Empirical PAC-Bayes Bounds for Markov Chains

Vahe Karagulyan
ESSEC Business School, Cergy

Pierre Alquier
ESSEC Business School, Singapore

Abstract

The core of generalization theory was developed for independent observations. Some PAC and PAC-Bayes bounds are available for data that exhibit a temporal dependence. However, there are constants in these bounds that depend on properties of the data-generating process: mixing coefficients, mixing time, spectral gap... Such constants are unknown in practice. In this paper, we prove a new PAC-Bayes bound for Markov chains. This bound depends on a quantity called the *pseudo-spectral gap*, γ_{ps}. The main novelty is that we can provide an empirical bound on γ_{ps} when the state space is finite. Thus, we obtain the first fully empirical PAC-Bayes bound for Markov chains. This extends beyond the finite case, although this requires additional assumptions. On simulated experiments, the empirical version of the bound is essentially as tight as the one that depends on γ_{ps}.

1 INTRODUCTION

The PAC-Bayes theory is a flexible framework to derive generalization guarantees for learning algorithms. Since the seminal paper by McAllester (1998), it has found applications across a wide range of domains: from classification and regression to deep learning and variational inference. It has become a popular method in the machine learning community. One of the most successful applications of PAC-Bayes bounds was to obtain non-vacuous generalization bounds for deep neural networks Dziugaite and Roy (2017). We refer the reader to Alquier (2024); Hellström et al. (2025) for a recent overview on PAC-Bayes bounds.

The original paper by McAllester (1998) was written in the context of i.i.d. observations. Since then, PAC-Bayes bounds have been extended to handle more challenging settings, such as data with temporal or spatial dependencies. We mention Alquier and Wintenberger (2012); Banerjee et al. (2021); Haußmann et al. (2021); Eringis et al. (2021) among others. However, all these bounds involve constants characterizing the dependence in the data generating process: for example, the bound of Alquier and Wintenberger (2012) depends on weak-dependence coefficients, while the one of Banerjee et al. (2021) depends on α-mixing coefficients, etc. The strategy adopted by the authors was to assume *a priori* upper bounds on these constants. However, if this assumption is not correct, the PAC-Bayes bound is no longer valid. It would be much more satisfactory to estimate these constants and make the bounds fully empirical.

In this paper, we provide fully empirical PAC-Bayes bounds for a fundamental class of processes: Markov chains. We prove a PAC-Bayes bound that holds when the data is the trajectory of a Markov chain. This bound depends on a parameter γ_{ps} called the pseudo-spectral gap of the transition operator of the chain (Paulin, 2015). We then provide an empirical bound on γ_{ps} when the state space is finite, using tools from Wolfer and Kontorovich (2024). Putting everything together leads to fully-empirical PAC-Bayes bounds. Empirical bounds on γ_{ps} can be provided beyond the finite case, but this will in general require more assumptions on the chain. For example, we provide such a bound when the data is an autoregressive process.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

1.1 Related Works

PAC-Bayes bounds: while the original PAC-Bayes bounds were proven by McAllester (1998, 1999), our proof will follow the alternative approach by Catoni (2003). This approach is summarized in Alquier (2024). Important references include Catoni (2004, 2007); Hellström et al. (2025).

Empirical PAC-Bayes bounds: the bounds in the above references are empirical, that is, they depend on the data but not on the unknown data-generating process. There were many attempts to make these bounds tighter Seeger (2002); Maurer (2004); Kuzborskij et al. (2024) or to extend them in various directions Seldin et al. (2012b); Alquier and Guedj (2018); Rodriguez-Galvez et al. (2024). A natural way to make the bounds tighter relies on the application of Bernstein's inequality: this gives bounds that are in principle tighter, but that depend on the variance of the loss function under the data-generating process (Catoni, 2003). This quantity is unknown in practice. In order to make the PAC-Bayes-Bernstein bound practical, it is necessary to provide an empirical upper bound on the variance term. The first occurrence of such an "empirical-PAC-Bayes-Bernstein" is due to Seldin et al. (2012a) and these bounds were refined in later works (Tolstikhin and Seldin, 2013; Mhammedi et al., 2019; Wu et al., 2021; Wu and Seldin, 2022; Jang et al., 2023).

PAC-Bayes bounds for dependent observations: PAC-Bayes bounds for Markov chains were proven by Fard et al. (2011) in the setting of Markov decision processes, the bounds depend on ergodic coefficients that are unknown in practice. Also for Markov chains, Banerjee et al. (2021) proved bounds that depend explicitly on α-mixing coefficients (also unknown in practice). Some bounds in Alquier and Guedj (2018) which also depend on α-mixing coefficients hold for a more general class of stochastic processes. We also mention Haußmann et al. (2021) for continuous dynamical systems and Eringis et al. (2021) for linear time-invariant (LTI) systems.

There are other type of generalization bounds beyond PAC-Bayes, such as stability bounds and bounds based on the Rademacher complexity. Such results were also extended from the i.i.d. setting to time series under various assumptions: Yu (1994); Gamarnik (1999); Meir (2000); Steinwart et al. (2009); Mohri and Rostamizadeh (2010); Modha and Masry (2002); Steinwart et al. (2009); Shalizi and Kontorovich (2013); Kuznetsov and Mohri (2015, 2017); McDonald et al. (2017); Kuznetsov and Mohri (2020); Abeles et al. (2024). These bounds also depend on mixing or weak-dependence coefficients. In the case of Markov chains, some bounds also depend on the mixing time of the chain, on its spectral gap or on its pseudo-spectral gap Garnier et al. (2023); Alquier and Kengne (2025).

Estimation of the mixing coefficients: there were recently attempts to estimate the mixing coefficients (McDonald et al., 2015; Khaleghi and Lugosi, 2023). Although it is possible to estimate the mixing coefficients α and β, the results of Khaleghi and Lugosi (2023) do not provide confidence intervals on this estimation, and thus, cannot be used to derive empirical PAC-Bayes bounds. Some recent progress has been made in estimating the β-mixing coefficients of Markov chains (Grünewälder and Khaleghi, 2024; Wolfer and Alquier, 2024).

In the case of Markov chains, the estimation of the mixing time t_{mix}, the spectral gap and the pseudo-spectral gap was also studied thoroughly in the past years when the state space is finite (Hsu et al., 2015; Levin and Peres, 2016; Hsu et al., 2019; Wolfer and Kontorovich, 2019, 2024).

1.2 Contributions and Organization of the Paper

In this paper, we derive a PAC-Bayes bound for Markov chains that depends on a spectral quantity called the pseudo-spectral gap: γ_{ps} (the definition will be given below). The main tool in the proof is a Bernstein inequality for Markov chains due to Paulin (2015). Recently, Wolfer and Kontorovich (2024) derived estimators of γ_{ps} for finite-state Markov chains, together with confidence intervals. From this we derive empirical versions of our PAC-Bayes bound. We also provide an example in which such an empirical bound can be obtained while the state space is infinite.

In the end of this introduction, we introduce the notation used in the paper. We also remind important notions on Markov chains and define the pseudo-spectral gap γ_{ps}. In Section 2, we provide the (non-empirical) version of the PAC-Bayes bound for Markov chains. Then, in Section 3, we study various settings in which this bound can be made empirical. Finally, in Section 4, we exemplify our results in the problem of learning the best predictor in a finite set and provide numerical evaluations of the bound in this context. The proofs are gathered in the supplement, together with additional experiments and a

discussion on other possible approaches.

1.3 Problem Formulation

Let $\mathcal{U}$ denote the object space and $\mathcal{Y}$ the label space. Suppose, we are given object $\times$ label observations $(U_1, Y_1), (U_2, Y_2), \ldots, (U_n, Y_n)$. Usually, it is assumed that the pairs (U_t, Y_t) are i.i.d. from a distribution Q. In such a setting, $U_1, \ldots, U_n$ would be i.i.d. from the first marginal distribution Q_U of Q. Here, we want to allow some temporal dependence between the objects (U_t), so we will not assume that they are independent. Instead, we will assume they form a stationary Markov chain (the definition will be reminded below). As in the i.i.d. case, we will assume that the label Y_t depends *only* on U_t. In other words, there is a regular condition probability distribution $Q(\cdot, \cdot)$ such that the distribution of Y_t given $(U_1, \ldots, U_t)$ is given by $Q(U_t, \cdot)$. For short, let $\mathcal{S} = ((U_1, Y_1), \ldots, (U_n, Y_n))$ denote the sample.

We consider a parametrized set of predictors $\mathcal{F} = \{f_\theta : \mathcal{U} \to \mathcal{Y} \mid \theta \in \Theta\}$. To measure the prediction error, we use a loss function $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$, which is assumed to be bounded throughout the paper $\ell(\cdot, \cdot) \le c$. To measure the accuracy of the prediction, we will use the classical notion of risk:

$$R(\theta) = \mathbb{E}_{\mathcal{S}} \left[\frac{1}{n} \sum_{t=1}^{n} \ell\left(f_\theta(U_t), Y_t\right) \right]$$

(where $\mathbb{E}_{\mathcal{S}}$ denotes the expectation with respect to the sample, see Subsection 1.5 below), and the empirical risk

$$r(\theta) = \frac{1}{n} \sum_{t=1}^{n} \ell\left(f_\theta(U_t), Y_t\right).$$

1.4 Definitions and Reminders on Markov Chains

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, where Ω is a set equipped with a σ-algebra $\mathcal{A}$ and a probability measure $\mathbb{P}$. A $\mathcal{U}$-valued sequence of random variables $\{U_t\}_{t \ge 1}$ is said to be a Markov chain on $(\Omega, \mathcal{A}, \mathbb{P})$ if it satisfies the Markov property $\mathbb{P}(U_t \in A \mid U_{t-1}, U_{t-2}, \ldots, U_1) = \mathbb{P}(U_t \in A \mid U_{t-1})$ for any $t \ge 1$ and $A \in \mathcal{A}$. We will also assume the chain to be homogeneous: $\mathbb{P}(U_t \in A \mid U_{t-1} = u)$ does not depend on t. It is then common to introduce the notation $P(u, A) := \mathbb{P}(U_t \in A \mid U_{t-1} = u)$. P is called the transition kernel of the chain. Note that for every u, $P(u, \cdot)$ is a probability measure.

We say that a distribution π on $\mathcal{U}$ is a stationary distribution for the chain if

$$\int_{u \in \mathcal{U}} \pi(du) P(u, \mathrm{d}x) = \pi(\mathrm{d}x).$$

If $U_1 \sim \pi$, then any $U_t \sim \pi$ for $t \ge 1$ and the chain is said to be stationary. We will essentially work with stationary chains.

1.4.1 Asymptotic Behavior and Ergodicity

In this paper, we will work under the assumption that P is ergodic, which in particular implies it has a unique invariant distribution π. We will moreover assume it satisfies a spectral property defined below. This property will be enough to ensure that, for t large enough, the distribution of U_t is close enough to π regardless of the distribution of U_1. First, it will be helpful to remind a stronger, but more classical condition.

A chain is called uniformly ergodic with a rate $0 \le \rho < 1$, if there exists $C > 0$ such that

$$\sup_{u \in \mathcal{U}} \|P^k(u, \cdot) - \pi(\cdot)\|_{TV} \le C\rho^k.$$

For example, when $\mathcal{U}$ is a finite set, any chain that is irreducible has a unique invariant distribution π. If it is also aperiodic, then it is uniformly ergodic, see for example Douc et al. (2018).

Definition 1.1. *The mixing time t_{mix} of the chain is defined by $t_{mix} := t_{mix}(1/4)$ where*

$$t_{mix}(\varepsilon) := \inf \left\{ k : \sup_{u \in \mathcal{U}} \|P^k(u, \cdot) - \pi(\cdot)\|_{TV} \le \varepsilon \right\}.$$

The mixing time of a Markov chain measures how quickly the chain forgets its initial state and becomes close to its stationary distribution. If $t_{\mathrm{mix}} = +\infty$, then the chain is not uniformly ergodic.

1.4.2 Pseudo-Spectral Gap

Let $L^2(\pi)$ be the Hilbert space of complex valued measurable functions on Ω that are square integrable with respect to π. Consider $L^2(\pi)$ equipped with the inner product $\langle f, g \rangle_\pi = \int f g^* \, \mathrm{d}\pi$, and norm $\|f\|_{2,\pi} := \langle f, f \rangle_\pi^{1/2} = \left(\mathbb{E}_\pi \left[f^2\right]\right)^{1/2}$, then P defines a linear operator on $L^2(\pi)$ given by

$$(Pf)(u) := \mathbb{E}_{V \sim P(u, \cdot)}[f(V)] = \int_v P(u, \mathrm{d}v) f(v)$$

for any $f \in L^2(\pi)$. The operator P acts on measures to the left: for a probability measure ν, νP is also a probability measure given by $\nu P(A) :=$

$\int_u P(u, A)\nu(\mathrm{d}u)$ for every $A \in \mathcal{A}$.

As mentioned above, we assume P admits a unique invariant π. Such a kernel is said to be reversible if

$$\pi(\mathrm{d}u)\, P(u, \mathrm{d}v) = \pi(\mathrm{d}v)\, P(v, \mathrm{d}u).$$

Reversibility of the chain is equivalent to the linear operator P being self-adjoint on $L^2(\pi)$. When P is not self-adjoint (or the chain is not reversible), the chain is said to be non-reversible. We define the time reversal kernel P^* of P by

$$P^*(u, \mathrm{d}v) \;:=\; \frac{\pi(\mathrm{d}v)\, P(v, \mathrm{d}u)}{\pi(\mathrm{d}u)},$$

and the linear operator P^* is the adjoint of P in $L^2(\pi)$. In particular, if P is reversible, then $P^* = P$. Suppose π is the stationary distribution of the chain, and I is the identity operator. The spectrum $\mathrm{sp}(P)$ is defined as the set of all $\tau \in \mathbb{C}\backslash 0$ such that $(\tau I - P)^{-1}$ does not exist or is not a bounded linear operator in $L^2(\pi)$.

For a transition kernel P, $\tau \in \mathrm{sp}(P)$ satisfy $|\tau| \leq 1$. Moreover, $\tau = 1$ is necessarily an eigenvalue of P, and thus $1 \in \mathrm{sp}(P)$. When P is in addition reversible, $\mathrm{sp}(P) \subset \mathbb{R}$. In this case, if the multiplicity of the eigenvalue $\tau = 1$ is 1, the spectral gap of P is defined as

$$\gamma(P) := 1 - \sup\{\tau \in \mathrm{sp}(P) : \tau \neq 1\}.$$

When the multiplicity of the eigenvalue $\tau = 1$ is larger than 1, we define $\gamma(P) = 0$.

Another spectral characterization of a chain is the notion of a pseudo-spectral gap, proposed by Paulin (2015). It is more general, as it will allow us to consider chains that are not reversible.

