Bernoulli **21**(4), 2015, 2073–2092
DOI: 10.3150/14-BEJ635

Some remarks on MCMC estimation of spectra of integral operators

RADOSŁAW ADAMCZAK* and WITOLD BEDNORZ**

Institute of Mathematics, University of Warsaw, ul. Banacha 2, 02-097 Warszawa, Poland.
E-mail: *R.Adamczak@mimuw.edu.pl; **W.Bednorz@mimuw.edu.pl

We prove a law of large numbers for empirical approximations of the spectrum of a kernel integral operator by the spectrum of random matrices based on a sample drawn from a Markov chain, which complements the results by V. Koltchinskii and E. Giné for i.i.d. sequences. In a special case of Mercer's kernels and geometrically ergodic chains, we also provide exponential inequalities, quantifying the speed of convergence.

Keywords: approximation of spectra; kernel operators; MCMC algorithms; random matrices

1. Introduction

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space. Consider a probability measure π on $(\mathcal{X}, \mathcal{F})$ and a symmetric measurable kernel $h: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, square integrable with respect to $\pi \otimes \pi$. With h one can associate the kernel linear operator defined by the formula

$$\mathbf{H} f(x) = \int_{\mathcal{X}} h(x, y) f(y) \pi(\mathrm{d}y). \tag{1}$$

This is a Hilbert–Schmidt self-adjoint operator on $L_2(\pi)$ and as such it possesses a real spectrum consisting of a square summable sequence of eigenvalues. In [14], Koltchinskii and Giné investigated the problem of approximating the spectrum of $\mathbf{H}$ by the spectra of certain finite dimensional random operators constructed with the help of the function h and a sequence of i.i.d. random variables $(X_n)_{n\ge 0}$, distributed according to π. More precisely, they define a sequence of random matrices

$$\tilde{\mathbf{H}}_n = \frac{1}{n}\bigl(h(X_i, X_j)\bigr)_{0\le i,j\le n-1} \tag{2}$$

and

$$\mathbf{H}_n = \frac{1}{n}\bigl((1-\delta_{ij})h(X_i, X_j)\bigr)_{0\le i,j\le n-1} = \tilde{\mathbf{H}}_n - \frac{1}{n}\operatorname{diag}\bigl(\bigl(h(X_i, X_i)\bigr)_{i=0}^{n-1}\bigr) \tag{3}$$

(above δ_{ij} is the Kronecker's symbol) and show that with probability one the spectrum of $\mathbf{H}_n$ (completed to an infinite sequence with zeros) converges in a certain metric to that of $\mathbf{H}$. They also show by simple examples that in general one cannot replace $\mathbf{H}_n$ with $\tilde{\mathbf{H}}_n$. Moreover, under some stronger assumptions, they provide rates of convergence as well as infinite-dimensional limit theorems.

Besides intrinsic mathematical interest, the original motivation in [14] came from the limiting theory of U-statistics. A U-statistic of degree 2, based on a kernel h and a sequence $\mathbf{X} = (X_n)_{n\geq 0}$ is a random variable of the form

$$U_n(h) = U_n(h, \mathbf{X}) = \frac{1}{n(n-1)} \sum_{0\leq i\neq j\leq n-1} h(X_i, X_j). \tag{4}$$

It is well known, that under certain assumptions and proper normalization, the law of $U_n(h)$ converges to a random variable of the form $\sum_i \lambda_i(g_i^2 - 1)$, where g_i's are i.i.d. standard Gaussian variables and λ_i's are the eigenvalues of $\mathbf{H}$. Thus, the approximate knowledge of the spectrum of $\mathbf{H}$ allows for approximate sampling from the limiting spectral distribution of corresponding U-statistics. Since the publication of [14], empirical approximations of spectra found further applications, for example, in machine learning, especially in the theory of spectral clustering on manifolds and in the Kernel Principal Component Analysis (see, e.g., [23–25, 28]).

Although the authors of [14] do not develop specific applications, their results can be interpreted as a Monte Carlo method for approximating the spectrum of a kernel operator. However, such an approach would require access to an i.i.d. sample from the distribution π, whereas for many situations of interest the density of the underlying probability measure is known only up to constants. In such situations, random samples approximating π can be often obtained via Markov Chain Monte Carlo (MCMC) methods, which rely on simulating a Markov chain with a simple transition function and invariant measure π. By the ergodic theorem, after sufficiently many steps the value of the chain will be distributed approximately as π. There are two popular ways of using such samples with estimators. One of them is to generate sufficiently many independent samples and to plug them in the estimator. Another one is to use the estimator directly on the dependent sample coming from the Markov chain. While the former approach requires analysis of the stability of the estimated quantity with respect to a small perturbation of the probability measure, the latter one requires laws of large numbers in the dependent setting, which would justify using the estimator directly on the Markov chain.

The objective of this paper is to provide such a law of large numbers, together with some probability bounds for the problem of approximation of the spectrum of an integral operator. Our motivation is manifold. First, we believe that extending the results of Koltchinskii and Giné to a dependent setting is an interesting probabilistic problem in its own right. At the same time, it indicates a possibility of having practical MCMC methods of estimating spectra. Of course, a practical implementation of this approach would require overcoming additional obstacles related, for example, to numerical inaccuracy; however, the law of large numbers and probabilistic bounds provide its theoretical justification. Additionally, our results suggest that it should be possible to justify the validity of at least some of the aforementioned machine learning methods in a dependent case, which may more accurately model real-life situations.

As a tool, we also develop a law of large numbers for U-statistics of Markov chains started from an arbitrary initial distribution, which complements results from [1,4,5,8].

The organization of the paper is as follows. First, in Section 2 we formulate our results, next in Section 3 we present basic notation and preliminary facts concerning Markov chains (in particular the regeneration method) as well as tools from linear algebra which will be used in the proofs. In Section 4, we prove the law of large numbers for U-statistics, and in Sections 5 and 6 we provide the proofs of our main results. Finally, in the last section we discuss the optimality of our assumptions.

2. Main results

We will work with a measurable space $(\mathcal{X}, \mathcal{F})$, where $\mathcal{F}$ is a countably generated σ-field. Let $\mathbf{X} = (X_n)_{n\geq 0}$ be a Harris ergodic Markov chain with transition function $P : \mathcal{X} \times \mathcal{F} \to [0, 1]$ and let π be its unique invariant probability measure (we refer to [17,21] for the general theory of Markov chains on not necessarily countable spaces). We will consider a symmetric measurable kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and the corresponding kernel type operator $\mathbf{H}$ given by (1). Let $\tilde{\mathbf{H}}_n$ and $\mathbf{H}_n$ be random matrices given by (2) and (3) respectively.

Since the infinite-dimensional operators we will consider will always be Hilbert–Schmidt, their spectra may be identified with an infinite sequence $\lambda = (\lambda_n)_{n\geq 0} \in \ell_2$, where ℓ_2 is the Hilbert space of all square summable sequences. There is clearly some ambiguity here related to the ordering of eigenvalues, but thanks to the choice of the metric we are about to make, it will not pose a problem in the sequel, so we may disregard it.

Since we want to approximate the spectrum of $\mathbf{H}$ by a spectrum of a finite-dimensional operator, just as in [14] we will always identify the finite spectrum of the latter with an element of ℓ_2, by appending to it an infinite sequence of zeros. We will denote the spectrum of an operator or a matrix K, by $\lambda(K)$.

The metric we will use to compare spectra will be the δ_2 metric defined as

$$\delta_2(x, y) = \inf_{\sigma\in\mathcal{P}} \left(\sum_{i=0}^{\infty} (x_i - y_{\sigma(i)})^2 \right)^{1/2},$$

where $\mathcal{P}$ is the set of all permutations of natural numbers. It is easy to see that δ_2 is a pseudometric on ℓ_2.

