Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains, and applications

Quentin Duchemin, Yohann de Castro, Claire Lacour

Concentration inequality for U-statistics for uniformly ergodic Markov chains

Quentin Duchemin
LAMA, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France.
quentin.duchemin@univ-eiffel.fr
&
Yohann De Castro
Institut Camille Jordan, École Centrale de Lyon, Lyon, France
yohann.de-castro@ec-lyon.fr
&
Claire Lacour
LAMA, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France.
claire.lacour@univ-eiffel.fr

February 17, 2021

Abstract

We prove a new concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. Working with bounded π-*canonical* kernels, we show that we can recover the convergence rate of [4] who proved a concentration result for U-statistics of independent random variables and canonical kernels. Our proof relies on an inductive analysis where we use martingale techniques, uniform ergodicity, Nummelin splitting and Bernstein's type inequality.

Our result allows us to conduct three applications. First, we establish a new exponential inequality for the estimation of spectra of trace class integral operators with MCMC methods. The novelty is that this result holds for kernels with positive and negative eigenvalues, which is new as far as we know.

In addition, we investigate generalization performance of online algorithms working with pairwise loss functions and Markov chain samples. We provide an online-to-batch conversion result by showing how we can extract a low risk hypothesis from the sequence of hypotheses generated by any online learner.

We finally give a non-asymptotic analysis of a goodness-of-fit test on the density of the invariant measure of a Markov chain. We identify the classes of alternatives over which our test based on the L_2 distance has a prescribed power.

This work was supported by grants from Région Ile-de-France.

Contents

1 Introduction **3**
- 1.1 Main results . . . 4
- 1.2 Three applications of the main results . . . 5
- 1.3 Outline . . . 5

2 Assumptions and notations **6**
- 2.1 Uniform ergodicity . . . 6
- 2.2 Upper-bounded Markov kernel . . . 6
- 2.3 Exponential integrability of the regeneration time . . . 6
- 2.4 π-canonical and bounded kernels . . . 8
- 2.5 Additional technical assumption . . . 8
- 2.6 Examples of Markov chains satisfying the Assumptions . . . 9

3 Three applications **10**
- 3.1 Estimation of spectra of signed integral operator with MCMC algorithms . . . 10
- 3.2 Online Learning with Pairwise Loss Functions . . . 13
- 3.3 Adaptive goodness-of-fit tests in a density model . . . 17

4 Proof of Theorem 1 **22**
- 4.1 Concentration of the first term of the decomposition of the U-statistic . . . 23
- 4.2 Reasoning by descending induction with a logarithmic depth . . . 34
- 4.3 Bounding the remaining statistic with uniform ergodicity . . . 36

5 Proof of Theorem 2 **39**

6 Deviation inequality for the spectrum of signed integral operators **39**

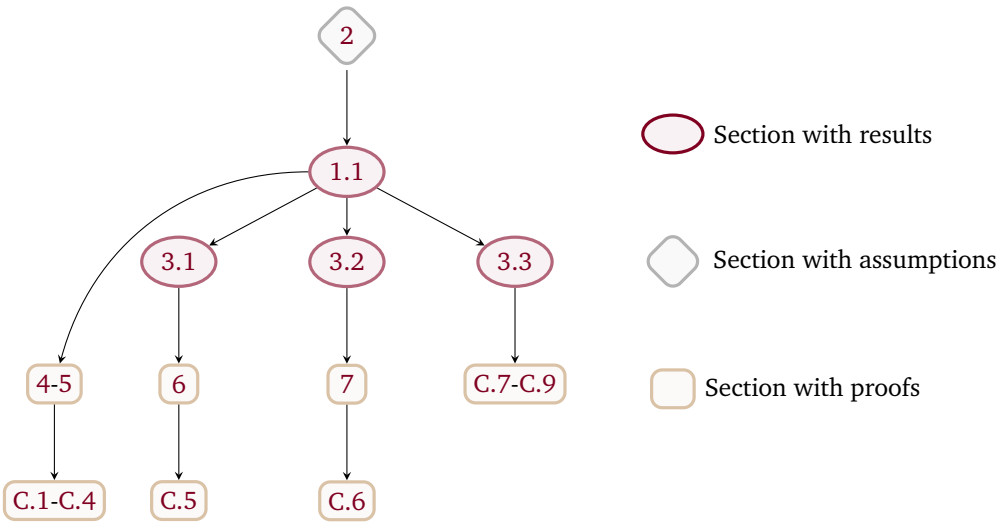
7 Proofs of Theorems 4 and 5 **42**
- 7.1 Proof of Theorem 4 . . . 42
- 7.2 Proof of Theorem 5 . . . 46

A Definitions and properties for Markov chains **53**

B Connections with the literature **54**

C Additional proofs **55**

We summarize the structure of our paper with the following diagram.



1 Introduction

Concentration of measure has been intensely studied during the last decades since it finds application in large span of topics such as model selection (see [46] and [42]), statistical learning (see [14]), online learning (see [63]) or random graphs (see [17] and [16]). Important contributions in this field are those concerning U-statistics. A U-statistic of order m is a sum of the form

$$\sum_{1\le i_1<\cdots<i_m\le n} h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m}),$$

where $X_1,\dots,X_n$ are independent random variables taking values in a measurable space (E,Σ) and with respective laws P_i and where $h_{i_1,\dots,i_m}$ are measurable functions of m variables $h_{i_1,\dots,i_m}: E^m \to \mathbb{R}$.

One important exponential inequality for U-statistics was provided by [4] using a Rademacher chaos approach. Their result holds for bounded and canonical kernels, namely satisfying for all $1 \le i_1 < \cdots < i_m \le n$ and for all $x_1,\dots,x_m \in E$,

$$\left\|h_{i_1,\dots,i_m}\right\|_\infty < \infty \quad \text{and} \quad \forall j \in [1,n],\ \mathbb{E}_{X_j}\left[h_{i_1,\dots,i_m}(x_1,\dots,x_{j-1},X_j,x_{j+1},\dots,x_m)\right] = 0\,.$$

Houdré and Reynaud-Bouret in [35] improved the constants in the exponential inequality of [4] for U-statistics of order two. If the kernels are unbounded but if we assume that the random variables $h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m})$ have sufficiently light tails (e.g. sub-Gaussian), Giné, Latala and Zinn in [30] proved that exponential inequality can still be obtained for the associated U-statistics. It is known that with heavy-tailed distribution for $h_{i_1,\dots,i_m}(X_{i_1},\dots,X_{i_m})$ we cannot expect to get exponential inequalities anymore. Nevertheless working with kernels that have finite p-th moment for some $p \in (1,2]$, Joly and Lugosi in [38] construct an estimator of the mean of the U-process using the median-of-means technique that performs as well as the classical U-statistic with bounded kernels.

All the above mentioned results consider that the random variables $(X_i)_{i\ge1}$ are independent. This condition can be prohibitive for practical applications since modelization of real phenomena often involves some dependence structure. The simplest and the most widely used tool to incorporate such dependence is Markov chain. One can give the example of Reinforcement Learning (see [60]) or Biology (see [59]). Recent works provide extensions of the classical concentration results to the Markovian settings as [22, 37, 50, 1, 14]. However, there are only few results for the non-asymptotic behaviour of tails of U-statistics in a dependent framework. The first results were provided in [9] and [34] where exponential inequalities for U-statistics of order $m \ge 2$ of time series under mixing conditions are proved. Those works were improved by [56] where a Bernstein's type inequality for V and U statistics is provided under conditions on the time dependent process that are easier to check in practice. Their result holds for geometrically α-mixing stationary sequences which includes in particular geometrically (and hence uniformly) ergodic Markov chains (see [39, p.6]). If [56] has the advantage of considering U-statistics of arbitrary order $m \ge 2$, their result holds only for state space like $\mathbb{R}^d$ with $d \ge 1$. Moreover, they consider a unique kernel h (i.e. $h_{i_1,\dots,i_m} = h$ for any $i_1,\dots,i_m$) which is assumed to be symmetric continuous, integrable and that satisfies some smoothness assumption.

On the contrary, our result holds for general state space and possibly different kernels that are not assumed symmetric or smooth. We rather work with bounded kernels that are π-canonical. This latter notion was first introduced in [27] who proved a variance inequality for U-statistics of ergodic Markov chains. In Section B, we give a concrete connection between our result and [56].

1.1 Main results

We consider a Markov chain $(X_i)_{i\geq 1}$ with transition kernel $P : E \times E \to \mathbb{R}$ taking values in a measurable space (E, Σ), and we introduce bounded functions $h_{i,j} : E^2 \to \mathbb{R}$. Our assumptions on the Markov chain $(X_i)_{i\geq 1}$ are fully-provided in Section 2 and include in particular the uniform ergodicity of the chain and an "upper-bounded" transition kernel P (see Assumption 2). The invariant distribution of the chain $(X_i)_{i\geq 1}$ will be denoted π. We further assume that kernel functions are π-canonical, namely

$$\forall i, j \in [n], \quad \forall x, y \in E, \quad \mathbb{E}_\pi h_{i,j}(X, x) = \mathbb{E}_\pi h_{i,j}(X, y),$$

and we denote this common expectation $\mathbb{E}_\pi h_{i,j}(X, \cdot)$.

Under those assumptions, Theorem 1 gives an exponential inequality for the U-statistic

$$U_{\text{stat}}(n) = \sum_{1\leq i<j\leq n} \big(h_{i,j}(X_i, X_j) - \mathbb{E}_\pi[h_{i,j}(X, \cdot)]\big).$$

The proof of Theorem 1 can be found in Section 4.

Theorem 1 *Let $n \geq 2$. We suppose Assumptions 1, 2, 3 and 4 described in Section 2. Then, there exist two constants $\beta, \kappa > 0$ such that for any $u > 0$ it holds with probability at least $1 - \beta e^{-u}\log(n)$,*

$$U_{\text{stat}}(n) \leq \kappa \log(n)\Big([An]\sqrt{u} + \big[A + B\sqrt{n}\big]u + \big[2A\sqrt{n}\big]u^{3/2} + A\big[u^2 + n\big]\Big),$$

where

$$A := 2\max_{i,j} \|h_{i,j}\|_\infty, \quad \text{and} \quad B^2 := \max\left[\max_i \left\| \sum_{j=i+1}^{n} \mathbb{E}_{X\sim\pi}\big[p_{i,j}^2(\cdot, X)\big] \right\|_\infty , \max_j \left\| \sum_{i=1}^{j-1} \mathbb{E}_{X\sim\pi}[p_{i,j}^2(X, \cdot)] \right\|_\infty \right],$$

and for all $i, j \in [n]$, $p_{i,j}(X_i, X_j) := h_{i,j}(X_i, X_j) - \mathbb{E}_{X'\sim\pi}[h_{i,j}(X_i, X')]$. The constant $\kappa > 0$ only depends on constants related to the Markov chain $(X_i)_{i\geq 1}$, namely δ_M, $\|T_1\|_{\psi_1}$, $\|T_2\|_{\psi_1}$, L, m and ρ. The constant $\beta > 0$ only depends on ρ. See Section 2 for the definitions of those constants.

Note that the kernels $h_{i,j}$ do not need to be symmetric and that we do not consider any assumption on the initial measure of the Markov chain $(X_i)_{i\geq 1}$ if the kernels $h_{i,j}$ do not depend on i (see Assumption 4). Let us mention that depending on whether Assumption 4.(i) or Assumption 4.(ii) holds, we do not obtain exactly the same bounds and we take the maximum of both results in Theorem 1. By bounding coarsely the constant B in the proof of Theorem 1, we show in Section 5 that under milder conditions the following result holds.

Theorem 2 *Let $n \geq 2$. We suppose Assumptions 1, 2.(i), 3 and 4 described in Section 2. Then there exist constants $\beta, \kappa > 0$ such that for any $u \geq 1$, it holds with probability at least $1 - \beta e^{-u}\log n$,*

$$\frac{2}{n(n-1)} U_{\text{stat}}(n) \leq \kappa \max_{i,j} \|h_{i,j}\|_\infty \log n \left\{ \frac{u}{n} + \left[\frac{u}{n}\right]^2 \right\},$$

where the constant $\kappa > 0$ only depends on constants related to the Markov chain $(X_i)_{i\geq 1}$, namely δ_M, $\|T_1\|_{\psi_1}$, $\|T_2\|_{\psi_1}$, L, m and ρ. The constant $\beta > 0$ only depends on ρ.

Note that in Theorem 2, we do not ask Assumption 2.(ii) to hold. Theorem 2 shows

$$\frac{2}{n(n-1)} U_{\text{stat}}(n) = \mathcal{O}_{\mathbb{P}}\left(\frac{\log(n)\log\log n}{n} \right),$$

where $\mathcal{O}_{\mathbb{P}}$ denotes stochastic boundedness. Up to a $\log(n)\log\log n$ multiplicative term, we uncover the optimal rate of Hoeffding's inequality for canonical U-statistics of order 2, see [38].

1.2 Three applications of the main results

Theorems 1 and 2 are cornerstones to uncover new results that we referred to as *applications* for brevity. These "applications" may be understood as new results part of active areas of research in Probability, Statistics and Machine Learning. Although Theorems 1 and 2 are key steps in these applications, the results are not a direct consequence of them, and extra analysis has been put to achieve these results. The three applications are the following.

- **Estimation of spectra of signed integral operator with MCMC algorithms** (Section 3.1)
  We study convergence of sequence of spectra of kernel matrices towards spectrum of integral operator. Previous important works may include [2] and, as far as we know, they all assume that the kernel is of positive type. For the first time, this paper proves a non-asymptotic result of convergence of spectra for kernels that are not of positive-type (*i.e.,* giving an integral operator with positive and negative eigenvalues). We further prove that *independent Hastings algorithms* are valid sampling schemes to apply our result.

- **Online Learning with Pairwise Loss Functions** (Section 3.2)
  Inspired by ranking problems where one aims at comparing pairs of individuals, we study algorithm with pairwise loss functions. We assume that data is coming *on the fly*, which is referred to as online algorithm. To propose realistic scenario, we consider that data is coming following a Markovian dynamic (instead of considering an i.i.d. scheme). As far as we know, we are the first to study *online algorithms under Markov sampling schemes*. Our contribution is three-fold: we introduce a new *average paired empirical risk*, denoted $\mathscr{M}^n$, that can be computed in practice; we give non-asymptotic error bounds between $\mathscr{M}^n$ and the true average risk; and we build an hypothesis selection procedure that outputs an hypothesis achieving this average risk.

- **Adaptive goodness-of-fit tests in a density model** (Section 3.3)
  Several works have already proposed goodness-of-fit tests for the density of the stationary distribution of a sequence of dependent random variables. In [43], a test based on an L^2-type distance between the nonparametrically estimated conditional density and its model-based parametric counterpart is proposed. In [5] a Kolmogorov-type test is considered. [13] derive a test procedure for τ-mixing sequences using Stein discrepancy computed in a reproducing kernel Hilbert space. In all the above mentioned papers, asymptotic properties of the test statistic are derived but no non-asymptotic analysis of the methods is conducted. As far as we know, this paper is the first to provide a non-asymptotic condition on the classes of alternatives ensuring that the statistical test reaches a prescribed power working in a dependent framework.

1.3 Outline

In Section 2, we introduce some notations and we present in details the assumptions under which our concentration results from Theorems 1 and 2 hold.

In Section 3, we give some applications of our result. We start by providing a convergence result for the estimation of spectra of integral operators with MCMC algorithms (see Section 3.1). We show that independent Hastings algorithms satisfy under mild conditions the assumptions of Section 2 and we illustrate our result with the estimation of the spectra of some Mercer kernels. For the second application of our concentration inequality, we investigate the generalization performance of online algorithms with pairwise loss functions in a Markovian framework (see Section 3.2). We motivate the study of such problems and we provide an online-to-batch conversion result. In a third and final application, we propose a goodness-of-fit test for the density of the invariant density of a Markov chain (see Section 3.3). We give an explicit condition on the set of alternatives to ensure that the statistical test proposed reaches a prescribed power. The proofs of Theorems 1 and 2 are provided respectively in Sections 4 and 5, while the proofs related to the three applications from Section 3 are given in Section 6, Section 7 and Sections C.7-C.9 respectively.

A reminder of the useful definitions and properties of Markov chains on a general state space can be found in Section A.

2 Assumptions and notations

2.1 Uniform ergodicity

Assumption 1 *The Markov chain $(X_i)_{i\geq 1}$ is ψ-irreducible for some maximal irreducibility measure ψ on Σ (see [48, Section 4.2]). Moreover, there exist $\delta_m > 0$ and some integer $m \geq 1$ such that*

$$\forall x \in E,\ \forall A \in \Sigma, \quad \delta_m \mu(A) \leq P^m(x,A).$$

for some probability measure μ.

For the reader familiar with the theory of Markov chains, Assumption 1 states that the whole space E is a small set which is equivalent to the uniform ergodicity of the Markov chain $(X_i)_{i\geq 1}$ (see [48, Theorem 16.0.2]), namely there exist constants $0 < \rho < 1$ and $L > 0$ such that

$$\|P^n(x,\cdot) - \pi\|_{TV} \leq L\rho^n, \qquad \forall n \geq 0,\ \pi\text{—a.e } x \in E,$$

where π is the unique invariant distribution of the chain $(X_i)_{i\geq 1}$ and for any measure ω on (E,Σ), $\|\omega\|_{TV} := \sup_{A\in\Sigma} |\omega(A)|$ is the total variation norm of ω. From [26, section 2.3]), we also know that the Markov chain $(X_i)_{i\geq 1}$ admits a spectral gap $1-\lambda > 0$ with $\lambda \in [0,1)$ (thanks to uniform ergodicity). We refer to Section A.2 for a reminder on the spectral gap of Markov chains.

2.2 Upper-bounded Markov kernel

Assumption 2 can be read as a reverse Doeblin's condition and allows us to achieve a change of measure in expectations in our proof to work with i.i.d. random variables with distribution ν.

Assumption 2 *There exists $\delta_M > 0$ such that*

$$(i) \qquad \forall x \in E,\ \forall A \in \Sigma, \quad P(x,A) \leq \delta_M \nu(A),$$

for some probability measure ν. We may also assume that

$$(ii) \qquad \forall A \in \Sigma, \quad \int_{x\in E} \nu(x)P(x,A) = \nu(A).$$

If Assumption 1 and 2.(ii) hold, then $\nu = \pi$ is the unique (finite) invariant measure of the Markov chain. We recall that if Assumption 2.(ii) does not hold, one can still invoke Theorem 2 with possible counterpart of larger constants. For this reason we keep on purpose the notation ν in our proof to handle frameworks where Assumption 2.(ii) does not hold. Assumption 2.(i) has already been used in the literature (see [44, Section 4.2]) and was introduced in [19]. This condition can typically require the state space to be compact as highlighted in [44].

Let us describe another situation where Assumption 2.(i) holds. Consider that $(E, \|\cdot\|)$ is a normed space and that for all $x \in E$, $P(x,dy)$ has density $p(x,\cdot)$ with respect to some measure η on (E,Σ). We further assume that there exists an integrable function $u : E \to \mathbb{R}_+$ such that

$$\forall x,y \in E, \quad p(x,y) \leq u(y).$$

Then considering for ν the probability measure with density $u/\|u\|_1$ with respect to η and $\delta_M = \|u\|_1$, Assumption 2.(i) holds.

2.3 Exponential integrability of the regeneration time

We introduce some additional notations which will be useful to apply Talagrand concentration result from [55]. Note that this section is inspired from [1] and [48, Theorem 17.3.1]. We assume that Assumption 1 is satisfied and we extend the Markov chain $(X_i)_{i\geq 1}$ to a new (so called *split*) chain $(\widetilde{X}_n, R_n) \in E \times \{0,1\}$ (see Section A.3 for a construction of the split chain), satisfying the following properties.

- $(\widetilde{X}_n)_n$ is again a Markov chain with transition kernel P with the same initial distribution as $(X_n)_n$. We recall that π is the invariant distribution on the E.
- if we define $T_1 = \inf\{n > 0 : R_{nm} = 1\}$,
$$T_{i+1} = \inf\{n > 0 : R_{(T_1+\cdots+T_i+n)m} = 1\},$$
then $T_1, T_2, \dots$ are well defined and independent. Moreover $T_2, T_3, \dots$ are i.i.d.
- if we define $S_i = T_1 + \cdots + T_i$, then the "blocks"
$$Y_0 = (\widetilde{X}_1, \dots, \widetilde{X}_{mT_1+m-1}), \quad \text{and} \quad Y_i = (\widetilde{X}_{m(S_i+1)}, \dots, \widetilde{X}_{m(S_{i+1}+1)-1}),\ i > 0,$$
form a one-dependent sequence (i.e. for all i, $\sigma((Y_j)_{j<i})$ and $\sigma((Y_j)_{j>i})$ are independent). Moreover, the sequence $Y_1, Y_2, \dots$ is stationary and if $m = 1$ the variables $Y_0, Y_1, \dots$ are independent. In consequence, for any measurable space $(S, \mathscr{B})$ and measurable functions $f : S \to \mathbb{R}$, the variables
$$Z_i = Z_i(f) = \sum_{j=m(S_i+1)}^{m(S_{i+1}+1)-1} f(\widetilde{X}_j), \quad i \geq 1,$$
constitute a one-dependent sequence (an i.i.d. sequence if $m = 1$). Additionally, if f is π-integrable (recall that π is the unique stationary measure for the chain), then
$$\mathbb{E}[Z_i] = \delta_m^{-1} m \int f \, d\pi.$$
- the distribution of T_1 depends only on π, P, δ_m, μ, whereas the law of T_2 only on P, δ_m and μ.

Remark Let us highlight that $(\widetilde{X}_n)_n$ is a Markov chain with transition kernel P and same initial distribution as $(X_n)_n$. Hence for our purposes of estimating the tail probabilities, we will identify $(X_n)_n$ and $(\widetilde{X}_n)_n$.

To derive a concentration inequality, we use the exponential integrability of the regeneration times which is ensured if the chain is uniformly ergodic as stated by the following Proposition.

Proposition 1 *If Assumption 1 holds, then*
$$\|T_1\|_{\psi_1} < \infty \quad \textit{and} \quad \|T_2\|_{\psi_1} < \infty, \tag{1}$$
where $\|\cdot\|_{\psi_1}$ *is the* 1*-Orlicz norm introduced in Definition 1. We denote* $\tau := \max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1})$.

Proof of Proposition 1.
Since the split chain has the same distribution as the original Markov chain, we get that $(\tilde{X}_i)_i$ is ψ-irreducible for some measure ψ and uniformly ergodic. From [48, Theorem 16.0.2], Assumption 1 ensures that for every measurable set $A \subset E \times \{0, 1\}$ such that $\psi(A) > 0$, there exists some $\kappa_A > 1$ such that
$$\sup_x \mathbb{E}[\kappa_A^{\tau_A} | \tilde{X}_1 = x] < \infty,$$
where $\tau_A := \inf\{n \geq 1 \ : \ \tilde{X}_n \in A\}$ is the first hitting time of the set A. Let us recall that T_1 and T_2 are defined as hitting times of the atom of the split chain $E \times \{1\}$ which is accessible (i.e. the atom has a positive ψ-measure). Hence, there exist $C > 0$ and $\kappa > 1$ such that,
$$\sup_x \mathbb{E}[\kappa^{\tau_{E\times\{1\}}} | \tilde{X}_1 = x] = \sup_x \mathbb{E}[\exp(\tau_{E\times\{1\}} \log(\kappa)) | \tilde{X}_1 = x] \leq C.$$
Considering $k \geq 1$ such that $C^{1/k} \leq 2$, a straight forward application of Jensen inequality gives that $\max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq k/\log(\kappa)$.
■

Definition 1 *For* $\alpha > 0$*, define the function* $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ *with the formula* $\psi_\alpha(x) = \exp(x\alpha) - 1$*. Then for a random variable* X*, the* α*-Orlicz norm is given by*
$$\|X\|_{\psi_\alpha} = \inf\{\gamma > 0 : \mathbb{E}[\psi_\alpha(|X|/\gamma)] \leq 1\}.$$

2.4 π-canonical and bounded kernels

With Assumption 3, we introduce the notion of π-*canonical* kernel which is the counterpart of the canonical property from [29].

Assumption 3 *Let us denote $\mathscr{B}(\mathbb{R})$ the Borel algebra on $\mathbb{R}$. For all $i,j \in [n]$, we assume that $h_{i,j} : (E^2, \Sigma\otimes\Sigma) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is measurable and is π-canonical, namely*

$$\forall x, y \in E, \quad \mathbb{E}_\pi[h_{i,j}(X,x)] = \mathbb{E}_\pi[h_{i,j}(X,y)].$$

This common expectation will be denoted $\mathbb{E}_\pi[h_{i,j}(X,\cdot)]$.
Moreover, we assume that for all $i,j \in [n]$, $\|h_{i,j}\|_\infty < \infty$.

Remarks:

• A large span of kernels are π-canonical. This is the case of translation-invariant kernels which have been widely studied in the Machine Learning community. Another example of π-canonical kernel is a rotation invariant kernel when $E = \mathbb{S}^{d-1} := \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ with π also rotation invariant (see [17] or [16]).

• The notion of π-canonical kernels is the counterpart of canonical kernels in the i.i.d. framework (see for example [35]). Note that we are not the first to introduce the notion of π-canonical kernels working with Markov chains. In [27], Fort and al. provide a variance inequality for U-statistics whose underlying sequence of random variables is an ergodic Markov Chain. Their results holds for π-canonical kernels as stated with [27, Assumption A2].

• Note that if the kernels $h_{i,j}$ are not π-canonical, the U-statistic decomposes into a linear term and a π-canonical U-statistic. This is called the *Hoeffding decomposition* (see [29, p.176]) and takes the following form

$$\begin{aligned}
&\sum_{i\neq j}\big(h_{i,j}(X_i,X_j) - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}[h_{i,j}(X,Y)]\big)\\
&= \sum_{i\neq j}\widetilde{h}_{i,j}(X_i,X_j) - \mathbb{E}_\pi\big[\widetilde{h}_{i,j}(X,\cdot)\big] + \sum_{i\neq j}\Big(\mathbb{E}_{X\sim\pi}\big[h_{i,j}(X,X_j)\big] - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\big[h_{i,j}(X,Y)\big]\Big)\\
&\quad + \sum_{i\neq j}\Big(\mathbb{E}_{X\sim\pi}\big[h_{i,j}(X_i,X)\big] - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}\big[h_{i,j}(X,Y)\big]\Big),
\end{aligned}$$

where for all j, the kernel $\widetilde{h}_{i,j}$ is π-canonical with

$$\forall x, y \in E, \quad \widetilde{h}_{i,j}(x,y) = h_{i,j}(x,y) - \mathbb{E}_{X\sim\pi}\big[h_{i,j}(x,X)\big] - \mathbb{E}_{X\sim\pi}\big[h_{i,j}(X,y)\big].$$

We will use this method several times in the proofs of our applications (for example in (24)).

2.5 Additional technical assumption

In the case where the kernels $h_{i,j}$ depend on both i and j, we need Assumption 4.(ii) to prove Theorem 1. Assumption 4.(ii) is a mild condition on the initial distribution of the Markov chain that is used when we apply Bernstein's inequality for Markov chains from Proposition 5.

Assumption 4 *At least one of the following conditions holds.*

(i) *For all $i,j \in [n]$, $h_{i,j} \equiv h_{1,j}$, i.e. the kernel function $h_{i,j}$ does not depend on i.*

(ii) *The initial distribution of the Markov chain $(X_i)_{i\geq 1}$, denoted χ, is absolutely continuous with respect to the invariant measure π and its density, denoted by $\frac{d\chi}{d\pi}$, has finite p-moment for some $p \in (1,\infty]$, i.e*

$$\infty > \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} := \begin{cases} \left[\int \left|\frac{d\chi}{d\pi}\right|^p d\pi\right]^{1/p} & \text{if } p < \infty,\\ \operatorname{ess\,sup}\left|\frac{d\chi}{d\pi}\right| & \text{if } p = \infty. \end{cases}$$

In the following, we will denote $q = \frac{p}{p-1} \in [1,\infty)$ (with $q = 1$ if $p = +\infty$) which satisfies $\frac{1}{p} + \frac{1}{q} = 1$.

2.6 Examples of Markov chains satisfying the Assumptions

Example 1: Finite state space. For Markov chains with finite state space, Assumption 2.(i) holds trivially. Hence, in such framework the result of Theorem 2 holds for any uniformly ergodic Markov chain. In particular, this is true for any aperiodic and irreducible Markov chains using [7, Lemma 7.3.(ii)].

Example 2: AR(1) process. Let us consider the process $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^k$ defined by

$$X_0 \in \mathbb{R}^k \text{ and for all } n \in \mathbb{N}, \quad X_{n+1} = H(X_n) + Z_n,$$

where $(Z_n)_{n\in\mathbb{N}}$ are i.i.d random variables in $\mathbb{R}^k$ and $H : \mathbb{R}^k \to \mathbb{R}^k$ is an application. Such a process is called an auto-regressive process of order 1, noted AR(1). We show that under mild conditions, our result can be applied to AR(1) processes. Before providing a result from [21] giving conditions ensuring the uniform ergodicity of an AR(1) process, let us introduce some notations.

Notations:

We consider $k, d \in \mathbb{N}^* := \mathbb{N}\backslash\{0\}$ and we denote $\|\cdot\|_k$ (resp. $\|\cdot\|_d$) the euclidean norm on $\mathbb{R}^k$ (resp. on $\mathbb{R}^d$). We need the preliminary notations.

- λ_{Leb} is the Lebesgue measure on $\mathbb{R}^k$ and $L^2(\mathbb{R}^k, \lambda_{Leb})$ is the space of square integrable functions from $\mathbb{R}^k$ to $\mathbb{R}$.
- If v is linear map from $\mathbb{R}^k$ to $\mathbb{R}^d$, we denote $\|v\|$ its norm defined by

$$\|v\| := \sup_{\|x\|_k=1} \|v(x)\|_d.$$

If some function $G : \mathbb{R}^k \to \mathbb{R}^d$ is differentiable on $\mathbb{R}^k$, we define

$$\|dG\|_2 := \sqrt{\int_{\mathbb{R}^k} \|dG(x)\|^2 d\lambda_{Leb}(x)}.$$

We can now state the following result from [21].

Proposition 2 *Let us consider $\mathbb{R}^k$ endowed with its Borel sigma-algebra. Suppose that the random variables Z_n are i.i.d. with a distribution equivalent to the Lebesgue measure λ_{Leb} on $\mathbb{R}^k$ and with density f_Z with respect λ_{Leb}. Assume that*

- *H is bounded and continuously differentiable with $\|H\|_\infty + \|dH\|_2 < \infty$.*
- *f_Z is continuously differentiable and $\|f_Z\|_\infty + \|f_Z\|_2 + \|df_Z\|_2 < \infty$.*

Then, the Markov chain $(X_n)_n$ satisfying $X_{n+1} = H(X_n) + Z_n$ is uniformly ergodic.

We keep the assumptions of Proposition 2 and we consider some $B > 0$ such that $\|H\|_\infty \leq B$. Assuming furter that $y \mapsto \sup_{z\in[-B,B]} f_Z(y-z)$ is integrable on E with respect to λ_{Leb}, we get that Assumption 2.(i) holds (see the remark following Assumption 2). The previous condition on f_Z is for example satisfied for Gaussian distributions. We deduce that Theorem 2 can be applied in such settings that can typically be found in nonlinear filtering problem (see [19, Section 4]).

Example 3: ARCH process. Let us consider $E = \mathbb{R}$. The ARCH model is

$$X_{n+1} = H(X_n) + G(X_n)Z_{n+1},$$

where H and G are continuous functions, and $(Z_n)_n$ are i.i.d. centered normal random variables with variance $\sigma^2 > 0$. Assuming that $\inf_x |G(x)| \geq a > 0$, we know that the Markov chain (X_n) is irreducible and aperiodic (see [3, Lemma 1]). Assuming further that $\|H\|_\infty \leq B < \infty$ and that $\|G\|_\infty \leq A$, we can show that Assumptions 1 and 2.(i) hold. Let us first remark that the transition kernel P of the Markov

chain $(X_n)_n$ is such that for any $x \in \mathbb{R}$, $P(x, dy)$ has density $p(x, \cdot)$ with respect to the Lebesgue measure with

$$p(x,y) = \left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{(y-H(x))^2}{2\sigma^2 G(x)^2}\right).$$

We deduce that for any $x, y \in \mathbb{R}$ we have

$$p(x,y) \geq \left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{(y-H(x))^2}{2\sigma^2 a^2}\right) \geq g_m(y),$$

where

$$g_m(y) = \left(2\pi\sigma^2\right)^{-1} \times \begin{cases} \exp\left(-\frac{(y-B)^2}{2\sigma^2 a^2}\right) & \text{if } y < -B \\ \exp\left(-\frac{2B^2}{\sigma^2 a^2}\right) & \text{if } |y| \leq B \\ \exp\left(-\frac{(y+B)^2}{2\sigma^2 a^2}\right) & \text{if } y > B \end{cases}$$

With a similar approach, we get

$$p(x,y) \leq \left(2\pi\sigma^2\right)^{-1} \exp\left(-\frac{(y-H(x))^2}{2\sigma^2 A^2}\right) \leq g_M(y),$$

where

$$g_M(y) = \left(2\pi\sigma^2\right)^{-1} \times \begin{cases} \exp\left(-\frac{(y+B)^2}{2\sigma^2 a^2}\right) & \text{if } y < -B \\ 1 & \text{if } |y| \leq B \\ \exp\left(-\frac{(y-B)^2}{2\sigma^2 a^2}\right) & \text{if } y > B \end{cases}$$

We deduce that considering $\delta_m = \|g_m\|_1$, $\delta_M = \|g_M\|_1$ and μ (resp. ν) with density $g_m/\|g_m\|_1$ (resp. $g_M/\|g_M\|_1$) with respect to the Lebesgue measure on $\mathbb{R}$, Assumptions 1 and 2.(i) hold.

