma^{-1} e^{-kp(U)}\; e^{-H(\mathscr{C}(A_k))} \le \mathbb{P}(\mathscr{C}(A_k)) \le \gamma e^{-kp(U)}\; e^{-H(\mathscr{C}(A_k))} \tag{2.28}$$

Two other well-known properties of Gibbs measures in $d = 1$ which will be used often are listed below.

Proposition 2.29. *For the unique Gibbs measure $\mathbb{P}$ corresponding to the interaction $U \in \mathcal{U}$, there are constants $\rho < 1$ and $c > 0$, such that for all $A_k \in \Omega_{C_k}$ and for all $\eta \in \Omega$,*

$$\mathbb{P}(\sigma_{C_k} = A_k) \le \rho^k\,. \tag{2.30}$$

and

$$c^{-1}\mathbb{P}(\sigma_{C_k} = A_k) \le \mathbb{P}(\sigma_{C_k} = A_k|\eta_{\mathbb{Z}\setminus C_k}) \le \mathbb{P}(\sigma_{C_k} = A_k)c \tag{2.31}$$

2.5 Useful lemmas

In the proofs of our theorems we will make frequently use of the following results.

Lemma 2.32. *For $q \geq 0$ the function $\frac{p(qU)}{q}$ is non increasing.*

Proof. From the definition of $p(U)$ and $s(U)$ and from the thermodynamic relation (2.25), which is equivalent to $s = p - q\frac{dp}{dq}$, it follows immediately

$$\frac{d}{dq}\left(\frac{p(qU)}{q}\right) = -\frac{s(qU)}{q^2}\,.$$

The claim is then a consequence of the positivity of the entropy. □

In order to state the next lemma, we need the following notation which will be used through the whole paper.

Definition 2.33. *Let a_k and b_k be two sequences of positive numbers. Then we write*

$$a_k \approx b_k$$

if $\log(a_k) - \log(b_k)$ *is a bounded sequence and*

$$a_k \preceq b_k$$

if

$$a_k \leq c_k$$

with $c_k \approx b_k$.

Note that we have that $\approx$ and $\preceq$ "behave" as ordinary equalities and inequalities and are "compatible" with usual equalities and inequalities. E.g., if $a_k \preceq b_k$ and $b_k \approx c_k$, then $a_k \preceq c_k$, if $a_k \approx b_k$ and $b_k \leq c_k$ then $a_k \preceq c_k$, etc.

Lemma 2.34. *Define*

$$\alpha = p(U) - \frac{p(2U)}{2} \tag{2.35}$$

We have $\alpha > 0$ and

$$\sum_{A_k \in \Omega_{C_k}} [\mathbb{P}(\sigma_{C_k} = A_k)]^2 \approx e^{-2k\alpha} \tag{2.36}$$

while for $s > 2$

$$\sum_{A_k \in \Omega_{C_k}} [\mathbb{P}(\sigma_{C_k} = A_k)]^s \preceq e^{-sk\alpha} \tag{2.37}$$

Proof. The positivity of α follows from lemma 2.32. From proposition 2.27 we obtain

$$\sum_{A_k \in \Omega_{C_k}} [\mathbb{P}(\sigma_{C_k} = A_k)]^2 \approx \sum_{A_k \in \Omega_{C_k}} e^{-2kp(U)} e^{-2H(\mathscr{C}(A_k))} \approx e^{-2k[p(U) - \frac{p(2U)}{2}]} = e^{-2\alpha k}$$

For $s > 2$ we have

$$\sum_{A_k \in \Omega_{C_k}} \mathbb{P}(\sigma_{C_k} = A_k)^s \approx \sum_{A_k \in \Omega_{C_k}} e^{-skp(U)} e^{-sH(\mathscr{C}(A_k))} \approx e^{-sk[p(U) - \frac{p(sU)}{s}]} \leq e^{-s\alpha k}$$

where in the last inequality we have used the monotonicity property of lemma 2.32. □

3 The average number of shift matches

We will focus on the quantity $N(\sigma, n, k)$ and we will study how the number of shift-matchings behaves when the size of the matching, k, is varied as function of the string length, n. It is clear that when $k = k(n)$ is very large (say of the order of n) then there will be no matching of size k with probability close to one, in the limit $n \to \infty$. On the other hand, if $k = k(n)$ is too small, then the number of shift-matchings will be very large with probability close to one. We want to identify a scale $k^*(n)$ such that $N(\sigma, n, k^*(n))$ will have a non trivial distribution. Our first result concerns the average of $N(\sigma, n, k)$. Define

$$k^*(n) = \frac{\ln n}{\alpha} \,, \tag{3.1}$$

with α as in (2.35). For sequences $k'(n)$ and $k(n)$ we write $k(n) >> k'(n)$ if $k(n) - k'(n) \to \infty$ as $n \to \infty$.

Then we have the following result.

