Three rates of convergence or separation via U-statistics in a dependent framework

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

June 28, 2021

Abstract

Despite the ubiquity of U-statistics in modern Probability and Statistics, their non-asymptotic analysis in a dependent framework may have been overlooked. In a recent work, a new concentration inequality for U-statistics of order two for uniformly ergodic Markov chains has been proved. In this paper, we put this theoretical breakthrough into action by pushing further the current state of knowledge in three different active fields of research.

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 some 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.

1 Introduction

For the last twenty years, the phenomenon of the concentration of measure has received much attention. The main interesting feature of concentration inequalities is that, unlike central limit theorems or large deviations inequalities, they are nonasymptotic. Among others, Pascal Massart, Michel Ledoux and Gabor Lugosi produced a serie of works that led to a large span of powerfull inequalities. Their results have found application in model selection (see [29] and [25]), statistical learning (see [9]), online learning (see [39]) or random graphs (see [12] and [11]). Most of the concentration inequalities are formulated for U-statistics of order m, which are defined as a sum of the form

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

where $X_1,\dots,X_n$ are random variables taking values in a measurable space (E,Σ) (with E Polish) and where $h_{i_1,\dots,i_m}$ are measurable functions of m variables $h_{i_1,\dots,i_m}: E^m \to \mathbb{R}$. The pioneering works considered independent random variables $(X_i)_{i\geq 1}$, an assumption that can be prohibitive for practical applications which often involve some dependence structure. To cope with this issue, some researchers left the independent setting by working with Markov chains or by adopting some mixing conditions, see e.g. [15, 23, 33, 1, 9]. The previous mentioned papers considered U-statistics of order $m=1$ and the non-asymptotic behaviour of tails of U-statistics of order $m\geq 2$ in a dependent framework remains so far very little touched. Recently, the two papers [14] and [35] made a first step to fill this gap. While [35] considers U-statistics of arbitrary order with smooth and symmetric kernels and works with mixing conditions, [14] is focused on U-statistics of order 2 for uniformly ergodic Markov chains and bounded kernels. Hence their results are complementary since they hold under different assumptions, but the intersection is non-empty (we refer to [14, Section 1] for a comparison between both concentration inequalities).

Our paper is in the line of work of [28] where concentration of measure is applied to tackle problems arising from model selection. In our paper, we shed light on the large number of potential theoretical breakthroughs allowed by a better understanding of the non-asymptotic tail behavior of U-statistics of order 2 in a dependent framework. Based on the concentration inequality from [14], we present new theoretical results in three different branches of statistics ranging from online learning to goodness-of-fit tests. In Section 1.1, we present the assumptions and the main result of [14] while in Section 1.2 we describe in details our three contributions.

Let us highlight that we work with the concentration result from [14] rather than the one from [35] since the result from [14] is valid for any initial distribution of the chain. We give further details at the end of Section 1.2.

1.1 Concentration inequality for U-statistics of uniformly ergodic Markov chains

In this section, we present the concentration result from [14] for U-statistics of uniformly ergodic Markov chains that will be useful in our proofs. 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 with a unique invariant distribution π. Denoting $\mathscr{B}(\mathbb{R})$ the Borel algebra on $\mathbb{R}$, we consider some measurable function $h:(E^2,\Sigma\otimes\Sigma)\to(\mathbb{R},\mathscr{B}(\mathbb{R}))$ and we are interested in the following U-statistic

$$U_{stat}(n) := \sum_{1\leq i\neq j\leq n} \Big(h(X_i,X_j) - \mathbb{E}_{(X,Y)\sim\pi\otimes\pi}[h(X,Y)] \Big).$$

We will work under the following set of assumptions.

Assumption 1 *The Markov chain $(X_i)_{i\geq 1}$ is ψ-irreducible for some maximal irreducibility measure ψ on Σ (see [31, 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 [31, Theorem 16.0.2]), namely there exist constants $0<\rho<1$ and $L>0$ such that

$$\|P^n(x,\cdot)-\pi\|_{TV} \leq L\rho^n, \qquad \forall n\geq 0,\ \pi\text{−a.e } x\in E, \tag{1}$$

where π is the unique invariant distribution of the chain $(X_i)_{i\geq 1}$ and for any measure ω on (E,Σ), we define $\|\omega\|_{TV} := \sup_{A\in\Sigma}|\omega(A)|$ is the total variation norm of ω. Assumption 1 also implies that the regeneration times associated to the split chain are exponentially integrable, meaning that their Orlicz norm with respect to the function $\psi_1(x)=\exp(x)-1$ are bounded by some constant $\tau>0$. We refer to [14, Section 2.3] for details.

Assumption 2 can be read as a reverse Doeblin's condition and is used in [14] as a decoupling tool. [14] gives several natural examples for which this condition holds.

Assumption 2 *There exist $\delta_M>0$ and some probability measure ν such that*

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

The last assumption introduces the notion of π-*canonical* kernel, which is the counterpart in the Markovian setting of the canonical (or degenerate) property of the independent framework.

Assumption 3 *Denoting π the invariant distribution of the Markov chain $(X_i)_{i\geq 1}$, we assume that $h:(E^2,\Sigma\otimes\Sigma)\to(\mathbb{R},\mathcal{B}(\mathbb{R}))$ is measurable, bounded and is π-canonical, namely*

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

This common expectation will be denoted $\mathbb{E}_\pi[h]$.

Let us mention that 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 [25]. 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 [12] or [11]). Note also that if the kernel h is not π-canonical, the U-statistic decomposes into a linear term and a π-canonical U-statistic. This is called the *Hoeffding decomposition* (see [20, p.176]) and takes the following form

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

where the kernel $\widetilde{h}$ is π-canonical with

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

We will use this method several times in our proofs (for example in (12)).

We are now ready to state the result from [14] that is one key theoretical tool to derive our three contributions presented in the next section.

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

$$\frac{2}{n(n-1)}U_{\text{stat}}(n)\leq \kappa\|h\|_\infty \log n\left\{\frac{u}{n}+\left[\frac{u}{n}\right]^2\right\}.$$

1.2 Our contributions

Theorem 1 is the cornerstone 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 Theorem 1 is one key element in our proofs, our contributions are not a direct consequence of it, 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 2)
  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)
  Motivated 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 4)
  Several works have already proposed goodness-of-fit tests for the density of the stationary distribution of a sequence of dependent random variables. In [26], a test based on an L^2-type distance between the nonparametrically estimated conditional density and its model-based parametric counterpart is proposed. In [3] a Kolmogorov-type test is considered. [8] 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.