3 Three applications

3.1 Estimation of spectra of signed integral operator with MCMC algorithms

3.1.1 MCMC estimation of spectra of signed integral operators

Let us consider $(X_n)_{n\geq 1}$ a Markov chain on E satisfying the assumptions of Theorem 2 with invariant distribution π, and some symmetric kernel $h : E \times E \to \mathbb{R}$, square integrable with respect to $\pi \otimes \pi$. We can associate to h the kernel linear operator $\mathbf{H}$ defined by

$$\mathbf{H}f(x) := \int_E h(x,y) f(y) d\pi(y).$$

This is a Hilbert Schmidt operator on $L^2(\pi)$ and thus it has a real spectrum consisting of a square summable sequence of eigenvalues. In the following, we will denote the eigenvalues of $\mathbf{H}$ by $\lambda(\mathbf{H}) := (\lambda_1, \lambda_2, \dots)$. For some $n \in \mathbb{N}^*$, we consider

$$\widetilde{\mathbf{H}}_n := \frac{1}{n}\left(h(X_i, X_j)\right)_{1\leq i,j\leq n} \text{ and } \mathbf{H}_n := \frac{1}{n}\left((1-\delta_{i,j})h(X_i, X_j)\right)_{1\leq i,j\leq n},$$

with respective eigenvalues $\lambda(\widetilde{\mathbf{H}}_n)$ and $\lambda(\mathbf{H}_n)$. We introduce the rearrangement distance δ_2 with Definition 2 that will be useful to compare two spectra.

Definition 2 *Given two sequences x, y of reals – completing finite sequences by zeros – such that*

$$\sum_i x_i^2 + y_i^2 < \infty,$$

we define the ℓ_2 rearrangement distance $\delta_2(x,y)$ as

$$\delta_2^2(x,y) := \inf_{\sigma\in\mathfrak{S}} \sum_i (x_i - y_{\sigma(i)})^2,$$

where $\mathfrak{S}$ is the set of permutations with finite support.

Theorem 3 gives conditions ensuring that the spectrum of $\mathbf{H}_n$ (resp. $\widetilde{\mathbf{H}}_n$) converges towards the spectrum of the integral operator $\mathbf{H}$ as $n \to \infty$. The proof of Theorem 3 is postponed to Section 6.

Theorem 3 *We consider a Markov chain $(X_i)_{i\geq 1}$ on E satisfying Assumptions 1 and 2.(i) described in Section 2 with invariant distribution π. We assume that*

- *the integral operator $\mathbf{H}$ is trace-class, i.e. $S := \sum_{r\geq 1} |\lambda_r| < \infty$. We set $\Lambda := \sup_{r\geq 1} |\lambda_r| < \infty$.*
- *there exist continuous functions $\varphi_r : E \to \mathbb{R}$, $r \in I$ (where $I = \mathbb{N}$ or $I = 1,\dots,N$) that form an orthonormal basis of $L^2(\pi)$ such that it holds pointwise*

$$h(x,y) = \sum_{r\in I} \lambda_r \varphi_r(x)\varphi_r(y),$$

with $\sup_{r\geq 1} \|\varphi_r\|_\infty \leq \Upsilon$.

Then there exists a constant $C > 0$ such for any $t > 0$, it holds,

$$\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \geq \frac{C\log n}{n} + 2 \sum_{i>\lceil n^{1/4}\rceil, i\in I} \lambda_i^2 + t \right)$$
$$\leq 32\sqrt{n}\exp\left(-\mathscr{C}\min\left(nt^2, \sqrt{n}t\right)\right) + \beta\log(n)\exp\left(-\frac{n}{\log n}\min\left(\mathscr{B}t, (\mathscr{B}t)^{1/2}\right)\right).$$

where for some universal constant $K > 0$, we have $\mathscr{B} = \left(K\Upsilon^2\kappa S\right)^{-1}$, $\mathscr{C} = \left(KS^2\Upsilon^4\right)$. $\kappa > 0$ and $\beta > 0$ are constants depending on the Markov chain.

Remark In [2], Adamczak and Bednorz studied the convergence properties of MCMC methods to estimate the spectrum of integral operators with bounded *positive* kernels. They show a sub-exponential tail behavior for the δ_2 distance between the spectrum of $\mathbf{H}$ and the one of the random matrix $\mathbf{H}_n$. If their result holds for geometrically ergodic Markov chains, they assume that the eigenvalues of $\mathbf{H}$ are non-negative. Hence, working with stronger conditions on the Markov chain $(X_i)_i$, Theorem 3 proves a new concentration inequality for the δ_2 distance between $\lambda(\mathbf{H})$ and $\lambda(\mathbf{H}_n)$ that holds for **arbitrary signs of the eigenvalues** of $\mathbf{H}$.

3.1.2 Admissible sampling schemes: Independent Hastings algorithm

One can use the previous result to estimate the spectrum of the integral operator $\mathbf{H}$ using MCMC methods. To do so, we need to make sure that the Markov chain used for the MCMC method satisfies the conditions of Theorem 2. It is for example well known that Metropolis random walks on $\mathbb{R}$ are not uniformly ergodic (see [48]). In the following, we show that an independent Hastings algorithm can be used on bounded state space to generate a uniformly ergodic chain with the desired invariant distribution. We provide another application of our result estimating the spectrum of some Mercer kernels on the d-dimensional sphere.

Independent Hastings algorithm on bounded state space. Let us consider $E \subset \mathbb{R}^k$ a bounded subset of $\mathbb{R}^k$ equipped with the Borel σ-algebra $\mathscr{B}(E)$. We consider a density π which is only known up to a factor and a probability density q with respect to the Lebesgue measure λ_{Leb} on E, satisfying $\pi(y), q(y) > 0$

for all $y \in E$. In the independent Hastings algorithm, a candidate transition generated according to the law $q\lambda_{Leb}$ is then accepted with probability $\alpha(x,y)$ given by

$$\alpha(x,y) = \min\left(1, \frac{\pi(y)q(x)}{\pi(x)q(y)}\right).$$

With an approach similar to Theorem 2.1 from [47], Proposition 3 shows that under some conditions on the densities π and q, the independent Hastings algorithm satisfies the Assumptions 1 and 2.(i).

Proposition 3 *Let us assume that* $\sup_{x\in E} q(x) < \infty$ *and that there exists* $\beta > 0$ *such that*

$$\frac{q(y)}{\pi(y)} > \beta, \quad \forall y \in E.$$

Then, the independent Hastings algorithm satisfies the Assumptions 1 and 2.(i).

Proof of Proposition 3.
We denote P the transition kernel of the Markov chain generated with the independent Hastings algorithm. For any $x \in E$, the density with respect to λ_{Leb} of the absolutely continuous part of $P(x,dy)$ is $p(x,\cdot) = q(\cdot)\alpha(x,\cdot)$, while the singular part is given by $\mathbb{1}_x(\cdot)\left(\int_{z\in E} q(z)\alpha(x,z)d\lambda_{Leb}(z)\right)$. For fixed $x \in E$, we have either $\alpha(x,y) = 1$ in which case $p(x,y) = q(y) \geq \beta\pi(y)$, or else

$$p(x,y) = q(y)\frac{\pi(y)q(x)}{\pi(x)q(y)} = q(x)\frac{\pi(y)}{\pi(x)} \geq \beta\pi(y).$$

We deduce that for any $x \in E$, it holds

$$P(x,A) \geq \beta \int_{y\in A} \pi(y)d\lambda_{Leb}(y),$$

which proves that the chain is uniformly ergodic (see Definition 8). Hence Assumption 1 is satisfied. Assumption 2.(i) trivially holds since E is bounded and $\sup_{y\in E} q(y) < \infty$. ■

3.1.3 Estimation of the spectrum of Mercer kernels

In this example, we illustrate Theorem 3 by computing the eigenvalues of an integral operator naturally associated with a Mercer kernel using a MCMC algorithm. A function $K : E \times E \to \mathbb{R}$ is called a Mercer kernel if E is a compact metric space and $K : E \times E \to \mathbb{R}$ is a continuous symmetric and positive definite function. It is well known that if K is a Mercer kernel, then the integral operator L_K associated with K is a compact and bounded linear operator, self-adjoint and semi-definite positive. The spectral theorem implies that if K is a Mercer kernel, then there is a complete orthonormal system $(\varphi_1, \varphi_2, \dots)$ of eigenvectors of L_K. The eigenvalues $(\lambda_1, \lambda_2, \dots)$ are real and non-negative. The Mercer Theorem (see for instance see [12, Theorem 4.49]) shows that the eigen-structure of L_K can be used to get a representation of the Mercer kernel K as a sum of a convergent sequence of product functions for the uniform norm.

To illustrate our purpose, we consider the d-dimensional sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. We consider a positive definite kernel on $\mathbb{S}^{d-1}$ defined by $\forall x, y \in \mathbb{S}^{d-1}, \quad K(x,y) = \psi(x^\top y)$ where $\psi : [-1,1] \to \mathbb{R}$ is continuous. From the Funk-Hecke Theorem (see e.g [49, p.30]), we know that the eigenvalues of the Mercer kernel K are

$$\lambda_k = \omega(\mathbb{S}^{d-2}) \int_{-1}^{1} \psi(t)P_k(d;t)\left(1-t^2\right)^{\frac{d-3}{2}} dt, \tag{2}$$

where $P_k(d;t)$ is the Legendre polynomial of degree k in dimension d. For any $k \in \mathbb{N}$, the multiplicity of the eigenvalue λ_k is the dimension of the space of spherical harmonics of degree k. In Figure 1, we plot the non-zero eigenvalues using function $\psi : t \mapsto (1+t)^2$. We plot both the true eigenvalues and the ones computed using a MCMC approach. To build the Markov chain $(X_i)_{i\geq 1}$, we start by sampling randomly X_1 on $\mathbb{S}^{d-1}$. Then, for any $i \in \{2, \dots, n\}$, we sample

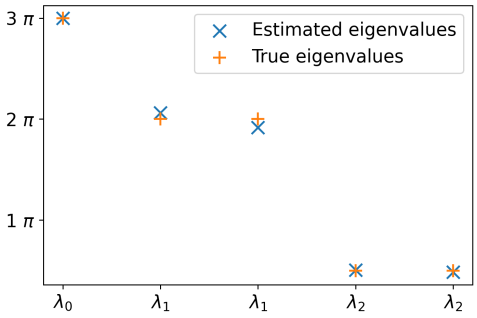
- a unit vector $Y_i \in \mathbb{S}^{d-1}$ uniformly, orthogonal to X_{i-1}.
- a real $r_i \in [-1,1]$ encoding the distance between X_{i-1} and X_i. r_i is sampled from a distribution $f_{\mathscr{L}} : [-1,1] \to [0,1]$. We take $f_{\mathscr{L}}$ proportional to $r \mapsto f_{(5,1)}(\frac{r+2}{4})$ where $f_{(5,1)}$ is the pdf of the Beta distribution with parameter $(5,1)$.

then X_i is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1-r_i^2} \times Y_i \,.$$

Since $\min_{r\in[-1,1]} f_{\mathscr{L}}(r) > 0$ and $\|f_{\mathscr{L}}\|_\infty < \infty$, Assumptions 1 and 2.(i) hold (see for example [16]).

Figure 1: Consider function $\psi : t \mapsto (1+t)^2$, $d = 2$ and $n = 1000$. The true eigenvalues can be computed using (2), but in this case, we know the exact values of the three non-zero eigenvalues namely $\lambda_0 = 3\pi$, $\lambda_1 = 2\pi$ and $\lambda_2 = \pi/2$. Their respective multiplicities are 1, 2 and 2. The estimated eigenvalues are the eigenvalues of the matrix $\mathbf{H}_n = \frac{1}{n}\left((1-\delta_{i,j})\psi(X_i^\top X_j)\right)_{1\le i,j\le n}$ where the n points $X_1, X_2, \ldots, X_n$ are sampled on the Euclidean sphere $\mathbb{S}^{d-1}$ using a Markovian dynamic.



3.2 Online Learning with Pairwise Loss Functions

3.2.1 Brief introduction to online learning and motivations

Presentation of the traditional online learning setting Online learning is an active field of research in Machine Learning in which data becomes available in a sequential order and is used to update the best predictor for future data at each step. This method aims at learning some function $f : E \to \mathscr{Y}$ where E is the space of inputs and $\mathscr{Y}$ is the space of outputs. At each time step t, we observe a new example $(x_t, y_t) \in E \times \mathscr{Y}$. Traditionally, the random variables (x_t, y_t) are supposed i.i.d. with common joint probability distribution $(x,y) \mapsto p(x,y)$ on $E \times \mathscr{Y}$. In this setting, the loss function is given as $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$, such that $\ell(f(x), y)$ measures the difference between the predicted value $f(x)$ and true value y. The goal is to select at each time step t a function $h_t : E \to \mathscr{Y}$ in a fixed set $\mathscr{H}$ based on the observed examples until time t (namely $(x_i, y_i)_{1\le i\le t}$) such that h_t has "small" risk $\mathscr{R}$ defined by

$$\mathscr{R}(h) = \mathbb{E}_{(X,Y)\sim p}\big[\ell(h(X), Y)\big],$$

where h is any measurable mapping from E to $\mathscr{Y}$.

Online learning is used when data is coming *on the fly* and we do not want to wait for the acquisition of the complete dataset to take a decision. In such cases, online learning algorithms allow to dynamically adapt to new patterns in the data.

Online learning with pairwise loss functions In some cases, the framework provided in the previous paragraph is not appropriated to solve the task at stake. Consider the example of ranking problems. The state space is E and there exists a function $f : E \to \mathbb{R}$ which assigns to each state $x \in E$ a label $f(x) \in \mathbb{R}$. f naturally defines a partial order on E. At each time step t, we observe an example $x_t \in E$ together with its label $f(x_t)$ and we suppose that the random variables $(x_t)_t$ are i.i.d. with common distribution p. Our goal is to learn the partial order of the items in E induced by the function f. More precisely, we consider a space $\mathscr{H} \subset \{h : E \times E \to \mathbb{R}\}$, called the set of hypotheses. An *ideal* hypothesis $h \in \mathscr{H}$ would satisfy

$$\forall x, u \in E, \quad f(x) \ge f(u) \Leftrightarrow (h(x,u) \ge 0 \text{ and } h(u,x) \le 0)\,.$$

We consider a loss function $\ell : \mathscr{H} \times E \times E \to \mathbb{R}$ such that $\ell(h,x,u)$ measures the ranking error induced by h and a typical choice is the 0-1 loss

$$\ell(h,x,u) = \mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}}.$$

U-statistics naturally arise in such settings as for example in [15] where Clémençon and al. study the consistency of the empirical risk minimizer of ranking problems using the theory of U-processes in an i.i.d. framework.

Example: `Bipartite ranking problems`

We describe the concrete problem of bipartite ranking. We consider that we have as input a training set of examples. Each example is described by some feature vector and is associated with a binary label. Typically one can consider that we have access to health data of an individual along time. We know at each time step her/his health status x_t and his label which is 0 if the individual is healthy and 1 if she/he is sick. In the bipartite ranking problem, we want to learn a `scorer` *which maps any feature vector describing the health status of the individual to a real number such that sick states have a higher score than healthy ones. Following the health status of individuals is time-consuming and we cannot afford to wait for the end of the data acquisition process to understand the relationship between the feature vector describing the health status of the individual and her/his sickness. In such settings where data is coming on the fly, online algorithms are common tools that allow to learn a scorer function along time. At each time step the scorer function is updated based on the new measurement provided.*

Using online learning with a Markovian dynamic Up to now, online algorithms have been widely studied in the i.i.d. framework. In this work, we aim at providing some theoretical results related to online learning methods with pairwise loss functions in a Markovian framework.

The theoretical analysis of Machine learning algorithms with an underlying Markovian distribution of the data has become a very active field of research. The first papers to study online learning with samples drawn from non-identical distributions were [57] and [58] where online learning for least square regression and off-line support vector machines are investigated. In [65], the generalization performance of the empirical risk minimization algorithm is studied with uniformly ergodic Markov chain samples. In [64], generalization performance bounds of online support vector machine classification learning algorithms with uniformly ergodic Markov chain samples are proved. Hence the analysis of online algorithms with dependent samples is recent and several works make the assumption that the sequence is a uniformly ergodic Markov chain.

With the upcoming application, we are the first - as far as we know - to study online algorithms with pairwise loss functions and Markov chain samples. Moreover, our result holds for any online algorithm and thus covers a large span of settings. We motivate the Markovian assumption on the example of the previous paragraph.

Example (continued): `Interest to consider online algorithms with Markovian dynamic`

The health status of the individual at time $n+1$ is not independent from the past and a simple way to model this time evolution would be to consider that it only depends on the last measured health status namely the feature vector x_n. This is a Markovian assumption on the sequence of observed health status of the individual.

We have explained why pairwise loss functions capture ranking problems and naturally arise in several Machine Learning problems such as metric learning or bipartite ranking (see for example [15]). We have shown the interest to provide a theoretical analysis of online learning learning with pairwise loss functions with a Markovian assumption on the distribution of the sequence of examples and this is the goal of the next section.

3.2.2 Online-to-batch conversion for pairwise loss functions with Markov chains

We consider a reversible Markov chain $(X_i)_{i\geq 1}$ with state space E satisfying Assumptions 1 and 2.(i), with invariant distribution π. We assume that we have a function $f : E \to \mathbb{R}$ which defines the ordering of the objects in E. We aim at finding a relevant approximation of the ordering of the objects in E by selecting a function h (called a *hypothesis* function) in a space $\mathscr{H}$ based on the observation of the random sequence $(X_i, f(X_i))_{1\leq i\leq n}$. To measure the performance of a given hypothesis $h : E \times E \to \mathbb{R}$, we use a pairwise loss function of the form $\ell(h,(X,f(X)),(U,f(U)))$. Typically, one could use the *misranking loss* defined by

$$\ell(h,x,u) = \mathbb{1}_{\{(f(x)-f(u))h(x,u)<0\}},$$

which is 1 if the examples are ranked in the wrong order and 0 otherwise. The goal of the learning problem is to find a hypothesis h which minimizes the *expected misranking risk*

$$\mathscr{R}(h) := \mathbb{E}_{(X,X')\sim\pi\otimes\pi}\Big[\ell(h,X,X')\Big].$$

We show that the investigation of the generalization performance of online algorithms with pairwise loss functions provided by [63] can be extended to a Markovian framework. Our contribution is two fold.

- Firstly, we prove that with high probability, the average risk of the sequence of hypotheses generated by an arbitrary online learner is bounded by some easily computable statistic.
- This first technical result is then used to show how we can extract a low risk hypothesis from a given sequence of hypotheses selected by an online learner. This is an *online-to-batch* conversion for pairwise loss functions with a Markovian assumption on the distribution of the observed states.

Given a sequence of hypotheses $(h_i)_{1\leq i\leq n} \in \mathscr{H}^n$ generated by any online algorithm, we define the *average paired empirical risk* $\mathscr{M}^n$ (see (3)) averaging the *paired empirical risks* M_t (see (4)) of hypotheses h_{t-b_n} when paired with X_t as follows

$$\mathscr{M}^n := \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} M_t, \tag{3}$$

$$\text{and} \quad M_t := \frac{1}{t-b_n}\sum_{i=1}^{t-b_n} \ell(h_{t-b_n},X_t,X_i), \tag{4}$$

where

$$c_n = \lceil c\times n\rceil \text{ for some } c\in(0,1) \quad \text{and} \quad b_n = \lfloor q\log(n)\rfloor, \tag{5}$$

for an arbitrarily chosen $q > \frac{1}{\log(1/\rho)}$ where ρ is a constant related to the uniform ergodicity of the Markov chain, see Definition 8.

M_t is the *paired empirical risk* of hypothesis h_{t-b_n} with X_t. It measures the performance of the hypothesis h_{t-b_n} on the example X_t when paired with examples seen before time $t-b_n$. $\mathscr{M}^n$ is the mean value of a proportion $1-c$ of these paired empirical risks. Hence the parameter $c\in(0,1)$ controls the proportion of hypotheses h_{t-b_n} whose paired empirical risk M_t does not appear in the average paired empirical risk value $\mathscr{M}^n$. The parameter b_n controls the time gaps between elements of pairs (X_t,X_i) appearing in (4) in such way that their joint law is close to the product law $\pi\otimes\pi$ (mixing of the chain is met). From a pragmatic point of view,

- we discard the first hypotheses that are not reliable, namely we do not consider hypothesis h_i for $i\leq c_n-b_n$. These first hypotheses are considered as not reliable since the online learner selected them based on a too small number of observed examples.
- since h_{t-b_n} is learned from $X_1,\dots,X_{t-b_n}$, we test the performance of h_{t-b_n} on X_t (and not on some X_i with $t-b_n+1\leq i<t$) to ensure that the distribution of X_t conditionally on $\sigma(X_1,\dots,X_{t-b_n})$ is approximately the invariant distribution of the chain π (see Assumption 1 and Definition 8). Stated otherwise, this ensures that sufficient mixing has occurred.

Note that we assume n large enough to ensure that $c_n-b_n\geq 1$. For any $\eta>0$, we denote $\mathscr{N}(\mathscr{H},\eta)$ the L_∞ η-covering number for the hypothesis class $\mathscr{H}$ (see Definition 3).

Definition 3 *(See [62, Chapter 5.1]) Let us consider some $\eta>0$. A L_∞ η-cover of a set $\mathscr{H}$ is a set $\{g_1,\dots,g_N\}\subset\mathscr{H}$ such that for any $h\in\mathscr{H}$, there exists some $i\in\{1,\dots,N\}$ such that $\|g_i-h\|_\infty\leq\eta$. The L_∞ η-covering number $\mathscr{N}(\mathscr{H},\eta)$ is the cardinality of the smallest L_∞ η-cover of the set $\mathscr{H}$.*

Theorem 4 bounds the average risk of the sequence of hypotheses in terms of its empirical counterpart $\mathscr{M}^n$ and is proved in Section 7.1.

Theorem 4 *Assume that the Markov chain $(X_i)_{i\geq 1}$ is reversible and satisfies Assumption 1. Assume the hypothesis space $(\mathscr{H}, \|\cdot\|_\infty)$ is compact. Let $h_0, h_1, \ldots, h_n \in \mathscr{H}$ be the ensemble of hypotheses generated by an arbitrary online algorithm working with a pairwise loss function ℓ such that,*

$$\ell(h, x_1, x_2) = \varphi(f(x_1) - f(x_2), h(x_1, x_2)),$$

where $\varphi : \mathbb{R}\times\mathbb{R} \to [0,1]$ is a Lipschitz function w.r.t. the second variable with a finite Lipschitz constant $Lip(\varphi)$. Let $\xi > 0$ be an arbitrary positive number and let us consider $q = \frac{\xi+1}{\log(1/\rho)}$ for the definition of b_n (see (5)). Then for all $c > 0$ and for all $\varepsilon > 0$ such that $\varepsilon \underset{n\to\infty}{=} o\left(n^{\xi}\right)$, we have for sufficiently large n

$$\mathbb{P}\left(\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n}) - \mathscr{M}^n\right| \geq \varepsilon\right) \leq 2\left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8Lip(\varphi)}\right) + 1\right] b_n \exp\left(-\frac{(c_n - b_n)C(m,\tau)\varepsilon^2}{16 b_n^2}\right),$$

where $C(m,\tau)^{-1} = 7\times 10^3 \times m^2\tau^2$.

Theorem 4 shows that average paired empirical risk $\mathscr{M}^n$ (see (3)) is close to average risk given by

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n}).$$

Quantitative errors bounds can be given assuming that the L_∞-metric entropy (l.h.s of the next equation) satisfies

$$\log\mathscr{N}(\mathscr{H}, \eta) = \mathscr{O}(\eta^{-\beta}),$$

where β is an exponent, depending on the dimension of state space E and the regularity of hypotheses of $\mathscr{H}$, that can be computed in some situations (Lipschitz function, higher order smoothness classes, see [62, Chapter 5.1] for instance). In this case, Theorem 4 shows

$$\left|\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n}) - \mathscr{M}^n\right| = \mathscr{O}_{\mathbb{P}}\left[\frac{\log^{\frac{2}{2+\beta}} n}{n^{\frac{1}{2+\beta}}}\right].$$

3.2.3 Batch hypothesis selection

Theorem 4 is a result on the performance of online learning algorithms. We will use this result to study the generalization performance of such online algorithms in the batch setting (see Theorem 5). Hence we are now interested in *selecting a good hypothesis from the ensemble of hypotheses generated by the online learner* namely that has a small empirical risk.

We measure the risk for h_{t-b_n} on the last $n-t$ examples of the sequence $X_1, \ldots, X_n$, and penalize each h_{t-b_n} based on the number of examples on which it is evaluated. More precisely, let us define the empirical risk of hypothesis h_{t-b_n} on $\{X_{t+1}, \ldots, X_n\}$ as

$$\widehat{\mathscr{R}}(h_{t-b_n}, t+1) := \binom{n-t}{2}^{-1} \sum_{k>i, i\geq t+1}^{n} \ell(h_{t-b_n}, X_i, X_k).$$

For a confidence parameter $\gamma \in (0,1)$ that will be specified in Theorem 5, the hypothesis $\widehat{h}$ is chosen to minimize the following penalized empirical risk,

$$\widehat{h} = h_{\widehat{t}-b_n} \quad \text{and} \quad \widehat{t} \in \arg\min_{c_n \leq t \leq n-1}\left(\widehat{\mathscr{R}}(h_{t-b_n}, t+1) + c_\gamma(n-t)\right), \tag{6}$$

where

$$c_\gamma(x) = \sqrt{\frac{C(m,\tau)^{-1}}{x}\log\frac{64(n-c_n)(n-c_n+1)}{\gamma}},$$

with $C(m,\tau)^{-1} = 7\times 10^3 \times m^2\tau^2$.

Theorem 5 proves that the model selection mechanism previously described select a hypothesis $\widehat{h}$ from the hypotheses of an arbitrary online learner whose risk is bounded relative to $\mathscr{M}^n$. The proof of Theorem 5 is postponed to Section 7.2.

Theorem 5 *Assume that the Markov chain $(X_i)_{i\geq 1}$ is reversible and satisfies Assumptions 1 and 2.(i). Let $h_0,\dots,h_n$ be the set of hypotheses generated by an arbitrary online algorithm $\mathscr{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 4. Let $\xi>0$ be an arbitrary positive number and let us consider $q=\frac{\xi+1}{\log(1/\rho)}$ for the definition of b_n (see (5)). For all $\varepsilon>0$ such that $\varepsilon \underset{n\to\infty}{=} o\left(n^{\xi}\right)$, if the hypothesis is selected via (6) with the confidence γ chosen as*

$$\gamma = 64(n-c_n+1)\exp\left(-(n-c_n)\varepsilon^2 C(m,\tau)/128\right),$$

then, when n is sufficiently large, we have

$$\mathbb{P}\left(\mathscr{R}(\widehat{h})\geq \mathscr{M}^n+\varepsilon\right)\leq 32\left[\mathscr{N}\left(\mathscr{H},\frac{\varepsilon}{16\mathrm{Lip}(\varphi)}\right)+1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{(16b_n)^2}+2\log n\right).$$

Remark The computation of the regularization term $c_\gamma(n-t)$ requires the knowledge of the constants m and τ related to the Markov chain. In the case where Assumption 1 holds with $m=1$, the chain regenerates at each time step with probability δ_m, and τ can be made explicit. The random variables $(T_i)_i$ have geometric distribution with parameter of success δ_m and using the moment generating function, one can show that $\|T_i\|_{\psi_1}=\left[\ln\left(\frac{2-\delta_m}{2(1-\delta_m)}\right)\right]^{-1}$.

3.3 Adaptive goodness-of-fit tests in a density model

3.3.1 Goodness-of-fit tests and review of the literature

In its original formulation, the goodness-of-fit test aims at determining if a given distribution q matches some unknown distribution p from samples $(X_i)_{i\geq 1}$ drawn independently from p. Classical approaches to solve the goodness of fit problem use the empirical process theory. Most of the popular tests such as the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics are based on the empirical distribution function of the samples. Other traditional approaches may require space partitioning or closed-form integrals [6], [8]. In [54], a non-parametric method is proposed with a test based on a kernel density estimator. In the last decade, a lot of effort has been put into finding more efficient goodness of fit tests. The motivation was mainly coming from graphical models where the distributions are known up to a normalization factor that is often computationally intractable. To address this problem, several tests have been proposed based on Reproducing Kernel Hilbert Space (RKHS) embedding. A large span of them use classes of Stein transformed RKHS functions ([45],[32]). For example in [13], a goodness-of-fit test is proposed for both i.i.d or non i.i.d samples. The test statistic uses the squared Stein discrepancy, which is naturally estimated by a V-statistic. One drawback of such approach is that the theoretical results provided are only asymptotic. This paper is part of a large list of works that proposed a goodness-of-fit test and where the use of U-statistics naturally emerge (see [45],[23],[11],[24],[25], [28]). To conduct a non-asymptotic analysis of the goodness of fit tests proposed for non i.i.d samples, a concentration result for U-statistics with dependent random variables is much needed.

3.3.2 Goodness-of-fit test for the density of the invariant measure of a Markov chain

In this section, we provide a goodness-of-fit test for Markov chains whose invariant distribution has density with respect to the Lebesgue measure λ_{Leb} on $\mathbb{R}$. Our work is inspired from [28] where Fromont and Laurent tackled the goodness-of-fit test with i.i.d samples. Conducting a non-asymptotic theoretical study of our test, we are able to identify the classes of alternatives over which our method has a prescribed power.

Let $X_1,\dots,X_n$ be a Markov chain with invariant distribution π with density f with respect to the Lebesgue measure on $\mathbb{R}$. Let f_0 be some given density in $L^2(\mathbb{R})$ and let α be in $]0,1[$. Assuming that f belongs to $L^2(\mathbb{R})$, we construct a level α test of the null hypothesis "$f=f_0$" against the alternative "$f\neq f_0$" from the observation $(X_1,\dots,X_n)$. The test is based on the estimation of $\|f-f_0\|_2^2$ that is $\|f\|_2^2+\|f_0\|_2^2-2\langle f,f_0\rangle$. $\langle f,f_0\rangle$ is usually estimated by the empirical estimator $\sum_{i=1}^n f_0(X_i)/n$ and the cornerstone of our approach is to find a way to estimate $\|f\|_2^2$. We follow the work of [28] and we introduce a set $\{S_m, m\in\mathscr{M}\}$ of linear subspaces of $L^2(\mathbb{R})$. For all m in $\mathscr{M}$, let $\{p_l, l\in\mathscr{L}_m\}$ be some orthonormal basis of S_m. The

variable

$$\widehat{\theta}_m = \frac{1}{n(n-1)} \sum_{l \in \mathscr{L}_m} \sum_{i \neq j=1}^{n} p_l(X_i) p_l(X_j)$$

estimates $\|\Pi_{S_m}(f)\|_2^2$ where Π_{S_m} denotes the orthogonal projection onto S_m. Then $\|f - f_0\|_2^2$ can be approximated by

$$\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n} \sum_{i=1}^{n} f_0(X_i),$$

for any m in $\mathscr{M}$. Denoting by $t_m(u)$ the $(1-u)$ quantile of the law of $\widehat{T}_m$ under the hypothesis "$f = f_0$" and considering

$$u_\alpha = \sup_{u \in]0,1[} \mathbb{P}_{f_0}\left(\sup_{m \in \mathscr{M}} (\widehat{T}_m - t_m(u)) > 0 \right) \leq \alpha,$$

we introduce the test statistic T_α defined by

$$T_\alpha = \sup_{m \in \mathscr{M}} (\widehat{T}_m - t_m(u_\alpha)). \tag{7}$$

The test consists in rejecting the null hypothesis if T_α is positive. This approach can be read as a multiple testing procedure. Indeed, for each m in $\mathscr{M}$, we construct a level u_α test of the null hypothesis "$f = f_0$" by rejecting this hypothesis if $\widehat{T}_m$ is larger than its $(1-u_\alpha)$ quantile under the hypothesis "$f = f_0$". We thus obtain a collection of tests and we decide to reject the null hypothesis if for some of the tests of the collection this hypothesis is rejected.

Now we define the different collection of linear subspaces $\{S_m, m \in \mathscr{M}\}$ that we will use in the following. We will focus on constant piecewise functions, scaling functions and, in the case of compactly supported densities, trigonometric polynomials.

- For all D in $\mathbb{N}^*$ and $k \in \mathbb{Z}$, let
$$I_{D,k} = \sqrt{D} \mathbb{1}_{[k/D,(k+1)/D[}.$$
For all $D \in \mathbb{N}^*$, we define $S_{(1,D)}$ as the space generated by the functions $\{I_{D,k}, k \in \mathbb{Z}\}$ and
$$\widehat{\theta}_{(1,D)} = \frac{1}{n(n-1)} \sum_{k \in \mathbb{Z}} \sum_{i \neq j=1}^{n} \mathbb{1}_{D,k}(X_i) \mathbb{1}_{D,k}(X_j).$$

- Let us consider a pair of compactly supported orthonormal wavelets (φ, ψ) such that for all $J \in \mathbb{N}$, $\{\varphi_{J,k} = 2^{J/2}\varphi(2^J \cdot - k), k \in \mathbb{Z}\} \cup \{\psi_{j,k} = 2^{j/2}\psi(2^j \cdot - k), j \in \mathbb{N}, j \geq J, k \in \mathbb{Z}\}$ is an orthonormal basis of $L_2(\mathbb{R})$. For all $J \in \mathbb{N}$ and $D = 2^J$, we define $S_{(2,D)}$ as the space generated by the scaling functions $\{\varphi_{J,k}, k \in \mathbb{Z}\}$ and
$$\widehat{\theta}_{(2,D)} = \frac{1}{n(n-1)} \sum_{k \in \mathbb{Z}} \sum_{i \neq j=1}^{n} \varphi_{J,k}(X_i) \varphi_{J,k}(X_j).$$

- Let us consider the Fourier basis of $L_2([0,1])$ given by
$$\begin{aligned} g_0(x) &= \mathbb{1}_{[0,1]}(x), \\ g_{2p-1}(x) &= \sqrt{2}\cos(2\pi p x)\mathbb{1}_{[0,1]}(x) \quad \forall p \geq 1, \\ g_{2p}(x) &= \sqrt{2}\sin(2\pi p x)\mathbb{1}_{[0,1]}(x) \quad \forall p \geq 1. \end{aligned}$$
For all $D \in \mathbb{N}^*$, we define $S_{(3,D)}$ as the space generated by the functions $\{g_l, l = 0, \ldots, D\}$ and
$$\widehat{\theta}_{(3,D)} = \frac{1}{n(n-1)} \sum_{l=0}^{D} \sum_{i \neq j=1}^{n} g_l(X_i) g_l(X_j).$$

We denote $\mathbb{D}_1 = \mathbb{D}_3 = \mathbb{N}\backslash\{0\}$ and $\mathbb{D}_2 = \{2^J, J \in \mathbb{N}\}$. For l in $\{1,2,3\}$, D in $\mathbb{D}_l$, $\Pi_{S_{(l,D)}}$ denotes the orthogonal projection onto $S_{(l,D)}$ in $L^2(\mathbb{R})$. For all l in $\{1,2,3\}$, we take $\mathscr{D}_l \subset \mathbb{D}_l$ with $\cup_{l\in\{1,2,3\}}\mathscr{D}_l \neq \emptyset$ and $\mathscr{D}_3 = \emptyset$ if the X_i's are not included in $[0,1]$. Let $\mathscr{M} = \{(l,D), l \in \{1,2,3\}, D \in \mathscr{D}_l\}$.