Theorem 3.2. *Let $\{k(n)\}_{n\in\mathbb{N}}$ be a sequence of integers. Then we have*

1. *If $k^*(n) >> k(n)$, then $\lim_{n\to\infty} \mathbb{E}\big(N(\sigma, n, k(n))\big) = \infty$.*
2. *If $k(n) >> k^*(n)$, then $\lim_{n\to\infty} \mathbb{E}\big(N(\sigma, n, k(n))\big) = 0$.*
3. *If $k(n) - k^*(n)$ is a bounded sequence, then we have*

$$0 < \liminf_{n\to\infty} \mathbb{E}(N(\sigma, n, k(n)) \le \limsup_{n\to\infty} \mathbb{E}(N(\sigma, n, k(n)) < \infty \tag{3.3}$$

Proof. We will assume (without loss of generality) that the sequence is such that

$$\lim_{n\to\infty} \frac{k(n)}{n} = 0.$$

We may rewrite $N(\sigma, n, k)$ by summing over all possible patterns of length k

$$N(\sigma, n, k) = \sum_{i=0}^{n-k} \sum_{j=i+1}^{n-k} \sum_{A_k \in \Omega_{C_k}} 1\!\!I\{(\tau_i\sigma)_{C_k} = (\tau_j\sigma)_{C_k} = A_k\} \,.$$

We split the above sum into two sums, one (S_0) corresponding to absence of overlap between $(\tau_i\sigma)_{C_k}$ and $(\tau_j\sigma)_{C_k}$ (i.e. the indices i and j are more than k far apart) and one (S_1) where there is overlap:

$$S_0 = \sum_{i=0}^{n-2k} \sum_{j=i+1+k}^{n-k} \sum_{A_k \in \Omega_{C_k}} 1\!\!I\{(\tau_i\sigma)_{C_k} = (\tau_j\sigma)_{C_k} = A_k\} \,,$$

$$S_1 = \sum_{i=0}^{n-k} \sum_{j=i+1}^{i+k} \sum_{A_k \in \Omega_{C_k}} 1\!\!I\{(\tau_i\sigma)_{C_k} = (\tau_j\sigma)_{C_k} = A_k\} \,.$$

We have of course $\mathbb{E}(N(\sigma,n,k)) = \mathbb{E}(S_0) + \mathbb{E}(S_1)$. In order to prove the first statement of the theorem it suffices to show that $\mathbb{E}(S_0)$ diverges under the hypothesis $k^*(n) >> k(n)$. Using translation invariance one has

$$\begin{aligned}
\mathbb{E}(S_0) &= \sum_{l=k}^{n-k}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}\mathbb{P}\big(\sigma_{C_k}=(\tau_l\sigma)_{C_k}=A_k\big)\\
&= \sum_{l=k}^{n-k}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}\mathbb{P}(\sigma_{C_k}=A_k)\,\mathbb{P}((\tau_l\sigma)_{C_k}=A_k\,|\,\sigma_{C_k}=A_k)\;.
\end{aligned}$$

Because of the mixing conditions (2.20) we have

$$\mathbb{E}(S_0) \;=\; \sum_{l=k}^{n-k}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}[\mathbb{P}(\sigma_{C_k}=A_k)]^2+\Delta(n,k) \tag{3.4}$$

where the error $\Delta(n,k)$ is bounded by

$$|\Delta(n,k)| \le \mathcal{O}(1)\sum_{l=k}^{n-k}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}\mathbb{P}(\sigma_{C_k}=A_k)^2e^{-c(l-k)}$$

Using the mixing property (2.20) and Lemma 2.34 the error can be bounded by

$$|\Delta(n,k)| \le \mathcal{O}(1)e^{-2\alpha k}\sum_{m=0}^{n-2k}(n-2k-m+1)e^{-cm} \le \mathcal{O}(1)e^{-2\alpha k} \tag{3.5}$$

On the other hand, applying lemma 2.34, we have that

$$\sum_{l=k+1}^{n-k}(n-k+1-l)\sum_{A_k}\mathbb{P}(A_k)^2 \approx (n-2k)^2e^{-2\alpha k} \tag{3.6}$$

Combining together (3.4), (3.5) and (3.6), we obtain

$$(n-2k)^2e^{-2\alpha k} \preceq \mathbb{E}(N(\sigma,n,k)) \tag{3.7}$$

which proves statement 1 of the theorem.

To prove statement 2 we have to control $\mathbb{E}(S_1)$, which is the contribution to $\mathbb{E}(N(\sigma,n,k)$ due to self-overlapping cylinders. Using translation invariance we have

$$\mathbb{E}(S_1) = \sum_{l=1}^{k-1}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}\mathbb{P}\big(\sigma_{C_k}=(\tau_l\sigma)_{C_k}=A_k\big)\;.$$

We further split this in two sums, namely $\mathbb{E}(S_1) = \mathbb{E}(S_1') + \mathbb{E}(S_1'')$ with

$$\mathbb{E}(S_1') = \sum_{l=1}^{\lfloor k/2\rfloor}(n-k+1-l)\sum_{A_k\in\Omega_{C_k}}\mathbb{P}\big(\sigma_{C_k}=(\tau_l\sigma)_{C_k}=A_k\big)\,, \tag{3.8}$$

$$\mathbb{E}(S_1'') = \sum_{l=\lfloor k/2\rfloor+1}^{k-1} (n-k+1-l) \sum_{A_k\in\Omega_{C_k}} \mathbb{P}\big(\sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k\big)\,. \tag{3.9}$$

Let us consider first $\mathbb{E}(S_1'')$, i.e., $\lfloor k/2\rfloor < l < k$. In this case the overlap between C_k and $\tau_l C_k$ imposes that the sum over cylinders of length k can be reduced to a sum over cylinders of length l. In the notation of subsection 2.1, we have the following inequality

$$\mathbb{1}(\sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k) \le \mathbb{1}(\sigma_{C_{l+k}} = A_k(1,l)A_k(1,l)A_k(1,k-l)) \tag{3.10}$$