In what follows, we will always use the notation $\mu f = \int f \, \mathrm{d}\mu$ for a measure μ and a function f.

Our first result is the following.

Theorem 2.1. *Let $\mathbf{X} = (X_n)_{n\geq 0}$ be a Harris ergodic Markov chain on $(\mathcal{X}, \mathcal{F})$ with invariant probability measure π and let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric measurable function. Assume that there exists $F : \mathcal{X} \to \mathbb{R}$, such that $\pi F^2 < \infty$ and $|h(x, y)| \leq F(x)F(y)$ for all $x, y \in \mathcal{X}$.*

Let $\mathbf{H}: L^2(\pi) \to L^2(\pi)$ *be the linear operator given by* (1) *and* $\tilde{\mathbf{H}}_n$, $\mathbf{H}_n$ *be defined by* (2), (3), *respectively. Then for every initial measure* μ *of the chain* $\mathbf{X}$, *with probability one*,

$$\delta_2\big(\lambda(\tilde{\mathbf{H}}_n), \lambda(\mathbf{H})\big), \delta_2\big(\lambda(\mathbf{H}_n), \lambda(\mathbf{H})\big) \to 0.$$

Let us now briefly comment on the hypotheses of the above theorem. Our main assumption is the majorization of the form $|h(x, y)| \le F(x)F(y)$ for some $F : \mathcal{X} \to \mathbb{R}$ with $\pi F^2 < \infty$. There are two main reasons for considering this type of assumptions. The first one is technical. As shown in [1], the law of large numbers for U-statistics (which we will use in the proofs) of mixing sequences may fail if one assumes just integrability of the kernel, which intuitively is related to the fact that the behaviour of the random variable $h(X_i, X_{i+1})$ may depend on the behaviour of h on $\pi^{\otimes 2}$-negligible sets (since X_i, X_{i+1} are dependent). As we will see in Section 7, in our setting a similar phenomenon occurs, in particular the law of large numbers for the spectra may fail if one assumes only that $\pi^{\otimes 2}h^2 < \infty$. The second reason is the fact that in the theory of Markov chains, one often proves ergodicity by means of drift conditions and pointwise assumptions related to the drift functions $V : \mathcal{X} \to [0, \infty)$ (see, e.g., [7,11,12,17]). The drift conditions are expressed only in terms of the drift function and the transition function P. While it is not always easy to check integrability of a general function with respect to the stationary measure, the drift criteria provide certain integrability for the drift function. Thus, one can often construct the majorant F in terms of the function V.

Let us also stress that we require that the inequality between h and F hold pointwise and not just $\pi^{\otimes 2}$ a.s. Again, the reason is related to the dependencies between the variables X_i. From the point of the MCMC applications, it is crucial to allow the Markov chain to start from arbitrary initial conditions and the distribution of the chain approaches the stationary measure only in the limit. As a consequence, it is not enough to assume a $\pi^{\otimes 2}$-a.s. bound. In Section 7, we will illustrate these remarks with examples.

Finally, let us note that the above theorem provides convergence of spectra also for the random operator $\tilde{\mathbf{H}}_n$, which as we have mentioned and as was noted in [14] is not the case in general, even in the i.i.d. setting. To see this, it is enough to choose a function h vanishing everywhere on $\mathcal{X} \times \mathcal{X}$ except for the diagonal, for absolutely continuous π and such that $\int h(x, x)\pi(\mathrm{d}x) = \infty$. The validity of the law of large numbers for the spectrum of $\tilde{\mathbf{H}}_n$ in our case is of course again a consequence of our assumptions on h and F, which preclude such counterexamples.

Let us now pass to our second result, which is a tail inequality for the approximation of spectra. For this, we will work in a more restrictive, analytic framework, we will also impose stronger ergodicity assumptions on the chain.

Recall that a Harris ergodic Markov chain with transition function P and invariant measure π, is geometrically ergodic if there exists $0 < \rho < 1$ such that for every $x \in \mathcal{X}$ and some constant $M(x)$, we have for every $n \ge 0$,

$$\big\| P^n(x, \cdot) - \pi \big\|_{\mathrm{TV}} \le M(x)\rho^n, \tag{5}$$

where $\| \cdot \|_{\mathrm{TV}}$ is the total-variation distance and P^n is the n-step transition function of the chain.

Theorem 2.2. *Let π be a probability measure on $(\mathcal{X}, \mathcal{F})$, where $\mathcal{X}$ is a metric space and $\mathcal{F}$ the Borel σ-field. Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a bounded function and* $\mathbf{H}$ *the corresponding kernel operator defined by* (1). *Assume that there exist continuous functions $\phi_n : \mathcal{X} \to \mathbb{R}$, $n \in I$ (where $I = \{0, \dots, R\}$ or $I = \mathbb{N}$) which form an orthonormal system in $L_2(\pi)$ and a sequence of non-negative numbers $\lambda = (\lambda_n)_{n \in I} \in \ell_2(I)$ such that we have a point-wise equality*

$$h(x, y) = \sum_{n \in I} \lambda_n \phi_n(x) \phi_n(y),$$

with the series converging absolutely and almost uniformly on $\mathcal{X} \times \mathcal{X}$. Assume furthermore that $\mathbf{X} = (X_n)_{n \geq 0}$ is a geometrically ergodic Markov chain with invariant measure π, started at a point z. Then

$$\mathbb{P}\big(\delta_2\big(\lambda(\tilde{\mathbf{H}}_n), \lambda(\mathbf{H})\big) \geq t\big) \leq 2 \exp\left(-\frac{1}{L} n \min\left(\frac{t^2}{\sup_{x \in \mathcal{X}} h(x, x)^2}, \frac{t}{\sup_{x \in \mathcal{X}} h(x, x)}\right)\right),$$

where the constant L depends only on the transition function P and the starting point z.

In the above formulation, we do not specify the dependence of the constants in the inequality on the parameters of the Markov chain. This will be done in Section 6 via drift conditions.

We state Theorem 2.2 for chains started from a point. In fact, it holds also for chains started from more general measures μ satisfying some mild conditions. Since to formulate this condition we would need to introduce the regeneration technique for Markov chains, such a formulation is deferred to Remark 6.3 in Section 6.

We remark that the assumptions concerning the function h are satisfied for continuous positive definite kernels on a large class of topological spaces. In the case of compact spaces this fact is known as Mercer's theorem (see, e.g., [18,26]). Since there are many generalizations of this result, with subtle differences, and a discussion of this topic is beyond the scope of this article we prefer to formulate the theorem in an abstract form.

We remark that similar inequalities in the i.i.d. case were considered, for example, in [15,16] under weaker assumptions than the boundedness of h (instead some exponential integrability was assumed). However, those estimates consider a weaker metric between spectra and, when specialized to the case of bounded kernels, involve additional logarithmic factors. Thus, Theorem 2.2 (in a version for chains started not necessarily from a point) improves on their result for bounded kernels even in the i.i.d. case.

Let us also mention that in our case one can also obtain results for unbounded kernels, under appropriate drift conditions involving the function h (using, e.g., results from [3]). However, their formulation would be much more involved, so we restrict to the special case of uniformly bounded kernels.

We would like to stress the important role of positive definiteness in Theorem 2.2. As will be shown in the proof, thanks to this assumption we can replace the operator $\tilde{\mathbf{H}}_n$ by a sum of the form $\sum_{i=1}^{n-1} f(X_i) \otimes f(X_i)$ for some $f : \mathcal{X} \to L_2(\pi)$ which is an $L_2(\pi)$-valued additive functional of the Markov chain $\mathbf{X}$ (similar ideas in the i.i.d. case were used, e.g., in [15,16,25]). This allows to apply the regeneration technique for obtaining concentration inequalities for additive functionals of Markov chains.

