Toy examples for effective concentration bounds

Benoît Kloeckner

Toy examples for effective concentration bounds

Benoît R. Kloeckner *

February 17, 2021

In this note we prove a spectral gap for various Markov chains on various functional spaces. While proving that a spectral gap exists is relatively common, explicit estimates seems somewhat rare.

These estimates are then used to apply the concentration inequalities of [Klo17] (most of the present material was part of Section 3 of that article, which has been reduced to its core in the published version).

Let us recall briefly the notation and concentration inequalities from [Klo17].

Let $(X_k)_{k\geq 0}$ be a Markov chain taking value in a general state space Ω with a unique stationary measure μ_0, and let $\varphi : \Omega \to \mathbb{R}$ be a function (the "observable"). We are interested in the speed of the convergence of the empirical average $\hat{\mu}_n(\varphi) := \frac{1}{n}\sum_{k=1}^{n} \varphi(X_k)$ to $\mu_0(\varphi)$. We denote by μ the law of X_0, which can be arbitrary.

Assumption 1. *The observable φ belongs to a function space $\mathscr{X}$ satisfying*

i. its norm $\|\cdot\|$ dominates the uniform norm: $\|\cdot\| \geq \|\cdot\|_\infty$,

ii. $\mathscr{X}$ is a Banach algebra, i.e. for all $f, g \in \mathscr{X}$ we have $\|fg\| \leq \|f\|\|g\|$,

iii. $\mathscr{X}$ contains the constant functions and $\|\mathbf{1}\| = 1$ (where $\mathbf{1}$ denotes the constant function with value 1).

To the transition kernel M is associated an averaging operator acting on $\mathscr{X}$:

$$\mathrm{L}_0 f(x) = \int_\Omega f(y)\, \mathrm{d}m_x(y).$$

Since each m_x is a probability measure, L_0 has 1 as eigenvalue, with eigenfunction $\mathbf{1}$.

*Université Paris-Est, Laboratoire d'Analyse et de Matématiques Appliquées (UMR 8050), UPEM, UPEC, CNRS, F-94010, Créteil, France

Assumption 2. *The Markov chain* M *satisfies the following:*

i. L_0 *acts as a bounded operator from* $\mathscr{X}$ *to itself, and its operator norm* $\|\mathrm{L}_0\|$ *is equal to* 1.

ii. L_0 *is contracting with gap* $\delta_0 > 0$*, i.e. there is a closed hyperplane* $G_0 \subset \mathscr{X}$ *such that*

$$\|\mathrm{L}_0 f\| \leq (1-\delta_0)\|f\| \qquad \forall f \in G_0.$$

The second hypothesis is a particular case of a *spectral gap*: it implies in particular that 1 is a simple isolated eigenvalue.

In [Klo17] the following two results where proved (plus a Berry-Esseen bound that we will not use here).

Theorem A. *Assuming assumptions 1 and 2, for all* $n \geq 1 + \frac{\log 100}{-\log(1-\delta_0/13)}$ *it holds:*

$$\mathbb{P}_\mu\Big[|\hat{\mu}_n(\varphi) - \mu_0(\varphi)| \geq a\Big] \leq \begin{cases} 2.488 \exp\Big(-n \dfrac{\delta_0}{13.44\delta_0 + 8.324} \dfrac{a^2}{\|\varphi\|^2}\Big) & \text{if } \dfrac{a}{\|\varphi\|} \leq \dfrac{\delta_0}{3} \\ 2.624 \exp\Big(-n \dfrac{0.98\delta_0^2}{12+13\delta_0}\Big(\dfrac{a}{\|\varphi\|} - 0.254\delta_0\Big)\Big) & \text{otherwise.} \end{cases}$$

(We will often use the strenghtened hypothesis $n \geq 60/\delta_0$ for simplicity.)

Theorem B. *Assuming assumptions 1 and 2, for all* $n \geq \frac{60}{\delta_0}$*, all* $U \geq \sigma^2(\varphi)$ *and all* $a \leq \frac{U}{\|\varphi\|} \log\Big(1 + \frac{\delta_0^2}{12+13\delta_0}\Big)$ *it holds:*

$$\mathbb{P}_\mu\Big[|\hat{\mu}_n(\varphi) - \mu_0(\varphi)| \geq a\Big] \leq 2.637 \exp\left(-n \cdot \left(\frac{a^2}{2U} - 10(1+\delta_0^{-1})^2 \frac{\|\varphi\|^3 a^3}{U^3}\right)\right).$$

Above, we use the notation $\sigma^2(\varphi) = \mu_0(\varphi^2) - (\mu_0\varphi)^2 + 2\sum_{k\geq 1} \mu_0(\varphi \mathrm{L}_0^k \bar{\varphi})$ where $\bar{\varphi} = \varphi - \mu_0(\varphi)$. This "dynamical variance" is precisely the variance appearing in the CLT.

