Optimal Gaussian concentration bounds for stochastic chains of unbounded memory

Jean-Rene Chazottes, Sandro Gallo, Daniel Takahashi

Optimal Gaussian concentration bounds for stochastic chains of unbounded memory

Jean-René Chazottes [*1], Sandro Gallo [†2], and Daniel Y. Takahashi [‡3]

[1]Centre de Physique Théorique, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris, France
[2]Departamento de Estatística, Universidade Federal de São Carlos (UFSCar), São Paulo, Brazil
[3]Instituto do Cérebro, Universidade Federal do Rio Grande do Norte (UFRN), Natal, Brazil

January 17, 2020

Abstract

We obtain explicit and optimal Gaussian concentration bounds (GCBs) for stochastic chains of unbounded memory (SCUMs) on countable alphabets. These stochastic processes are also known as "chains with complete connections" or "g-measures". We prove that a GCB holds when the sum of oscillations of the kernel is less than one, or when the variation of the kernel is summable, i.e., belongs to $\ell^1(\mathbb{N})$. The proof is based on maximal coupling. Our conditions are optimal in the sense that we exhibit examples of SCUMs that do not have GCB and the sum of oscillations is strictly larger than one, or the variation belongs to $\ell^{1+\epsilon}(\mathbb{N})$ for any $\epsilon > 0$. These examples are based on the existence of phase transitions. We also extend the validity of GCB to a class of functions which can depend on infinitely many coordinates. We illustrate our results by three main applications. First, we derive a Dvoretzky-Kiefer-Wolfowitz type inequality which gives a uniform control on the fluctuations of the empirical measure. Second, in the finite-alphabet case, we apply our bounds to obtain an upper bound on the $\bar{d}$-distance between two stationary SCUMs and, as a by-product, we obtain new bounds on the speed of Markovian approximation in $\bar{d}$. Third, we obtain exponential rate of convergence for Birkhoff sums of a certain class of observables.

Keywords: concentration inequalities, maximal coupling, chains of infinite order, g-measures, categorical time series, empirical distribution, Dvoretzky-Kiefer-Wolfowitz type inequality, $\bar{d}$-distance, Markovian approximation.

[*]Email: jeanrene@cpht.polytechnique.fr
[†]Email: sandro.gallo@ufscar.br
[‡]Email: takahashiyd@gmail.com

Contents

1 Introduction

Stochastic chains with unbounded memory (SCUMs) are a natural generalization of Markov chains. Their dynamics is provided by a family of probability kernels that describe the probability of observing a symbol at any time given (possibly) the entire past. Such processes first appeared in [41], and, since then, have been intensively studied in different fields under different names. In the literature of stochastic processes [14] coined the name chains with complete connections, while [26] later called the same objects chains of infinite order. In symbolic dynamical systems, stationary SCUMs are studied under the name of g-measures [30, 35, 45, 28]. In applied statistics literature, SCUMs have been used to model various natural phenomena, including some popular stochastic processes, *e.g.*, categorical time series and binary autoregressive models [38, 31, 19, 44]. SCUMs

are also natural dynamical counterpart of Gibbs measures on the lattice $\mathbb{Z}$ in statistical physics and the family of probability kernels has been called left interval specifications [18]. Different fields investigated SCUMs using different techniques, making this family of stochastic processes a rich object to be studied.

One of the main interests in SCUMs comes from the fact that they exhibit different mixing properties depending on the characteristics of the probability kernels. For instance, kernels with strong dependence on the past can have two or more shift invariant measures compatible with the kernel [3, 27, 21, 20, 12, 1]. Weak dependence on the past leads to uniqueness of the compatible measure. Different uniqueness conditions and the respective mixing properties have been studied [14, 26, 8, 4, 18, 23, 22].

In the present paper, we investigate the relationship between the characteristics of the probability kernel and the existence or non-existence of a Gaussian concentration bound (GCB) for the associated SCUMs. We prove that when the kernel has sum of oscillations less than one, or has summable variation, the respective SCUM satisfy a GCB. Moreover, we show that both conditions are tight by exhibiting processes that do not satisfy GCB whenever the oscillation is strictly larger than one, or the variation belongs to $\ell^{1+\epsilon}(\mathbb{N})$ for any $\epsilon > 0$.

We prove that a reason for the failure of GCB comes from the non-uniqueness of measures that are compatible with the same kernel. Our proof has an interest in its own by providing a method to prove that a GCB cannot be satisfied.

Our bounds are explicit and involve constants that are straightforwardly calculated from the kernels. We apply our inequalities in some important examples. Whenever possible, we make comparisons with other papers obtaining GCB for non-independent processes in the literature [36, 42, 34, 33]. Finally, as a simple application of our results, we use the relationship between GCB and transportation cost inequalities to obtain new bounds for $\bar{d}$-distance between SCUMs. As a corollary, we obtain the speed of Markovian approximation in $\bar{d}$-distance for SCUMs, in cases not covered in [8, 5, 24]. We also prove a Dvoretzky-Kiefer-Wolfowitz type inequality for SCUMs with summable variation.

Notation and necessary definitions are given in Section 2. Section 3 contains our main results and in Section 4 we present some consequences of these results. We end the paper with the proofs of our results in Section 5.

2 Definitions and notation

Let A be a countable set ("alphabet") endowed with the discrete topology. The alphabet A can be finite or infinite, if not stated explicitly. We denote by T the shift on $A^{\mathbb{Z}}$, that is, $(T\omega)_i = \omega_{i+1}$, $i \in \mathbb{Z}$. We equip this space with the σ- algebra generated by the cylinder sets $[a_{-n+1}, \dots, a_{n-1}] = \{\omega \in A^{\mathbb{Z}} : \omega_i = a_i, |i| \leq n-1\}$, $a_i \in A$, $n \in \mathbb{N}$. It comprises all Borel sets of $A^{\mathbb{Z}}$.

For $i, j \in \mathbb{Z}$ such that $i < j$ we write $[\![i, j]\!] = [i, j] \cap \mathbb{Z}$. Define $\mathcal{X}^- = A^{[\![-\infty,-1]\!]}$. For $i < j$, we indicate the vector $(\omega_i, \dots, \omega_j)$ by writing ω_i^j. We also use the convention that if $i > j$, ω_i^j is the empty string. For $x \in \mathcal{X}^-$, $n \geq 0$, and $\sigma \in A^{[\![0,n]\!]}$, $z = x\sigma$ is a concatenation of the respective symbols, such that $(\dots, z_{-1}, z_0, \dots, z_n) = (\dots, x_{-1}, \sigma_0, \dots, \sigma_n)$. For all $S \subset \mathbb{Z}$ and $\sigma \in A^S$ we define the projection function associated to all indices $i, j \in S$, $i \leq j$, by $\pi_i^j(\sigma) = \sigma_i^j$. Throughout the paper x, y, z will denote left-infinite sequences and ω and η will denote right- (or bi-) infinite sequences.

2.1 Kernels and SCUMs

To define the measures of interest of this paper, the stochastic chains of unbounded memory, we first need to define what is a probability kernel.

Definition 2.1 (Probability kernel)**.** *For all $n \in \mathbb{Z}$, a probability kernel g_n is a function $g_n : A \times A^{[\![-\infty,n-1]\!]} \to [0,1]$ such that for all $x \in A^{[\![-\infty,n-1]\!]}$, $\sum_{s\in A} g_n(s|x) = 1$. Because we will only consider shift-invariant kernels, with some abuse of notation, we will always refer to function g instead of g_n regardless of the index set inside the function.*

Let us first introduce SCUMs started from a fixed past, which will play an important role in the present work.

Definition 2.2 (Probability measure started with a fixed past)**.** *For $x \in \mathcal{X}^-$ and $\sigma \in A^{[\![0,\dots,k]\!]}$, we define $P^{x\sigma}$ as the probability measure specified by g when started with $x\sigma \in A^{[\![-\infty,k]\!]}$, that is, for all $\omega \in A^{\mathbb{N}}$ and all $n \geq k+1$*

$$P^{x\sigma}([\omega_{k+1}^n]) = \prod_{j=k+1}^{n} g\big(\omega_j | x\sigma\omega_{k+1}^{j-1}\big).$$

Sometimes we will write $P_g^{x\sigma}$ when it is not clear from the context to which kernel the measure corresponds.

Now, we introduce the definition of SCUMs compatible with a kernel, which is similar to the definition of a Gibbs measure compatible with a specification. We denote by $\mathcal{F}_i^j$ the σ-algebra generated by the cylinders with base in the interval $[\![i,j]\!]$. We use the shorthand notation $\mathcal{F}_k = \mathcal{F}_0^k$, $k \geq 0$.

Definition 2.3 (Probability measure compatible with a kernel)**.** *We say that a probability measure μ on $A^{\mathbb{Z}}$ is compatible with g if, for all $n \in \mathbb{Z}$, $a \in A$ and μ-a.e. $x \in A^{[\![-\infty,n-1]\!]}$, we have*

$$\mu([a]|\mathcal{F}_{-\infty}^{n-1})(x) = g(a|x).$$

A stationary stochastic process $(X_n)_{n\in\mathbb{Z}}$, where the random variables take values in A, is charaterized by a shift-invariant probability measure μ on $A^{\mathbb{Z}}$, that is, a measure satisfying $\mu \circ T^{-1} = \mu$. The canonical process $(X_n)_{n\in\mathbb{Z}}$ corresponding to a measure μ compatible with a kernel is called a stochastic chains of unbounded memory (SCUM) compatible with g. Equivalently, we say that a SCUM $(X_n)_{n\in\mathbb{Z}}$ is compatible with g if it satisfies

$$\mathbb{E}_\mu[\mathbb{1}_{\{a\}}(X_n)|X_\infty^{n-1} = x] = g(a|x)$$

for all n, $a \in A$, and μ-a.e $x \in A^{[\![-\infty,n-1]\!]}$, where $\mathbb{1}_{\{a\}}(\cdot)$ is the indicator function of the symbol a.

2.2 Regularity assumptions on kernels

We will use two main ways to quantify how the kernel g depends on the past. The first quantity is the oscillation of g of order $j \geq 1$ defined by

$$\mathrm{Osc}_j(g) := \sup\left\{\frac{1}{2}\sum_{a\in A}|g(a|z) - g(a|z')| : z, z' \in \mathcal{X}^-, z_k = z'_k, \forall k \neq -j\right\}.$$

The second quantity is the variation of g of order $j \geq 1$ defined by

$$\mathrm{Var}_j(g) := \sup\left\{\frac{1}{2}\sum_{a\in A}|g(a|z)-g(a|z')| : z,z' \in \mathcal{X}^-, z_k = z'_k, \forall k \geq -j\right\}.$$

We also define $\mathrm{Var}_0(g) := \sup_{z,z'\in\mathcal{X}^-}\frac{1}{2}\sum_{a\in A}|g(a|z)-g(a|z')|$. Note that the usual definition of oscillation ([27, 18, for instance]) and variation ([26, 30, for instance]) are, respectively,

$$\mathrm{osc}_j(g) := \sup\left\{|g(a|z)-g(a|z')| : a\in A, z,z'\in\mathcal{X}^-, z_k = z'_k, \forall k \neq -j\right\}$$

and

$$\mathrm{var}_j(g) := \sup\left\{|g(a|z)-g(a|z')| : a\in A, z,z'\in\mathcal{X}^-, z_k = z'_k, \forall k \geq -j\right\}.$$

When the alphabet is finite the definitions are equivalent since we have $\mathrm{var}_j(g) \leq \mathrm{Var}_j(g) \leq |A|\,\mathrm{var}_j(g)$ and $\mathrm{osc}_j(g) \leq \mathrm{Osc}_j(g) \leq |A|\,\mathrm{osc}_j(g)$. We use $\mathrm{Osc}_j(g)$ and $\mathrm{Var}_j(g)$ as they are more convenient to state our results when $|A| = \infty$. Given a kernel g, the following quantities will play a central role:

$$\Delta(g) := 1 - \sum_{j=1}^{\infty}\mathrm{Osc}_j(g) \tag{2.1}$$

and

$$\Gamma(g) := \prod_{j=0}^{\infty}(1-\mathrm{Var}_j(g)). \tag{2.2}$$

Remark 2.1 (Relation with existence/uniqueness criteria of the literature)**.** *A natural question to ask is whether there is a unique shift-invariant measure compatible with a given kernel g. If $\Delta(g) > 0$, Theorem 4.6 in [18] states that there is at most one compatible measure, which is therefore shift-invariant. In the case of finite alphabet, the assumption $\Gamma(g) > 0$ implies uniqueness of a shift-invariant compatible measure (see [26, 30] for instance). In the case of countably infinite alphabets, the conditions for uniqueness of the compatible measure are not based on $\mathrm{var}_j(g)$ anymore, and it is not obvious how to compare the assumption $\Gamma(g) > 0$ with other assumptions of the literature. For our purpose, we only discuss the conditions of uniqueness and existence when needed in the proofs.*

2.3 Gaussian concentration bound

Let $n \geq 0$ and $f : A^{n+1} \to \mathbb{R}$. Define

$$\delta_j f = \sup\left\{\left|f(\omega_0^{j-1} a\, \omega_{j+1}^n) - f(\omega_0^{j-1} b\, \omega_{j+1}^n)\right| : a,b\in A, \omega\in A^{n+1}\right\}$$

for $0 \leq j \leq n$. We denote by $\underline{\delta} f$ the column vector of size $n+1$ whose j-th coordinate is $\delta_{j-1} f$. Also let

$$\|\underline{\delta} f\|_2^2 = \sum_{j=0}^{\infty}(\delta_j f)^2 = \sum_{j=0}^{n}(\delta_j f)^2. \tag{2.3}$$

More generally, for $p \in \mathbb{N}$ and $\underline{v} = (v_0, v_1, \ldots)$ with $v_i \in \mathbb{R}$, define

$$\|\underline{v}\|_p^p = \sum_{j=0}^{\infty}|v_j|^p.$$

We say that $\underline{v} \in \ell^p(\mathbb{N})$ if $\|\underline{v}\|_p < +\infty$.
For $f : A^{\mathbb{Z}} \to \mathbb{R}$ μ-integrable, we use the notation $\mathbb{E}_\mu[f] = \int f\,\mathrm{d}\mu$.

Definition 2.4 (Gaussian concentration bound)**.** *A probability measure μ on $A^{\mathbb{Z}}$ or on $A^{\mathbb{N}}$ is said to satisfy a Gaussian concentration bound (GCB for short) if there exists a constant $C > 0$ such that, for all integers $n \geq 0$ and for all $f : A^{n+1} \to \mathbb{R}$, we have*

$$\mathbb{E}_\mu \left[\mathrm{e}^{f - \mathbb{E}_\mu[f]} \right] \leq \mathrm{e}^{C \|\underline{\delta} f\|_2^2} \tag{2.4}$$

where $\|\underline{\delta} f\|^2$ is defined in (2.3).