Theorem 6 describes classes of alternatives over which the corresponding test has a prescribed power. We refer to Section C.7 for the proof of Theorem 6.

Theorem 6 *Let $X_1,\dots,X_n$ a Markov chain on $\mathbb{R}$ satisfying the Assumptions 1 and 2.(i), with invariant measure π and initial distribution χ. We assume that π has density f with respect the Lebesgue measure on $\mathbb{R}$ and let f_0 be some given density. Let T_α be the test statistic defined by (7). Assume that f_0 and f belong to $L_\infty(\mathbb{R})$ and that there exist $p_1, p_2 \in (1,+\infty]$ such that*

$$C_\chi := \left\| \frac{1}{f}\frac{d\chi}{d\lambda_{Leb}} \right\|_{f\lambda_{Leb},p_1} \vee \left\| \frac{1}{f_0}\frac{d\chi}{d\lambda_{Leb}} \right\|_{f_0\lambda_{Leb},p_2} < \infty,$$

where we used the notations of Assumption 4. We fix some γ in $]0,1[$. For any $\varepsilon \in]0,2[$, there exist some positive constants C_1, C_2, C_3 such that, setting for all $m = (l,D)$ in $\mathscr{M}$,

$$V_m(\gamma) = C_1\|f\|_\infty \frac{\log(3C_\chi/\gamma)}{\varepsilon n} + C_2\left(\|f\|_\infty \log(D+1) + \|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)}{n} + C_3\left(\|f\|_\infty + 1\right) DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),$$

with

$$R(n,u) = \log n\left\{\frac{u}{n} + \left[\frac{u}{n}\right]^2\right\},$$

if f satisfies

$$\|f - f_0\|_2^2 > (1+\varepsilon)\inf_{m\in\mathscr{M}}\left\{\|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha) + V_m(\gamma)\right\}, \tag{8}$$

then

$$\mathbb{P}_f\left(T_\alpha \leq 0\right) \leq \gamma.$$

In order to make the condition (8) more explicit and to study its sharpness, we define the uniform separation rate which provides for any $\gamma \in (0,1)$ the smallest distance between the set of null hypotheses and the set of alternatives to ensure that the power of our statistic test with level α is at least $1-\gamma$.

Definition 4 *Given $\gamma \in]0,1[$ and a class of functions $\mathscr{B} \subset L_2(\mathbb{R})$, we define the uniform separation rate $\rho(\Phi_\alpha, \mathscr{B}, \gamma)$ of a level α test Φ_α of the null hypothesis "$f \in \mathscr{F}$" over the class $\mathscr{B}$ as the smallest number ρ such that the test guarantees a power at least equal to $(1-\gamma)$ for all alternatives $f \in \mathscr{B}$ at a distance ρ from $\mathscr{F}$. Stated otherwise, denoting by $d_2(f,\mathscr{F})$ the L_2-distance between f and $\mathscr{F}$ and by $\mathbb{P}_f$ the distribution of the observation $(X_1,\dots,X_n)$,*

$$\rho(\Phi_\alpha, \mathscr{B}, \gamma) = \inf\left\{\rho > 0, \forall f \in \mathscr{B}, d_2(f,\mathscr{F}) \geq \rho \implies \mathbb{P}_f(\Phi_\alpha \text{ rejects}) \geq 1-\gamma\right\}.$$

In the following, we derive on explicit upper bound on the uniform separation rates of the test proposed above over several classes of alternatives. For $s > 0, P > 0, M > 0$ and $l \in \{1,2,3\}$, we introduce

$$\mathscr{B}_s^{(l)}(P,M) = \left\{f \in L_2(\mathbb{R}) \mid \forall D \in \mathscr{D}_l, \quad \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2 D^{-2s}, \|f\|_\infty \leq M\right\}.$$

These sets of functions include some Hölder balls or Besov bodies with smoothness s, as highlighted in [28, Section 2.3]. Corollary 1 gives an upper bound for the uniform separation rate of our testing procedure over the classes $\mathscr{B}_s^{(l)}(P,M)$ and is proved in Section C.9.

Corollary 1 *Let T_α be the test statistic defined by (7). Assume that for $l \in \{1,2,3\}$, $\mathscr{D}_l$ is $\{2^J, 0 \leq J \leq \log_2\left(n/(\log(n)\log\log n)^2\right)\}$ or $\emptyset$. For all $s > 0$, $M > 0, P > 0$ and $l \in \{1,2,3\}$ such that $\mathscr{D}_l \neq \emptyset$, there exists some positive constant $C = C(s,\alpha,\gamma,M,\|f_0\|_\infty)$ such that the uniform separation rate of the test $\mathbb{1}_{T_\alpha>0}$ over $\mathscr{B}_s^{(l)}(P,M)$ satisfies for n large enough*

$$\rho\left(\mathbb{1}_{T_\alpha>0}, \mathscr{B}_s^{(l)}(P,M), \gamma\right) \leq C' P^{\frac{1}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}}.$$

Remark In Corollary 1, the condition *n large enough* corresponds to

$$\left(\log(n)\frac{\log\log n}{n}\right)^{1/2} \leq P \leq \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}.$$

For the problem of testing the null hypothesis "$f = \mathbb{1}_{[0,1]}$" against the alternative $f = \mathbb{1}_{[0,1]} + g$ with $g \neq 0$ and $g \in B_s(P)$ where $B_s(P)$ is a class of smooth functions (like some Hölder, Sobolev or Besov ball in $L_2([0,1])$) with unknown smoothness parameter s, Ingster in [36] established in the case where the random variables $(X_i)_{i\geq 1}$ are i.i.d. that the adaptive minimax rate of testing is of order $(\sqrt{\log\log n}/n)^{2s/(4s+1)}$. From Corollary 1, we see that our procedure leads to a rate which is close (at least for sufficiently large smoothness parameter s) to the one derived by Ingster in the i.i.d. framework since the upper bound on the uniform separation rate from Corollary 1 can be read (up to a log factor) as $([\log\log n]/n)^{\frac{2s}{4s+2}}$.

3.3.3 Simulations

We propose to test our method on three practical examples.[1] In all our simulations, we use Markov chains of length $n = 100$. We choose different alternatives to test our method and we use i.i.d. samples from these distributions. We chose a level $\alpha = 5\%$ for all our experiments. All tests are conducted as follows.

1. We start by the estimation of the $(1-u)$ quantiles $t_m(u)$ of the variables $\widehat{T}_m = \widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^n f_0(X_i)$ under the hypothesis "$f = f_0$" for u varying on a regular grid of $]0,\alpha[$. We sample 5,000 sequences of length $n = 100$ with i.i.d. random variables with distribution f_0. We end up with an estimation $\widehat{t}_m(u)$ of $t_m(u)$ for any u in the grid and any $m \in \mathscr{M}$.

2. Then, we estimate the value of u_α. We sample again 5,000 sequences of length $n = 100$ with i.i.d. random variables with distribution f_0. We use them to estimate the probabilities $\mathbb{P}_{f_0}(\sup_{m\in\mathscr{M}}(\widehat{T}_m - \widehat{t}_m(u)) > 0)$ for any u in the grid and we keep the larger value of u such that the corresponding probability is still larger than α. The selected value of the grid is called u_α. Thanks to the first step, we have the estimates $\widehat{t}_m(u_\alpha)$ of $t_m(u_\alpha)$ for any $m \in \mathscr{M}$.

3. Finally, we sample 5,000 Markov chains with length $n = 100$ with invariant distribution f. For each sequence, we can compute $\widehat{T}_m$. Dividing by 5,000 the number of sequences for which $\sup_{m\in\mathscr{M}}(\widehat{T}_m - \widehat{t}_m(u_\alpha)) > 0$, we get an estimation of the power of the test.

To define comparison points, we compare the power of our test with the classical Kolmorogorov-Smirnov test (KS test) and the Chi-squared test (χ^2 test). The rejection region associated with a test of level 5% is set by *sampling under the null* for both the KS test and the χ^2 test.

Example 1: AR(1) process Let us consider some $\theta \in (0,1)$. Then, we define the AR(1) process $(X_i)_{i\geq 1}$ starting from $X_1 = 0$ with for any $n \geq 1$,

$$X_{n+1} = \theta X_n + \xi_{n+1},$$

where $(\xi_n)_n$ are i.i.d. random variables with distribution $\mathcal{N}(0,\tau^2)$ with $\tau > 0$. From Example 1 of Section 2.6, we know that Assumptions 1 and 2.(i) hold. The invariant measure π of the Markov chain $(X_i)_{i\geq 1}$ is $\mathcal{N}\left(0, \frac{\tau^2}{1-\theta^2}\right)$, i.e. π has density f with respect to the Lebesgue measure on $\mathbb{R}$ with

$$\forall y \in \mathbb{R}, \quad f(y) = \frac{\sqrt{1-\theta^2}}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(1-\theta^2)y^2}{2\tau^2}\right).$$

We focus on the following alternatives

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

[1] The code is available at https://github.com/quentin-duchemin/goodness-of-fit-MC.

(μ,σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f}-\mathbf{f}_{\mu,\sigma^2}\|_2$
$(2,1.5)$	0.99	0.85	0.98	0.39
$(0,1)$	0.97	0.9	0.8	0.2
$(-0.2,1.2)$	0.86	0.63	0.84	0.17
$(0,1.2)$	0.81	0.64	0.82	0.16
$(0,2)$	0.1	0.03	0.29	0.06

Table 1: Estimated power of tests for Markov chains with size $n = 100$. We worked with $\tau = 1$, $\theta = 0.8$ and $\mathscr{M} = \{(1,i) : i \in \{1,\dots,10\}\}$. Hence, the invariant distribution of the chain is approximately $\mathcal{N}(0,2.8)$. For the χ^2 test, we work on the interval $[-5,5]$ that we split into 100 regular parts.

Example 2: Markov chain generated from independent Metropolis Hasting algorithm Let us consider the probability measure π with density f with respect to the Lebesgue measure on $[-3,3]$ where

$$\forall x \in [-3,3], \quad f(x) = \frac{1}{Z} e^{-x^2} \left(3 + \sin(5x) + \sin(2x)\right),$$

with Z a normalization constant such that $\int_{-3}^{3} f(x)dx = 1$. To construct a Markov chain with invariant measure π, we use an independent Metropolis-Hasting algorithm with proposal density $q(x) \propto \exp(-x^2/6)$. Using Proposition 3, we get that the above built Markov chain $(X_i)_{i\geq 1}$ satisfies Assumptions 1 and 2.(i). We focus on the following alternatives

$$g_{\mu,\sigma^2}(x) = \frac{1}{Z(\mu,\sigma^2)} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{[-3,3]}(x),$$

where $Z(\mu,\sigma^2)$ is a normalization constant such that $\int g_{\mu,\sigma^2}(x)dx = 1$. Table 2 shows the estimated powers for our test, the KS test and the χ^2 test.

(μ,σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f}-\mathbf{g}_{\mu,\sigma^2}\|_2$
$(0,1)$	0.96	0.91	0.9	0.29
$(0,0.7^2)$	0.93	0.84	0.95	0.23
$(0.3,0.7^2)$	0.92	0.87	0.93	0.19

Table 2: Estimated powers of the tests for Markov chains with size $n = 100$. We used $\mathscr{M} = \{(1,i) : i \in \{1,\dots,10\}\}$. For the χ^2 test, we work on the interval $[-3,3]$ that we split into 100 regular parts.

Example 3: ARCH process Let us consider some $\theta \in (-1,1)$. We are interested in the simple threshold auto-regressive model $(X_n)_{n\geq 1}$ defined by $X_1 = 0$ and for any $n \geq 1$,

$$X_{n+1} = \theta|X_n| + (1-\theta^2)^{1/2}\xi_{n+1},$$

where the random variables $(\xi_n)_{n\geq 2}$ are i.i.d. with standard Gaussian distribution. From Example 3 of Section 2.6, we know that Assumptions 1 and 2.(i) hold. The transition kernel of the Markov chain $(X_i)_{i\geq 1}$ is

$$\forall x, y \in \mathbb{R}, \quad P(x,y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\theta|x|)^2}{2(1-\theta^2)}\right).$$

The invariant distribution π of the Markov chain has density f with respect to the Lebesgue measure on $\mathbb{R}$ with

$$\forall y \in \mathbb{R}, \quad f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \Phi\left(\frac{\theta y}{(1-\theta^2)^{1/2}}\right),$$

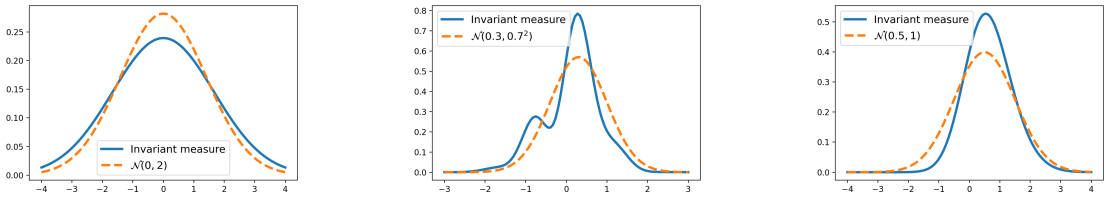
where Φ is the standard normal cumulative distribution function. We focus on the following alternatives

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Table 3 shows the estimated powers for our test, the KS test and the χ^2 test.

(μ, σ^2)	**Our test**	χ^2 test	**KS** test	$\|\mathbf{f} - \mathbf{f}_{\mu,\sigma^2}\|_2$
$(0, 1)$	0.98	0.85	0.95	0.3
$(1, 0.8^2)$	0.95	0.79	0.88	0.22
$(0.5, 1)$	0.41	0.07	0.55	0.14
$(0.6, 0.8^2)$	0.44	0.16	0.22	0.036

Table 3: Estimated powers of the tests for Markov chains with size $n = 100$. We used $\theta = 0.8$ and $\mathscr{M} = \{(1, i) : i \in \{1, \dots, 10\}\}$. For the χ^2 test, we work on the interval $[-20, 20]$ that we split into 100 regular parts.



(a) Example 1 (b) Example 2 (c) Example 3

Figure 2: In solid line, we plot the density of the invariant of the Markov chain for the three examples of our simulations. In dotted line, we plot the density of the alternative that gives the smaller power on our experiments.

4 Proof of Theorem 1

In Section 5, we explain succinctly how to easily obtain the proof of Theorem 2 from the one of Theorem 1. Note that in Theorem 2, Assumption 2.(ii) is not required and thus ν may be different from π. In order to make the reader easily understand that the proof of Theorem 2 closely follows the one of Theorem 1, we keep on purpose the distinction between ν and π in this section.

Let us recall that Theorem 1 requires either a mild condition on the initial distribution of the Markov chain through or the fact that the kernels $h_{i,j}$ do not depend on i (see Assumption 4). One only needs to consider different Bernstein concentration inequalities for sums of functions of Markov chains to go from one result to the other. In this section, we give the proof of Theorem 1 in the case where Assumption 4.(i) holds. We specify the part of the proof that should be changed to get the result when $h_{i,j}$ may depend on both i and j and when Assumption 4.(ii) holds. We make this easily identifiable using the symbol  .

Our proof is inspired from [29, Section 3.4.3] where a Bernstein-type inequality is shown for U-statistics of order 2 in the independence setting. Their proof relies on the *canonical* property of the kernel functions which endowed the U-statistic with a martingale structure. We want to use a similar argument and we decompose $U_{\text{stat}}(n)$ to recover the martingale property for each term (except for the last one). Considering for any $l \geq 1$ the σ-algebra $G_l = \sigma(X_1, \dots, X_l)$, the notation $\mathbb{E}_l$ refers to the conditional expectation with

respect to G_l. Then we decompose $U_{\text{stat}}(n)$ as follows,

$$U_{\text{stat}}(n) = \sum_{k=1}^{t_n} \sum_{i<j} \left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{j-k}[h_{i,j}(X_i, X_j)] \right) + \sum_{i<j} \left(\mathbb{E}_{j-t_n}[h_{i,j}(X_i, X_j)] - \mathbb{E}_{\pi}[h_{i,j}(X, \cdot)] \right), \tag{9}$$

where t_n is an integer that scales logarithmically with n and that will be specified latter. By convention, we assume here that for all $k < 1$, $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot]$. Hence the first term that we will consider is given by

$$U_n = \sum_{1 \le i < j \le n} h^{(0)}_{i,j}(X_i, X_{j-1}, X_j),$$

where for all $x, y, z \in E$,

$$h^{(0)}_{i,j}(x, y, z) = h_{i,j}(x, z) - \int_w h_{i,j}(x, w) P(y, dw).$$

We provide a detailed proof of a concentration result for U_n by taking advantage of its martingale structure following the work of [29, Section 3.4.3]. Reasoning by induction, we show that the $t_n - 1$ following terms involved in the decomposition (9) of $U_{\text{stat}}(n)$ can be handled using a similar approach. Since the last term of the decomposition (9) has not a martingale property, another argument is required. We deal with the last term exploiting the uniform ergodicity of the Markov chain $(X_i)_{i \ge 1}$ which is guaranteed by Assumption 1 (see [52, Theorem 8]).

4.1 Concentration of the first term of the decomposition of the U-statistic

Martingale structure of the U-statistic Defining $Y_j = \sum_{i=1}^{j-1} h^{(0)}_{i,j}(X_i, X_{j-1}, X_j)$, U_n can be written as $U_n = \sum_{j=2}^{n} Y_j$. Since

$$\mathbb{E}_{j-1}[Y_j] = \mathbb{E}[Y_j \mid X_1, \ldots, X_{j-1}] = 0,$$

we know that $(U_k)_{k \ge 2}$ is a martingale relative to the σ-algebras G_l, $l \ge 2$. This martingale can be extended to $n = 0$ and $n = 1$ by taking $U_0 = U_1 = 0$, $G_0 = \{\emptyset, E\}$, $G_1 = \sigma(X_1)$. We will use the martingale structure of $(U_n)_n$ through the following Lemma.

Lemma 1 *(see [29, Lemma 3.4.6])*
Let $(U_m, G_m), m \in \mathbb{N}$, be a martingale with respect to a filtration G_m such that $U_0 = U_1 = 0$. For each $m \ge 1$ and $k \ge 2$, define the angle brackets $A^k_m = A^k_m(U)$ of the martingale U by

$$A^k_m = \sum_{i=1}^{m} \mathbb{E}_{i-1}[(U_i - U_{i-1})^k]$$

(and note $A^k_1 = 0$ for all k). Suppose that for $\alpha > 0$ and all $i \ge 1$, $\mathbb{E}[e^{\alpha |U_i - U_{i-1}|}] < \infty$. Then

$$\left(\varepsilon_m := e^{\alpha U_m - \sum_{k \ge 2} \alpha^k A^k_m / k!}, G_m \right), \ m \in \mathbb{N},$$

is a supermartingale. In particular, $\mathbb{E}[\varepsilon_m] \le \mathbb{E}[\varepsilon_1] = 1$, so that, if $A^k_m \le w^k_m$ for constants $w^k_m \ge 0$; then

$$\mathbb{E}[e^{\alpha U_m}] \le e^{\sum_{k \ge 2} \alpha^k w^k_m / k!}.$$

We will also use the following convexity result several times.

Lemma 2 *For all $\theta_1, \theta_2, \varepsilon \ge 0$,*

$$(\theta_1 + \theta_2)^k \le (1 + \varepsilon)^{k-1} \theta_1^k + (1 + \varepsilon^{-1})^{k-1} \theta_2^k.$$

Proof of Lemma 2. Notice that for all $\theta_1, \theta_2 \geq 0$ and $0 < \varepsilon \leq 1$ by convexity,

$$\left(\frac{\theta_1+\theta_2}{1+\varepsilon}\right)^k = \left(\frac{\theta_1}{1+\varepsilon} + \frac{\varepsilon\varepsilon^{-1}\theta_2}{1+\varepsilon}\right)^k \leq \frac{1}{1+\varepsilon}\theta_1^k + \frac{\varepsilon}{1+\varepsilon}\varepsilon^{-k}\theta_2^k,$$

so that

$$(\theta_1+\theta_2)^k \leq (1+\varepsilon)^{k-1}\theta_1^k + \varepsilon^{-(k-1)}(1+\varepsilon)^{k-1}\theta_2^k = (1+\varepsilon)^{k-1}\theta_1^k + (1+\varepsilon^{-1})^{k-1}\theta_2^k.$$

By symmetry, this inequality holds for all $\varepsilon \geq 0$, that is, for all $\theta_1, \theta_2, \varepsilon \geq 0$,

$$(\theta_1+\theta_2)^k \leq (1+\varepsilon)^{k-1}\theta_1^k + (1+\varepsilon^{-1})^{k-1}\theta_2^k.$$

■

For all $k \geq 2$ and $n \geq 1$, we have :

$$\begin{aligned}
A_n^k &= \sum_{j=2}^n \mathbb{E}_{j-1}\left[\sum_{i=1}^{j-1} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j)\right]^k \\
&\leq V_n^k := \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} h_{i,j}^{(0)}(X_i, X_{j-1}, X_j)\right|^k \\
&= \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\Big(h_{i,j}(X_i,X_j) - \mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] + \mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] - \mathbb{E}_{j-1}[h_{i,j}(X_i,X_j)]\Big)\right|^k \\
&= \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\big(p_{i,j}(X_i,X_j) + m_{i,j}(X_i,X_{j-1})\big)\right|^k,
\end{aligned}$$

where

$$p_{i,j}(x,z) = h_{i,j}(x,z) - \mathbb{E}_{X'\sim\nu}[h_{i,j}(x,X')],$$

and

$$m_{i,j}(x,y) = \mathbb{E}_{X'\sim\nu}[h_{i,j}(x,X')] - \int_z h_{i,j}(x,z)P(y,dz).$$

Using Lemma 2 with $\varepsilon = 1/2$, we deduce that

$$\begin{aligned}
V_n^k &\leq \sum_{j=2}^n \mathbb{E}_{j-1}\left(\left|\sum_{i=1}^{j-1} p_{i,j}(X_i,X_j)\right| + \left|\sum_{i=1}^{j-1} m_{i,j}(X_i,X_{j-1})\right|\right)^k \\
&\leq \left(\frac{3}{2}\right)^{k-1}\sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i,X_j)\right|^k + 3^{k-1}\sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} m_{i,j}(X_i,X_{j-1})\right|^k.
\end{aligned}$$

Let us remark that

$$\begin{aligned}
\sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} m_{i,j}(X_i,X_{j-1})\right|^k &= \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\Big(\mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] - \mathbb{E}_{j-1}[h_{i,j}(X_i,X_j)]\Big)\right|^k \\
&= \sum_{j=2}^n \left|\sum_{i=1}^{j-1}\Big(\mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] - \mathbb{E}_{j-1}[h_{i,j}(X_i,X_j)]\Big)\right|^k \\
&= \sum_{j=2}^n \left|\mathbb{E}_{j-1}\left[\sum_{i=1}^{j-1}\Big(\mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] - h_{i,j}(X_i,X_j)\Big)\right]\right|^k \\
&\leq \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\Big(\mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] - h_{i,j}(X_i,X_j)\Big)\right|^k,
\end{aligned}$$

where the last inequality comes from Jensen's inequality. We obtain the following upper-bound for V_n^k,

$$V_n^k \le 2\times 3^{k-1}\sum_{j=2}^{n}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j)\right|^k$$
$$\le 2\times 3^{k-1}\delta_M\sum_{j=2}^{n}\mathbb{E}_{X_j'}\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k,$$

where the random variables $(X_j')_j$ are i.i.d. with distribution ν (see Assumption 2). $\mathbb{E}_{X_j'}$ denotes the expectation on the random variable X_j'.

Lemma 3 *(see [29, Ex.1 Section 3.4]) Let Z_j be independent random variables with respective probability laws P_j. Let $k>1$, and consider functions $f_1,\dots,f_N$ where for all $j\in[N]$, $f_j\in L^k(P_j)$. Then the duality of L^p spaces and the independence of the variables Z_j imply that*

$$\left(\sum_{j=1}^{N}\mathbb{E}\left[|f_j(Z_j)|^k\right]\right)^{1/k}=\sup_{\sum_{j=1}^{N}\mathbb{E}|\xi_j(Z_j)|^{k/(k-1)}=1}\sum_{j=1}^{N}\mathbb{E}\left[f_j(Z_j)\xi_j(Z_j)\right],$$

where the sup runs over $\xi_j\in L^{k/(k-1)}(P_j)$.

Then by the duality result of Lemma 3,

$$\left(V_n^k\right)^{1/k}\le\left(2\delta_M\times 3^{k-1}\sum_{j=2}^{n}\mathbb{E}_{X_j'}\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k\right)^{1/k}$$
$$\le(2\delta_M)^{1/k}\sup_{\xi\in\mathscr{L}_k}\sum_{j=2}^{n}\sum_{i=1}^{j-1}\mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]$$
$$\text{where }\mathscr{L}_k=\left\{\xi=(\xi_2,\dots,\xi_n)\,s.t.\,\forall 2\le j\le n,\ \xi_j\in L^{k/(k-1)}(\nu)\ with\ \sum_{j=2}^{n}\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1\right\}.$$
$$=(2\delta_M)^{1/k}\sup_{\xi\in\mathscr{L}_k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]$$

Let us denote F the subset of the set $\mathscr{F}(E,\mathbb{R})$ of all measurable functions from (E,Σ) to $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ that are bounded by A. We set $S:=E\times F^{n-1}$. For all $i\in[n]$, we define W_i by

$$W_i:=\left(X_i,\underbrace{0,\dots,0}_{(i-1)\text{ times}},p_{i,i+1}(X_i,\cdot),p_{i,i+2}(X_i,\cdot),\dots,p_{i,n}(X_i,\cdot)\right)\in S.$$

Hence for all $i\in[n]$, W_i is $\sigma(X_i)$-measurable. We define for any $\xi=(\xi_2,\dots,\xi_n)\in\prod_{i=2}^{n}L^{k/(k-1)}(\nu)$ the function

$$\forall w=(x,p_2,\dots,p_n)\in S,\quad f_\xi(w)=\sum_{j=2}^{n}\int p_j(y)\xi_j(y)d\nu(y).$$

Then setting $\mathscr{F}=\{f_\xi:\sum_{j=2}^{n}\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1\}$, we have

$$(V_n^k)^{1/k}\le(2\delta_M)^{1/k}\sup_{f_\xi\in\mathscr{F}}\sum_{i=1}^{n-1}f_\xi(W_i).$$

By the separability of the L^p spaces of finite measures, $\mathscr{F}$ can be replaced by a countable subset $\mathscr{F}_0$. To upper-bound the tail probabilities of U_n, we will bound the variable V_n^k on sets of large probability using Talagrand's inequality. Then we will use Lemma 1 on these sets by means of optional stopping.

Application of Talagrand's inequality for Markov chains The proof of Lemma 4 is provided in Section C.2.

Lemma 4 *Let us denote*

$$Z = \sup_{f_\xi \in \mathscr{F}} \sum_{i=1}^{n-1} f_\xi(W_i)$$

and

$$\sigma_k^2 = \mathbb{E}\left[\sum_{i=1}^{n-1} \sup_{f_\xi \in \mathscr{F}} f_\xi(W_i)^2\right] \quad \text{and} \quad b_k = \sup_{w \in S} \sup_{f_\xi \in \mathscr{F}} |f_\xi(w)|.$$

Then it holds for any $t > 0$,

$$\mathbb{P}\left(Z > \mathbb{E}[Z] + t\right) \leq \exp\left(-\frac{1}{8\|\Gamma\|^2} \min\left(\frac{t^2}{4\sigma_k^2}, \frac{t}{b_k}\right)\right),$$

where $\|\Gamma\| \leq \frac{2L}{1-\rho}$.

Using Lemma 4, we deduce that for any $t > 0$,

$$\mathbb{P}\left((V_n^k)^{1/k} \geq (2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k} t\right) \leq \exp\left(-\frac{1}{8\|\Gamma\|^2} \min\left(\frac{t^2}{4\sigma_k^2}, \frac{t}{b_k}\right)\right),$$

which implies that for any $x \geq 0$,

$$\mathbb{P}\left((V_n^k)^{1/k} \geq (2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k} 2\sigma_k \sqrt{x} + (2\delta_M)^{1/k} b_k x\right) \leq \exp\left(-\frac{x}{8\|\Gamma\|^2}\right).$$

Using the change of variable $x = k8\|\Gamma\|^2 u$ with $u \geq 0$ in the previous inequality leads to

$$\mathbb{P}\left(\bigcup_{k=2}^{\infty} (V_n^k)^{1/k} \geq (2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k} \sigma_k 3\|\Gamma\| \sqrt{ku} + (2\delta_M)^{1/k} k8\|\Gamma\|^2 b_k u\right) \leq 1.62 e^{-u},$$

because

$$1 \wedge \sum_{k=2}^{\infty} \exp(-ku) \leq 1 \wedge \frac{1}{e^u(e^u - 1)} = \left(e^u \wedge \frac{1}{e^u - 1}\right) e^{-u} \leq \frac{1+\sqrt{5}}{2} e^{-u} \leq 1.62 e^{-u}.$$

Bounding b_k. Using Hölder's inequality we have,

$$\begin{aligned}
b_k &= \sup_{w \in S} \sup_{f_\xi \in \mathscr{F}} |f_\xi(w)| \\
&= \sup_{(p_2,\ldots,p_n) \in F^{n-1}} \sup_{\xi \in \mathscr{L}_k} \sum_{j=2}^{n} \mathbb{E}[p_j(X_j')\xi_j(X_j')] \\
&\leq \sup_{(p_2,\ldots,p_n) \in F^{n-1}} \sup_{\sum_{j=2}^n \mathbb{E}|\xi_j(X_j')|^{k/(k-1)} = 1} \sum_{j=2}^{n} \left(\mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k} \left(\mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
&\leq \sup_{(p_2,\ldots,p_n) \in F^{n-1}} \sup_{\sum_{j=2}^n \mathbb{E}|\xi_j(X_j')|^{k/(k-1)} = 1} \left(\sum_{j=2}^{n} \mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k} \left(\sum_{j=2}^{n} \mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
&\leq \sup_{(p_2,\ldots,p_n) \in F^{n-1}} \left(\sum_{j=2}^{n} \mathbb{E}\left|p_j(X_j')\right|^k\right)^{1/k} \\
&\leq ((nA^2)A^{k-2})^{1/k},
\end{aligned}$$

where

$$A := 2 \max_{i,j} \|h_{i,j}\|_\infty \text{ which satisfies } \max_{i,j} \|p_{i,j}\|_\infty \leq A.$$

Here, we used that F is the set of measurable functions from (E, Σ) to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ bounded by A.

Bounding the variance.

$$\begin{aligned}
\sigma_k^2 &= \mathbb{E}\left[\sum_{i=1}^{n-1} \sup_{f_\xi \in \mathscr{F}} f_\xi(W_i)^2\right] \\
&= \sum_{i=1}^{n-1} \mathbb{E}\left[\sup_{\xi\in\mathscr{L}_k}\left(\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right)^2\right] \\
&= \sum_{i=1}^{n-1} \mathbb{E}\left[\left(\sup_{\xi\in\mathscr{L}_k}\left|\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right|\right)^2\right] \\
&\le n\left(B^2A^{k-2}\right)^{2/k},
\end{aligned}$$

where the last inequality comes from the following (where we use twice Holder's inequality),

$$\begin{aligned}
&\sup_{\xi\in\mathscr{L}_k}\left|\sum_{j=i+1}^{n} \mathbb{E}_{X_j'}\left[p_{i,j}(X_i,X_j')\xi_j(X_j')\right]\right| \\
\le\ & \sup_{\xi\in\mathscr{L}_k}\sum_{j=i+1}^{n}\left(\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k}\left(\mathbb{E}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
\le\ & \sup_{\sum_{j=2}^n \mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1}\left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k}\left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|\xi_j(X_j')\right|^{k/(k-1)}\right)^{(k-1)/k} \\
\le\ & \left(\sum_{j=i+1}^{n}\mathbb{E}_{X_j'}\left|p_{i,j}(X_i,X_j')\right|^k\right)^{1/k} \\
\le\ & \left(B^2A^{k-2}\right)^{1/k},
\end{aligned}$$

where

$$B^2 := \max\left[\max_i\left\|\sum_{j=i+1}^{n}\mathbb{E}_{X\sim\nu}\left[p_{i,j}^2(\cdot,X)\right]\right\|_\infty,\ \max_j\left\|\sum_{i=1}^{j-1}\mathbb{E}_{X\sim\pi}[p_{i,j}^2(X,\cdot)]\right\|_\infty\right]. \tag{10}$$

Using Lemma 2 twice and the bounds obtained on b_k and σ_k^2 gives for $u>0$,

$$\begin{aligned}
&\left[(2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k}\sigma_k 3\|\Gamma\|\sqrt{ku} + (2\delta_M)^{1/k}k8\|\Gamma\|^2 b_k u\right]^k \\
\le\ & \left[(2\delta_M)^{1/k}\mathbb{E}[Z] + (2\delta_M)^{1/k}3\|\Gamma\|(B^2A^{k-2})^{1/k}\sqrt{nku} + (2\delta_M)^{1/k}8\|\Gamma\|^2((nA^2)A^{k-2})^{1/k}ku\right]^k \\
\le\ & (1+\varepsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + (1+\varepsilon^{-1})^{k-1}\Bigg[(2\delta_M)^{1/k}8\|\Gamma\|^2((nA^2)A^{k-2})^{1/k}ku \\
&\quad + (2\delta_M)^{1/k}3\|\Gamma\|(B^2A^{k-2})^{1/k}\sqrt{nku}\Bigg]^k \\
\le\ & (1+\varepsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + 2\delta_M(1+\varepsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k(nA^2)A^{k-2}(ku)^k \\
&\quad + (1+\varepsilon)^{k-1}(1+\varepsilon^{-1})^{k-1}2\delta_M\left(3\|\Gamma\|\right)^k B^2A^{k-2}(nku)^{k/2}.
\end{aligned}$$

So, setting

$$\begin{aligned}
w_n^k := &\left((1+\varepsilon)^{k-1}2\delta_M\left(\mathbb{E}[Z]\right)^k + 2\delta_M(1+\varepsilon^{-1})^{2k-2}\left(8\|\Gamma\|^2\right)^k(nA^2)A^{k-2}(ku)^k\right. \\
&\left. + (1+\varepsilon)^{k-1}(1+\varepsilon^{-1})^{k-1}2\delta_M\left(3\|\Gamma\|\right)^k B^2A^{k-2}(nku)^{k/2},\right.
\end{aligned}$$

we have

$$\mathbb{P}\left(V_n^k \le w_n^k \quad \forall k \ge 2\right) \ge 1 - 1.62e^{-u}, \tag{11}$$

where the dependence in u of w_n^k is leaved implicit.