While it is well known that the presence of a spectral gap ensures classical limit theorems, these results turn explicit contraction estimates into explicit non-asymptotic results. The main goal of this note is to compute lower bounds on δ_0 for several pairs of Markov chains and functional spaces. We shall apply the above result for illustration, and compare to previous results when available.

1 Preliminary lemma

In each example below we will use the following lemma which, in the spirit of Doeblin-Fortet and Lasota-Yorke inequalities, enables to turn an exponential contraction in the "regularity part" of a functional norm into a spectral gap.

Lemma 1.1. *Consider a normed space $\mathscr{X}$ of (Borel measurable, bounded) functions $\Omega \to \mathbb{R}$, with norm $\|\cdot\| = \|\cdot\|_\infty + V(\cdot)$ where V is a semi-norm (usually quantifying some regularity of the argument, such as* Lip *or* BV*).*

Assume that for some constant $C > 0$, for all probability μ on Ω and for all $f \in \mathscr{X}$ such that $\mu(f) = 0$, $\|f\|_\infty \leq CV(f)$.

Let $\mathrm{L}_0 \in \mathscr{B}(\mathscr{X})$ and assume that for some $\theta \in (0,1)$ and all $f \in \mathscr{X}$:

$$\|\mathrm{L}_0 f\|_\infty \leq \|f\|_\infty \quad \textit{and} \quad V(\mathrm{L}_0 f) \leq \theta V(f)$$

and having eigenvalue 1 with an eigenprobability μ_0, i.e. $\mathrm{L}_0^* \mu_0 = \mu_0$.

Then L_0 is contracting with gap at least

$$\delta_0 = \frac{1-\theta}{1+C\theta}.$$

The condition $\|f\|_\infty \leq CV(f)$ is often valid in practice (assuming Ω has finite diameter for spaces such as $\mathrm{Lip}(\Omega)$): the condition that $\mu(f) = 0$ implies that f vanishes (if functions in $\mathscr{X}$ are continuous) or at least takes both non-positive and non-negative values, and $V(f)$ usually bounds the variations of f, implying a bound on its uniform norm.

Proof. Let $f \in \ker \mu_0$; then $\|\mathrm{L}_0 f\|_\infty \leq \|f\|_\infty$ and $\mathrm{L}_0 f \in \ker \mu_0$, so that $\|\mathrm{L}_0 f\|_\infty \leq CV(\mathrm{L}_0 f) \leq C\theta V(f)$.

Denote by $t \in [0,1]$ the number such that $\|f\|_\infty = t\|f\|$ (and therefore $V(f) = (1-t)\|f\|$). The above two controls on $\|\mathrm{L}_0(f)\|_\infty$ can then be written as $\|\mathrm{L}_0(f)\|_\infty \leq \min\Big(t, C\theta(1-t)\Big)\|f\|$ and using $V(\mathrm{L}_0 f) \leq \theta V(f)$ again we get

$$\|\mathrm{L}_0(f)\| \leq \min\Big(t + \theta(1-t), (C+1)\theta(1-t)\Big)\|f\|$$
$$\|(\mathrm{L}_0)_{|\ker \mu_0}\| \leq \max_{t\in[0,1]} \min\Big(t + \theta(1-t), (C+1)\theta(1-t)\Big).$$

The maximum is reached when $t+\theta(1-t) = (C+1)\theta(1-t)$, i.e. when $t = C\theta/(1+C\theta)$, at which point the value in the minimum is $(C+1)\theta/(C\theta+1) \in (0,1)$. We get contraction with gap $1-(C+1)\theta/(C\theta+1)$, as claimed. □

2 Chains with Doeblin's minorization

We start with a warm-up in the simplest example of a Banach Algebra of functions, the space of measurable bounded functions $L^\infty(\Omega)$.[1] To fit our framework, we will need to endow $L^\infty(\Omega)$ with the norm $\|\cdot\|_S = \|f\|_\infty + S(f)$ where

$$S(f) := \sup_{x,y\in\Omega} |f(x) - f(y)| = \sup f - \inf f$$

[1] We do not have a single reference measure here, which is why we consider genuinely bounded functions rather than essentially bounded functions.

measures how "spread out" f is, which we need to manage separately from the magnitude of f. Of course, this norm is equivalent to the uniform norm, and it is easily checked what we still get a Banach Algebra.