3. Notation and preliminary facts

3.1. Markov chains

We will now present basic facts related to the regeneration technique for Markov chains on general state spaces. This technique was independently discovered by Nummelin [20] and Athreya–Ney [6] and relies on a decomposition of the trajectory of a Markov chain into one-dependent paths of random length. Instead of providing the technical details of the construction, we will just present its properties, which will be used in the proof. The technical details can be found in many monographs on Markov chains; we recommend [10,17,21].

Let thus $(\mathcal{X}, \mathcal{F})$ be a state space, with $\mathcal{F}$ countably generated and assume that P is a Markov chain transition function on $\mathcal{X}$. Assume also that the corresponding Markov chain $\mathbf{X} = (X_n)_{n\geq 0}$ is Harris ergodic. Then there exists a set $C \in \mathcal{F}$ with $\pi(C) > 0$, a positive integer m, $\delta > 0$ and a probability measure ν on $(\mathcal{X}, \mathcal{F})$, such that for all $x \in C$, $A \in \mathcal{F}$,

$$P^m(x, A) \geq \delta\nu(A). \tag{6}$$

Using the set C for any probability measure μ one can define two sequences of random variables $(\tilde{X}_n)_{n\geq 0}$, $(Y_n)_{n\geq 0}$ (on some probability space) with the following properties:

(A0) $(\tilde{X}_n)_{n\geq 0}$ is a Markov chain, $\tilde{X}_0 \sim \mu$.
(A1) $Y_n \in \{0, 1\}$.
(A2) The stopping times $T_0 = \inf\{k \geq 0\colon\ Y_k = 1\}$, $T_i = \inf\{k > T_{i-1}\colon\ Y_k = 0\}$ are almost surely finite. Moreover, $T_0, T_1 - T_0, T_2 - T_1, \ldots$ are independent random variables, whereas $T_1 - T_0, T_2 - T_1, \ldots$ are i.i.d. and their distribution depends only on P (and not on μ). Moreover, $\mathbb{E}(T_1 - T_0) < \infty$.
(A3) The blocks $Z_i = (\tilde{X}_{m(T_i+1)}, \tilde{X}_{m(T_i+1)+1}, \ldots, \tilde{X}_{mT_{i+1}+m-1})$ form a one-dependent stationary sequence of random variables with values in $(\mathcal{Z}, \mathcal{S})$, where $\mathcal{Z} = \bigcup_{k=1}^{\infty} \mathcal{X}^k$, $\mathcal{S} = \sigma(\bigcup_{k=1}^{\infty} \mathcal{F}^{\otimes k})$ (i.e., for all k, the σ-fields $\sigma(Z_i\colon\ i < k)$ and $\sigma(Z_i\colon\ i > k)$ are independent).
(A4) For any $f \in L_1(\pi)$ and all k,

$$\mathbb{E} \sum_{i=m(T_k+1)}^{mT_{k+1}+m-1} f(\tilde{X}_i) = m\mathbb{E}(T_1 - T_0)\pi f.$$

As already mentioned, in the proofs we will use only the above properties and so we do not present the general construction of the chain $(\tilde{X}_n)_{n\geq 0}$. Let us however briefly describe the intuition hidden behind it in the special case of $m = 1$. Informally, if one attempts to generate the chain then one draws $\tilde{X}_0$ according to the measure μ, and next if at step n one has $\tilde{X}_n = x$, then for $x \notin C$, the next variable $\tilde{X}_{n+1}$ is drawn from the distribution $P(x, \cdot)$ and one sets $Y_n = 0$. If $x \in C$ then one tosses a coin with heads probability equal to δ. If one gets heads, then $\tilde{X}_{n+1}$ is generated according to ν and Y_n is set to one, otherwise Y_n is set to zero and $\tilde{X}_{n+1}$ is generated according to the probability measure

$$Q(x, \cdot) = \frac{P(x, \cdot) - \delta\nu(\cdot)}{1 - \delta}.$$

It is straightforward but slightly tedious to formalize this intuition and prove that for Harris ergodic chains it gives properties (A0)–(A4). For general m, one can still repeat this construction for the m-step transition function to define the chain $(\tilde{X}_{nm})_{n\geq 0}$ and then fill in the intermediate variables in such a way that properties (A0)–(A4) are still satisfied (note that for $m = 1$ the blocks Z_i of property (A3) are in fact independent, which is not necessarily the case for general m). We refer the reader to [10,17,21] for the details.

Since the Markov chain $(X_n)_{n\geq 0}$, started from μ has the same distribution as $(\tilde{X}_n)_{n\geq 0}$ above, to prove a limit theorem for $(X_n)_{n\geq 0}$ it is enough to do it for $(\tilde{X}_n)_{n\geq 0}$ for which one can exploit the additional structure given by the auxiliary variables $(Y_n)_{n\geq 0}$, which often allows to reduce the proof to the corresponding limit theorem in the one-dependent or independent case. This strategy has been adopted for many problems, including the law of large numbers, the central limit theorem or the law of the iterated logarithm. We again refer to [10,17,21] for a detailed exposition. As a consequence, for the purpose of proving limit theorems, we can identify the sequences $(X_n)_{n\geq 0}$ and $(\tilde{X}_n)_{n\geq 0}$. In what follows, we will adopt this convention (in particular we will drop the tilde in $\tilde{X}_n$).

In the proofs, we will use the strong law of large numbers for Markov chains, which can be easily proved using the regeneration method (see [17,21]).

Theorem 3.1. *Let $\mathbf{X} = (X_n)_{n\geq 0}$ be a Harris ergodic Markov chain on $(\mathcal{X}, \mathcal{F})$, with invariant probability measure π and let $f : \mathcal{X} \to \mathbb{R}$ be a π-integrable function. Then with probability one, as $n \to \infty$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \pi f.$$

3.2. Linear algebra

The main linear-algebraic result we will need is the Hoffman–Wielandt inequality. To prove the law of large numbers, it will be sufficient to use its original finite-dimensional version. However, for the exponential inequality we will use the infinite-dimensional version proved in [9].

Theorem 3.2 (Hoffman–Wielandt inequality). *If A, B are normal Hilbert–Schmidt operators on some Hilbert space, then*

$$\delta_2\big(\lambda(A), \lambda(B)\big) \leq \|A - B\|_{\mathrm{HS}}.$$

4. Strong law of large numbers for U-statistics of Markov chains

Recall the notation (4). The aim of this section is to prove the following.

Proposition 4.1. *Let $\mathbf{X} = (X_n)_{n\geq 0}$ be a Harris ergodic Markov chain on $(\mathcal{X}, \mathcal{F})$ with invariant probability measure π and let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric measurable function. Assume that there exists a π-integrable $F : \mathcal{X} \to \mathbb{R}_+$, such that $|h(x, y)| \leq F(x)F(y)$ for all $x, y \in \mathcal{X}$. Then for every initial probability μ of the chain $\mathbf{X}$, with probability one,*

$$U_n(h) \to \pi h$$

as $n \to \infty$.