As previously mentioned, we work with the concentration result from [14] rather than the one from [35] since the result from [35] only holds for stationary chains if the kernel h is π-canonical (see Assumption 3). Stationarity may be seen as a really strong assumption which would make our main results from Section 2 of little interest since MCMC methods are used when we are not able to directly sample from the distribution π. Regarding Sections 4 and 3, the control of the tail behaviour of U-statistics used in the proofs of our results needs to hold for any initial distribution of the chain.

1.3 Outline

We start by providing a convergence result for the estimation of spectra of integral operators with MCMC algorithms (see Section 2). We show that independent Hastings algorithms satisfy under mild conditions the assumptions of Section 1.1 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). 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 4). We give an explicit condition on the set of alternatives to ensure that the statistical test proposed reaches a prescribed power. The proofs related to the three applications are given in Section A, Section B and Sections C.1-C.3 respectively.

2 Estimation of spectra of signed integral operator with MCMC algorithms

2.1 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 1 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

$$\mathbf{H}f(x) := \int_E h(x,y)f(y)d\pi(y). \tag{2}$$

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 1 that will be useful to compare two spectra.

Definition 1 *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 2 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 2 is postponed to Section A.

Theorem 2 *We consider a Markov chain $(X_i)_{i\geq 1}$ on E satisfying Assumptions 1 and 2 described in Section 1.1 with invariant distribution π. We assume that*

- *the integral operator $\mathbf{H}$ is trace-class, i.e. $S := \sum_{r\geq 1} |\lambda_r| < \infty$.*
- *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 the constants from Theorem 1 and depend 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 2 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 $\mathbf{H}$.

2.2 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 1. It is for example well known that Metropolis random walks on $\mathbb{R}$ are not uniformly ergodic (see [31]). 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.

2.2.1 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 [30], Proposition 1 shows that under some conditions on the densities π and q, the independent Hastings algorithm satisfies the Assumptions 1 and 2.

Proposition 1 *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.

Proof of Proposition 1.

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 Equation (1)). Hence Assumption 1 is satisfied. Assumption 2 trivially holds since E is bounded and $\sup_{y\in E} q(y) < \infty$. ■

From Proposition 1 and Theorem 2, we deduce that one can use a MCMC approach to estimate the spectrum of a signed trace-class integral operator $\mathbf{H}$ as defined in (2) where E is a bounded subset of $\mathbb{R}^k$. More precisely, if the density π of (2) is known up to a factor and if there exists some probability density q with respect to λ_{Leb} satisfying the assumptions of Proposition 1, the Independent Hastings algorithm provides a Markov chain that satisfies Assumptions 1 and 2. Hence the non-asymptotic bound from Theorem 2 holds. We put this methodology into action in the new section by estimating the spectrum of some Mercer kernels on the d-dimensional sphere.

2.3 Estimation of the spectrum of Mercer kernels

In this example, we illustrate Theorem 2 by computing the eigenvalues of an integral operator naturally associated with a Mercer kernel using a MCMC algorithm. A function $h : E \times E \to \mathbb{R}$ is called a Mercer kernel if E is a compact metric space and $h : E \times E \to \mathbb{R}$ is a continuous symmetric and positive definite function. It is well known that if h is a Mercer kernel, then the integral operator L_h associated with h is a compact and bounded linear operator, self-adjoint and semi-definite positive. The spectral theorem implies that if h is a Mercer kernel, then there is a complete orthonormal system $(\varphi_1, \varphi_2, \dots)$ of eigenvectors of L_h. The eigenvalues $(\lambda_1, \lambda_2, \dots)$ are real and non-negative. The Mercer Theorem (see for instance see [7, Theorem 4.49]) shows that the eigen-structure of L_h can be used to get a representation of the Mercer kernel h as a sum of a convergent sequence of product functions for the uniform norm. In this context, Theorem 2 allows to derive the convergence rate in the δ_2 metric of the estimated spectrum towards the one of the integral operator $\mathbf{H}$ as presented in Proposition 2.

Proposition 2 *We keep the notations and the assumptions of Theorem 2. We assume further that there exists $s > 0$, a (Sobolev) regularity parameter, such that for some constant $C(s) > 0$,*

$$\forall R > 1, \quad \sum_{i>R} \lambda_i^2 \leq C(s) R^{-2s}.$$

Then it holds

$$\delta_2(\lambda(\mathbf{H}), \lambda(\mathbf{H}_n))^2 = \begin{cases} \mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{\log n}{n}}\right) & \text{if } s \geq 1 \\ \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n^{s/2}}\right) & \text{if } s \in (0,1) \end{cases} .$$

Proof.

Proposition 2 directly follows from Theorem 2 by choosing $t = \sqrt{\frac{\log n}{n}}$. □

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 $\forall x, y \in \mathbb{S}^{d-1}, \quad h(x,y) = \psi(x^\top y)$ where $\psi : [-1,1] \to \mathbb{R}$ is continuous. From the Funk-Hecke Theorem (see e.g [32, p.30]), we know that the eigenvalues of the Mercer kernel h are

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

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. 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 $f_{\mathscr{L}} : [-1,1] \to [0,1]$.

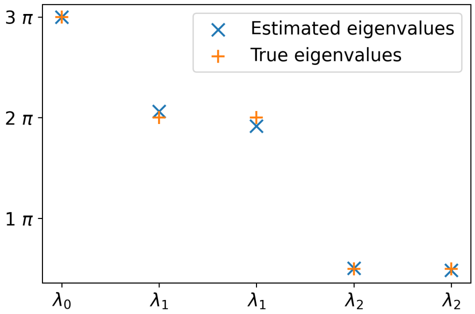
then X_i is defined by

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

By assuming that $\min_{r\in[-1,1]} f_{\mathscr{L}}(r) > 0$ and $\|f_{\mathscr{L}}\|_\infty < \infty$, Assumptions 1 and 2 hold and the invariant distribution of the chain $(X_i)_{i\geq 1}$ is the Haar measure on $\mathbb{S}^{d-1}$ (see for example [11]).

In Figure 1, we plot the non-zero eigenvalues using function $\psi : t \mapsto (1+t)^2$ and taking $f_{\mathscr{L}}$ proportional to $r \mapsto f_{(5,1)}(\frac{r+2}{4})$ where $f_{(5,1)}$ is the pdf of the Beta distribution with parameter $(5,1)$. We plot both the true eigenvalues and the ones computed using a MCMC approach.