A key-point in this definition is that C does neither depend on n nor on f. Since f is bounded, this inequality implies that, for all $\theta \in \mathbb{R}$, we have

$$\mathbb{E}_\mu \left[\mathrm{e}^{\theta(f - \mathbb{E}_\mu[f])} \right] \leq \mathrm{e}^{C\theta^2 \|\underline{\delta} f\|_2^2}$$

and using a standard argument (usually referred to as Chernoff bounding method, see [39]), we deduce that, for all $u > 0$,

$$\mu(|f - \mathbb{E}_\mu[f]| > u) \leq 2 \exp\left(- \frac{u^2}{4C \|\underline{\delta} f\|_2^2} \right). \tag{2.5}$$

The formulation of Definition 2.4 is made in such a way that we can take a probability measure with a fixed past as defined above. Also, if we have a shift-invariant probability measure μ, then it is indifferent to work either with $A^{\mathbb{Z}}$ or $A^{\mathbb{N}}$.

3 Main results and examples

3.1 GCB under a condition on the oscillation of the kernel

Our first result is a GCB for a probability measure started with a fixed past in the sense of Definition 2.2). The bound is uniform in the past $x \in \mathcal{X}^-$.

Theorem 3.1. *Let g be a kernel such that $\Delta(g) > 0$. Then, for all $n \geq 0$, for all $f : A^{n+1} \to \mathbb{R}$, and for all $\theta \in \mathbb{R}$, we have*

$$\sup_{x \in \mathcal{X}^-} \mathbb{E}_{P^x} \left[\mathrm{e}^{\theta(f - \mathbb{E}_{P^x}[f])} \right] \leq \mathrm{e}^{\frac{\theta^2 \Delta(g)^{-2}}{8} \|\underline{\delta} f\|_2^2} \tag{3.1}$$

where $\Delta(g)$ is defined in (2.1). As a consequence, for all $u > 0$, we have

$$\sup_{x \in \mathcal{X}^-} P^x(|f - \mathbb{E}_{P^x}[f]| > u) \leq 2 \exp\left(- \frac{2u^2}{\Delta(g)^{-2} \|\underline{\delta} f\|_2^2} \right). \tag{3.2}$$

Let us illustrate this theorem with two examples.

Example 3.1 (Binary autogregressive process)**.** *Consider a function $\psi : \mathbb{R} \to (0,1)$ such that $\psi(r) + \psi(-r) = 1$ and an absolutely summable sequence of real numbers $(\xi_j)_{j \geq 0}$. Then the kernel $g : \{-1,+1\} \times \{-1,+1\}^{[\![-\infty,-1]\!]} \to (0,1)$ is defined as*

$$g(a|x) = \psi\left(a \sum_{j=1}^{\infty} \xi_j x_{-j} + a\xi_0 \right).$$

The process generated by this kernel is called a binary auto-regressive process [31]. If ψ is differentiable, we have that $\mathrm{Osc}_j(g) \le 2(\sup \psi')|\xi_j|$, hence we have $\Delta(g) \ge 1 - 2(\sup \psi') \sum_{j=1}^{\infty} |\xi_j|$. For instance, if $\psi(u) = (1 + \mathrm{e}^{-2u})^{-1}$ then $\Delta(g) \ge 1 - \sum_{j=1}^{\infty} |\xi_j|$.

Example 3.2 (Poisson regression for count time series)**.** *Let $A = \mathbb{N}$ and $(\xi_j)_{j \ge 0}$ be a sequence of non-positive absolutely summable real numbers, and a constant $c > 0$. For all $x \in \mathbb{N}^{[\![-\infty,-1]\!]}$, let*

$$v(x) = \exp\left(\sum_{j=1}^{\infty} \xi_j \min\{x_{-j}, c\} \right).$$

For all $a \in \mathbb{N}$ and $x \in \mathcal{X}^-$, the kernel of a Poisson regression model is defined as [31]

$$g(a|x) = \frac{\mathrm{e}^{-v(x)}\, v(x)^a}{a!}.$$

Applying the mean value theorem to $\psi(r) = \mathrm{e}^{-\mathrm{e}^r}\, \mathrm{e}^{ra}/a!$, and maximizing on $r \in (-\infty, 0]$ for each $a \in \mathbb{N}$, we obtain $\mathrm{Osc}_j(g) \le \mathrm{e}^{-1} \sum_{a \ge 0} \frac{1}{a!} |\xi_j| = |\xi_j|$. Therefore, $\Delta(g) \ge 1 - \sum_{j=1}^{\infty} |\xi_j|$.

We also have a theorem for stationary SCUMs.

Theorem 3.2. *If μ is a shift-invariant measure compatible with a kernel g on a finite alphabet such that $\Delta(g) > 0$, then inequalities (3.1) and (3.2) hold with μ in place of P^x, with the same constant.*

Remark 3.1. *In the preceding theorem, we assume that $|A| < \infty$ because, when $|A| = \infty$, the space $A^{\mathbb{Z}}$ is no longer compact, so we loose compactness of the space of probability measures on $A^{\mathbb{Z}}$ on which we rely in our proof.*

3.2 GCB under a condition on the variation of the kernel

We have the analog of Theorem 3.1 under a natural condition on the variation.

Theorem 3.3. *Let g be a kernel such that $\Gamma(g) > 0$. Then, for all $n \ge 0$, for all $f : A^{n+1} \to \mathbb{R}$, and for all $\theta \in \mathbb{R}$, we have*

$$\sup_{x \in \mathcal{X}^-} \mathbb{E}_{P^x}\left[\mathrm{e}^{\theta(f - \mathbb{E}_{P^x}[f])} \right] \le \mathrm{e}^{\frac{\theta^2 \Gamma(g)^{-2}}{8} \|\underline{\delta} f\|_2^2}$$

where $\Gamma(g)$ is defined in (2.2). As a consequence, for all $u > 0$, we have

$$\sup_{x \in \mathcal{X}^-} P^x(|f - \mathbb{E}_{P^x}[f]| > u) \le 2 \exp\left(-\frac{2u^2}{\Gamma(g)^{-2} \|\underline{\delta} f\|_2^2} \right).$$

We give a class of examples illustrating this theorem.

Example 3.3 (Convex mixture of Markov chains)**.** *Let $(\lambda_j)_{j \ge 1}$ be a sequence of non-negative real numbers such that $\sum_{j=1}^{\infty} \lambda_j = 1$. Let A be a countable set. Define a family of Markov kernels $p^{[k]} : A \times A^{[\![-k,-1]\!]} \to [0,1], k \ge 0$, that is, for all $x \in \mathcal{X}^-$, $\sum_{a \in A} p^{[k]}(a|x_{-k}^{-1}) = 1$. The kernel for mixture of Markov chains is defined, for all $a \in A$ and $x \in \mathcal{X}^-$, as*

$$g(a|x) = \sum_{j=1}^{\infty} \lambda_j\, p^{[j]}(a|x_{-j}^{-1}).$$

We have $\sum_{j=1}^{\infty} \mathrm{Var}_j(g) \leq \sum_{j=1}^{\infty} j\lambda_j$. *This result is quite general since a large class of kernels, including all kernels* g *on finite alphabet with* $\lim_j \mathrm{Var}_j(g) = 0$, *can be represented as a convex mixture of Markov chains [29].*

The next result complements Theorem 3.3 in the case of stationary SCUMs.

Theorem 3.4. *If* μ *is a shift-invariant measure compatible with a kernel* g *such that* $\Gamma(g) > 0$, *then inequality* (2.4) *hold with* μ *in place of* P^x, *with the same constant.*

3.3 Optimality of the bounds

Here we show that Theorems 3.2 and 3.4 are optimal already for binary alphabets. Theorems 3.5 and 3.6 below give necessary conditions to get GCB for a large class of processes that exhibit phase transition. Our optimality results are simple consequences of these theorems.

The following result shows that, for kernels satisfying strong regularity conditions, phase transition is a fundamental barrier for GCB.

Theorem 3.5. *Let* g *be a kernel such that* $\inf_{a\in A, x\in\mathcal{X}^-} g(a|x) > 0$ *and* $\lim_j \mathrm{Var}_j(g) = 0$. *If* g *has two (or more) distinct ergodic compatible measures, then they do not satisfy GCB.*

[27] proved that, for all $\epsilon > 0$, there are examples of g such that $\Delta(g) + \epsilon < 0$ and exhibit multiple shift-invariant ergodic compatible measures. Because of Thereom 3.5, this implies that the shift- invariant ergodic compatible measures do not satisfy GCB. This shows optimality of Theorem 3.2, a fact that we now state as a corollary of Theorem 3.5.

Corollary 3.1. *For any* $\epsilon > 0$, *there is a kernel* g *on a binary alphabet and a compatible shift-invariant probability measure* μ *such that* $\sum_{j=1}^{\infty} \mathrm{Osc}_j(g) \in (1, 1+\epsilon]$ *and* μ *does not satisfy GCB. Moreover,* g *can be chosen to satisfy* $\lim_j \mathrm{Var}_j(g) = 0$ *and* $\inf_{a\in A, x\in\mathcal{X}^-} g(a|x) > 0$.

To demonstrate the optimality of Theorem 3.4, we settle the question of GCB for renewal measures, a particular class of SCUMs. Let $(q_j)_{j\geq 0}$ with $q_j \in (0,1)$. Given $x \in \mathcal{X}^-$, let $\ell(x) = \inf\{k \geq 0 : x_{-k-1} = 1\}$ and $\ell(\ldots 00) = \infty$. We define the renewal kernel $\tilde{g} : \{0,1\} \times \{0,1\}^{[\![-\infty,-1]\!]} \to (0,1)$ by taking $\tilde{g}(1|x) = q_{\ell(x)}$. Obviously, if $q_\infty = 0$ then the degenerate measure δ_{0^∞} is stationary and compatible, and trivially satisfies GCB. However, we call renewal measure the stationary measure μ compatible with g satisfying $\mu([a]) > 0$ for any $a \in \{0,1\}$, when it exists. It is not difficult to see that this measure will actually consists of a sequence of i.i.d. concatenation of blocks of the form $0^i 1, i \geq 1$. The probability that the distance between two consecutive 1's equals $n \geq 1$, denoted f_n, is

$$f_n := P^{x1}_{\tilde{g}}(0^{n-1}1) = q_{n-1}\prod_{i=0}^{n-2}(1-q_i), \quad \forall n \geq 1, \forall x \in \mathcal{X}^- \tag{3.3}$$

with the convention $\prod_{i=0}^{-1} = 1$. The probability distribution $(f_n)_{n\geq 1}$ is usually called inter-arrival distribution in the literature. Then, the renewal measure exists if and only if the expected distance between consecutive ones, $\sum_{n\geq 1} n f_n$, is finite, which is equivalent to

$$\sum_{j\geq 1}\prod_{i=0}^{j-1}(1-q_i) < \infty. \tag{3.4}$$

We have the following result.

Theorem 3.6. *The renewal measure satisfies a GCB if, and only if,* $\sum_n f_n r^n < \infty$ *for some* $r > 1$.

Consider now the particular case in which $q_j = j^{-\alpha}$ for $j \geq 2$ with $\alpha \in (0,1)$ so that (3.4) is satisfied and therefore the renewal process exists. A simple calculation shows that in this case f_n is stretched exponential, and therefore, by Theorem 3.6, the renewal process does not satisfy GCB. Moreover, $\mathrm{Var}_j(g) = q_j$, therefore $\lim_j \mathrm{Var}_j(g) = 0$ and $\inf_{x \in \mathcal{X}^-} g(a|x) > 0$ for some $a \in A$. Hence, if we choose $\alpha = (1+\epsilon/2)(1+\epsilon)^{-1}$ the variation will not be summable, but $\sum_{j=1}^{\infty} \mathrm{Var}_j(g)^{1+\epsilon} < \infty$, proving that Theorem 3.4 is optimal, a fact that we state as a corollary of Theorem 3.6.

Corollary 3.2. *For any* $\epsilon > 0$*, there is kernel* g *on a binary alphabet and a compatible shift-invariant probability measure* μ *such that*

$$\sum_{j=1}^{\infty} \mathrm{Var}_j(g)^{1+\epsilon} < \infty, \ \sum_{j=1}^{\infty} \mathrm{Var}_j(g) = \infty$$

and μ *does not satisfy GCB. Moreover,* g *can be chosen to satisfy* $\inf_{x \in \mathcal{X}^-} g(a|x) > 0$ *for some* $a \in A$.

Remark 3.2. *Note that the two kernels used to obtain examples of processes which do not satisfy GCB exhibit phase transition since for the renewal process, when* $q_j = j^{-\alpha}$ *we have that* $q_\infty = 0$ *(to get* $\mathrm{Var}_k(g) \to 0$*), in which case the Dirac measure* δ_{0^∞} *is also compatible. However, this kernel does not fall into the class considered by Theorem 3.5 because it does not satisfy* $\inf_{a \in A, x \in \mathcal{X}^-} g(a|x) > 0$.

Remark 3.3. *Kernels satisfying* $\inf_{a \in A, x \in \mathcal{X}^-} g(a|x) > 0$ *are said to be "strongly non-null". If we restrict to strongly non-null kernels* g *instead of assuming the slightly weaker condition*

$$\inf_{x \in \mathcal{X}^-} g(a|x) > 0$$

for some $a \in A$*, then we don't know if the summable variation condition is tight for the validity of GCB. Nevertheless, even if we restrict to strongly non-null kernels* g*, GCB does not hold in general beyond square summable variation because of Theorem 3.5 and the existence of examples with phase transition for strongly non-null kernels such that* $\sum_{j=1}^{\infty} \mathrm{Var}_j(g)^{2+\epsilon} < \infty, \epsilon > 0$*, see [1].*

3.4 GCB for a more general class of functions

Denote by $\mathrm{C}(A^{\mathbb{N}})$ the set of real-valued continuous functions on $A^{\mathbb{N}}$ that we equip with the supremum norm. We define two of its subspaces, namely the set of bounded continuous functions, denoted $\mathrm{BC}(A^{\mathbb{N}})$, and the set of uniformly continuous functions, denoted $\mathrm{UC}(A^{\mathbb{N}})$. As $A^{\mathbb{N}}$ is in general not compact, $\mathrm{UC}(A^{\mathbb{N}})$ intersects but does not contain $\mathrm{BC}(A^{\mathbb{N}})$, nor does $\mathrm{BC}(A^{\mathbb{N}}) \supseteq \mathrm{UC}(A^{\mathbb{N}})$. Obviously, the set of all functions $f : A^{n+1} \to \mathbb{R}$, for $n = 0, 1, \ldots$ is contained in each of these three spaces. We have $\mathrm{C}(A^{\mathbb{N}}) = \mathrm{UC}(A^{\mathbb{N}}) = \mathrm{BC}(A^{\mathbb{N}})$ if and only if $A^{\mathbb{N}}$ is compact, which holds if and only if A is finite. For $f \in \mathrm{C}(A^{\mathbb{N}})$, let

$$\mathrm{var}_n(f) = \sup\{|f(\omega) - f(\omega')| : \omega_i = \omega'_i, i = 0, \ldots, n\}, n \geq 0.$$

One can easily check that $\mathrm{var}_n(f) \to 0$ if and only if $f \in \mathrm{UC}(A^{\mathbb{N}})$.