We remark that in the literature there are several results concerning laws of large numbers for U-statistics under dependence. In [1], such a result is obtained for a class of ergodic stationary sequences, under assumption of the same nature as ours. However, we need the above version, since for MCMC applications it is important to consider Markov chains started from a point (as the very purpose of MCMC algorithms is to simulate the stationary distribution, which is not directly accessible). Results of this type have been obtained recently, for example, in [4,8]; however, they require higher order ergodicity of the chain. We would like to add that the results in [8] are not expressed in terms of point-wise bounds on the kernel h but rather in terms of integrability of certain functionals on the paths of the Markov chains. Thus, in general they are not comparable to Proposition 4.1. On the one hand they may be applicable to kernels which are not bounded by tensor products, on the other hand the verification of assumptions may be more difficult.

To prove Proposition 4.1, we will use the following result which is a simple corollary to Theorem 5.2. in [1] (we remark that this theorem is stated for $\mathcal{Z} = \mathbb{R}$, but it is easy to see that its proof works for an arbitrary measurable space).

Lemma 4.2. *Let $\mathbf{Z} = (Z_k)_{k\geq 0}$ be a one-dependent stationary sequence of $(\mathcal{Z}, \mathcal{S})$-valued random variables and let $H : \mathcal{Z}^2 \to \mathbb{R}$ be a symmetric measurable function. Assume that there exists $F : \mathcal{Z} \to \mathbb{R}_+$ such that $|H(x, y)| \leq F(x)F(y)$ for all $x, y \in \mathcal{Z}$ and $\mathbb{E}F(Z_0) < \infty$. Then with probability one*

$$U_n(H, \mathbf{Z}) \to \mathbb{E}H(Z_0, Z_2)$$

as $n \to \infty$.

Proof of Proposition 4.1. Define $N_n = \sup\{k : mT_k + m - 1 \leq n - 1\}$ (with the convention that $\sup \varnothing = 0$). By the law of large numbers and property (A2), we have as $n \to \infty$,

$$\frac{n}{N_n} \to m\mathbb{E}(T_1 - T_0), \qquad \mathbb{P}_\mu\text{-a.s.} \tag{7}$$

Recall the space $\mathcal{Z}$ defined in property (A3). In what follows, we will use the following convention regarding its elements: for $\mathbf{x} = (x_1, \ldots, x_k) \in \mathcal{Z}$ we set $|\mathbf{x}| = k$. Let $H : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be the kernel defined by

$$H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \sum_{j=1}^{|\mathbf{y}|} h(x_i, y_j).$$

Note that for $\tilde{F} : \mathcal{Z} \to \mathbb{R}_+$, given by

$$\tilde{F}(\mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} F(x_i),$$

we have $|H(\mathbf{x}, \mathbf{y})| \le \tilde{F}(\mathbf{x})\tilde{F}(\mathbf{y})$. Moreover, by property (A4) we have $\mathbb{E}\tilde{F}(Z_0) < \infty$.

By properties (A3), (A4) and the Fubini theorem, we also get

$$\mathbb{E}H(Z_k, Z_l) = \big(m\mathbb{E}(T_1 - T_0)\big)^2 \pi^{\otimes 2} h$$

if $|k - l| \ge 2$.

Thus, by Lemma 4.2, (7) and the above equality, we get

$$\frac{1}{n(n-1)} \sum_{0 \le i \ne j \le N_n - 1} H(Z_i, Z_j) \to \pi^{\otimes 2} h. \tag{8}$$

Define also $\tilde{H} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ as $\tilde{H}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \sum_{j=1}^{|\mathbf{y}|} |h(x_i, y_j)|$.

In view of (8), to prove the proposition it remains to show that with probability one the sequences

$$I_n = \frac{1}{n(n-1)} \sum_{i=0}^{mT_0+m-1} \sum_{j=m(T_0+1)}^{n-1} \big|h(X_i, X_j)\big|,$$

$$II_n = \frac{1}{n(n-1)} \sum_{i=0}^{N_n} \tilde{H}(Z_i, Z_i),$$

$$III_n = \frac{1}{n(n-1)} \sum_{i=0}^{N_n - 1} \tilde{H}(Z_{N_n}, Z_i),$$

converge a.s. to 0 as $n \to \infty$.

Note that

$$I_n \le \sum_{i=0}^{mT_0+m-1} F(X_i) \frac{1}{n(n-1)} \sum_{j=0}^{n-1} F(X_j) \to 0 \qquad \text{a.s.}$$

since by Theorem 3.1, $n^{-1}\sum_{j=0}^{n-1} F(X_j) \to \pi F$ a.s.

As for II_n, we have

$$\mathbb{E}\tilde{H}(Z_i, Z_i)^{1/2} \le \mathbb{E} \sum_{i=m(T_i+1)}^{mT_{i+1}+m-1} F(X_i) = m\mathbb{E}(T_1 - T_0)\pi F < \infty,$$

where we again used (A4).

Thus, using (A3) and (7) we get by the Marcinkiewicz law of large numbers that $II_n \to 0$ a.s.

To prove that $III_n \to 0$ a.s., note that

$$III_n = \frac{N_n(N_n+1)}{2n(n-1)} U_{N_n+1}(\tilde{H}, \mathbf{Z}) - \frac{N_n(N_n-1)}{2n(n-1)} U_{N_n}(\tilde{H}, \mathbf{Z}).$$

By Lemma 4.2 and (7), both terms on the right-hand side above converge a.s. to

$$2^{-1}\bigl(m\mathbb{E}(T_1 - T_0)\bigr)^{-2}\mathbb{E} \sum_{i=m(T_0+1)}^{mT_1+m-1} \sum_{j=m(T_2+1)}^{mT_3+m-1} \bigl|h(X_i, X_j)\bigr| \le 2^{-1}(\pi F)^2 < \infty,$$

where in the first inequality we used the assumption on h and F together with (A3), (A4) and the Fubini theorem.

This shows that indeed $III_n \to 0$ a.s. and ends the proof of Proposition 4.1. □

5. Proof of Theorem 2.1

To prove Theorem 2.1, we will need one more simple result, namely a Marcinkiewicz–Zygmund-type law of large numbers for Markov chains. Its proof is a standard application of the regeneration technique. Since we have not been able to find it in the literature, we provide it for completeness.

Lemma 5.1 (Marcinkiewicz–Zygmund LLN for Markov chains). *Let $\mathbf{X} = (X_n)_{n\ge 0}$ be a Harris ergodic Markov chain on $(\mathcal{X}, \mathcal{F})$ and let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function. Consider $p \in (0, 1)$ and assume that $\pi|f|^p < \infty$. Then for any initial measure of the chain, with probability one*

$$\frac{1}{n^{1/p}} \sum_{i=0}^{n-1} f(X_i) \to 0.$$

Proof. As in the proof of Proposition 4.1, define $N = N_n = \sup\{k\colon mT_k + m - 1 \le n - 1\}$ and recall (7). Define a function $F : \mathcal{Z} \to \mathbb{R}$ (where $\mathcal{Z}$ is defined in property (A3) of Section 3) with the formula

$$F(\mathbf{x}) = \sum_{i=1}^{|\mathbf{x}|} \bigl|f(x_i)\bigr|.$$

Then, by concavity of the function $t \mapsto |t|^p$ and property (A4) we get for $i \geq 1$,

$$\mathbb{E}F(Z_i)^p \leq \mathbb{E} \sum_{i=m(T_i+1)}^{mT_{i+1}+m-1} \left|f(X_i)\right|^p = m\big(\mathbb{E}(T_1 - T_0)\big)\pi|f|^p < \infty. \tag{9}$$

Now

$$\frac{1}{n^{1/p}}\left|\sum_{i=0}^{n-1} f(X_i)\right| \leq \frac{1}{n^{1/p}} \sum_{i=0}^{mT_0+m-1} \left|f(X_i)\right| + \frac{1}{n^{1/p}} \sum_{i=0}^{N_n} F(Z_i).$$

The first term on the right-hand side above converges a.s. to zero as $n \to \infty$. Moreover, since Z_i form a stationary one-dependent sequence by (7), (9) and the classical Marcinkiewicz–Zygmund LLN, the second term also converges a.s. to zero, which ends the proof of the lemma. □

The proof of Theorem 2.1 will mimic closely the corresponding proof by Koltchinskii and Giné, in fact one could keep the linear-algebraic part exactly the same, while replacing just the probabilistic ingredients (using Proposition 4.1 and Lemma 5.1). However, we will slightly change the exposition with respect to [14], which will allow to shorten the proof a little bit.