Definition 1.2. *The pseudo-spectral gap is defined by*

$$\gamma_{ps}(P) \;:=\; \max_{k \geq 1} \left\{ \frac{\gamma\big((P^*)^k P^k\big)}{k} \right\}.$$

As the transition kernel of the observations will always be P, we can safely write γ_{ps} instead of $\gamma_{ps}(P)$. In this paper, our main assumption on the observations is that they form a Markov chain with positive pseudo-spectral gap: $\gamma_{ps} > 0$.

We mention that this condition is more general than the classical uniform ergodicity condition. Indeed, Proposition 3.4 of Paulin (2015) states that, if a Markov chain is uniformly ergodic, and thus has $t_{mix} < +\infty$, then

$$\gamma_{ps} \geq \frac{1}{2t_{mix}} > 0.$$

There are many examples of chains with $\gamma_{ps} > 0$ that are not uniformly ergodic, such as the AR(1) process in Subsection 3.2 below.

Thanks to a result of Davydov (1968) that we remind in Appendix C, for a uniformly ergodic chain, the φ-mixing coefficients $\varphi(k)$ decrease to 0 exponentially fast, while for a chain which is not uniformly ergodic, $\varphi(k)$ does not converge to 0. Thus, the condition $\gamma_{ps} > 0$ is also weaker than the condition $\varphi(k) \to 0$. We are not aware of a direct relation between γ_{ps} and the β-mixing coefficients, see the discussion by Wolfer and Alquier (2024).

1.5 Notations

Expectation and probability with respect to the sample will be denoted by $\mathbb{E}_{\mathcal{S}}$ and $\mathbb{P}_{\mathcal{S}}$ respectively. In PAC-Bayes bounds, we also consider expectation and probabilities with respect to some random parameter θ sampled from various probability distributions. So it is important to keep the distribution in the notation. When θ is sampled from some ρ, we will respectively write $\mathbb{E}_{\theta\sim\rho}$ and $\mathbb{P}_{\theta\sim\rho}$ for the expectation and probability with respect to θ. Rigorously,

$$\mathbb{E}_{\theta\sim\rho}[f(\theta)] = \int f(\theta)\rho(\mathrm{d}\theta)$$

(when this integral is well-defined). We let $\mathcal{P}(\Theta)$ denote the set of all probability distributions on Θ (equipped with a σ-field). Given $\nu_1, \nu_2 \in \mathcal{P}(\Theta)$ we let $KL(\nu_1\|\nu_2)$ denote the Kullback-Leibler divergence between ν_1 and ν_2. PAC-Bayes bounds involve a reference measure in $\mathcal{P}(\Theta)$ called the prior, we will let μ denote the prior throughout the paper. For an integer K, $[K] := \{1, 2, \ldots, K\}$.

2 PAC-BAYES BOUNDS FOR MARKOV CHAINS

We first state a (non-empirical) PAC-Bayes bound in this setting. The proof follows the same steps of the classical PAC-Bayes bound of Catoni (2003) in the i.i.d. setting. To handle Markov data, we use known concentration results for Markov chains from Paulin (2015).

Theorem 2.1. *Assume $\{U_t\}_{t=1}^n$ is a stationary Markov chain with pseudo-spectral gap $\gamma_{ps} > 0$. Then for any constants $0 < \lambda < \frac{n}{10}$, $\delta \in (0, 1)$, and prior*

$\mu \in \mathcal{P}(\Theta)$,

$$\mathbb{P}_{\mathcal{S}}\Bigg(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda} + \frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda\gamma_{ps}}\Bigg) \geq 1-\delta.$$

Observe the effect of γ_{ps} on the bound: the larger γ_{ps}, the tighter the bound is. However, when $\gamma_{ps} \to 0$, the bound explodes to infinity. Prediction is easier with a larger γ_{ps}.

In practice, if we observe data generated by an unknown Markov process, γ_{ps} is usually unknown. A naive approach is to assume a lower bound on $\gamma_{ps} \geq \gamma_0$, say $\gamma_{ps} > 0.1$ (this is similar to the *a priori* upper bounds assumed on mixing coefficients in the previous works mentioned in the introduction). This approach is problematic for two reasons: first, if $\gamma_{ps} = 0.05$, then our generalization bound is wrong. Moreover, if $\gamma_{ps} = 0.9$, our bound is correct, but it is also excessively pessimistic.

An alternative is to assume $\gamma_{ps} \geq \gamma_0 = 1/n^a$ with $a \in (0,1)$: such an assumption will always be satisfied for large enough sample size n. Under such an assumption, we can simply upper bound $1/\gamma_{ps}$ by n^a in the theorem. For the term $1 + 1/(n\gamma_{ps}) \leq 1 + 1/n^{1-a} \leq 2$, this is actually not a bad upper bound. The problem comes from

$$\frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda\gamma_{ps}}$$

which will change the order of magnitude of the bound. It would of course be far better to replace γ_{ps} by a consistent estimator $\widehat{\gamma}_{ps}$. If we can give an accurate upper bound on $1/\gamma_{ps}$ in terms of $1/\widehat{\gamma}_{ps}$, we will obtain an empirical PAC-Bayes bound. This is the object of the next section.

3 EMPIRICAL PAC-BAYES BOUNDS

Assuming we have an estimator of the pseudo-spectral gap γ_{ps}, we can state the following corollary of Theorem 2.1.

Corollary 3.1. *Under the conditions of Theorem 2.1, fix $a \in (0,1)$ and assume that n is large enough to ensure $n \geq 1/\gamma_{ps}^{1/a}$. Assume we have an estimator $\widehat{\gamma}_{ps}$ of γ_{ps} such that, for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} - 1\right| \leq \varepsilon\right) \geq 1 - \alpha(n, \gamma_{ps}, \varepsilon). \tag{1}$$

Then, we have

$$\mathbb{P}_{\mathcal{S}}\Bigg(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{2\lambda c^2\left(1+\frac{1}{n^{1-a}}\right)}{n-10\lambda} + \frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda\widehat{\gamma}_{ps}}\left(1+\varepsilon\right)\Bigg) \geq 1-\delta-\alpha(n,\gamma_{ps},\varepsilon).$$

A condition for the corollary to be actually useful is that $\alpha(n, \gamma_{ps}, \varepsilon)$ is a nonincreasing function of γ_{ps} that satisfies $\alpha(n, n^{-1/a}, \varepsilon) \to 0$ when $n \to \infty$. Indeed, in this case,

$$\alpha(n, \gamma_{ps}, \varepsilon) \leq \alpha(n, n^{-1/a}, \varepsilon) \xrightarrow[n\to\infty]{} 0. \tag{2}$$

3.1 First Example of Estimation of γ_{ps} : the Finite State Space Case

For an ergodic Markov chain on a finite state-space, say $\mathrm{card}(\mathcal{U}) = d$, Wolfer and Kontorovich (2019) provided an estimator for the pseudo-spectral gap given by the following formula:

$$\widehat{\gamma}_{ps,[K]} = \max_{k\in[K]}\left\{\frac{\gamma\left(\left(\hat{P}^{\dagger}\right)^k \hat{P}^k\right)}{k}\right\} \tag{3}$$

where $\hat{P} = \hat{P}\left(U_1, U_2, \cdots, U_n\right)$ is a natural empirical estimator for P and K a positive integer.

Rewriting their result in a way that matches Corollary 3.1, we obtain the following proposition (the proof is provided in the appendix).

Proposition 3.1. *Under the conditions of Theorem 2.1, assuming the chain (U_t) is ergodic and the state-space is finite, that is $\mathrm{card}(\mathcal{U}) = d$, for any $\varepsilon > 0$, the estimator $\widehat{\gamma}_{ps} := \widehat{\gamma}_{ps,[K]}$ given by 3 with $K = \lceil 2/\varepsilon \rceil$ we have*

$$\mathbb{P}_{\mathcal{S}}\left(\left|\frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} - 1\right| \geq \varepsilon\right) \leq \frac{C_{ps} d}{\varepsilon\gamma_{ps}\sqrt{\pi_*}} e^{-n\varepsilon^2\gamma_{ps}^2\pi_* \min\left\{\gamma_{ps}, \frac{1}{C(P)}\right\}} \tag{4}$$

where $C(P) = \|P\|_\pi \min\{d, \|P\|_\pi\}$, with $\|P\|_\pi = \max\{\pi(i)/\pi(j), i, j \in [d]^2\}$, and $\pi_* = \min\{\pi(i), i \in [d]\}$.

Observe that, as we assume that the chain is ergodic, $\pi_* = \min\{\pi(i), i \in [d]\} > 0$. However, π_* can be arbitrarily small, which leads to less confident estimation of γ_{ps}. Then, note that, by taking a large enough, (2) is satisfied.

3.2 Example of Estimation of γ_{ps} in the Infinite Case

In the finite case, we estimated the pseudo-spectral gap without strong assumptions on P. As argued by Wolfer and Kontorovich (2019), this is not possible for infinite Markov chains, even in the countable case. Intuitively, this can be understood from Proposition 3.1: when the state space is countably infinite, we have necessarily $\pi_* = 0$, and thus, the statement of the proposition becomes vacuous.

Obtaining empirical bounds is feasible, however, only by imposing strong restrictions on P. In this subsection, we illustrate this fact in the situation where the inputs are sampled from an autoregressive process on the real line. That is, we assume that $(U_t)_{t\geq 1}$ is a stationary process with

$$U_t = aU_{t-1} + \zeta_t \tag{5}$$

where $-1 < a < 1$ and the ζ_t are i.i.d. from $\mathcal{N}(0,1)$. In other words, $P(x,\cdot) = \mathcal{N}(ax, 1)$. Such a process is known to be ergodic, but non-uniformly ergodic: $t_{mix} = +\infty$. The following propositions show that its pseudo-spectral gap has a simple form and can be estimated with confidence.

Proposition 3.2. *Let $(U_t)_{t\geq 1}$ be a stationary AR(1) process, defined by (5), then its pseudo-spectral gap is given by*

$$\gamma_{ps} = 1 - a^2 = \frac{1}{\mathrm{Var}(U_1)}.$$

Proposition 3.3. *Let $(U_t)_{t\geq 1}$ be a stationary AR(1) process, defined by (5), then for the estimator $\widehat{\gamma}_{ps}$ given by*

$$\widehat{\gamma}_{ps} := \min\left\{\frac{1}{\frac{1}{n}\sum_{t=1}^n U_t^2}, 1\right\}, \tag{6}$$

it holds

$$\mathbb{P}_{\mathcal{S}}\left(\left|\frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} - 1\right| \leq \varepsilon\right) \geq 1 - \exp\left(\frac{9}{4} - \frac{n\varepsilon^2\gamma_{ps}^3}{2304}\right).$$

In other words,

$$\mathbb{P}_{\mathcal{S}}\left(\left|\frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} - 1\right| \leq \frac{24}{\gamma_{ps}^{3/2}}\sqrt{\frac{9 + 4\log\frac{1}{\delta}}{n}}\right) \geq 1 - \delta.$$

The proof relies on more general results on the estimation of variances and covariances of time series in Nakakita et al. (2025). Combining Proposition 3.3 and Corollary 3.1, we obtain the following result.

Corollary 3.2. *Let $(U_t)_{t\geq 1}$ be a stationary AR(1) process, defined by (5), and $\widehat{\gamma}_{ps}$ defined by (6), then we have*

$$\mathbb{P}_{\mathcal{S}}\Bigg(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{2\lambda c^2\left(1 + \frac{1}{\gamma_{ps}n}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log\frac{1}{\delta}}{\lambda\widehat{\gamma}_{ps}}\left(1 + \frac{24}{\gamma_{ps}^{3/2}}\sqrt{\frac{9 + 4\log\frac{1}{\delta}}{n}}\right)\Bigg) \geq 1 - 2\delta.$$

For example, when n is large enough to ensure $n \geq 1/\gamma_{ps}^4$ then

$$\mathbb{P}_{\mathcal{S}}\Bigg(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{2\lambda c^2\left(1 + \frac{1}{n^{3/4}}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log\frac{1}{\delta}}{\lambda\widehat{\gamma}_{ps}}\left(1 + \frac{24}{n^{1/8}}\sqrt{9 + 4\log\frac{1}{\delta}}\right)\Bigg) \geq 1 - 2\delta.$$

The tools developed in Nakakita et al. (2025) allow to tackle more general situations, such as multivariate U_t's and the case where the variance of ζ_t is unknown.

3.3 Optimization with Respect to λ and Oracle Bounds

We discuss briefly here how to tune the parameter λ in the PAC-Bayes bound and how we can obtain oracle bounds. The procedure is relatively standard, so we provide only the bare minimum, together with references for more details.

Given a finite grid $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$ of possible values for λ, we can perform a union bound on Theorem 2.1. We obtain:

$$\mathbb{P}_{\mathcal{S}}\Bigg(\forall \rho \in \mathcal{P}(\Theta), \exists \lambda \in \Lambda\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{2\lambda c^2\left(1 + \frac{1}{n\gamma_{ps}}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log\frac{L}{\delta}}{\lambda\gamma_{ps}}\Bigg) \geq 1 - \delta.$$

Definition 3.1. *We put*

$$\hat{\rho} = \operatorname*{argmin}_{\rho}\left[\mathbb{E}_{\theta\sim\rho}\big[r(\theta)\big] + B\left(\rho, \frac{(1+\varepsilon)}{\widehat{\gamma}_{ps}}\right)\right]$$

where, for any probability distribution $\nu \in \mathcal{P}(\Theta)$ and any real number $u > 0$,

$$B(\nu, u) := \min_{\lambda \in \Lambda} \left\{ \frac{2\lambda c^2 \left(1 + \frac{u}{n}\right)}{n - 10\lambda} + u \frac{KL(\nu||\mu) + \log \frac{L}{\delta}}{\lambda} \right\}.$$

The excess risk of $\hat{\rho}$ is upper bounded in the following theorem.

Theorem 3.1. *Under the conditions of Theorem 2.1,*

$$\mathbb{E}_{\theta\sim\hat{\rho}}\big[R(\theta)\big] \leq \inf_{\rho} \left\{ \mathbb{E}_{\theta\sim\hat{\rho}}\big[R(\theta)\big] + 2B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps} - \varepsilon}\right) \right\}$$

with probability at least $1 - 2\delta - \alpha(n, \gamma_{ps}, \varepsilon)$.

Remark 3.1. *A well-chosen grid will contain a λ of the order of*

$$c\sqrt{\frac{2n(1+\varepsilon)KL(\rho\|\mu)}{\gamma_{ps} - \varepsilon}}$$

which gives $B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right)$ of the order of

$$2c\sqrt{\frac{2(1+\varepsilon)KL(\rho\|\mu)}{n(\gamma_{ps} - \varepsilon)}},$$

we refer the reader to Section 2.1.4 in Alquier (2024) for more details on the construction of the grid.