We can generalize Theorems 3.1, 3.2, 3.3 and 3.4, thanks to the following abstract result.

Theorem 3.7. *If a probability measure* μ *satisfies a Gaussian concentration bound for some constant* $C > 0$ *(Definition 2.4), then this bound remains true for all* $f \in \mathrm{UC}(A^{\mathbb{N}}) \cap \mathrm{BC}(A^{\mathbb{N}})$ *such that* $\|\underline{\delta} f\|_2 < +\infty$*, with the same constant* C.

We refer the reader to Section 4.3 for a natural application of Theorem 3.7.

3.5 Some comparisons with existing results

In the independent case, $\Gamma(g) = \Delta(g) = 1$, so we recover McDiarmid's inequality (with the optimal constant) [39]. In the Markov case $\Gamma(g) = 1 - \mathrm{Var}_0(g) = 1 - \mathrm{Osc}_1(g) = \Delta(g)$ and it is the Dobrushin ergodicity coefficient [13]. In general, $\Gamma(g) \neq \Delta(g)$ and the conditions are complementary to each other. Indeed, consider the kernel in Example 3.1 with $\psi(u) = 1/(1 + \exp(-2u))$, $\xi_i \geq \xi_j \geq 0$ for all $j > i \geq 1$, and $\xi_0 = \sum_{k=1}^{\infty} \xi_k$. We have

$$\frac{2\,\mathrm{e}^{2\xi_j}}{(1+\mathrm{e}^{2\xi_j})^2}\xi_j \leq \mathrm{Osc}_j(g) \leq \xi_j \quad \text{and} \quad \frac{2\,\mathrm{e}^{2\xi_1}}{(1+\mathrm{e}^{2\xi_1})^2}\sum_{k>j}\xi_k \leq \mathrm{Var}_j(g) \leq \sum_{k>j}\xi_k.$$

If $\xi_j = c/j^{1+\epsilon}$ with $\epsilon \in (0, 1]$ and small enough $c > 0$, we have $\Delta(g) > 0$ but $\Gamma(g) = 0$. On the other hand, if $\xi_j = C/j^{1+\epsilon}$ with $\epsilon \in (1, \infty)$ and large enough C we have $\Gamma(g) > 0$, but $\Delta(g) < 0$.

There are SCUMs that satisfy GCB, but are not covered by our results. Here is an example using the renewal kernel defined in Section 3.3. For this example, let $\alpha \in (0, 1)$ and consider the sequence $q_j = q_\infty + \alpha/j^\alpha$. For the renewal kernel we have $\mathrm{Osc}_j(\tilde{g}) = \mathrm{Var}_j(\tilde{g}) = q_j - q_\infty = \alpha/j^\alpha$. If $q_\infty > 0$ we easily find that the inter-arrival distribution defined in (3.3) is exponential, and therefore, by Theorem 3.6 this renewal process satisfy GCB. However, neither the oscillation nor the variation are summable, and therefore we have $\Delta(\tilde{g}) < 0$ and $\Gamma(\tilde{g}) = 0$.

Marton [36] proved a measure concentration property that is equivalent to a version of Theorem 3.4 in which $\|\underline{\delta} f\|_2^2$ is substituted by $n\|\underline{\delta} f\|_\infty^2$. Because $\|\underline{\delta} f\|_2^2 \leq (n+1)\|\underline{\delta} f\|_\infty^2$, our result gives an improvement. For example, consider $A = \{0, 1\}$, $\epsilon > 0$, and $f(x_0^n) = \sum_{j=0}^n x_j/(j+1)^{(1+\epsilon)/2}$. In this case, for all $n \geq 0$, $\|\underline{\delta} f\|_2^2 < C$ for some constant C, but $n\|\underline{\delta} f\|_\infty^2 = n + 1$. Perhaps more importantly, we offer a different proof. Marton's proof is based on a transportation cost inequality together with its tensorization, whereas our proof is based on martingale methods together with coupling inequalities. Because of the stationarity requirement in [36], we don't know if we can obtain Theorem 3.3 using the same method as in [36]. We also note that [25] obtained a version of Theorem 3.4 for finite alphabets, by a different approach than the one we use here, and with a suboptimal constant 2/9 instead of 2 as we obtained here.

To conclude, let us mention that [34] proved a GCB for Gibbs random fields satisfying the two-sided Dobrushin condition. On $\mathbb{Z}$, if a Gibbs specification satisfies the two-sided Dobrushin condition then it also satisfies $\Delta(g) > 0$ [17, Theorem 4.20]. Therefore, Theorem 3.2 implies GCB the result in [34] for one-dimensional Gibbs measures. We do not know whether the converse also holds.

3.6 Some open problems

There are only few results in the literature giving necessary conditions for the existence of GCB for dependent process. We think that answers to the following questions could help in the development of new tools to prove necessary conditions for GCB.

- Is there a SCUM on finite alphabet with a unique compatible stationary measure but which does not satisfy GCB?
- Is the summable variation condition tight for the GCB to hold among strongly non-null kernels?
- Do we have GCB when $\Delta(g) = 0$ and $\mathrm{Var}_j(g) \sim 1/j$?

- What is the rate of concentration for kernels with more than one compatible measure?

4 Applications

In this section we explore some consequences of the results presented above. We show (1) a new bound on the probability of deviation of the empirical distribution from the stationary distribution, (2) new bounds for the distance between two processes under the $\bar{d}$-distance, (3) an exponential bound for the rate convergence of Birkhoff sums for SCUMs. For further applications of GCB, the reader can check [7].

4.1 Dvoretzky-Kiefer-Wolfowitz type inequality

In statistics we are often interested in the empirical distribution: For $\sigma \in A^{[\![1,k]\!]}$ and $\omega \in A^{\mathbb{N}}$, let

$$\hat{\rho}_{n,k}(\sigma,\omega) = \frac{1}{n-k+2} \sum_{j=0}^{n-k+1} \mathbb{1}_{\{\sigma\}}(\pi_j^{j+k-1}\omega).$$

We will simply write $\hat{\rho}_{n,k}(\sigma)$ for the corresponding random variable. To estimate the probability of deviation from the expected value, it is natural to use Theorem 3.3 to obtain

$$\mu\left(|\hat{\rho}_{n,k}(\sigma) - \rho(\sigma)| > u\right) \leq 2\exp\left(-(n-k+2)\Gamma(g)^2u^2\right),$$

where $\hat{\rho}_{n,k}(\sigma) = \hat{\rho}_{n,k}(\sigma,\cdot)$ and $\rho(\sigma) := \mathbb{E}[\hat{\rho}_{n,k}(\sigma)] = \mu([\sigma])$. If we want to obtain a uniform bound, we should upper bound

$$\mu\left(\|\hat{\rho}_{n,k} - \rho\|_\infty > u\right) = \mu\left(\sup_\sigma |\hat{\rho}_{n,k}(\sigma) - \rho(\sigma)| > u\right).$$

In this case, it is tempting to use a union bound. However, when the cardinality of the set of symbols is large, we get a bad bound, and when $A = \mathbb{N}$ this approach fails. One possible solution is to concentrate directly the uniform deviation $\|\hat{\rho}_{n,k} - \rho\|_\infty$, which yields the following result.

Theorem 4.1. *Let g be a kernel and μ be a shift-invariant measure compatible with g. If $\Gamma(g) > 0$, we have, for all $u > 0$ and for all $n > 0$ and $0 < k \leq n$,*

$$\mu\left(\|\hat{\rho}_{n,k} - \rho\|_\infty > \frac{u + \sqrt{2k}}{\sqrt{(n-k+2)\Gamma(g)}}\right) \leq \exp\left(-\Gamma(g)\,u^2\right). \tag{4.1}$$

A similar result for $k = 1$ was obtained in [32] for Markov chains and hidden Markov models, but as far as we know, our result is the first of the literature for SCUMs. Because Theorem 4.1 gives uniform control on the empirical distributions, we can use these results to estimate quantities that can be written as functionals of empirical distribution, *e.g.*, entropy, kernels, and potentials of the processes.

4.2 Explicit upper bound for the $\bar{d}$-distance, and speed of Markovian approximation

Given two probability measures μ and ν on $A^{\mathbb{Z}}$, a coupling of μ and ν is a measure $\mathbb{P}$ on $A^{\mathbb{Z}} \times A^{\mathbb{Z}}$ satisfying, for all $B \in \mathcal{F}_{-\infty}^{+\infty}$

$$\mathbb{P}\big(B \times A^{\mathbb{Z}}\big) = \mu(B) \quad \text{and} \quad \mathbb{P}\big(B^{\mathbb{Z}} \times A\big) = \nu(B).$$

Let $\mathcal{J}_{\mu,\nu}$ denote the set of couplings of μ and ν. The $\bar{d}$-distance between μ and ν is then defined as

$$\bar{d}(\mu,\nu) = \inf_{\mathbb{P}\in\mathcal{J}_{\mu,\nu}} \mathbb{P}\left(\{(\eta,\omega) \in A^{\mathbb{Z}} \times A^{\mathbb{Z}} : \eta_0 \neq \omega_0\}\right).$$

It is natural to ask if given two "close" (in a sense to be made precise below) probability kernels g and h, whether we can upper bound the $\bar{d}$-distance between the respective compatible measures μ and ν. The following result gives such a bound.

Theorem 4.2. *Let μ be a shift-invariant measure compatible with a kernel g such that*

$$\inf_{a\in A, x\in X^-} g(a|x) > 0$$

and satisfying either the conditions of Theorem 3.2 or of Theorem 3.4. Let also ν be a shift-invariant measure compatible with a kernel h with $\lim_j \mathrm{Var}_j(h) = 0$ and $\inf_{a\in A, x\in\mathcal{X}^-} h(a|x) > 0$. The $\bar{d}$-distance between μ and ν is bounded by

$$\bar{d}(\mu,\nu) \leq \frac{1}{\mathcal{C}}\sqrt{\frac{1}{2}\mathbb{E}_\nu\left[\log\frac{h}{g}\right]} \tag{4.2}$$

where $\mathcal{C}$ equals $\Delta(g)$ or $\Gamma(g)$ depending on which condition g satisfies.

Remark 4.1. *The conditions*

$$\inf_{a\in A, x\in X^-} g(a|x) > 0, \quad \inf_{a\in A, x\in X^-} h(a|x) > 0$$

imply that the alphabet is finite. Although this condition can be weakened, it is the simplest way to guarantee that $|\log\frac{h}{g}| < \infty$, so that the upper bound on the right hand side of (4.2) remains meaningful.

In Theorem 4.2 we measure the closeness of h and g by $\mathbb{E}_\nu\left[\log\frac{h}{g}\right]$. [8] proved, without obtaining an explicit upper bound, that if the variation of kernel g satisfies $\sum_{n\geq r}\prod_{j=r}^{n}(1-(|A|/2)\,\mathrm{var}_j(g)) = \infty$, for some $r \geq 1$, then a small $\|g-h\|_\infty$ implies a small $\bar{d}(\mu,\nu)$. As a consequence of Theorem

4.2 we have

$$\begin{aligned}
\mathbb{E}_\nu\left[\log\frac{h}{g}\right] &= \int \log\frac{h}{g}\,\mathrm{d}\nu = \int \log\left(1+\frac{h-g}{g}\right)\mathrm{d}\nu \le \int \frac{h-g}{g}\,\mathrm{d}\nu \\
&= \int \sum_{a\in A} h(a|x)\,\frac{h(a|x)-g(a|x)}{g(a|x)}\,\mathrm{d}\nu(x) \\
&= \int \sum_{a\in A} \frac{(h(a|x)-g(a|x))^2}{g(a|x)}\,\mathrm{d}\nu(x) \\
&\le \frac{1}{\inf g}\int \sum_{a\in A} (h(a|x)-g(a|x))^2\,\mathrm{d}\nu(x) \\
&\le \frac{1}{\inf g}\Big(\sup_{x\in\mathcal{X}^-}\sum_{a\in A}|h(a|x)-g(a|x)|\Big)^2
\end{aligned}$$

where $\inf g := \inf_{a\in A, x\in X^-} g(a|x)$. Therefore, Theorem 4.2 yields

$$\bar{d}(\mu,\nu) \le \frac{1}{\mathcal{C}}\sqrt{\frac{1}{2\inf g}}\left(\sup_{x\in\mathcal{X}^-}\sum_{a\in A}|h(a|x)-g(a|x)|\right)^2. \tag{4.3}$$

Theorem 4.2 can also be used to upper bound the $\bar{d}$-distance between a measure μ with kernel g and a k-step Markov approximation of μ, $\mu^{[k]}, k \ge 1$. Consider that the Markov approximation has kernel

$$g^{[k]}(a|x_{-k}^{-1}) := g(a|x_{-k}^{-1}y_{-\infty}^{-k-1})$$

for some fixed $y \in \mathcal{X}^-$. [5, 16] showed that if the kernel g satisfies $\Gamma(g) > 0$, then there exists a sequence $\mu^{[k]}, k \ge 1$ such that $\bar{d}(\mu,\mu^{[k]}) \le C\,\mathrm{var}_k(g)$ where C is some positive constant. Later [24] extended this result, obtaining, via "coupling from the past" arguments, upper bounds in the case where $\sum_{n\ge1}\prod_{k=1}^{n}(1-\mathrm{var}_k(g)) = \infty$, but their results are not explicit, depending on the tail distribution of the time for success in the coupling. Such bounds have proved to be a valuable tool to obtain further properties of the measure μ [9, 15, for instance]. Here, if either $\Gamma(g) > 0$ or $\Delta(g) > 0$, using (4.3) we obtain the following result.