Proof of Theorem 2.1. Let us first notice that thanks to the Hoffman–Wielandt inequality and the assumption on h, we have

$$\delta_2\big(\lambda(\mathbf{H}_n), \lambda(\tilde{\mathbf{H}}_n)\big)^2 \leq \|\mathbf{H}_n - \tilde{\mathbf{H}}_n\|_{\mathrm{HS}}^2 \leq \frac{1}{n^2}\sum_{i=1}^{n} F(X_i)^4.$$

Since $\pi F^2 < \infty$, by Lemma 5.1 applied with $p = 1/2$, the right-hand side above converges a.s. to zero. Thus, it is enough to prove the theorem for the matrix $\tilde{\mathbf{H}}_n$.

Since $\pi^{\otimes 2}h^2 \leq (\pi F^2)^2 < \infty$, $\mathbf{H}$ is a Hilbert–Schmidt operator and so, by the spectral theorem, there exists an orthonormal system $(\phi_i)_{i\in I}$ in $L_2(\pi)$ (where $I = \{0, \ldots, R\}$ for some $R \in \mathbb{N}$ or $I = \mathbb{N}$) and a square summable sequence $(\lambda_i)_{i\in I}$ with non-increasing absolute values such that

$$h(x, y) = \sum_{i\in I} \lambda_i \phi_i(x)\phi_i(y), \tag{10}$$

where the equality holds in the $L_2(\pi^{\otimes 2})$ sense.

As in [14] assume first that $h(x, y) = \sum_{i=0}^{R} \lambda_i\phi_i(x)\phi_i(y)$ and the equality holds pointwise. Define for $n \geq 0$, the sequence of vectors in $\mathbb{R}^n$,

$$\Phi_i^n = \left(\frac{\phi_i(X_0)}{\sqrt{n}}, \ldots, \frac{\phi_i(X_{n-1})}{\sqrt{n}}\right), \qquad 0 \leq i \leq R$$

and note that for $u \in \mathbb{R}^n$,

$$\tilde{\mathbf{H}}_n u = \sum_{i=0}^{R} \lambda_i \langle \Phi_i^n, u \rangle \Phi_i^n,$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^n$.

Now consider the space $\mathbb{R}^{R+1}$ with the standard basis $e_0, \ldots, e_R$ and let $A_n : \mathbb{R}^{R+1} \to \mathbb{R}^n$ be the operator given by $A_n e_i = \Phi_i^n$, $i = 0, \ldots, R$. Define also an operator K on $\mathbb{R}^{R+1}$ as

$$Ku = \sum_{i=0}^{R} \lambda_i \langle e_i, u \rangle e_i.$$

Then, as one can easily check,

$$\tilde{\mathbf{H}}_n = A_n K A_n^T$$

and since for any two operators $K_1 : \mathbb{R}^a \to \mathbb{R}^b$ and $K_2 : \mathbb{R}^b \to \mathbb{R}^a$, the (algebraic) spectra of $K_1 K_2$ and $K_2 K_1$ are the same (recall our convention of completing the spectra with zeros to an infinite sequence), we get $\lambda(\tilde{\mathbf{H}}_n) = \lambda(K A_n^T A_n)$. Together with the obvious equality $\lambda(K) = \lambda(\mathbf{H})$, this gives

$$\delta_2\big(\lambda(\tilde{\mathbf{H}}_n), \lambda(\mathbf{H})\big) = \delta_2\big(\lambda\big(K A_n^T A_n\big), \lambda(K)\big). \tag{11}$$

But for each $i, j = 0, \ldots, R$ we have $\langle A_n^T A_n e_i, e_j \rangle = \langle A_n e_i, A_n e_j \rangle = \frac{1}{n} \sum_{k=0}^{n-1} \phi_i(X_k) \phi_j(X_k)$. Thus, by Theorem 3.1 with probability one, $\langle A_n^T A_n e_i, e_j \rangle \to \pi \phi_i \phi_j = \delta_{ij}$ and thus $K A_n^T A_n \to K$, which implies that the right-hand side of (11) converges to zero a.s. (note that $K A_n^T A_n$ in general is not a normal matrix, so we cannot use the Hoffman–Wielandt inequality, but we are working now in a fixed dimension $R + 1$ and so we can simply use the fact that the eigenvalues are continuous functions of the matrix entries; see, e.g., Appendix D in [13]). This proves the theorem in the special case of finite dimensional kernels.

Consider now an arbitrary kernel h, satisfying (10). Fix $\varepsilon > 0$. Since $\sum_{i \in I} \lambda_i^2 < \infty$, there exists R such that $\sum_{i \in I, i > R} \lambda_i^2 < \varepsilon$. Set $h_R(x, y) = \sum_{i=0}^{R} \lambda_i \phi_i(x) \phi_i(y)$ (by which we mean that the equality holds pointwise, for some particular fixed choice of representatives from the equivalence class of ϕ_i in $L^2(\pi)$). Let $\mathbf{H}^R$ be the kernel operator corresponding to h_R and $\tilde{\mathbf{H}}_n^R = (h_R(X_i, X_j))_{0 \le i, j \le n-1}$. Define moreover $\tilde{h}_R = h - h_R$. We have

$$\delta_2\big(\lambda(\mathbf{H}), \lambda\big(\mathbf{H}^R\big)\big)^2 = \sum_{i \in I, i > R}^{\infty} \lambda_i^2 < \varepsilon. \tag{12}$$

Define the function $F_1 = F + \sum_{i=0}^{R} \sqrt{|\lambda_i|} |\phi_i|$ (again we interpret this equality in the pointwise sense) and note that $F_1 \in L_2(\pi)$. Moreover, for all $x, y \in \mathcal{X}$, $|h_R(x, y)|, |\tilde{h}_R(x, y)| \le F_1(x) F_1(y)$. Thus, by the first part of the proof, we get

$$\delta_2\big(\lambda\big(\mathbf{H}^R\big), \lambda\big(\tilde{\mathbf{H}}_n^R\big)\big) \to 0 \qquad \text{a.s.}, \tag{13}$$

while by Proposition 4.1 and Lemma 5.1 we obtain that with probability one,

$$\lim_{n\to\infty} \|\tilde{\mathbf{H}}_n - \tilde{\mathbf{H}}_n^R\|_{\mathrm{HS}}^2 = \lim_{n\to\infty} \frac{1}{n^2} \sum_{i,j=0}^{n-1} \tilde{h}_R(X_i, X_j)^2$$

$$\leq \lim_{n\to\infty} U_n(h, \mathbf{X}) + \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=0}^{n-1} F_1(X_i)^4 = \pi^{\otimes 2}\tilde{h}_R^2 = \sum_{i\in I, i>R}^{\infty} \lambda_i^2 < \varepsilon.$$

Thus, by the Hoffman–Wielandt inequality,

$$\limsup_{n\to\infty} \delta_2\big(\lambda(\tilde{\mathbf{H}}_n), \lambda\big(\tilde{\mathbf{H}}_n^R\big)\big) \leq \varepsilon^{1/2} \qquad \text{a.s.}$$

In combination with (12) and (13), this implies that for every $\varepsilon > 0$,

$$\limsup_{n\to\infty} \delta_2\big(\lambda(\mathbf{H}), \lambda(\tilde{\mathbf{H}}_n)\big) \leq 2\varepsilon^{1/2} \qquad \text{a.s.}$$

and in consequence

$$\delta_2\big(\lambda(\mathbf{H}), \lambda(\tilde{\mathbf{H}}_n)\big) \to 0 \qquad \text{a.s.} \qquad \square$$

6. Proof of Theorem 2.2

Proof of Theorem 2.2. In what follows by $\langle\cdot,\cdot\rangle$ we will denote both the inner product in $L_2(\pi)$ and in finite-dimensional spaces, since the precise meaning will always be clear from the context, this should not lead to ambiguity. The letters C, c will denote absolute positive constants, whose values may differ between occurrences.