Remark 3.2. *As in Corollary 3.1, we could exemplify the theorem in the case where $\gamma_{ps} \geq 1/n^a$. However, the constraint $\varepsilon < \gamma_{ps}$ will require one to take $\varepsilon < 1/n^a$. For example, with $\varepsilon = 1/(2n^a)$ we obtain:*

$$\mathbb{E}_{\theta\sim\hat{\rho}}\big[R(\theta)\big] \leq \inf_{\rho} \left\{ \mathbb{E}_{\theta\sim\hat{\rho}}\big[R(\theta)\big] + 2B\left(\rho, 2n^a + 1\right) \right\}$$

with probability at least $1 - 2\delta - \alpha(n, 1/n^a, 1/2n^a)$. This also requires to check that $\alpha(n, 1/n^a, 1/2n^a) \to 0$. For a small enough, this is straightforward in the two examples we developed above.

4 APPLICATION: FINITE SET OF PREDICTORS

In this section, we exemplify the approach of Sections 2 and 3 in the case where the set of predictors is finite: $\text{card}(\Theta) = M < +\infty$. This case was studied extensively in the machine learning literature. We believe it is also of pedagogical interest as the bound takes a simpler form in this situation, the reader will also observe that the optimization with respect to λ is more explicit. We will then assess the tightness of the bound on simulated data.

4.1 PAC-Bayes Bound with a Finite Θ

We consider the posterior ρ that minimizes the expected empirical risk. This ρ corresponds to the Dirac mass on the empirical risk minimizer $\hat{\theta}_{\text{ERM}} = \text{argmin}_\theta\, r(\theta)$, since

$$\inf_{\rho} \mathbb{E}_{\theta\sim\rho}\big[r(\theta)\big] = \inf_{\rho} \left[\sum_{\theta_i} \rho(\theta_i) r(\theta_i) \right] = r(\hat{\theta}_{\text{ERM}}). \tag{7}$$

The next theorem, which is proven using the PAC-Bayes bound of Theorem 2.1, provides a generalization bound on $\hat{\theta}_{\text{ERM}}$.

Theorem 4.1. *Fix $\varepsilon > 0$. Let $\{U_t\}_{t=1}^n$ be a stationary Markov chain with pseudo-spectral gap $\gamma_{ps} > 0$. Suppose card$(\Theta) = M < \infty$, and μ is the uniform prior on Θ, then for any $\delta \in (0,1)$, as soon as n is large enough to ensure*

$$n > \frac{50(1+\varepsilon)\log\frac{M}{\delta}}{\varepsilon^2 c^2 \gamma_{ps}\left(1 + \frac{1}{\gamma_{ps} n}\right)},$$

we have

$$\mathbb{P}_{\mathcal{S}}\Bigg(R(\hat{\theta}_{ERM}) \leq r(\hat{\theta}_{ERM}) + \sqrt{\frac{8(1+\varepsilon)c^2 \log\frac{M}{\delta}}{\gamma_{ps} n}\left(1 + \frac{1}{\gamma_{ps} n}\right)} \Bigg) \geq 1 - \delta.$$

Remark 4.1. *The main difference between this bound and similar results in the i.i.d. case (like Example 2.1 of Alquier (2024)) is that n is replaced by $n\gamma_{ps}$. Thus, we can think of $n\gamma_{ps}$ as an "effective sample size". When γ_{ps} is close to one, Markov observations are almost as informative as i.i.d. observations.*

Remark 4.2. *When the initial prior μ is not uniform, the bound in the statement still holds after replacing $\log\frac{M}{\delta}$ with $\log\frac{1}{\mu(\hat{\theta}_{\text{ERM}})\,\delta}$. All other terms in the bound remain unchanged.*

We now mimic what was done in Section 3 in the general case, to make this bound empirical, by using an estimator $\widehat{\gamma}_{ps}$.

Corollary 4.1. *Under the conditions of Theorem 4.1, fix $a \in (0,1)$ and assume that n is large enough to ensure $n \geq 1/\gamma_{ps}^{1/a}$. Assume we have an estimator $\widehat{\gamma}_{ps}$ of γ_{ps} such that*

$$\mathbb{P}_{\mathcal{S}}\left(\left|\frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} - 1\right| \leq \varepsilon\right) \geq 1 - \alpha(n, \gamma_{ps}, \varepsilon), \qquad (8)$$

then

$$\mathbb{P}_{\mathcal{S}}\Bigg(R(\hat{\theta}_{ERM}) \leq r(\hat{\theta}_{ERM}) + \sqrt{\frac{8c^2 \log\frac{M}{\delta}}{\widehat{\gamma}_{ps} n}(1+\varepsilon)^2\left(1 + \frac{1}{n^{1-a}}\right)}\Bigg) \geq 1 - \delta - \alpha(n, \gamma_{ps}, \varepsilon).$$

4.2 Experiments

In this section we assess the accuracy of the empirical bound of Corollary 4.1. That is, we sample trajectories of Markov chains[1] with various state spaces and transition kernels P. On the contrary to a real-life situation, where P would be unknown, we can compute γ_{ps} numerically here. This allows to compare the non-empirical bound of Theorem 4.1 to the empirical bound of Corollary 4.1, using the estimator $\widehat{\gamma}_{ps}$ of Wolfer and Kontorovich (2019), see (3).

Setting: we will consider $\mathcal{U} = [d]$ for various values of d: 4, 10, 20, 50 and 100. We work with a binary classification problem, where $\mathcal{Y} = \{0,1\}$, and our set of predictors are simply thresholds $f_\theta(u) = \mathbf{1}(u \geq \theta)$ for $\theta \in \Theta = [d]$, that is, $\mathrm{card}(\Theta) = d$ in Theorem (4.1) and Corollary 4.1. We consider various with various sample sizes: $n \in \{10, 100, 1000, 10000\}$.

In order to study the behavior of the bound under various values of γ_{ps}, we ran simulations for a wide range of transition kernels. We designed these kernels as follows: we first fixed a transition kernel P with $\gamma_{ps}(P) \simeq 0$, and a transition kernel Q with $\gamma_{ps}(Q) = 1$. This is easily obtained with a rank one $Q = 1^T \cdot \pi$ where $1^T = (1, \ldots, 1)$ and π is the invariant distribution of P; we refer the reader to the supplement for the exact definition of P. We then defined the kernels

$$R_t := tP + (1-t)Q \qquad (9)$$

for an interpolation parameter $t \in [0,1]$, and ran experiments for each $t \in \{k/20, 0 \leq k \leq 20\}$. In each case, we simply sample trajectories $(U_1, \ldots, U_n)$ from the kernel R_t, then, the

[1]The code is available online: `https://github.com/v-ahe/empirical-pac-bayes-markov/tree/main`

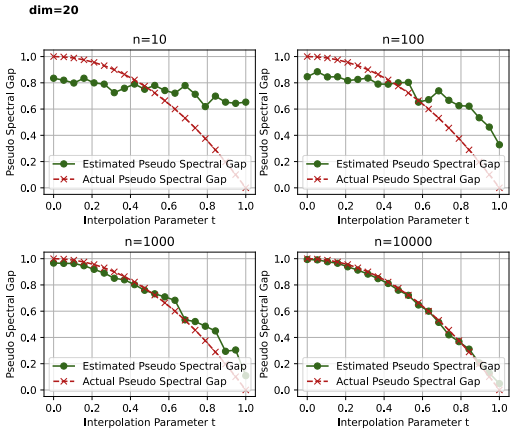

Figure 1: Estimation of $\gamma_{ps}(R_t)$ when $d = 20$. In red, the actual values of $\gamma_{ps}(R_t)$, as a function of our interpolation parameter t, see (9). In green, the value of the estimator $\widehat{\gamma}_{ps}$.

Tuning parameters: the parameter K in the definition of $\hat{\gamma}_{ps,[K]}$ was set as 20 in all our experiments[2]. The estimator $\hat{P}$ of the transition matrix of Wolfer and Kontorovich (2024) involves a smoothing parameter α that was set to 1. The empirical PAC-Bayes bound in Corollary 4.1 depends on two parameters a, set to 0.1, and ε, set to $0.1/n^{1/3}$.

Checking the estimator $\widehat{\gamma}_{ps}$: first, we ran sanity checks on the estimator $\widehat{\gamma}_{ps}$ of γ_{ps}. Figure 1 shows in red, the true $\gamma_{ps}(R_t)$ as a function of the interpolation parameter t, and in green, the estimator $\widehat{\gamma}_{ps}$, for four sample sizes n, and $d = 20$. This is essentially illustrative, as each point in these plots were obtained on one single experiment. We can still get some information from these plots: the estimation is poor for very small n and good for large n, as expected. More importantly, the estimator $\widehat{\gamma}_{ps}$ is far more accurate for small values of t, that is, for large values of γ_{ps}, as predicted by Proposition 3.1. We observed these findings are consistent when we ran more experiments.

[2]In most cases, we found that the maximum was reached for $k = 1$. Note however that, in the $d = 4$ setting, in a couple of cases the maximum was reached for $k = 4$. It can be understood intuitively why this is the case. Assume for simplicity that P is reversible, that is $P^\star = P$, and let λ denote the largest eigenvalue different from 1. Then $\gamma((P^\star)^k P^k) = \gamma(P^{2k}) = (1-\lambda^2)^k$ and thus $\gamma((P^\star)^k P^k)/k = (1-\lambda^2)^k/k$, which is obviously decreasing in k. Explicit examples where the maximum is reached for $k > 1$ are necessarily non-reversible. Such example can be found in Paulin (2016).

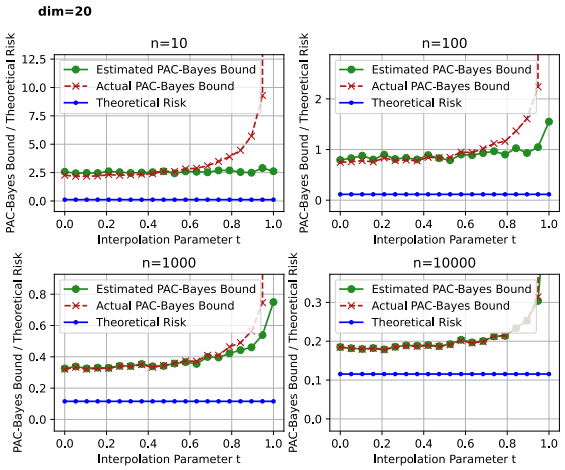

Figure 2: Value of the PAC-Bayes bounds evaluated on a single trajectory, for $d = 20$. In red, the non-empirical PAC-Bayes bound, as a function of t. In green, the empirical PAC-Bayes bound. In blue, the true value of the risk.

As mentioned earlier, the results in Figure 1 illustrative, as each point is based on a single trajectory. In order to confirm the good performances of the estimator $\widehat{\gamma}_{ps}$ of Wolfer and Kontorovich (2024) we sampled 100 trajectories in the case $n = 1000$, $d = 100$. We reported for each value of t the MSE of the estimator $\widehat{\gamma}_{ps}$ over all these 100 replications. The results are in Figure 3. They confirm that the estimator is accurate, and also that the estimation is more difficult for large values of t, that is, small values of γ_{ps}.

Checking the PAC-bayes bounds: for each value of n, d and t, we then sampled a trajectory, and computed both the non-empirical and the empirical PAC-Bayes bounds on each trajectory. Figure 2 shows, for the various n, as functions of t: in red, the value of the non-empirical PAC-Bayes bound over all replications, in green, the value of the empirical PAC-Bayes bound over all replications, and in blue, the actual value of the risk $R(\hat{\theta}_{\mathrm{ERM}})$. Here, we only show the results for $d = 20$, the plots for the other values of d are reported in the supplement. The take-home message is: for small sample size, the empirical bound is not a very good estimate of the non-empirical bound, but this is a regime where both bounds are vacuous anyway. For larger sample sizes, both bounds are non-vacuous and very similar. Finally, for very large t (very small γ_{ps}), the non-empirical bound becomes unreliable: it seems to confirm that very mild assumptions such as $n^a \geq 1/\gamma_{ps}$ are indeed unavoidable in practice.

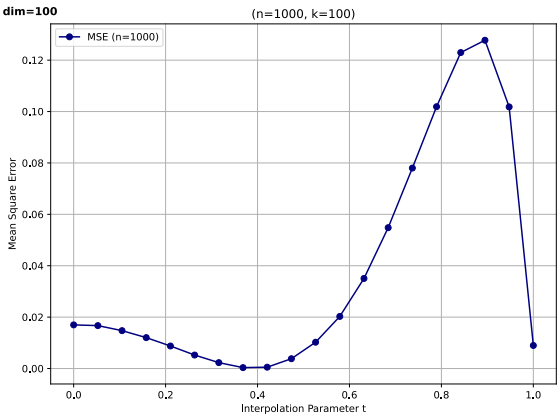

Figure 3: Mean Square Error of $k = 100$ estimations of $\widehat{\gamma}_{ps}$.

5 CONCLUSION

In this paper, we provided the first empirical PAC-Bayes bounds for Markov chains. The numerical results are encouraging: they show that, when the non-empirical bound is tight, the empirical bound is essentially as tight. This still relies on strong assumptions, and we believe that empirical bounds for time series beyond Markov chain are a very important research direction.

Content of the supplementary material: all the proofs are in Appendix A. More details on the experiments, together with the results for $d \in \{4, 10, 50, 100\}$ are provided in Appendix B. In Appendix C, we discuss how other PAC-Bayes bounds for time series from Alquier et al. (2013), which rely on φ-mixing coefficients rather than the pseudo-spectral gap, could also be made empirical in the context of Markov chains.

Acknowledgments

We thank Geoffrey Wolfer (Tokyo University of Agriculture and Technology) who introduced us to the pseudo-spectral gap and its estimation, and Mikołaj Kasprzak (ESSEC Business School) who spotted minor problems in the final version. We are grateful to the Anonymous Referees and the Area Chair for the very constructive feedback and the improvements they suggested. AI was used to track typos and formatting problems when preparing the camera-ready version, but it was not used at any other stage. All remaining mistakes are ours.

References

Baptiste Abeles, Eugenio Clerico, and Gergely Neu. Generalization bounds for mixing processes via delayed online-to-pac conversions. *arXiv preprint arXiv:2406.12600*, 2024.

Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends® in Machine Learning*, 17(2):174–303, 2024. ISSN 1935-8245.

Pierre Alquier and Benjamin Guedj. Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902, 2018.

Pierre Alquier and William Kengne. Minimax optimality of deep neural networks on dependent data via PAC-Bayes bounds. arXiv preprint arXiv:2410.21702, 2025.