Corollary 4.1. *For all $k \ge 1$ we have*

$$\bar{d}(\mu,\mu^{[k]}) \le \frac{1}{\mathcal{C}}\sqrt{\frac{1}{2\inf g}}\,\mathrm{Var}_k(g)$$

where $\mathcal{C}$ equals $\Delta(g)$ or $\Gamma(g)$ depending on which condition g satisfies.

Proof. Apply (4.3) with $\nu = \mu^{[k]}$. □

Remark 4.2. *A similar bound was obtained by [16] under the assumption that $\Gamma(g) > 0$, while Corollary 4.1 also holds if $\Delta(g) > 0$. We refer the reader to Subsection 3.5 where an example satisfying $\Delta(g) > 0$, but such that $\Gamma(g) = 0$, is provided, showing that our result is strictly more general than the one of [16].*

Example 4.1 (Bramson-Kalikow-Friedli model)**.** *Let $A = \{-1, +1\}$, $\varepsilon \in (0, 1/2)$, $(\lambda_j)_{j\geq 1}$ be a sequence of positive real numbers such that $\sum_{j=1}^{\infty} \lambda_j = 1$, and $(m_j)_{j\geq 1}$ be an increasing sequence of positive odd integers. Let also $\varphi : [-1, 1] \to [\varepsilon, 1 - \varepsilon]$ be a monotonically increasing function satisfying $\varphi(-s) + \varphi(s) = 1$ for $s \in [-1, 1]$. The Bramson-Kalikow-Friedli model is given by*

$$g(+1|x) = \sum_{j=1}^{\infty} \lambda_j \, \varphi\left(\frac{1}{m_j} \sum_{i=1}^{m_j} x_{-i}\right), \; x \in \mathcal{X}^-.$$

There always exists at least one compatible measure μ since the alphabet is finite and $\mathrm{Var}_j(g)$ vanishes in j. If $\varphi(s) = \varepsilon + (1 - 2\varepsilon)\, \mathbb{1}_{\mathbb{Z}_{<0}}(s)$, we get the original model introduced by Bramson and Kalikow [3]. They showed that the sequences $(\lambda_j)_{j\geq 1}$ and $(m_j)_{j\geq 1}$ can be chosen so that the corresponding kernel exhibits multiple compatible shift-invariant measures [3].

For all $k \geq 1$, the m_k-step Markov approximation $\mu^{[m_k]}$ is defined by the kernel

$$g^{[m_k]}(+1|x) = \sum_{j=1}^{k} \lambda_j \, \varphi\left(\frac{1}{m_j} \sum_{i=1}^{m_j} x_{-i}\right) + (1 - \varepsilon) \sum_{j>k} \lambda_j.$$

The sequence of measures $\mu^{[m_k]}, k \geq 1$ was used by [21] for their proof of phase transition of the Bramson-Kalikow model.

Suppose for now that there exists a shift-invariant measure μ compatible with g. We can easily derive a lower bound for $\bar{d}(\mu, \mu^{[m_k]})$. Indeed, from the definition of $\bar{d}$-distance, we have that

$$\bar{d}(\mu, \mu^{[m_k]}) \geq |\mu^{[m_k]}([1]) - \mu([1])|.$$

By symmetry and uniqueness of μ, we have that $\mu([1]) = 1/2$. A direct calculation then shows that

$$\mu^{[m_k]}([1]) \geq \varepsilon \sum_{j>k} \lambda_j + 1/2$$

whence

$$\bar{d}(\mu, \mu^{[m_k]}) \geq \varepsilon \sum_{j>k} \lambda_j \, .$$

If $\Gamma(g) > 0$ or $\Delta(g) > 0$, Corollary 4.1 allows us to show that this bound is actually of the right order in k since we have that for all $k \geq 1$,

$$\bar{d}(\mu, \mu^{[m_k]}) \leq \frac{1}{\Gamma(g)\sqrt{2\varepsilon}} \sum_{j>k} \lambda_j.$$

Thus, it only remains to give examples of kernels in which $\Gamma(g) > 0$ or $\Delta(g) > 0$. First, observe that, for any function φ, we have

$$\sum_{j\geq 1} \mathrm{Var}_j(g) \leq \sum_{j\geq 1} m_j \sum_{i\geq j} \lambda_i.$$

So if $\sum_{j=1}^{\infty} m_j \sum_{i\geq j} \lambda_i < \infty$, we have $\Gamma(g) > 0$ and $\mathrm{Var}_0(g) = 1 - 2\varepsilon > 0$, independently of the function φ. However, if $\sum_{j=1}^{\infty} m_j \sum_{i\geq j} \lambda_i = \infty$ we can still have examples in which $\Delta(g) > 0$ and

use Corollary 4.1. For example, take $\varphi(s) := \frac{1}{2} + \left(\frac{1}{2} - \epsilon\right) s$*, which was studied in [20]. In this case, a simple calculation shows that*

$$\sum_{j\geq 1} \mathrm{Osc}_j(g) \leq (1-2\varepsilon) \sum_{j\geq 1} \lambda_j < 1$$

and therefore $\Delta(\bar{g}) > 0$*, independently of the choice of the sequences* $(\lambda_j)_{j\geq 1}$ *and* $(m_j)_{j\geq 1}$*.*

4.3 Concentration of functions that depend on infinitely many coordinates

A natural application of Theorem 3.7 is the following. For $n \geq 1$, let $S_n\phi := \phi + \phi\circ T + \cdots + \phi \circ T^{n-1}$ where $\phi \in \mathrm{UC}(A^{\mathbb{N}}) \cap \mathrm{BC}(A^{\mathbb{N}})$ and satisfies $\|\underline{\delta}\phi\|_1 < +\infty$. Then we have, for all $n \geq 1$ and $u > 0$,

$$\mu\left(\left|\frac{S_n\phi}{n} - \int \phi \,\mathrm{d}\mu\right| > u\right) \leq 2\exp\left(-\frac{2nu^2}{\mathcal{C}\|\underline{\delta}\phi\|_1^2}\right) \tag{4.4}$$

where $\mathcal{C}$ equals $\Delta(g)$ or $\Gamma(g)$ depending on which condition g satisfies. The proof is as follows. Taking $f = S_n\phi$, we can check that

$$\|\underline{\delta} S_n\phi\|_2^2 \leq n\|\underline{\delta}\phi\|_1^2,\ n \geq 1. \tag{4.5}$$

Then we apply (2.5) to get (4.4). To prove (4.5), observe that $\delta_j S_n\phi \leq \sum_{i=0}^{n-1} \delta_{i+j}\phi$, and apply Young's inequality for (discrete) convolutions: if $\underline{v} \in \ell^p(\mathbb{N})$ and $\underline{w} \in \ell^q(\mathbb{N})$, for some $1 \leq p \leq q \leq +\infty$, then $\underline{v} * \underline{w} \in \ell^r(\mathbb{N})$ where $r \geq 1$ satisfies $1 + r^{-1} = p^{-1} + q^{-1}$, and $\|\underline{v} * \underline{w}\|_r \leq \|\underline{v}\|_p \|\underline{w}\|_q$. We use it with $r = 2, p = 2, q = 1$, $v_k = \mathbb{1}_{[\![0,n-1]\!]}(k)$ and $w_k = \delta_k h$. ([1])

5 Proofs of the results

5.1 Gaussian concentration bound using coupling

All the GCBs obtained in this work are consequences of an abstract GCB proved in [6] for finite alphabet processes. The point is then to have a good control on a certain "coupling matrix", which is what we do hereafter for stochastic chains with unbounded memory. We will state it in a form more adapted for our purpose.

Initially we write $f(\sigma_0^n) - \mathbb{E}[f(\sigma_0^n)]$ as a sum of martingale differences. Defining $V_k(\sigma) := \mathbb{E}_\mu[f|\mathcal{F}_k](\sigma) - \mathbb{E}_\mu[f|\mathcal{F}_{k-1}](\sigma)$, we have

$$f(\sigma_0^n) - \mathbb{E}_\mu[f(\sigma_0^n)] = \sum_{k=0}^{n} \left(\mathbb{E}_\mu[f|\mathcal{F}_k](\sigma) - \mathbb{E}_\mu[f|\mathcal{F}_{k-1}](\sigma)\right) = \sum_{k=0}^{n} V_k(\sigma).$$

Observe that $\mathbb{E}_\mu[f|\mathcal{F}_k] = \sum_{\omega_{k+1}^n} f(\sigma_0^k\omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_0^k)$, where $\mu(B|\sigma_0^k) := \mu(B \cap [\sigma_0^k])/\mu([\sigma_0^k])$ for all measurable sets B. Now, we will obtain an upper bound on $V_k(\sigma)$ based on coupling.

[1] Note that we don't have a convolution defined as usual, but one can readily check that the proof of Young's inequality works if we use $\sum_{i\geq 0} u_i v_{j+i}$ instead of $\sum_{i\geq 0} u_i v_{j-i}$.

Lemma 5.1. *For $\sigma \in A^{[\![0,\infty]\!]}$, $a, b \in A$, $n \geq 1$, and $j \geq 0$, let $\nu_j^{\sigma,a,b}$ be any coupling between $\mu(\cdot|\sigma_0^{j-1}a)$ and $\mu(\cdot|\sigma_0^{j-1}b)$. For all $k \in [\![0,n]\!]$ we have*

$$V_k(\sigma) \leq \delta_k f + \sup_{a,b\in A} \sum_{j=1}^{n-k-1} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j})\, \delta_{k+j} f.$$

Proof. Following [6] we have for $\sigma \in A^{n+1}$ We have

$$\begin{aligned}
V_k(\sigma) &= \sum_{\omega_{k+1}^n} f(\sigma_0^k \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_0^k) - \sum_{\omega_k^n} f(\sigma_0^{k-1}\omega_k^n)\mu([\omega_k^n]|\sigma_0^{k-1}) \\
&= \sum_{\omega_{k+1}^n} f(\sigma_0^k \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_0^k) \qquad (5.1)\\
&\quad - \sum_{\omega_k^n} f(\sigma_0^{k-1}\omega_k^n)\mu([\omega_{k+1}^n]|\sigma_0^{k-1}\omega_k)\mu([\omega_k]|\sigma_0^{k-1}) \\
&\leq \sup_{a\in A} \sum_{\omega_{k+1}^n} f(\sigma_0^{k-1} a\, \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_1^{k-1}a) \\
&\quad - \inf_{b\in A} \sum_{\omega_k^n} f(\sigma_0^{k-1} b\, \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_1^{k-1}b)\mu([\omega_k]|\sigma_0^{k-1}) \\
&\leq \sup_{a\in A} \sum_{\omega_{k+1}^n} f(\sigma_0^{k-1} a\, \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_0^{k-1}a) \\
&\quad - \inf_{b\in A} \sum_{\omega_{k+1}^n} f(\sigma_0^{k-1} b\, \omega_{k+1}^n)\mu([\omega_{k+1}^n]|\sigma_0^{k-1}b). \qquad (5.2)
\end{aligned}$$

Let $\eta_k := a$ and $\omega_k := b$. We have

$$\begin{aligned}
&|f(\sigma_1^{k-1} a\, \eta_{k+1}^n) - f(\sigma_1^{k-1} b\, \omega_{k+1}^n)| \\
&\leq \sum_{j=0}^{n-k} |f(\sigma_1^{k-1}\omega_k^{k-1+j}\eta_{k+j}^n) - f(\sigma_1^{k-1}\omega_k^{k+j}\eta_{k+j+1}^n)| \\
&\leq \sum_{j=0}^{n-k} \delta_{k+j} f\, \mathbb{1}_{\{\eta_{k+j}\neq\omega_{k+j}\}}. \qquad (5.3)
\end{aligned}$$

Hence, from (5.2) and (5.3), we have

$$\begin{aligned}
V_k(\sigma) &\leq \sup_{a,b\in A} \sum_{\omega_{k+1}^n} |f(\sigma_0^{k-1} a\, \eta_{k+1}^n) - f(\sigma_0^{k-1} b\, \omega_{k+1}^n)|\, \nu_k^{\sigma,a,b}([\eta_{k+1}^n, \omega_{k+1}^n]) \\
&\leq \sup_{a,b\in A} \sum_{\eta_{k+1}^n, \omega_{k+1}^n} \sum_{j=0}^{n-k} \delta_{k+j} \mathbb{1}_{\{\eta_{k+j}\neq\omega_{k+j}\}}\, \nu_k^{\sigma,a,b}([\eta_{k+1}^n, \omega_{k+1}^n]) \\
&\leq \delta_k f + \sup_{a,b\in A} \sum_{j=1}^{n-k} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j})\, \delta_{k+j} f.
\end{aligned}$$

This ends the proof of the lemma. □

At the end of the proof we used the notation $[a_m^n, b_m^n] := \{(\omega, \eta) \in A^{\mathbb{Z}} \times A^{\mathbb{Z}} : \omega_m^n = a_m^n, \eta_m^n = b_m^n\}$ as a natural extension for denoting cylinder sets in $A^{\mathbb{Z}} \times A^{\mathbb{Z}}$.