In fact, if the pattern A_k is such that the set $\{\sigma\in\Omega : \sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k\}$ is not empty, then we have equality in (3.10). Hence,

$$\begin{aligned}
\sum_{A_k\in\Omega_k} \mathbb{P}\left(\sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k\right) &= \sum_{A_l}\sum_{B_{k-l}} \mathbb{P}\left(\sigma_{C_k} = A_lB_{k-l}, (\tau_l\sigma)_{C_k} = A_lB_{k-l}\right) \\
&\le \sum_{A_l} \mathbb{P}(\sigma_{C_{l+k}} = A_lA_lA_l(1,k-l)) \\
&\preceq \sum_{A_l} \mathbb{P}(A_l)^2\mathbb{P}(A_l(1,k-l))
\end{aligned} \tag{3.11}$$

where in the first inequality we used the fact that contributions with $B_{k-l} \ne A_l(1,k-l)$ are zero. Therefore, using proposition 2.29, we obtain

$$\mathbb{E}(S_1'') \preceq \sum_{l=\lfloor k/2\rfloor+1}^{k} (n-k-l) \sum_{A_l} \mathbb{P}(A_l)^2\rho^{k-l}$$

From this we deduce, thanks to lemma 2.34,

$$\begin{aligned}
\mathbb{E}(S_1'') &\preceq (n-k) \sum_{l=\lfloor k/2\rfloor+1}^{k} e^{-2l\alpha}\rho^{k-l} \\
&\le (n-k)e^{-k\alpha} \sum_{l=\lfloor k/2\rfloor+1}^{k} \rho^{k-l} \\
&\le (n-k)e^{-k\alpha} \sum_{x=0}^{\infty} \rho^{x} \\
&\approx (n-k)e^{-k\alpha}\,.
\end{aligned} \tag{3.12}$$

We now treat $\mathbb{E}(S_1')$, i.e. the case with $1 \le l \le \lfloor k/2\rfloor$. Write $k = rl + q$ with r and s integers, $r \ge 2$, $0 \le q \le l-1$. If the set $\{\sigma : \sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k\}$ is not empty, then the pattern A_k has to consist of $r+1$ repetitions of the subpattern $A_k(1,l)$ followed by a subpattern $A_k(1,q)$, where q is such that $(r+1)l + q = k + l$. Hence

$$\mathbb{1}(\sigma_{C_k} = (\tau_l\sigma)_{C_k} = A_k) \le \mathbb{1}(\sigma_{C_{k+l}} = \underbrace{A_k(1,l)\cdots A_k(1,l)}_{r+1 \text{ times}} A_k(1,q)) \tag{3.13}$$

At this stage one could repeat the same approach as in the previous estimate for $\mathbb{E}(S_1'')$ by immediately employing proposition 2.29. However, this approach would not work because the repeating blocks are two small. To circumvent this, we observe that in the pattern

$[A_k(1,l)]^{r+1}A_k(1,q)$ there exists a piece of length $\lfloor k/2 \rfloor$ which occurs at least two times, and the remaining l symbols are fixed by that piece. Therefore, using proposition 2.29

$$\sum_{A_k \in \Omega_k} \mathbb{P}\left(\sigma_{C_k} = (\tau_l \sigma)_{C_k} = A_k\right) \leq \sum_{B_{k/2}} \mathbb{P}(B_{k/2})^2 \rho^l \, . \tag{3.14}$$

By inserting (3.14) in (3.8) and using lemma 2.34, we finally have

$$\mathbb{E}(S_1') \preceq (n-k)e^{-k\alpha} \, . \tag{3.15}$$

Combining together the estimates (3.6), (3.12) and (3.15) we obtain so far

$$\mathbb{E}(N(\sigma, n, k)) \preceq (n-k)e^{-k\alpha} + (n-2k)^2 e^{-2k\alpha} \tag{3.16}$$

from which statement 2 of the theorem follows.

Finally, combining (3.7),(3.16) gives statement 3 of the theorem.

□

4 Second moment estimate

In this section we will show that the random variable $N(\sigma, n, k(n))$ converges in probability to $+\infty$ in the regime where $k(n) << k^*(n)$ while it converges to 0 in the opposite regime $k(n) >> k^*(n)$. Finally, if the difference $k(n) - k^*(n)$ is bounded, then we show that $N(\sigma, n, k(n))$ is tight and does not converge to zero in distribution. These results will follow as an application of the method of first moment and second moment, respectively.

Theorem 4.1. *Let $\{k(n)\}_{n \in \mathbf{N}}$ be a sequence of integers. For every positive $m \in \mathbb{N}$*

1. *If $k^*(n) >> k(n)$, then $\lim_{n\to\infty} \mathbb{P}\big(N(\sigma, n, k(n)) \leq m\big) = 0$.*
2. *If $k(n) >> k^*(n)$, then $\lim_{n\to\infty} \mathbb{P}\big(N(\sigma, n, k(n)) \geq m\big) = 0$.*
3. *If $k(n) - k^*(n)$ is bounded, then $N(\sigma, n, k(n))$ is tight and does not converge to zero in distribution. More precisely, we have that there exists a constant $C > 0$ such that*

$$\limsup_{n\to\infty} \mathbb{P}\left(N(\sigma, n, k(n)) > m\right) \leq C/m \tag{4.2}$$

and

$$\liminf_{n\to\infty} \mathbb{P}\left(N(\sigma, n, k(n)) > 0\right) > 0 \tag{4.3}$$

Proof. We will assume, once more, without loss of generality that

$$\lim_{n\to\infty} \frac{k(n)}{n} = 0.$$

Statement 2 and (4.2) follow from theorem (3.2) and the Markov inequality. To prove statement 1, and (4.3), we use the Paley-Zigmund inequality, which gives that for all $0 \leq a \leq 1$

$$\mathbb{P}\big(N \geq a\mathbb{E}(N)\big) \geq (1-a)^2 \frac{\mathbb{E}(N)^2}{\mathbb{E}(N^2)} \tag{4.4}$$