Define $f : \mathcal{X} \to L_2(\pi)$ with the formula

$$f(x) = \sum_{i\in I} \sqrt{\lambda_i}\phi_i(x)\phi_i(\cdot).$$

Note that $\sum_{i\in I}(\sqrt{\lambda_i}\phi_i(x))^2 = h(x,x) < \infty$ and that ϕ_i form an orthonormal system in $L_2(\pi)$, so the above series indeed converges in $L_2(\pi)$. Consider now a random operator on $L_2(\pi)$ given by

$$K_n = \frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \otimes f(X_i),$$

that is, for all $u \in L_2(\pi)$

$$K_n u = \frac{1}{n}\sum_{i=0}^{n-1} \big\langle f(X_i), u\big\rangle f(X_i).$$

Note that K_n can be written as $A_n A_n^T$, where $A_n : \mathbb{R}^n \to L_2(\pi)$ is defined by $A_n e_i = n^{-1/2} f(X_i)$ ($e_0, \ldots, e_{n-1}$ being the standard basis in $\mathbb{R}^n$). Thus $\lambda(K_n) = \lambda(A_n^T A_n)$ (recall that we append spectra of finite dimensional operators with infinite sequences of zeros). But

$$\langle A_n^T A_n e_i, e_j \rangle = \langle A_n e_i, A_n e_j \rangle = \frac{1}{n}\langle f(X_i), f(X_j)\rangle = \frac{1}{n}\sum_{k\in I}\lambda_k \phi_k(X_i)\phi_k(X_j) = \frac{1}{n} h(X_i, X_j),$$

so $A_n^T A_n = \tilde{\mathbf{H}}_n$. Thus, our goal will be to bound the distance between the spectrum of K_n and the sequence λ.

The random operator K_n is a sum of independent random rank one operators, moreover, using the fact that ϕ_i form an orthonormal system in $L_2(\pi)$ one easily checks that

$$\| f(x) \otimes f(x) \|_{\mathrm{HS}} = h(x, x) \tag{14}$$

and

$$\mathbb{E}_\pi f(X_i) \otimes f(X_i) = \mathbf{H} \tag{15}$$

(where the expectation on the left-hand side is the Bochner integral in the Hilbert space of Hilbert–Schmidt operators).

Thus, we can apply to K_n classical results concerning concentration for sums of independent Banach space valued random variables [after passing to the block decomposition given by (A3)]. The inequality we will use is a version of Bernstein's ψ_1 inequality. To formulate it, let us first recall the definition of the Orlicz ψ_1 norm. For a Banach space valued random variable X, we define

$$\|X\|_{\psi_1} = \inf\{\rho > 0 \colon \mathbb{E}\exp(\|X\|/\rho) \le 2\}.$$

By exponential Chebyshev's inequality, we have

$$\mathbb{P}(|X| \ge t) \le 2\exp(-t/\|X\|_{\psi_1}) \tag{16}$$

for $t > 0$.

The following inequality is a simple corollary to Theorem 1.4. in [27].

Lemma 6.1. *Let U, U_i, $i = 1, \ldots, n$, be i.i.d. mean zero random variables with values in a Banach space $(B, \|\cdot\|)$. Assume that $\|U\|_{\psi_1} < \infty$. Then for all $t > 0$,*

$$\mathbb{P}\left(\left|\left\|\sum_{i=1}^n U_i\right\| - \mathbb{E}\left\|\sum_{i=1}^n U_i\right\|\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{n\|U\|_{\psi_1}^2}, \frac{t}{\|U\|_{\psi_1}}\right)\right),$$

where $c > 0$ is a universal constant.

It is well known (see, e.g., [17], Chapters 15, 16, or [7,22]) that for uniformly ergodic Markov chains we have $\|T_1 - T_0\|_{\psi_1} < \infty$ and if the chain is started from a point, then also $\|T_0\|_{\psi_1} < \infty$, which allows for the use of the above inequality in our setting.

Let us now define $g(x) = f(x) \otimes f(x)$, $U_i = \sum_{i=m(T_i+1)}^{mT_{i+1}+m-1} (g(X_i) - \pi g)$ and recall the definition $N = N_n = \sup\{k\colon\ mT_k + m - 1 \le n - 1\}$. Using properties (A0)–(A4) and (15) we get that $\mathbb{E}U_i = 0$ and

$$\begin{aligned}
\left\| \frac{1}{n}\sum_{i=0}^{n-1} g(X_i) - \mathbf{H}\right\|_{\mathrm{HS}} &= \frac{1}{n}\left\|\sum_{i=0}^{n-1}\bigl(g(X_i) - \pi g\bigr)\right\|_{\mathrm{HS}} \\
&\le \frac{1}{n}\left\|\sum_{i=0}^{(mT_0+m-1)\wedge(n-1)}\bigl(g(X_i) - \pi g\bigr)\right\|_{\mathrm{HS}} + \frac{1}{n}\left\|\sum_{i=0}^{N-1} U_i\right\|_{\mathrm{HS}} \\
&\quad + \frac{1}{n}\left\|\sum_{i=m(T_N+1)}^{n-1} g(X_i) - \pi g\right\|_{\mathrm{HS}} \\
&\le \frac{2m}{n}(T_0+1)\|g\|_\infty + \frac{1}{n}\left\|\sum_{i=0}^{N-1} U_i\right\|_{\mathrm{HS}} + \frac{2}{n}\bigl(n - m(T_N+1)\bigr)_+\|g\|_\infty,
\end{aligned}$$

where $\|g\|_\infty = \sup_{x\in\mathcal{X}} \|g(x)\|_{\mathrm{HS}}$.

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=0}^{n-1} g(X_i) - \mathbf{H}\right\|_{\mathrm{HS}} \ge t\right) \\
&\quad\le \mathbb{P}\bigl(2m(T_0+1)\|g\|_\infty \ge tn/3\bigr) + \mathbb{P}\left(\left\|\sum_{i=0}^{N-1} U_i\right\|_{\mathrm{HS}} \ge tn/3\right) \\
&\qquad + \mathbb{P}\bigl(2\bigl(n - m(T_N+1)\bigr)\|g\|_\infty \ge tn/3\bigr).
\end{aligned} \tag{17}$$

By (16),

$$\mathbb{P}\bigl(2m(T_0+1)\|g\|_\infty \ge tn/3\bigr) \le 2\exp\left(-\frac{nt}{6m\|T_0+1\|_{\psi_1}\|g\|_\infty}\right). \tag{18}$$