Upper-bounding U_n using the martingale structure Let

$$T+1 := \inf\{l \in \mathbb{N} : V_l^k \ge w_n^k \text{ for some } k \ge 2\}.$$

Then, the event $\{T \le l\}$ depends only on $X_1, \dots, X_l$ for all $l \ge 1$. Hence, T is a stopping time for the filtration $(\mathscr{G}_l)_l$ where $\mathscr{G}_l = \sigma((X_i)_{i\in[l]})$ and we deduce that $U_l^T := U_{l\wedge T}$ for $l = 0, \dots, n$ is a martingale with respect to $(\mathscr{G}_l)_l$ with $U_0^T = U_0 = 0$ and $U_1^T = U_1 = 0$. We remark that $U_j^T - U_{j-1}^T = U_j - U_{j-1}$ if $T \ge j$ and zero otherwise, and that $\{T \ge j\}$ is $\mathscr{G}_{j-1}$ measurable. Then, the angle brackets of this martingale admit the following bound:

$$\begin{aligned}
A_n^k(U^T) &= \sum_{j=2}^n \mathbb{E}_{j-1}[(U_j^T - U_{j-1}^T)^k] \\
&= \sum_{j=2}^n \mathbb{E}_{j-1}|U_j - U_{j-1}|^k \mathbb{1}_{T\ge j} \\
&= \sum_{j=2}^n \mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1} h(X_i, X_{j-1}, X_j)\right|^k \mathbb{1}_{T\ge j} \\
&= \sum_{j=2}^{n-1} V_j^k \mathbb{1}_{T=j} + V_n^k \mathbb{1}_{T\ge n} \\
&\le w_n^k \left(\sum_{j=2}^{n-1} \mathbb{1}_{T=j} + \mathbb{1}_{T\ge n}\right) \\
&\le w_n^k,
\end{aligned}$$

since, by definition of T, $V_j^k \le w_n^k$ for all k on $\{T \ge j\}$. Hence, Lemma 1 applied to the martingale U_n^T implies

$$\mathbb{E}e^{\alpha U_n^T} \le \exp\left(\sum_{k\ge2} \frac{\alpha^k}{k!} w_n^k\right).$$

Also, since V_n^k is nondecreasing in n for each k, inequality (11) implies that

$$\mathbb{P}(T < n) \le \mathbb{P}\left(V_n^k \ge w_n^k \quad \text{for some } k \ge 2\right) \le 1.62e^{-u}.$$

Thus we deduce that for all $s \ge 0$,

$$\mathbb{P}(U_n \ge s) \le \mathbb{P}(U_n^T \ge s, T \ge n) + \mathbb{P}(T < n) \le e^{-\alpha s} \exp\left(\sum_{k\ge2} \frac{\alpha^k}{k!} w_n^k\right) + 1.62e^{-u}. \tag{12}$$

The final step of the proof consists in simplifying $\exp\left(\sum_{k\ge2} \frac{\alpha^k}{k!} w_n^k\right)$.

$$\begin{aligned}
\sum_{k\ge2} \frac{\alpha^k}{k!} w_n^k &= 2\delta_M \sum_{k\ge2} \frac{\alpha^k}{k!} (1+\varepsilon)^{k-1} (\mathbb{E}[Z])^k \\
&\quad + 2\delta_M \sum_{k\ge2} \frac{\alpha^k}{k!} (2+\varepsilon+\varepsilon^{-1})^{k-1} (3\|\Gamma\|)^k B^2 A^{k-2} (nku)^{k/2} \\
&\quad + 2\delta_M \sum_{k\ge2} \frac{\alpha^k}{k!} (1+\varepsilon^{-1})^{2k-2} \left(8\|\Gamma\|^2\right)^k (nA^2) A^{k-2} (ku)^k \\
&:= a_1 + a_2 + a_3.
\end{aligned}$$

Bounding a_3. Using the inequality $k! \geq (k/e)^k$, we have,

$$\begin{aligned}
a_3 &\leq 2\delta_M \sum_{k\geq 2} \alpha^k (1+\varepsilon^{-1})^{2k-2} \left(8\|\Gamma\|^2\right)^k (nA^2) A^{k-2} (eu)^k \\
&= 2\delta_M \alpha^2 \left[\sqrt{n} A (1+\varepsilon^{-1}) 8\|\Gamma\|^2 eu\right]^2 \sum_{k\geq 2} \alpha^{k-2} (1+\varepsilon^{-1})^{2(k-2)} \left(8\|\Gamma\|^2\right)^{k-2} A^{k-2} (eu)^{k-2} \\
&= \frac{2\delta_M \alpha^2 \left[\sqrt{n} A (1+\varepsilon^{-1}) 8\|\Gamma\|^2 eu\right]^2}{1-\alpha(1+\varepsilon^{-1})^2 \left(8\|\Gamma\|^2\right) Aeu}, \quad \text{for } \alpha < \left((1+\varepsilon^{-1})^2 \left(8\|\Gamma\|^2\right) Aeu\right)^{-1}.
\end{aligned}$$

Bounding a_2. We use the inequality $k! \geq k^{k/2}$ because $(k/e)^k > k^{k/2}$ for $k \geq e^2$ and for k smaller, the inequality follows by direct verification. Hence,

$$\begin{aligned}
a_2 &\leq 2\delta_M \sum_{k\geq 2} \alpha^k (2+\varepsilon+\varepsilon^{-1})^{k-1} (3\|\Gamma\|)^k B^2 A^{k-2} (nu)^{k/2} \\
&= 2\delta_M (2+\varepsilon+\varepsilon^{-1}) \alpha^2 \left[3\|\Gamma\| B\sqrt{nu}\right]^2 \sum_{k\geq 2} \alpha^{k-2} (2+\varepsilon+\varepsilon^{-1})^{k-2} (3\|\Gamma\|)^{k-2} A^{k-2} (nu)^{(k-2)/2} \\
&= \frac{2\delta_M (2+\varepsilon+\varepsilon^{-1}) \alpha^2 \left[3\|\Gamma\| B\sqrt{nu}\right]^2}{1-\alpha(2+\varepsilon+\varepsilon^{-1}) (3\|\Gamma\|) A (nu)^{1/2}},
\end{aligned}$$

for $\alpha < \left((2+\varepsilon+\varepsilon^{-1}) (3\|\Gamma\|) A (nu)^{1/2}\right)^{-1}$.

Bounding a_1.

The way we bound a_1 is the only part of the proof that needs to be modified to get the concentration result when Assumption 4.(i) or Assumption 4.(ii) holds. This is where we can use different Bernstein concentration inequalities. Here we present the approach when $h_{i,j} \equiv h_{1,j}$, $\forall i,j$ (i.e. when Assumption 4.(i) is satisfied). We refer to Section C.4 for the details regarding the way we bound a_1 when Assumption 4.(ii) holds.

Using Jensen inequality and Lemma 3, we obtain

$$\begin{aligned}
(\mathbb{E}[Z])^k \leq \mathbb{E}[Z^k] & \\
&= \mathbb{E}\left[\left(\sup_{\xi\in\mathscr{L}_k} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}_{X'_j}[p_{i,j}(X_i, X'_j)\xi_j(X'_j)]\right)^k\right] \\
&= \mathbb{E}\left[\sum_{j=2}^{n} \mathbb{E}_{X'_j}\left[\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^k\right]\right] \\
&= \sum_{j=2}^{n} \mathbb{E}\left[\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^k\right],
\end{aligned}$$

where we recall that $\mathbb{E}_{X'_j}$ denotes the expectation on the random variable X'_j. One can remark that conditionally to X'_j, the quantity $\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)$ is a sum of function of the Markov chain $(X_i)_{i\geq 1}$. Hence to control this term, we apply a Bernstein inequality for Markov chains.

Let us consider some $j \in [n]$ and some $x \in E$. We define

$$\forall l \in \{0,\dots,n\}, \quad Z_l^j(x) = \sum_{i=m(S_l+1)}^{m(S_{l+1}+1)-1} p_{i,j}(X_i, x).$$

By convention, we set $p_{i,j} \equiv 0$ for any $i \geq j$. Let us consider $N_j = \sup\{i \in \mathbb{N} \,:\, mS_{i+1} + m - 1 \geq j - 1\}$. Then using twice Lemma 2, we have

$$\begin{aligned}
\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X_j')\right|^k &= \left|\sum_{l=0}^{N_j} Z_l^j(x) + \sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right|^k \\
&\leq \left(\frac{3}{2}\right)^{k-1} \left|\sum_{l=1}^{N_j} Z_l^j(x)\right|^k + 3^{k-1} \left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right|^k \\
&\leq \left(\frac{9}{4}\right)^{k-1} \left|\sum_{l=0}^{\lfloor N_j/2 \rfloor} Z_{2l}^j(x)\right|^k + \left(\frac{9}{2}\right)^{k-1} \left|\sum_{l=0}^{\lfloor (N_j-1)/2 \rfloor} Z_{2l+1}^j(x)\right|^k + 3^{k-1} \left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right|^k .
\end{aligned} \tag{13}$$

We have $|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)| \leq AmT_{N_j+1}$. So using the definition of the Orlicz norm and the fact that the random variables $(T_i)_{i\geq 2}$ are i.i.d., it holds for any $t \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right| \geq t\right) &\leq \mathbb{P}(T_{N_j+1} \geq \frac{t}{Am}) \\
&\leq \mathbb{P}(\max(T_1, T_2) \geq \frac{t}{Am}) \\
&\leq \mathbb{P}(T_1 \geq \frac{t}{Am}) + \mathbb{P}(T_2 \geq \frac{t}{Am}) \\
&\leq 4\exp(-\frac{t}{Am\tau}).
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right|^k\right] &= 4\int_0^{+\infty} \mathbb{P}\left(\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, x)\right|^k \geq t\right) dt \\
&\leq 4\int_0^{+\infty} \exp(-\frac{t^{1/k}}{Am\tau}) \\
&\leq 4(Am\tau)^k \int_0^{+\infty} \exp(-v) k v^{k-1} dv \\
&= 4(Am\tau)^k k!,
\end{aligned}$$

where we used that if G is an exponential random variable with parameter 1, then for any $p \in \mathbb{N}$, $\mathbb{E}[G^p] = p!$.

The random variable $Z_{2l}^j(x)$ is $\sigma(X_{m(S_{2l}+1)}, \ldots, X_{m(S_{2l+1}+1)-1})$-measurable. Let us insist that this holds because we consider that $h_{i,j} \equiv h_{1,j}$, $\forall i, j$ which implies that $p_{i,j} \equiv p_{1,j}$, $\forall i, j$. Hence for any $x \in E$, the random variables $(Z_{2l}^j(x))_l$ are independent (see Section 2.3). Moreover, one has that for any l, $\mathbb{E}[Z_{2l}^j(x)] = 0$. This is due to [48, Eq.(17.23) Theorem 17.3.1] together with Assumption 3 which gives that

$$\forall x' \in E, \quad \mathbb{E}_{X\sim\pi}[p_{i,j}(X, x')] = 0.$$

Let us finally notice that for any $x \in E$ and any $l \geq 0$, $|Z_{2l}^j(x)| \leq AmT_{2l+1}$, so $\|Z_{2l}^j(x)\|_{\psi_1} \leq Am \max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq Am\tau$. First, we use Lemma 5 to obtain that

$$\mathbb{E}\left|\sum_{l=0}^{\lfloor N_j/2 \rfloor} Z_{2l}^j(x)\right|^k \leq \mathbb{E} \max_{0\leq s\leq n-1} \left|\sum_{l=0}^{s} Z_{2l}^j(x)\right|^k \leq 2 \times 4^k \mathbb{E}\left|\sum_{l=0}^{n-1} Z_{2l}^j(x)\right|^k ,$$

where for the last inequality we gathered (15) with the left hand side of (14) from Lemma 5.

Lemma 5 *(see [18, Lemma 1.2.6])*
Let us consider some separable Banach space B endowed with the norm $\|\cdot\|$. Let X_i, $i \le n$, be independent centered B-valued random variables with norms L_p for some $p \ge 1$ and let ε_i be independent Rademacker random variables independent of the variables X_i. Then

$$2^{-p}\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i X_i\right\|^p \le \mathbb{E}\left\|\sum_{i=1}^{n} X_i\right\|^p \le 2^p\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i X_i\right\|^p, \tag{14}$$

and

$$\mathbb{E}\max_{k\le n}\left\|\sum_{i=1}^{k} X_i\right\|^p \le 2^{p+1}\mathbb{E}\left\|\sum_{i=1}^{n}\varepsilon_i X_i\right\|^p \tag{15}$$

Similarly, the random variables $(Z^j_{2l+1}(x))_l$ are independent and satisfy for any l, $\mathbb{E}[Z^j_{2l+1}(x)] = 0$. With an analogous approach, we get that

$$\mathbb{E}\left|\sum_{l=0}^{\lfloor (N_j-1)/2\rfloor} Z^j_{2l+1}(x)\right|^k \le \mathbb{E}\max_{0\le s\le n-1}\left|\sum_{l=0}^{s} Z^j_{2l+1}(x)\right|^k \le 2\times 4^k\mathbb{E}\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(x)\right|^k.$$

Let us denote for any $j \in [n]$, $\mathbb{E}_{|X'_j}$ the conditional expectation with respect to the σ-algebra $\sigma(X'_j)$. Coming back to (13), we proved that

$$\begin{aligned}
&\mathbb{E}_{|X'_j}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^k \\
&\le \left(\frac{9}{4}\right)^{k-1}\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{\lfloor N_j/2\rfloor} Z^j_{2l}(X'_j)\right|^k + \left(\frac{9}{2}\right)^{k-1}\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{\lfloor (N_j-1)/2\rfloor} Z^j_{2l+1}(X'_j)\right|^k + 3^{k-1}\mathbb{E}_{|X'_j}\left|\sum_{i=m(S_{N_j}+1)}^{j-1} p_{i,j}(X_i, X'_j)\right|^k \\
&\le 2\times 9^k\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(X'_j)\right|^k + 2\times 18^k\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z^j_{2l}(X'_j)\right|^k + 4(3Am\tau)^k k!.
\end{aligned} \tag{16}$$

It remains to bound the two expectations in (16). The two latter expectations will be control similarly and we give the details for the first one. We use the following Bernstein's inequality with the sequence of random variables $(Z^j_{2l+1}(x))_l$.

Lemma 6 *(Bernstein's ψ_1 inequality, [61, Lemma 2.2.11] and the subsequent remark).*
If $Y_1, \dots, Y_n$ are independent random variables such that $\mathbb{E}Y_i = 0$ and $\|Y_i\|_{\psi_1} \le \tau$, then for every $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Y_i\right| > t\right) \le 2\exp\left(-\frac{1}{K}\min\left(\frac{t^2}{n\tau^2}, \frac{t}{\tau}\right)\right),$$

for some universal constant $K > 0$ ($K = 8$ fits).

We obtain

$$\mathbb{P}\left(\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(x)\right| > t\right) \le 2\exp\left(-\frac{1}{K}\min\left(\frac{t^2}{nA^2m^2\tau^2}, \frac{t}{Am\tau}\right)\right).$$

We deduce that for any $x \in E$, any $j \in [n]$ and any $t \ge 0$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(x)\right|^k\right] \\
&= \int_0^\infty \mathbb{P}\left(\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(x)\right|^k > t\right) dt
\end{aligned}$$

$$= 2\int_0^\infty \exp\left(-\frac{1}{K}\min\left(\frac{t^{2/k}}{nA^2m^2\tau^2}, \frac{t^{1/k}}{Am\tau}\right)\right)dt.$$

Let us remark that

$$\frac{t^{2/k}}{A^2m^2n\tau^2} \leq \frac{t^{1/k}}{Am\tau} \Leftrightarrow t \leq (nA\tau m)^k.$$

Hence for any $j \in [n]$,

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(X_j')\right|^k\right] \\
&\leq 2\int_0^{(nA\tau m)^k} \exp\left(-\frac{t^{2/k}}{KnA^2m^2\tau^2}\right)dt + 2\int_0^\infty \exp\left(-\frac{t^{1/k}}{KAm\tau}\right)dt. \\
&\leq 2\int_0^{n/K} \exp(-v)\frac{k}{2}v^{k/2-1}\left(\sqrt{K}n^{1/2}A\tau m\right)^k dv + 2\int_0^\infty \exp(-v)\,kv^{k-1}(KAm\tau)^k dv. \\
&\leq 2\int_0^{n/K} \exp(-v)\frac{k}{2}v^{k/2-1}\left(\sqrt{K}n^{1/2}A\tau m\right)^k dv + 2k\times(k-1)!(KAm\tau)^k \\
&\leq k\left(\sqrt{K}n^{1/2}A\tau m\right)^k \int_0^{n/K} \exp(-v)\,v^{k/2-1}dv + 2k!(KAm\tau)^k,
\end{aligned}$$

where we used again that if G is an exponential random variable with parameter 1, then for any $p \in \mathbb{N}$, $\mathbb{E}[G^p] = p!$. Since for any real $l \geq 1$,

$$\begin{aligned}
&\int_0^{n/K} \exp(-v)\,v^{l-1}dv \\
&= \sum_{r=0}^{+\infty}\frac{(-1)^r}{r!}\int_0^{n/K} v^{r+l-1}dv \\
&= \sum_{r=0}^{+\infty}\frac{(-1)^r}{r!}\frac{1}{r+l}(n/K)^{r+l} \\
&\leq (n/K)^l\sum_{r=0}^{+\infty}\frac{(-1)^r}{r!}\frac{1}{l}(n/K)^r \\
&\leq \frac{(n/K)^l}{l}e^{-\frac{n}{K}},
\end{aligned}$$

we get that

$$k\left(\sqrt{K}n^{1/2}A\tau m\right)^k \int_0^{n/K} \exp(-v)\,v^{k/2-1}dv \leq 2\left(\sqrt{K}n^{1/2}A\tau m\right)^k e^{-n/K}(n/K)^{k/2} = 2\,(KnA\tau m)^k\,e^{-n/K}.$$

Hence we proved that for some universal constant $K > 1$,

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{l=0}^{n-1} Z_{2l+1}^j(x)\right|^k\right] &\leq 2\,(KnA\tau m)^k\,e^{-n/K} + 2k!(KAm\tau)^k \\
&\leq 4k!(K^2Am\tau)^k,
\end{aligned}$$

since for all $k \geq 2$, $e^{-n/K}(n/K)^k/(k!) \leq 1$. Using a similar approach, one can show the same bound for the

second expectation in (16). We proved that for some universal constant $K > 1$,

$$\begin{aligned}
\mathbb{E}[Z]^k &\le \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{i=1}^{j-1} p_{i,j}(X_i, X'_j)\right|^k\right] \\
&\le 2\times 9^k \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z^j_{2l+1}(X'_j)\right|^k\right] + 2\times 18^k \sum_{j=2}^{n} \mathbb{E}\left[\mathbb{E}_{|X'_j}\left|\sum_{l=0}^{n-1} Z^j_{2l}(X'_j)\right|^k\right] + 4\sum_{j=2}^{n}(3Am\tau)^k k! \\
&\le 2n\times 18^k \times 4k!(K^2Am\tau)^k + 4n(3Am\tau)^k k! \\
&= 16n\times k!(KAm\tau)^k,
\end{aligned}$$

where in the last inequality, we still call K the universal constant defined by $18K^2$.

$$\begin{aligned}
a_1 &= 2\delta_M \sum_{k\ge 2} \frac{\alpha^k}{k!}(1+\varepsilon)^{k-1}(\mathbb{E}[Z])^k \\
&\le 32\delta_M n \sum_{k\ge 2} \alpha^k (1+\varepsilon)^{k-1}(KAm\tau)^k \\
&\le 32\delta_M n\alpha^2(1+\varepsilon)[KAm\tau]^2 \sum_{k\ge 2} \alpha^{k-2}(1+\varepsilon)^{k-2}(KAm\tau)^{k-2} \\
&\le \frac{32\delta_M n\alpha^2(1+\varepsilon)[KAm\tau]^2}{1-\alpha(1+\varepsilon)KAm\tau},
\end{aligned}$$

for $0 < \alpha < ((1+\varepsilon)KAm\tau)^{-1}$. Putting altogether we obtain

$$\exp\left(\sum_{k\ge 2}\frac{\alpha^k}{k!}w_n^k\right) \le \exp\left(\frac{\alpha^2 W^2}{1-\alpha c}\right),$$

where

$$\begin{aligned}
W = {} & 6\sqrt{\delta_M}(1+\varepsilon)^{1/2} n^{1/2} KA\tau m \\
& + \sqrt{2\delta_M}(2+\varepsilon+\varepsilon^{-1})^{1/2} 3\|\Gamma\| B\sqrt{nu} + \sqrt{2\delta_M} A(1+\varepsilon^{-1}) 8\|\Gamma\|^2 \sqrt{n} eu,
\end{aligned}$$

and

$$c = \max\left[(1+\varepsilon)KA\tau m\ ,\ (2+\varepsilon+\varepsilon^{-1})\,(3\|\Gamma\|)A(nu)^{1/2}\ ,\ (1+\varepsilon^{-1})^2\left(8\|\Gamma\|^2\right)Aeu\right].$$

Using this estimate in (12) and taking $s = 2W\sqrt{u} + cu$ and $\alpha = \sqrt{u}/(W + c\sqrt{u})$ in this inequality yields

$$\mathbb{P}\left(U_n \ge 2W\sqrt{u} + cu\right) \le e^{-u} + 1.62e^{-u} \le (1+e)e^{-u}.$$

By taking $\varepsilon = 1/2$, we deduce that for any $u \ge 0$, it holds with probability at least $1-(1+e)e^{-u}$

$$\begin{aligned}
& \sum_{i<j} h_j^{(0)}(X_i, X_{j-1}, X_j) \\
\le\ & 12\sqrt{\delta_M} KA\tau m\sqrt{nu} + 18\sqrt{\delta_M}\|\Gamma\| B\sqrt{n}u + 100\sqrt{\delta_M}\|\Gamma\|^2 A\sqrt{n}eu^{3/2} \\
& + 3KA\tau mu + 27A\|\Gamma\|\sqrt{n}u^{3/2} + 72A\|\Gamma\|^2 eu^2,
\end{aligned}$$

Denoting

$$\kappa := \max\left(12\sqrt{\delta_M}KA\tau m\ ,\ 18\sqrt{\delta_M}\|\Gamma\|, 100\sqrt{\delta_M}\|\Gamma\|^2 e, 3K\tau m, 72\|\Gamma\|^2 e\right)$$

we have with probability at least $1-(1+e)e^{-u}$

$$\sum_{i<j} h_j^{(0)}(X_i, X_{j-1}, X_j) \le \kappa\left(A\sqrt{n}\sqrt{u} + (A + B\sqrt{n})u + 2A\sqrt{n}u^{3/2} + Au^2\right).$$

4.2 Reasoning by descending induction with a logarithmic depth

As previously explained, we apply a proof similar the one of the previous subsection on the $t_n := \lceil \log n \rceil$ first terms of the decomposition (9). Let us give the key elements to justify such approach by considering the second term of the decomposition (9), namely

$$\begin{aligned}
&\sum_{i<j} \big(\mathbb{E}_{j-1}\big[h_{i,j}(X_i,X_j)\big] - \mathbb{E}_{j-2}\big[h_{i,j}(X_i,X_j)\big] \big) \\
= &\sum_{i=1}^{n-2}\sum_{j=i+2}^{n} h^{(1)}_{i,j}(X_i,X_{j-2},X_{j-1}) + \sum_{i=1}^{n-1} u_i(X_i) \\
= &\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} h^{(1)}_{i,j}(X_i,X_{j-1},X_j) + \sum_{i=1}^{n-1} u_i(X_i), \qquad (17)
\end{aligned}$$

where

$$h^{(1)}_{i,j}(x,y,z) = \int_w h_{i,j}(x,w)P(z,dw) - \int_w h_{i,j}(x,w)P^2(y,dw)$$

and

$$u_i(x) = \int_w h_{i,i+1}(x,w)P(x,dw) - \mathbb{E}_{i-1}[h_{i,i+1}(X_i,X_{i+1})].$$

We can upper-bound directly $\left|\sum_{i=1}^{n-1} u_i(X_i)\right|$ by $2n \max_{i,j} \|h_{i,j}\|_\infty$ and we aim at proving a concentration result for the term

$$U^{(1)}_{n-1} := \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} h^{(1)}_{i,j}(X_i,X_{j-1},X_j) = \sum_{j=2}^{n-1}\sum_{i=1}^{j-1} h^{(1)}_{i,j}(X_i,X_{j-1},X_j),$$

using an approach similar to the one of the previous subsection. One can use exactly the same sketch of proof.

- Martingale structure
  Using the notation $Y^{(1)}_j = \sum_{i=1}^{j-1} h^{(1)}_{i,j}(X_i,X_{j-1},X_j)$, we have $U^{(1)}_{n-1} = \sum_{j=2}^{n-1} Y^{(1)}_j$ which shows that $(U^{(1)}_n)_n$ is a martingale with respect to the σ-algebras $(G_l)_l$. Indeed, we have $\mathbb{E}_{j-1}[Y^{(1)}_j] = 0$.

- Talagrand's inequality
  To upper-bound $(V^k_n)_n$, we split it as previously namely

$$\begin{aligned}
V^k_n &:= \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} h^{(1)}_{i,j}(X_i,X_{j-1},X_j) \right|^k \\
&= \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} \Big(I^{(1)}_{i,j}(X_i,X_j) - \mathbb{E}_{j-1}[I^{(1)}_{i,j}(X_i,X_j)] \Big) \right|^k,
\end{aligned}$$

  where

$$I^{(1)}_{i,j}(x,z) = \int_w h_{i,j}(x,w)P(z,dw).$$

  Using as previously Lemma 2 with $\varepsilon = 1/2$, we get

$$\begin{aligned}
V^k_n &= \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} \Big(I^{(1)}_{i,j}(X_i,X_j) - \mathbb{E}_{X'\sim\nu}[I^{(1)}_{i,j}(X_i,X')] + \mathbb{E}_{X'\sim\nu}[I^{(1)}_{i,j}(X_i,X')] - \mathbb{E}_{j-1}[I^{(1)}_{i,j}(X_i,X_j)] \Big) \right|^k \\
&\le (3/2)^{k-1} \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} \Big(I^{(1)}_{i,j}(X_i,X_j) - \mathbb{E}_{X'\sim\nu}[I^{(1)}_{i,j}(X_i,X')] \Big) \right|^k \\
&\quad + 3^{k-1} \sum_{j=2}^{n-1} \mathbb{E}_{j-1} \left| \sum_{i=1}^{j-1} \Big(\mathbb{E}_{X'\sim\nu}[I^{(1)}_{i,j}(X_i,X')] - \mathbb{E}_{j-1}[I^{(1)}_{i,j}(X_i,X_j)] \Big) \right|^k.
\end{aligned}$$

Again, basic computations and Jensen's inequality lead to

$$\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}\Big(\mathbb{E}_{X'\sim\nu}[I_{i,j}^{(1)}(X_i,X')]-\mathbb{E}_{j-1}[I_{i,j}^{(1)}(X_i,X_j)]\Big)\right|^k=\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}p_{i,j}^{(1)}(X_i,X_j)\right|^k,$$

where

$$p_{i,j}^{(1)}(x,z):=I_{i,j}^{(1)}(x,z)-\mathbb{E}_{X'\sim\nu}[I_{i,j}^{(1)}(x,X')].$$

Hence, using Assumption 1 and Lemma 5 exactly like in the previous section, we get

$$\begin{aligned}V_n^k&=2\times 3^{k-1}\sum_{j=2}^{n-1}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}p_{i,j}^{(1)}(X_i,X_j)\right|^k\\&\leq 2\times 3^{k-1}\delta_M\sum_{j=2}^{n-1}\mathbb{E}_{X_j'}\left|\sum_{i=1}^{j-1}p_{i,j}^{(1)}(X_i,X_j')\right|^k.\end{aligned}$$

Then, one can use the same duality trick to show that the V_n^k can be controlled using the supremum of a sum of functions of the Markov chain $(X_i)_{i\geq 1}$ using [55, Theorem 3].

- Bounding $\exp(w_n^k\alpha^k/k!)$

  The terms a_2 and a_3 can be bounded in a similar way. For the term a_1, we only need to show that $p_{i,j}^{(1)}$ satisfies $\mathbb{E}_{X_i\sim\pi}|p_{i,j}^{(1)}(X_i,z)]=0,\ \forall z\in E$ in order to apply as previously a Bernstein's type inequality.

$$\begin{aligned}&\mathbb{E}_{X_i\sim\pi}|p_{i,j}^{(1)}(X_i,z)]\\&=\int_{x_i}d\pi(x_i)\int_w h_{i,j}(x_i,w)P(z,dw)-\mathbb{E}_{X\sim\pi}\mathbb{E}_{X'\sim\nu}[I_{i,j}^{(1)}(X,X')]\\&=\mathbb{E}_\pi[h_{i,j}(X,\cdot)]-\mathbb{E}_\pi[h_{i,j}(X,\cdot)]\quad\text{(Using Assumption 3)}\\&=0.\end{aligned}$$

- Conclusion of the proof

  The key element to conclude is to compute the constants A and B. We note A_1 and B_1 the counterparts of the constants A and B. One can easily see that $A_1=A$. Let us give details about B_1.

  For any $x\in E$,

$$\begin{aligned}&\mathbb{E}_{X'\sim\nu}\Big[(p_{i,j}^{(1)})^2(x,X')\Big]\\&=\int_z\Big(I_{i,j}^{(1)}(x,z)-\mathbb{E}_{X'\sim\nu}[I_{i,j}^{(1)}(x,X')]\Big)^2\,d\nu(z)\\&=\int_z\left(\int_w h_{i,j}(x,w)P(z,dw)-\int_w\int_a h_{i,j}(x,w)P(a,dw)d\nu(a)\right)^2 d\nu(z)\\&=\int_z\left(\int_w h_{i,j}(x,w)P(z,dw)\right)^2 d\nu(z)+\int_z\left(\int_w\int_a h_{i,j}(x,w)P(a,dw)d\nu(a)\right)^2 d\nu(z)\\&\quad-2\left(\int_z\int_w h_{i,j}(x,w)P(z,dw)d\nu(z)\right)\times\left(\int_w\int_a h_{i,j}(x,w)P(a,dw)d\nu(a)\right)\\&=\int_z\left(\int_w h_{i,j}(x,w)P(z,dw)\right)^2 d\nu(z)-\left(\int_z\int_w h_{i,j}(x,w)P(z,dw)d\nu(z)\right)^2\\&\leq\int_w h_{i,j}(x,w)^2\left[\int_z P(z,dw)d\nu(z)\right]-\left(\int_w h_{i,j}(x,w)\int_z P(z,dw)d\nu(z)\right)^2\quad\text{(Using Jensen inequality)}\\&=\mathrm{Var}_{X'\sim\mathbb{Q}}(h_{i,j}(x,X')),\end{aligned}$$

where $\mathrm{Var}_{X'\sim\mathbb{Q}}$ denotes the variance with respect to the measure $\mathbb{Q}$ defined by $\forall A\in\Sigma,\quad \mathbb{Q}(A)=\int_z P(z,A)d\nu(z)$. From Assumption 2.$(ii)$, we have that $\forall A\in\Sigma,\quad \mathbb{Q}(A)=\nu(A)$. Hence, we get that for any $x\in E$,

$$\mathbb{E}_{X'\sim\nu}\left[(p_{i,j}^{(1)})^2(x,X')\right]\leq \mathrm{Var}_{X'\sim\nu}(h_{i,j}(x,X'))=\mathbb{E}_{X'\sim\nu}[p_{i,j}^2(x,X')].$$

Noticing that Assumption 2.(ii) gives that $\pi=\nu$, we get that

$$B_1^2:=\max\left[\max_i\left\|\sum_{j=i+1}^n\mathbb{E}_{X\sim\nu}\left[(p_{i,j}^{(1)})^2(\cdot,X)\right]\right\|_\infty,\ \max_j\left\|\sum_{i=1}^{j-1}\mathbb{E}_{X\sim\pi}[(p_{i,j}^{(1)})^2(X,\cdot)]\right\|_\infty\right]\leq B^2. \quad (18)$$

This allows us to get a concentration inequality similar to the one of the previous subsection, namely for any $u>0$, it holds with probability at least $1-(1+e)e^{-u}$,

$$\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}h_{i,j}^{(1)}(X_i,X_{j-1},X_j)\leq\kappa\left(A\sqrt{n}\sqrt{u}+(A+B\sqrt{n})u+2A\sqrt{n}u^{3/2}+Au^2\right)$$

Going back to (17), we get that for any $u>0$, it holds with probability at least $1-(1+e)e^{-u}$,

$$\begin{aligned}&\sum_{i<j}\left(\mathbb{E}_{j-1}\left[h_{i,j}(X_i,X_j)\right]-\mathbb{E}_{j-2}\left[h_{i,j}(X_i,X_j)\right]\right)\\ \leq\ &\kappa\left(A\sqrt{n}\sqrt{u}+(A+B\sqrt{n})u+2A\sqrt{n}u^{3/2}+Au^2\right)+nA \quad (19)\end{aligned}$$

One can do the same analysis for the t_n first terms in the decomposition (9). Hence for any $u>0$, it holds with probability at least $1-(1+e)e^{-u}t_n$,

$$\begin{aligned}&\sum_{k=1}^{t_n}\sum_{i<j}\left(\mathbb{E}_{j-k+1}[h_{i,j}(X_i,X_j)]-\mathbb{E}_{j-k}[h_{i,j}(X_i,X_j)]\right)\\ \leq\ &\kappa t_n\left(A\sqrt{n}\sqrt{u}+(A+B\sqrt{n})u+2A\sqrt{n}u^{3/2}+Au^2\right)+nt_nA. \quad (20)\end{aligned}$$

4.3 Bounding the remaining statistic with uniform ergodicity

In the previous steps of the proof, we decompose U_n in t_n+1 terms (see (9)). The martingale structure of the first t_n terms of this decomposition allowed us to derive a concentration inequality for each of them. It remains to control the last term of this decomposition, namely

$$\sum_{i<j}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right]-\mathbb{E}_\pi[h_{i,j}(X,\cdot)]\right),$$

where $t_n=\lfloor q\log n\rfloor$ with $q>(\log(1/\rho))^{-1}$. In the following, we assume that $t_n\leq n$, otherwise the last term of the decomposition (9) is an empty sum. Using our convention which states that for all $k<1$, $\mathbb{E}_k[\cdot]:=\mathbb{E}[\cdot]$, we need to control

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right]-\mathbb{E}_\pi[h_{i,j}(X,\cdot)]\right).$$

Let us consider $(\widetilde{X}_j)_j$ i.i.d random variables with distribution π, and independent of $(X_i)_{i\geq1}$. Using Assumption 3, we have

$$\begin{aligned}&\left|\sum_{i<j}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right]-\mathbb{E}_\pi[h_{i,j}(X,\cdot)]\right)\right|\\ \leq\ &\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left|\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)-h_{i,j}(X_i,\widetilde{X}_j)\right]\right|\\ \leq\ &(1)+(2)+(3)+(4),\end{aligned}$$

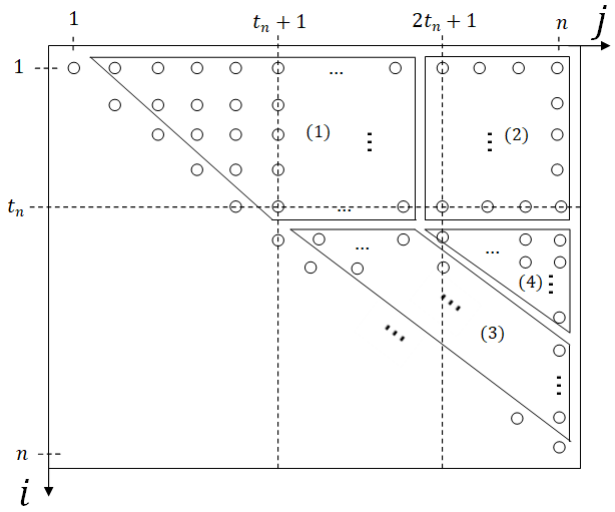


Figure 3: Visualization of the partition of the set $\{(i,j) \in [n]^2 \ : \ i < j\}$ that we use to control the last term of the decomposition (9) of the U-statistic.

with, denoting $H_{ij} = \mathbb{E}_{j-t_n}\left[h_{i,j}(X_i, X_j) - h_{i,j}(X_i, \widetilde{X}_j)\right]$,

$$(1) := \sum_{i=1}^{t_n} \sum_{j=i+1}^{2t_n} \left|H_{ij}\right| \leq 2t_n^2 A,$$

$$(2) := \sum_{i=1}^{t_n} \sum_{j=(2t_n+1)}^{n} \left|H_{ij}\right|,$$

$$(3) := \sum_{i=t_n+1}^{n-1} \sum_{j=i+1}^{n \wedge (t_n+i-1)} \left|H_{ij}\right| \leq n t_n A,$$

$$(4) = \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sum_{j=t_n+i}^{n} \left|H_{ij}\right|.$$

Figure 3 presents a visualization of the way we are partitioning the sum.