We fix now a sequence $k_n \uparrow \infty$ such that $k_n^*/k_n \to \infty$ Consider the auxiliary random variable

$$\mathcal{N}_n := \sum_{i,j=0,|i-j|>2k_n}^{n-k_n} 1\!\!1\left((\tau_i\sigma)_{C_{k_n}} = (\tau_j\sigma)_{C_{k_n}}\right) \tag{4.5}$$

Clearly, to obtain statement 1, it is sufficient that $\mathcal{N}_n$ goes to infinity with probability one. On the other hand, using the first moment computations of the previous section, we have

$$\mathbb{E}(\mathcal{N}_n) \approx n^2 e^{-2\alpha k_n} \tag{4.6}$$

So, in order to use the Paley-Zigmund inequality, it is sufficient to show that

$$\mathbb{E}(\mathcal{N}_n^2) \preceq \xi_n^4 \tag{4.7}$$

where we introduced the notation

$$\xi_n := n e^{-\alpha k_n} \tag{4.8}$$

Remark that $\xi_n \to \infty$ for our choice of k_n (as in statement 1).

Indeed, if we have (4.7) in the regime $k^*(n) >> k(n)$, then the ratio

$$\frac{\mathbb{E}(\mathcal{N}^2)}{(\mathbb{E}(\mathcal{N}))^2}$$

remains bounded from above as $n \to \infty$, and hence, using (4.4), $\mathcal{N}_n$ diverges with probability at least $\delta > 0$. Therefore, in that case, by ergodicity, $N(\sigma, n, k_n) \geq \mathcal{N}_n$ goes to infinity with probability one, since the set of σ such that $N(\sigma, n, k_n)$ goes to infinity is translation invariant.

To see how statement (4.3) follows from (4.7) in the regime where $k(n) - k^*(n)$ is bounded, use the (more classical) second moment inequality

$$\mathbb{P}(\mathcal{N} > 0) \geq \frac{(\mathbb{E}(\mathcal{N}))^2}{\mathbb{E}(\mathcal{N}^2)}$$

combined with

$$N(\sigma, n, k(n)) \geq \mathcal{N}$$

We now proceed with the proof of (4.7). We have,

$$\mathbb{E}(\mathcal{N}_n^2) = \sum_{i,j,r,s,|i-j|>2k_n,|r-s|>2k_n} \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n})_i (A_{k_n})_j (B_{k_n})_r (B_{k_n})_s\right) \tag{4.9}$$

where we use the abbreviate notation $(A_{k_n})_i$ for the event $(\tau_i\sigma)_{C_{k_n}} = A_{k_n}$. Similarly, if we have a word of length l say, consisting of p symbols of A_p followed by $l-p$ symbols of B_{l-p}, we write $(A_p B_{l-p})_i$ for the event that this word appears at location i, i.e., the event $(\tau_i\sigma)_{C_l} = A_p B_{l-p}$.

The sum in the rhs of (4.9) will be split into different sums, according to the amount of overlap in the set of indices $\{i, j, r, s\}$. By this we mean the following: we say that there is *overlap* between two indices i, j if $|i-j| < k_n$. The number of overlaps of a set of indices $\{i, j, r, s\}$ is denoted by $\theta(i, j, r, s)$ and is the number of unordered pairs of indices which have overlap. Since we restrict in the sum (4.9) to $|i-j| > 2k_n, |r-s| > 2k_n$, it follows from the triangular inequality that in that case $\theta(i,j,r,s) \leq 2$. Therefore, we split the sum into three cases

$$\sum_{i,j,r,s,|i-j|>2k_n,|r-s|>2k_n} \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n})_i(A_{k_n})_j(B_{k_n})_r(B_{k_n})_s\right) = S_0 + S_1 + S_2 \tag{4.10}$$

where

$$S_p = \sum_{(i,j,r,s)\in K_{k,p}} \sum_{A,B} \mathbb{P}\left((A_{k_n})_i(A_{k_n})_j(B_{k_n})_r(B_{k_n})_s\right) \tag{4.11}$$

where we abbreviated

$$K_{k_n,p} = \{(i,j,r,s) : |i-j| > 2k_n, |r-s| > 2k_n, \theta(i,j,r,s) = p\} \tag{4.12}$$

to be the set of indices such that the overlap is p.

1. Zero Overlap: S_0
We use lemma 2.34, and notation (4.8)

$$S_0 \preceq \sum_{i,j,r,s} \sum_{A_{k_n},B_{k_n}} \mathbb{P}(A_{k_n})^2 \mathbb{P}(B_{k_n})^2 \preceq \xi_n^4 \tag{4.13}$$

2. One overlap: S_1
We treat the case $|i-r| < k_n$, $i < r < j < s$. The other cases are treated in exactly the same way. Put $A_{k_n} = [a_1 a_2, \ldots a_{k_n}]$, $B_{k_n} = [b_1 b_2, \ldots b_{k_n}]$. The intersection $(A_{k_n})_i \cap (B_{k_n})_r$ is non-empty if and only if $a_r = b_1$, $a_{r+1} = b_2, \ldots a_{k_n} = b_{k_n-r+1}$, i.e., the last $k_n - r + 1$ symbols of A_{k_n} are equal to the first $k_n - r + 1$ symbols of B_{k_n}.