Moreover, by Lemma 3 in [2] we have for all $t > 0$,

$$\mathbb{P}\bigl(2\bigl(n - m(T_N+1)\bigr) \ge t\bigr) \le 2\exp\left(-c\frac{t}{m\tau\log\tau}\right),$$

where $\tau = \max\{\|T_0+1\|_{\psi_1}, \|T_1 - T_0\|_{\psi_1}\}$ (we remark that the notation and the definition of splitting times in [2] are slightly different than ours, in particular the Markov chain there is indexed by $\mathbb{N}\setminus\{0\}$ and not by $\mathbb{N}$, however it is easy to see that the simple proof of Lemma 3 can be carried over to our setting). Thus,

$$\mathbb{P}\bigl(2\bigl(n - m(T_N+1)\bigr)\|g\|_\infty \ge tn/3\bigr) \le 2\exp\left(-c\frac{nt}{m\|g\|_\infty\tau\log\tau}\right). \tag{19}$$

To handle the middle term in the decomposition (17), we will apply Lemma 6.1 to the random variables U_i. Since for $m > 1$, these variables are only one-dependent; moreover, the number of full blocks Z_i in the sequence $X_0, \ldots, X_{n-1}$ is random, there are two technical steps, which have to be carried out first, namely we have to split the sum $\sum_{i=0}^{N-1} U_i$ into odd and even terms and use a Lévy type inequality to handle the random number of summands. We have

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=0}^{N-1} U_i\right\|_{\mathrm{HS}} \geq t/3\right) \\
&\quad \leq \mathbb{P}\left(\left\|\sum_{0\leq i\leq N-1, 2|i} U_i\right\|_{\mathrm{HS}} \geq t/6\right) + \mathbb{P}\left(\left\|\sum_{0\leq i\leq N-1, \neg 2|i} U_i\right\|_{\mathrm{HS}} \geq t/6\right) \\
&\quad \leq C\mathbb{P}\left(\left\|\sum_{i=0}^{\lfloor n/m\rfloor -1} \tilde{U}_i\right\|_{\mathrm{HS}} \geq t/C\right),
\end{aligned}$$

where C is a universal constant and $\tilde{U}_i$ is a sequence of independent random variables, distributed as U_0. In the last inequality, we used the fact that $Nm \leq n$ and a Lévy-type inequality for i.i.d. Banach-space valued random variables due to Montgomery–Smith [19], which asserts that for a sequence W_i of i.i.d. Banach space-valued variables

$$\mathbb{P}\left(\max_{k\leq n}\left\|\sum_{i=1}^{k} W_i\right\| \geq t\right) \leq C\mathbb{P}\left(\left\|\sum_{i=1}^{n} W_i\right\| \geq t/C\right).$$

Now, we have

$$\|U_i\|_{\psi_1} \leq 2m\|g\|_\infty \|T_1 - T_0\|_{\psi_1}$$

and so Lemma 6.1 gives

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{i=0}^{N-1} U_i\right\|_{\mathrm{HS}} \geq C\mathbb{E}\left\|\sum_{i=0}^{\lfloor n/m\rfloor -1} \tilde{U}_i\right\|_{\mathrm{HS}} + s/3\right) \\
&\quad \leq 2\exp\left(-c\min\left(\frac{s^2}{nm\|g\|_\infty^2\|T_1 - T_0\|_{\psi_1}^2}, \frac{s}{m\|g\|_\infty\|T_1 - T_0\|_{\psi_1}}\right)\right).
\end{aligned}$$

Using the above bound together with (17), (18) and (19) we arrive (after adjusting the constants) at

$$\begin{aligned}
&\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=0}^{n-1} g(X_i) - \mathbf{H}\right\|_{\mathrm{HS}} \geq Cn^{-1}\mathbb{E}\left\|\sum_{i=0}^{\lfloor n/m\rfloor -1} \tilde{U}_i\right\|_{\mathrm{HS}} + t\right) \\
&\quad \leq 2\exp\left(-cn\min\left(\frac{t^2}{m\|g\|_\infty^2\tau^2}, \frac{t}{m\|g\|_\infty\tau\log\tau}\right)\right).
\end{aligned}$$

Using the fact that the norm $\|\cdot\|_{\mathrm{HS}}$ is Hilbertian and $\mathbb{E}\tilde{U}_i = 0$, we obtain

$$\mathbb{E}\Bigg\|\sum_{i=0}^{\lfloor n/m\rfloor-1}\tilde{U}_i\Bigg\|_{\mathrm{HS}}^2 = \sum_{i=0}^{\lfloor n/m\rfloor-1}\mathbb{E}\|\tilde{U}_i\|_{\mathrm{HS}}^2 \le \frac{4n}{m}m^2\mathbb{E}(T_1-T_0)^2\|g\|_\infty^2 \le Cnm\tau^2\|g\|_\infty^2,$$

which combined with the previous inequality gives

$$\begin{aligned}&\mathbb{P}\Bigg(\Bigg\|\frac{1}{n}\sum_{i=0}^{n-1}g(X_i)-\mathbf{H}\Bigg\|_{\mathrm{HS}} \ge C\sqrt{\frac{m}{n}}\tau\|g\|_\infty + t\Bigg)\\ &\quad\le 2\exp\bigg(-cn\min\bigg(\frac{t^2}{m\|g\|_\infty^2\tau^2},\frac{t}{m\|g\|_\infty\tau\log\tau}\bigg)\bigg).\end{aligned}$$

It is easy to see that by adjusting the value of the absolute constant this is equivalent to

$$\mathbb{P}\Bigg(\Bigg\|\frac{1}{n}\sum_{i=0}^{n-1}g(X_i)-\mathbf{H}\Bigg\|_{\mathrm{HS}} \ge t\Bigg) \le 2\exp\bigg(-cn\min\bigg(\frac{t^2}{m\|g\|_\infty^2\tau^2},\frac{t}{m\|g\|_\infty\tau\log\tau}\bigg)\bigg). \tag{20}$$

Since by (14) $\|g\|_\infty = \sup_{x\in\mathcal{X}}|h(x,x)|$, to finish the proof of Theorem 2.2 it is enough to combine the above inequality with Theorem 3.2. □

Remark 6.2. Let us mention that a Markov chain is geometrically ergodic iff it satisfies the following drift condition (see Theorem 16.0.1. in [17]). There exists $\lambda \in (0,1)$, $b \in \mathbb{R}_+$ and $V : \mathcal{X} \to [1,\infty)$ such that for some set C, satisfying (6) and $\pi(C) > 0$,

$$P^m V - V \le -\lambda V + b\mathbf{1}_C$$

and $K := \sup_{x\in C} V(x) < \infty$. Finding appropriate drift functions is in fact the most common way of proving geometric ergodicity.

It turns out that one can bound the quantity τ appearing in the estimate (20) in terms of the parameters of the drift condition and (6). Such an estimate follows directly from Propositions 6, 7 from [3] (obtained with help of previous important estimates from [7]). Namely for a chain started from a point x, we have

$$\begin{aligned}\tau &\le 2\log\bigg(\frac{\log(6/(2-\delta))}{\log(6/(2-\delta))}\bigg)\\ &\quad\times\max\bigg(\frac{\log(V(x)\mathbf{1}_{C^c}(x)+(b(1-\lambda)^{-1}+K)\mathbf{1}_C(x))}{\log 2},\frac{\log(b(1-\lambda)^{-1}+K)}{\log 2},1\bigg)\\ &\quad\times\frac{1}{\log(1/(1-\lambda))},\end{aligned}$$

where δ is the parameter from (6).

Remark 6.3. It is clearly seen from the proof of Theorem 2.2 that the chain does not have to be started from a point. It is sufficient to assume that the stopping time T_0 is exponentially integrable under the starting measure μ. This will be the case, for example, if the function V in the drift conditions is μ-integrable (as follows by Proposition 4.1. (ii) in [7]).