Let us upper-bound (2) and (4) to conclude the proof. First note that for $i \leq j - t_n$, it holds

$$\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i, X_j)\right] = \int_z h_{i,j}(X_i, z) P^{t_n}(X_{j-t_n}, dz).$$

We start by upper-bounding (2),

$$
\begin{aligned}
(2) &= \sum_{i=1}^{t_n} \sum_{j=2t_n+1}^{n} \left|\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j) - h_{i,j}(X_i,\widetilde{X}_j)\right]\right| \\
&\leq \sum_{i=1}^{t_n} \sup_{x_i} \sum_{j=2t_n+1}^{n} \left|\mathbb{E}_{j-t_n}\left[h_{i,j}(x_i,X_j) - h_{i,j}(x_i,\widetilde{X}_j)\right]\right| \\
&\leq \sum_{i=1}^{t_n} \sup_{x_i} \sum_{j=2t_n+1}^{n} \left|\int_z h_{i,j}(x_i,z)\left(P^{t_n}(X_{j-t_n},dz) - d\pi(z)\right)\right| \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \sum_{i=1}^{t_n} \sum_{j=(2t_n+1)}^{n} \int_z \left|P^{t_n}(X_{j-t_n},dz) - d\pi(z)\right| \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \sum_{i=1}^{t_n} \sum_{j=2t_n+1}^{n} L\rho^{t_n} \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty n t_n L\rho^{t_n}.
\end{aligned}
$$

where $L > 0$ and $0 < \rho < 1$ are constants related to the uniform ergodicity of the Markov chain (see Definition 8). With analogous computations, we upper-bound the term (4)

$$
\begin{aligned}
(4) &= \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sum_{j=t_n+i}^{n} \left|\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j) - h_{i,j}(X_i,\widetilde{X}_j)\right]\right| \\
&\leq \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sup_{x_i} \sum_{j=t_n+i}^{n} \left|\mathbb{E}_{j-t_n}\left[h_{i,j}(x_i,X_j) - h_{i,j}(x_i,\widetilde{X}_j)\right]\right| \\
&\leq \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sup_{x_i} \sum_{j=t_n+i}^{n} \left|\int_z h_{i,j}(x_i,z)\left(P^{t_n}(X_{j-t_n},dz) - d\pi(z)\right)\right| \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sum_{j=t_n+i}^{n} \int_z \left|P^{t_n}(X_{j-t_n},dz) - d\pi(z)\right| \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \sum_{i=t_n+1}^{(n-1)\wedge(n-t_n)} \sum_{j=t_n+i}^{n} L\rho^{t_n} \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty n^2 L\rho^{t_n}.
\end{aligned}
$$

We deduce that

$$
\begin{aligned}
&\sum_{i<j}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}_\pi[h_{i,j}(X_i,\cdot)]\right) \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \left(n^2 L\rho^{t_n} + n t_n L\rho^{t_n} + 4t_n^2 + 2n t_n\right) \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \left(2n^2 L\rho^{t_n} + 6n t_n\right) \quad \text{(Using } t_n \leq n\text{)} \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \left(2Ln^2 n^{q\log(\rho)} + 6nq\log(n)\right).
\end{aligned}
$$

We deduce that

$$
\begin{aligned}
&\frac{2}{n(n-1)}\sum_{i<j}\left(\mathbb{E}_{j-t_n}\left[h_{i,j}(X_i,X_j)\right] - \mathbb{E}_\pi[h_{i,j}(X,\cdot)]\right) \\
&\leq \max_{i,j} \|h_{i,j}\|_\infty \left(4Ln^{q\log(\rho)} + 24q\frac{\log(n)}{n}\right) \\
&\leq A\left(\frac{2L}{n} + 12q\frac{\log(n)}{n}\right), \qquad (21)
\end{aligned}
$$

because $\rho^{q\log(n)} = n^{q\log(\rho)} \leq n^{-1}$. Indeed $1+q\log(\rho) < 0$ because we choose q such that $q > (\log(1/\rho))^{-1}$. Coupling this result with the concentration result (20) concludes the proof of Theorem 1. Hence, for any $u > 0$ it holds with probability at least $1-(1+Ke)e^{-u}t_n$,

$$\begin{aligned}
&\frac{2}{n(n-1)}\sum_{i<j}\big(h_{i,j}(X_i,X_j)-\mathbb{E}_\pi[h_{i,j}(X,\cdot)]\big)\\
\leq\ &\frac{2\kappa t_n}{n(n-1)}\left(A\sqrt{n}\sqrt{u}+(A+B\sqrt{n})u+2A\sqrt{n}u^{3/2}+Au^2\right)+A\left(\frac{4t_n}{n}+(2L+12q)\frac{\log(n)}{n}\right)\\
\leq\ &\frac{2\kappa t_n}{n(n-1)}\left(A\sqrt{n}\sqrt{u}+(A+B\sqrt{n})u+2A\sqrt{n}u^{3/2}+Au^2\right)+A(2L+16q)\frac{\log(n)}{n}.
\end{aligned}$$

5 Proof of Theorem 2

In the proof of Theorem 1, we kept on purpose the distinction between ν and π to clearly justify that our approach still works even if Assumption 2.(ii) does not hold. Hence to prove Theorem 2, one only needs to follow closely the steps of the proof of Theorem 1 and to bound coarsely the constants B^2 and B_1^2 in (10) and in (18) by nA^2.

6 Deviation inequality for the spectrum of signed integral operators

As shown in Section C.5, Theorem 3 is a direct consequence of the concentration result provided by Theorem 7.

Theorem 7 *We keep notations of Section 3.1. Assume that $(X_n)_{n\geq 1}$ is a Markov chain on E satisfying Assumptions 1 and 2.(i) described in Section 2 with invariant distribution π. Let us consider some symmetric kernel $h : E\times E\to\mathbb{R}$, square integrable with respect to $\pi\otimes\pi$. Let us consider some $R\in\mathbb{N}^*$. We assume that there exist continuous functions $\varphi_r : E\to\mathbb{R}$, $r\in I$ (where $I=\mathbb{N}$ or $I=1,\dots,N$) that form an orthonormal basis of $L^2(\pi)$ such that it holds pointwise*

$$h(x,y)=\sum_{r\in I}\lambda_r\varphi_r(x)\varphi_r(y),$$

with

$$\Lambda_R := \sup_{r\in I,\, r\leq R}|\lambda_r| \text{ and } \|\varphi_r\|_\infty\leq\Upsilon_R,\quad \forall r\in I,\ r\leq R.$$

We also define $h_R(x,y)=\sum_{r\in I,\, r\leq R}\lambda_r\varphi_r(x)\varphi_r(y)$ and we assume that $\|h_R\|_\infty, \|h-h_R\|_\infty<\infty$. Then there exists a universal constant $K>0$ such that for any $t>0$, it holds

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2\geq\left(\|h_R\|_\infty^2+\kappa\|h-h_R\|_\infty^2\right)\frac{\log n}{n}+2\sum_{i>R,i\in I}\lambda_i^2+t\right)\\
\leq\ &16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right)+\beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{c}\right]\wedge\left[\frac{t}{c}\right]^{1/2}\right\}\right)\\
+\ &16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda_R^2\Upsilon_R^4}\right).
\end{aligned}$$

where $c=\kappa\|h-h_R\|_\infty$ with $\kappa>0$ depending on δ_M, $\|T_1\|_{\psi_1}$, $\|T_2\|_{\psi_1}$, L, m and ρ. β depends only on ρ.

Proof of Theorem 7. For any integer $R \geq 1$, we denote

$$\begin{aligned} X_{n,R} &:= \frac{1}{\sqrt{n}} \left(\varphi_r(X_i)\right)_{1\leq i\leq n,\, 1\leq r\leq R} \in \mathbb{R}^{n\times R} \\ A_{n,R} &:= \left(X_{n,R}^\top X_{n,R}\right)^{1/2} \in \mathbb{R}^{R\times R} \\ K_R &:= \mathrm{Diag}(\lambda_1, \ldots, \lambda_R) \\ \widetilde{\mathbf{H}}_n^R &:= X_{n,R} K_R X_{n,R}^\top \\ \mathbf{H}_n^R &:= \left((1-\delta_{i,j})\left(\widetilde{\mathbf{H}}_n^R\right)_{i,j}\right)_{1\leq i,j\leq n}. \end{aligned}$$

We remark that $A_{n,R}^2 = I_R + E_{R,n}$ where $\left(E_{R,n}\right)_{r,s} = (1/n)\sum_{i=1}^n \left(\varphi_r(X_i)\varphi_s(X_i) - \delta_{r,s}\right)$ for all $r,s \in [R]$. Denoting $\lambda(\mathbf{H}^R) = (\lambda_1, \ldots, \lambda_R)$, we have

$$\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 \leq 4\left[\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}^R))^2 + \delta_2(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R))^2 + \delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))^2 + \delta_2(\lambda(\mathbf{H}_n^R), \lambda(\mathbf{H}_n))^2\right].$$

Bounding $\delta_2\left(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R)\right)$. Using a singular value decomposition of $X_{n,R}$, one can show that $\lambda(X_{n,R}K_R X_{n,R}^\top) = \lambda(A_{n,R}K_R A_{n,R})$ which leads to

$$\begin{aligned} \delta_2\left(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R)\right) &= \delta_2\left(\lambda(K_R), \lambda(X_{n,R}K_R X_{n,R}^\top)\right) \\ &= \delta_2\left(\lambda(K_R), \lambda(A_{n,R}K_R A_{n,R})\right) \\ &\leq \|K_R - A_{n,R}K_R A_{n,R}\|_F, \end{aligned}$$

where the least inequality follows from Hoffman-Wielandt inequality. Using Equation (4.8) from [40, page 127], we get

$$\delta_2\left(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R)\right)^2 \leq 2\|K_R E_{R,n}\|_F^2 = 2\sum_{1\leq r,s\leq R} \lambda_s^2 \left(\frac{1}{n}\sum_{i=1}^n \varphi_r(X_i)\varphi_s(X_i) - \delta_{r,s}\right)^2. \tag{22}$$

Hence,

$$\begin{aligned} & \mathbb{P}\left(\delta_2\left(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R)\right)^2 \geq t\right) \\ \leq & \sum_{1\leq s,r\leq R} \mathbb{P}\left(\sqrt{2}|\lambda_s| \left|\frac{1}{n}\sum_{i=1}^n \varphi_r(X_i)\varphi_s(X_i) - \delta_{r,s}\right| \geq \sqrt{t}/R\right) \\ \leq & \sum_{1\leq s,r\leq R, \lambda_s\neq 0} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \varphi_r(X_i)\varphi_s(X_i) - \delta_{r,s}\right| \geq \sqrt{t}/(\sqrt{2}R|\lambda_s|)\right) \\ \leq & \sum_{1\leq s,r\leq R, \lambda_s\neq 0} 16\exp\left(-\left(Km^2\tau^2\right)^{-1} \frac{nt}{R^2|\lambda_s|^2\Upsilon_R^4}\right) \\ = & \; 16R^2\exp\left(-\left(Km^2\tau^2\right)^{-1} \frac{nt}{R^2\Lambda_R^2\Upsilon_R^4}\right), \end{aligned}$$

where the last inequality follows from Proposition 4 and where $K > 0$ is a universal constant.

Bounding $\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))^2$.

$$\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))^2 \leq \|\widetilde{\mathbf{H}}_n^R - \mathbf{H}_n^R\|_F^2 = \frac{1}{n^2}\left(\sum_{i=1}^n h_R^2(X_i, X_i)\right) \leq \frac{\|h_R\|_\infty^2}{n}$$

Bounding $\delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2$.

$$\delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2 \le \|\widetilde{\mathbf{H}}_n^R - \widetilde{\mathbf{H}}_n\|_F^2 = \frac{1}{n^2}\left(\sum_{1\le i,j\le n,\, i\neq j}(h-h_R)(X_i,X_j)^2\right).$$

Let us consider,

$$\forall x,y\in E,\quad m_R(x,y) := (h-h_R)^2(x,y) - s_R(x) - s_R(y) - \mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)],$$

where $s_R(x) = \mathbb{E}_\pi[(h-h_R)^2(x,X)] - \mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)]$. One can check that for any $x\in E$, $\mathbb{E}_\pi[m_R(x,X)] = \mathbb{E}_\pi[m_R(X,x)] = 0$. Hence, m_R is π-canonical.

$$\frac{1}{n(n-1)}\left(\sum_{1\le i,j\le n,\, i\neq j}(h-h_R)(X_i,X_j)^2\right) \tag{23}$$

$$= \frac{1}{n(n-1)}\sum_{1\le i,j\le n,\, i\neq j} m_R(X_i,X_j) + \frac{2}{n}\sum_{i=1}^n s_R(X_i) + \mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)]. \tag{24}$$

Using Theorem 2, we get that there exist two constants $\beta,\kappa>0$ such that for any $u\ge 1$, it holds with probability at least $1-\beta e^{-u}\log(n)$,

$$\frac{1}{n(n-1)}\sum_{1\le i,j\le n,\, i\neq j} m_R(X_i,X_j) \le \kappa\|h-h_R\|_\infty \log n\left\{\frac{u}{n}\vee\left\lceil\frac{u}{n}\right\rceil^2\right\}.$$

Let us now consider some $t>0$ such that

$$\kappa\|h-h_R\|_\infty \log n\left\{\frac{u}{n}\vee\left\lceil\frac{u}{n}\right\rceil^2\right\} \le t. \tag{25}$$

The condition (25) is equivalent to

$$u \le n\left\{\frac{t}{\kappa\|h-h_R\|_\infty\log n}\wedge\left(\frac{t}{\kappa\|h-h_R\|_\infty\log n}\right)^{1/2}\right\},$$

which is satisfied in particular if t and u are such that

$$u = \frac{n}{\log n}\left\{\left\lceil\frac{t}{c}\right\rceil\wedge\left\lceil\frac{t}{c}\right\rceil^{1/2}\right\},$$

where $c=\kappa\|h-h_R\|_\infty$. One can finally notice that for this choice of u, the condition $u\ge 1$ holds in particular for n large enough in order to have $n/\log n \ge \kappa\|h-h_R\|_\infty t^{-1}$.

We deduce from this analysis that for any $t>0$, we have for n large enough to satisfy $n/\log n \ge \kappa\|h-h_R\|_\infty t^{-1}$,

$$\mathbb{P}\left(\frac{1}{n(n-1)}\sum_{1\le i,j\le n,\, i\neq j} m_R(X_i,X_j)\ge t\right) \le \beta\log(n)\exp\left(-\frac{n}{\log n}\left\{\left\lceil\frac{t}{c}\right\rceil\wedge\left\lceil\frac{t}{c}\right\rceil^{1/2}\right\}\right).$$

Using Proposition 4, we get that for some universal constant $K>0$,

$$\mathbb{P}\left(\frac{2}{n}\left|\sum_{i=1}^n s_R(X_i)\right|\ge t\right) \le 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right).$$

We deduce that for some universal constant $K>0$ it holds

$$\mathbb{P}\left(\frac{1}{n^2}\left(\sum_{1\le i,j\le n,\, i\neq j}(h-h_R)(X_i,X_j)^2\right) - \mathbb{E}_{\pi\otimes\pi}\left[(h-h_R)^2\right]\ge t\right)$$

$$\le 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right) + \beta\log(n)\exp\left(-\frac{n}{4\log n}\left\{\left\lceil\frac{t}{c}\right\rceil\wedge\left\lceil\frac{t}{c}\right\rceil^{1/2}\right\}\right).$$

Since $\mathbb{E}_{\pi\otimes\pi}\left[(h-h_R)^2(X,Y)\right]=\sum_{i>R,i\in I}\lambda_i^2$, we deduce that

$$\begin{aligned}&\mathbb{P}\left(\delta_2(\lambda(\mathbf{H}_n^R),\lambda(\mathbf{H}_n))^2-\sum_{i>R,i\in I}\lambda_i^2\geq t\right)\\ \leq\;& 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right)+\beta\log(n)\exp\left(-\frac{n}{4\log n}\left\{\left[\frac{t}{c}\right]\wedge\left[\frac{t}{c}\right]^{1/2}\right\}\right).\end{aligned}$$

Hence we proved that for any $u>0$ such that $n/\log n\geq\kappa\|h-h_R\|_\infty u^{-1}$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2\geq\frac{\|h_R\|_\infty^2}{n}+2\sum_{i>R,i\in I}\lambda_i^2+u\right)\\ \leq\;& 16\exp\left(-n\frac{u^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right)+\beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{u}{c}\right]\wedge\left[\frac{u}{c}\right]^{1/2}\right\}\right)\\ +\;& 16R^2\exp\left(-\frac{nu}{Km^2\tau^2R^2\Lambda_R^2\Upsilon_R^4}\right).\end{aligned}$$

Considering $t>0$ and applying the previous inequality with $u=t+\frac{\kappa\|h-h_R\|_\infty\log n}{n}$, we get

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2\geq\left(\|h_R\|_\infty^2+\kappa\|h-h_R\|_\infty^2\right)\frac{\log n}{n}+2\sum_{i>R,i\in I}\lambda_i^2+t\right)\\ \leq\;& 16\exp\left(-n\frac{t^2}{Km^2\tau^2\|h-h_R\|_\infty^2}\right)+\beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{c}\right]\wedge\left[\frac{t}{c}\right]^{1/2}\right\}\right)\\ +\;& 16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda_R^2\Upsilon_R^4}\right).\end{aligned}$$

This concludes the proof of Theorem 7.

7 Proofs of Theorems 4 and 5

In this section, for any $k\geq 0$ we denote $\mathbb{E}_k$ the conditional expectation with respect to the σ-algebra $\sigma(X_1,\dots,X_k)$.

7.1 Proof of Theorem 4

By definition of $\mathscr{M}^n$, we want to bound

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n})-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}M_t\geq\varepsilon\right),$$

which takes the form

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathscr{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]\right]-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[M_t-\mathbb{E}_{t-b_n}[M_t]\right]\geq\varepsilon\right)\\ \leq\;& \mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathscr{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]\right]\geq\varepsilon/2\right)+\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_{t-b_n}[M_t]-M_t\right]\geq\varepsilon/2\right).\end{aligned}\tag{26}$$

Step 1: Martingale difference We first deal with the second term of (26). Note that we can write

$$\sum_{t=c_n}^{n-1}\left[\mathbb{E}_{t-b_n}[M_t]-M_t\right]=\sum_{t=c_n}^{n-1}\sum_{k=1}^{b_n}[\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t]]=\sum_{k=1}^{b_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-k}[M_t]-\mathbb{E}_{t-k+1}[M_t]].$$

Let us consider some $k \in \{1,\dots,b_n\}$, then we have that $V_t^{(k)} = (\mathbb{E}_{t-k}[M_t] - \mathbb{E}_{t-k+1}[M_t])/(n-c_n)$ is a martingale difference sequence, i.e. $\mathbb{E}_{t-k}[V_t^{(k)}] = 0$. Since the loss function is bounded in $[0,1]$, we have $|V_t^{(k)}| \le 2/(n-c_n)$, $t = 1,\dots,n$. Therefore by the Hoeffding-Azuma inequality, $\sum_t V_t^{(k)}$ can be bounded such that

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}[\mathbb{E}_{t-k}[M_t] - \mathbb{E}_{t-k+1}[M_t]] \ge \frac{\varepsilon}{2b_n}\right) \le \exp\left(-\frac{(1-c)n\varepsilon^2}{8b_n^2}\right).$$

We deduce that

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathbb{E}_{t-b_n}[M_t] - M_t\big] \ge \varepsilon/2\right) \le b_n \exp\left(-\frac{(1-c)n\varepsilon^2}{8b_n^2}\right). \tag{27}$$

Step 2: Symmetrization by a ghost sample In this step we bound the first term in (26). Let us start by introducing a ghost sample $\{\xi_j\}_{1\le j\le n}$, where the random variables ξ_j i.i.d with distribution π. Recall the definition of M_t and define $\widetilde{M}_t$ as

$$M_t = \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\ell(h_{t-b_n}, X_t, X_i), \qquad \widetilde{M}_t = \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\ell(h_{t-b_n}, X_t, \xi_i).$$

The difference between $\widetilde{M}_t$ and M_t is that M_t is the sum of the loss incurred by h_{t-b_n} on the current instance X_t and all the previous examples X_j, $j = 1,\dots,t-b_n$ on which h_{t-b_n} is trained, while $\widetilde{M}_t$ is the loss incurred by the same hypothesis h_{t-b_n} on the current instance X_t and an independent set of examples ξ_j, $j = 1,\dots,t-b_n$.

First remark that we have

$$\begin{aligned}
&\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathscr{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[M_t]\big] \\
= \;& \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathscr{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[\widetilde{M}_t]\big] + \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\big].
\end{aligned}$$

Since ℓ is in $[0,1]$, the first term can be bounded directly using the uniform ergodicity of the Markov chain $(X_i)_i$ as follows

$$\begin{aligned}
&\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathscr{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[\widetilde{M}_t]\big] \\
= \;& \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\big(d\pi(x)\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n}, x, X)] - P^{b_n}(X_{t-b_n}, dx)\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n}, x, X)]\big) \\
= \;& \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\mathbb{E}_{X\sim\pi}[\ell(h_{t-b_n}, x, X)]\big(d\pi(x) - P^{b_n}(X_{t-b_n}, dx)\big) \\
\le \;& \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\int_{x\in E}\big|d\pi(x) - P^{b_n}(X_{t-b_n}, dx)\big| \\
\le \;& L\rho^{b_n},
\end{aligned}$$

where we used Definition 8.

It remains to control

$$\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\big],$$

and we follow an approach similar to [63]. Let us remind that M_t and $\widetilde{M}_t$ depend on the hypothesis h_{t-b_n} and let us define $L_t(h_{t-b_n}) = \left[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\right]$. We have

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1} L_t(h_{t-b_n}) \geq \varepsilon\right) \\
\leq\ &\mathbb{P}\left(\sup_{\widehat{h}_{c_n-b_n},\ldots,\widehat{h}_{n-1-b_n}} \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} L_t(\widehat{h}_{t-b_n}) \geq \varepsilon\right) \\
\leq\ &\sum_{t=c_n}^{n-1} \mathbb{P}\left(\sup_{\widehat{h}\in\mathscr{H}} L_t(\widehat{h}) \geq \varepsilon\right). \qquad (28)
\end{aligned}$$

To bound the right hand side of (28) we give first the following Lemma.

Lemma 7 *Given any function $f \in \mathscr{H}$ and any $t \geq c_n$,*

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(L_t(f) \geq \varepsilon\right) \leq 16\exp\left(-(t-b_n)C(m,\tau)\varepsilon^2\right).$$

Proof of Lemma 7.
Note that

$$\begin{aligned}
L_t(f) &= \mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t] \\
&= \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\left(\mathbb{E}_{t-b_n}[\ell(f,X_t,\xi_i)] - \mathbb{E}_{t-b_n}[\ell(f,X_t,X_i)]\right) \\
&= \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\mathbb{E}_{\xi\sim\pi}\left[\mathbb{E}_{X_t\sim P^{b_n}(X_{t-b_n},\cdot)}\{\ell(f,X_t,\xi)\}\right] - \mathbb{E}_{X_t\sim P^{b_n}(X_{t-b_n},\cdot)}\{\ell(f,X_t,X_i)\}.
\end{aligned}$$

Hence, denoting $m(f,X_{t-b_n},x) = \mathbb{E}_{X_t\sim P^{b_n}(X_{t-b_n},\cdot)}\{\ell(f,X_t,x)\}$, we get

$$L_t(f) \leq \frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\left\{\mathbb{E}_{\xi\sim\pi}\left[m(f,X_{t-b_n},\xi)\right] - m(f,X_{t-b_n},X_i)\right\}.$$

By the reversibility of the chain $(X_i)_{i\geq 1}$, we know that the sequence $(X_{t-b_n},X_{t-b_n-1},\ldots,X_1)$ conditionally on X_{t-b_n} is a Markov chain with invariant distribution π. Applying Proposition 4 we get that

$$\begin{aligned}
&\mathbb{P}\left(L_t(f) \geq \varepsilon \mid X_{t-b_n}\right) \\
\leq\ &\mathbb{P}\left(\frac{1}{t-b_n}\sum_{i=1}^{t-b_n}\left\{\mathbb{E}_{\xi\sim\pi}\left[m(f,X_{t-b_n},\xi_i)\right] - m(f,X_{t-b_n},X_i)\right\} \geq \varepsilon \mid X_{t-b_n}\right) \\
\leq\ &16\exp\left(-(t-b_n)C(m,\tau)\varepsilon^2\right),
\end{aligned}$$

for some constant $C(m,\tau) > 0$ depending only on m and τ. Then we deduce that

$$\begin{aligned}
\mathbb{P}\left(L_t(f) \geq \varepsilon\right) &= \mathbb{E}\left[\mathbb{E}\left\{\mathbb{1}_{L_t(f)\geq\varepsilon} \mid X_{t-b_n}\right\}\right] \\
&= \mathbb{E}\left[\mathbb{P}\left\{L_t(f) \geq \varepsilon \mid X_{t-b_n}\right\}\right] \\
&\leq 16\exp\left(-(t-b_n)C(m,\tau)\varepsilon^2\right),
\end{aligned}$$

which concludes the proof of Lemma 7.
■

The following two Lemmas are key elements to prove Lemma 10. Their proofs are strictly analogous to the proofs of Lemmas 6, 7 and 8 from [63].

Lemma 8 *(See [63, Lemma 6]) For any two functions $h_1, h_2 \in \mathscr{H}$, the following equation holds*

$$|L_t(h_1) - L_t(h_2)| \leq 2\mathrm{Lip}(\varphi)\|h_1 - h_2\|_\infty.$$

Lemma 9 *Let $\mathscr{H} = S_1 \cup \cdots \cup S_l$ and $\varepsilon > 0$. Then*

$$\mathbb{P}\left(\sup_{h\in\mathscr{H}} L_t(h) \geq \varepsilon\right) \leq \sum_{j=1}^{l} \mathbb{P}\left(\sup_{h\in S_j} L_t(h) \geq \varepsilon\right).$$

Lemma 10 *(See [63, Lemma 8]) For any $c_n \leq t \leq n$, it holds*

$$\mathbb{P}\left(\sup_{h\in\mathscr{H}} L_t(h) \geq \varepsilon\right) \leq 16\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{4\mathrm{Lip}(\varphi)}\right)\exp\left(-\frac{(t-b_n)C(m,\tau)\varepsilon^2}{4}\right).$$

Combining Lemma 10 and (28), we have

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1} L_t(h_{t-b_n}) \geq \varepsilon\right) \leq 16\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{4\mathrm{Lip}(\varphi)}\right) n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{4}\right).$$

We deduce that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathscr{R}(h_{t-b_n}) - \mathbb{E}_{t-b_n}[M_t]\right] \geq \varepsilon/2\right)\\
\leq\ &\mathbb{P}\left(L\rho^{b_n} + \frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\left[\mathbb{E}_{t-b_n}[\widetilde{M}_t] - \mathbb{E}_{t-b_n}[M_t]\right] \geq \varepsilon/2\right)\\
\leq\ &16\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8\mathrm{Lip}(\varphi)}\right) n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\left(\varepsilon/2 - L\rho^{b_n}\right)^2}{4}\right).
\end{aligned}$$

Step 3: Conclusion of the proof. From the previous inequality and (27), we get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n}) - \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} M_t \geq \varepsilon\right)\\
\leq\ &b_n\exp\left(-\frac{(1-c)n\varepsilon^2}{8b_n^2}\right) + 16\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8\mathrm{Lip}(\varphi)}\right) n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\left(\varepsilon/2 - L\rho^{b_n}\right)^2}{4}\right).
\end{aligned}$$

Note that $(c_n - b_n)\varepsilon\rho^{b_n} \underset{n\to\infty}{=} o\left(n\varepsilon n^{q\log(\rho)}\right) \underset{n\to\infty}{=} o\left(n^{1+\xi+q\log(\rho)}\right)$ because by assumption $\varepsilon \underset{n\to\infty}{=} o\left(n^{\xi}\right)$. However, by choice of q we have

$$1 + \xi + q\log(\rho) = 1 + \xi + \frac{1+\xi}{\log(1/\rho)}\log(\rho) = 0,$$

and we finally get that $(c_n - b_n)\varepsilon\rho^{b_n} \underset{n\to\infty}{=} o(1)$. We deduce that for n large enough it holds

$$\exp\left(-\frac{(c_n-b_n)C(m,\tau)\left(\varepsilon/2 - L\rho^{b_n}\right)^2}{4}\right) \leq 2\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16}\right).$$

Then, noticing that

$$\exp\left(-\frac{(1-c)n\varepsilon^2}{8b_n^2}\right) \underset{n\to\infty}{=} \mathscr{O}\left(\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16b_n^2}\right)\right),$$

we finally get for n large enough

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\mathscr{R}(h_{t-b_n}) - \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} M_t \geq \varepsilon\right)\\
\leq\ &\left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8\mathrm{Lip}(\varphi)}\right) + 1\right] b_n\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16b_n^2}\right).
\end{aligned}$$

7.2 Proof of Theorem 5

Let us recall that for any $1 \le t \le n-2$, $\widehat{\mathscr{R}}(h_{t-b_n}, t+1) = \binom{n-t}{2}^{-1} \sum_{k>i, i\ge t+1}^{n} \ell(h_{t-b_n}, X_i, X_k)$. We define

$$\ell(h,x) := \mathbb{E}_\pi[\ell(h,X,x)] - \mathscr{R}(h), \text{ and } \widetilde{\ell}(h,x,y) = \ell(h,x,y) - \ell(h,x) - \ell(h,y) - \mathscr{R}(h).$$

Then for any $t \in \{b_n+1, \dots, n-2\}$ we have the following decomposition

$$\widehat{\mathscr{R}}(h_{t-b_n}, t+1) - \mathscr{R}(h_{t-b_n}) = \binom{n-t}{2}^{-1} \sum_{k>i, i\ge t+1}^{n} \widetilde{\ell}(h_{t-b_n}, X_i, X_k) + \frac{2}{n-t} \sum_{i=t+1}^{n} \ell(h_{t-b_n}, X_i). \tag{29}$$

One can check that for any $x \in E$, $\mathbb{E}_\pi\left[\widetilde{\ell}(h,X,x)\right] = \mathbb{E}_\pi\left[\widetilde{\ell}(h,x,X)\right] = 0$. Moreover, for any hypothesis $h \in \mathscr{H}$, $\|\widetilde{\ell}(h,\cdot,\cdot)\|_\infty \le 4$ (because the loss function ℓ takes its value in $[0,1]$). Hence, for any fixed hypothesis $h \in \mathscr{H}$, the kernel $\widetilde{\ell}(h,\cdot,\cdot)$ satisfies Assumption 3. Applying Theorem 2, we know that there exist constants $\beta, \kappa > 0$ such that for any $t \in \{b_n+1, \dots, n-2\}$ and for any $\gamma \in (0,1)$, it holds with probability at least $1-\gamma$,

$$\left| \binom{n-t}{2}^{-1} \sum_{k>i, i\ge t+1}^{n} \widetilde{\ell}(h_{t-b_n}, X_i, X_k) \right| \le \kappa \frac{\log(n-t-1)}{n-t-1} \log((\beta \vee e^1)\log(n-t+1)/\gamma)^2.$$

Note that we used that for $u = \log\left((\beta \vee e^1)\log(n-t+1)/\gamma\right) \ge 1$ it holds

$$\log n \left\{ \frac{u}{n} \vee \left\lceil \frac{u}{n} \right\rceil^2 \right\} \le \frac{\log n}{n} u^2.$$

Using Proposition 4, we also have that for any $t \in \{b_n+1, \dots, n-2\}$ and any $\varepsilon > 0$,

$$\mathbb{P}\left(\left| \frac{2}{n-t} \sum_{i=t+1}^{n} \ell(h_{t-b_n}, X_i) \right| > \varepsilon \right) \le 32 \exp\left(-C(m,\tau)(n-t)\varepsilon^2\right),$$

where $C(m,\tau) = (Km^2\tau^2)^{-1} > 0$ for some universal constant K (one can check from the proof of Proposition 4 that $K = 7 \times 10^3$ fits). We get that for any $t \in \{b_n+1, \dots, n-2\}$ and any $\gamma \in (0,1)$, it holds with probability at least $1-\gamma$,

$$\left| \frac{2}{n-t} \sum_{i=t+1}^{n} \ell(h_{t-b_n}, X_i) \right| \le \frac{\log(32/\gamma)^{1/2} C(m,\tau)^{-1/2}}{\sqrt{n-t}}.$$

We deduce that for any $t \in \{b_n+1, \dots, n-2\}$ and any fixed $\gamma \in (0,1)$, it holds with probability at least $1-\gamma$,

$$\left| \widehat{\mathscr{R}}(h_{t-b_n}, t+1) - \mathscr{R}(h_{t-b_n}) \right| \le C(m,\tau)^{-1/2} \sqrt{\frac{\log(64/\gamma)}{n-t}},$$

i.e.

$$\mathbb{P}\left(\left| \widehat{\mathscr{R}}(h_{t-b_n}, t+1) - \mathscr{R}(h_{t-b_n}) \right| \ge c_\gamma (n-t) \right) \le \frac{\gamma}{(n-c_n)(n-c_n+1)}. \tag{30}$$

Based on the selection procedure of the hypothesis $\widehat{h}$ defined in (6), the concentration result (30) allows us to show that $\mathscr{R}(\widehat{h})$ is close to $\min_{c_n \le t \le n-1} \mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ with high probability. This is stated by Lemma 11 which is proved in Section C.6.