Figure 1: Consider function $\psi : t \mapsto (1+t)^2$, $d = 2$ and $n = 1000$. The true eigenvalues can be computed using (3), 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 $\mathbf{H}_n = \frac{1}{n}\left((1-\delta_{i,j})\psi(X_i^\top X_j)\right)_{1\leq i,j\leq n}$ where the n points $X_1, X_2, \dots, X_n$ are sampled on the Euclidean sphere $\mathbb{S}^{d-1}$ using a Markovian dynamic.



3 Online Learning with Pairwise Loss Functions

3.1 Brief introduction to online learning and motivations

3.1.1 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 $(x_i, y_i)_{1\leq i\leq t}$) such that h_t has "small" risk $\mathscr{R}$ defined by

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

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.

3.1.2 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

$$\forall x, u \in E, \quad f(x) \geq f(u) \Longleftrightarrow (h(x,u) \geq 0 \text{ and } h(u,x) \leq 0).$$

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 [10] 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.*

3.1.3 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 [36] and [37] where online learning for least square regression and off-line support vector machines are investigated. In [41], the generalization performance of the empirical risk minimization algorithm is studied with uniformly ergodic Markov chain samples. In [40], 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 [10]). 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 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 Assumption 1 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 $\ell(h,X,U)$. 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 [39] 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 (4)) averaging the *paired empirical risks* M_t (see (5)) of hypotheses h_{t-b_n} when paired with X_t as follows

$$\mathscr{M}^n := \frac{1}{n-c_n}\sum_{t=c_n}^{n-1} M_t, \tag{4}$$

$$\text{and} \quad M_t := \frac{1}{t-b_n}\sum_{i=1}^{t-b_n} \ell(h_{t-b_n},X_t,X_i), \tag{5}$$

where

$$c_n = \lceil c \times n \rceil \text{ for some } c \in (0,1) \quad \text{ and } \quad b_n = \lfloor q\log(n) \rfloor, \tag{6}$$

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 Equation (1).

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 (5) 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 Equation (1)). 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 2).

Definition 2 *(See [38, 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 3 bounds the average risk of the sequence of hypotheses in terms of its empirical counterpart $\mathscr{M}^n$ and is proved in Section B.1.

Theorem 3 *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 $h_0, h_1, \dots, h_n \in \mathscr{H}$ 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 (6)). 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 $C(m,\tau)^{-1} = 7 \times 10^3 \times m^2\tau^2$. We refer to Assumption 1 and the following remark (or to [14, Section 2]) for the definitions of the constants m and τ that depend on the Markov chain $(X_i)_{i\geq 1}$.

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

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

where θ 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 [38, Chapter 5.1] for instance). In this case, Theorem 3 shows

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

3.3 Batch hypothesis selection

Theorem 3 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 4). 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 $X_1, \dots, X_n$, 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 $\{X_{t+1}, \dots, X_n\}$ 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 4, the hypothesis $\widehat{h}$ is chosen to minimize the following penalized empirical risk,

$$\widehat{h} = h_{\widehat{t}-b_n} \quad \text{and} \quad \widehat{t} \in \arg\min_{c_n \leq t \leq n-1}\left(\widehat{\mathscr{R}}(h_{t-b_n}, t+1) + c_\gamma(n-t)\right), \tag{7}$$

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 4 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 4 is postponed to Section B.2.

Theorem 4 *Assume that the Markov chain $(X_i)_{i\geq 1}$ is reversible and satisfies Assumptions 1 and 2. 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 3. 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 (6)). For all $\varepsilon > 0$ such that $\varepsilon \underset{n\to\infty}{=} o\left(n^{\xi}\right)$, if the hypothesis is selected via (7) 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

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

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

4.1 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 [4], [5]. In [34], 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 ([27],[21]). For example in [8], 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 [27],[16],[6],[17],[18], [19]). 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.

4.2 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 [19] 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 [19] 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 $\mathcal{M}$. Denoting by $t_m(u)$ the $(1-u)$ quantile of the law of $\widehat{T}_m$ under the hypothesis "$f=f_0$" and considering

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

we introduce the test statistic T_α defined by

$$T_\alpha = \sup_{m\in\mathcal{M}} (\widehat{T}_m - t_m(u_\alpha)). \tag{8}$$

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 $\mathcal{M}$, 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 $\{S_m, m\in\mathcal{M}\}$ 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
$$\widehat{\theta}_{(1,D)} = \frac{1}{n(n-1)}\sum_{k\in\mathbb{Z}}\sum_{i\neq j=1}^{n} I_{D,k}(X_i)I_{D,k}(X_j).$$

- 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 $\mathcal{D}_l\subset\mathbb{D}_l$ with $\cup_{l\in\{1,2,3\}}\mathcal{D}_l\neq\emptyset$ and $\mathcal{D}_3=\emptyset$ if the X_i's are not included in $[0,1]$. Let $\mathcal{M} = \{(l,D), l\in\{1,2,3\}, D\in\mathcal{D}_l\}$.

Theorem 5 describes classes of alternatives over which the corresponding test has a prescribed power. We work under the additional Assumption 4. We refer to Section C.1 for the proof of Theorem 5.

Assumption 4 *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 $\frac{1}{p} + \frac{1}{q} = 1$.

Theorem 5 *Let $X_1, \ldots, X_n$ a Markov chain on $\mathbb{R}$ satisfying the Assumptions 1, 2 and 4 with invariant measure π. 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 (8). 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 $\mathcal{M}$,

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

with

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

if f satisfies

$$\|f - f_0\|_2^2 > (1+\varepsilon) \inf_{m\in\mathcal{M}} \left\{\|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha) + V_m(\gamma)\right\}, \tag{9}$$

then

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

In order to make the condition (9) 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 3 *Given $\gamma \in]0,1[$ and a class of functions $\mathcal{B} \subset L_2(\mathbb{R})$, we define the uniform separation rate $\rho(\Phi_\alpha, \mathcal{B}, \gamma)$ of a level α test Φ_α of the null hypothesis "$f \in \mathcal{F}$" over the class $\mathcal{B}$ as the smallest number ρ such that the test guarantees a power at least equal to $(1-\gamma)$ for all alternatives $f \in \mathcal{B}$ at a distance ρ from $\mathcal{F}$. Stated otherwise, denoting by $d_2(f,\mathcal{F})$ the L_2-distance between f and $\mathcal{F}$ and by $\mathbb{P}_f$ the distribution of the observation $(X_1,\ldots,X_n)$,*

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

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

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

These sets of functions include some Hölder balls or Besov bodies with smoothness s, as highlighted in [19, Section 2.3]. Corollary 1 gives an upper bound for the uniform separation rate of our testing procedure over the classes $\mathcal{B}_s^{(l)}(P,M)$ and is proved in Section C.3.