Observe that convergence of measures in duality to $L^\infty(\Omega)$ is convergence in total variation, and the most usual normalization is

$$d_{\mathrm{TV}}(\mu,\nu) := \sup_{S(f)=1} \Big|\mu(f) - \nu(f)\Big|.$$

For a transition kernel M, having an averaging operator L_0 with a spectral gap is a very strong condition, called *uniform ergodicity*.

Glynn and Ormoneit [GO02] and Kontoyiannis, Lastras-Montaño and Meyn [KLMM05] gave explicit concentration results for such chains, using the characterization of uniform ergodicity by the *Doeblin minorization condition*: there exist an integer $\ell \geq 1$, a positive number β and a probability measure ω on Ω such that for all $x \in \Omega$ and all Borel set $B \subset \Omega$:

$$m_x^\ell(B) \geq \beta\omega(B) \tag{1}$$

where m_x^ℓ is the law of X_ℓ conditionally to $X_0 = x$.

We shall look at the case $\ell = 1$, which fits better in our context. For arbitrary value of ℓ, one can in practice apply the result to each extracted chain $(X_{k_0+k\ell})_{k\geq 0}$.

Proposition 2.1. *If* M *satisfies Doeblin's minorization condition* (1) *with* $\ell = 1$*, then its averaging operator* L_0 *is contracting on* $L^\infty(\Omega)$ *with gap* $\beta/(2-\beta)$*.*

Proof. This is simply the classical maximal coupling method in a functional guise. For each $x \in \Omega$ decompose m_x into $\beta\omega$ and $r_x := m_x - \beta\omega$ (which is a positive measure of mass $1-\beta$). Recall that we denote by μ_0 the stationary measure of M. For all $f \in L^\infty(\Omega)$ we have:

$$\begin{aligned}
\mathrm{L}_0 f(x) &= \beta\omega(f) + r_x(f) \\
\mathrm{L}_0 f(x) - \mathrm{L}_0 f(y) &= \int (r_x(f) - r_y(f))\,\mathrm{d}\mu_0(y) \\
\Big|\mathrm{L}_0 f(x) - \mathrm{L}_0 f(y)\Big| &\leq \int (1-\beta) S(f)\,\mathrm{d}\mu_0(y) \\
S(\mathrm{L}_0 f) &\leq (1-\beta) S(f).
\end{aligned}$$

We can thus apply Lemma 1.1 with $C = 1$ and $\theta = 1-\beta$, obtaining a spectral gap of size $\beta/(2-\beta)$. □

Corollary 2.2. *If* M *satisfies Doeblin's minorization condition* (1) *with* $\ell = 1$ *and* $\varphi : \Omega \to [-1,1]$*, for all* $n \geq 120/\beta$ *and all* $a \leq \beta/2$ *it holds*

$$\mathbb{P}_\mu\Big[|\hat{\mu}_n(\varphi) - \mu_0(\varphi)| \geq a\Big] \leq 2.5 \exp\Big(-na^2 \cdot \frac{\beta}{150 + 47\beta}\Big).$$

Proof. We have here $\|\varphi\|_S \leq 3$ and, by Lemma 2.1, $\delta_0 \geq \beta/(2-\beta) \geq \beta/2$. It then suffices to apply Theorem A and round constants up. □

The exponent is proportional to β, which is the correct rate and improves on [GO02] and [KLMM05] which get a β^2; but Paulin obtains better constants in this case [Pau15] (Corollary 2.10), and we do not study this example further.

3 Discrete hypercube

Let us consider the same toy example as Joulin and Ollivier [JO10], the lazy random walk (aka Gibbs sampler, aka Glauber dynamics) on the discrete hypercube $\{0,1\}^N$: the transition kernel M chooses randomly uniformly a slot $i \in \{1,\dots,N\}$ and replaces it with the result of a fair coin toss, i.e.

$$m_x = \frac{1}{2}\delta_x + \sum_{y\sim x} \frac{1}{2N}\delta_y.$$

We consider two kind of observables: the "polarization"$\rho : \{0,1\}^N \to \mathbb{R}$ giving the proportion of 1's in its argument, and the characteristic function $\mathbf{1}_S$ of a subset $S \subset \{0,1\}^N$. In this second example, we will in particular consider the simple case $S = [0] := \{(0, x_2, \dots, x_N) \colon x_i \in \{0,1\}\}$.

3.1 Spectral gap estimates

The discrete hypercube $\{0,1\}^N$ is endowed with the Hamming metric: if $x = (x_1,\dots,x_N)$ and $y = (y_1,\dots,y_N)$, then $d(x,y)$ is the number of indexes i such that $x_i \neq y_i$. Two elements at distance 1 are said to be adjacent, denoted by $x \sim y$.