Pierre Alquier and Olivier Wintenberger. Model selection for weakly dependent time series forecasting. *Bernoulli*, 18(3):883 – 913, 2012.

Pierre Alquier, Xiaoyin Li, and Olivier Wintenberger. Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling*, 1:65–93, 2013.

Imon Banerjee, Vinayak A. Rao, and Harsha Honnappa. PAC-Bayes bounds on variational tempered posteriors for Markov models. *Entropy*, 23 (3), 2021.

Stéphane Boucheron, Pascal Massart, and Gábor Lugosi. *Concentration inequalities*. Oxford University Press, 2006.

Olivier Catoni. A PAC-Bayesian approach to adaptive classification. preprint LPMA 840, 2003.

Olivier Catoni. *Statistical learning theory and stochastic optimization*. Saint-Flour Summer School on Probability Theory 2001 (Jean Picard ed.), Lecture Notes in Mathematics. Springer, 2004.

Olivier Catoni. *PAC-Bayesian supervised classification: The thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes – Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

Yu A Davydov. Convergence of distributions generated by stationary stochastic processes. *Theory of Probability & Its Applications*, 13(4):691–696, 1968.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*, volume 4. Springer, 2018.

Paul Doukhan. Mixing. In *Mixing: Properties and Examples*, pages 15–23. Springer, 1995.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *CoRR*, abs/1703.11008, 2017. URL `https://arxiv.org/abs/1703.11008`.

Deividas Eringis, John-Josef Leth, Zheng-Hua Tan, Rafal Wisniewski, Alireza Fakhrizadeh Esfahani, and Mihaly Petreczky. PAC-Bayesian theory for stochastic LTI systems. In *2021 60th IEEE Conference on Decision and Control (CDC)*, IEEE Conference on Decision and Control. Proceedings, pages 6626–6633. IEEE (Institute of Electrical and Electronics Engineers), 2021.

Mahdi Milani Fard, Joelle Pineau, and Csaba Szepesvári. PAC-Bayesian policy evaluation for reinforcement learning. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 195–202, 2011.

David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 308–317, 1999.

Remy Garnier, Raphaël Langhendries, and Joseph Rynkiewicz. Hold-out estimates of prediction models for Markov processes. *Statistics*, 57(2):458–481, 2023.

Steffen Grünewälder and Azadeh Khaleghi. Estimating the mixing coefficients of geometrically ergodic Markov processes. *arXiv:2402.07296*, 2024.

Manuel Haußmann, Sebastian Gerwinn, Andreas Look, Barbara Rakitsch, and Melih Kandemir. Learning partially known stochastic dynamics with empirical PAC-Bayes. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 478–486. PMLR, 13–15 Apr 2021.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj, Maxim Raginsky, et al. Generalization bounds: Perspectives from information theory and PAC-Bayes. *Foundations and Trends® in Machine Learning*, 18(1):1–223, 2025.

Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. *Advances in neural information processing systems*, 28, 2015.

Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible Markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439 – 2480, 2019.

Kyoungseok Jang, Kwang-Sung Jun, Ilja Kuzborskij, and Francesco Orabona. Tighter PAC-Bayes bounds through coin-betting. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2240–2264. PMLR, 12–15 Jul 2023.

Azadeh Khaleghi and Gabor Lugosi. Inferring the mixing properties of a stationary ergodic process from a single sample-path. *IEEE Transactions on Information Theory*, 69(6):4014–4026, 2023.

Walter Frederick Kibble. An extension of a theorem of Mehler's on Hermite polynomials. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 41, pages 12–15. Cambridge University Press, 1945.

Ilja Kuzborskij, Kwang-Sung Jun, Yulian Wu, Kyoungseok Jang, and Francesco Orabona. Better-than-KL PAC-Bayes bounds. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3325–3352. PMLR, 2024.

Vitaly Kuznetsov and Mehryar Mohri. Learning theory and algorithms for forecasting non-stationary time series. *Advances in neural information processing systems*, 28, 2015.

Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

Vitaly Kuznetsov and Mehryar Mohri. Discrepancy-based theory and algorithms for forecasting non-stationary time series. *Annals of Mathematics and Artificial Intelligence*, 88(4):367–399, 2020.

David A Levin and Yuval Peres. Estimating the spectral gap of a reversible Markov chain from a short trajectory. *arXiv preprint arXiv:1612.05330*, 2016.

Andreas Maurer. A note on the PAC-Bayesian theorem. *arXiv preprint cs/0411099*, 2004.

David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 230–234, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, COLT '99, page 164–170, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131674.

Daniel J. McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Estimating β-mixing coefficients via histograms. *Electronic Journal of Statistics*, 9 (2):2855 – 2883, 2015.

Daniel J McDonald, Cosma Rohilla Shalizi, and Mark Schervish. Nonparametric risk bounds for time-series forecasting. *Journal of Machine Learning Research*, 18(32):1–40, 2017.

Ron Meir. Nonparametric time series prediction through adaptive model selection. *Machine learning*, 39(1):5–34, 2000.

Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Dharmendra S Modha and Elias Masry. Memory-universal prediction of stationary random processes. *IEEE transactions on information theory*, 44(1):117–133, 2002.

Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary ϕ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11 (26):789–814, 2010.

Shogo Nakakita, Pierre Alquier, and Masaaki Imaizumi. Corrigendum to "Dimension-free bounds for sums of dependent matrices and operators with heavy-tailed distributions". *Electronic Journal of Statistics*, 19(2):3273–3291, 2025.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015.

Daniel Paulin. Mixing and concentration by Ricci curvature. *Journal of Functional Analysis*, 270(5): 1623–1662, 2016.

Emmanuel Rio. Inégalités de Hoeffding pour les fonctions lipschitziennes de suites dépendantes. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 330(10):905–908, 2000.

Borja Rodriguez-Galvez, Ragnar Thobaben, and Mikael Skoglund. More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity. *Journal of Machine Learning Research*, 25(110):1–43, 2024.

Matthias Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Jour-*

nal of machine learning research, 3(Oct):233–269, 2002.

Yevgeny Seldin, Nicolò Cesa-Bianchi, Peter Auer, François Laviolette, and John Shawe-Taylor. PAC-Bayes-Bernstein inequality for martingales and its application to multiarmed bandits. In Dorota Glowacka, Louis Dorard, and John Shawe-Taylor, editors, *Proceedings of the Workshop on On-line Trading of Exploration and Exploitation 2*, volume 26 of *Proceedings of Machine Learning Research*, pages 98–111, Bellevue, Washington, USA, 02 Jul 2012a. PMLR.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012b.

Cosma Shalizi and Aryeh Kontorovich. Predictive PAC learning and process decompositions. *Advances in neural information processing systems*, 26, 2013.

Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

Ilya O. Tolstikhin and Yevgeny Seldin. PAC-Bayes-empirical-Bernstein inequality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Geoffrey Wolfer and Pierre Alquier. Optimistic estimation of convergence in Markov chains with the average-mixing time. *arXiv preprint arXiv:2402.10506*, 2024.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic Markov chains. In *Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.

Geoffrey Wolfer and Aryeh Kontorovich. Improved estimation of relaxation time in nonreversible Markov chains. *Annals of Applied Probability*, 34 (1A):249–276, 2024.

Yi-Shan Wu and Yevgeny Seldin. Split-kl and PAC-Bayes-split-kl inequalities for ternary random variables. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 11369–11381. Curran Associates, Inc., 2022.

Yi-Shan Wu, Andres Masegosa, Stephan Lorenzen, Christian Igel, and Yevgeny Seldin. Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12625–12636. Curran Associates, Inc., 2021.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. YES
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. NOT APPLICABLE
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. YES

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. YES
   (b) Complete proofs of all theoretical results. YES
   (c) Clear explanations of any assumptions. YES

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). YES
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). YES
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). YES
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). NOT APPLICABLE

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. NOT APPLICABLE
   (b) The license information of the assets, if applicable. NOT APPLICABLE
   (c) New assets either in the supplemental material or as a URL, if applicable. NOT APPLICABLE
   (d) Information about consent from data providers/curators. NOT APPLICABLE
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. NOT APPLICABLE

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots.NOT APPLICABLE
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. NOT APPLICABLE
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. NOT APPLICABLE

Supplementary Material

A PROOFS

A.1 A Preliminary Remark on the Observations

In the introduction of the paper, we define the distribution of the pairs (U_t, Y_t) by saying that the (U_t) are sampled from a Markov chain with transition kernel P, and that the distribution of Y_t given $(U_1, Y_1), \dots, (U_{t-1}, Y_{t-1}), U_t$ is given by $Q(U_t, \cdot)$. Observe that this is simply equivalent to stating that $[(U_t, Y_t)]_{t\geq 1}$ is a Markov chain on the space $\mathcal{U} \times \mathcal{Y}$ with transition kernel $\bar{P}$ given by

$$\bar{P}((u,y), \mathrm{d}(u',y')) = P(u, \mathrm{d}u')Q(u', \mathrm{d}y'). \tag{10}$$

In the proof of Theorem 2.1, we will actually use this fact. In particular, the pseudo-spectral gap of $\bar{P}$ will appear in the proof, while the assumption on Theorem 2.1 is on the pseudo-spectral gap γ_{ps} of P: it turns out that this will not lead to any complication, as these two quantities are equal. We start by proving this fact.

Lemma A.1. *Assume P is the transition kernel of an ergodic chain. Let $\bar{P}$ be given by* (10)*. The pseudo-spectral gap of $\bar{P}$ is equal to the pseudo-spectral gap γ_{ps} of P.*

This might be well known among Markov chains specialists, but we did not find it in any classical textbook. Thus, we preferred to provide a complete proof.

Proof. The proof goes in four steps. In the first step, we will write the (cumbersome but) explicit formulas for $(\bar{P}^*)^k(\bar{P})^k$. In a second step, we prove that any τ in the spectrum of $(P^*)^k(P^k)$ is also in the spectrum of $(\bar{P}^*)^k(\bar{P})^k$. In the third step, we prove that any $\tau < 1$ in the spectrum of $(\bar{P}^*)^k(\bar{P}^k)$ also belongs to the spectrum of $(P^*)^k(P^k)$. In the fourth and last step, we study the case of $\tau = 1$, which is necessarily an eigenvalue for both operators: it appears that its multiplicity is 1 for $(\bar{P}^*)^k(\bar{P})^k$ if and only if its multiplicity is also 1 for $(P^*)^k(P^k)$.

Step 1: first, observe that as P has a unique stationary distribution π, $\bar{P}$ has a unique stationary distribution $\bar{\pi}(\mathrm{d}(u,y)) = \pi(\mathrm{d}u)Q(u,\mathrm{d}y)$. Thus, we have:

$$\begin{aligned}
\bar{P}^*((u,y), \mathrm{d}(u',y')) &= \frac{\bar{P}((u',y'), \mathrm{d}(u,y))\bar{\pi}(\mathrm{d}(u',y'))}{\bar{\pi}(\mathrm{d}(u,y))} \\
&= \frac{P(u', \mathrm{d}u)Q(u, \mathrm{d}y)\pi(\mathrm{d}u')Q(u', \mathrm{d}y')}{\pi(\mathrm{d}u)Q(u, \mathrm{d}y)} \\
&= \frac{P(u', \mathrm{d}u)\pi(\mathrm{d}u')Q(u', \mathrm{d}y')}{\pi(\mathrm{d}u)} \\
&= P^*(u, \mathrm{d}u')Q(u', \mathrm{d}y').
\end{aligned}$$

Then, we use recursion to prove that $(\bar{P}^*)^k((u,y), \mathrm{d}(u',y')) = (P^*)^k(u, \mathrm{d}u')Q(u', \mathrm{d}y')$ and

$\bar{P}^k((u,y),\mathrm{d}(u',y')) = P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y')$. The proof is similar for both. For example,

$$\begin{aligned}
\bar{P}^k((u,y),\mathrm{d}(u',y')) &= \int_{(v,x)} \bar{P}^{k-1}((u,y),\mathrm{d}(v,x))\bar{P}^1((v,x),\mathrm{d}(v,x)) \\
&= \int_v \int_x P^{k-1}(u,\mathrm{d}v)Q(v,\mathrm{d}x)P(v,\mathrm{d}u')Q(u',\mathrm{d}y') \\
&= \int_v P^{k-1}(u,\mathrm{d}v)P(v,\mathrm{d}u')Q(u',\mathrm{d}y') \underbrace{\int_x Q(v,\mathrm{d}x)}_{=1} \\
&= P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y').
\end{aligned}$$

We are now ready to work with $(\bar{P}^*)^k(\bar{P})^k$:

$$\begin{aligned}
(\bar{P}^*)^k(\bar{P})^k((u,y),\mathrm{d}(u',y')) &= \int_{(v,x)} (P^*)^k(u,\mathrm{d}v)Q(v,\mathrm{d}x)P^k(v,\mathrm{d}u')Q(u',\mathrm{d}y') \\
&= \int_v (P^*)^k(u,\mathrm{d}v)P^k(v,\mathrm{d}u')Q(u',\mathrm{d}y') \int_x Q(v,\mathrm{d}x) \\
&= (P^*)^k P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y').
\end{aligned}$$

The conclusion of the first step is thus:

$$(\bar{P}^*)^k(\bar{P})^k((u,y),\mathrm{d}(u',y')) = (P^*)^k P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y'). \tag{11}$$

Step 2: let $\tau \in \mathrm{sp}((P^*)^k P^k)$, that is, there are sequences of functions $(F_n(u))$ and $(G_n(u))$ with

$$[(P^*)^k P^k - \tau I]F_n = G_n,$$

$$\|F_n\|_\pi^2 = \int_u F_n(u)^2 \pi(\mathrm{d}u) = 1 \text{ and}$$

$$\|G_n\|_\pi^2 \xrightarrow[n\to\infty]{} 0.$$

Then, put $f_n(u,y) = F_n(u)$ and $g_n(u,y) = G_n(u)$, and observe that

$$\begin{aligned}
[(\bar{P}^*)^k\bar{P}^k - \tau I]f_n(u,y) &= \int_{(u',y')} (\bar{P}^*)^k(\bar{P})^k((u,y),\mathrm{d}(u',y'))f_n(u',y') - \tau f_n(u,y) \\
&= \int_{(u',y')} (P^*)^k P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y')F_n(u') - \tau F_n(u)
\end{aligned}$$

where we used both (11) and the definition of f_n: $f_n(u,y) = F_n(u)$. Thus,

$$\begin{aligned}
[(\bar{P}^*)^k\bar{P}^k - \tau I]f_n &= \int_{u'}\int_{y'} (P^*)^k P^k(u,\mathrm{d}u')Q(u',\mathrm{d}y')F_n(u') - \tau F_n(u) \\
&= \int_{u'} (P^*)^k P^k(u,\mathrm{d}u')F_n(u') \int_{y'} Q(u',\mathrm{d}y') - \tau F_n(u) \\
&= [(P^*)^k P^k - \tau I]F_n(u) \\
&= G_n(u) \\
&= g_n(u,y).
\end{aligned}$$

Moreover,

$$\begin{aligned}\|f_n\|_{\bar{\pi}}^2 &= \int_{(u,y)} f_n(u,y)^2 \bar{\pi}(\mathrm{d}(u,y)) \\ &= \int_u \int_y F_n(u)^2 \pi(\mathrm{d}u) Q(u,\mathrm{d}y) \\ &= \int_u F_n(u)^2 \pi(\mathrm{d}u) \int_y Q(u,\mathrm{d}y) \\ &= \|F_n\|_{\pi}^2 \\ &= 1\end{aligned}$$

and with exactly the same argument,

$$\|g_n\|_{\bar{\pi}}^2 = \|G_n\|_{\pi}^2 \xrightarrow[n\to\infty]{} 0.$$

This proves that $\tau \in \mathrm{sp}((\bar{P}^*)^k \bar{P}^k)$.