Lemma 11 *Let $h_0, \dots, h_{n-1}$ be the set of hypotheses generated by an arbitrary online algorithm $\mathscr{A}$ working with a pairwise loss ℓ which satisfies the conditions given in Theorem 4. Then for any $\gamma \in (0,1)$, we have*

$$\mathbb{P}\left(\mathscr{R}(\widehat{h}) > \min_{c_n \le t < n-1} (\mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)) \right) \le \gamma.$$

To conclude the proof, we need to show that $\min_{c_n \le t \le n-1} \mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ is close to $\mathscr{M}^n$.
First we remark that

$$\begin{aligned}
& \min_{c_n \le t \le n-1} \mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t) \\
= & \min_{c_n \le t \le n-1} \min_{t \le i \le n-1} \mathscr{R}(h_{i-b_n}) + 2c_\gamma(n-i) \\
\le & \min_{c_n \le t \le n-1} \frac{1}{n-t} \sum_{i=t}^{n-1} \left(\mathscr{R}(h_{i-b_n}) + 2c_\gamma(n-i) \right) \\
\le & \min_{c_n \le t \le n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) + \frac{2}{n-t} \sum_{i=t}^{n-1} \sqrt{\frac{C(m,\tau)^{-1}}{n-i} \log \frac{64(n-c_n)(n-c_n+1)}{\gamma}} \right) \\
\le & \min_{c_n \le t \le n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) + \frac{2}{n-t} \sum_{i=t}^{n-1} \sqrt{\frac{2C(m,\tau)^{-1}}{n-i} \log \frac{64(n-c_n+1)}{\gamma}} \right) \\
\le & \min_{c_n \le t \le n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t} \log \frac{64(n-c_n+1)}{\gamma}} \right),
\end{aligned}$$

where the last inequality holds because $\sum_{i=1}^{n-t} \sqrt{1/i} \le 2\sqrt{n-t}$. Indeed, $x \mapsto 1/\sqrt{x}$ is a decreasing and continuous function and a classical serie/integral approach leads to

$$\sum_{i=1}^{n-t} \sqrt{1/i} \le 1 + \int_1^{n-t} \frac{1}{\sqrt{x}} dx = 1 + \left[2\sqrt{x} \right]_1^{n-t} \le 2\sqrt{n-t}.$$

We define $\mathscr{M}_t^n := \frac{1}{n-t} \sum_{m=t}^{n-1} M_m$. From Theorem 4, one can see that for each $t = c_n, \dots, n-1$,

$$\mathbb{P}\left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) \ge \mathscr{M}_t^n + \varepsilon \right) \le \left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8\mathrm{Lip}(\varphi)} \right) + 1 \right] b_n \exp\left(-\frac{(t-b_n)C(m,\tau)\varepsilon^2}{16 b_n^2} \right).$$

Let us set

$$K_t = \mathscr{M}_t^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t} \log \frac{64(n-c_n+1)}{\gamma}} + \varepsilon.$$

Using the fact that if $\min(a_1, a_2) \le \min(b_1, b_2)$ then either $a_1 \le b_1$ or $a_2 \le b_2$, we can write

$$\begin{aligned}
& \mathbb{P}\left(\min_{c_n \le t \le n-1} \mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t) \ge \min_{c_n \le t \le n-1} K_t \right) \\
\le & \mathbb{P}\left(\min_{c_n \le t \le n-1} \left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t} \log \frac{64(n-c_n+1)}{\gamma}} \right) \ge \min_{c_n \le t \le n-1} K_t \right) \\
\le & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t} \log \frac{64(n-c_n+1)}{\gamma}} \ge K_t \right) \\
= & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\frac{1}{n-t} \sum_{i=t}^{n-1} \mathscr{R}(h_{i-b_n}) \ge \mathscr{M}_t^n + \varepsilon \right) \\
\le & (n-c_n)\left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8Lip(\varphi)} \right) + 1 \right] b_n \exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16 b_n^2} \right) \\
\le & \left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8Lip(\varphi)} \right) + 1 \right] \exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16 b_n^2} + 2\log n \right).
\end{aligned}$$

Using Lemma 11, we get

$$
\begin{aligned}
&\mathbb{P}\left(\mathscr{R}(\widehat{h}) \geq \min_{c_n \leq t \leq n-1} \mathscr{M}_t^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-t}\log\frac{64(n-c_n+1)}{\gamma}} + \varepsilon\right) \\
\leq\ & \gamma + \left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8Lip(\varphi)}\right) + 1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16b_n^2} + 2\log n\right),
\end{aligned}
$$

which gives in particular

$$
\begin{aligned}
&\mathbb{P}\left(\mathscr{R}(\widehat{h}) \geq \mathscr{M}^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} + \varepsilon\right) \\
\leq\ & \gamma + \left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{8Lip(\varphi)}\right) + 1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{16b_n^2} + 2\log n\right).
\end{aligned}
$$

We substitute ε with $\varepsilon/2$ and we choose γ such that $4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} = \varepsilon/2$ with n large enough to ensure that $\gamma < 1$. We have for any $c > 0$,

$$
\begin{aligned}
&\mathbb{P}\left(\mathscr{R}(\widehat{h}) \geq \mathscr{M}^n + 4\sqrt{\frac{2C(m,\tau)^{-1}}{n-c_n}\log\frac{64(n-c_n+1)}{\gamma}} + \frac{\varepsilon}{2}\right) \\
\leq\ & 64(n-c_n+1)\exp\left(-\frac{(n-c_n)C(m,\tau)\varepsilon^2}{128}\right) + \left[32\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{16Lip(\varphi)}\right) + 1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{(16b_n)^2} + 2\log n\right) \\
\leq\ & 32\left[\mathscr{N}\left(\mathscr{H}, \frac{\varepsilon}{16Lip(\varphi)}\right) + 1\right]\exp\left(-\frac{(c_n-b_n)C(m,\tau)\varepsilon^2}{(16b_n)^2} + 2\log n\right),
\end{aligned}
$$

where these inequalities hold for n large enough.

References

[1] R. Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13, 10 2007.

[2] R. Adamczak and W. Bednorz. Some remarks on MCMC estimation of spectra of integral operators. *Bernoulli*, 21(4):2073–2092, Nov 2015.

[3] P. Ango Nze. Critères d'ergodicité géométrique ou arithmétique de modèles linéaires pertubés à représentation Markovienne. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 326(3):371 – 376, 1998.

[4] M. A. Arcones and E. Gine. Limit theorems for U-processes. *Ann. Probab.*, 21(3):1494–1542, 07 1993.

[5] J. Bai. Testing parametric conditional distributions of dynamic models. *The Review of Economics and Statistics*, 85(3):531–549, 2003.

[6] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. *Metrika*, 35(1):339–348, 1988.

[7] E. Behrends. *Introduction to Markov chains*, volume 228. Springer, 2000.

[8] J. Beirlant, L. Györfi, and G. Lugosi. On the asymptotic normality of the l1- and l2-errors in histogram density estimation. *Canadian Journal of Statistics*, 22:309 – 318, 12 2008.

[9] I. Borisov and N. Volodko. A note on exponential inequalities for the distribution tails of canonical von Mises statistics of dependent observations. *Statistics & Probability Letters*, 96:287–291, Jan 2015.

[10] T. R. Boucher. A Hoeffding inequality for Markov chains using a generalized inverse. *Statistics & probability letters*, 79(8):1105–1107, 2009.

[11] C. Butucea et al. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *The Annals of Statistics*, 35(5):1907–1930, 2007.

[12] A. Christmann and I. Steinwart. Support Vector Machines. *Support Vector Machines: Information Science and Statistics.*, 01 2008.

[13] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. JMLR: Workshop and Conference Proceedings, 2016.

[14] S. Clémençon, G. Ciolek, and P. Bertail. Concentration inequalities for regenerative and Harris recurrent Markov chains with applications to statistical learning. In *Séminaire généraliste de l'équipe de Probabilités et Statistiques*, Nancy, France, May 2017.

[15] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Empirical Minimization of U-statistics. *Ann. Statist.*, 36(2):844–874, 04 2008.

[16] Y. De Castro and Q. Duchemin. Markov Random Geometric Graph (MRGG): A Growth Model for Temporal Dynamic Networks. working paper or preprint, June 2020.

[17] Y. De Castro, C. Lacour, and T. M. Pham Ngoc. Adaptive estimation of nonparametric geometric graphs. *Mathematical Statistics and Learning*, 2020.

[18] V. de la Pena and E. Giné. Decoupling, from dependence to independence, randomly stopped processes, u-statistics and processes, martingales and beyond. *Journal of the American Statistical Association*, 95, 09 2000.

[19] P. Del Moral and A. Guionnet. Central limit theorem for nonlinear filtering and interacting particle systems. *Ann. Appl. Probab.*, 9(2):275–297, 05 1999.

[20] R. DeVore and G. Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1993.

[21] P. Doukhan and M. Ghindes. Estimations dans le processus: "$x_{n+1} = f(x_n) + \varepsilon_n$". *C.R. Acad. Sci. Paris Sér. A*, 291:61–64, 01 1980.

[22] J. Fan, B. Jiang, and Q. Sun. Hoeffding's lemma for Markov chains and its applications to statistical learning. *The Journal of Machine Learning Research*, 2018.

[23] Y. Fan. Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. *Journal of Multivariate Analysis*, 62(1):36 – 63, 1997.

[24] Y. Fan and A. Ullah. On goodness-of-fit tests for weakly dependent processes using kernel method. *Journal of Nonparametric Statistics*, 11(1-3):337–360, 1999.

[25] T. Fernández and A. Gretton. A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2966–2975, 2019.

[26] D. Ferré, L. Hervé, and J. Ledoux. Limit theorems for stationary Markov processes with L2-spectral gap. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(2):396–423, May 2012.

[27] G. Fort, E. Moulines, P. Priouret, and P. Vandekerkhove. A simple variance inequality for U-statistics of a Markov chain with applications. *Statistics and Probability Letters*, 82(6):1193–1201, 2012.

[28] M. Fromont and B. Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 04 2006.

[29] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

[30] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. *High Dimensional Probability II*, page 13–38, 2000.

[31] P. W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics & probability letters*, 56(2):143–146, 2002.

[32] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 1292–1301. JMLR.org, 2017.

[33] A. H. Guide. *Infinite dimensional analysis*. Springer, 2006.

[34] F. Han. An exponential inequality for U-statistics under mixing conditions. *Journal of Theoretical Probability*, 31(1):556–578, Nov 2016.

[35] C. Houdré and P. Reynaud-Bouret. Exponential inequalities for U-statistics of order two with constants. *Stochastic Inequalities and Applications. Progress in Probability*, 56, 01 2002.

[36] Y. I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.*, 2(2):85–114, 1993.

[37] B. Jiang, Q. Sun, and J. Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

[38] E. Joly and G. Lugosi. Robust estimation of U-statistics. *Stochastic Processes and their Applications*, In Memoriam: Evarist Giné:3760–3773, 2016.

[39] G. L. Jones. On the Markov chain central limit theorem. *Probab. Surveys*, 1:299–320, 2004.

[40] V. Koltchinskii and E. Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6, 02 2000.

[41] S. Kwapien and W. Woyczynski. *Random Series and Stochastic Integrals: Single and Multiple: Single and Multiple*. Probability and Its Applications. Birkhäuser Boston, 2002.

[42] M. Lerasle, N. M. Magalhães, and P. Reynaud-Bouret. Optimal kernel selection for density estimation. *Progress in Probability*, page 425–460, 2016.

[43] F. Li and G. Tkacz. A Consistent Bootstrap Test for Conditional Density Functions with Time-Dependent Data. Staff working papers, Bank of Canada, 2001.

[44] F. Lindsten, R. Douc, and E. Moulines. Uniform ergodicity of the particle Gibbs sampler. *Scandinavian Journal of Statistics*, 42(3):775–797, Feb 2015.

[45] Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284, 2016.

[46] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[47] K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24(1):101–121, 02 1996.

[48] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*, volume 92. 01 1993.

[49] C. Müller. *Analysis of spherical symmetries in Euclidean spaces*, volume 129. Springer Science & Business Media, 2012.

[50] D. Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(0), 2015.

[51] S. Rao et al. A Hoeffding inequality for Markov chains. *Electronic Communications in Probability*, 24, 2019.

[52] G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1(0):20–71, 2004.

[53] R. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1970.

[54] R. Rudzkis and A. Bakshaev. Goodness of fit tests based on kernel density estimators. *Informatica*, 24(3):447–460, 2013.

[55] P.-M. Samson et al. Concentration of measure inequalities for Markov chains and Phi-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

[56] Y. Shen, F. Han, and D. Witten. Exponential inequalities for dependent V-statistics via random Fourier features. *Electronic Journal of Probability*, 25(0), 2020.

[57] S. Smale and D.-X. Zhou. Online learning with Markov sampling. *Analysis and Applications*, 7(01):87–113, 2009.

[58] I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

[59] M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular biology and evolution*, 18(6):1001–1013, 2001.

[60] R. S. Sutton and A. G. Barto. Reinforcement learning: An introduction. *Cambridge, MA: MIT Press*, 2011.

[61] A. Van Der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer New York, 2013.

[62] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[63] Y. Wang, R. Khardon, D. Pechyony, and R. Jones. Online learning with pairwise loss functions. *CoRR*, 2013.

[64] J. Xu, Y. Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu. The generalization ability of online SVM classification based on Markov sampling. *IEEE transactions on neural networks and learning systems*, 26(3):628–639, 2014.

[65] B. Zou, H. Zhang, and Z. Xu. Learning from uniformly ergodic Markov chains. *Journal of Complexity*, 25(2):188–200, 2009.

A Definitions and properties for Markov chains

We recall some definitions and well-known results on Markov chains that will be useful for the next sections.

A.1 Ergodic and reversible Markov chains

Definition 5 *[52, section 3.2] (φ-irreducible Markov chains)*
The Markov chain $(X_i)_{i\geq 1}$ is said φ-irreducible if there exists a non-zero σ-finite measure φ on E such that for all $A\in\Sigma$ with $\varphi(A)>0$, and for all $x\in E$, there exists a positive integer $n=n(x,A)$ such that $P^n(x,A)>0$ (where $P^n(x,\cdot)$ denotes the distribution of X_{n+1} conditioned on $X_1=x$).

Definition 6 *[52, section 3.2] (Aperiodic Markov chains)*
The Markov chain $(X_i)_{i\geq 1}$ with invariant distribution π is aperiodic if there do not exist $m\geq 2$ and disjoint subsets $A_1,\dots,A_m\subset E$ with $P(x,A_{i+1})=1$ for all $x\in A_i$ $(1\leq i\leq m-1)$, and $P(x,A_1)=1$ for all $x\in A_m$, such that $\pi(A_1)>0$ (and hence $\pi(A_i)>0$ for all i).

Definition 7 *[52, section 3.4] (Geometric ergodicity)*
The Markov chain $(X_i)_{i\geq 1}$ is said geometrically ergodic if there exists an invariant distribution π, $\rho\in(0,1)$ and $C:E\to[1,\infty)$ such that

$$\|P^n(x,\cdot)-\pi\|_{TV}\leq C(x)\rho^n,\qquad \forall n\geq 0,\ \pi\text{—a.e } x\in E,$$

where $\|\mu\|_{TV}:=\sup_{A\in\Sigma}|\mu(A)|$.

Definition 8 *[52, section 3.3] (Uniform ergodicity)*
The Markov chain $(X_i)_{i\geq 1}$ is said uniformly ergodic if there exists an invariant distribution π and constants $0<\rho<1$ and $L>0$ such that

$$\|P^n(x,\cdot)-\pi\|_{TV}\leq L\rho^n,\qquad \forall n\geq 0,\ \pi\text{—a.e } x\in E,$$

where $\|\mu\|_{TV}:=\sup_{A\in\Sigma}|\mu(A)|$.

Remark A Markov chain geometrically or uniformly ergodic admits a unique invariant distribution.

A.2 Spectral gap

This section is largely inspired from [22]. Let us consider that the Markov chain $(X_i)_{i\geq 1}$ admits a unique invariant distribution π on E.
For any real-valued, Σ-measurable function $h:E\to\mathbb{R}$, we define $\pi(h):=\int h(x)d\pi(x)$. The set

$$\mathscr{L}_2(E,\Sigma,\pi):=\{h:\pi(h^2)<\infty\}$$

is a Hilbert space endowed with the inner product

$$\langle h_1,h_2\rangle_\pi=\int h_1(x)h_2(x)d\pi(x),\ \forall h_1,h_2\in\mathscr{L}^2(E,\Sigma,\pi).$$

The map

$$\|\cdot\|_\pi : h\in\mathscr{L}_2(E,\Sigma,\pi)\mapsto\|h\|_\pi=\sqrt{\langle h,h\rangle_\pi},$$

is a norm on $\mathscr{L}_2(E,\Sigma,\pi)$. $\|\cdot\|_\pi$ naturally allows to define the norm of a linear operator T on $\mathscr{L}_2(E,\Sigma,\pi)$ as

$$N_\pi(T)=\sup\{\|Th\|_\pi:\|h\|_\pi=1\}.$$

To each transition probability kernel $P(x,B)$ with $x\in E$ and $B\in\Sigma$ invariant with respect to π, we can associate a bounded linear operator $h\mapsto\int h(y)P(\cdot,dy)$ on $\mathscr{L}_2(E,\Sigma,\pi)$. Denoting this operator P, we get

$$Ph(x)=\int h(y)P(x,dy),\ \forall x\in E,\ \forall h\in\mathscr{L}_2(E,\Sigma,\pi).$$

Let $\mathscr{L}_2^0(\pi):=\{h\in\mathscr{L}_2(E,\Sigma,\pi)\ :\ \pi(h)=0\}$. We define the absolute spectral gap of a Markov operator.

Definition 9 *(Spectral gap) A Markov operator P reversible admits a spectral gap $1-\lambda$ if*

$$\lambda := \sup\left\{ \frac{\|Ph\|_\pi}{\|h\|_\pi} \ : \ h \in \mathscr{L}_2^0(\pi),\, h \neq 0 \right\} < 1.$$

Remark Uniform ergocity ensures that the chain admits a spectral gap (see [26, Section 2.3]).

A.3 The splitting method

We describe the construction of the split chain that we use in our proof. Let us consider some Markov chain $(X_n)_n$ on the probability space $(E, \Sigma, \mathbb{P})$ with transition kernel P. We assume that there exists a small set $C \in \Sigma$, there exist a positive integer m, a constant $\delta_m > 0$, and a probability measure μ on E with the following minorisation condition holds

$$\forall x \in C, \quad \forall A \in \Sigma, \quad P^m(x, A) \geq \delta_m \mu(A).$$

We give only the construction of the split chain in the case where $m = 1$ and we refer to [48] for further details on the construction of the split chain. Each point x in E is splitted in $x_0 = (x, 0) \in E_0 = E \times \{0\}$ and $x_1 = (x, 1) \in E_1 = E \times \{1\}$. Each set B in Σ is splitted in $B_0 = B \times \{0\}$ and $B_1 = B \times \{1\}$. Thus, we have defined a new probability space (E^*, Σ^*) where $E^* = E_0 \cup E_1$ and $\Sigma^* = \sigma(B_0, B_1 : B \in \Sigma)$. A measure η on Σ splits into two measures on E_0 and E_1 by defining η^* on (E^*, Σ^*) through

$$\begin{aligned} \eta^*(B_0) &= \eta(B \cap C)(1 - \delta_m) + \eta(B \cap C^c) \\ \eta^*(B_1) &= \eta(B \cap C)\delta_m. \end{aligned}$$

Now we define a new transition probability $P^*(\cdot, \cdot)$ on (E^*, Σ^*) with

$$P^*((x, e), \cdot) = \begin{cases} P(x, \cdot)^* & \text{if } x \notin C \text{ and } e = 0 \\ (1 - \delta_m)^{-1}\left(P(x, \cdot)^* - \delta_m \mu^*\right) & \text{if } x \in C \text{ and } e = 0 \\ \mu^* & \text{if } e = 1 \end{cases}.$$

The split chain associated with $(X_n)_n$ is then defined as the Markov chain $((\widetilde{X}_n, R_n))_n$ on E^* with transition kernel P^*. After the previous rigorous construction of the split chain, we can give the following interpretation of the method. Each time the chain reaches C, there is a possibility for the chain to regenerate. Each time the chain is at $x \in C$, a coin is tossed with probability of success δ_m. If the toss is successful, then the chain is moved according to the probability distribution μ, otherwise, according to $(1 - \delta_m)^{-1}(P(x, \cdot) - \delta_m \mu(\cdot))$. Overall, the dynamic of the chain is not affected by this coin toss, but at each time the toss is successful, the chains regenerates with regeneration distribution μ independent from x. In particular if Assumption 1 holds, the whole space is small and if in addition $m = 1$, the chain regenerates with probability δ_m at each time step.

B Connections with the literature

Let us establish a clear connection between Theorem 2 and the exponential inequality from [56]. We consider an integer $n \in \mathbb{N}^*$ and a Markov chain $(X_i)_{i \geq 1}$ geometrically ergodic on $\mathbb{R}^d$ with invariant measure π. As mentioned in [39, p.6], $(X_i)_{i \geq 1}$ is in particular geometrically α-mixing (see [56, Section 2]) with coefficient

$$\alpha(i) \leq \gamma_1 \exp(-\gamma_2 i), \quad \text{for all } i \geq 1,$$

where γ_1, γ_2 are two positive absolute constants. We consider a kernel $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ π-canonical, continuous, integrable and satisfying for some $q \geq 1$, $\int_{\mathbb{R}^{2d}} |\mathscr{F}h(u)| \|u\|_2^q du < \infty$. Then Eq.(2.4) from [56] states that there exists a constant $C > 0$ such that for any $u > 0$, it holds with probability at least $1 - 6e^{-u}$

$$\frac{2}{n(n-1)} U_{\text{stat}}(n) \leq 4C \|\mathscr{F}h\|_{L^1} \left\{ A_n^{1/2} \frac{u}{n} + C \log^4(n) \left[\frac{u}{n} \right]^2 \right\},$$

where $A_n^{1/2} = 4\left(\frac{64\gamma_1^{1/3}}{1-\exp(-\gamma_2/3)} + \frac{\log^4(n)}{n}\right)$ and $U_{\text{stat}}(n) = \sum_{1\le i<j\le n}\big(h(X_i,X_j) - \mathbb{E}_\pi[h(X,\cdot)]\big)$.
Hence, when we consider a unique kernel function with strong regularity conditions, assumptions on the Markov chain $(X_i)_{i\ge1}$ can be weakened and [56, Theorem 2.1] provides a result that is close to the concentration inequality of Theorem 2. In a framework where both Theorem 2 and [56, Theorem 2.1] hold, noticing that for any $u \ge 1$,

$$\log(n)\frac{u}{n} + \log n\left[\frac{u}{n}\right]^2 \le \mathbf{log(n)}\frac{u}{n} + \log^4(n)\left[\frac{u}{n}\right]^2,$$

we deduce that our bound is asymptotically at least as good as the one from [56] (up to a possible log factor marked in bold).

C Additional proofs

C.1 Hoeffding inequality for uniformly ergodic Markov chains

A large number of different Hoeffding inequalities for Markov chains can be found in the literature such as in [22],[10],[51] or [31]. We need such concentration result in our applications and our goal is to work with the milder assumptions possible. More precisely, we want to consider a uniformly ergodic Markov chain without any assumption on the initial distribution of the chain. As stated in [10, Lemma 1], for a ψ-irreducible and aperiodic Markov chain, a Hoeffding inequality with exponents independent of the initial state of the chain exists if and only if the chain is uniformly ergodic. Some Hoeffding inequalities for uniformly ergodic Markov chains without condition on the initial distribution already exist (see [31] or [10]), but they require n to be large enough to hold and can involve quantities related to the chain that we did not use so far (such that the Drazin inverse of $I-P$). This is the reason why we propose here to prove briefly a different Hoeffding inequality for uniformly ergodic Markov chains that holds for any sample size n, any initial distribution of the chain and that only uses the notations from Section 2.

Proposition 4 *Let $(X_i)_{i\ge1}$ be a Markov chain on E uniformly ergodic (namely satisfying Assumption 1) with invariant distribution π and let us consider some function $f : E \to \mathbb{R}$ such that $\mathbb{E}_{X\sim\pi}[f(X)] = 0$ and $\|f\|_\infty \le A$. Then it holds for any $t \ge 0$*

$$\mathbb{P}\left(\left|\sum_{i=1}^n f(X_i)\right| \ge t\right) \le 16\exp\left(-\frac{1}{K}\min\left(\frac{t^2(\mathbb{E}T_2)}{nA^2m^2\tau^2}, \frac{t}{Am\tau}\right)\right),$$

for some universal constant K.
In particular, since by definition of τ and of the Orlicz norm we have $\mathbb{E}[T_2] \le \left(\mathbb{E}[e^{T_2/\tau}]-1\right)\tau \le \tau$, it holds for any $t \ge 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^n f(X_i)\right| \ge t\right) \le 16\exp\left(-\frac{1}{K(m,\tau)}\frac{t^2}{nA^2}\right),$$

where $K(m,\tau) = 2Km^2\tau^2$ for some universal constant $K > 0$.

Proof of Proposition 4.
Let us consider $N = \sup\{i \in \mathbb{N} \,:\, mS_{i+1} + m - 1 \ge n\}$. Then,

$$\begin{aligned}\left|\sum_{i=1}^n f(X_i)\right| &= \left|\sum_{l=0}^N Z_l + \sum_{i=m(S_N+1)}^n f(X_i)\right| \\ &\le \left|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\right| + \left|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\right| + \left|\sum_{i=m(S_N+1)}^n f(X_i)\right|. \end{aligned} \tag{31}$$

We have $|\sum_{i=m(S_N+1)}^{n} f(X_i)| \leq AmT_{N+1}$. So using the definition of the Orlicz norm and the fact that the random variables $(T_i)_{i\geq 2}$ are i.i.d., it holds for any $t \geq 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=m(S_N+1)}^{n} f(X_i)\right| \geq t\right) &\leq \mathbb{P}(T_{N+1} \geq \frac{t}{Am}) \\ &\leq \mathbb{P}(\max(T_1, T_2) \geq \frac{t}{Am}) \\ &\leq \mathbb{P}(T_1 \geq \frac{t}{Am}) + \mathbb{P}(T_2 \geq \frac{t}{Am}) \\ &\leq 4\exp\left(-\frac{t}{Am\tau}\right).\end{aligned}$$

In order to control the first two terms in (31), we need to describe the tail behaviour of the random variable N with Lemma 12.

Lemma 12 *(See [1, Lemma 5])*
If $\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1} \leq \tau$, then

$$\mathbb{P}(N > R) \leq 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right),$$

where $R = \lfloor 3n/(\mathbb{E}T_2)\rfloor$.

The random variable Z_{2l} is $\sigma(X_{m(S_{2l}+1)}, \dots, X_{m(S_{2l+1}+1)-1})$-measurable. Hence the random variables $(Z_{2l})_l$ are independent (see Section 2.3). Moreover, one has that for any l, $\mathbb{E}[Z_{2l}] = 0$. This is due to [48, Eq.(17.23) Theorem 17.3.1] together with the assumption that $\mathbb{E}_{X\sim\pi}[f(X)] = 0$. Let us finally notice for any $l \geq 0$, $|Z_{2l}| \leq AmT_{2l+1}$, so $\|Z_{2l}\|_{\psi_1} \leq Am\max(\|T_1\|_{\psi_1}, \|T_2\|_{\psi_1}) \leq Am\tau$. One can similarly get that $(Z_{2l+1})_l$ are independent with $\mathbb{E}[Z_{2l+1}] = 0$ and $\|Z_{2l+1}\|_{\psi_1} \leq Am\tau$ for all $l \in \mathbb{N}$. Using these facts we have for any $t \geq 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\right| + \left|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\right| \geq t\right) \\ &\leq \mathbb{P}\left(\left|\sum_{l=0}^{\lfloor N/2\rfloor} Z_{2l}\right| + \left|\sum_{l=0}^{\lfloor (N-1)/2\rfloor} Z_{2l+1}\right| \geq t\ , N \leq R\right) + 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right) \\ &\leq \mathbb{P}\left(\max_{0\leq s\leq \lfloor R/2\rfloor}\left|\sum_{l=0}^{s} Z_{2l}\right| \geq t/2\right) + \mathbb{P}\left(\max_{0\leq s\leq \lfloor (R-1)/2\rfloor}\left|\sum_{l=0}^{s} Z_{2l+1}\right| \geq t/2\right) + 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right) \\ &\leq 3\mathbb{P}\left(\left|\sum_{l=0}^{\lfloor R/2\rfloor} Z_{2l}\right| \geq t/6\right) + 3\mathbb{P}\left(\left|\sum_{l=0}^{\lfloor (R-1)/2\rfloor} Z_{2l+1}\right| \geq t/6\right) + 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right) \quad \text{(Using Lemma 13)} \\ &\leq 12\exp\left(-\frac{1}{8}\min\left(\frac{t^2}{36RA^2m^2\tau^2}, \frac{t}{6Am\tau}\right)\right) + 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right),\end{aligned}$$

where we used Lemma 6 in the last inequality.

Lemma 13 *(see [41, Proposition 1.1.1]) If $X_1, X_2, \dots$ are independent Banach space valued random variables (not necessarily identically distributed), and if $S_k = \sum_{i=1}^{k} X_i$, then*

$$\mathbb{P}\left(\max_{1\leq j\leq k}\|S_j\| > t\right) \leq 3\max_{1\leq j\leq k}\mathbb{P}\left(\|S_j\| > t/3\right).$$

Gathering the previous results, we obtain that for any $t \geq 0$

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} f(X_i)\right| \geq t\right) \leq 12\exp\left(-\frac{1}{8}\min\left(\frac{t^2(\mathbb{E}T_2)}{36\times 12\times nA^2m^2\tau^2}, \frac{t}{12Am\tau}\right)\right) + 2\exp\left(-\frac{n\mathbb{E}T_2}{8\tau^2}\right) + 4\exp\left(-\frac{t}{2Am\tau}\right).$$

Since the left hand side of the previous inequality is zero for $t \geq nA$, and since $m \geq 1$, we obtain Proposition 4.

■

C.2 Talagrand inequality for Markov chains

In the section, we show that in the proof of Theorem 1, we can use the concentration inequality for the supremum of an empirical process of [55, Theorem 3].
Let us consider the sequence of random variables $W = (W_1, \dots, W_n)$ on the probability space $(\Omega, \mathscr{A}, \mathbb{P})$ taking values in the measurable space $S = E \times F^{n-1}$ where F is the subset of the set $\mathscr{F}(E, \mathbb{R})$ of all measurable functions from (E, Σ) to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ that are bounded by A. Note that

$$\{0_{\mathscr{F}(E,\mathbb{R})}\} \cup \{p_{i,j}(x, \cdot) \,:\, x \in E,\ i, j \in [n]\} \subset F.$$

We define

$$\mathscr{P} := \left\{ D \in \mathscr{P}(F) \,:\, \forall i \in [n-1], \forall j \in \{i+1, \dots, n\},\ f_{i,j}^{-1}(D) \in \Sigma \right\},$$

where $\mathscr{P}(F)$ is the powerset of F and where $\forall i \in [n-1],\ \forall j \in \{i+1, \dots, n\}$,

$$\begin{aligned} f_{i,j} : (E, \Sigma) &\to (F, \mathscr{P}(F)) \\ x &\mapsto p_{i,j}(x, \cdot). \end{aligned}$$

Then we have the following straightforward result.

Lemma 14 *$\mathscr{P}$ is a σ-algebra on F.*

In the following, we endow the space F with the σ-algebra $\mathscr{P}$ and we consider on S the product σ-algebra given by

$$\mathscr{S} := \sigma\left(\left\{ C \times D_2 \times \cdots \times D_n \,:\, C \in \Sigma,\ D_j \in \mathscr{P}\ \forall j \in \{2, \dots, n\} \right\} \right).$$

For all $i \in [n]$, we define W_i by

$$W_i := \left(X_i, \underbrace{0, \dots, 0}_{(i-1) \text{ times}}, p_{i,i+1}(X_i, \cdot), p_{i,i+2}(X_i, \cdot), \dots, p_{i,n}(X_i, \cdot) \right).$$

Hence for all $i \in [n]$, W_i is $\sigma(X_i)$-measurable. Let us consider for any $i \in [n-1]$,

$$\begin{aligned} \Phi_i : (E, \Sigma) &\to (S, \mathscr{S}) \\ x &\mapsto \left(x, \underbrace{0_F, \dots, 0_F}_{(i-1) \text{ times}}, p_{i,i+1}(x, \cdot), \dots, p_{i,n}(x, \cdot) \right). \end{aligned}$$

Then, one can directly see that for all $i \in [n-1]$, $W_i = \Phi_i(X_i)$ and by construction of $\mathscr{P}$ and $\mathscr{S}$, Φ_i is measurable. Indeed, each coordinate of Φ_i is measurable by construction of $\mathscr{P}$ and this ensures that Φ_i is measurable thanks to the following Lemma.