Therefore, we obtain that the sum over the patterns A_{k_n}, B_{k_n} in S_1 equals

$$\begin{aligned}
& \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n})_i(A_{k_n})_j(B_{k_n})_r(B_{k_n})_s\right) \\
= & \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n}B_{k_n}(k_n-r,k_n))_i(A_{k_n})_j(A_{k_n}(r,k_n)B_{k_n}(k_n-r,k_n))_s\right) \\
\preceq & \sum_{A_{k_n},B_{k_n}} \mathbb{P}(A_{k_n}(r,k_n))^3\, \mathbb{P}(A_{k_n}(1,r-1))^2\, \mathbb{P}(B_{k_n}(k_n-r,k_n))^2 \\
\preceq & \; e^{-3(k_n-r)\alpha} e^{-2r\alpha} e^{-2r\alpha}\,.
\end{aligned} \tag{4.14}$$

Summing over the indices $(i,j,r,s) \in K(k_n,1)$ then gives

$$S_1 \preceq n^3 e^{-3\alpha k_n} \sum_{r\leq k_n} e^{-r\alpha} \preceq \xi_n^3 \tag{4.15}$$

3. Two overlaps: S_2
We treat the case $i < r < j < s$ and $r - i < k_n, s - j < k_n$. Other cases are treated in the same way. Put $l_1 := i + k_n - r + 1, p_1 = j + k_n - s + 1$. We suppose $l_1 > p_1$. Then the last l_1 symbols of A_{k_n} have to equal the first l_1 symbols of B_{k_n}, otherwise the intersection $(A_{k_n})_i(A_{k_n})_j(B_{k_n})_r(B_{k_n})_s$ is empty. Therefore, we obtain that the sum over the patterns

A_{k_n}, B_{k_n} in S_2 equals

$$\begin{aligned}
& \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n})_i(A_{k_n})_j(B_{k_n})_r(B_{k_n})_s\right) \\
= & \sum_{A_{k_n},B_{k_n}} \mathbb{P}\left((A_{k_n}B_{k_n-l_1})_i(A_{k_n}B_{k_n-p_1})_j\right) \\
\preceq & \sum_{A_{k_n},B_{k_n}} \mathbb{P}(A_{k_n})^2\mathbb{P}(B_{k_n-l_1})^2\rho^{l_1-p_1} \\
\preceq & \; e^{-2k\alpha}e^{-2(k-l_1)\alpha}\rho^{l_1-p_1} \qquad (4.16)
\end{aligned}$$

Summing over the indices in $K(k,2)$ then gives

$$S_2 \preceq n^2 e^{-2k_n\alpha} \sum_{l_1<k_n} e^{-2\alpha(k_n-l_1)} \sum_{p_1<l_1} \rho^{l_1-p_1} \preceq \xi_n^2 \qquad (4.17)$$

Using the bounds (4.13), (4.15) and (4.17) in eq. (4.9) and (4.10) we deduce (4.7) and then, as explained below, the statement 1 of the theorem follows from Paley-Zigmund inequality. This complete the proof.

□

The following result relates theorem (4.1) and the behavior of the maximal shift-matching, and is the analogue of Theorem 1 in [6] (which is however convergence almost surely for more general comparison of sequences based on scores, but for independent sequences).

Proposition 4.18. *Let $M(\sigma,n)$ be defined as in definition 2.4. Then we have that $M(\sigma,n)/\log n$ converges in probability to α, defined in* (2.35)*.*

Proof. Use the relations of proposition 2.8. We have

$$\mathbb{P}\left(\frac{M(\sigma,n)}{\alpha\log n} \geq (1+\epsilon)\right) \leq \mathbb{P}\left(N(\sigma,n,\lfloor\alpha(1+\epsilon)\log n\rfloor) \geq 1\right)$$

and

$$\mathbb{P}\left(\frac{M(\sigma,n)}{\alpha\log n} < (1-\epsilon)\right) \leq \mathbb{P}\left(N(\sigma,n,\lceil\alpha(1-\epsilon)\log n\rceil) = 0\right)$$

So the result follows from theorem 4.1. □

5 Two independent strings

In this section we study the number of matches with shift when two *independent* sequences σ and η are considered. The marginal distributions of σ and η are denoted with $\mathbb{P}$ and $\mathbb{Q}$, which are chosen to be Gibbs measure with exponentially decaying translation invariant interactions $U(X,\sigma)$ and $V(X,\eta)$, respectively. We assume the two strings belong to the same alphabet $\mathcal{A}$. In analogy with the case of one string we give the following definition.

Definition 5.1 (Number of shift-matches for 2 strings). *For every couple of configurations $\sigma,\eta \in \Omega\times\Omega$ and for every $n\in\mathbb{N}$, $k\in\mathbb{N}$, with $k<n$, we define the number of matches with shift of length k as*

$$N(\sigma,\eta,n,k) = \sum_{i=0}^{n-k}\sum_{j=0,j\neq i}^{n-k} 1\!\!I\{(\tau_i\sigma)_{C_k} = (\tau_j\eta)_{C_k}\}. \qquad (5.2)$$

Of course, in the case $\sigma = \eta$ we recover the previous definition (2.2), i.e. $\tilde{N}(\sigma, \sigma, n, k) = N(\sigma, n, k)$.

5.1 Identical marginal distribution

We treat here the case $\mathbb{Q} = \mathbb{P}$, that is the two sequences σ and η are chosen independently from the same Gibbs distribution $\mathbb{P}$ with interaction $U(X, \sigma)$. Then the results of the previous section are generalized as follows.