Corollary 1 *Let T_α be the test statistic defined by (8). Assume that for $l \in \{1,2,3\}$, $\mathscr{D}_l$ is $\{2^J, 0 \le J \le \log_2\left(n/(\log(n)\log\log n)^2\right)\}$ 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*

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

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

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

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 [22] established in the case where the random variables $(X_i)_{i\ge1}$ 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 $([\log\log n]/n)^{\frac{2s}{4s+2}}$.

4.3 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. With Figure 2, we provide a visualization of the density of the invariant distribution of the Markov chain and of the density of the alternative that gives the smaller power on our experiments.

4.3.1 Example 1: AR(1) process

Let us consider some $\theta \in (0,1)$. Then, we define the AR(1) process $(X_i)_{i\ge1}$ starting from $X_1 = 0$ with for any $n \ge 1$,

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

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

where $(\xi_n)_n$ are i.i.d. random variables with distribution $\mathcal{N}(0,\tau^2)$ with $\tau > 0$. From Example 1 of [14, Section 2.6], we know that Assumptions 1 and 2 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

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

Table 1 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	$\|\mathbf{f}-\mathbf{f}_{\mu,\sigma^2}\|_2$
$(2,1.5)$	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 powers of the tests for Markov chains with size $n = 100$. We worked with $\tau = 1$, $\theta = 0.8$ and $\mathcal{M} = \{(1,i) : i \in \{1,\ldots,10\}\}$. 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.

4.3.2 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 1, we get that the above built Markov chain $(X_i)_{i\geq 1}$ satisfies Assumptions 1 and 2. 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 $\mathcal{M} = \{(1,i) : i \in \{1,\ldots,10\}\}$. For the χ^2 test, we work on the interval $[-3,3]$ that we split into 100 regular parts.

4.3.3 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 [14, Section 2.6], we know that Assumptions 1 and 2 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

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

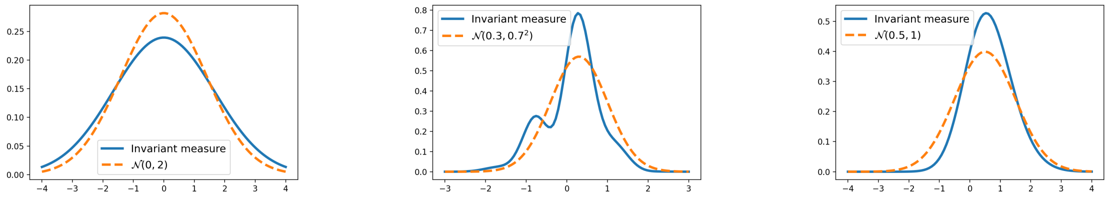
where Φ is the standard normal cumulative distribution function. We focus on the following alternatives

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

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 $\mathscr{M} = \{(1, i) \,:\, i \in \{1, \ldots, 10\}\}$. 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 measure 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.

Acknowledgements

This work was supported by a grant from Région Ile de France.

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

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

[5] 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.

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

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

[8] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2606–2615. JMLR.org, 2016.

[9] 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.

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

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

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

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

[14] Q. Duchemin, Y. de Castro, and C. Lacour. Concentration inequality for U-statistics of order two for uniformly ergodic markov chains. `https://arxiv.org/abs/2011.11435`, 2021.

[15] 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.

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

[17] 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.

[18] 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.

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

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

[21] 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.

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

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

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

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

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

[27] 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.

[28] P. Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

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

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

[31] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*, volume 92. Cambridge University Press, 01 1993.

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

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

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

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

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

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

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

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

[40] 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.

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

SUPPLEMENTARY MATERIAL: "Three rates of convergence or separation via U-statistics in a dependent framework"

Guidelines for the supplementary material.

- Sections A, B and C: Proofs
  Sections A, B and C provide respectively the proofs of our main results from Sections 2, 3 and 4.

- Section D: Technical Lemmas
  This section contains some Lemmas useful for our proofs.

A Proofs for Section 2

A.1 Deviation inequality for the spectrum of signed integral operators

As shown in Section A.2, Theorem 2 is a direct consequence of the concentration result provided by Theorem 6.

Theorem 6 *We keep notations of Section 2. Assume that $(X_n)_{n\geq 1}$ is a Markov chain on E satisfying Assumptions 1 and 2 described in Section 1.1 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, τ, L, m and ρ. β depends only on ρ.

Proof of Theorem 6. For any integer $R \geq 1$, we denote

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

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 $\lambda(\mathbf{H}^R) = (\lambda_1,\dots,\lambda_R)$, we have

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

A.1.1 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 $\lambda(X_{n,R}K_RX_{n,R}^\top) = \lambda(A_{n,R}K_RA_{n,R})$ which leads to

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

where the least inequality follows from Hoffman-Wielandt inequality. Using Equation (4.8) from [24, page 127], we get

$$\delta_2\left(\lambda(\mathbf{H}^R), \lambda(\widetilde{\mathbf{H}}_n^R)\right)^2 \leq 2\|K_RE_{R,n}\|_F^2 = 2\sum_{1\leq r,s\leq R} \lambda_s^2 \left(\frac{1}{n}\sum_{i=1}^n \varphi_r(X_i)\varphi_s(X_i) - \delta_{r,s}\right)^2. \tag{10}$$

Hence,

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

where the last inequality follows from Proposition 3 and where $K > 0$ is a universal constant.

A.1.2 Bounding $\delta_2(\lambda(\widetilde{\mathbf{H}}_n^R), \lambda(\mathbf{H}_n^R))^2$

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

A.1.3 Bounding $\delta_2(\lambda(\mathbf{H}_n^R), \lambda(\mathbf{H}_n))^2$

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

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.

$$\begin{aligned}&\frac{1}{n(n-1)}\left(\sum_{1\leq i,j\leq n,\, i\neq j} (h-h_R)(X_i,X_j)^2\right) &(11)\\ = \quad &\frac{1}{n(n-1)}\sum_{1\leq i,j\leq n,\, i\neq j} m_R(X_i,X_j) + \frac{2}{n}\sum_{i=1}^n s_R(X_i) + \mathbb{E}_{\pi\otimes\pi}[(h-h_R)^2(X,Y)]. &(12)\end{aligned}$$

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

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

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

$$\kappa \|h - h_R\|_\infty \log n \left\{ \frac{u}{n} \vee \left[\frac{u}{n} \right]^2 \right\} \leq t. \tag{13}$$