Theorem 5.1. *For $\sigma \in A^{[\![0,\infty]\!]}$, $a, b \in A$, $n \geq 1$, and $k \geq 1$, let $\nu_k^{\sigma,a,b}$ be any coupling between $\mu(\cdot|\sigma_0^{k-1}a)$ and $\mu(\cdot|\sigma_0^{k-1}b)$. Also, define*

$$r = \sum_{j=1}^{\infty} \sup_k \sup_\sigma \sup_{a,b} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j}).$$

For all $\theta \in \mathbb{R}$, $n \geq 1$ and $f : A^n \to \mathbb{R}$ we have

$$\mathbb{E}_\mu \left[\mathrm{e}^{\theta(f - \mathbb{E}_\mu f)} \right] \leq \exp\left(\frac{\theta^2 (1+r)^2}{8} \|\underline{\delta} f\|_2^2 \right).$$

As a consequence, we get, for all $u > 0$,

$$\mu(|f - \mathbb{E}_\mu[f]| > u) \leq 2 \exp\left(- \frac{2u^2}{(1+r)^2 \|\underline{\delta} f\|_2^2} \right).$$

Proof. Define

$$U_k(\sigma) = \sup_{a \in A} \sum_{\omega_{k+1}^n} f(\sigma_0^{k-1} a\, \omega_{k+1}^n) \mu([\omega_{k+1}^n] | \sigma_0^{k-1} a) - \mathbb{E}_\mu[f|\mathcal{F}_{k-1}](\sigma)$$

and

$$L_k(\sigma) = \inf_{b \in A} \sum_{\omega_{k+1}^n} f(\sigma_0^{k-1} b\, \omega_{k+1}^n) \mu([\omega_{k+1}^n] | \sigma_0^{k-1} b) - \mathbb{E}_\mu[f|\mathcal{F}_{k-1}](\sigma).$$

For $k, j \geq 0$, let us also define

$$D_{k,k+j} := \sup_\sigma \sup_{a,b \in A} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j}).$$

From Lemma 5.1, we have

$$U_k - L_k \leq \delta_k f + \sum_{j=1}^{n-k} \sup_\sigma \sup_{a,b \in A} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j})\, \delta_{k+j} f = \sum_{j=0}^{n-k} D_{k,k+j}\, \delta_{k+j} f\,.$$

Now observe that $L_k \leq V_k \leq L_k + (U_k - L_k)$, and thus, using Lemma 2.3 of [11] and then proceeding as in the proof of Theorem 1 of [6] we get for all $\theta > 0$

$$\mathbb{E}_\mu \left[\mathrm{e}^{\theta(f - \mathbb{E}_\mu[f])} \right] \leq \mathrm{e}^{\frac{\theta^2 \|D\|_2^2}{8} \|\underline{\delta} f\|_2^2}$$

and, for all $u > 0$,

$$\mu\left(|f - \mathbb{E}_\mu[f]| \geq u\right) \leq 2 \exp\left(- \frac{2u^2}{\|D\|_2^2\, \|\underline{\delta} f\|_2^2} \right).$$

Using the inequality $\|D\|_2^2 \leq \|D\|_1 \|D\|_\infty \leq (1+r)^2$, we conclude the proof of the theorem. □

5.2 One-step maximal coupling

Here we introduce what is called the one-step maximal coupling. We will assume without loss of generality that $A = \{1, \dots, |A|\}$, when A is finite, and $A = \mathbb{N}$ when A is infinite. We will define a probability kernel on $p : A \times A \times \mathcal{X}^- \times \mathcal{X}^- \to [0,1]$ as follows. For $(s, s') \in A \times A$ and $(x, x') \in \mathcal{X}^- \times \mathcal{X}^-$, we put

$$\sum_{c \geq s} \sum_{d \geq s'} p(c, d|x, x') = \sum_{c \geq s} g(c|x) \wedge \sum_{d \geq s'} g(d|x').$$

Let us denote by $\mathbb{P}$ the measure specified by the kernel p. The following equalities

$$\sum_{c \geq s} \sum_{d \geq 1} p(c, d|x, x') = \sum_{c \geq s} g(c|x) \quad \text{and} \quad \sum_{c \geq 1} \sum_{d \geq s'} p(c, d|x, x') = \sum_{d \geq s'} g(d|x')$$

imply that $\mathbb{P}$ is a coupling of two copies of the process specified by g. It is called one-step maximal coupling because it maximizes the probability of agreement (diagonal of the coupling) at each step, given any pair of pasts,

$$p(s, s|x, x') = g(s|x) \wedge g(s|x').$$

In particular, notice that

$$\sum_{c \neq d} p(c, d|x, x') = 1 - \sum_{s \in A} g(s|x) \wedge g(s|x') = \frac{1}{2} \sum_{s \in A} |g(s|x) - g(s|x')|.$$

5.3 Bounding the coupling error by oscillation

We have the following important lemma.

Lemma 5.2. *Take any $x \in \mathcal{X}^-$. For all $a, b \in A$, let $\mathbb{P}^{x,a,b}$ be the one-step maximal coupling between P^{xa} and P^{xb}. Then, for all $j \geq 1$, we have*

$$\mathbb{P}^{x,a,b}(\eta_j \neq \omega_j) \leq \mathrm{Osc}_j(g) + \sum_{k=1}^{j-1} \mathrm{Osc}_{j-k}(g)\, \mathbb{P}^{x,a,b}(\eta_k \neq \omega_k).$$

Proof. We want to compute $\mathbb{P}^{x,a,b}(\eta_i \neq \omega_i)$, which equals

$$\sum_{y_1^{i-1}, z_1^{i-1}} \mathbb{P}^{x,a,b}([y_1^{i-1}, z_1^{i-1}])\, \mathbb{P}^{x,a,b}(\eta_i \neq \omega_i | [y_1^{i-1}, z_1^{i-1}]).$$

Under the maximal coupling we have

$$\mathbb{P}^{x,a,b}(\eta_i \neq \omega_i | [y_1^{i-1}, z_1^{i-1}]) \leq \frac{1}{2} \sum_{s \in A} \left| g(s|xay_1^{i-1}) - g(s|xbz_1^{i-1}) \right|.$$

Let $y_0 := a$ and $z_0 := b$. We get for each $s \in A$

$$\begin{aligned} \left| g(s|xay_1^{i-1}) - g(s|xbz_1^{i-1}) \right| &= \Big| \sum_{j=0}^{i} g(s|xz_0^{j-1}y_j^{i-1}) - g(s|xz_0^j y_{j+1}^{i-1}) \Big| \\ &\leq \sum_{j=0}^{i-1} \left| g(s|xz_0^{j-1}y_j^{i-1}) - g(s|xz_0^j y_{j+1}^{i-1}) \right|. \end{aligned}$$

Putting $y_0 := a$ and $z_0 := b$, we have the bound

$$\frac{1}{2}\sum_{s\in A}\left|g(s|xay_1^{i-1}) - g(s|xbz_1^{i-1})\right| \leq \sum_{j=0}^{i-1} \mathbb{1}_{\{y_j\neq z_j\}}\operatorname{Osc}_{i-j}(g).$$

Hence

$$\begin{aligned}
\mathbb{P}^{x,a,b}(\eta_i \neq \omega_i) &\leq \sum_{y_1^{i-1},z_1^{i-1}} \mathbb{P}^{x,a,b}([y_1^{i-1},z_1^{i-1}]) \sum_{j=0}^{i-1} \mathbb{1}_{\{y_j\neq z_j\}}\operatorname{Osc}_{i-j}(g) \\
&= \sum_{j=0}^{i-1}\sum_{y_1^{i-1},z_1^{i-1}} \mathbb{P}^{x,a,b}([y_1^{i-1},z_1^{i-1}])\mathbb{1}_{\{y_j\neq z_j\}}\operatorname{Osc}_{i-j}(g) \\
&= \sum_{j=0}^{i-1}\sum_{y_j\neq z_j} \mathbb{P}^{x,a,b}([y_j,z_j])\operatorname{Osc}_{i-j}(g) \\
&\leq \operatorname{Osc}_i(g) + \sum_{j=1}^{i-1}\mathbb{P}^{x,a,b}(\eta_j\neq\omega_j)\operatorname{Osc}_{i-j}(g)
\end{aligned}$$

which concludes the proof. □

The following result is a straightforward consequence of Lemma 5.2.

Proposition 5.1. *For all $a, b \in A$ and $x \in \mathcal{X}^-$, let $\mathbb{P}^{x,a,b}$ be the one-step maximal coupling between P^{xa} and P^{xb}. If $\Delta(g) > 0$, then, for all $n \geq 1$, we have*

$$\sum_{j=1}^{n}\sup_{\substack{a,b\in A\\ x\in\mathcal{X}^-}} \mathbb{P}^{x,a,b}(\eta_j\neq\omega_j) \leq \frac{1-\Delta(g)}{\Delta(g)}.$$

Proof. From Lemma 5.2, we have

$$\sup_{\substack{a,b\in A\\ x\in\mathcal{X}^-}} \mathbb{P}^{x,a,b}\left(\eta_j\neq\omega_j\right) \leq \operatorname{Osc}_j(g) + \sum_{k=1}^{j-1}\operatorname{Osc}_{j-k}(g)\sup_{\substack{a,b\in A\\ x\in\mathcal{X}^-}} \mathbb{P}^{x,a,b}\left(\eta_k\neq\omega_k\right). \tag{5.4}$$

Define vectors α and β such that for $i \geq 1$, $\alpha_i = \sup_{\substack{a,b\in A\\ x\in\mathcal{X}^-}} \mathbb{P}^{x,a,b}\left(\eta_i\neq\omega_i\right)$ and $\beta_i = \operatorname{Osc}_i(g)$. We also define a matrix L such that for $i > j$, $L_{ij} = \operatorname{Osc}_{i-j}(g)$ and $L_{ij} = 0$ otherwise. Hence, from (5.4) we have

$$\alpha \leq L\alpha + \beta.$$

Therefore,

$$\|\alpha\|_1 \leq \frac{\|\beta\|_1}{1-\|L\|_1} = \frac{\sum_{j=1}^{\infty}\operatorname{Osc}_j(g)}{1-\sum_{j=1}^{\infty}\operatorname{Osc}_j(g)}$$

as we wanted to prove. □

5.4 Proofs of Theorems 3.1 and 3.2

Proof of Theorem 3.1. In the statement of Theorem 5.1, for $\sigma \in A^{[\![0,\infty]\!]}$, $a, b \in A$, $n \geq 1$, and $k \geq 1$, let $\nu_k^{\sigma,a,b}$ be the one-step maximal coupling between $P^{x\sigma_0^{k-1}a} = P^x(\cdot|\sigma_0^{k-1}a)$ and $P^{x\sigma_0^{k-1}b} = P^x(\cdot|\sigma_0^{k-1}b)$. In this case, for all $x \in \mathcal{X}^-$, Proposition 5.1 implies that $1 + r \leq \Delta(g)^{-1}$, proving Theorem 3.1. □

In order to prove Theorem 3.2, we will need the following lemma. Recall that T denotes the shift operator defined by $(Tx)_i = x_{i+1}$ and define the sequence of shifted measures $P_j^x = P^x \circ T^{-j}, j \geq 1$ in which $P_j^x = P^x$ is the measure compatible with g started from the fixed past $x \in \mathcal{X}^-$.

Lemma 5.3. *Let g be a kernel such that $\mathrm{Var}_k(g) \to 0$ as k diverges and fix a past $x \in \mathcal{X}^-$. Then, the sequence of measures $P_j^x, j \geq 1$ converges, by subsequence, to a measure μ^x which is compatible with g.*

Proof. Let μ^x be a weak accumulation point of $(P_j^x)_j$, which is a measure on $A^{\mathbb{Z}}$. For any $n \in \mathbb{Z}$, by the reverse martingale theorem, there exists a set $\mathcal{S}_n \subset A^{[\![-\infty,n-1]\!]}$ of full μ^x-measure for which, for each $a \in A$

$$\mu^x([a]|\mathcal{F}_{n-\ell}^{n-1})(y) \xrightarrow[\ell\to\infty]{} \mu^x([a]|\mathcal{F}_{-\infty}^{n-1})(y), \quad \forall y \in \mathcal{S}_n.$$

We will consider this convergence on the smaller set $\mathcal{S}_n \cap \mathcal{T}_n$ where $\mathcal{T}_n \subset A^{[\![-\infty,n-1]\!]}$ denotes the set of y's satisfying $\mu^x([y_{n-\ell}^{n-1}]) > 0$ for any $\ell \geq 1$. We can prove that $\mu^x(\mathcal{T}_n) = 1$ for any $n \in \mathbb{Z}$. Indeed, if we let, for any $y \in A^{[\![-\infty,n-1]\!]}$, $m(y) := \inf\{\ell \geq 1 : \mu^x([y_{n-\ell}^{n-1}]) = 0\}$, we have

$$\mu^x(\mathcal{T}_n^c) = \sum_{i\geq 1} \mu^x(\{y \in A^{[\![-\infty,n-1]\!]} : m(y) = i\}) = \sum_{i\geq 1} \sum_{y_{n-i}^{n-1} : m(y_{n-i}^{n-1})=i} \mu^x([y_{n-i}^{n-1}]) = 0.$$

This means in particular that $\mu^x(\mathcal{S}_n \cap \mathcal{T}_n) = 1$.

Now, on $\mathcal{S}_n \cap \mathcal{T}_n$ we can write $\mu^x([a]|\mathcal{F}_{n-\ell}^{n-1})(y) = \mu^x([y_{n-\ell}^{n-1}a])/\mu^x([y_{n-\ell}^{n-1}])$. So to conclude the proof of compatibility of μ^x with g, it is enough to prove that this ratio converges to $g(a|y)$ as ℓ tends to $+\infty$, for any $y \in \mathcal{S}_n \cap \mathcal{T}_n$. This is proven below using the facts that P_j^x converges by subsequence to μ^x, and that $\mathrm{Var}_k(g) \to 0$.

The convergence by subsequence tells us that there exists a subsequence $n_j, j \geq 1$ such that $P_{n_j}^x \xrightarrow{j} \mu^x$, where for $n_j \geq n - \ell$,

$$P_{n_j}^x([y_{n-\ell}^{n-1}a]) = \sum_{z_{-n_j}^{n-\ell-1}} P_{n_j}^x([z_{-n_j}^{n-\ell-1}y_{n-\ell}^{n-1}])\, g(a|xz_{-n_j}^{n-\ell-1}y_{n-\ell}^{n-1}).$$

On the other hand, recalling the definition of $\mathrm{Var}_k(g), k \geq 0$, observe that

$$|g(a|y) - g(a|xz_{-j}^{n-\ell-1}y_{n-\ell}^{n-1})| \leq \mathrm{Var}_\ell(g).$$

We now compute

$$\begin{aligned}
\mu^x(y_{n-\ell}^{n-1}a) &= \lim_j P^x_{n_j}([y_{n-\ell}^{n-1}a]) \\
&= \lim_j \sum_{z_{-j}^{n-\ell-1}} P^x_j([z_{-j}^{n-\ell-1}y_{n-\ell}^{n-1}])g(a|xz_{-j}^{n-\ell-1}y_{n-\ell}^{n-1}) \\
&\leq (g(a|y) + \mathrm{Var}_\ell(g)) \lim_j \sum_{z_{-j}^{n-\ell-1}} P^x_{n_j}([z_{-j}^{n-\ell-1}y_{n-\ell}^{n-1}]) \\
&\leq (g(a|y) + \mathrm{Var}_\ell(g)) \lim_j P^x_{n_j}([y_{n-\ell}^{n-1}]) \\
&\leq (g(a|y) + \mathrm{Var}_\ell(g))\, \mu^x([y_{n-\ell}^{n-1}]).
\end{aligned}$$

Naturally, we similarly obtain $\mu^x([y_{n-\ell}^{n-1}a]) \geq (g(a|y) - \mathrm{Var}_\ell(g))\mu^x([y_{n-\ell}^{n-1}])$. Now using that $\mathrm{Var}_\ell(g) \to 0$, we obtain that for any $y \in \mathcal{S}_n \cap \mathcal{T}_n$

$$\mu^x([a]|\mathcal{F}_{n-\ell}^{n-1})(y) = \frac{\mu^x([y_{n-\ell}^{n-1}a])}{\mu^x([y_{n-\ell}^{n-1}])} \to g(a|y).$$