Theorem 5.3. *Let $\{k(n)\}_{n\in\mathbb{N}}$ be a sequence of integers.*

1. *If $k^*(n) >> k(n)$, then $\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}\times\mathbb{P}}\left[N(\sigma, \eta, n, k(n))\right] = \infty$.*
2. *If $k^*(n) << k(n)$, then $\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}\times\mathbb{P}}\left[N(\sigma, \eta, n, k(n))\right] = 0$.*
3. *If $k(n) - k^*(n)$ is a bounded sequence, then we have*

$$0 < \liminf_{n\to\infty} \mathbb{E}_{\mathbb{P}\times\mathbb{P}}(N(\sigma, \eta, n, k(n)) \leq \limsup_{n\to\infty} \mathbb{E}_{\mathbb{P}\times\mathbb{P}}(N(\sigma, \eta, n, k(n)) < \infty \tag{5.4}$$

Proof. Because of independence we immediately have

$$\begin{aligned} \mathbb{E}_{\mathbb{P}\times\mathbb{P}}[N(\sigma, \eta, n, k)] &= \sum_{i\neq j=0}^{n-k} \sum_{A_k\in\Omega_k} \mathbb{P}((\tau_i\sigma)_{C_k} = A_k)\, \mathbb{P}((\tau_j\eta)_{C_k} = A_k) \\ &= (n-k)^2 \sum_{A_k\in\Omega_{C_k}} \mathbb{P}(A_k)^2 \\ &\approx (n-k)^2 e^{-2k\alpha}. \end{aligned} \tag{5.5}$$

□

Theorem 5.6. *Let $\{k(n)\}_{n\in\mathbf{N}}$ be a sequence of integers. For every positive $m \in \mathbb{N}$*

1. *If $k^*(n) >> k(n)$, then $\lim_{n\to\infty} \mathbb{P}\times\mathbb{P}[N(\sigma, \eta, n, k(n)) \leq \epsilon] = 0$.*
2. *If $k^*(n) << k(n)$, then $\lim_{n\to\infty} \mathbb{P}\times\mathbb{P}[N(\sigma, \eta, n, k(n)) \geq \epsilon] = 0$.*
3. *If $k(n) - k^*(n)$ is bounded, then $N(\sigma, \eta, n, k(n))$ is tight and does not converge to zero in distribution. More precisely, we have that there exists a constant $C > 0$ such that*

$$\limsup_{n\to\infty} \mathbb{P}\times\mathbb{P}\left(N(\sigma, \eta, n, k(n)) > m\right) \leq C/m \tag{5.7}$$

and

$$\liminf_{n\to\infty} \mathbb{P}\times\mathbb{P}\left(N(\sigma, \eta, n, k(n)) > 0\right) > 0 \tag{5.8}$$

Proof. The strategy of the proof is the as in theorem (4.1). Thus we need to control the second moment to show that $\mathbb{E}(N^2) \approx (\mathbb{E}(N))^2$. We start from

$$\begin{aligned} &\mathbb{E}_{\mathbb{P}\times\mathbb{P}}(N^2(\sigma, \eta, n, k)) = \\ &= \sum_{i_1,j_1,i_2,j_2=1}^{n-k} \sum_{A_k,B_k\in\Omega_k} \mathbb{P}((\tau_{i_1}\sigma)_{C_k} = A_k, (\tau_{i_2}\sigma)_{C_k} = B_k)\, \mathbb{P}((\tau_{j_1}\eta)_{C_k} = A_k, (\tau_{j_2}\eta)_{C_k} = B_k) \end{aligned} \tag{5.9}$$

Using translational invariance and defining new indices $l_1 = i_2 - i_1$ and $l_2 = j_2 - j_1$ we have

$$\mathbb{E}_{\mathbb{P}\times\mathbb{P}}(N^2(\sigma,\eta,n,k)) = \sum_{A_k,B_k\in\Omega_k} \left(\sum_{l_1=1}^{n-k}(n-k+1-l_1)\mathbb{P}(\sigma_{C_k}=A_k\ ,(\tau_{l_1}\sigma)_{C_k}=B_k) \right. \left. \sum_{l_2=1}^{n-k}(n-k+1-l_2)\mathbb{P}(\eta_{C_k}=A_k\ ,(\tau_{l_2}\eta)_{C_k}=B_k)\right)$$

We have to distinguish three kind of contributions in the previous sums:

1. Zero overlap, i.e. $l_1 > k, l_2 > k$. Then

$$\begin{aligned} & \sum_{A_k,B_k\in\Omega_k} \left(\sum_{l_1=k+1}^{n-k}(n-k+1-l_1)\mathbb{P}(\sigma_{C_k}=A_k\ ,(\tau_{l_1}\sigma)_{C_k}=B_k)\right. \\ & \qquad \left.\sum_{l_2=k+1}^{n-k}(n-k+1-l_2)\mathbb{P}(\eta_{C_k}=A_k\ ,(\tau_{l_2}\eta)_{C_k}=B_k)\right) \\ \approx\ & (n-k)^4 \sum_{A_k,B_k\in\Omega_k} \mathbb{P}(A_k)^2\mathbb{P}(B_k)^2 \\ \approx\ & (n-k)^4 e^{-4k\alpha} \end{aligned} \tag{5.10}$$