We denote by E the set of tuples $\epsilon = (\epsilon_i)_{1\leq i\leq N}$ such that exactly one of the ϵ_i is 1. Identifying $\{0,1\}$ with $\mathbb{Z}/2\mathbb{Z}$, an edge thus writes $(x, x+\epsilon)$ for some $x \in \{0,1\}^N$ and some $\epsilon \in E$.

We shall consider several function spaces to showcase the flexibility of the spectral method; since the space $\{0,1\}^N$ is finite, we always consider the space of all functions $\{0,1\}^N \to \mathbb{R}$, and it is the considered norm which will matter. Let us define:

- $\|f\|_L = \|f\|_\infty + \mathrm{Lip}(f)$: this is the standard Lipschitz norm;
- $\|f\|_{dL} = \|f\|_\infty + N\,\mathrm{Lip}(f)$: this is the Lipschitz norm with a weigth to the regularity part equal to the diameter;
- $\|f\|_W = \|f\|_\infty + W(f)$ where

  $$W(f) = \sup_{x\in\{0,1\}^N} \sum_{\epsilon\in E} |f(x+\epsilon) - f(x)|;$$

  this norm stays small for functions having large variations only in few directions (small "local total variation").

We shall use later the following non-trivial comparison with $\|\cdot\|_S$.

Lemma 3.1 (Fedor Petrov [Pet17])**.** *For all $f : \{0,1\}^N \to \mathbb{R}$ we have*

$$\max f - \min f \leq W(f).$$

Proof. Without lost of generality, we can assume $W(f) \leq 1$ and $f(0,0,\dots,0) = 0$, and reduce to proving $f(1,1,\dots,1) \leq 1$.

Define the *cost* of a path $x^0, x^2, \dots, x^k$ as the number $\sum_{i=0}^{k-1} |f(x^{i+1}) - f(x^i)|$, and let Σ be the sum of the costs of all paths of length N from $(0,0,\dots,0)$ to $(1,1,\dots,1)$. We shall prove that $\Sigma \leq N!$, and since there are $N!$ such paths one of them will have cost at most 1, proving the lemma.

We call "level" of $x \in \{0,1\}^N$ the number of 1s among the coordinates of x, and denote it by $|x|$. For each $i \in \{0,1,\dots,N-1\}$, define $p_i = \frac{i!(N-i)!}{N+1}$. Then all p_i are positive and $p_i + p_{i+1} = i!(N-i-1)!$ is precisely the number of paths that use any given edge from level i to level $i+1$.

The contribution to Σ of an edge $(x, x+\epsilon)$ from level i to level $i+1$ is thus $i!(N-i-1)!|f(x+\epsilon) - f(x)|$, which we split into two parts, one $p_i|f(x+\epsilon) - f(x)|$ attributed to x and the other $p_{i+1}|f(x+\epsilon) - f(x)|$ to $x+\epsilon$. It follows

$$\Sigma \leq \sum_{x \in \{0,1\}^N} p_{|x|} W(f) \leq \sum_{i=0}^{N} p_i \binom{N}{i} = \sum_{i=0}^{N-1} (p_i + p_{i+1}) \binom{N-1}{i} = N(N-1)! = N!$$

as desired. □

We get the following gap estimates.

Theorem 3.2. *Each of the norm $\|\cdot\|_L$, $\|\cdot\|_{dL}$ and $\|\cdot\|_W$ turns the space of all functions $\{0,1\}^N \to \mathbb{R}$ into a Banach algebra where $\mathbf{1}$ has norm 1.*

Moreover the averaging operator L_0 of the transition kernel M has operator norm 1, and is contracting with gap respectively $1/N^2$, $1/(2N-1)$ and $1/(4N-1)$ in the norms $\|\cdot\|_L$, $\|\cdot\|_{dL}$ and $\|\cdot\|_W$.

Proof. Each norm considered here has the form $\|\cdot\| = \|\cdot\|_\infty + V(\cdot)$ for some semi-norm V such that $V(fg) \leq \|f\|_\infty V(g) + V(f)\|f\|_\infty$; it follows that the considered spaces are Banach algebras. All the other properties but the contraction are trivial.

To prove the contraction, we simply apply Lemma 1.1. First, it is well-known that for all $\varphi : \{0,1\}^N \to \mathbb{R}$,

$$\mathrm{Lip}(\mathrm{L}_0\varphi) \leq (1 - 1/N)\,\mathrm{Lip}(\varphi)$$

(in the parlance of [Oll09], M is positively curved with $\kappa = 1/N$).