This concludes the proof of compatibility of μ^x with g. □

We are now ready for the proof of Theorem 3.2

Proof of Theorem 3.2. Let P^x be the measure of the chain with kernel g starting from a fixed past $x \in \mathcal{X}^-$. Let T be the shift operator defined by $(Tx)_i = x_{i+1}$. We define $P^x_j = P^x \circ T^{-j}$. If g satisfies $\Delta(g) > 0$ and $|A| < \infty$, then by Theorem 4.6 of [18] g has a unique compatible measure, which therefore has to be μ. Because $A^{\mathbb{Z}}$ is compact (recall that we assume A finite for this theorem), there is a subsequence of $P^x_j, j \geq 1$, that converges weakly, and since $\Delta(g) > 0$ we have $\mathrm{Var}_j(g) \to 0$ which, by Lemma 5.3, implies that the limit is a measure compatible with g. Since the only compatible measure is μ, all convergent subsequence converge weakly to μ. This means that the complete sequence, $P^x_j, j \geq 1$, converges weakly to μ, *i.e.*, $\mathbb{E}_{P^x_j}[h] \to \mathbb{E}_\mu[h]$ for all bounded continuous functions $h : A^{[\![0,\infty]\!]} \to \mathbb{R}$. Moreover, we have that, for all $B \in \mathcal{F}_n$, $P^x_j(B) \to \mu(B)$. For $\sigma \in A^{[\![0,n+j]\!]}$, let $f_j(\sigma) := f(\sigma_j^{j+n})$. Fix $u \in \mathbb{R}$. For all $\epsilon > 0$ there is a j_0 such that for all $j \geq j_0$

$$\begin{aligned}
\mu(|f - \mathbb{E}_\mu[f]| > u) &\leq P^x_j\Big(|f - \mathbb{E}_\mu[f]| > u \Big) + \epsilon \\
&\leq P^x_j\Big(\Big|f - \mathbb{E}_{P^x_j}[f]\Big| > u - \epsilon \Big) + \epsilon \\
&= P^x\Big(|f_j - \mathbb{E}_{P^x}[f_j]| > u - \epsilon \Big) + \epsilon \\
&\leq 2\exp\left(-\frac{2(u-\epsilon)^2}{\Delta(g)^{-2}\|\underline{\delta} f\|_2^2} \right) + \epsilon
\end{aligned}$$

where the last line is a consequence of Theorem 3.1 and because $\|\underline{\delta} f\|_2 = \|\underline{\delta} f_j\|_2$. Taking $\epsilon \to 0$, we conclude the proof. □

5.5 Bounding the coupling error by variation

Fix $y, z \in A^{[\![-\infty,0]\!]}$. Let $\mathbb{P}^{y,z}$ be the one-step maximal coupling between measures P^y and P^z with kernel g. We want to obtain an upper bound $\mathbb{P}^{y,z}(\eta_j \neq \omega_j)$. To achieve this, we will use an auxiliary process. Given $x \in \mathcal{X}^-$, let $\ell(x) = \inf\{k \geq 1 : x_{-k} = 1\}$ and $\ell(\dots 00) = \infty$. Consider the kernel h associated with g as $h^g : \{0,1\} \times \{0,1\}^{[\![-\infty,-1]\!]} \to (0,1)$ where $h^g(1|x) = q_{\ell(x)}$ and $q_j = \mathrm{Var}_j(g)$. For any $x \in \{0,1\}^{\mathcal{X}^-}$ consider the measure P_h^{x1} constructed as in Definition 2.2 using h in place of g. It is the undelayed renewal measure, and will be our auxiliary process. Recall the definition of the projection functions π_i^j, $i \leq j$ given in Section 2, and put $\pi_i := \pi_i^i$ as the projection on the single coordinate i.

Lemma 5.4. *We have that, for all $y, z \in A^{[\![-\infty,0]\!]}$ and all $j \geq 0$,*

$$\mathbb{P}^{y,z}(\eta_j \neq \omega_j) \leq P_h^{x1}(\pi_j = 1).$$

Proof. For any η, ω in $A^{\mathbb{Z}}$, let $\sigma_j(\eta,\omega) := \mathbb{1}_{\{\eta_j \neq \omega_j\}}$ for $j \in \mathbb{Z}$. By definition, for all $j \geq 0$, we have

$$\sup_{y,z \in A^{[\![-\infty,0]\!]}} \mathbb{P}^{y,z}\left(\sigma_j = 1 | \sigma_1^{j-1} = 0, \sigma_0 = 1\right) \leq \mathrm{Var}_{j+1}(g) = q_{j+1}.$$

By stochastic domination, we conclude that

$$\mathbb{P}^{y,z}(\sigma_j = 1) \leq P_h^{x1}(\pi_j = 1)$$

as we wanted to prove. □

Lemma 5.5. *Let $\gamma_1(g) = \mathrm{Var}_0(g)$ and for $k \geq 2$, $\gamma_k(g) := \mathrm{Var}_{k-1}(g) \prod_{i=0}^{k-2}(1 - \mathrm{Var}_i(g))$. If P_h^{x1} be the measure specified by renewal kernel h^g starting with $x_0 = 1$ then, for all $j \geq 1$, we have the renewal equation*

$$P_h^{x1}(\pi_j = 1) = \gamma_j(g) + \sum_{k=1}^{j-1} \gamma_{j-k}(g)\, P_h^{x1}(\pi_k = 1).$$

Proof. We want to compute (for any $i \geq 2$ we denote by 0^i the string $00\dots0$ of i consecutive 0)

$$P_h^{x1}(\pi_j = 1) = P_h^{x1}(\pi_j = 1, \pi_1^{j-1} = 0^{j-1}) + \sum_{k=1}^{j-1} P_h^{x1}(\pi_j = 1, \pi_{j-k+1}^{j-1} = 0^{k-1}, \pi_{j-k} = 1).$$

For $k \geq 1$, we have that

$$\begin{aligned} P_h^{x1}(\pi_j = 1, \pi_{j-k+1}^{j-1} = 0^{k-1}, \pi_{j-k} = 1) &= P_h^{x1}(\pi_j = 1 | \pi_{j-k+1}^{j-1} = 0^{k-1}, \pi_{j-k} = 1) \\ &\quad \times \prod_{i=j-k+1}^{j-1} P_h^{x1}(\pi_i = 0 | \pi_{j-k+1}^{i-1} = 0^{k-1}, \pi_{j-k} = 1) \\ &= \mathrm{Var}_{k-1}(g) \prod_{i=0}^{k-2}(1 - \mathrm{Var}_i(g))\, P_h^{x1}(\pi_{j-k} = 1) \\ &= \gamma_k(g)\, P_h^{x1}(\pi_{j-k} = 1) \end{aligned}$$

where we used the convention $\prod_{i=0}^{-1} = 1$. Similarly,

$$P_h^{x1}\big(\pi_j = 1, \pi_1^{j-1} = 0^{j-1}\big) = \gamma_j(g).$$

Therefore, we have

$$P_h^{x1}(\pi_j = 1) \le \gamma_j(g) + \sum_{k=1}^{j-1} \gamma_k(g)\, P_h^{x1}(\pi_{j-k} = 1).$$

Using the symmetry between the indices k and $j-k$ in the summation, we conclude the proof. □

The next result is a direct consequence of Lemmas 5.4 and 5.5.

Proposition 5.2. *For all $y, z \in A^{\mathbb{[}-\infty,0\mathbb{]}}$, let $\mathbb{P}^{y,z}$ be the one-step maximal coupling between P^y and P^z. If $\Gamma(g) := \prod_{j=0}^{\infty}(1 - \mathrm{Var}_j(g)) > 0$, then, for all $n \ge 1$, we have*

$$\sum_{j=1}^{n} \sup_{y,z \in A^{\mathbb{[}-\infty,0\mathbb{]}}} \mathbb{P}^{y,z}(\eta_j \neq \omega_j) \le \frac{1 - \Gamma(g)}{\Gamma(g)}.$$

Proof. From Lemma 5.5, we have

$$P_h^{x1}\left(\pi_j = 1\right) = \gamma_j(g) + \sum_{k=1}^{j-1} \gamma_{j-k}(g) P_h^{x1}\left(\pi_k = 1\right). \tag{5.5}$$

Define vector α such that for $j \ge 1$, $\alpha_j = P_h^{x1}\left(\pi_j = 1\right)$. We also define a matrix L such that for $i > j$, $L_{ij} = \gamma_{i-j}(g)$ and $L_{ij} = 0$ otherwise. Therefore, from (5.5) we have

$$\alpha \le L\alpha + \beta.$$

Therefore,

$$\|\alpha\|_1 \le \frac{\|\gamma\|_1}{1 - \|L\|_1} = \frac{\sum_{j=1}^{\infty} \mathrm{Var}_{j-1}(g) \prod_{k=0}^{j-2}(1 - \mathrm{Var}_k(g))}{1 - \sum_{j=1}^{\infty} \mathrm{Var}_{j-1}(g) \prod_{k=0}^{j-2}(1 - \mathrm{Var}_k(g))}. \tag{5.6}$$

Because

$$\mathrm{Var}_{j-1}(g) \prod_{k=0}^{j-2}(1 - \mathrm{Var}_k(g)) = \prod_{k=0}^{j-2}(1 - \mathrm{Var}_k(g)) - \prod_{k=0}^{j-1}(1 - \mathrm{Var}_k(g)),$$

we have

$$\sum_{j=1}^{\infty} \mathrm{Var}_{j-1}(g) \prod_{k=0}^{j-2}(1 - \mathrm{Var}_k(g)) = \mathrm{Var}_0(g) - \prod_{k=0}^{\infty}(1 - \mathrm{Var}_k(g)) \le 1 - \prod_{k=1}^{\infty}(1 - \mathrm{Var}_k(g)).$$

Therefore, from (5.6), we get

$$\|\alpha\|_1 \le \frac{1 - \prod_{k=0}^{\infty}(1 - \mathrm{Var}_k(g))}{\prod_{k=0}^{\infty}(1 - \mathrm{Var}_k(g))}.$$

From Lemma 5.4, we have that

$$\sum_{j=1}^{n} \sup_{y,z \in A^{\mathbb{[}-\infty,0\mathbb{]}}} \mathbb{P}^{y,z}(\eta_j \neq \omega_j) \le \sum_{j=1}^{n} P_h^{x1}(\pi_j = 1),$$

which concludes the proof. □

5.6 Proofs of Theorems 3.3 and 3.4

Proof of Theorem 3.3. We proceed exactly as in the proof of Theorem 3.1, substituting $\Delta(g)$ by $\Gamma(g)$. □

Proof of Theorem 3.4. Let μ be a measure compatible with g. Let also $\sigma \in A^{[\![0,\infty]\!]}$, $a, b \in A$, $n \geq 1$, and $k \geq 1$. Let first define a coupling $\nu_k^{\sigma,a,b}$ between $\mu(\cdot|\sigma_0^{k-1}a)$ and $\mu(\cdot|\sigma_0^{k-1}b)$. For all $x, y \in \mathcal{X}^-$, let $\mathbb{P}_k^{x,y,\sigma,a,b}$ be the one-step maximal coupling between $P^x(\cdot|\sigma_0^{k-1}a)$ and $P^y(\cdot|\sigma_0^{k-1}b)$. We define

$$\nu_k^{\sigma,a,b}(\cdot) = \int_{\mathcal{X}^-}\int_{\mathcal{X}^-} \mathbb{P}_k^{x,y,\sigma,a,b}(\cdot)\,\mu\big(\,\mathrm{d}x|\sigma_0^{k-1}a\big)\,\mu\big(\,\mathrm{d}y|\sigma_0^{k-1}a\big).$$

From the definition of the above coupling, we have

$$\begin{aligned}\sup_k \sup_\sigma \sup_{a,b} \nu_k^{\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j}) &\leq \sup_k \sup_\sigma \sup_{a,b} \sup_{x,y} \mathbb{P}_k^{x,y,\sigma,a,b}(\eta_{k+j} \neq \omega_{k+j}) \\ &\leq \sup_{y,z \in A^{[\![-\infty,0]\!]}} \mathbb{P}^{y,z}(\eta_j \neq \omega_j)\end{aligned}$$

where $\mathbb{P}^{y,z}$ is the one-step maximal coupling between P^y and P^z. Using Proposition 5.2 and Theorem 5.1, we conclude that $1 + r \leq \Gamma(g)^{-1}$ and GCB holds, as we wanted to show. □

5.7 Proof of Theorem 3.5

We first need to define two properties: the positive divergence property, and the blowing-up property.

Definition 5.1. *We say that an ergodic measure μ on $A^{\mathbb{Z}}$ satisfies the positive divergence property if for any ergodic measure ν on $A^{\mathbb{Z}}$ different from μ we have*

$$\liminf_n \frac{1}{n+1}\,\mathbb{E}_{\nu_n}\left[\log\frac{\nu_n}{\mu_n}\right] > 0$$

where $\mu_n = \mu|_{\mathcal{F}_n}$ and $\nu_n = \nu|_{\mathcal{F}_n}$.

We now state two propositions that we will use to prove the next theorem.