2. One overlap. We treat the case $l_1 \le k$ and $l_2 > k$ (other cases are treated similarly). We have

$$\begin{aligned} & \sum_{A_k,B_k\in\Omega_k} \left(\sum_{l_1=1}^{k}(n-k+1-l_1)\mathbb{P}(\sigma_{C_k}=A_k\ ,(\tau_{l_1}\sigma)_{C_k}=B_k)\right. \\ & \qquad \left.\sum_{l_2=k+1}^{n-k}(n-k+1-l_2)\mathbb{P}(\eta_{C_k}=A_k\ ,(\tau_{l_2}\eta)_{C_k}=B_k)\right) \\ \approx\ & (n-k)^3 \sum_{l_1=1}^{k} \sum_{D_{l_1},E_{k-l_1},F_{l_1}} \mathbb{P}(D_{l_1}E_{k-l_1}F_{l_1})\mathbb{P}(D_{l_1}E_{k-l_1})\mathbb{P}(E_{k-l_1}F_{l_1}) \\ \approx\ & (n-k)^3 \sum_{l_1=1}^{k} \sum_{D_{l_1},E_{k-l_1},F_{l_1}} \mathbb{P}(D_{l_1})^2\mathbb{P}(E_{k-l_1})^3\mathbb{P}(F_{l_1})^2 \\ \preceq\ & (n-k)^3 \sum_{l_1=1}^{k} e^{-2l_1\alpha}e^{-2l_1\alpha}e^{-3(k-l_1)\alpha} \\ \le\ & (n-k)^3 e^{-3k\alpha} \end{aligned} \tag{5.11}$$

3. Two overlaps, We treat the case $l_1 < l_2 \le k$ (other cases are treated similarly). We

have

$$\begin{aligned}
&\sum_{A_k,B_k\in\Omega_k}\left(\sum_{l_1=1}^{k}(n-k+1-l_1)\mathbb{P}(\sigma_{C_k}=A_k\ ,(\tau_{l_1}\sigma)_{C_k}=B_k)\right.\\
&\qquad\left.\sum_{l_2=1}^{k}(n-k+1-l_2)\mathbb{P}(\eta_{C_k}=A_k\ ,(\tau_{l_2}\eta)_{C_k}=B_k)\right)\\
\approx\ &(n-k)^2\sum_{l_1,l_2=1}^{k}\ \sum_{D_{l_1},E_{l_2-l_1},F_{k-l_2},G_{l_1},H_{l_2-l_1}}\mathbb{P}(D_{l_1}E_{l_2-l_1}F_{k-l_2}G_{l_1})\mathbb{P}(D_{l_1}E_{l_2-l_1}F_{k-l_2}G_{l_1}H_{l_2-l_1})\\
\approx\ &(n-k)^2\sum_{l_1,l_2=1}^{k}\sum_{D_{l_1}}\mathbb{P}(D_{l_1})^2\sum_{E_{l_2-l_1}}\mathbb{P}(E_{l_2-l_1})^2\sum_{F_{k-l_2}}\mathbb{P}(F_{k-l_2})^2\sum_{G_{l_1}}\mathbb{P}(G_{l_1})^2\sum_{H_{l_2-l_1}}\mathbb{P}(H_{l_2-l_1})\\
\preceq\ &(n-k)^2\sum_{l_1,l_2=1}^{k}e^{-2l_1\alpha}e^{-2(l_2-l_1)\alpha}e^{-2(k-l_2)\alpha}e^{-2l_1\alpha}\\
\leq\ &(n-k)^2e^{-2k\alpha}
\end{aligned}\tag{5.12}$$

Combining together (5.10), (5.11) and (5.12) and similar expression for other cases with one and two overlap we obtain the second moment condition $\mathbb{E}(N^2)\preceq(\mathbb{E}(N))^2$.

□

5.2 Different marginal distributions

In the case $\mathbb{P}\neq\mathbb{Q}$, the first moment is controlled in an analogous way, but the second moment analysis is different, and in fact as we will show in an example, it can happen for some scale $k_n\to\infty$ that

1. $\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}(N(\sigma,\eta,n,k_n))\to\infty$ as $n\to\infty$
2. $\mathbb{P}\times\mathbb{Q}(N(\sigma,\eta,n,k_n)=0)>e^{-\delta}$ for some $\delta>0$ independent of n.

This means that in order to decide whether $N(\sigma,\eta,n,k_n)$ goes to infinity $\mathbb{P}\times\mathbb{Q}$ almost surely, it is not sufficient to have $\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}(N(\sigma,\eta,n,k_n))\to\infty$.

We start with the case $\mathbb{P}$ and $\mathbb{Q}$ Gibbs measures with potentials U,V resp. and define

$$\tilde{\alpha}=\frac{1}{2}p(U)+\frac{1}{2}p(V)-\frac{1}{2}p(U+V)>0\tag{5.13}$$

and

$$\tilde{k}^*=\frac{\log n}{\tilde{\alpha}}\tag{5.14}$$

then we have

Theorem 5.15. *Let $\{k(n)\}_{n\in\mathbb{N}}$ be a sequence of integers.*

1. *If $\tilde{k}^*(n)>>k(n)$, then* $\lim_{n\to\infty}\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}\big(N(\sigma,\eta,n,k(n))\big)=\infty$.
2. *If $\tilde{k}^*(n)<<k(n)$, then* $\lim_{n\to\infty}\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}\big(N(\sigma,\eta,n,k(n))\big)=0$.