Remark 6.4. We also note that the absolute constant c in (20) can be given explicitly, since Lemma 6.1 with explicit constants is known [27], the constant from Lemma 3 in [2] can be easily read from the proof and the Lévy type inequality by Montgomery–Smith is also given with explicit constants [19]. We do not pursue this direction here. See [3] for related inequalities for additive functionals of Markov chains with explicit constants.

7. Discussion of optimality. Counterexamples

We would like to conclude with an example of a square integrable kernel h and a uniformly ergodic Markov chain for which the conclusion of Theorem 2.1 fails and the empirical counterpart of the spectrum almost surely is not convergent to the spectrum of $\mathbf{H}$. The example uses directly the construction of [1], where a counterexample to the law of large numbers for U-statistics was given. We adapt it to our setting and provide the details for the sake of completeness.

Let thus $\varepsilon_0, \varepsilon_1, \ldots$ be i.i.d. random variables with distribution $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = 0) = 1/2$ and $Y_0, Y_1, \ldots$ – i.i.d. random variables distributed uniformly on the interval $(0, 1)$, independent of the sequence (ε_i). Define $X_0 = x$,

$$X_{n+1} = \begin{cases} X_n & \text{if } \varepsilon_n = 0, \\ Y_{n+1} & \text{if } \varepsilon_n = 1. \end{cases}$$

It is easy to see that (6) is satisfied with $m = 1$, $\delta = 1/2$, $C = (0, 1)$ and ν being the Lebesgue measure on $(0, 1)$, thus (see [17]) the chain is uniformly ergodic (i.e., it is geometrically ergodic and the function $M(x)$ in (5) is bounded by a constant independent of x). The unique stationary measure for the chain, π is in this case the Lebesgue measure. Consider now a function $h : (0, 1)^2 \to \mathbb{R}$ given by $h(x, y) = 0$ if $x \neq y$ and $h(x, x) = 1/x^3$.

Of course $h = 0$ $\pi \otimes \pi$-a.s. and so $\mathbf{H} = 0$. Let now $i_0 < i_1 < i_2 < \cdots$ be defined as $i_0 = 0$, $i_{n+1} = \min\{i > i_n : \varepsilon_i = 0, \varepsilon_{i+1} = 1\}$. Then for $k > 0$, $X_{i_k} = X_{i_k+1}$, moreover conditionally on $(\varepsilon_i)_{i \geq 0}$, X_{i_k} are i.i.d., distributed according to π. Since the absolute value of the largest eigenvalue of a matrix is not smaller than the absolute value of its maximal entry, both $\mathbf{H}_n$ and $\tilde{\mathbf{H}}_n$ have at least one eigenvalue, which in absolute value exceeds $n^{-1} \max_{0 \leq i \leq n-2} h(X_i, X_{i+1})$. Moreover, by the law of large numbers $i_n/n \to 4$ a.s., so using the conditional independence of X_i, the Borel–Cantelli lemma and the fact that $\mathbb{P}(X_{i_k} \leq t) = t$ for $t \in (0, 1)$, we get

$$\begin{aligned} \limsup_{n\to\infty} \max \lambda(\mathbf{H}_n), \limsup_{n\to\infty} \max \lambda(\tilde{\mathbf{H}}_n) &\geq \limsup_{k\to\infty} \frac{1}{i_k + 2} h(X_{i_k}, X_{i_k+1}) \\ &= \limsup_{k\to\infty} \frac{1}{i_k + 2} \frac{1}{X_{i_k}^3} = \infty \qquad \text{a.s.} \end{aligned}$$

This shows that the law of large numbers for spectra fails in this case.

Acknowledgements

Research partially supported by NCN Grant N N201 608740. We would like to thank the anonymous referee for all the remarks concerning the submitted version of the manuscript.

References

[1] Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T. and Weiss, B. (1996). Strong laws for L- and U-statistics. *Trans. Amer. Math. Soc.* **348** 2845–2866. MR1363941

[2] Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985

[3] Adamczak, R. and Bednorz, W. (2013). Exponential concentration inequalities for additive functionals of Markov chains. Available at arXiv:1201.3569.

[4] Andrieu, C., Jasra, A., Doucet, A. and Del Moral, P. (2011). On nonlinear Markov chain Monte Carlo. *Bernoulli* **17** 987–1014. MR2817614

[5] Arcones, M.A. (1998). The law of large numbers for U-statistics under absolute regularity. *Electron. Commun. Probab.* **3** 13–19 (electronic). MR1624866

[6] Athreya, K.B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501. MR0511425

[7] Baxendale, P.H. (2005). Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *Ann. Appl. Probab.* **15** 700–738. MR2114987

[8] Bertail, P. and Clémençon, S. (2011). A renewal approach to Markovian U-statistics. *Math. Methods Statist.* **20** 79–105. MR2882153

[9] Bhatia, R. and Elsner, L. (1994). The Hoffman–Wielandt inequality in infinite dimensions. *Proc. Indian Acad. Sci. Math. Sci.* **104** 483–494. MR1314392

[10] Chen, X. (1999). Limit theorems for functionals of ergodic Markov chains with general state space. *Mem. Amer. Math. Soc.* **139** xiv+203. MR1491814

[11] Douc, R., Fort, G., Moulines, E. and Soulier, P. (2004). Practical drift conditions for subgeometric rates of convergence. *Ann. Appl. Probab.* **14** 1353–1377. MR2071426

[12] Douc, R., Guillin, A. and Moulines, E. (2008). Bounds on regeneration times and limit theorems for subgeometric Markov chains. *Ann. Inst. H. Poincaré Probab. Statist.* **44** 239–257. MR2446322

[13] Horn, R.A. and Johnson, C.R. (2013). *Matrix Analysis*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2978290

[14] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185

[15] Mendelson, S. and Pajor, A. (2005). Ellipsoid approximation using random vectors. In *Learning Theory. Lecture Notes in Computer Science* **3559** 429–443. Berlin: Springer. MR2203278

[16] Mendelson, S. and Pajor, A. (2006). On singular values of matrices with independent rows. *Bernoulli* **12** 761–773. MR2265341

[17] Meyn, S. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2509253

[18] Minh, H.Q., Niyogi, P. and Yao, Y. (2006). Mercer's theorem, feature maps, and smoothing. In *Learning Theory. Lecture Notes in Computer Science* **4005** 154–168. Berlin: Springer. MR2280604

[19] Montgomery-Smith, S.J. (1993). Comparison of sums of independent identically distributed random vectors. *Probab. Math. Statist.* **14** 281–285 (1994). MR1321767

[20] Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Z. Wahrsch. Verw. Gebiete* **43** 309–318. MR0501353

[21] Nummelin, E. (1984). *General Irreducible Markov Chains and Nonnegative Operators*. *Cambridge Tracts in Mathematics* **83**. Cambridge: Cambridge Univ. Press. MR0776608

[22] Roberts, G.O. and Rosenthal, J.S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[23] Rosasco, L., Belkin, M. and De Vito, E. (2010). On learning with integral operators. *J. Mach. Learn. Res.* **11** 905–934. MR2600634

[24] Shawe-Taylor, J., Williams, C.K.I., Cristianini, N. and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory* **51** 2510–2522. MR2246374

[25] Smale, S. and Zhou, D.-X. (2009). Geometry on probability spaces. *Constr. Approx.* **30** 311–323. MR2558684

[26] Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *J. Complexity* **21** 337–349. MR2138444

[27] Vershynin, R. (2000). On large random almost Euclidean bases. *Acta Math. Univ. Comenian.* (*N.S.*) **69** 137–144. MR1819516

[28] von Luxburg, U., Belkin, M. and Bousquet, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807

Received November 2013