In the case of $\|\cdot\|_L$, we get $\theta = 1 - 1/N$ and $C = N$ (since a function of vanishing average must take positive and negative values, and $\mathrm{diam}\{0,1\}^N = N$), hence a contraction with gap $1/N^2$. In the case of $\|\cdot\|_{dL}$, the normalizing factor gives $C = 1$ (and we still have $\theta = 1 - 1/N$), hence a spectral gap of size $1/(2N-1)$.

To deal with $\|\cdot\|_W$, we first show that in Lemma 1.1 we can take $\theta = 1 - 1/(2N)$.

$$\begin{aligned} W(\mathrm{L}_0\varphi) &= \sup_x \sum_{\epsilon \in E} \left| \frac{1}{2}\varphi(x+\epsilon) + \frac{1}{2N} \sum_{\eta \in E} \varphi(x+\eta+\epsilon) - \frac{1}{2}\varphi(x) - \frac{1}{2N} \sum_{\eta \in E} \varphi(x+\eta) \right| \\ &= \sup_x \sum_{\epsilon \in E} \left| \left(\frac{1}{2} - \frac{1}{2N}\right)\varphi(x+\epsilon) + \frac{1}{2N} \sum_{\eta \neq \epsilon} \varphi(x+\eta+\epsilon) \right. \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad - \Big(\frac{1}{2} - \frac{1}{2N}\Big)\varphi(x) - \frac{1}{2N}\sum_{\eta\neq\epsilon}\varphi(x+\eta)\Big| \\
&\leq \sup_x \frac{N-1}{2N}\sum_{\epsilon\in E}|\varphi(x+\epsilon)-\varphi(x)| + \frac{1}{2N}\sum_{\epsilon\in E}\sum_{\eta\neq\epsilon}|\varphi(x+\epsilon+\eta)-\varphi(x+\eta)| \\
&\leq \frac{N-1}{2N}W(\varphi) + \frac{1}{2N}\sup_x \sum_{y\sim x}\sum_{\epsilon\in E}|\varphi(y+\epsilon)-\varphi(y)|.
\end{aligned}
$$

Hence we obtain $W(\mathrm{L}_0\varphi) \leq \big(1-\frac{1}{2N}\big)W(\varphi)$.

Then Lemma 3.1 shows that we can take $C = 1$, providing a spectral gap of size $1/(4N-1)$. □

3.2 Concentration inequalities

Let us combine 3.2 with A and B to obtain explicit concentration estimates. We will not compute the explicit constants, and concentrate on the dependency with the parameters a and N.

Consider first the "polarization" observable $\rho : \{0,1\}^N \to \mathbb{R}$, where $\rho(x)$ is the proportion of 1's in the word x. We have

$$
\|\rho\|_L = 1 + \frac{1}{N}, \qquad \|\rho\|_{dL} = 2, \qquad \|\rho\|_W = 2.
$$

To use Theorem A with optimal efficiency, assuming a will be small enough, we need to maximize $\delta_0/\|\rho\|^2$. Here, we shall thus use the norm $\|\cdot\|_{dL}$. For $a \lesssim N$, Theorem A shows that we need at most $O(N/a^2)$ iterations to have a good convergence to the actual mean; meanwhile Joulin and Ollivier only need $O(1/a^2)$, but for concentration around the expectancy of the empiric process, not around the expectancy with respect to the stationary measure. Without burn-in, one also needs to bound the bias, which approaches zero in time $O(N/a)$ according to the bound of Joulin and Ollivier, for a total run time of $O(N/a + 1/a^2)$. With burn-in, they need a run time of $O(N + 1/a^2)$.

For $1/N \lesssim a \lesssim 1$, we enter our exponential regime while staying inside Joulin-Ollivier's Gaussian window; Theorem A shows we need no more than $O(N^2/a)$ iterations, while [JO10] still gives a bound of $O(N + 1/a^2)$.

In this example, Joulin and Ollivier get a sharper result; this seems to be explained in one part by the fact that we do not get to decouple the bias from the convergence of expectancies, and in another part by our need to have a Banach algebra, hence to include the uniform norm in our norm.

Consider now the potential $\mathbf{1}_S$, the indicator function for a (non-trivial) set S. This function is only 1-Lipschitz, so that we have $\|\mathbf{1}_S\|_L = 2$ and $\|\mathbf{1}_S\|_{dL} = 1 + N$. If we insist on using a Lipschitz norm, the unormalized one is thus better and with $\delta_0 = 1/N^2$ Theorem A shows that we need (in the Gaussian regime) $O(N^2/a^2)$ iterations to ensure the error is probably less than a, which is the same order of magnitude than given by [JO10] with a worse constant, ~ 34 instead of 8. But here we have two ways to improve on this bound.