Step 3: let us now assume $\tau \in \mathrm{sp}((\bar{P}^*)^k \bar{P}^k)$ and $\tau \neq 0$, that is, there are sequences $(f_n(u,y))$ and $(g_n(u,y))$ with

$$\begin{aligned}[(\bar{P}^*)^k \bar{P}^k - \tau I] f_n &= g_n, \\ \|f_n\|_{\bar{\pi}}^2 &= 1 \text{ and} \\ \|g_n\|_{\bar{\pi}}^2 &\xrightarrow[n\to\infty]{} 0.\end{aligned}$$

As

$$(\bar{P}^*)^k \bar{P}^k f_n(u,y) = \int_{(u',y')} (\bar{P}^*)^k \bar{P}^k ((u,y), \mathrm{d}(u',y')) f_n(u',y') = \int_{(u',y')} (P^*)^k P^k(u,\mathrm{d}u') Q(u',\mathrm{d}y') f_n(u',y')$$

does not depend on y, the equality $[(\bar{P}^*)^k \bar{P}^k - \tau I] f_n = g_n$ implies that $\tau f_n(u,y) + g_n(u,y)$ also does not depend on y, that is, we can for example write:

$$\tau f_n(u,y) + g_n(u,y) = \int [\tau f_n(u,y') + g_n(u,y')] Q(u,\mathrm{d}y') \tag{12}$$

and as a consequence

$$\tau f_n(u,y) - \tau \int f_n(u,y') Q(u,\mathrm{d}y') + g_n(u,y) = \int g_n(u,y') Q(u,\mathrm{d}y'). \tag{13}$$

Put

$$F_n(u) = \int_{y'} f_n(u,y') Q(u,\mathrm{d}y').$$

Then

$$\begin{aligned}[(P^*)^k P^k - \tau I] F_n(u) &= \int_{u'} (P^*)^k P^k(u,\mathrm{d}u') F_n(u') - \tau F_n(u) \\ &= \int_{u'} (P^*)^k P^k(u,\mathrm{d}u') \int_{y'} f_n(u',y') Q(u',\mathrm{d}y') - \tau \int_{y'} f_n(u,y') Q(u,\mathrm{d}y') \\ &= \int_{u'} \int_{y'} (P^*)^k P^k(u,\mathrm{d}u') Q(u',\mathrm{d}y') f_n(u',y') - \tau \int_{y'} f_n(u,y') Q(u,\mathrm{d}y') \\ &= \int_{(u',y')} (\bar{P}^*)^k \bar{P}^k ((u,y), \mathrm{d}(u',y')) f_n(u',y') - \tau \int_{y'} f_n(u,y') Q(u,\mathrm{d}y')\end{aligned}$$

where we used again (11), and thus

$$\begin{aligned}[(P^*)^k P^k - \tau I]F_n(u) &= (\bar{P}^*)^k \bar{P}^k f_n(u,y) - \tau \int_{y'} f_n(u,y')Q(u,\mathrm{d}y') \\ &= [(\bar{P}^*)^k \bar{P}^k - \tau I]f_n(u,y) + \tau f_n(u,y) - \tau \int_{y'} f_n(u,y')Q(u,\mathrm{d}y') \\ &= g_n(u,y) + \tau f_n(u,y) - \tau \int_{y'} f_n(u,y')Q(u,\mathrm{d}y') \\ &= \int g_n(u,y')Q(u,\mathrm{d}y')\end{aligned}$$

according to (13). We can put $G_n(u) = \int g_n(u,y')Q(u,\mathrm{d}y')$, we thus have $[(P^*)^k P^k - \tau I]F_n(u) = G_n(u)$. Using Jensen,

$$\|G_n\|_\pi^2 = \int \left(\int g_n(u,y')Q(u,\mathrm{d}y')\right)^2 \pi(\mathrm{d}u) \le \int_u \int_{y'} g_n(u,y')^2 Q(u,\mathrm{d}y')\pi(\mathrm{d}u) = \|g_n\|_{\bar{\pi}}^2 \to 0.$$

Then, observe that

$$\begin{aligned}\|\tau F_n - G_n\|_\pi^2 &= \int_u \left(\int_{y'} \tau f_n(u,y')Q(u,\mathrm{d}y') - \int_{y'} g_n(u,y')Q(u,\mathrm{d}y')\right)^2 \pi(\mathrm{d}u) \\ &= \int_u \int_y \left(\int_{y'} \tau f_n(u,y')Q(u,\mathrm{d}y') - \int_{y'} g_n(u,y')Q(u,\mathrm{d}y')\right)^2 \pi(\mathrm{d}u)Q(u,\mathrm{d}y) \\ &= \int_u \int_y \left(\tau f_n(u,y) - g_n(u,y)\right)^2 \pi(\mathrm{d}u)Q(u,\mathrm{d}y)\end{aligned}$$

where we used (12), and thus

$$\|\tau F_n - G_n\|_\pi^2 = \|\tau f_n - g_n\|_{\bar{\pi}}^2,$$

that is

$$\|\tau F_n - G_n\|_\pi = \|\tau f_n - g_n\|_{\bar{\pi}}$$

and finally

$$\|\tau F_n\|_\pi \ge \|\tau F_n - G_n\|_\pi - \|G_n\|_\pi = \|\tau f_n - g_n\|_{\bar{\pi}} - \|G_n\|_\pi \xrightarrow[n\to\infty]{} \tau.$$

Besides, using Jensen again, we have

$$\|F_n\|_\pi^2 = \int \left(\int f_n(u,y')Q(u,\mathrm{d}y')\right)^2 \pi(\mathrm{d}u) \le \int_u \int_{y'} f_n(u,y')^2 Q(u,\mathrm{d}y')\pi(\mathrm{d}u) = \|f_n\|_{\bar{\pi}}^2 \to 1.$$

Therefore, thanks to the sandwich theorem, and as $\tau \neq 0$, we can divide both sides by τ to get $\|F_n\|_\pi \to 1$. So we found sequences (F_n) and (G_n) such that $[(P^*)^k P^k - \tau I]F_n(u) = G_n(u)$, $\|F_n\|_\pi \to 1$ and $\|G_n\|_\pi \to 0$, which proves that $\tau \in \mathrm{sp}((P^*)^k P^k)$.

Step 4: as $(P^*)^k P^k$ and $(\bar{P}^*)^k \bar{P}^k$ are both Markov kernels, they both admit $\tau = 1$ as an eigenvalue, associated to the constant eigenfunction $F(u) = 1$ for $(P^*)^k P^k$ and $f(u,y) = 1$ for $(\bar{P}^*)^k \bar{P}^k$. In case $(P^*)^k P^k$ admits another eigenfunction $F_1(u)$ associated to $\tau = 1$, it is necessarily non-constant, and we easily show that $f_1(u,y) = F_1(u)$ is then an eigenfunction of $(\bar{P}^*)^k \bar{P}^k$ associated to $\tau = 1$ (computations similar to step 2), which is non-constant, and thus different from f. On the other hand, if $f_1(u,y)$ is a non-constant eigenfunction $(\bar{P}^*)^k \bar{P}^k$ associated to $\tau = 1$, the equation $(\bar{P}^*)^k \bar{P}^k f_1 = f_1$ implies that $f_1(u,y)$ does not depend on y (the left-hand side does not depend on y), which allows to define $F_1(u) = f_1(u,y)$. We then check immediately that $F_1(u)$ is an eigenfunction of $(P^*)^k P^k$ associated to $\tau = 1$ and non constant, thus, different from $F(u)$.

Conclusion of the proof: from step 4, the eigenvalue 1 has the same multiplicity for $(P^*)^k P^k$ and $(\bar{P}^*)^k \bar{P}^k$. If this multiplicity is larger than 1, we have $\gamma\left((\bar{P}^*)^k \bar{P}^k\right) = \gamma\left((P^*)^k P^k\right) = 0$. Let us now

assume that this multiplicity is equal to 1. Put $\mathrm{sp}_1((P^*)^k P^k) = \{\tau \in \mathrm{sp}((P^*)^k P^k) : \tau \neq 1\} \subset [0,1)$ and $\mathrm{sp}_1((\bar{P}^*)^k \bar{P}^k) = \{\tau \in \mathrm{sp}((\bar{P}^*)^k \bar{P}^k) : \tau \neq 1\} \subset [0,1)$. From step 2, $\mathrm{sp}_1((P^*)^k P^k) \subset \mathrm{sp}_1((\bar{P}^*)^k \bar{P}^k)$. From step 3, $\mathrm{sp}_1((\bar{P}^*)^k \bar{P}^k) \subset \mathrm{sp}_1((P^*)^k P^k) \cup \{0\}$. This proves that $\sup \mathrm{sp}_1((\bar{P}^*)^k \bar{P}^k) = \sup \mathrm{sp}_1((P^*)^k P^k)$. Thus

$$\gamma\left((\bar{P}^*)^k \bar{P}^k\right) = 1 - \sup \mathrm{sp}_1((\bar{P}^*)^k \bar{P}^k) = 1 - \sup \mathrm{sp}_1((P^*)^k P^k) = \gamma\left((P^*)^k P^k\right).$$

This concludes the proof. □

The variance of $\ell(f_\theta(U_t), Y_t)$ under the stationary distribution $\bar{\pi}$ will appear in some of the proofs. In order to keep formulas short enough, let us introduce a short notation for this quantity.

Definition A.1. *We put, for any $\theta \in \Theta$,*

$$V_{\ell(\theta)} := \int_{(u,y)} \left[\ell(f_\theta(u), y) - R(\theta)\right]^2 \bar{\pi}(\mathrm{d}(u,y)).$$

Remark A.1. *In this paper, using the assumption $\ell \leq c$, we will always use the upper bound $V_{\ell(\theta)} \leq c^2$. This bound can be poor in some situations. In the i.i.d. setting, important efforts were made to provide tighter empirical bounds for $V_{\ell(\theta)}$, we mentioned Seldin et al. (2012a) (and many more papers) in the introduction. In this work, our primary objective was to provide empirical upper bounds on γ_{ps}, but we mention that providing tight empirical upper bounds on $V_{\ell(\theta)}$ in the Markov case would be extremely useful making our bounds tighter.*

A.2 Proof of the Result in Section 2

Proof of Theorem 2.1. The theorem assumes that $(U_t)_t$ is a Markov chain with spectral gap $\gamma_{ps} > 0$. Lemma A.1 ensures that the pairs $((U_t, Y_t))_t$ also form a Markov chain with the same spectral gap. Using Theorem 3.4 of Paulin (2015) (or, more precisely, the last inequality in the proof of Theorem 3.4), we have for every $s < \frac{\gamma_{ps}}{10}$

$$\mathbb{E}_{\mathcal{S}}\Big[\exp\left(sn\left(R(\theta) - r(\theta)\right)\right)\Big] \leq \exp\left(\frac{2s^2\left(n + \frac{1}{\gamma_{ps}}\right)V_{\ell(\theta)}}{\gamma_{ps} - 10s}\right).$$

By plugging $s = \frac{\tilde{\lambda}}{n}$ for some $\tilde{\lambda} > 0$, it becomes

$$\mathbb{E}_{\mathcal{S}}\Big[\exp\Big(\tilde{\lambda}\left(R(\theta) - r(\theta)\right)\Big)\Big] \leq \exp\left(\frac{2\tilde{\lambda}^2\left(n + \frac{1}{\gamma_{ps}}\right)V_{\ell(\theta)}}{n^2(\gamma_{ps} - \frac{10\tilde{\lambda}}{n})}\right).$$

For the convenience of writing put

$$A(\theta, \gamma_{ps}, \tilde{\lambda}) = \exp\left(\frac{2\tilde{\lambda}^2\left(n + \frac{1}{\gamma_{ps}}\right)V_{\ell(\theta)}}{n^2(\gamma_{ps} - \frac{10\tilde{\lambda}}{n})}\right).$$

By rearranging terms and integrating both sides with respect to μ we get

$$\mathbb{E}_{\theta\sim\mu}\mathbb{E}_{\mathcal{S}}\left[\exp\Big(\tilde{\lambda}\left(R(\theta) - r(\theta)\right) - A(\theta, \gamma_{ps}, \tilde{\lambda})\Big)\right] \leq 1.$$

We can change the order of expectations due to Fubini–Tonelli's theorem, hence

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\theta\sim\mu}\left[\exp\Big(\tilde{\lambda}\left(R(\theta) - r(\theta)\right) - A(\theta, \gamma_{ps}, \tilde{\lambda})\Big)\right] \leq 1\,,$$

then using Donsker and Varadhan's variational formula (see for example Lemma 2.2 page 28 of Alquier (2024)) with $h(\theta) = \tilde{\lambda}R(\theta) - \tilde{\lambda}r(\theta) - A(\theta, \gamma_{ps}, \tilde{\lambda})$ we arrive to

$$\mathbb{E}_{\mathcal{S}}\left[\exp\left(\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[\tilde{\lambda}\left(R(\theta)-r(\theta)\right)-A(\theta,\gamma_{ps},\tilde{\lambda})\right]+KL(\rho||\mu)\right\}\right)\right]\leq 1.$$

Now let us transition to a probability bound, that is for any $s > 0$

$$\begin{aligned}&\mathop{\mathbb{P}}_{\mathcal{S}}\left(\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[\tilde{\lambda}\left(R(\theta)-r(\theta)\right)-A(\theta,\gamma_{ps},\tilde{\lambda})\right]+KL(\rho||\mu)\right\}> s\right)\\&\leq e^{-s}\mathbb{E}_{\mathcal{S}}\left[\exp\left(\sup_{\rho\in\mathcal{P}(\Theta)}\left\{\mathbb{E}_{\theta\sim\rho}\left[\tilde{\lambda}\Big(R(\theta)-r(\theta)\Big)-A(\theta,\gamma_{ps},\tilde{\lambda})\right]+KL(\rho||\mu)\right\}\right)\right]\leq\exp(-s).\end{aligned}$$