3. If $k(n)-\tilde{k}^*(n)$ is a bounded sequence, then we have

$$0<\liminf_{n\to\infty}\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}(N(\sigma,\eta,n,k(n))\leq\limsup_{n\to\infty}\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}(N(\sigma,\eta,n,k(n))<\infty \tag{5.16}$$

Proof. Start by rewriting

$$N(\sigma,\eta,n,k)=\sum_{i=0}^{n-k}\sum_{j=0,j\neq i}^{n-k}\sum_{A_k\in\Omega_k}\mathbb{1}\{(\tau_i\sigma)_{C_k}=A_k\;,(\tau_j\eta)_{C_k}=A_k\}\;.$$

Taking into account the independence of the measures $\mathbb{P}$ and $\mathbb{Q}$, we obtain:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}\times\mathbb{Q}}(N(\sigma,\eta,n,k)) &= \sum_{i\neq j=0}^{n-k}\sum_{A_k\in\Omega_k}\mathbb{P}((\tau_i\sigma)_{C_k}=A_k)\,\mathbb{Q}((\tau_j\eta)_{C_k}=A_k)\\ &\approx (n-k)^2\sum_{A_k\in\Omega_{C_k}}e^{-kp(U)}e^{-kH_U(\mathscr{C}(A_k))}e^{-kp(V)}e^{-kH_V(\mathscr{C}(A_k))}\\ &\approx (n-k)^2e^{-k[p(U)+p(V)-p(U+V)]}\\ &= (n-k)^2e^{-2k\tilde{\alpha}} \end{aligned}\tag{5.17}$$

where in the second line we made use of translation invariance and proposition 2.27. □

In case 1 of theorem 5.15, we will not in general be able to conclude that $N(\sigma,\eta,n,k(n))$ goes to infinity almost surely as $n\to\infty$. Indeed, if we compute the second moment, we find terms analogous to the case $\mathbb{P}=\mathbb{Q}$, of which now we have to take the $\mathbb{P}\times\mathbb{Q}$ expectation. In particular, the one overlap contribution will contain a term of the order

$$(n-k)^3\sum_{E_k}\mathbb{P}(E_k)\mathbb{Q}(E_k)^2$$

If $\mathbb{P}\neq\mathbb{Q}$ this term may however not be dominated by $n^4e^{-4k\tilde{\alpha}}$. Indeed, the inequality

$$\sum_{E_k}\mathbb{P}(E_k)\mathbb{Q}(E_k)^2\leq\left(\sum_{E_k}\mathbb{P}(E_k)\mathbb{Q}(E_k)\right)^{3/2}$$

is not valid in general. In particular, if $\mathbb{P}$ gives uniform measure to cylinders E_k and $\mathbb{Q}$ concentrates on one particular cylinder, then this inequality will be violated.

As an example, inspired by this, we choose $\mathbb{P}$ to be a Gibbs measure with potential U, and $\mathbb{Q}=\delta_a$, where δ_a denotes the Dirac measure concentrating on the configuration $\eta(x)=a$ for all $x\in\mathbb{Z}$ (which is strictly speaking no Gibbs measure, but a limit of Gibbs measures). In that case $\mathbb{P}\times\mathbb{Q}$ almost surely,

$$N(\sigma,\eta,n,k(n))=n\sum_{i=1}^{n-k}\mathbb{1}\left((\tau_i\sigma)_{C_k}=[a]_k\right)$$

where $[a]_k$ denotes a block of k successive a's. Therefore,

$$\mathbb{P}\times\mathbb{Q}\left(N(\sigma,\eta,n,k(n))=0\right)=\mathbb{P}(\Theta_{[a]_k}(\sigma)\geq n-k)$$

where.

$$\Theta_{[a]_k}(\sigma) = \inf\{j > 0 : \sigma_j = a, \sigma_{j+1} = a, \ldots, \sigma_{j+k-1} = a\}$$

is the hitting time of the pattern $[a]_k$ in the configuration σ. For this hitting time we have the exponential law [2] which gives

$$\mathbb{P}(\Theta_{[a]_k}(\sigma) \geq n) \geq e^{-\lambda \mathbb{P}([a]_k)n}$$

with λ a positive constant not depending on n. Now we choose the scale k_n such that the first moment of $N(\sigma, \eta, n, k(n))$ diverges as $n \to \infty$, i.e., such that

$$n^2 \mathbb{P}([a]_{k_n}) \to \infty$$

Furthermore, we impose that

$$\mathbb{P}([a]_{k_n})n \leq \delta$$

for all n. In that case

$$\mathbb{P}(\Theta_{[a]_{k_n}}(\sigma) \geq n) \geq e^{-\lambda \mathbb{P}([a]_{k_n}))n} \geq e^{-\lambda\delta}$$

which implies $N(\sigma, \eta, n, k_n)$ does not go to infinity $\mathbb{P} \times \mathbb{Q}$ almost surely.

References

[1] M. Abadi, Exponential approximation for hitting times in mixing processes, Math. Phys. Electron. J. **7** (2001).

[2] M. Abadi, J.R. Chazottes, F. Redig and E. Verbitskiy, Exponential distribution for the occurrence of rare patterns in Gibbsian random fields, Comm. Math. Phys. **246**, 269-294 (2004).

[3] R. Bowen, *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, Lecture Notes in Math. **470**, Springer, 1975.

[4] P. Collet, A. Galves, B. Schmitt, *Fluctuations of repetition times for gibbsian sources*, Nonlinearity **12**, 1225–1237 (1999).

[5] A. Dembo, S. Karlin, O. Zeitouni, Limit distribution of maximal non-aligned two-sequence segmental score. Ann. Probab. **22**, 2022–2039 (1994).

[6] A. Dembo, S. Karlin, O. Zeitouni, Critical phenomena for sequence matching with scoring. Ann. Probab. **22** 1993–2021 (1994).

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

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

[9] N.R. Hansen, Local alignment of Markov chains. Ann. Appl. Probab. **16**, 1262–1296 (2006).

[10] D. Ruelle, Thermodynamic formalism. The mathematical structures of classical equilibrium statistical mechanics. Encyclopædia of Mathematics and its Applications **5**. Addison-Wesley Publishing Co., Reading, Mass., 1978