The first one is to use Theorem B. When $S = [0] := \{0x_2x_3\cdots x_N \in \{0,1\}^N\}$, the dynamical variance can be computed explicitly (distinguish the cases when the first digit has been changed an odd or even number of times, and observe that at each step the probability of changing the first digit is $1/2N$):

$$\mu_0(\mathbf{1}_{[0]}^2) - (\mu_0\mathbf{1}_{[0]})^2 = 1/4 \quad \text{and} \quad \sum_{k\geq 1}\mu_0(\mathbf{1}_{[0]}\mathrm{L}_0^k\bar{\mathbf{1}}_{[0]}) = \frac{1}{4}\sum_{k\geq 1}\left(\frac{N-1}{N}\right)^k = \frac{N-1}{4}.$$

This gives $\sigma^2(\mathbf{1}_S) \simeq N/2$. Switching back to the norm $\|\cdot\|_{dL}$, when $a \lesssim 1/N^2$ and $n \geq 60N^2$, in Theorem B the positive term in the exponential is negligible compared to the main term which is $-na^2/N$. In particular $O(N/a^2)$ iterations suffice to get a small probability for a deviation at least a: compared to Joulin and Ollivier, we gain one power of N in this regime (and the optimal constant 1 in the leading term of the rate) but only for very small values of a.[2] This choice of S might seem very specific, but for less regular S the gain should be greater for sufficiently smaller a. For example, if S contains half the vertices and every vertex $x \in \{0,1\}^N$ has exactly $2Np$ neighbors with the same $\mathbf{1}_S$ value, the above computation of variance gives $\sigma^2(\mathbf{1}_S) = \frac{1}{4} + \frac{1-2p}{4p}$. We shall call a family of sets $S_N \in \{0,1\}^N$ "scrambled" when the indicator functions $\mathbf{1}_{S_N}$ have bounded variance (independently of N) with respect to the lazy random walk; by abuse, we shall speak of a scrambled set for a member of such a family. For scrambled sets taking $n = O(1/a^2)$ is sufficient: there is no dependency on the dimension. A further study of scrambled sets seems an interesting direction of work.

The second way to improve our first estimate is to use the norm $\|\cdot\|_W$ in Theorem A. Then $\|\mathbf{1}_{[0]}\|_W = 2$ and $\delta_0 \simeq 1/N$. For $a \lesssim 1/N$, Theorem A ensures that we need only $O(N/a^2)$ iterations to have a good convergence to the actual mean, which is again the optimal order of magnitude (since it corresponds to the CLT) but obtained on a much larger window than with Theorem B. This extends to all observables with $W(\varphi) \lesssim 1$; observe that this domain of applicability is quite complementary to the domain of applicability of the previous paragraph.

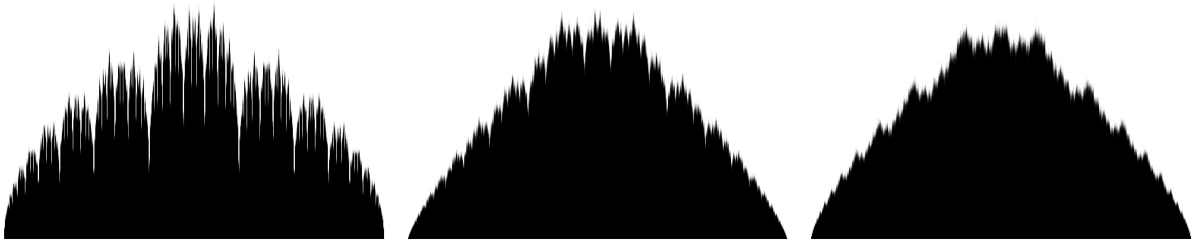
4 Bernoulli convolutions and observables of bounded variation

We now consider the "Bernoulli convolution" of parameter $\lambda \in (0,1)$, defined as the law β_λ of the random variable

$$\sum_{k\geq 1}\epsilon_k\lambda^k$$

where the ϵ_k are independent variables taking the value 1 with probability $1/2$ and the value -1 with probability $1/2$ (see [Klo17] for a brief account, and [PSS00] for more information on these measures, which are the object of intense scrutiny for decades).

[2] If we want to consider a of the order of $1/N$, we can then take $U \simeq N^2$ to enlarge the window, at the cost of a weaker leading term. We get a bound similar to the one of Joulin-Ollivier, possibly with a smaller constant (depending on the value of a).