Lemma 15 *(See [33, Lemma 4.49]) Let $(X, \Sigma), (X_1, \Sigma_1)$ and (X_2, Σ_2) be measurable spaces, and let $f_1 : X \to X_1$ and $f_2 : X \to X_2$. Define $f : X \to X_1 \times X_2$ by $f(x) = (f_1(x), f_2(x))$. Then $f : (X, \Sigma) \to (X_1 \times X_2, \Sigma_1 \otimes \Sigma_2)$ is measurable if and only if the two functions $f_1 : (X, \Sigma) \to (X_1, \Sigma_1)$ and $f_2 : (X, \Sigma) \to (X_2, \Sigma_2)$ are both measurable.*

Then it holds for any $i \in \{2, \dots, n-1\}$ and any $G \in \mathscr{S}$,

$$\begin{aligned} &\mathbb{P}(W_i \in G \mid W_{i-1}) \\ &= \mathbb{P}(\Phi_i(X_i) \in G \mid W_{i-1}) \\ &= \mathbb{P}(\Phi_i(X_i) \in G \mid X_{i-1}) \\ &= \mathbb{P}\left(X_i \in \Phi_i^{-1}(G) \mid X_{i-1}\right) \\ &= P\left(X_{i-1}, \Phi_i^{-1}(G)\right) \\ &= \left[(\Phi_i)_{\#} P(X_{i-1}, \cdot)\right](G), \end{aligned} \tag{32}$$

where $(\Phi_i)_{\#}P(X_{i-1},\cdot)$ denotes the pushforward measure of the measure $P(X_{i-1},\cdot)$ by the measurable map Φ_i. We deduce that W_i is non-homogeneous Markov chain. Moreover, (32) proves that the transition kernel of the Markov chain $(W_k)_k$ from state $i-1$ to state i is given by $K^{(i-1,i)}$ where for all $(x,p_2,\dots,p_n)\in S$ and for all $G\in\mathscr{S}$,

$$K^{(i-1,i)}((x,p_2,\dots,p_n),G)=[(\Phi_i)_{\#}P(x,\cdot)](G).$$

One can easily generalize this notation. Let us consider some $i,j\in[n]$ with $i<j$ and let us denote $K^{(i,j)}$ the transition kernel of the Markov chain $(W_k)_k$ from state i to state j. Then for all $x\in E$, for all $p_2,\dots,p_n\in F$ and for all $G\in\mathscr{S}$,

$$K^{(i,j)}((x,p_2,\dots,p_n),G)=\left[(\Phi_j)_{\#}P^{j-i}(x,\cdot)\right](G),$$

We introduce the mixing matrix $\Gamma=(\gamma_{i,j})_{1\le i,j\le n-1}$ where coefficients are defined by

$$\gamma_{i,j}:=\sup_{w_i\in S}\sup_{z_i\in S}\|\mathscr{L}(W_j|W_i=w_i)-\mathscr{L}(W_j|W_i=z_i)\|_{TV}.$$

For any $w\in S=E\times F^{n-1}$, we denote $w^{(1)}$ the first coordinate of the vector w. Hence, $w^{(1)}$ is an element of E. Then

$$\begin{aligned}
\gamma_{i,j}&=\sup_{w_i\in S}\sup_{z_i\in S}\sup_{G\in\mathscr{S}}\left|\left[(\Phi_j)_{\#}P^{j-i}(w_i^{(1)},\cdot)\right](G)-\left[(\Phi_j)_{\#}P^{j-i}(z_i^{(1)},\cdot)\right](G)\right|\\
&=\sup_{w_i\in S}\sup_{z_i\in S}\sup_{G\in\mathscr{S}}\left|P^{j-i}\left(w_i^{(1)},\Phi_j^{-1}(G)\right)-P^{j-i}\left(z_i^{(1)},\Phi_j^{-1}(G)\right)\right|\\
&\le\sup_{w_i\in S}\sup_{z_i\in S}\sup_{C\in\Sigma}\left|P^{j-i}\left(w_i^{(1)},C\right)-P^{j-i}\left(z_i^{(1)},C\right)\right|\\
&=\sup_{x_i\in E}\sup_{x_i'\in E}\sup_{C\in\Sigma}\left|P^{j-i}(x_i,C)-P^{j-i}\left(x_i',C\right)\right|\\
&=\sup_{x_i\in E}\sup_{x_i'\in E}\|P^{j-i}(x_i,\cdot)-\pi(\cdot)+\pi(\cdot)-P^{j-i}(x_i',\cdot)\|_{TV}\\
&\le\sup_{x_i\in E}\|P^{j-i}(x_i,\cdot)-\pi(\cdot)\|_{TV}+\sup_{x_i'\in E}\|P^{j-i}(x_i',\cdot)-\pi(\cdot)\|_{TV}\\
&\le 2L\rho^{j-i},
\end{aligned}$$

where in the first inequality we used that $\Phi_j:(E,\Sigma)\to(S,\mathscr{S})$ is measurable and in the last inequality we used the uniform ergodicity of the Markov chain $(X_i)_{i\ge1}$. We deduce that

$$\|\Gamma\|\le 2L\left\|Id+\sum_{l=1}^{n-1}\rho^l N_l\right\|,$$

where $N_l=\left(n_{i,j}^{(l)}\right)_{1\le i,j\le n-1}$ represents the nilpotent matrix of order l defined by

$$n_{i,j}^{(l)}=\begin{cases}1 & \text{if } j-i=l\\ 0 & \text{otherwise.}\end{cases}$$

Since for each $1\le l\le n-1$, $\|N_l\|\le1$, it follows from the triangular inequality that

$$\|\Gamma\|\le 2L\sum_{l=0}^{n-1}\rho^l\le\frac{2L}{1-\rho}.$$

To conclude the proof and get the concentration result stated in Lemma 4, one only needs to apply [55, Theorem 3] with the class of functions $\mathscr{F}$ and with the Markov chain $(W_k)_k$. Let us recall that $\mathscr{F}$ is defined by $\mathscr{F}=\{f_\xi:\sum_{j=2}^n\mathbb{E}|\xi_j(X_j')|^{k/(k-1)}=1\}$ where for any $\xi=(\xi_2,\dots,\xi_n)\in\prod_{i=2}^n L^{k/(k-1)}(\nu)$,

$$\forall w=(x,p_2,\dots,p_n)\in E\times F^{n-1},\quad f_\xi(w)=\sum_{j=2}^n\int p_j(y)\xi_j(y)d\nu(y).$$

C.3 Bernstein's inequality for non-stationary Markov chains

This subsection is dedicated to the proof of Proposition 5 which is used in the proof of Theorem 1. A Bernstein type concentration inequality was proved in [37] for stationary Markov chains. Following the work of [22], we extend the previous Bernstein inequality to the framework of non-stationary Markov chains with initial distribution that satisfies Assumption 4.(ii). Contrary to Proposition 4, this concentration handles the setting where the sum involves different functions $(f_i)_i$. This is of interest for Theorem 1 when we allow the kernel functions $h_{i,j}$ to depend on both i and j.

Proposition 5 *Suppose that the sequence $(X_i)_{i\geq 1}$ is a Markov chain satisfying Assumptions 1 and 4.(ii) with invariant distribution π and with an absolute spectral gap $1-\lambda > 0$. Let us consider some $n \in \mathbb{N}^*$ and bounded real valued functions $(f_i)_{1\leq i\leq n}$ such that for any $i \in \{1,\dots,n\}$, $\int f_i(x)d\pi(x) = 0$ and $\|f_i\|_\infty \leq c$ for some $c > 0$. Let $\sigma^2 = \sum_{i=1}^n \int f_i^2(x)d\pi(x)/n$. Then for any $\varepsilon \geq 0$ it holds*

$$\mathbb{P}\left(\sum_{i=1}^n f_i(X_i) \geq \varepsilon\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \exp\left(-\frac{\varepsilon^2/(2q)}{A_2 n\sigma^2 + A_1 c\varepsilon}\right),$$

where $A_2 := \frac{1+\lambda}{1-\lambda}$ and $A_1 := \frac{1}{3}\mathbb{1}_{\lambda=0} + \frac{5}{1-\lambda}\mathbb{1}_{\lambda>0}$. q is the constant introduced in Assumption 4.(ii). Stated otherwise, for any $u > 0$ it holds

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_i(X_i) > \frac{2quA_1\|f\|_\infty}{n} + \sqrt{\frac{2quA_2\sigma^2}{n}}\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} e^{-u},$$

C.3.1 Proof of Proposition 5.

Let us recall that we denote indifferently $\mathbb{P}_\chi$ or $\mathbb{P}$ the probability distribution of the Markov chain $(X_i)_{i\geq 1}$ when the distribution of the first state X_1 is χ, whereas $\mathbb{P}_\pi$ refers to the distribution of the Markov chain when the distribution of the first state X_1 is the invariant measure π.
In [37], they proved that for any $0 \leq t < (1-\lambda)/(5c)$, it holds

$$\mathbb{E}_\pi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \leq \exp\left(\frac{n\sigma^2}{c^2}(e^{tc} - tc - 1) + \frac{n\sigma^2\lambda t^2}{1-\lambda-5ct}\right).$$

We deduce that for any $0 \leq t < (1-\lambda)/(5cq)$,

$$\begin{aligned}
\mathbb{E}_\chi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] &\leq \mathbb{E}_\pi\left[\frac{d\chi}{d\pi}(X_1)e^{t\sum_{i=1}^n f_i(X_i)}\right] \\
&\leq \left\{\mathbb{E}_\pi\left[\left|\frac{d\chi}{d\pi}(X_1)\right|^p\right]\right\}^{1/p}\left\{\mathbb{E}_\pi\left[e^{qt\sum_{i=1}^n f_i(X_i)}\right]\right\}^{1/q} \\
&= \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\left\{\mathbb{E}_\pi\left[e^{qt\sum_{i=1}^n f_i(X_i)}\right]\right\}^{1/q} \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\left\{\exp\left(\frac{n\sigma^2}{c^2}(e^{tqc} - tqc - 1) + \frac{n\sigma^2\lambda q^2t^2}{1-\lambda-5cqt}\right)\right\}^{1/q} \\
&= \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\exp\left(\frac{n\sigma^2}{qc^2}(e^{tqc} - tqc - 1) + \frac{n\sigma^2\lambda qt^2}{1-\lambda-5cqt}\right). \qquad (33)
\end{aligned}$$

Let us define

$$g_1(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{n\sigma^2}{qc^2}(e^{tqc} - tqc - 1) & \text{if } t \geq 0 \end{cases}$$

and

$$g_2(t) = \begin{cases} 0 & \text{if } t < 0 \\ \frac{n\sigma^2\lambda q t^2}{1-\lambda-5cqt} & \text{if } 0 \le t < \frac{1-\lambda}{5cq} \\ +\infty & \text{if } t \ge \frac{1-\lambda}{5cq} \end{cases}.$$

In order to lower-bound the convex conjugate of the function $g_1 + g_2$, we will need the convex conjugate of g_1 and g_2 which are provided by Lemma 16. The proof of Lemma 16 can be found in Section C.3.2.

Lemma 16 *g_1 and g_2 are closed proper convex functions with convex conjugates*

$$\forall \varepsilon_1 \in \mathbb{R}, \quad g_1^*(\varepsilon_1) = \begin{cases} \frac{n\sigma^2}{qc^2} h_1\left(\frac{\varepsilon_1 c}{n\sigma^2}\right) & \text{if } \varepsilon_1 \ge 0 \\ +\infty & \text{if } \varepsilon_1 < 0 \end{cases} \tag{34}$$

with $h_1(u) = (1+u)\log(1+u) - u \ge \frac{u^2}{2(1+u/3)}$ for any $u \ge 0$, and

$$\forall \varepsilon_2 \in \mathbb{R}, \quad g_2^*(\varepsilon_2) = \begin{cases} \frac{(1-\lambda)\varepsilon^2}{qn\sigma^2\lambda} h_2\left(\frac{5c\varepsilon_2}{n\sigma^2\lambda}\right) & \text{if } \varepsilon_2 \ge 0 \\ +\infty & \text{if } \varepsilon_2 < 0 \end{cases} \tag{35}$$

with $h_2(u) = \left(\frac{(\sqrt{u+1}-1}{u}\right)^2 \ge \frac{1}{2(u+2)}$.

Since $g_1(t) = O(t^2)$ and $g_2(t) = O(t^2)$ as $t \to 0^+$, $t\varepsilon - g_1(t) - g_2(t) > 0$ for small enough $t > 0$, and $t\varepsilon - g_1(t) - g_2(t) \le 0$ for $t \le 0$. Hence

$$(g_1+g_2)^*(\varepsilon) = \sup_{0 \le t < (1-\lambda)/(5cq)} \varepsilon t - g_1(t) - g_2(t) = \sup_{t \in \mathbb{R}} \varepsilon t - g_1(t) - g_2(t).$$

• If $\lambda > 0$, then by the Moreau-Rockafellar formula [53, Theorem 16.4], the convex conjugate of $g_1 + g_2$ is the infimal convolution of their conjugates g_1^* and g_2^*, namely

$$(g_1+g_2)^*(\varepsilon) = \inf\left\{g_1^*(\varepsilon_1) + g_2^*(\varepsilon_2) : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \in \mathbb{R}\right\}.$$

Using (34) and (35), this reads as

$$(g_1+g_2)^*(\varepsilon) = \inf\left\{\frac{n\sigma^2}{qc^2} h_1\left(\frac{\varepsilon_1 c}{n\sigma^2}\right) + \frac{(1-\lambda)\varepsilon_2^2}{qn\sigma^2\lambda} h_2\left(\frac{5c\varepsilon_2}{n\sigma^2\lambda}\right) : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \ge 0\right\}.$$

Bounding $h_1(u) \ge \frac{u^2}{2(1+u/3)}$ and $h_2(u) \ge \frac{1}{2(u+2)}$, we have

$$\begin{aligned}(g_1+g_2)^*(\varepsilon) &\ge \inf\left\{\frac{1}{qc^2}\frac{c^2\varepsilon_1^2}{2(n\sigma^2 + c\varepsilon_1/3)} + \frac{(1-\lambda)\varepsilon_2^2}{qn\sigma^2\lambda}\frac{1}{\frac{10c\varepsilon_2}{n\sigma^2\lambda} + 4} : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \ge 0\right\} \\ &\ge \inf\left\{\frac{\varepsilon_1^2}{2q(n\sigma^2 + c\varepsilon_1/3)} + \frac{(1-\lambda)\varepsilon_2^2}{2q}\frac{1}{5c\varepsilon_2 + 2n\sigma^2\lambda} : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \ge 0\right\}.\end{aligned}$$

Using the fact that $\varepsilon_1^2/a + \varepsilon_2^2/b \ge (\varepsilon_1+\varepsilon_2)^2/(a+b)$ for any non-negative $\varepsilon_1, \varepsilon_2$ and positive a, b yield

$$\begin{aligned}(g_1+g_2)^*(\varepsilon) &\ge \inf\left\{\frac{(\varepsilon_1+\varepsilon_2)^2}{2q(n\sigma^2 + c\varepsilon_1/3) + \frac{2q}{(1-\lambda)}(5c\varepsilon_2 + 2n\sigma^2\lambda)} : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \ge 0\right\} \\ &= \inf\left\{\frac{\varepsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda}n\sigma^2 + c\varepsilon_1/3 + \frac{5c\varepsilon}{1-\lambda} - \frac{5c\varepsilon_1}{1-\lambda}} : \varepsilon = \varepsilon_1 + \varepsilon_2,\ \varepsilon_1, \varepsilon_2 \ge 0\right\} \\ &\ge \frac{\varepsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda}n\sigma^2 + \frac{5c\varepsilon}{1-\lambda}},\end{aligned}$$

where we used for the last inequality that for any $\varepsilon_1 \ge 0$,

$$c\varepsilon_1/3 - \frac{5c\varepsilon_1}{1-\lambda} = \frac{c\varepsilon_1}{3(1-\lambda)}(1-\lambda-15) < 0.$$

• If $\lambda = 0$,

$$(g_1+g_2)^*(\varepsilon) = g_1^*(\varepsilon) = \frac{n\sigma^2}{qc^2} h_1\left(\frac{\varepsilon_1 c}{n\sigma^2}\right) \geq \frac{\varepsilon^2/(2q)}{n\sigma^2 + c\varepsilon/3}.$$

We deduce from the previous computations that for any $t, \varepsilon \geq 0$ it holds

$$\begin{aligned}
&\mathbb{P}_\chi\left(\sum_{i=1}^n f_i(X_i) \geq \varepsilon\right) \\
&\leq e^{-\varepsilon t}\mathbb{E}_\chi\left[e^{t\sum_{i=1}^n f_i(X_i)}\right] \text{ using Markov's inequality} \\
&\leq e^{-\varepsilon t}\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \exp\left(\frac{n\sigma^2}{qc^2}(e^{tqc} - tqc - 1) + \frac{n\sigma^2\lambda q t^2}{1-\lambda-5cqt}\right) \text{ using (33)} \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \exp\left(-(g_1+g_2)^*(\varepsilon)\right) \\
&\leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} \times \begin{cases} \exp\left(-\frac{\varepsilon^2/(2q)}{\frac{1+\lambda}{1-\lambda}n\sigma^2 + \frac{5c\varepsilon}{1-\lambda}}\right) & \text{if } \lambda > 0 \\ \exp\left(-\frac{\varepsilon^2/(2q)}{n\sigma^2 + c\varepsilon/3}\right) & \text{if } \lambda = 0 \end{cases}.
\end{aligned}$$

C.3.2 Proof of Lemma 16

The convex conjugate of g_1 is usual and follows from easy computations. We focus on the convex conjugate of g_2 which requires non-trivial computations.
Let $f_\varepsilon(t) = \varepsilon t - \frac{n\sigma^2\lambda q t^2}{1-\lambda-5cqt}$ for any $0 \leq t < (1-\lambda)/(5cq)$. We have for any $0 \leq t < (1-\lambda)/(5cq)$,

$$f_\varepsilon'(t) = \varepsilon - \frac{2n\sigma^2\lambda q t(1-\lambda-5cqt) + 5cq^2 n\sigma^2\lambda t^2}{(1-\lambda-5cqt)^2}.$$

Hence, for $0 \leq t < (1-\lambda)/(5cq)$ such that $f_\varepsilon'(t) = 0$ we have

$$\begin{aligned}
& & \varepsilon(1-\lambda-5cqt)^2 - 2n\sigma^2\lambda q t(1-\lambda-5cqt) - 5cq^2 n\sigma^2\lambda t^2 &= 0 \\
\iff & \varepsilon(1-\lambda)^2 - 10\varepsilon(1-\lambda)cqt + 25\varepsilon c^2q^2t^2 - 2n\sigma^2\lambda q(1-\lambda)t + 10n\sigma^2cq^2\lambda t^2 - 5n\sigma^2cq^2\lambda t^2 &= 0 \\
\iff & \varepsilon(1-\lambda)^2 - 10\varepsilon(1-\lambda)cqt + 25\varepsilon c^2q^2t^2 - 2n\sigma^2\lambda q(1-\lambda)t + 5n\sigma^2cq^2\lambda t^2 &= 0.
\end{aligned}$$

We are looking for the roots of a polynomial of degree 2 in t. The discriminant is

$$\begin{aligned}
\Delta &= (10\varepsilon(1-\lambda)cq + 2n\sigma^2\lambda q(1-\lambda))^2 - 4\varepsilon(1-\lambda)^2(25\varepsilon c^2q^2 + 5n\sigma^2cq^2\lambda) \\
&= 4(1-\lambda)^2q^2\left[(5\varepsilon c + n\sigma^2\lambda)^2 - \varepsilon(25c^2\varepsilon + 5n\sigma^2c\lambda)\right] \\
&= 4(1-\lambda)^2q^2\left[25\varepsilon^2c^2 + 10n\sigma^2\lambda\varepsilon c + n^2\sigma^4\lambda^2 - 25c^2\varepsilon^2 - 5n\sigma^2c\lambda\varepsilon\right] \\
&= 4(1-\lambda)^2q^2\left[5n\sigma^2\lambda\varepsilon c + n^2\sigma^4\lambda^2\right] \\
&= 4(1-\lambda)^2q^2n^2\sigma^4\lambda^2[u+1],
\end{aligned}$$

where $u = \frac{5c\varepsilon}{n\sigma^2\lambda}$.
Hence, the roots of the polynomial of interest are of the form

$$\begin{aligned}
&\frac{10\varepsilon(1-\lambda)cq + 2n\sigma^2\lambda q(1-\lambda) \pm \sqrt{\Delta}}{2[25\varepsilon c^2q^2 + 5n\sigma^2cq^2\lambda]} \\
&= \frac{2(1-\lambda)qn\sigma^2\lambda\left[\frac{5\varepsilon c}{n\sigma^2\lambda} + 1\right] \pm \sqrt{\Delta}}{10q^2cn\sigma^2\lambda\left[\frac{5\varepsilon c}{n\sigma^2\lambda} + 1\right]} \\
&= \frac{1-\lambda}{5cq} \times \frac{u+1 \pm \sqrt{u+1}}{u+1}.
\end{aligned}$$

We deduce that the polynomial has a root in the interval $[0, \frac{1-\lambda}{5cq})$ which is given by

$$t^* := \frac{1-\lambda}{5cq} \times \frac{u+1-\sqrt{u+1}}{u+1},$$

and one can check that this critical point corresponds to a maximum of the function f_ε. We deduce that for any $\varepsilon > 0$,

$$\begin{aligned}
g_2^*(\varepsilon) &= \varepsilon t^* - \frac{n\sigma^2\lambda q (t^*)^2}{1-\lambda-5cqt^*} \\
&= t^* \left\{ \varepsilon - \frac{n\sigma^2\lambda q \frac{1-\lambda}{5cq} \times \frac{u+1-\sqrt{u+1}}{u+1}}{1-\lambda-5cq\frac{1-\lambda}{5cq} \times \frac{u+1-\sqrt{u+1}}{u+1}} \right\} = t^* \left\{ \varepsilon - \frac{n\sigma^2\lambda q \left(u+1-\sqrt{u+1}\right)}{5cq(u+1)-5cq\left(u+1-\sqrt{u+1}\right)} \right\} \\
&= t^* \left\{ \varepsilon - \frac{n\sigma^2\lambda q \left(u+1-\sqrt{u+1}\right)}{5cq\sqrt{u+1}} \right\} = t^* \left\{ \varepsilon - \frac{n\sigma^2\lambda}{5c} \times \frac{\left(u+1-\sqrt{u+1}\right)}{\sqrt{u+1}} \right\} \\
&= t^* \left\{ \varepsilon - \frac{\varepsilon\left(u+1-\sqrt{u+1}\right)}{u\sqrt{u+1}} \right\} = t^*\varepsilon \left\{ \frac{u\sqrt{u+1}-u-1+\sqrt{u+1}}{u\sqrt{u+1}} \right\} = \frac{1-\lambda}{5cq}\varepsilon \times \frac{u+1-\sqrt{u+1}}{u+1} \left\{ \frac{u-\sqrt{u+1}+1}{u} \right\} \\
&= \frac{(1-\lambda)\varepsilon}{q}\frac{1}{5c} \times \left(\sqrt{u+1}-1\right) \left\{ \frac{\sqrt{u+1}-1}{u} \right\} = \frac{(1-\lambda)\varepsilon^2}{qn\sigma^2\lambda} \frac{(\sqrt{u+1}-1)^2}{u^2} = \frac{(1-\lambda)\varepsilon^2}{qn\sigma^2\lambda} h_2(u),
\end{aligned}$$

where $h_2(u) = \frac{(\sqrt{u+1}-1)^2}{u^2}$. However, the function $u \mapsto \sqrt{u+1}$ is analytic on $]0,+\infty[$ and for any $v \in]0,+\infty[$,

$$\sqrt{1+v} = \sum_{k=0}^{\infty} \frac{v^k}{k!} a_k,$$

where $a_0 = 1$ and for all $k \in \mathbb{N}^*$, $a_k = \frac{1}{2}(\frac{1}{2}-1)\ldots(\frac{1}{2}-k+1)$. Hence, we have

$$\begin{aligned}
\frac{\sqrt{v+1}-1}{v} &= \sum_{k=1}^{\infty} \frac{v^{k-1}}{k!} \frac{1}{2}(\frac{1}{2}-1)\ldots(\frac{1}{2}-k+1) = \sum_{k=0}^{\infty} \frac{v^k}{(k+1)!} \frac{1}{2}(\frac{1}{2}-1)\ldots(\frac{1}{2}-k) \\
&= \frac{1}{2}\sum_{k=0}^{\infty} \frac{v^k}{(k+1)!} b_k = \frac{1}{2}\sum_{k=0}^{\infty} \frac{(v/2)^k}{k!} b_k \underbrace{\frac{2^k}{k+1}}_{\geq 1} \\
&\geq \frac{1}{2}\sum_{k=0}^{\infty} \frac{(v/2)^k}{k!} b_k = \frac{1}{2}(v/2+1)^{-1/2} = \frac{1}{\sqrt{2}} \frac{1}{\sqrt{v+2}},
\end{aligned}$$

where we have denoted $b_0 = 1$ and for all $k \in \mathbb{N}^*$, $b_k = \left(-\frac{1}{2}\right)\left(-\frac{1}{2}-1\right)\ldots\left(-\frac{1}{2}-k+1\right)$. Hence we proved that for any $\varepsilon > 0$,

$$g_2^*(\varepsilon) = \frac{(1-\lambda)\varepsilon^2}{qn\sigma^2\lambda} h_2(u) \geq \frac{(1-\lambda)\varepsilon^2}{2qn\sigma^2\lambda} \times \frac{1}{u+2} \text{ with } u = \frac{5c\varepsilon}{n\sigma^2\lambda}.$$

C.4 Complement for the proof of Theorem 1

In this section, we only provide the part of the proof of Theorem 1 that needs to be modified to get the result when the kernels $h_{i,j}$ depend on both i and j and when Assumption 4.(ii) holds. Keeping the notations of Theorem 1, we only want to bound a_1 using a different concentration result that can allow to deal with kernel functions $h_{i,j}$ that might depend on both i and j.

Let us recall that

$$\begin{aligned}
(\mathbb{E}[Z])^k &\leq \mathbb{E}[Z^k] \quad \text{(Using Jensen's inequality)} \\
&= \mathbb{E}\left[\left(\sup_{f_\xi\in\mathscr{F}}\sum_{i=1}^{n-1} f_\xi(X_i)\right)^k\right] \\
&= \mathbb{E}\left[\left(\sup_{f_\xi\in\mathscr{F}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}_{j-1}[p_{i,j}(X_i,X_j')\xi_j(X_j')]\right)^k\right] \\
&= \mathbb{E}\left[\sum_{j=2}^{n}\mathbb{E}_{j-1}\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k\right] \quad \text{(Using Lemma 3)} \\
&= \mathbb{E}\left[\sum_{j=2}^{n}\mathbb{E}_{|X'}\left|\sum_{i=1}^{j-1}p_{i,j}(X_i,X_j')\right|^k\right].
\end{aligned}$$

Thus we have

$$a_1 = \frac{2\delta_M}{1+\varepsilon}\mathbb{E}\sum_{j=2}^{n}\left(\mathbb{E}_{|X'}\left[e^{\alpha(1+\varepsilon)K|C_j|}\right]-\alpha(1+\varepsilon)K\mathbb{E}_{|X'}\left[|C_j|\right]-1\right),$$

where $C_j = \sum_{i=1}^{j-1} p_{i,j}(X_i,X_j')$ and where the notation $\mathbb{E}_{|X'}$ refers to the expectation conditionally to the σ-algebra $\sigma(X_2',\dots,X_n')$.
Now we use a symmetrization trick: since $e^x - x - 1 \geq 0$ for all x and since $e^{a|x|} + e^{-a|x|} = e^{ax} + e^{-ax}$, adding $\mathbb{E}_{|X'}[\exp\left(-\alpha(1+\varepsilon)K|C_j|\right)] + \alpha(1+\varepsilon)K\mathbb{E}_{|X'}[|C_j|] - 1$ to a_1 gives

$$a_1 \leq \frac{2\delta_M}{1+\varepsilon}\mathbb{E}\sum_{j=2}^{n}\left(\mathbb{E}_{|X'}[e^{\alpha(1+\varepsilon)KC_j}]-1+\mathbb{E}_{|X'}[e^{-\alpha(1+\varepsilon)KC_j}]-1\right). \tag{36}$$

Let us consider some $j \in \{2,\dots,n\}$. Conditionally on $\sigma(X_2',\dots,X_n')$, C_j is a sum of bounded functions (by A) depending on the Markov chain. We denote

$$v_j(X_j') = \sum_{i=1}^{j-1}\mathbb{E}_{X_i\sim\pi}[p_{i,j}^2(X_i,X_j')|X_j'] \leq B^2$$

and $V = \sum_{j=2}^{n}\mathbb{E}v_j^k(X_j') \leq C^2B^{2(k-1)}$ (with $C^2 = \sum_{j=2}^{n}\sum_{i=1}^{j-1}\mathbb{E}[p_{i,j}^2(X_i,X_j')]$).
Remark that

$$\begin{aligned}
&\mathbb{E}_{X_i\sim\pi}[p_{i,j}(X_i,X_j')|X_j'] \\
=\ &\mathbb{E}_{X_i\sim\pi}\left[h_{i,j}(X_i,X_j') - \mathbb{E}_{X'\sim\nu}[h_{i,j}(X_i,X')] \mid X_j'\right] \\
=\ &\int_{x'}\left(\int_{x_i}(h_{i,j}(x_i,X_j') - h_{i,j}(x_i,x'))d\pi(x_i)\right)d\nu(x') \\
=\ &0,
\end{aligned}$$

where the last equality comes from Assumption 3. We use a Bernstein inequality for Markov chain (see Proposition 5 in Section C.3). Notice from Taylor expansion that $(1-p/3)(e^p-p-1) \leq p^2/2$ for all $p \geq 0$. Applying (33) with $t = \alpha(1+\varepsilon)K$ and $c = A$ (using the notations of the proof of Proposition 5), we get that for $\alpha < [(1+\varepsilon)K\sqrt{q}(A\sqrt{q}/3 + B\sqrt{3/2})]^{-1} \wedge [(1-\lambda)^{-1/2}(1+\varepsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2} + \sqrt{3\lambda}B\right)]^{-1}$,

$$\begin{aligned}
&\mathbb{E}_{|X'}[e^{\alpha(1+\varepsilon)K|C_j|}] \\
\leq\ &2\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}\times\mathbb{E}_{|X'}\left[\exp\left(\frac{\alpha^2(1+\varepsilon)^2K^2qv_j(X_j')}{2-2Aq\alpha(1+\varepsilon)K/3} + \frac{v_j(X_j')\lambda\alpha^2(1+\varepsilon)^2K^2q}{1-\lambda-5\alpha(1+\varepsilon)KAq}\right)\right].
\end{aligned}$$

Considering $\alpha < [(1+\varepsilon)K\sqrt{q}(A\sqrt{q}/3+B\sqrt{3/2})]^{-1} \wedge [(1-\lambda)^{-1/2}(1+\varepsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2}+\sqrt{3\lambda}B\right)]^{-1}$, $\varepsilon < 1$ and using (36), this leads to

$$
\begin{aligned}
\frac{a_1}{2\left\|\frac{d\chi}{d\pi}\right\|_{\pi,p}} &\leq \frac{2\delta_M}{1+\varepsilon}\sum_{j=2}^{n}\mathbb{E}\left[\exp\left(\frac{\alpha^2(1+\varepsilon)^2K^2q\nu_j(X_j')}{2-2Aq\alpha(1+\varepsilon)K/3}+\frac{\nu_j(X_j')\lambda\alpha^2(1+\varepsilon)^2K^2q}{1-\lambda-5\alpha(1+\varepsilon)KAq}\right)-1\right] \\
&= \frac{2\delta_M}{1+\varepsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{\alpha^2(1+\varepsilon)^2K^2q\nu_j(X_j')}{2-2Aq\alpha(1+\varepsilon)K/3}+\frac{\nu_j(X_j')\lambda\alpha^2(1+\varepsilon)^2K^2q}{1-\lambda-5\alpha(1+\varepsilon)KAq}\right)^k \\
&= \frac{2\delta_M}{1+\varepsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}\left(\frac{3}{2}\right)^{k-1}\left(\frac{\alpha^2(1+\varepsilon)^2K^2q\nu_j(X_j')}{2-2Aq\alpha(1+\varepsilon)K/3}\right)^k \\
&\quad + \frac{2\delta_M}{1+\varepsilon}\sum_{j=2}^{n}\sum_{k=1}^{\infty}\frac{1}{k!}3^{k-1}\left(\frac{\nu_j(X_j')\lambda\alpha^2(1+\varepsilon)^2K^2q}{1-\lambda-5\alpha(1+\varepsilon)KAq}\right)^k \quad \text{(Using Lemma 2)} \\
&\leq \frac{\delta_M}{3(1+\varepsilon)}\sum_{k=1}^{\infty}\frac{3^k\alpha^{2k}(1+\varepsilon)^{2k}K^{2k}q^kV}{(4-4Aq\alpha(1+\varepsilon)K/3)^k}+\frac{2\delta_M}{3(1+\varepsilon)}\sum_{k=1}^{\infty}\frac{3^kV\lambda^k\alpha^{2k}(1+\varepsilon)^{2k}K^{2k}q^k}{(1-\lambda-5\alpha(1+\varepsilon)KAq)^k} \\
&\leq \frac{\delta_M}{3(1+\varepsilon)}\sum_{k=1}^{\infty}\frac{3^k\alpha^{2k}(1+\varepsilon)^{2k}K^{2k}q^kC^2B^{2(k-1)}}{(2-2Aq\alpha(1+\varepsilon)K/3)^k}+\frac{2\delta_M}{3(1+\varepsilon)}\sum_{k=1}^{\infty}\frac{3^kC^2B^{2(k-1)}\lambda^k\alpha^{2k}(1+\varepsilon)^{2k}K^{2k}q^k}{(1-\lambda-5\alpha(1+\varepsilon)KAq)^k} \\
&= \frac{(1+\varepsilon)C^2\alpha^2K^2\delta_Mq}{2-2Aq\alpha(1+\varepsilon)K/3-3\alpha^2(1+\varepsilon)^2K^2B^2q}+\frac{2\delta_MC^2\lambda\alpha^2(1+\varepsilon)K^2q}{1-\lambda-5\alpha(1+\varepsilon)KAq-3B^2\lambda\alpha^2(1+\varepsilon)^2K^2q} \\
&= \frac{(1+\varepsilon)C^2\alpha^2K^2\delta_Mq/2}{1-Aq\alpha(1+\varepsilon)K/3-3\alpha^2(1+\varepsilon)^2K^2B^2q/2} \\
&\quad + \frac{2\delta_MC^2\lambda\alpha^2(1+\varepsilon)K^2q(1-\lambda)^{-1}}{1-5(1-\lambda)^{-1}\alpha(1+\varepsilon)KAq-3B^2\lambda(1-\lambda)^{-1}\alpha^2(1+\varepsilon)^2K^2q} \\
&\leq \frac{(1+\varepsilon)C^2\alpha^2K^2\delta_Mq/2}{1-\alpha(1+\varepsilon)K\sqrt{q}(A\sqrt{q}/3+B\sqrt{3/2})} \\
&\quad + \frac{2\delta_MC^2\lambda\alpha^2(1+\varepsilon)K^2q(1-\lambda)^{-1}}{1-\alpha(1-\lambda)^{-1/2}(1+\varepsilon)K\sqrt{q}\left(5A\sqrt{q}(1-\lambda)^{-1/2}+\sqrt{3\lambda}B\right)}.
\end{aligned}
$$

From this bound on a_1, one can follow exactly the steps of the proof of Theorem 1 to conclude the proof.