The condition (13) is equivalent to

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

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

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

where $c = \kappa \|h - h_R\|_\infty$. One can finally notice that for this choice of u, the condition $u \geq 1$ holds in particular for n large enough in order to have $n / \log n \geq \kappa \|h - h_R\|_\infty t^{-1}$.
We deduce from this analysis that for any $t > 0$, we have for n large enough to satisfy $n / \log n \geq \kappa \|h - h_R\|_\infty t^{-1}$,

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

Using Proposition 3, we get that for some universal constant $K > 0$,

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

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

$$\begin{aligned} & \mathbb{P}\left(\frac{1}{n^2} \left(\sum_{1 \leq i,j \leq n,\, i \neq j} (h - h_R)(X_i, X_j)^2 \right) - \mathbb{E}_{\pi \otimes \pi}\left[(h - h_R)^2 \right] \geq t \right) \\ \leq \quad & 16 \exp\left(-n \frac{t^2}{K m^2 \tau^2 \|h - h_R\|_\infty^2} \right) + \beta \log(n) \exp\left(-\frac{n}{4 \log n} \left\{ \left[\frac{t}{c} \right] \wedge \left[\frac{t}{c} \right]^{1/2} \right\} \right). \end{aligned}$$

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

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

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

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

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 6.

A.2 Proof of Theorem 2.

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 6 we get for any $t > 0$,

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

where $\Lambda := \sup_{r\geq 1}|\lambda_r| < \infty$. 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 $\mathscr{C} = \left(Km^2\tau^2S^2\Upsilon^4\right)^{-1}$.

B Proofs for Section 3

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

B.1 Proof of Theorem 3

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

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathscr{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]\big]-\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[M_t-\mathbb{E}_{t-b_n}[M_t]\big]\geq\varepsilon\right)\\
\leq\ &\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathscr{R}(h_{t-b_n})-\mathbb{E}_{t-b_n}[M_t]\big]\geq\varepsilon/2\right)+\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathbb{E}_{t-b_n}[M_t]-M_t\big]\geq\varepsilon/2\right). \qquad (14)
\end{aligned}$$

B.1.1 Step 1: Martingale difference

We first deal with the second term of (14). Note that we can write

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

Let us consider some $k\in\{1,\ldots,b_n\}$, 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 $|V_t^{(k)}|\leq 2/(n-c_n),\ \ t=1,\ldots,n$. Therefore by the Hoeffding-Azuma inequality, $\sum_t V_t^{(k)}$ can be bounded such that

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

We deduce that

$$\mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}\big[\mathbb{E}_{t-b_n}[M_t]-M_t\big]\geq\varepsilon/2\right)\leq b_n\exp\left(-\frac{(1-c)n\varepsilon^2}{8b_n^2}\right). \qquad (15)$$

B.1.2 Step 2: Symmetrization by a ghost sample

In this step we bound the first term in (14). Let us start by introducing a ghost sample $\{\xi_j\}_{1\leq j\leq n}$, 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 $X_j,\ j=1,\ldots,t-b_n$ 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 $\xi_j,\ j=1,\ldots,t-b_n$.
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

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

where we used Equation (1).
It remains to control

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

and we follow an approach similar to [39]. 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

$$\begin{aligned}
& \mathbb{P}\left(\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(h_{t-b_n}) \geq \varepsilon\right) \\
\leq & \mathbb{P}\left(\sup_{\widehat{h}_{c_n-b_n},\dots,\widehat{h}_{n-1-b_n}}\frac{1}{n-c_n}\sum_{t=c_n}^{n-1}L_t(\widehat{h}_{t-b_n}) \geq \varepsilon\right) \\
\leq & \sum_{t=c_n}^{n-1}\mathbb{P}\left(\sup_{\widehat{h}\in\mathscr{H}}L_t(\widehat{h}) \geq \varepsilon\right). \qquad (16)
\end{aligned}$$

To bound the right hand side of (16) we give first the following Lemma.

Lemma 1 *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 1.
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 $(X_{t-b_n},X_{t-b_n-1},\dots,X_1)$ conditionally on X_{t-b_n} is a Markov chain with invariant distribution π. Applying Proposition 3 (sse Section D) 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 1.

■

The following two Lemmas are key elements to prove Lemma 4. Their proofs are strictly analogous to the proofs of Lemmas 6, 7 and 8 from [39].

Lemma 2 *(See [39, 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 3 *Let $\mathscr{H}=S_1\cup\dots\cup S_l$ 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 4 *(See [39, 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 4 and (16), 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

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

B.1.3 Step 3: Conclusion of the proof.

From the previous inequality and (15), we get

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

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

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

we finally get for n large enough

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

B.2 Proof of Theorem 4

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

$$\widehat{\mathscr{R}}(h_{t-b_n},t+1)-\mathscr{R}(h_{t-b_n})=\binom{n-t}{2}^{-1}\sum_{k>i,i\geq t+1}^{n}\widetilde{\ell}(h_{t-b_n},X_i,X_k)+\frac{2}{n-t}\sum_{i=t+1}^{n}\ell(h_{t-b_n},X_i).\tag{17}$$

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 $h\in\mathscr{H}$, $\|\widetilde{\ell}(h,\cdot,\cdot)\|_\infty\leq 4$ (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 1, 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 $1-\gamma$,

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

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

$$\log n\left\{\frac{u}{n}\vee\left[\frac{u}{n}\right]^2\right\}\leq\frac{\log n}{n}u^2.$$

Using Proposition 3, 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 3 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.

$$\mathbb{P}\left(\left|\widehat{\mathscr{R}}(h_{t-b_n},t+1) - \mathscr{R}(h_{t-b_n})\right| \ge c_\gamma(n-t)\right) \le \frac{\gamma}{(n-c_n)(n-c_n+1)}. \tag{18}$$

Based on the selection procedure of the hypothesis $\widehat{h}$ defined in (7), the concentration result (18) allows us to show that $\mathscr{R}(\widehat{h})$ is close to $\min\limits_{c_n\le t\le n-1}\mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ with high probability. This is stated by Lemma 5 which is proved in Section B.3.

Lemma 5 *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 3. Then for any $\gamma \in (0,1)$, we have*

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

To conclude the proof, we need to show that $\min\limits_{c_n\le t\le n-1}\mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)$ 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 3, one can see that for each $t = c_n,\dots,n-1$,

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

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 $\min(a_1, a_2) \leq \min(b_1, b_2)$ then either $a_1 \leq b_1$ or $a_2 \leq b_2$, we can write

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

Using Lemma 5, we get

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

which gives in particular

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

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 $c > 0$,

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

where these inequalities hold for n large enough.