(a) $\lambda = \frac{\sqrt{5}-1}{2} \simeq 0.618$ (b) $\lambda = \frac{e}{4} \simeq 0.680$ (c) $\lambda = \frac{2}{3} \simeq 0.667$

Figure 1: Histogram of the empirical distribution of the Markov chain associated to (T_0, T_1), with $X_0 = 0$, binned in 500 subintervals (averaged image over 30 independent runs of 10^6 points each). Parameter λ is the inverse of a Pisot number on the left, a very well approximable irrational at the center, rational on the right.

One can realize naturally β_λ as the stationary law of the Markov transition kernel $\mathsf{M} = (m_x)_{x\in\mathbb{R}}$ defined by

$$m_x = \frac{1}{2}\delta_{T_0(x)} + \frac{1}{2}\delta_{T_1(x)}$$

where $T_0(x) = \lambda x - \lambda$ and $T_1(x) = \lambda x + \lambda$ (this is a particular case of an Iterated Function System).

In order to evaluate $\beta_\lambda(\varphi)$ by a MCMC method, one cannot use the methods developed for ergodic Markov chains since, conditionally to $X_0 = x$, the law m_x^k of X_k is atomic and thus singular with respect to β_λ: $d_{\mathrm{TV}}(m_x^k, \beta_\lambda) = 1$ for all k. The convergence only holds for observables satisfying some regularity assumption, and it is natural to ask what regularity is needed.

For a Lipschitz observable φ one only need to observe that M has positive curvature in the sense of Ollivier (this is easy using the coupling $\frac{1}{2}\delta_{(T_0(x),T_0(y))} + \frac{1}{2}\delta_{(T_1(x),T_1(y))}$ of m_x and m_y) and apply [JO10]. But what if φ is not Lipschitz (or has large Lipschitz constant)? We shall consider observables of bounded variation, a regularity which has the great advantage over Lipschitz to include the characteristic functions of intervals.

Definition 4.1. Given an interval $I \subset \mathbb{R}$, we consider the Banach space $\mathrm{BV}(I)$ of *bounded variation* functions $I \to \mathbb{R}$, defined by the norm $\|\cdot\|_{\mathrm{BV}} = \|\cdot\|_\infty + \mathrm{var}(\cdot, I)$ where

$$\mathrm{var}(f, I) := \sup_{x_0<x_1<\cdots<x_p\in I} \sum_{j=1}^{p} |f(x_j) - f(x_{j-1})|$$

(the uniform norm is usually replaced by the L^1 norm, but when I is bounded our choice is equivalent up to a constant, it does not single out the Lebesgue measure, and most importantly it ensures that $\mathrm{BV}(I)$ is a Banach algebra).

Important features of total variation are:

- its extensiveness: $\operatorname{var}(f, I) \geq \operatorname{var}(f, J) + \operatorname{var}(f, K)$ whenever J, K are disjoint subintervals of I,
- its invariance under monotonic maps: $\operatorname{var}(f \circ T, I) = \operatorname{var}(f, T(I))$ whenever T is monotonic.

It turns out that the averaging operator L_0 of the transition kernel M has a spectral gap for all λ, but is not a contraction when $\lambda > 1/2$ (i.e. we have an inequality $\|\mathrm{L}_0^n(f)\| \leq C(1-\delta_0)^n\|f\|$ on a closed hyperplane, but with $C > 1$). In yet other words, an iterate of L_0 is a contraction, and to apply directly Theorems A and B we need to consider an extracted Markov chain $(X_{\ell k})_{k\geq 0}$ for some ℓ.

Let I_λ be the attractor of the IFS (T_0, T_1), i.e. the interval whose endpoints are the fixed points of T_0 and T_1:

$$I_\lambda = \Big[\frac{-\lambda}{1-\lambda}, \frac{\lambda}{1-\lambda}\Big].$$

Given a word $\omega = \omega_1\omega_2\ldots\omega_k$ in the letters 0 and 1, we define

$$T_\omega = T_{\omega_1} \circ T_{\omega_2} \circ \cdots \circ T_{\omega_k} : I_\lambda \to I_\lambda.$$

Theorem 4.2. *If $\lambda^\ell < \frac{1}{2}$, then L_0^ℓ has a spectral gap on $\mathrm{BV}(I_\lambda)$ of size $1/(2^{\ell+1}-1)$ and constant 1.*

Proof. Let I_λ^-, I_λ^+ be the left and right halves of I_λ, i.e. $I_\lambda^- = \big[\frac{-\lambda}{1-\lambda}, 0\big)$ and $I_\lambda^+ = \big(0, \frac{\lambda}{1-\lambda}\big]$.