By denoting $\delta = \exp(-s)$, and rewriting $A(\theta, \gamma_{ps}, \tilde{\lambda})$ explicitly, we obtain

$$\mathbb{P}_{\mathcal{S}}\left(\forall\rho\in\mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right]\leq\mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right]+\mathbb{E}_{\theta\sim\rho}\left[\frac{2\tilde{\lambda}\left(n+\frac{1}{\gamma_{ps}}\right)V_{\ell(\theta)}}{n^2(\gamma_{ps}-\frac{10\tilde{\lambda}}{n})}\right]+\frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\tilde{\lambda}}\right)\geq 1-\delta.$$

With a bounded loss $\ell(\cdot,\cdot)\leq c$, we are able to bound the variance term $V_{\ell(\theta)} = \mathrm{var}_{\pi'}\left[\ell(f_\theta(U_t), Y_t)\right] \leq c^2$, and replace the expectation on right hand side by its upper bound:

$$\mathbb{P}_{\mathcal{S}}\left(\forall\rho\in\mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right]\leq\mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right]+\frac{2\tilde{\lambda}c^2\left(n+\frac{1}{\gamma_{ps}}\right)}{n^2(\gamma_{ps}-\frac{10\tilde{\lambda}}{n})}+\frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\tilde{\lambda}}\right)\geq 1-\delta\ .$$

Finally, put $\lambda = \tilde{\lambda}/\gamma_{ps}$ to obtain:

$$\mathbb{P}_{\mathcal{S}}\left(\forall\rho\in\mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right]\leq\mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right]+\frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda}+\frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda\gamma_{ps}}\right)\geq 1-\delta\ .$$

□

A.3 Proof of the Results in Section 3

Proof of Proposition 3.1. We start with Theorem 5.3 from Wolfer and Kontorovich (2019). The results are in the format of bounding the sample complexity n, and we restate them as concentration results. The last part of the proof of the theorem can be translated to the following concentration form:

$$\begin{aligned}&\mathbb{P}\Big(|\gamma_{ps}-\widehat{\gamma}_{ps}|\geq\varepsilon\Big)\\&\qquad\leq\sum_{k=1}^{\lceil\frac{2}{\varepsilon}\rceil}2d\exp\left(-C\frac{n\varepsilon^2\pi_*k}{\|P\|_\pi\|P^k\|_1}\right)+\sum_{k=1}^{\lceil\frac{2}{\varepsilon}\rceil}\frac{d}{\sqrt{\pi_*}}\exp\left(-\frac{n\varepsilon^2\pi_*\gamma_{ps}}{C}\right)+\frac{Kd}{\sqrt{2\pi_*}}\exp\left(-\frac{n\varepsilon^2\pi_*\gamma_{ps}}{CK^2}\right)\end{aligned}$$

We simplify it to the following confidence interval:

$$\mathbb{P}\Big(|\gamma_{ps}-\widehat{\gamma}_{ps}|\geq\varepsilon\Big)\leq\frac{C_{\mathrm{ps}}d}{\varepsilon\sqrt{\pi_*}}\exp\left(-n\varepsilon^2\pi_*\min\left\{\gamma_{ps},\min_{1\leq k\leq\lceil\frac{\varepsilon}{2}\rceil}\left\{\frac{1}{\|P\|_\pi\|P^k\|_1}\right\}\right\}\right)$$

Or more concisely:

$$\mathbb{P}\left(|\gamma_{ps}-\widehat{\gamma}_{ps}|\geq\varepsilon\right)\leq\frac{C_{\mathrm{ps}}d}{\varepsilon\sqrt{\pi_*}}e^{-n\varepsilon^2\pi_*\min\left\{\gamma_{ps},\frac{1}{C(P)}\right\}}\tag{14}$$

Dividing both sides of the inequality by γ_{ps} and re-defining ε as ε/γ_{ps}, we obtain the result. □

Proof of Proposition 3.2. First, observe that (U_t) is a Gaussian process, so that the vector (U_{t-1}, U_t) is a Gaussian vector:

$$(U_{t-1}, U_t) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{1-a^2} & \frac{a}{1-a^2} \\ \frac{a}{1-a^2} & \frac{1}{1-a^2} \end{pmatrix}\right).$$

Thus, if we put $W_t = U_t\sqrt{1-a^2}$, it defines a Markov chain which is also a Gaussian process with

$$(W_{t-1}, W_t) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}\right).$$

Let $p(w, w')$ denote the joint density of (W_{t-1}, W_t) and $p(w)$ denote the marginal density of W_t. The transition kernel P_W of W can be written through its density: $P_W(w, \mathrm{d}w') = [p(w, w')/p(w)]\mathrm{d}w'$. Note that the symmetry $p(w, w') = p(w', w)$ leads to $p(w)P_W(w, \mathrm{d}w')\mathrm{d}w = p(w')P_W(w', \mathrm{d}w)\mathrm{d}w'$, that is, $P_W^* = P_W$. In other words, from the diagonalization of P_W we will directly obtain the diagonalization of $P_W^* P_W = P_W^2$, from which we will deduce the diagonalization of P.

We will now use Mehler's formula, in the form stated by Kibble (1945):

$$p(w, w') = p(w)p(w')\sum_{n=0}^{\infty} \frac{a^n}{n!}\mathrm{He}_n(w)\mathrm{He}_n(w')$$

where (He_n) are the Hermite polynomials satisfying:

$$\int_{\mathbb{R}} \mathrm{He}_n(x)\mathrm{He}_m(x)\frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{d}x = n!\delta_{nm}.$$

Plugging this in the formula for P_W, we obtain:

$$\begin{aligned} P_W(w, \mathrm{d}w') &= \frac{p(w)p(w')\sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)\mathrm{He}_n(w')}{p(w)}\mathrm{d}w' \\ &= p(w')\sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)\mathrm{He}_n(w')\mathrm{d}w'. \end{aligned}$$

This gives the diagonalization of P_W. The eigenfunctions of P_W are the (He_n) with corresponding eigenvalues a^n:

$$\begin{aligned} \int \mathrm{He}_m(w')P_W(w, \mathrm{d}w') &= \int \mathrm{He}_m(w')p(w')\sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)\mathrm{He}_n(w')\mathrm{d}w' \\ &= \sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)\int p(w')\mathrm{He}_n(w')\mathrm{He}_m(w')\mathrm{d}w' \\ &= \sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)\int \frac{\mathrm{e}^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\mathrm{He}_n(w')\mathrm{He}_m(w')\mathrm{d}w' \\ &= \sum_{n=0}^{\infty}\frac{a^n}{n!}\mathrm{He}_n(w)n!\delta_{nm} \\ &= a^m\mathrm{He}_m(w). \end{aligned}$$

The last thing to note is that we want the diagonalization of P, not of P_W. However, observe that $P(u, \mathrm{d}u') = P_W(u/\sqrt{1-a^2}, \mathrm{d}u'/\sqrt{1-a^2})$ which obviously has different eigenfunctions $\mathrm{He}_n(\cdot/\sqrt{1-a^2})$ but the same eigvenvalues a^n:

$$\begin{aligned} \int \mathrm{He}_m(u'/\sqrt{1-a^2})P(u, \mathrm{d}u') &= \int \mathrm{He}_m(u'/\sqrt{1-a^2})P_W(u/\sqrt{1-a^2}, \mathrm{d}u'/\sqrt{1-a^2}) \\ &= \int \mathrm{He}_m(w')P_W(u/\sqrt{1-a^2}, \mathrm{d}w') \qquad (\text{ by c.o.v. } w' = u'/\sqrt{1-a^2}) \\ &= a^m\mathrm{He}_m(u/\sqrt{1-a^2}). \end{aligned}$$

Thus, the eigenvalues of P are $\{1, a, a^2, a^3, \dots\}$ and the eigenvalues of P^*P are $\{1, a^2, a^4, a^6, \dots\}$, that is, $\gamma_{ps} = 1 - a^2$. □

Proof of Proposition 3.3. We follow Nakakita et al. (2025): if we can prove that the sequence $(U_1, \dots, U_n)$ satisfies a log-Sobolev inequality with constant $K > 0$, then, Theorem 1 in Nakakita et al. (2025) gives, with probability at least $1 - \exp(-t)$,

$$\left| \frac{1}{n} \sum_{t=1}^{n} U_t^2 - \mathbb{E}(U_t^2) \right| \le 12\sqrt{K} |\mathbb{E}(U_t^2)| \sqrt{\frac{9 + 4t}{n}}.$$

However, Section 3.2 in Nakakita et al. (2025) shows that, if the sequence (ζ_t) satisfies a log-Sobolev inequality with constant $K_\zeta > 0$, then the sequence $(U_1, \dots, U_n)$ satisfies a log-Sobolev inequality with constant

$$K = K_\zeta \frac{1 - a^2}{(1 - |a|)^2}.$$

Moreover, Theorem 5.4 in Boucheron et al. (2006)) actually states that the sequence (ζ_t) satisfies a log-Sobolev inequality with constant $K_\zeta = 1$. From the previous proposition, $\mathbb{E}(U_t^2) = \frac{1}{1-a^2} = \frac{1}{\gamma_{ps}}$. Putting $\delta = \exp(-t)$, we obtain:

$$\left| \frac{1}{n} \sum_{t=1}^{n} U_t^2 - \frac{1}{\gamma_{ps}} \right| \le 12 \sqrt{\frac{1 - a^2}{(1 - |a|)^2}} \frac{1}{\gamma_{ps}} \sqrt{\frac{9 + 4 \log \frac{1}{\delta}}{n}}.$$

Observe that $\gamma_{ps} = 1 - a^2 \le 1$ by definition. Thus,

$$\left| \frac{1}{\widehat{\gamma}_{ps}} - \frac{1}{\gamma_{ps}} \right| = \left| \frac{1}{\min\left(1, \left(\frac{1}{n} \sum_{t=1}^{n} U_t^2\right)^{-1}\right)} - \frac{1}{\gamma_{ps}} \right| \le \left| \frac{1}{n} \sum_{t=1}^{n} U_t^2 - \frac{1}{\gamma_{ps}} \right|,$$

and hence

$$\left| \frac{1}{\widehat{\gamma}_{ps}} - \frac{1}{\gamma_{ps}} \right| \le \frac{12}{(1 - |a|)\sqrt{\gamma_{ps}}} \sqrt{\frac{9 + 4 \log \frac{1}{\delta}}{n}}.$$

Multiply both sides by $\widehat{\gamma}_{ps} \le 1$ to get

$$\left| 1 - \frac{\widehat{\gamma}_{ps}}{\gamma_{ps}} \right| \le \frac{12}{(1 - \sqrt{1 - \gamma_{ps}})\sqrt{\gamma_{ps}}} \sqrt{\frac{9 + 4 \log \frac{1}{\delta}}{n}} \le \frac{24}{\gamma_{ps}^{3/2}} \sqrt{\frac{9 + 4 \log \frac{1}{\delta}}{n}}.$$

□

Lemma A.2. *The statement of Theorem 2.1 remains valid when $r(\theta)$ and $R(\theta)$ are interchanged, that is:*

$$\mathbb{P}_{\mathcal{S}} \left(\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta \sim \rho}\left[r(\theta)\right] \le \mathbb{E}_{\theta \sim \rho}\left[R(\theta)\right] + \frac{2\lambda c^2 \left(1 + \frac{1}{n\gamma_{ps}}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log \frac{1}{\delta}}{\lambda \gamma_{ps}} \right) \ge 1 - \delta.$$

Proof of Lemma A.2. The inequality in Theorem 2.1 is originally stated with the following form

$$\mathbb{E}_{\theta \sim \rho}[R(\theta)] \le \mathbb{E}_{\theta \sim \rho}[r(\theta)] + \frac{2\lambda c^2 \left(1 + \frac{1}{n\gamma_{ps}}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log \frac{1}{\delta}}{\lambda \gamma_{ps}}.$$

We observe that replacing the loss function $\ell(\cdot, \cdot)$ by its negative, i.e., defining $\tilde{\ell} := -\ell$, allows us to reverse the inequality. Since all derivations in the proof of Theorem 2.1 depend linearly on ℓ, the same steps apply with $\tilde{\ell}$, leading to

$$\mathbb{E}_{\theta \sim \rho}[-R(\theta)] \le \mathbb{E}_{\theta \sim \rho}[-r(\theta)] + \frac{2\lambda c^2 \left(1 + \frac{1}{n\gamma_{ps}}\right)}{n - 10\lambda} + \frac{KL(\rho||\mu) + \log \frac{1}{\delta}}{\lambda \gamma_{ps}},$$

which results in the claimed bound:

$$\mathbb{E}_{\theta\sim\rho}[r(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[R(\theta)] + \frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda} + \frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda\gamma_{ps}}.$$

□

It is straightforward to notice that all other developments after Theorem 2.1 can be done in analogous manner also for the interchanged version of the bound.

Proof of Theorem 3.1. We rewrite Proposition 3.1 in two ways:

$$\gamma_{ps} - \widehat{\gamma}_{ps} \leq |\gamma_{ps} - \widehat{\gamma}_{ps}| \leq \left|\frac{\widehat{\gamma}_{ps}-\gamma_{ps}}{\widehat{\gamma}_{ps}}\right| \leq \left|1-\frac{\gamma_{ps}}{\widehat{\gamma}_{ps}}\right| \leq \varepsilon \qquad \text{[using that } \widehat{\gamma}_{ps}\leq 1\text{]} \tag{15}$$

$$\frac{1}{\gamma_{ps}} \leq \frac{1+\varepsilon}{\widehat{\gamma}_{ps}}. \tag{16}$$

Both of them simultaneously hold with probability $1-\alpha(n,\gamma_{ps},\varepsilon)$.