B.3 Proof of Lemma 5

Let

$$T^* := \arg\min_{c_n \le t < n-1} (\mathscr{R}(h_{t-b_n}) + 2c_\gamma(n-t)),$$

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

$$\widehat{T} := \arg\min_{c_n \le t < n-1} (\widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t)),$$

where $\widehat{h}$ coincides with $h_{\widehat{T}-b_n}$. Using this notation and since

$$\widehat{\mathscr{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*),$$

holds with certainty, we have

$$\begin{aligned}
& \mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + \mathscr{E}\right) \\
= \; & \mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{\widehat{T}-b_n} + c_\gamma(n-\widehat{T}) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right) \\
\le \; & \mathbb{P}\left(\bigcup_{c_n \le t \le n-1} \left\{\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right\}\right) \\
\le \; & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right),
\end{aligned}$$

where $\mathscr{E}$ is a positive-valued random variable to be specified. Now we remark that if

$$\widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*), \tag{19}$$

holds, then at least one of the following three conditions must hold

$$\begin{array}{lcl}
(i) & \widehat{\mathscr{R}}_{t-b_n} & \le \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t) \\
(ii) & \widehat{\mathscr{R}}^* & > \mathscr{R}(h^*) + c_\gamma(n-T^*) \\
(iii) & \mathscr{R}(h_{t-b_n}) - \mathscr{R}(h^*) & \le 2c_\gamma(n-T^*).
\end{array}$$

Stated otherwise, if (19) 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

$$\begin{aligned}
& \mathbb{P}\left(\mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}, \widehat{\mathscr{R}}_{t-b_n} + c_\gamma(n-t) \le \widehat{\mathscr{R}}^* + c_\gamma(n-T^*)\right) \\
\le \; & \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} \le \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + \mathbb{P}\left(\widehat{\mathscr{R}}^* > \mathscr{R}(h^*) + c_\gamma(n-T^*)\right) \\
& + \mathbb{P}\left(\mathscr{R}(h_{t-b_n}) - \mathscr{R}(h^*) \le 2c_\gamma(n-T^*)\,,\, \mathscr{R}(h_{t-b_n}) > \mathscr{R}(h^*) + \mathscr{E}\right).
\end{aligned}$$

By choosing $\mathscr{E} = 2c_\gamma(n-T^*)$, the last term in the previous inequality is zero and we can write

$$
\begin{aligned}
& \mathbb{P}\left(\mathscr{R}(\widehat{h}) > \mathscr{R}(h^*) + 2c_\gamma(n-T^*)\right) \\
\leq\ & \sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} \leq \mathscr{R}(h_{t-b_n}) - c_\gamma(n-t)\right) + (n-c_n)\mathbb{P}\left(\widehat{\mathscr{R}}^* > \mathscr{R}(h^*) + c_\gamma(n-T^*)\right) \\
\leq\ & (n-c_n)\frac{\gamma}{(n-c_n)(n-c_n+1)} + (n-c_n)\left\{\sum_{t=c_n}^{n-1} \mathbb{P}\left(\widehat{\mathscr{R}}_{t-b_n} > \mathscr{R}(h_{t-b_n}) + c_\gamma(n-t)\right)\right\} \text{ (Using (18))} \\
\leq\ & \frac{\gamma}{n-c_n+1} + (n-c_n)^2\frac{\gamma}{(n-c_n)(n-c_n+1)} \text{ (Using (18))} \\
\leq\ & \frac{\gamma}{n-c_n+1} + (n-c_n)\frac{\gamma}{n-c_n+1} = \gamma.
\end{aligned}
$$

C Proofs for Section 4

C.1 Proof of Theorem 5

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

$$
\begin{aligned}
\mathbb{P}_f(T_\alpha \leq 0) \leq \inf_{m \in \mathscr{M}} \mathbb{P}_f\Big(& U_n(H_m) + (P_n - P)(2\Pi_{S_m}(f) - 2f) + (P_n - P)(2f - 2f_0) + \|f - f_0\|_2^2 \\
& \leq \|f - \Pi_{S_m}(f)\|_2^2 + t_m(u_\alpha)\Big).
\end{aligned} \tag{20}
$$

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 $m \in \mathscr{M}$.

C.1.1 Control of $U_n(H_m)$

H_m is π-canonical and a direct application of Theorem 1 leads to the following Lemma (the proof of Lemma 6 is postponed to Section C.2).

Lemma 6 *Let us assume that the invariant distribution of the Markov chain $(X_i)_{i\geq 1}$ 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 13 and $Z_m=\frac{1}{n(n-1)}\sum_{i\neq j=1}^n H_m(X_i,X_j)$, 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 $(X_i)_{i\geq 1}$ while C also depends on φ) such that, for all $l\in\{1,2,3\}$, $D\in\mathbb{D}_l$ and $u\geq 1$, it holds with probability at least $1-\beta e^{-u}\log n$,*

$$|Z_{(l,D)}|\leq C\left(\|f\|_\infty+1\right)DR(n,u),$$

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 $m=(l,D)\in\mathscr{M}$,

$$\mathbb{P}_f\left(U_n(H_m)\leq -C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\right)\leq\gamma/3. \tag{21}$$

From (20) and (21) we get that

$$\begin{aligned}\mathbb{P}_f(T_\alpha\leq 0)\leq\frac{\gamma}{3}+\inf_{m\in\mathscr{M}}\mathbb{P}_f\Bigg((P_n-P)(2\Pi_{S_m}(f)-2f)+(P_n-P)(2f-2f_0)+\|f-f_0\|_2^2\\ \leq\|f-\Pi_{S_m}(f)\|_2^2+t_m(u_\alpha)+C\left(\|f\|_\infty+1\right)DR\left(n,\log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\Bigg).\end{aligned} \tag{22}$$

C.1.2 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 $\mathbb{D}_l$,

$$\left|2\Pi_{S_{(l,D)}}(f)(X_i)-2f(X_i)\right|\leq C'\|f\|_\infty.$$

Indeed,

- when $l=1$, for any $k\in\mathbb{Z}$,

$$\langle\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[},f\rangle=\int\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)f(x)dx\leq D^{-1/2}\|f\|_\infty.$$

Hence,

$$\begin{aligned}\sup_x|\Pi_{S_{(1,D)}}(f)(x)|&\leq\sup_x\sum_{k\in\mathbb{Z}}\left|\langle\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[},f\rangle\right|\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)\\ &\leq\quad D^{-1/2}\|f\|_\infty\sup_x\sum_{k\in\mathbb{Z}}\sqrt{D}\mathbb{1}_{[k/D,(k+1)/D[}(x)=\|f\|_\infty.\end{aligned}$$