Proposition 5.3. *Let g be a kernel such that $\inf_{a\in A, x\in\mathcal{X}^-} g(a|x) > 0$ and $\lim_j \mathrm{Var}_j(g) = 0$. Suppose that there are two distinct ergodic measures μ and μ compatible with g. Then the positive divergence property does not hold.*

Proof. Let μ be an ergodic measure compatible with a kernel g, and ν another ergodic measure compatible with a kernel h. Assume that $\lim_j \mathrm{Var}_j(g) = \lim_j \mathrm{Var}_j(h) = 0$ and $\inf_{a\in A, x\in\mathcal{X}^-} g(a|x) > 0$, $\inf_{a\in A, x\in\mathcal{X}^-} h(a|x) > 0$. We have

$$\begin{aligned}&\frac{1}{n+1}\,\mathbb{E}_{\nu_n}\left[\log\frac{\nu_n}{\mu_n}\right] = \frac{1}{n+1}\int \log\frac{\nu([x_0^n])}{\mu([x_0^n])}\,\nu(\mathrm{d}x_{-\infty}^n) \\ &= \int \frac{1}{n+1}\left(\sum_{j=1}^n \log\frac{\nu\big([x_j]|x_0^{j-1}\big)}{\mu\big([x_j]|x_0^{j-1}\big)} + \log\frac{\nu([x_0])}{\mu([x_0])}\right)\nu(\mathrm{d}x_{-\infty}^n) \\ &= \int \frac{1}{n+1}\sum_{j=1}^n \log\frac{\nu\big([x_0]|x_{-j}^{-1}\big)}{\mu\big([x_0]|x_{-j}^{-1}\big)}\,\nu(\mathrm{d}x_{-\infty}^0) + \frac{1}{n+1}\sum_{x_0\in A}\log\frac{\nu([x_0])}{\mu([x_0])}\,\nu([x_0])\end{aligned}$$

where the last equality uses shift-invariance of the measures. By uniform continuity of g and h we have

$$\log \frac{\nu\big([x_0]|x_{-j}^{-1}\big)}{\mu\big([x_0]|x_{-j}^{-1}\big)} \xrightarrow{j\to\infty} \log \frac{h(x)}{g(x)}$$

uniformly in x, and therefore

$$\frac{1}{n+1}\sum_{j=1}^{n} \log \frac{\nu\big([x_0]|x_{-j}^{-1}\big)}{\mu\big([x_0]|x_{-j}^{-1}\big)} \xrightarrow{n\to\infty} \log \frac{h(x)}{g(x)} \tag{5.7}$$

uniformly in x by Cesàro lemma. By the dominated convergence theorem we conclude that

$$\lim_n \frac{1}{n+1}\,\mathbb{E}_{\nu_n}\left[\log \frac{\nu_n}{\mu_n}\right] = \mathbb{E}_{\nu}\left[\log \frac{h}{g}\right].$$

Therefore, if $g = h$ and if there are multiple ergodic measures compatible with g, then there cannot have the positive divergence property. Indeed, if ν is an ergodic measure compatible with g but different from μ, then the r.h.s. of (5.7) is equal to 0, which violates the positive divergence property. □

For all $n \geq 0$, define the normalised Hamming distance between ω and σ on A^{n+1} by

$$\bar{d}_n(\sigma,\omega) = \frac{1}{n+1}\sum_{i=0}^{n} \mathbb{1}_{\{\sigma_i \neq \omega_i\}}. \tag{5.8}$$

For $F \subset A^{n+1}$ and $\epsilon > 0$, $\langle F\rangle_\epsilon$ denotes the ϵ-blowup of F, that is

$$\langle F\rangle_\epsilon = \{\sigma \in A^{n+1} : \bar{d}_n(\sigma,\omega) \leq \epsilon \text{ for some } \omega_0^n \in F\}.$$

Definition 5.2. *An ergodic measure μ has the blowing-up property if given $\epsilon > 0$ there is a $\delta > 0$ and n_0 such that if $n \geq n_0$ then $\mu(\langle F\rangle_\epsilon) \geq 1-\epsilon$, for any subset $F \subset A^{n+1}$ for which $\mu(F) \geq \mathrm{e}^{-(n+1)\delta}$.*

We make a slight abuse of notation by writing $\mu(F)$ instead of $\mu_n(F)$, or, stated differently, we use the same notation for a subset of A^{n+1} and the union of cylinders it generates.

Proposition 5.4. *Suppose that μ is a probability measure which satisfies GCB with a constant C. For any $n \geq 0$ and any $F \subset A^{n+1}$ such that $\mu(F) > 0$, we have*

$$\mu(\langle F\rangle_\epsilon) \geq 1 - \exp\left(-\frac{n+1}{4C}\left(\epsilon - 2\sqrt{\frac{C\log(\mu(F)^{-1})}{n+1}}\right)^2\right) \tag{5.9}$$

whenever $\epsilon > 2\sqrt{\frac{C\log(\mu(F)^{-1})}{n+1}}$. In particular, μ has the blowing-up property.

Proof. Let $n \geq 0$ and $f(\omega_0^n) = \inf_{\sigma_0^n \in F}\sum_{i=0}^{n} \mathbb{1}_{\{\sigma_i \neq \omega_i\}}$. It is obvious that $\delta_i f = 1$ for $i = 0,\ldots,n$. Since μ satisfies GCB with a constant C by assumption, we get from (2.5)

$$\mu(f > \mathbb{E}_\mu[f] + u) \leq \exp\left(-\frac{u^2}{4C(n+1)}\right), \quad u > 0.$$

We now derive an upper bound for $\mathbb{E}_\mu[f]$. We use (2.4) with $-\theta f$, where $\theta > 0$ will be fixed later on, to get

$$\exp(\theta \mathbb{E}_\mu[f])\, \mathbb{E}_\mu\big[\exp(-\theta f)\big] \leq \exp\big(C\theta^2(n+1)\big).$$

But, by the very definition of f, we have

$$\mathbb{E}_\mu\big[\exp(-\theta f)\big] \geq \mathbb{E}_\mu\big[\exp(-\theta f)\mathbb{1}_F\big] = \mu(F).$$

Hence, combining the two previous inequalities, taking the logarithm, and dividing out by θ, we obtain

$$\mathbb{E}_\mu[f] \leq \inf_{\theta>0}\left\{C(n+1)\theta + \frac{1}{\theta}\log\big(\mu(F)^{-1}\big)\right\}$$

which gives

$$\mathbb{E}_\mu[f] \leq 2\sqrt{C(n+1)\log(\mu(F)^{-1})}.$$

To finish the proof of (5.9), observe that $\mu(f > \epsilon) = \mu(\langle F\rangle_\epsilon^c)$.
Now, if we fix $\epsilon > 0$ and take F such that $\mu(F) \geq \exp(-(n+1)\delta)$, for some $\delta > 0$ to be chosen later on, subject to the condition $\epsilon > 2\sqrt{C\delta}$, we get from (5.9) that, for all $n \geq 0$,

$$\mu(\langle F\rangle_\epsilon) \geq 1 - \exp\left(-\frac{n+1}{4C}\left(\epsilon - 2\sqrt{C\delta}\right)^2\right).$$

We now take $\delta = \epsilon^2/(4C)$ which gives

$$\mu(\langle F\rangle_\epsilon) \geq 1 - \epsilon$$

for all $n \geq n_0 := \lfloor 4\epsilon^{-2}\log(\epsilon^{-1})\rfloor$. We thus proved that GCB implies the blowing-up property. □

We are ready to prove the following result, which is of independent interest.

Proof of Theorem 3.5. If $\inf_{a\in A, x\in\mathcal{X}^-} g(a|x) > 0$, then the alphabet has to be finite (see Remark 4.1). It is proved in [37] that, for finite alphabet ergodic stationary processes, the blowing-up property implies the positive divergence property. But by Proposition 5.3, we cannot have the latter property since we assume that there are at least two ergodic measures compatible with the kernel. Hence the blowing-up properties does not hold. But then, by Proposition 5.4, we cannot have GCB. □

5.8 Proof of Theorem 3.6

Recall the definitions of $\tilde{g}$, the kernel of the renewal measure μ, and the distribution $f_n, n \geq 1$ of the distance between consecutive 1's. In order to prove Theorem 3.6 we will use a well-known relation between the renewal process and an $\mathbb{N}$-valued Markov chain. Indeed, let $F : \mathbb{N}^{\mathbb{N}} \to \{0,1\}^{\mathbb{N}}$ be the deterministic coordinate-wise function defined by $\big(F(\sigma)\big)_i = \mathbb{1}_{[0]}(\sigma_i), i \in \mathbb{N}$. We refer the reader to [40], where in particular it is explained that $\mu = \nu \circ F^{-1}$ where ν is the Markov measure with transition matrix Q given by

$$Q(m,0) = 1 - Q(m,m+1) = \frac{f_{m+1}}{\sum_{i\geq m+1} f_i}, \qquad m \geq 0.$$

Proof of Theorem 3.6. We start by proving sufficiency. Suppose first that $\sum_n f_n r^n < \infty$ for some $r > 1$. Then the time τ^0 separating two consecutive 0's for the Markov measure ν has distribution $f_n, n \geq 1$, by construction. Therefore, $\mathbb{E}_{\nu^0}[r^{\tau^0}] < \infty$ for the same r, where $\mathbb{E}_{\nu^0}$ denotes the expectation with respect to the measure of the Markov chain initiated at state 0. Following [40], this characterizes ν as a geometrically ergodic Markov measure (in fact, it is equivalent, see [40, Theorem 15.1.4]). Using the result of [10], we conclude that ν satisfies GCB, and, as a coordinate-wise image of ν, the renewal process μ also has GCB. This last step is a consequence of [33, Theorem 7.1].

We now prove necessity. Suppose that μ satisfies GCB. Then, for some sufficiently small $c > 0$,

$$\begin{aligned}\mu([0^{n+1}]) &= \mu\left(\left\{\omega : \frac{1}{n+1}\sum_{i=0}^{n}\omega_i = 0\right\}\right)\\ &\leq \mu\left(\left\{\omega : \frac{1}{n+1}\sum_{i=0}^{n}\omega_i - \mu([1]) \leq -\frac{\mu([1])}{2}\right\}\right)\\ &\leq \mathrm{e}^{-cn}.\end{aligned}$$

On the other hand, by stationarity

$$\mu([0^{n+1}]) = \sum_{i\geq n+1}\mu([10^i]) = \mu([1])\sum_{i\geq n+1}\sum_{j\geq i} f_j \geq \mu([1])\sum_{i\geq n+1} f_i.$$

This means that $\mu([1])\sum_{i\geq n+1} f_i \leq \mathrm{e}^{-cn}$ which implies that $\sum_n f_n r^n < \infty$ for some $r > 1$. □

5.9 Proof of Theorem 3.7

Take an arbitrary $\eta \in A^{\mathbb{N}}$ and for $n \geq 0$ define $f_n(\omega) := f(\omega_0^n \eta_{n+1}^{\infty})$. By construction we have $\|f - f_n\|_\infty \leq \mathrm{var}_n(f) \to 0$. Now, for each i, $\delta_i(f - f_n)$ goes to 0 when $n \to \infty$ since for all $n \geq i$ it is easy to check that

$$\delta_i(f - f_n) \leq 2\,\mathrm{var}_n(f).$$

We have the inequality

$$(\delta_i(f - f_n))^2 \leq 4(\delta_i f)^2,\ \forall i, n.$$

Therefore, since $\|\underline{\delta} f\|_2 < \infty$ by assumption, we can use the dominated convergence theorem (for the counting measure on the set of nonnegative integers) to get

$$\|\underline{\delta}(f - f_n)\|_2 \to 0.$$

Finally, using GCB for f_n, and the obvious fact that $\delta_i(f + g) \leq \delta_i f + \delta_i g$, we get

$$\begin{aligned}\mathbb{E}_\mu\left[\mathrm{e}^{f-\mathbb{E}_\mu[f]}\right] &\leq \mathbb{E}_\mu\left[\mathrm{e}^{f_n-\mathbb{E}_\mu[f_n]}\right]\mathrm{e}^{2\|f-f_n\|_\infty}\\ &\leq \mathrm{e}^{C\|\underline{\delta} f_n\|_2^2}\,\mathrm{e}^{2\|f-f_n\|_\infty}\\ &\leq \mathrm{e}^{C\|\underline{\delta} f\|_2^2}\,\mathrm{e}^{2C\|\underline{\delta}(f)\|_2\|\underline{\delta}(f-f_n)\|_2}\,\mathrm{e}^{C\|\underline{\delta}(f-f_n)\|_2^2}\,\mathrm{e}^{2\|f-f_n\|_\infty}\end{aligned}$$

where the third inequality follows by writting $f_n = f_n - f + f$ and expanding $(\delta_i(f_n - f + f))^2$. The result follows by letting n tend to infinity.

5.10 Proof of Theorem 4.1

Define $f = \|\hat{\rho}_{n,k} - \rho\|_\infty$. For all $n \geq 1$, we have $\|\underline{\delta} f\|_2 = 1$, hence, from Theorem 3.1, we have

$$\mu(\|\hat{\rho}_{n,k} - \rho\|_\infty - \mathbb{E}_\mu[\|\hat{\rho}_{n,k} - \rho\|_\infty] > u) \leq \exp\left(-2(n-k+2)\Gamma(g)^2 u^2\right). \tag{5.10}$$

Therefore, to prove Theorem 4.1, we only need to find a good upper bound for $\mathbb{E}_\mu[\|\hat{\rho}_{n,k} - \rho\|_\infty]$. Here, we follow the argument used in [32]. By Jensen's inequality, and since $\mathbb{E}_\mu[\hat{\rho}_{n,k}(\sigma)] = \rho(\sigma)$, we have

$$\begin{aligned}
\left(\mathbb{E}_\mu[\|\hat{\rho}_{n,k} - \rho\|_\infty]\right)^2 &\leq \mathbb{E}_\mu\left[\|\hat{\rho}_{n,k} - \rho\|_\infty^2\right] \\
&\leq \mathbb{E}_\mu\left[\sum_{a \in A} (\hat{\rho}_{n,k}(\sigma) - \rho(\sigma))^2\right] \\
&\leq \sum_{\sigma \in A^{[\![1,k]\!]}} \left(\mathbb{E}_\mu\left[\hat{\rho}_{n,k}(\sigma)^2\right] - \rho(\sigma)^2\right).
\end{aligned} \tag{5.11}$$

We remind that, for all $S \subset \mathbb{Z}$ and $\sigma \in A^S$, we define the projection function associated to all indices $i, j \in S$, $j \leq i$, by $\pi_j^i(\sigma) = \sigma_j^i$. For all $\sigma \in A^{[\![1,k]\!]}$, we have

$$\begin{aligned}
&\mathbb{E}_\mu\left[\hat{\rho}_{n,k}(\sigma)^2\right] \\
&= \frac{1}{(n-k+2)^2}\,\mathbb{E}_\mu\left[\left(\sum_{i=0}^{n-k+1} \mathbb{1}_{\{\sigma\}} \circ \pi_i^{i+k-1}\right)^2\right] \\
&= \frac{1}{(n-k+2)^2}\,\mathbb{E}_\mu\left[\sum_{i=0}^{n-k+1} \mathbb{1}_{\{\sigma\}} \circ \pi_i^{i+k-1} + 2\sum_{j=1}^{n-k+1}\sum_{i=0}^{j-1} (\mathbb{1}_{\{\sigma\}} \circ \pi_i^{i+k-1})(\mathbb{1}_{\{\sigma\}} \circ \pi_j^{j+k-1})\right] \\
&= \frac{\rho(\sigma)}{n-k+2} + \frac{2}{(n-k+2)^2}\sum_{j=1}^{n-k+1}\sum_{i=0}^{j-1} \mu(\pi_i^{i+k-1} = \sigma, \pi_j^{j+k-1} = \sigma) \\
&= \frac{\rho(\sigma)}{n-k+2} + \frac{2}{(n-k+2)^2}\sum_{j=1}^{n-k+1}\sum_{i=0}^{j-1} \rho(\sigma)\mu(\pi_j^{j+k-1} = \sigma | \pi_i^{i+k-1} = \sigma) \\
&\leq \frac{\rho(\sigma)}{n-k+2} + \frac{2}{(n-k+2)^2}\sum_{j=1}^{n-k+1}\sum_{i=0}^{j-1} \rho(\sigma)\big(\rho(\sigma) + |\mu(\pi_j^{j+k-1} = \sigma | \pi_i^{i+k-1} = \sigma) - \rho(\sigma)|\big).
\end{aligned}$$