Recall that

$$B(\nu,u) = \min_{\lambda\in\Lambda}\left\{\frac{2\lambda c^2\left(1+\frac{u}{n}\right)}{n-10\lambda} + u\frac{KL(\nu||\mu)+\log\frac{L}{\delta}}{\lambda}\right\}$$

and

$$\hat{\rho} = \operatorname*{argmin}_{\rho}\left[\mathbb{E}_{\theta\sim\rho}\big[r(\theta)\big] + B\left(\rho, \frac{(1+\varepsilon)}{\widehat{\gamma}_{ps}}\right)\right].$$

The function B is a non-decreasing function on a second variable, hence applying (16) to the bound in Theorem 2.1, by union bound argument we have with probability $1-\delta-\alpha$ for all $\rho\in\mathcal{P}(\Theta)$

$$\mathbb{E}_{\theta\sim\rho}[R(\theta)] \leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + B\left(\rho, \frac{1+\varepsilon}{\widehat{\gamma}_{ps}}\right).$$

Particularly

$$\begin{aligned}
\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] &\leq \mathbb{E}_{\theta\sim\hat{\rho}}[r(\theta)] + B\left(\hat{\rho}, \frac{1+\varepsilon}{\widehat{\gamma}_{ps}}\right) && \text{[w.p. } 1-\delta-\alpha\text{]}\\
&\leq \mathbb{E}_{\theta\sim\rho}[r(\theta)] + B\left(\rho, \frac{1+\varepsilon}{\widehat{\gamma}_{ps}}\right) && \text{[for all } \rho\text{, by definition of } \hat{\rho}\text{]}\\
&\leq \mathbb{E}_{\theta\sim\rho}[R(\theta)] + B\left(\rho, \frac{1}{\gamma_{ps}}\right) + B\left(\rho, \frac{1+\varepsilon}{\widehat{\gamma}_{ps}}\right) && \text{[w.p. } 1-2\delta-\alpha \text{ by Lemma A.2]}\\
&\leq \mathbb{E}_{\theta\sim\rho}[R(\theta)] + B\left(\rho, \frac{1}{\gamma_{ps}}\right) + B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right) && \text{[by (15)]}\\
&\leq \mathbb{E}_{\theta\sim\rho}[R(\theta)] + 2B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right).
\end{aligned}$$

Thus with probability at least $1-2\delta-\alpha(n,\gamma_{ps},\varepsilon)$, we have

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \leq \inf_{\substack{\rho\in\mathcal{P}(\Theta)\\ \lambda\in\Lambda}}\left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + 2B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right)\right].$$

□

Remark A.2. *Instead of relaxing the last line of the proof, it is also possible to leave it in the exact form by rewriting it as*

$$B\left(\rho, \frac{1}{\gamma_{ps}}\right) + B\left(\rho, \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right) = B\left(\rho, \frac{1}{\gamma_{ps}} + \frac{1+\varepsilon}{\gamma_{ps}-\varepsilon}\right) = B\left(\rho, \frac{2\gamma_{ps}+\varepsilon(1-\gamma_{ps})}{\gamma_{ps}(\gamma_{ps}-\varepsilon)}\right).$$

Thus it will result in the slightly tighter, but maybe less readable:

$$\mathbb{E}_{\theta\sim\hat{\rho}}[R(\theta)] \le \inf_{\substack{\forall \rho\in\mathcal{P}\\ \lambda\in\Lambda}} \left[\mathbb{E}_{\theta\sim\rho}[R(\theta)] + 2B\left(\rho, \frac{2\gamma_{ps}+\varepsilon(1-\gamma_{ps})}{\gamma_{ps}(\gamma_{ps}-\varepsilon)}\right)\right].$$

A.4 Proof of the Results in Section 4

Proof of Theorem 4.1. We start by an application of Theorem 2.1, with probability at least $1-\delta$ on the sample $\mathcal{S}$,

$$\forall \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\big[R(\theta)\big] \le \mathbb{E}_{\theta\sim\rho}\big[r(\theta)\big] + B(\rho, \gamma_{ps}, \lambda).$$

In particular, this holds for any ρ in the set of Dirac masses $\{\delta_\theta \mid \theta\in\Theta\}$. Thus, we have, with probability at least $1-\delta$,

$$\forall \theta \in \Theta,\ \mathbb{E}_{\theta\sim\delta_\theta}\big[R(\theta)\big] \le \mathbb{E}_{\theta\sim\delta_\theta}\big[r(\theta)\big] + B(\delta_\theta, \gamma_{ps}, \lambda)$$

Having $\mathbb{E}_{\theta\sim\delta_\theta}\big[R(\theta)\big] = R(\theta)$, $\mathbb{E}_{\theta\sim\delta_\theta}\big[r(\theta)\big] = r(\theta)$, and

$$KL(\rho||\mu) = \sum_{\vartheta} \log\left(\frac{\delta_\theta(\vartheta)}{\mu(\vartheta)}\right)\delta_\theta(\vartheta) = \log\frac{1}{\mu(\theta)}$$

we get that for any $\theta\in\Theta$

$$R(\theta) \le r(\theta) + \frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda} + \frac{\log\frac{1}{\mu(\theta)\delta}}{\lambda\gamma_{ps}}$$

with probability $1-\delta$.
In particular, by putting $\theta = \hat{\theta}_{\mathrm{ERM}}$ we obtain

$$R(\hat{\theta}_{\mathrm{ERM}}) \le \min_\theta r(\theta) + \frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda} + \frac{\log\frac{1}{\mu(\theta)\delta}}{\lambda\gamma_{ps}}$$

With the assumption that μ is uniform

$$R(\hat{\theta}_{\mathrm{ERM}}) \le \min_\theta r(\theta) + \frac{2\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n-10\lambda} + \frac{\log\frac{M}{\delta}}{\lambda\gamma_{ps}}. \tag{17}$$

Assume that $10\lambda < n\varepsilon/(1+\varepsilon)$, then:

$$R(\hat{\theta}_{\mathrm{ERM}}) \le \min_\theta r(\theta) + \frac{2(1+\varepsilon)\lambda c^2\left(1+\frac{1}{n\gamma_{ps}}\right)}{n} + \frac{\log\frac{M}{\delta}}{\lambda\gamma_{ps}}. \tag{18}$$

Then we minimize the right-hand side by choosing the optimal λ. It is achieved for

$$\lambda_{op} = \sqrt{\frac{n\log\frac{M}{\delta}}{2(1+\varepsilon)c^2\gamma_{ps}\left(1+\frac{1}{\gamma_{ps}n}\right)}}$$

in which case the bound settles into its final form:

$$R(\hat{\theta}_{\mathrm{ERM}}) \le \min_\theta r(\theta) + \sqrt{\frac{8(1+\varepsilon)c^2\log\frac{M}{\delta}}{\gamma_{ps}n}\left(1+\frac{1}{\gamma_{ps}n}\right)}.$$

Note that our choice of λ is only compatible with $10\lambda < n\varepsilon/(1+\varepsilon)$ when:

$$n > \frac{50(1+\varepsilon)\gamma_{ps}\log\frac{M}{\delta}}{\varepsilon^2 c^2 \gamma_{ps}\left(1+\frac{1}{\gamma_{ps}n}\right)}.$$

□

B ADDITIONAL DETAILS ON THE EXPERIMENTS

We described in the paper the construction of the transition matrices in our experiments. We remind that $R_t := tP + (1-t)Q$ where P satisfies $\gamma_{ps}(P) \simeq 0$, and Q is such that $\gamma_{ps}(Q) = 1$. We actually took $Q = 1^T \cdot \pi$ where $1^T = (1, \ldots, 1)$ and π is the invariant distribution of P. That is, a Markov chain whose transition kernel Q is simply an i.i.d. sequence from π. The fact that both P and Q have the same invariant distribution π ensures that the invariant distribution of R_t is also π, indeed:

$$\pi R_t = \pi[tP + (1-t)Q] = t\pi P + (1-t)\pi Q = t\pi + (1-t)\pi = \pi.$$

All this was detailed in the main body of the paper, but it remains to give the definition of P.

Our choice for $d = 4$ is

$$P = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ p & 0 & 1-p & 0 \\ 0 & q & 0 & 1-q \end{bmatrix}.$$

We have fixed parameters $p = 0.01$ and $q = 0.001$. The heuristic reason behind such parameters is that, when the chain reaches state 3 and 4, it will stay stuck in this state for a very long time, making the convergence to the stationary distribution very slow. And indeed, we observed that γ_{ps} is very close to 0 for this chain.

For larger d, we generalized the construction in the following way:

$$P = \begin{bmatrix} 1/d & 1/d & 1/d & 1/d & \cdots & 1/d \\ 1/d & 1/d & 1/d & 1/d & \cdots & 1/d \\ p & 0 & 1-p & 0 & \cdots & 0 \\ 0 & q & 0 & 1-q & \cdots & 0 \\ 1/d & 1/d & 1/d & 1/d & \cdots & 1/d \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/d & 1/d & 1/d & 1/d & \cdots & 1/d \end{bmatrix}.$$

In the main body of the paper, we assessed the estimator $\widehat{\gamma}_{ps}$ when $d = 20$ in Figure 1, and the accuracy of the PAC-Bayes bound, also with $d = 20$, in Figure 2.

We now provide similar results when $d = 4$ (Figures 4 and 8 respectively), $d = 10$ (Figures 5 and 9), $d = 50$ (Figures 6 and 10) and finally $d = 100$ (Figures 7 and 11). The results remain essentially unchanged, note however that the estimation of γ_{ps} becomes more challenging when d is very large.

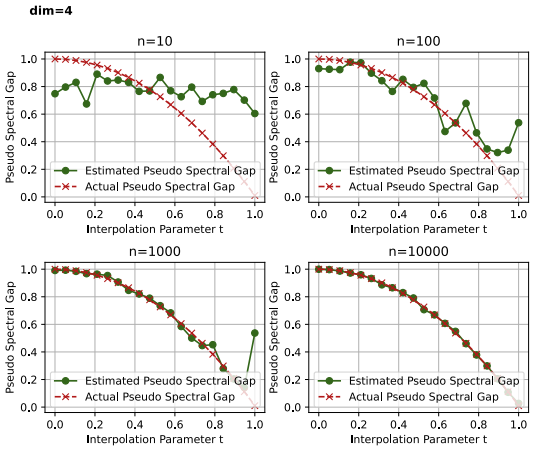

Figure 4: Estimation of γ_{ps} when $d = 4$.

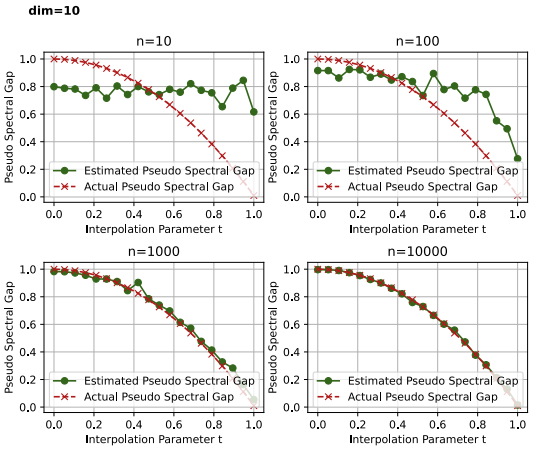

Figure 5: Estimation of γ_{ps} when $d = 10$.

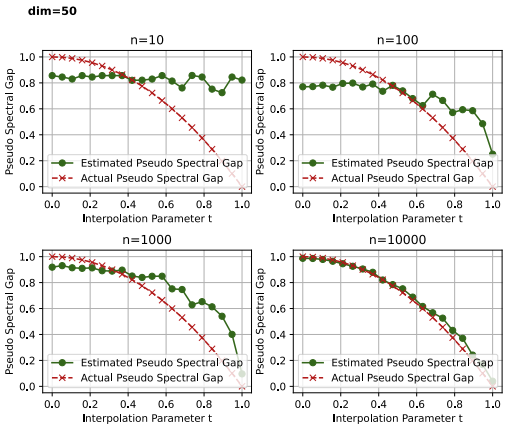

Figure 6: Estimation of γ_{ps} when $d = 50$.

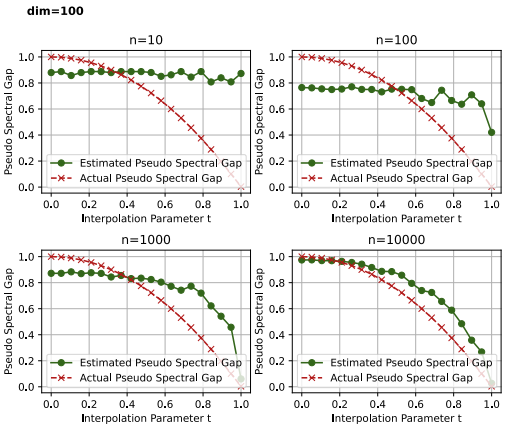

Figure 7: Estimation of γ_{ps} when $d = 100$.

C PAC-BAYES FOR TIME SERIES WITH φ-MIXING

In this final section, we discuss another PAC-Bayes bounds for time series due to Alquier et al. (2013). This bound holds under an assumption on the φ-mixing coefficients of the series, that do not require the series to be a Markov chain. This bound is not empirical. We show that, under the additional assumption that the series is actually a Markov chain, we can upper-bound the φ-mixing coefficients by a function of the pseudo-spectral gap, and thus make the bound empirical.

C.1 Hoeffding-Type PAC-Bayes Bound for Time Series

Our starting point follows the framework introduced by Rio (2000). Consider a sequence of metric spaces (E_t, d_t) for $t = 1, \dots, n$, each with diameter denoted by Δ_t. Let $E^n := E_1 \times E_2 \times \cdots \times E_n$. A real-valued

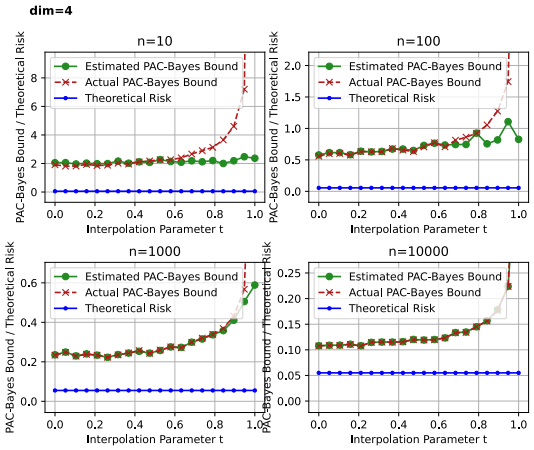

Figure 8: PAC-Bayes bounds for $R(\hat{\theta}_{\mathrm{ERM}})$ when $d = 4$.

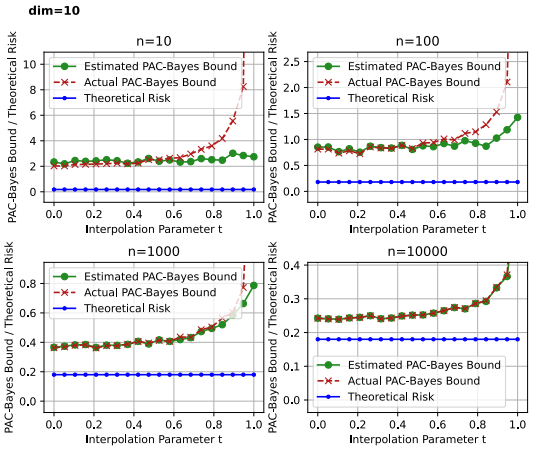

Figure 9: PAC-Bayes bounds for $R(\hat{\theta}_{\mathrm{ERM}})$ when $d = 10$.