- when $l=2$, $D=2^J$ for some $J\in\mathbb{N}$ and we have for any $k\in\mathbb{Z}$,

$$\langle\varphi_{J,k},f\rangle=\int 2^{J/2}\varphi(2^Jx-k)f(x)dx\leq\|f\|_\infty\int 2^{J/2}|\varphi(2^Jx)|dx\leq 2^{-J/2}\|f\|_\infty\|\varphi\|_1.$$

Hence,

$$\begin{aligned}\sup_x|\Pi_{S_{(2,D)}}(f)(x)|&\leq\sup_x\sum_{k\in\mathbb{Z}}\left|\langle\varphi_{J,k},f\rangle\right|\times|\varphi_{J,k}(x)|\\ &\leq\quad 2^{-J/2}\|f\|_\infty\|\varphi\|_1\sup_x\sum_{k\in\mathbb{Z}}|2^{J/2}\varphi(2^Jx-k)|\leq c\|f\|_\infty\|\varphi\|_1,\end{aligned}$$

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 [13, Page 269], that one can take C' such that for all D in $\mathbb{D}_3$,

$$|2\Pi_{S_{(3,D)}}(f)(X_i) - 2f(X_i)| \leq C'\|f\|_\infty \log(D+1).$$

Since

$$\mathbb{E}_{X\sim\pi}\left(2\Pi_{S_m}(f)(X) - 2f(X)\right)^2 \leq 4\|f\|_\infty \|\Pi_{S_m}(f) - f\|_2^2,$$

we can deduce using Proposition 4 (see Section D.2) that for all $m = (l, D) \in \mathcal{M}$,

$$\mathbb{P}_f\Bigg((P_n - P)(2\Pi_{S_m}(f) - 2f) < -\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty \log(D+1)}{n} \\ - 2\sqrt{\frac{2\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{n}}\|\Pi_{S_m}(f) - f\|_2\Bigg) \leq \frac{\gamma}{3}.$$

Considering some $\varepsilon \in]0,2[$, we use the inequality $\forall a, b \in \mathbb{R}$, $2ab \leq 4a^2/\varepsilon + \varepsilon b^2/4$ and we obtain that for any $m = (l, D) \in \mathcal{M}$,

$$\mathbb{P}_f\Bigg((P_n - P)(2\Pi_{S_m}(f) - 2f) + \frac{\varepsilon}{4}\|\Pi_{S_m}(f) - f\|_2^2 < -\frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty \log(D+1)}{n} \\ - \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}\Bigg) \leq \frac{\gamma}{3}. \tag{23}$$

The control of $(P_n - P)(2f - 2f_0)$ is computed in the same way and we get

$$\mathbb{P}_f\Bigg((P_n - P)(2f - 2f_0) + \frac{\varepsilon}{4}\|f - f_0\|_2^2 < -\frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty + \|f_0\|_\infty)}{n} \\ - \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n}\Bigg) \leq \frac{\gamma}{3}. \tag{24}$$

Finally, we deduce from (22), (23) and (24) that if there exists some $m = (l, D)$ in $\mathcal{M}$ such that

$$\left(1 - \frac{\varepsilon}{4}\right)\|f - f_0\|_2^2 > \left(1 + \frac{\varepsilon}{4}\right)\|f - \Pi_{S_m}(f)\|_2^2 + \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n} + \frac{4\log(3C_\chi/\gamma)qA_1(\|f\|_\infty + \|f_0\|_\infty)}{n} \\ + \frac{8\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\varepsilon n} + \frac{2C'\log(3C_\chi/\gamma)qA_1\|f\|_\infty \log(D+1)}{n} \\ + t_m(u_\alpha) + C\left(\|f\|_\infty + 1\right) DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right),$$

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

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

To conclude the proof of Theorem 5, 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

$$\begin{aligned}\|f-f_0\|_2^2 > (1+\eta)\Bigg\{ &\|f-\Pi_{S_m}(f)\|_2^2 + \frac{16\log(3C_\chi/\gamma)qA_2\|f\|_\infty}{\eta n} \\ &+ 4\big(\|f\|_\infty(C'\log(D+1)+1)+\|f_0\|_\infty\big)\frac{\log(3C_\chi/\gamma)qA_1}{n} \\ &+ t_m(u_\alpha) + C\left(\|f\|_\infty+1\right)DR\left(n, \log\left\{\frac{3\beta\log n}{\gamma}\right\}\right)\Bigg\},\end{aligned}$$

then

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

C.2 Proof of Lemma 6

Lemma 6 will follow from Theorem 1 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

$$\sum_{k\in\mathcal{L}_m} |\langle p_k, f\rangle|^2 \le \|f\|_2^2 = \int f(x)f(x)dx \le \|f\|_\infty, \tag{25}$$

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

$$\begin{aligned}|H_m(x,y)| &\le \sum_{k\in\mathcal{L}_m}|p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m}|p_k(x)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|p_k(y)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le \sum_{k\in\mathbb{Z}} D\mathbb{1}_{]k/D,(k+1)/D[}(x)\mathbb{1}_{]k/D,(k+1)/D[}(y) \\ &\qquad + 2\sup_z\sum_{k\in\mathbb{Z}}\sqrt{D}|\mathbb{1}_{]k/D,(k+1)/D[}(z)| \times |\langle\sqrt{D}\mathbb{1}_{]k/D,(k+1)/D[}, f\rangle| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le D + 2\|f\|_\infty + \|f\|_\infty,\end{aligned}$$

where in the last inequality we used (25).

- 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 $x, y \in \mathbb{R}$,

$$\begin{aligned}|H_m(x,y)| &\le \sum_{k\in\mathcal{L}_m}|p_k(x)p_k(y)| + \sum_{k\in\mathcal{L}_m}|p_k(x)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|p_k(y)\langle p_k,f\rangle| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le \sum_{k\in\mathbb{Z}}2^J\varphi(2^Jx-k)\varphi(2^Jy-k) + 2\sup_z\sum_{k\in\mathbb{Z}}2^{-J/2}\|f\|_\infty\|\varphi\|_1 2^{J/2}|\varphi(2^{J/2}z-k)| + \sum_{k\in\mathcal{L}_m}|\langle p_k,f\rangle|^2 \\ &\le c2^J + c'\|\varphi\|_1\|f\|_\infty + \|f\|_\infty \\ &= cD + c'\|\varphi\|_1\|f\|_\infty + \|f\|_\infty,\end{aligned}$$

for some constants $c, c' > 0$. In the last inequality we used (25) 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 $J \in \mathbb{Z}$, $2^J x - k$ falls into the support of φ.