Now let $j^* = \max\{j, i+k\}$. For all $a \in A$,

$$
\begin{aligned}
&|\mu(\pi_j^{j+k-1} = \sigma | \pi_i^{i+k-1} = \sigma) - \rho(\sigma)| \\
&\leq \sup_{\tilde{\sigma} \in A^{[\![1,k]\!]}} |\mu(\pi_{j^*}^{j+k-1} = \pi_{j^*}^{j+k-1}(\sigma) | \pi_i^{i+k-1} = \sigma) - \mu(\pi_{j^*}^{j+k-1} = \pi_{j^*}^{j+k-1}(\sigma) | \pi_i^{i+k-1} = \tilde{\sigma})| \\
&= \sup_{\tilde{\sigma} \in A^{[\![1,k]\!]}} |\mu(\pi_{j^*-i-k}^{j-i-1} = \pi_{j^*-i-k}^{j-i-1}(\sigma) | \pi_{-k}^{-1} = \sigma) - \mu(\pi_{j^*-i-k}^{j-i-1} = \pi_{j^*-i-k}^{j-i-1}(\sigma) | \pi_{-k}^{-1} = \tilde{\sigma})| \\
&\leq \sup_{x,y \in \mathcal{X}^-} |P^x(\pi_{j^*-i-k}^{j-i-1} = \pi_{j^*-i-k}^{j-i-1}(\sigma)) - P^y(\pi_{j^*-i-k}^{j-i-1} = \pi_{j^*-i-k}^{j-i-1}(\sigma))| \\
&\leq \sup_{x,y \in \mathcal{X}^-} \mathbb{P}^{x,y}(\eta_{j^*-i-k}^{j-i-1} \neq \omega_{j^*-i-k}^{j-i-1}) \\
&\leq \sup_{x,y \in \mathcal{X}^-} \sum_{\ell=j^*-i-k}^{j-i-1} \mathbb{P}^{x,y}(\eta_\ell \neq \omega_\ell)
\end{aligned}
$$

where $\mathbb{P}^{x,y}$ is the one-step maximal coupling between P^x and P^y. Observe that $j^* - i - k = \max\{j - i - k, 0\}$. Coming back to the estimation of $\mathbb{E}_\mu\left[\hat{\rho}_{n,k}(\sigma)^2\right]$, we have

$$
\begin{aligned}
&\mathbb{E}_\mu\left[\hat{\rho}_{n,k}(\sigma)^2\right] \\
&\leq \frac{\rho(\sigma)}{n-k+2} + \frac{2}{(n-k+2)^2} \sum_{j=1}^{n-k+1} \sum_{i=0}^{j-1} \rho(\sigma) \Bigg(\rho(\sigma) + \sup_{x,y \in \mathcal{X}^-} \sum_{\ell=\max\{j-i-k,0\}}^{j-i-1} \mathbb{P}^{x,y}(\eta_\ell \neq \omega_\ell) \Bigg) \\
&\leq \frac{\rho(\sigma)}{n-k+2} + \rho(\sigma)^2 + \frac{2\rho(\sigma)(n-k+1)}{(n-k+2)^2} \sum_{i=0}^{n-k} \sum_{\ell=\max\{n-2k-i+1,0\}}^{n-k-i} \sup_{x,y \in \mathcal{X}^-} \mathbb{P}^{x,y}(\eta_\ell \neq \omega_\ell)\Big) \\
&\leq \frac{\rho(\sigma)}{n-k+2} + \rho(\sigma)^2 + \frac{2\rho(\sigma)k}{n-k+2} \sum_{i=0}^{n-k} \sup_{x,y \in \mathcal{X}^-} \mathbb{P}^{x,y}(\eta_i \neq \omega_i) \\
&\leq \frac{\rho(\sigma)}{n-k+2} + \rho(\sigma)^2 + \frac{2\rho(\sigma)k}{n-k+2} \sum_{i=0}^{\infty} \sup_{x,y \in \mathcal{X}^-} \mathbb{P}^{x,y}(\eta_i \neq \omega_i) \\
&\leq \rho(\sigma)^2 + \frac{2\rho(\sigma)k}{(n-k+2)\Gamma(g)}
\end{aligned}
$$

where we used Proposition 5.2 in the last inequality. Finally, we obtain from (5.11) that

$$
\mathbb{E}_\mu[\|\hat{\rho}_{n,k} - \rho\|_\infty] \leq \sqrt{\frac{2k}{(n-k+2)\Gamma(g)}}
$$

which is the desired bound. Combining this bound with (5.10) and rescaling u in an obvious way, we finally obtain (4.1).

5.11 Proof of Theorems 4.2

Recall the definition (5.8) of the Hamming distance between $\omega, \sigma \in A^{[\![0,n]\!]}$. The $\bar{d}$-distance between two probability measures μ_n, ν_n on $A^{[\![0,n]\!]}$ is

$$
\bar{d}(\mu_n, \nu_n) = \inf \sum_{\sigma, \tilde{\sigma} \in A^{n+1}} \bar{d}_n(\sigma, \tilde{\sigma})\, \mathbb{P}_n(\sigma, \tilde{\sigma})
$$

where the infimum is taken over all couplings $\mathbb{P}_n$ of μ_n and ν_n.

Consider functions $f : A^{n+1} \to \mathbb{R}$ such that, for $j \in [\![0, n]\!]$, $\delta_j f \leq 1/(n+1)$. Such functions are 1- Lipschitz with respect to the Hamming distance because for all $\sigma, \eta \in A^{n+1}$

$$|f(\sigma) - f(\eta)| \leq \sum_{j=0}^{n} (\delta_j f)\, \mathbb{1}_{\{\sigma_j \neq \eta_j\}} \leq \frac{1}{n+1} \sum_{j=0}^{n} \mathbb{1}_{\{\sigma_j \neq \eta_j\}} = \bar{d}_n(\sigma, \eta).$$

Let g be a kernel and μ a compatible measure satisfying the conditions of the theorem. Then, Theorems 3.2 and 3.4 state that, for such functions, for all $\theta \in \mathbb{R}$ and $n \geq 0$,

$$\mathbb{E}_\mu \left[\mathrm{e}^{\theta(f - \mathbb{E}_\mu[f])} \right] \leq \exp\left(\frac{\mathcal{C}^{-2}(n+1)^{-1}\theta^2}{8} \right).$$

The main observation is that this is equivalent, according to [2, Theorem 3.1], to having

$$\bar{d}_n(\nu_n, \mu_n) \leq \frac{1}{\mathcal{C}} \sqrt{\frac{1}{2(n+1)}\, \mathbb{E}_{\nu_n} \left[\log \frac{\nu_n}{\mu_n} \right]}$$

where $\mu_n = \mu|_{\mathcal{F}_n}$ and ν_n is any probability measure on A^{n+1}. Consider now the measure ν compatible with h as given in the statement of the theorem and let ν_n be $\nu|_{\mathcal{F}_n}$. We have by stationarity [43] that

$$\bar{d}(\mu, \nu) = \lim_n \bar{d}_n(\mu_n, \nu_n).$$

So the proof of the proposition is concluded since we have that (recall the proof of Theorem 3.5 above)

$$\lim_n \frac{1}{n+1}\, \mathbb{E}_{\nu_n} \left[\log \frac{\nu_n}{\mu_n} \right] = \mathbb{E}_\nu \left[\log \frac{h}{g} \right].$$

Acknowledgements. SG thanks CNRS, FAPESP (19805/2014 and 2017/07084-6) as well as CNPq Universal (439422/2018-3) for financial support. DYT thanks École Polytechnique for a 2-month fellowship. SG and DYT thank the CPHT for its hospitality during several stays.

References

[1] N. Berger, C. Hoffman, and V. Sidoravicius. Non-uniqueness for specifications in $\ell^{2+\epsilon}$. *Ergodic Theory and Dynamical Systems*, 38(4):1342–1352, 2018.

[2] S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 1999.

[3] M. Bramson and S. Kalikow. Nonuniqueness in g-functions. *Israel Journal of Mathematics*, 84(1-2):153–160, 1993.

[4] X. Bressaud, R. Fernández, and A. Galves. Decay of correlations for non-Hölderian dynamics. A coupling approach. *Electron. J. Probab.*, 4:no. 3, 19 pp. (electronic), 1999.

[5] X. Bressaud, R. Fernández, and A. Galves. Speed of $\overline{d}$-convergence for Markov approximations of chains with complete connections. A coupling approach. *Stochastic Process. Appl.*, 83(1):127–138, 1999.

[6] J.-R. Chazottes, P. Collet, C. Külske, and F. Redig. Concentration inequalities for random fields via coupling. *Probab. Theory Related Fields*, 137(1-2):201–225, 2007.

[7] J.-R. Chazottes, P. Collet, and F. Redig. On concentration inequalities and their applications for Gibbs measures in lattice systems. *J. Stat. Phys.*, 169(3):504–546, 2017.

[8] Z. Coelho and A. Quas. Criteria for $\bar{d}$-continuity. *Transactions of the American Mathematical Society*, 350(8):3257–3268, 1998.

[9] P. Collet, D. Duarte, and A. Galves. Bootstrap central limit theorem for chains of infinite order via Markov approximations. *Markov Process. Related Fields*, 11(3):443–464, 2005.

[10] J. Dedecker and S. Gouëzel. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20, 2015.

[11] L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[12] J. Dias and S. Friedli. Uniqueness vs. non-uniqueness for complete connections with modified majority rules. *Probability Theory and Related Fields*, 164(3-4):893–929, 2016.

[13] R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[14] W. Doeblin and R. Fortet. Sur des chaînes à liaisons complètes. *Bulletin de la Société Mathématique de France*, 65:132–148, 1937.

[15] D. Duarte, A. Galves, and N. L. Garcia. Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc. (N.S.)*, 37(4):581–592, 2006.

[16] R. Fernández and A. Galves. Markov approximations of chains of infinite order. *Bull. Braz. Math. Soc. (N.S.)*, 33(3):295–306, 2002. Fifth Brazilian School in Probability (Ubatuba, 2001).

[17] R. Fernández and G. Maillard. Chains with complete connections and one-dimensional gibbs measures. *Electronic Journal of Probability*, 9:145–176, 2004.

[18] R. Fernández and G. Maillard. Chains with complete connections: general theory, uniqueness, loss of memory and mixing properties. *Journal of Statistical Physics*, 118(3-4):555–588, 2005.

[19] K. Fokianos and L. Truquet. On categorical time series models with covariates. *Stochastic Processes and their Applications*, 129(9):3446–3462, 2019.

[20] S. Friedli. A note on the Bramson–Kalikow process. *Brazilian Journal of Probability and Statistics*, 29(2):427–442, 2015.

[21] C. Gallesco, S. Gallo, and D. Y. Takahashi. Explicit estimates in the Bramson−Kalikow model. *Nonlinearity*, 27(9):2281–2296, 2014.

[22] C. Gallesco, S. Gallo, and D. Y. Takahashi. Dynamic uniqueness for stochastic chains with unbounded memory. *Stochastic Processes and their Applications*, 128(2):689–706, 2018.

[23] S. Gallo and N. L. Garcia. Perfect simulation for locally continuous chains of infinite order. *Stochastic Processes and their Applications*, 123(11):3877–3902, 2013.

[24] S. Gallo, M. Lerasle, and D. Y. Takahashi. Markov approximation of chains of infinite order in the $\bar{d}$-metric. *Markov Processes and Related Fields*, 19(1):51–82, 2013.

[25] S. Gallo and D. Y. Takahashi. Attractive regular stochastic chains: perfect simulation and phase transition. *Ergodic Theory and Dynamical Systems*, 34(5):1567–1586, 2014.

[26] T. E. Harris. On chains of infinite order. *Pacific Journal of Mathematics*, 5:707–724, 1955.

[27] P. Hulse. An example of non-unique g-measures. *Ergodic Theory and Dynamical Systems*, 26(2):439–445, 2006.

[28] A. Johansson and A. Öberg. Square summability of variations of g-functions and uniqueness of g-measures. *Mathematical Research Letters*, 10(5-6):587–601, 2003.

[29] S. Kalikow. Random Markov processes and uniform martingales. *Israel Journal of Mathematics*, 71(1):33–54, 1990.

[30] M. Keane. Strongly mixing g-measures. *Inventiones Mathematicae*, 16(4):309–324, 1972.

[31] B. Kedem and K. Fokianos. *Regression models for time series analysis*, volume 488. John Wiley & Sons, 2005.

[32] A. Kontorovich and R. Weiss. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51(4):1100–1113, 2014.

[33] L. A. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

[34] C. Külske. Concentration inequalities for functions of Gibbs fields with application to diffraction and random gibbs measures. *Communications in mathematical physics*, 239(1-2):29–51, 2003.

[35] F. Ledrappier. Principe variationnel et systèmes dynamiques symboliques. *Probability Theory and Related Fields*, 30(3):185–202, 1974.

[36] K. Marton. Measure concentration for a class of random processes. *Probability Theory and Related Fields*, 110(3):427–439, 1998.

[37] K. Marton and P. C. Shields. The positive-divergence and blowing-up properties. *Israel J. Math.*, 86(1-3):331–348, 1994.

[38] P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[39] C. McDiarmid. *On the method of bounded differences*, pages 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989.

[40] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[41] O. Onicescu and G. Mihoc. Sur les chaînes de variables statistiques. *Bulletin des Sciences Mathématiques*, 59(2):174–192, 1935.

[42] P.-M. Samson. Concentration of measure inequalities for markov chains and Φ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

[43] P. C. Shields. *The ergodic theory of discrete sample paths*, volume 13 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1996.

[44] L. Truquet. Coupling and perturbation techniques for categorical time series. *arXiv preprint arXiv:1907.13533*, 2019.

[45] P. Walters. Ruelle's operator theorem and g-measures. *Transactions of the American Mathematical Society*, 214:375–387, 1975.