Let $f \in \mathrm{BV}(I_\lambda)$ and observe that the condition $\lambda^\ell < \frac{1}{2}$ ensures that $T_{00\ldots0}(I_\lambda)$ and $T_{11\ldots1}(I_\lambda)$ are disjoint (they have length $< \frac{1}{2}|I_\lambda|$ and each contains an endpoint of I_λ). Then:

$$\begin{aligned}
\operatorname{var}(\mathrm{L}_0^\ell f, I_\lambda) &\leq \frac{1}{2^\ell} \sum_{\omega\in\{0,1\}^\ell} \operatorname{var}(f\circ T_\omega, I_\lambda) \leq \frac{1}{2^\ell} \sum_{\omega\in\{0,1\}^\ell} \operatorname{var}(f, T_\omega(I_\lambda)) \\
&\leq \frac{1}{2^\ell}\Big(\operatorname{var}(f, T_{00\ldots0}(I_\lambda)) + \operatorname{var}(f, T_{11\ldots1}(I_\lambda) + \sum_{\substack{\omega\neq 00\ldots0\\ \neq 11\ldots1}} \operatorname{var}(f, I_\lambda)\Big) \\
&\leq \frac{1}{2^\ell}\Big(\operatorname{var}(f, I_\lambda) + (2^\ell - 2)\operatorname{var}(f, I_\lambda)\Big) \\
\operatorname{var}(\mathrm{L}_0^\ell f, I_\lambda) &\leq (1 - 2^{-\ell})\operatorname{var}(f, I_\lambda).
\end{aligned}$$

Applying Lemma 1.1 with $C = 1$ and $\theta = 1 - 2^{-\ell}$ yields the claim. □

This enables us to apply our result to estimate $\beta_\lambda(\varphi)$ for any φ of bounded variation. For example, Theorem A yields the following.

Corollary 4.3. *Let $\lambda \in (\frac{1}{2}, 1)$ and let ℓ be an integer such that $\lambda^\ell < \frac{1}{2}$. Consider a Markov chain $(X_k)_{k\geq 0}$ with transition probability $2^{-\ell}$ from $x \in I_\lambda$ to $T_\omega(x)$, for each $\omega \in \{0,1\}^\ell$. For any starting distribution $X_0 \sim \mu$, any $\varphi \in \mathrm{BV}(I_\lambda)$, any positive $a < \|\varphi\|_{\mathrm{BV}}/3(2^{\ell+1}-1)$ and any $n \geq 120\cdot 2^\ell$ we have*

$$\mathbb{P}_\mu\Big[|\hat\mu_n(\varphi) - \mu_0(\varphi)| \geq a\Big] \leq 2.488\exp\Big(-\frac{na^2}{\|\varphi\|_{\mathrm{BV}}^2(16.65\cdot 2^\ell + 5.12)}\Big).$$

To the best of our knowledge, this example could not be handled effectively by previously known results. For example [GD12] needs the observable to be at least C^2 to have explicit estimates, and they do not give a concentration inequality.

References

[GD12] David M Gómez and Pablo Dartnell, *Simple monte carlo integration with respect to Bernoulli convolutions*, Applications of Mathematics **57** (2012), no. 6, 617–626. 4

[GO02] Peter W Glynn and Dirk Ormoneit, *Hoeffding's inequality for uniformly ergodic Markov chains*, Statistics & probability letters **56** (2002), no. 2, 143–146. 2, 2

[JO10] Aldéric Joulin and Yann Ollivier, *Curvature, concentration and error estimates for Markov chain Monte Carlo*, Ann. Probab. **38** (2010), no. 6, 2418–2442. MR 2683634 3, 3.2, 4

[KLMM05] Ioannis Kontoyiannis, Luis A Lastras-Montano, and Sean P Meyn, *Relative entropy and exponential deviation bounds for general Markov chains*, International Symposium on Information Theory, 2005, IEEE, 2005, pp. 1563–1567. 2, 2

[Klo17] Benoît R. Kloeckner, *Effective limit theorems for Markov chains with a spectral gap*, arXiv:1703.09623, 2017. (document), 4

[Oll09] Yann Ollivier, *Ricci curvature of Markov chains on metric spaces*, J. Funct. Anal. **256** (2009), no. 3, 810–864. MR 2484937 3.1

[Pau15] Daniel Paulin, *Concentration inequalities for Markov chains by Marton couplings and spectral methods*, Electronic Journal of Probability **20** (2015). 2

[Pet17] Fedor Petrov, *Answer to "diameter of a weighted Hamming cube"*, MathOverflow, 2017, https://mathoverflow.net/a/286346/4961. 3.1

[PSS00] Yuval Peres, Wilhelm Schlag, and Boris Solomyak, *Sixty years of Bernoulli convolutions*, Progress in probability (2000), 39–68. 4