- If $l = 3$ then we easily get for any $x, y \in [0,1]$,

$$\begin{aligned}|H_m(x,y)| &\leq \sum_{k\in\mathscr{L}_m} |p_k(x)p_k(y)| + \sum_{k\in\mathscr{L}_m} |p_k(x)\langle p_k, f\rangle| + \sum_{k\in\mathscr{L}_m} |p_k(y)\langle p_k, f\rangle| + \sum_{k\in\mathscr{L}_m} |\langle p_k, f\rangle|^2 \\ &\leq 2D + 4D\|f\|_\infty + \|f\|_\infty.\end{aligned}$$

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 6.

C.3 Proof of Corollary 1

Step 1: We start by providing an upper bound on $t_m(u_\alpha)$ with Lemma 7.

Lemma 7 *There exists a constant $C(\alpha) > 0$ such that for any $m = (l,D) \in \mathscr{M}$ it holds,*

$$t_m(u_\alpha) \leq W_m(\alpha),$$

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 7.

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 $|\mathscr{M}| \leq 3(1+\log_2 n)$. So, setting $\alpha_n = \alpha/(3(1+\log_2 n))$,

$$\begin{aligned}\mathbb{P}_{f_0}\Big(\sup_{m\in\mathscr{M}}(\widehat{T}_m - t_m(\alpha_n)) > 0\Big) &\leq \sum_{m\in\mathscr{M}} \mathbb{P}_{f_0}(\widehat{T}_m - t_m(\alpha_n) > 0) \\ &\leq \sum_{m\in\mathscr{M}} \alpha/(3(1+\log_2 n)) \\ &\leq \alpha.\end{aligned}$$

By definition of u_α, this implies that $\alpha_n \leq u_\alpha$ and for all $m \in \mathscr{M}$,

$$t_m(u_\alpha) \leq t_m(\alpha_n).$$

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 5 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 5.
From Lemma 6, there exist $C, \beta > 0$ such that for any $m = (l,D) \in \mathscr{M}$, it holds

$$\mathbb{P}_{f_0}\left(U_n(H_m) \leq C(\|f_0\|_\infty + 1)DR\left(n, \log\left\{\frac{2\beta\log n}{\alpha_n}\right\}\right)\right) \leq \alpha_n/2. \tag{26}$$

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 4 (see Section D.2) that for all $m = (l, D) \in \mathscr{M}$,

$$\mathbb{P}_{f_0}\Bigg((P_n - P)(2\Pi_{S_m}(f_0) - 2f_0) > \frac{2C' \log(2C_\chi/\alpha_n) q A_1 \|f_0\|_\infty \log(D+1)}{n} \\ + 2\sqrt{\frac{2\log(2C_\chi/\alpha_n) q A_2 \|f_0\|_\infty}{n}} \|\Pi_{S_m}(f_0) - f_0\|_2\Bigg) \leq \frac{\alpha_n}{2}.$$

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

$$\mathbb{P}_{f_0}\left(\widehat{T}_m > C(\|f_0\|_\infty + 1) DR\left(n, \log\left\{\frac{2\beta \log n}{\alpha_n}\right\}\right) + \frac{C''\|f_0\|_\infty \log(2C_\chi/\alpha_n)\log(n)}{n}\right) \leq \alpha_n.$$

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 $c(\alpha) > 0$,

$$t_m(\alpha_n) \leq c(\alpha) C(\|f_0\|_\infty + 1) DR(n, \log\log n) + c(\alpha)\frac{C''\|f_0\|_\infty \log\log n}{n}.$$

■

Step 2: Proof of Corollary 1.
Let us fix $\gamma \in]0,1[$ and $l \in \{1,2,3\}$. From Theorem 5 and Lemma 7, 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

$$\inf_{D \in \mathscr{D}_l} \Bigg\{P^2 D^{-2s} + C(\alpha)(\|f_0\|_\infty + 1)\left[DR(n, \log\log n) + \frac{\log\log n}{n}\right] + C_1\|f\|_\infty \frac{\log(3C_\chi/\gamma)}{\varepsilon n} \\ + C_2(\|f\|_\infty \log(D+1) + \|f_0\|_\infty)\frac{\log(3C_\chi/\gamma)}{n} + C_3(\|f\|_\infty + 1) DR\left(n, \log\left\{\frac{3\beta \log n}{\gamma}\right\}\right)\Bigg\}.$$

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

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

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

$$\left(\log(n)\frac{\log\log n}{n}\right)^{1/2} \leq P \leq \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}, \tag{27}$$

then the uniform separation rate of the test $\mathbb{1}_{T_\alpha>0}$ over $\mathscr{B}_s^{(l)}(P,M)$ satisfies

$$\rho\left(\mathbb{1}_{T_\alpha>0}, \mathscr{B}_s^{(l)}(P,M), \gamma\right) \leq C' P^{\frac{1}{2s+1}} \left(\frac{\log(n)\log\log n}{n}\right)^{\frac{s}{2s+1}}. \tag{28}$$

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 (27) (i.e. for $P \approx \frac{n^s}{(\log(n)\log\log n)^{2s+1/2}}$), the upper-bound in (28) still tends to zero as n goes to $+\infty$ for any possible values of the smoothness parameter s.

D Concentration Lemmas for Markov chains

D.1 Hoeffding inequality for uniformly ergodic Markov chains

Proposition 3 is an Hoeffding bound for uniformly ergodic Markov chains. A proof can be found in the Appendix of [14].

Proposition 3 *Let $(X_i)_{i\geq1}$ 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 $\|f\|_\infty \leq A$. Then it holds for any $t \geq 0$*

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

where $K(m,\tau) = 2Km^2\tau^2$ for some universal constant $K > 0$. We refer to Assumption 1 and the following remark (or to [14, Section 2]) for the definitions of the constants m and τ.

D.2 Bernstein's inequality for non-stationary Markov chains

Proposition 4 is an extension of the Bernstein type concentration inequality from [23] to non-stationary Markov chains. A proof can be found in the Appendix of [14].

Proposition 4 *Suppose that the sequence $(X_i)_{i\geq1}$ is a Markov chain satisfying Assumptions 1 and 4 with invariant distribution π and with an absolute spectral gap $1-\lambda > 0$. Let us consider some $n \in \mathbb{N}\backslash\{0\}$ 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. Stated otherwise, for any $u > 0$ it holds

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n f_i(X_i) > \frac{2quA_1c}{n} + \sqrt{\frac{2quA_2\sigma^2}{n}}\right) \leq \left\|\frac{d\chi}{d\pi}\right\|_{\pi,p} e^{-u}.$$