C.5 Proof of Theorem 3.

We consider any $R \in \mathbb{N}^*$. We remark that for any $x, y \in E$,

$$
\begin{aligned}
|h_R(x,y)| &= \left|\sum_{r=1}^{R}\lambda_r\varphi_r(x)\varphi_r(y)\right| \\
&\leq \left(\sum_{r=1}^{R}|\lambda_r|\varphi_r(x)^2\right)^{1/2}\times\left(\sum_{r=1}^{R}|\lambda_r|\varphi_r(y)^2\right)^{1/2} \quad \text{(Using Cauchy-Schwarz inequality)} \\
&\leq \Upsilon^2 S,
\end{aligned}
$$

which proves that $\|h_R\|_\infty \leq \Upsilon^2 S$. Similar computations lead to $\|h - h_R\|_\infty \leq \Upsilon^2 S$.

Using Theorem 7 we get for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \frac{\Upsilon^4 S^2(1+\kappa)\log n}{n} + 2\sum_{i>R, i\in I}\lambda_i^2 + t\right)\\
\leq\ & 16\exp\left(-n\frac{t^2}{Km^2\tau^2S^2\Upsilon^4}\right) + \beta\log(n)\exp\left(-\frac{n}{16\log n}\left\{\left[\frac{t}{\kappa\Upsilon^2 S}\right]\wedge\left[\frac{t}{\kappa\Upsilon^2 S}\right]^{1/2}\right\}\right)\\
+\ & 16R^2\exp\left(-\frac{nt}{Km^2\tau^2R^2\Lambda^2\Upsilon^4}\right).
\end{aligned}$$

Choosing $R^2 = \lceil\sqrt{n}\rceil$, we get

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{4}\delta_2(\lambda(\mathbf{H}),\lambda(\mathbf{H}_n))^2 \geq \frac{\Upsilon^4 S^2(1+\kappa)\log n}{n} + 2\sum_{i>\lceil n^{1/4}\rceil, i\in I}\lambda_i^2 + t\right)\\
\leq\ & 32\sqrt{n}\exp\left(-\mathscr{C}\min\left(nt^2, \sqrt{n}t\right)\right) + \beta\log(n)\exp\left(-\frac{n}{\log n}\min\left(\mathscr{B}t, (\mathscr{B}t)^{1/2}\right)\right),
\end{aligned}$$

where $\mathscr{B} = \left(K\Upsilon^2\kappa S\right)^{-1}$ and $\mathscr{C} = \left(Km^2\tau^2S^2\Upsilon^4\right)^{-1}$.

C.6 Proof of Lemma 11

Let

$$T^* := \arg\min_{c_n\leq t<n-1}(\mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)),$$

and $h^* = h_{T^*-b_n}$ is the corresponding hypothesis that minimizes the penalized true risk and let $\widehat{\mathscr{R}}^* = \widehat{\mathscr{R}}(h^*, T^*+1)$ to be the penalized empirical risk of $h_{T^*-b_n}$. Set, for brevity

$$\widehat{\mathscr{R}}_{t-b_n} = \widehat{\mathscr{R}}(h_{t-b_n}, t+1),$$

and let

$$\widehat{T} := \arg\min_{c_n\leq t<n-1}(\widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t)),$$

where $\widehat{h}$ coincides with $h_{\widehat{T}-b_n}$. Using this notation and since

$$\widehat{\mathscr{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*),$$

holds with certainty, we have

$$\begin{aligned}
&\mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + \mathscr{E}\right)\\
=\ & \mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right)\\
\leq\ & \mathbb{P}\left(\bigcup_{c_n\leq t\leq n-1}\left\{\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right\}\right)\\
\leq\ & \sum_{t=c_n}^{n-1}\mathbb{P}\left(\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right),
\end{aligned}$$

where $\mathscr{E}$ is a positive-valued random variable to be specified. Now we remark that if

$$\widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*), \tag{37}$$

holds, then at least one of the following three conditions must hold

$$\begin{array}{lll}
(i) & \widehat{\mathscr{R}}_{t-b_n} & \leq \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t)\\
(ii) & \widehat{\mathscr{R}}^* & > \mathscr{R}(h^*) + c_\gamma(n-T^*)\\
(iii) & \mathscr{R}(h_{t-b_n}) - \mathscr{R}(h^*) & \leq 2c_\gamma(n-T^*).
\end{array}$$

Stated otherwise, if (37) holds for some $t \in \{c_n, \dots, n-1\}$ then

- either $t = T^*$ and (iii) holds trivially.
- or $t \neq T^*$ which can occur because
  - $\widehat{\mathscr{R}}_{t-b_n}$ underestimates $\mathscr{R}(h_{t-b_n})$ and (i) holds.
  - $\widehat{\mathscr{R}}^*$ overestimates $\mathscr{R}(h^*)$ and (ii) holds.
  - n is too small to statistically distinguish $\mathscr{R}(h_{t-b_n})$ and $\mathscr{R}(h^*)$, and (iii) holds.

Therefore, for any fixed t, we have

$$\begin{aligned} & \mathbb{P}\left(\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \leq \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right) \\ \leq \; & \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} \leq \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + \mathbb{P}\left(\widehat{\mathscr{R}}^* > \mathscr{R}(h^*) + c_\gamma(n-T^*)\right) \\ & + \mathbb{P}\left(\mathscr{R}(h_{t-b_n}) - \mathscr{R}(h^*) \leq 2c_\gamma(n-T^*)\,,\; \mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}\right). \end{aligned}$$

By choosing $\mathscr{E} = 2c_\gamma(n-T^*)$, the last term in the previous inequality is zero and we can write

$$\begin{aligned} & \mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + 2c_\gamma(n-T^*)\right) \\ \leq \; & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} \leq \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + (n-c_n)\mathbb{P}\left(\widehat{\mathscr{R}}^* > \mathscr{R}(h^*) + c_\gamma(n-T^*)\right) \\ \leq \; & (n-c_n)\frac{\gamma}{(n-c_n)(n-c_n+1)} + (n-c_n)\left\{\sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} > \mathscr{R}(h_{t-b_n}) + c_\gamma(n-t)\right)\right\} \text{ (Using (30))} \\ \leq \; & \frac{\gamma}{n-c_n+1} + (n-c_n)^2\frac{\gamma}{(n-c_n)(n-c_n+1)} \text{ (Using (30))} \\ \leq \; & \frac{\gamma}{n-c_n+1} + (n-c_n)\frac{\gamma}{n-c_n+1} = \gamma. \end{aligned}$$

C.7 Proof of Theorem 6

In the following, $\mathbb{P}_g$ will denote the distribution of the Markov chain if the invariant distribution of the chain is assumed to have a density g with respect to the Lebesgue measure on $\mathbb{R}$. We consider $q = q_1 \vee q_2$ where $q_1, q_2 \in [1, \infty)$ are such that $\frac{1}{p_1} + \frac{1}{q_1} = 1$ and $\frac{1}{p_2} + \frac{1}{q_2} = 1$.

The main tool of the proof is the Hoeffding (also called canonical) decomposition of the U-statistics $\widehat{\theta}_m$. We introduce the processes U_n and P_n defined by

$$U_n(h) = \frac{1}{n(n-1)} \sum_{i \neq j=1}^{n} h(X_i, X_j), \quad P_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

We also define $P(h) = \langle h, f \rangle$. By setting, for all $m \in \mathscr{M}$,

$$H_m(x,y) = \sum_{l \in \mathscr{L}_m} (p_l(x) - a_l)(p_l(y) - a_l),$$

with $a_l = \langle f, p_l \rangle$, we obtain the decomposition

$$\widehat{\theta}_m = U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) + \|\Pi_{S_m}(f)\|_2^2.$$

Let us consider β in $]0,1[$. Since

$$\mathbb{P}_f(T_\alpha \leq 0) = \mathbb{P}_f\left(\sup_{m \in \mathscr{M}} (\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_0(X_i) - t_m(u_\alpha)) \leq 0\right),$$

we have

$$\mathbb{P}_f(T_\alpha \leq 0) \leq \inf_{m \in \mathscr{M}} \mathbb{P}_f\left(\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^{n} f_0(X_i) - t_m(u_\alpha) \leq 0\right).$$

Since $\|f-\Pi_{S_m}(f)\|_2^2 = \|f\|_2^2 - \|\Pi_{S_m}(f)\|_2^2$, it holds

$$\begin{aligned}
&\widehat{\theta}_m + \|f_0\|_2^2 - \frac{2}{n}\sum_{i=1}^n f_0(X_i) \\
= \;& U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) - \|f-\Pi_{S_m}(f)\|_2^2 + \|f\|_2^2 + \|f_0\|_2^2 - 2P_n(f_0) \\
= \;& U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f)) - \|f-\Pi_{S_m}(f)\|_2^2 + \|f-f_0\|_2^2 + 2P(f_0) - 2P_n(f_0),
\end{aligned}$$

which leads to

$$\begin{aligned}
\mathbb{P}_f(T_\alpha \le 0) \le \inf_{m\in\mathscr{M}} \mathbb{P}_f\Big(& U_n(H_m) + (P_n-P)(2\Pi_{S_m}(f)-2f) + (P_n-P)(2f-2f_0) + \|f-f_0\|_2^2 \\
& \le \|f-\Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha)\Big). \qquad (38)
\end{aligned}$$

We then need to control $U_n(H_m), (P_n-P)(2\Pi_{S_m}(f)-2f)$, $(P_n-P)(2f-2f_0)$ for every $m\in\mathscr{M}$.

Control of $U_n(H_m)$. H_m is π-canonical and a direct application of Theorem 2 leads to the following Lemma (the proof of Lemma 17 is postponed to Section C.8).

Lemma 17 *Let us assume that the invariant distribution of the Markov chain $(X_i)_{i\ge1}$ has density f with respect to the Lebesgue measure on $\mathbb{R}$. For all $m=(l,D)$ with $l\in\{1,2,3\}$ and $D\in\mathbb{D}_l$, introduce $\{p_l, l\in\mathscr{L}_m\}$ defined as in page 18 and $Z_m = \frac{1}{n(n-1)}\sum_{i\ne j=1}^n H_m(X_i,X_j)$, with $H_m(x,y) = \sum_{l\in\mathscr{L}_m}(p_l(x)-\langle f,p_l\rangle)(p_l(y)-\langle f,p_l\rangle)$. There exist some constants $C,\beta>0$ (both depending on the Markov chain $(X_i)_{i\ge1}$ while C also depends on φ) such that, for all $l\in\{1,2,3\}$, $D\in\mathbb{D}_l$ and $u\ge1$, it holds with probability at least $1-\beta e^{-u}\log n$,*

$$|Z_{(l,D)}| \le C\,(\|f\|_\infty+1)\,DR(n,u),$$

where $R(n,u) = \log n\left\{\frac{u}{n} + \left[\frac{u}{n}\right]^2\right\}$.

We deduce that there exist $C,\beta>0$ such that for any $\gamma\in(0,1\wedge(e^{-1}3\beta\log n))$ and any $m=(l,D)\in\mathscr{M}$,

$$\mathbb{P}_f\left(U_n(H_m) \le -C\,(\|f\|_\infty+1)\,DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right) \le \gamma/3. \qquad (39)$$

From (38) and (39) we get that

$$\begin{aligned}
\mathbb{P}_f(T_\alpha\le0) \le \frac{\gamma}{3} + \inf_{m\in\mathscr{M}} \mathbb{P}_f\Big(& (P_n-P)(2\Pi_{S_m}(f)-2f) + (P_n-P)(2f-2f_0) + \|f-f_0\|_2^2 \\
& \le \|f-\Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha) + C\,(\|f\|_\infty+1)\,DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\Big). \qquad (40)
\end{aligned}$$

Control of $(P_n-P)(2\Pi_{S_m}(f)-2f)$. It is easy to check that there exists some constant $C'>0$ such that for all l in $\{1,2\}$, D in $\mathbb{D}_l$,

$$\left|2\Pi_{S_{(l,D)}}(f)(X_i) - 2f(X_i)\right| \le C'\|f\|_\infty.$$

Indeed,

- when $l=1$, for any $k\in\mathbb{Z}$,

$$\langle\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}, f\rangle = \int \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)f(x)dx \le D^{-1/2}\|f\|_\infty.$$

  Hence,

$$\begin{aligned}
&\sup_x |\Pi_{S_{(1,D)}}(f)(x)| \le \sup_x \sum_{k\in\mathbb{Z}} \left|\langle\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}, f\rangle\right| \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x) \\
\le\;& D^{-1/2}\|f\|_\infty \sup_x \sum_{k\in\mathbb{Z}} \sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x) = \|f\|_\infty.
\end{aligned}$$

- when $l=2$, $D=2^J$ for some $J\in\mathbb{N}$ and we have for any $k\in\mathbb{Z}$,

$$\langle\varphi_{J,k},f\rangle=\int 2^{J/2}\varphi(2^Jx-k)f(x)dx\le\|f\|_\infty\int 2^{J/2}|\varphi(2^Jx)|dx\le 2^{-J/2}\|f\|_\infty\|\varphi\|_1.$$

Hence,

$$\begin{aligned}\sup_x|\Pi_{S_{(2,D)}}(f)(x)|&\le\sup_x\sum_{k\in\mathbb{Z}}\left|\langle\varphi_{J,k},f\rangle\right|\times|\varphi_{J,k}(x)|\\ &\le\ 2^{-J/2}\|f\|_\infty\|\varphi\|_1\sup_x\sum_{k\in\mathbb{Z}}|2^{J/2}\varphi(2^Jx-k)|\le c\|f\|_\infty\|\varphi\|_1,\end{aligned}$$

where $c>0$ is a constant depending only on φ since φ is bounded and compactly supported. Stated otherwise, there is only a finite number of integers $k\in\mathbb{Z}$ (which is independent of x and J) such that for any $x\in\mathbb{R}$ and any $J\in\mathbb{Z}$, 2^Jx-k falls into the support of φ.

Moreover, it is proved in [20, Page 269], that one can take C' such that for all D in $\mathbb{D}_3$,

$$|2\Pi_{S_{(3,D)}}(f)(X_i)-2f(X_i)|\le C'\|f\|_\infty\log(D+1).$$

Since

$$\mathbb{E}_{X\sim\pi}\left(2\Pi_{S_m}(f)(X)-2f(X)\right)^2\le 4\|f\|_\infty\|\Pi_{S_m}(f)-f\|_2^2,$$

we can deduce using Proposition 5 (see Section C.3) that for all $m=(l,D)\in\mathscr{M}$,

$$\begin{aligned}\mathbb{P}_f\Bigg((P_n-P)(2\Pi_{S_m}(f)-2f)<-\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty\log(D+1)}{n}\\ -2\sqrt{\frac{2\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{n}}\|\Pi_{S_m}(f)-f\|_2\Bigg)\le\frac{\gamma}{3}.\end{aligned}$$

Considering some $\varepsilon\in]0,2[$, we use the inequality $\forall a,b\in\mathbb{R}$, $2ab\le 4a^2/\varepsilon+\varepsilon b^2/4$ and we obtain that for any $m=(l,D)\in\mathscr{M}$,

$$\begin{aligned}\mathbb{P}_f\Bigg((P_n-P)(2\Pi_{S_m}(f)-2f)+\frac{\varepsilon}{4}\|\Pi_{S_m}(f)-f\|_2^2<-\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty\log(D+1)}{n}\\ -\frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}\Bigg)\le\frac{\gamma}{3}.\end{aligned}\tag{41}$$

The control of $(P_n-P)(2f-2f_0)$ is computed in the same way and we get

$$\begin{aligned}\mathbb{P}_f\Bigg((P_n-P)(2f-2f_0)+\frac{\varepsilon}{4}\|f-f_0\|_2^2<-\frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty+\|f_0\|_\infty)}{n}\\ -\frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}\Bigg)\le\frac{\gamma}{3}.\end{aligned}\tag{42}$$

Finally, we deduce from (40), (41) and (42) that if there exists some $m=(l,D)$ in $\mathscr{M}$ such that

$$\begin{aligned}\left(1-\frac{\varepsilon}{4}\right)\|f-f_0\|_2^2>\left(1+\frac{\varepsilon}{4}\right)\|f-\Pi_{S_m}(f)\|_2^2+\frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}+\frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty+\|f_0\|_\infty)}{n}\\ +\frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}+\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty\log(D+1)}{n}\\ +t_m(u_\alpha)+C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),\end{aligned}$$

i.e. such that

$$\left(1-\frac{\varepsilon}{4}\right)\|f-f_0\|_2^2 > \left(1+\frac{\varepsilon}{4}\right)\|f-\Pi_{S_m}(f)\|_2^2 + \frac{16\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n} + 4\left(\|f\|_\infty(C'\log(D+1)+1)+\|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)qA_1}{n} + t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),$$

then

$$\mathbb{P}_f\left(T_\alpha \le 0\right) \le \gamma.$$

To conclude the proof of Theorem 6, it suffices to notice that for any $\varepsilon \in]0,2[$, choosing $\eta > 0$ such that $1+\eta = \frac{1+\frac{\varepsilon}{4}}{1-\frac{\varepsilon}{4}}$ leads to $\varepsilon = \frac{4\eta}{2+\eta}$. One can immediately check that the condition $\varepsilon \in]0,2[$ is equivalent to $\eta \in]0,2[$. Noticing further that $\frac{1}{\varepsilon} = \frac{2+\eta}{4\eta} < \frac{2+2}{4\eta} = \frac{1}{\eta}$, we deduce that for any $\eta \in]0,2[$, if

$$\|f-f_0\|_2^2 > (1+\eta)\left\{\|f-\Pi_{S_m}(f)\|_2^2 + \frac{16\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\eta n} + 4\left(\|f\|_\infty(C'\log(D+1)+1)+\|f_0\|_\infty\right)\frac{\log(3C_\chi/\gamma)qA_1}{n} + t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right\},$$

then

$$\mathbb{P}_f\left(T_\alpha \le 0\right) \le \gamma.$$

C.8 Proof of Lemma 17

Lemma 17 will follow from Theorem 2 if we can show that the function H_m is bounded. Let us denote $m = (l,D)$ for some $l \in \{1,2,3\}$ and $D \in \mathbb{D}_l$. Let us first remark that the Bessel's inequality states that

$$\sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \le \|f\|_2^2 = \int f(x)f(x)dx \le \|f\|_\infty, \tag{43}$$

since $\int f(x)dx = 1$ and $f(x) \ge 0, \quad \forall x$.

- If $l = 1$, then we notice that for any $k \in \mathbb{Z}$,

$$\begin{aligned}|\langle\sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}, f\rangle| &= \left|\int \sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}(x)f(x)dx\right| \\ &\le \|f\|_\infty\sqrt{D}\int \mathbb{1}_{]k/D,(k+1)/D[}(x)dx \\ &\le D^{-1/2}\|f\|_\infty.\end{aligned}$$

Then for any $x, y \in \mathbb{R}$ it holds

$$\begin{aligned}|H_m(x,y)| &\le \sum_{k\in\mathcal{L}_m}|p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m}|p_k(x)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|p_k(y)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le \sum_{k\in\mathbb{Z}} D\mathbb{1}_{]k/D,(k+1)/D[}(x)\mathbb{1}_{]k/D,(k+1)/D[}(y) \\ &\quad + 2\sup_z\sum_{k\in\mathbb{Z}}\sqrt{D}|\mathbb{1}_{]k/D,(k+1)/D[}(z)| \times |\langle\sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}, f\rangle| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le D + 2\|f\|_\infty + \|f\|_\infty,\end{aligned}$$

where in the last inequality we used (43).

- If $l=2$ then $D=2^J$ for some $J\in\mathbb{N}$ and we have for any $k\in\mathbb{Z}$,

$$\langle\varphi_{J,k},f\rangle=\int 2^{J/2}\varphi(2^Jx-k)f(x)dx\le\|f\|_\infty\int 2^{J/2}|\varphi(2^Jx)|dx\le 2^{-J/2}\|f\|_\infty\|\varphi\|_1.$$

We get that for any $x,y\in\mathbb{R}$,

$$\begin{aligned}|H_m(x,y)|&\le\sum_{k\in\mathscr{L}_m}|p_k(x)p_k(y)|+\sum_{k\in\mathscr{L}_m}|p_k(x)\langle p_k,f\rangle|+\sum_{k\in\mathscr{L}_m}|p_k(y)\langle p_k,f\rangle|+\sum_{k\in\mathscr{L}_m}|\langle p_k,f\rangle|^2\\&\le\sum_{k\in\mathbb{Z}}2^J\varphi(2^Jx-k)\varphi(2^Jy-k)+2\sup_z\sum_{k\in\mathbb{Z}}2^{-J/2}\|f\|_\infty\|\varphi\|_1 2^{J/2}|\varphi(2^{J/2}z-k)|+\sum_{k\in\mathscr{L}_m}|\langle p_k,f\rangle|^2\\&\le c2^J+c'\|\varphi\|_1\|f\|_\infty+\|f\|_\infty\\&=cD+c'\|\varphi\|_1\|f\|_\infty+\|f\|_\infty,\end{aligned}$$

for some constants $c,c'>0$. In the last inequality we used (43) and the fact φ is bounded and compactly supported. Indeed, this implies that there is only a finite number of integers $k\in\mathbb{Z}$ (which is independent of x and J) such that for any $x\in\mathbb{R}$ and any $J\in\mathbb{Z}$, 2^Jx-k falls into the support of φ.
- If $l=3$ then we easily get for any $x,y\in[0,1]$,

$$\begin{aligned}|H_m(x,y)|&\le\sum_{k\in\mathscr{L}_m}|p_k(x)p_k(y)|+\sum_{k\in\mathscr{L}_m}|p_k(x)\langle p_k,f\rangle|+\sum_{k\in\mathscr{L}_m}|p_k(y)\langle p_k,f\rangle|+\sum_{k\in\mathscr{L}_m}|\langle p_k,f\rangle|^2\\&\le 2D+4D\|f\|_\infty+\|f\|_\infty.\end{aligned}$$

We deduce that in any case, H_m is bounded $c(1+\|f\|_\infty)D$ for some constant $c>0$ (depending only on φ) which concludes the proof of Lemma 17.

C.9 Proof of Corollary 1

Step 1: We start by providing an upper bound on $t_m(u_\alpha)$ with Lemma 18.

Lemma 18 *There exists a constant $C(\alpha)>0$ such that for any $m=(l,D)\in\mathscr{M}$ it holds,*

$$t_m(u_\alpha)\le W_m(\alpha),$$

where

$$W_m(\alpha)=C(\alpha)(\|f_0\|_\infty+1)\left[DR(n,\log\log n)+\frac{\log\log n}{n}\right].$$

Proof of Lemma 18.
Let us recall that $t_m(u)$ denotes the $(1-u)$ quantile of the distribution of $\widehat{T}_m$ under the null hypothesis. One can easily see that $|\mathscr{M}|\le 3(1+\log_2 n)$. So, setting $\alpha_n=\alpha/(3(1+\log_2 n))$,

$$\begin{aligned}\mathbb{P}_{f_0}(\sup_{m\in\mathscr{M}}(\widehat{T}_m-t_m(\alpha_n))>0)&\le\sum_{m\in\mathscr{M}}\mathbb{P}_{f_0}(\widehat{T}_m-t_m(\alpha_n)>0)\\&\le\sum_{m\in\mathscr{M}}\alpha/(3(1+\log_2 n))\\&\le\alpha.\end{aligned}$$

By definition of u_α, this implies that $\alpha_n\le u_\alpha$ and for all $m\in\mathscr{M}$,

$$t_m(u_\alpha)\le t_m(\alpha_n).$$

Hence it suffices to upper bound $t_m(\alpha_n)$. Let $m=(l,D)\in\mathscr{M}$. We use the same notation as in the proof of Theorem 6 to obtain that

$$\widehat{T}_m=U_n(H_m)+(P_n-P)(2\Pi_{S_m}(f))-2P_n(f_0)+\|f_0\|_2^2+\|\Pi_{S_m}(f)\|_2^2.$$

Under the null hypothesis, this reads as

$$\begin{aligned}\widehat{T}_m &= U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|f_0\|_2^2 + \|\Pi_{S_m}(f_0)\|_2^2 \\ &= U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|f_0 - \Pi_{S_m}(f_0)\|_2^2.\end{aligned}$$

We control $U_n(H_m)$ and $(P_n - P)(2\Pi_{S_m}(f_0) - 2f_0)$ exactly like in the proof of Theorem 6. From Lemma 17, there exist $C, \beta > 0$ such that for any $m = (l, D) \in \mathscr{M}$, it holds

$$\mathbb{P}_{f_0}\left(U_n(H_m) \leq C\left(\|f_0\|_\infty + 1\right) DR\left(n, \log\left\{\frac{2\beta \log n}{\alpha_n}\right\}\right)\right) \leq \alpha_n/2. \tag{44}$$

Moreover, since

$$|2\Pi_{S_{(l,D)}}(f_0)(X_i) - 2f_0(X_i)| \leq C'\|f_0\|_\infty \log(D+1),$$

and

$$\mathbb{E}_{X\sim\pi}\left(2\Pi_{S_m}(f_0)(X) - 2f_0(X)\right)^2 \leq 4\|f_0\|_\infty \|\Pi_{S_m}(f_0) - f_0\|_2^2,$$

we get using Proposition 5 (see Section C.3) that for all $m = (l, D) \in \mathscr{M}$,

$$\begin{aligned}\mathbb{P}_{f_0}\Bigg((P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) > \frac{2C'\log(2C_\chi/\alpha_n) q A_1 \|f_0\|_\infty \log(D+1)}{n} \\ + 2\sqrt{\frac{2\log(2C_\chi/\alpha_n) q A_2 \|f_0\|_\infty}{n}} \|\Pi_{S_m}(f_0) - f_0\|_2\Bigg) \leq \frac{\alpha_n}{2}.\end{aligned}$$

Using the inequality $\forall a, b \in \mathbb{R},\ 2ab \leq a^2 + b^2$, and the fact that for $n \geq 16$, $\log(D+1) \leq \log(n^2+1)$, we obtain that there exists $C'' > 0$ such that

$$\mathbb{P}_{f_0}\left((P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) - \|\Pi_{S_m}(f_0) - f_0\|_2^2 > \frac{C''\|f_0\|_\infty \log(2C_\chi/\alpha_n)\log(n)}{n}\right) \leq \frac{\alpha_n}{2}.$$

We deduce that it holds

$$\mathbb{P}_{f_0}\left(\widehat{T}_m > C\left(\|f_0\|_\infty + 1\right) DR\left(n, \log\left\{\frac{2\beta\log n}{\alpha_n}\right\}\right) + \frac{C''\|f_0\|_\infty \log(2C_\chi/\alpha_n)\log(n)}{n}\right) \leq \alpha_n.$$

Noticing that there exists some constant $c(\alpha) > 0$ such that

$$\log\left\{\frac{2\beta\log n}{\alpha_n}\right\} \vee \log(2C_\chi/\alpha_n) \leq c(\alpha)\log\log n,$$

we deduce by definition of $t_m(\alpha_n)$ that for some $c(\alpha) > 0$,

$$t_m(\alpha_n) \leq c(\alpha) C\left(\|f_0\|_\infty + 1\right) DR(n, \log\log n) + c(\alpha)\frac{C''\|f_0\|_\infty \log\log n}{n}.$$

■

Step 2: Proof of Corollary 1.
Let us fix $\gamma \in]0,1[$ and $l \in \{1,2,3\}$. From Theorem 6 and Lemma 18, we deduce that if f satisfies

$$\|f - f_0\|_2^2 > (1+\varepsilon) \inf_{D\in\mathscr{D}_l} \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V(l,D)(\gamma),$$

then

$$\mathbb{P}_f(T_\alpha \leq 0) \leq \gamma.$$

It is thus a matter of giving an upper bound for

$$\inf_{D\in\mathscr{D}_l}\left\{\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V_{(l,D)}(\gamma)\right\},$$

when f belongs to some specified classes of functions. Recall that

$$\mathscr{B}_s^{(l)}(P,M) = \{f \in L_2(\mathbb{R}) \mid \forall D \in \mathscr{D}_l, \|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2 D^{-2s}, \|f\|_\infty \leq M\}.$$

We now assume that f belongs to $\mathscr{B}_s^{(l)}(P,M)$. Since $\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 \leq P^2 D^{-2s}$, we only need an upper bound for

$$\inf_{D\in\mathscr{D}_l}\left\{P^2D^{-2s} + C(\alpha)(\|f_0\|_\infty + 1)\left[DR(n,\log\log n) + \frac{\log\log n}{n}\right] + C_1\|f\|_\infty\frac{\log(3C_\chi/\gamma)}{\varepsilon n}\right.$$
$$\left. + C_2(\|f\|_\infty\log(D+1) + \|f_0\|_\infty)\frac{\log(3C_\chi/\gamma)}{n} + C_3(\|f\|_\infty + 1)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right\}.$$

Using that f belongs to $\mathscr{B}_s^{(l)}(P,M)$ and the fact that

$$R(n,\log\log n) \vee R\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right) \lesssim \log(n)\frac{\log\log n}{n},$$

where $\lesssim$ states that the inequality holds up to some multiplicative constant independent of n, D and P, we deduce that we want to upper bound

$$\inf_{D\in\mathscr{D}_l}\left\{P^2D^{-2s} + D\log(n)\frac{\log\log n}{n} + \frac{\log\log n}{n} + \frac{\log(D+1)}{n}\right\}.$$

Since $\log(D+1) \leq D$ for all $D \in \mathscr{D}_l$, we only need to focus on

$$\inf_{D\in\mathscr{D}_l}\left\{P^2D^{-2s} + D\log(n)\frac{\log\log n}{n}\right\}.$$

$P^2D^{-2s} < D\log(n)\frac{\log\log n}{n}$ if and only if $D > \left(\frac{P^4n^2}{\log^2(n)(\log\log n)^2}\right)^{\frac{1}{4s+2}}$. Hence we define D_* by

$$\log_2(D_*) := \lfloor \log_2\left(\left(\frac{P^4n^2}{\log^2(n)(\log\log n)^2}\right)^{\frac{1}{4s+2}}\right)\rfloor + 1.$$

We consider three cases.

- If $D_* < 1$, then $P^2D^{-2s} < D\log(n)\frac{\log\log n}{n}$ for any $D \in \mathscr{D}_l$ and by choosing $D_0 = 1$ to upper bound the infimum we get

$$\inf_{D\in\mathscr{D}_l}\left\{\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V_{(l,D)}(\gamma)\right\} \leq \log(n)\frac{\log\log n}{n}.$$

- If $D_* > 2^{\lfloor \log_2(n/(\log(n)\log\log n)^2)\rfloor}$, then $P^2D^{-2s} > D\log(n)\frac{\log\log n}{n}$ for any $D \in \mathscr{D}_l$ and by choosing $D_0 = 2^{\log_2(\lfloor n/(\log(n)\log\log n)^2)\rfloor)}$ to upper bound the infimum we get

$$\inf_{D\in\mathscr{D}_l}\left\{\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V_{(l,D)}(\gamma)\right\} \lesssim 2P^2D_0^{-2s} \leq 2^{2s+1}P^2\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{2s}.$$

- Otherwise D_* belongs to $\mathscr{D}_l$ and we upper bound the infimum by choosing $D_0 = D_*$ and we get

$$\inf_{D\in\mathscr{D}_l}\left\{\|f - \Pi_{S_{(l,D)}}(f)\|_2^2 + W_{(l,D)}(\alpha) + V_{(l,D)}(\gamma)\right\} \lesssim 4P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}}.$$

The proof of Corollary 1 ends with simple computations that we provide below for the sake of completeness. Since

$$\log(n)\frac{\log\log n}{n} \leq P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}}$$
$$\Leftrightarrow \left(\log(n)\frac{\log\log n}{n}\right)^{1/2} \leq P.$$

and since

$$\begin{aligned}
& P^2\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{2s} && \leq P^{\frac{2}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{2s}{2s+1}} \\
\Leftrightarrow \quad & P\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s} && \leq P^{\frac{1}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}} \\
\Leftrightarrow \quad & P^{2s}\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s(2s+1)} && \leq \left(\frac{\log(n)\log\log n}{n}\right)^{s} \\
\Leftrightarrow \quad & P\left(\frac{(\log(n)\log\log n)^2}{n}\right)^{s+1/2} && \leq \left(\frac{\log(n)\log\log n}{n}\right)^{1/2} \\
\Leftrightarrow \quad & P && \leq \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}},
\end{aligned}$$

we deduce that if P is chosen such that

$$\left(\log(n)\frac{\log\log n}{n}\right)^{1/2} \leq P \leq \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}, \tag{45}$$

then the uniform separation rate of the test $\mathbb{1}_{T_\alpha>0}$ over $\mathscr{B}_s^{(l)}(P,M)$ satisfies

$$\rho\left(\mathbb{1}_{T_\alpha>0}, \mathscr{B}_s^{(l)}(P,M), \gamma\right) \leq C' P^{\frac{1}{2s+1}}\left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}}. \tag{46}$$

Remark This final statement can allow the reader to understand our choice for the size of the model $|\mathscr{M}|$ that we considered. Indeed, we chose for any $l \in \{1,2,3\}$, $\mathscr{D}_l = \{2^J, 0 \leq J \leq \log_2\left(n/(\log(n)\log\log n)^2\right)\}$ in order to ensure that for values of P saturing the right inequality in (45) (i.e. for $P \approx \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}$), the upper-bound in (46) still tends to zero as n goes to $+\infty$ for any possible values of the smoothness parameter s.