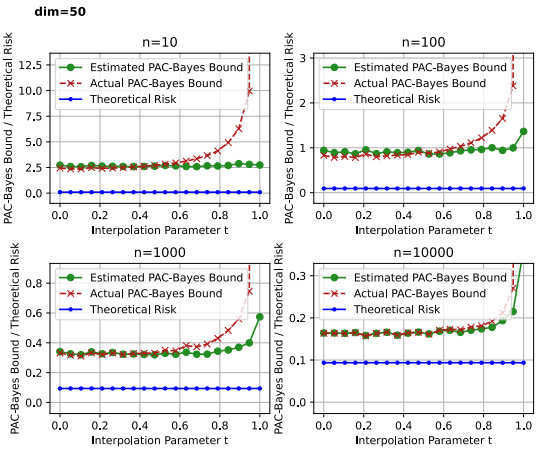

Figure 10: PAC-Bayes bounds for $R(\hat{\theta}_{\mathrm{ERM}})$ when $d = 50$.

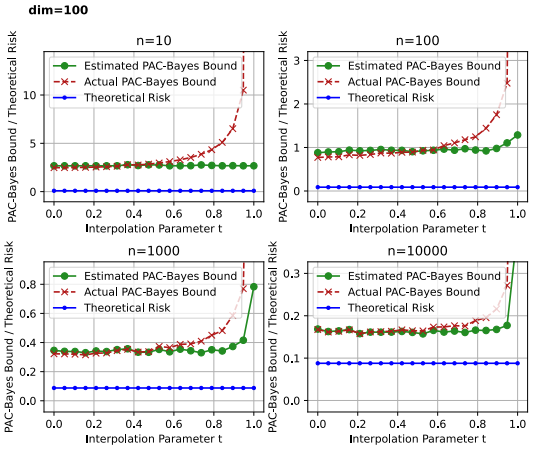

Figure 11: PAC-Bayes bounds for $R(\hat{\theta}_{\mathrm{ERM}})$ $d = 100$.

function $f : E^n \to \mathbb{R}$ is said to be M-Lipschitz if for all $(x_1, \ldots, x_n), (y_1, \ldots, y_n) \in E^n$, we have

$$\left|f(x_1, \ldots, x_n) - f(y_1, \ldots, y_n)\right| \leq M \sum_{t=1}^{n} d_t(x_t, y_t).$$

Now, let $(X_1, \ldots, X_n)$ be a sequence of random variables, and for each $t \in \{1, \ldots, n\}$, let $\mathcal{F}_t$ denote the σ-algebra generated by $X_1, \ldots, X_t$. The idea is to control how much the future (i.e., $X_{t+1}, \ldots, X_n$) can deviate from independence given the past $\mathcal{F}_t$. For each time step $t \in \{1, \ldots, n-1\}$ and any 1-Lipschitz function $g : E_{t+1} \times \cdots \times E_n \to \mathbb{R}$ a deviation measure is defined as

$$\Gamma_t(g) := \left\|\mathbb{E}\left[g(X_{t+1}, \ldots, X_n) \mid \mathcal{F}_t\right] - \mathbb{E}\left[g(X_{t+1}, \ldots, X_n)\right]\right\|_{\infty}, \tag{19}$$

and a sequence $(X_1, \ldots, X_n)$ is said to satisfy a γ-mixing condition if there exists a family of non-negative coefficients $(\gamma_{t,m})_{1 \leq t < m \leq n}$ such that, for all such g, we have:

$$\Gamma_t(g) \leq \sum_{m=t+1}^{n} \gamma_{t,m}. \tag{20}$$

Under Assumption (20), Rio (2000) provides a concentration inequality for 1-Lipschitz functions of dependent sequences. Specifically, if $(X_1, \dots, X_n)$ satisfies the γ-mixing condition described above and the underlying metric spaces have diameter Δ_t, then the following holds for any positive s:

$$\mathbb{E}\left[\exp\left(sf(X_1,\dots,X_n)\right)\right] \le \exp\left\{ s\mathbb{E}\left[f(X_1,\dots,X_n)\right] + \frac{s^2}{8}\sum_{t=1}^{n}\left(\Delta_t + 2\sum_{m>t}\gamma_{t,m}\right)^2\right\}. \tag{21}$$

This result generalizes classical Hoeffding-type inequalities to the dependent setting and is a foundation for many modern generalization bounds involving weakly dependent data.

Returning to the label prediction again, suppose the variables are object×label pairs. So we are considering $X_i = (U_i, Y_i)$ with the same definitions of labels Y_i, predictors f_θ, risk functions $r(\theta)$, $R(\theta)$, and prior distribution μ and same conditions as we had in the beginning of the paper, with the only difference that here U_i are drawn from arbitrary distributions. Recall that c is the uniform upper bound on the loss function ℓ, which also means that ℓ is $c-$Lipschitz. With this in mind, we have freedom to set our own trivial metrics as follows: for all $t = 1, \dots, n$ we define $d_t\big((u,y),(u',y')\big) := c\mathbf{1}_{(u,y)\neq(u',y')}$. This means that for all $t = 1, \dots, n$ diameter $\Delta_t = c$, and as a consequence function $n \cdot r(\theta)$ forces to be 1-Lipschitz in the space $E_1 \times \ldots \times E_n$.

$$\left| r_\theta\left((U_1,Y_1),(U_2,Y_2),\dots,(U_n,Y_n)\right) - r_\theta\left((U_1',Y_1'),(U_2',Y_2'),\dots,(U_n',Y_n')\right)\right| \le \frac{1}{n}(\Delta_1 + \ldots + \Delta_n) = c.$$

Moreover 1-Lipschitzianity of $n \cdot r(\theta)$ holds for any subspace of the form $E_{T+1} \times \ldots \times E_n$, since

$$\left| r_\theta\left((U_1,Y_1),(U_2,Y_2),\dots,(U_n,Y_n)\right) - r_\theta\left((U_1,Y_1),\dots,(U_{T+1}',Y_{T+1}'),\dots,(U_n',Y_n')\right)\right| \le \frac{1}{n}(\Delta_{T+1} + \ldots + \Delta_n) = \frac{n-T}{n}c.$$

Let us denote $C^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\Delta_t + 2\sum_{m>t}\gamma_{t,m}\right)^2$, and the inequality (21) will take a familiar form:

$$\mathbb{E}\left[\exp\left(sR(\theta) - r(\theta)\right)\right] \le e^{\frac{s^2}{8}nC^2}.$$

Thus, by following analogous steps of the proof of Theorem 2.1.

Theorem C.1. *Let $U_1, U_2, \dots, U_n$ be random variables, then for any constants $\lambda > 0$, $\delta \in (0,1)$, and prior $\mu \in \mathcal{P}(\Theta)$,*

$$\mathbb{P}_{\mathcal{S}}\left(\exists \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \le \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{\lambda C^2}{8n} + \frac{KL(\rho||\mu) + \log\frac{1}{\delta}}{\lambda}\right) \ge 1-\delta$$

where $C^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\Delta_t + 2\sum_{m>t}\gamma_{t,m}\right)^2$.

Note that this result is essentially Theorem 2 of Alquier et al. (2013).

C.2 Second PAC-Bayes Bound for Markov Chains and $\varphi-$mixing

As an application to Theorem (C.1), we investigate the case of stationary, ergodic, d-state Markov chains, with pseudo-spectral gap $\gamma_{ps} > 0$. In this case, we are able to get an upper bound on $C^2(\theta)$, and obtain a PAC-Bayes bound that depends on the γ_{ps}, and thus that can be made empirical.

Before we get to that, let us start with the definition of $\varphi-$mixing coefficients.

Definition C.1. *Suppose $(\Omega, \mathcal{E}, \mathbb{P})$ is a probability space, $\mathcal{A} \subset \mathcal{E}$ and $\mathcal{B} \subset \mathcal{E}$ are σ-fields, then*

$$\varphi(\mathcal{A},\mathcal{B}) := \sup_{\substack{A\in\mathcal{A}\\ B\in\mathcal{B}}} \left|\mathbb{P}(B \mid A) - \mathbb{P}(B)\right|.$$

We refer the reader to Doukhan (1995) for more details on this definition. In the time series setting $(U_1, U_2, U_3, ...)$ coefficients φ_k are defined as

$$\varphi(k) = \sup_{t\in\mathbb{N}} \varphi\big(\sigma(U_1,\dots,U_t), \sigma(U_{t+k}, U_{t+k+1},\dots)\big). \tag{22}$$

Following (Rio, 2000), the deviation measure $\gamma_{t,m}$ can be characterized by $\varphi-$mixing coefficients as follows

$$\gamma_{t,m} \leq \Delta_t \varphi(m-t).$$

Hence, we also arrive to the $\varphi-$mixing version of a PAC-Bayes bound.

Corollary C.1. *Let $U_1, U_2, \dots, U_n$ be random variables, and for each $t = 1, \dots, n$, let $\varphi(t)$ be defined as in (22), then for any constants $\lambda > 0$, $\delta \in (0,1)$, and prior $\mu \in \mathcal{P}(\Theta)$,*

$$\mathbb{P}_{\mathcal{S}}\left(\exists \rho \in \mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \leq \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{\lambda\Phi}{8n} + \frac{KL(\rho||\mu) + \log\frac{1}{\delta}}{\lambda}\right) \geq 1-\delta$$

where $\Phi = \frac{1}{n}\sum_{t=1}^{n}\left(\Delta_t + 2\left(\Delta_t\varphi(1) + \Delta_t\varphi(2) + \ldots + \Delta_t\varphi(n)\right)\right)^2$.

Now let us make the connection of φ-mixing, t_{mix}, and γ_{ps}. Given a stationary, ergodic Markov chain U, the coefficients $\varphi(k)$ are known to be expressed in terms of the distance to equilibrium:

$$\varphi(k) := \varphi_U(k) = \sup_{u\in\mathcal{U}} \|P^k(u,\cdot) - \pi(\cdot)\|_{TV},$$

as proven by Davydov (1968). This measure is non-increasing and is endowed with a sub-multiplicative property, namely

$$\begin{aligned} \varphi(t_1) &\leq \varphi(t_2) && [\text{when } t_1 \leq t_2] \\ \varphi(t_1+t_2) &\leq 2\varphi(t_1)\varphi(t_2) && [\text{for any } t_1, t_2] \end{aligned}$$

Denoting $a = \left\lceil \frac{k}{t_{\text{mix}}} \right\rceil$, and $\rho = \left(\frac{1}{2}\right)^{\frac{1}{t_{\text{mix}}}}$, then applying sub-multiplicative properties we derive

$$\varphi(k) \leq \varphi\left(a \cdot t_{\text{mix}}\right) \leq 2^{a-1}\varphi(t_{\text{mix}})^a \leq 2^{a-1}\left(\frac{1}{4}\right)^a = \left(\frac{1}{2}\right)^{a+1} \leq \frac{1}{2}\cdot\left(\frac{1}{2}\right)^{\frac{k}{t_{\text{mix}}}} = \frac{1}{2}\cdot\rho^k.$$

On the other hand using an upper bound on t_{mix} Paulin (2015), we have

$$b(\pi_*)\,\gamma_{ps} \ \leq\ \frac{1}{t_{\text{mix}}}$$

with $b(\pi_*) = \left(\ln\frac{1}{\pi_*} + 2\ln 2 + 1\right)^{-1}$. Subsequently, for ergodic Markov chains with pseudo-spectral gap γ_{ps} we are able to derive bounds on ρ and $\varphi(k)$ which depend on γ_{ps}.

$$\rho = \left(\frac{1}{2}\right)^{\frac{1}{t_{\text{mix}}}} \leq \left(\frac{1}{2}\right)^{b(\pi_*)\gamma_{ps}} \quad \text{and} \quad \varphi(k) \leq \frac{1}{2}\cdot\rho^k \leq \left(\frac{1}{2}\right)^{k\cdot b(\pi_*)\gamma_{ps}+1}$$

with $\alpha = \frac{C_{\text{ps}}d}{\varepsilon\sqrt{\pi_*}}e^{-n\varepsilon^2\pi_*\min\{\gamma_{ps}, \frac{1}{C(P)}\}}$.

Thus, applying the aforementioned bounds, we derive

$$\begin{aligned}\Phi &= \frac{1}{n}\sum_{t=1}^{n}\left(\Delta_t + 2\left(\Delta_t\varphi(1)+\Delta_t\varphi(2)+\ldots+\Delta_t\varphi(n)\right)\right)^2\\ &\le \frac{c^2}{n}\sum_{t=1}^{n}\left(1+2\left(\frac{1}{2}\rho+\frac{1}{2}\rho^2+\ldots+\frac{1}{2}\rho^n\right)\right)^2\\ &\le c^2\cdot\left(\frac{1-\rho^{n+1}}{1-\rho}\right)^2\\ &\le c^2\cdot\left(\frac{1-\left(\frac{1}{2}\right)^{(n+1)b(\pi_*)\gamma_{ps}}}{1-\left(\frac{1}{2}\right)^{b(\pi_*)\gamma_{ps}}}\right)^2 \qquad (23)\end{aligned}$$

where $b(\pi_*) = (\ln\frac{1}{\pi_*} + 2\ln 2 + 1)^{-1}$.

This brings us to the following thoerem.

Theorem C.2. *Assume $\{U_t\}_{t=1}^n$ be a stationary, ergodic, finite state Markov chain with pseudo-spectral gap $\gamma_{ps} > 0$, then for any constants $\lambda > 0$, $\delta \in (0,1)$, and prior $\mu \in \mathcal{P}(\Theta)$,*

$$\mathbb{P}_{\mathcal{S}}\left(\forall\rho\in\mathcal{P}(\Theta),\ \mathbb{E}_{\theta\sim\rho}\left[R(\theta)\right] \le \mathbb{E}_{\theta\sim\rho}\left[r(\theta)\right] + \frac{\lambda c^2}{8n}\left(\frac{1-\left(\frac{1}{2}\right)^{(n+1)b(\pi_*)\gamma_{ps}}}{1-\left(\frac{1}{2}\right)^{b(\pi_*)\gamma_{ps}}}\right)^2 + \frac{KL(\rho||\mu)+\log\frac{1}{\delta}}{\lambda}\right) \ge 1-\delta$$

where $b(\pi_*) = (\ln\frac{1}{\pi_*} + 2\ln 2 + 1)^{-1}$.

As before, one can substitute γ_{ps} with an empirical estimate $\widehat{\gamma}_{ps}$ to obtain an empirical bound.

It would be very nice to get an empirical version of the PAC-Bayes bounds with the φ coefficients without assuming the Markov property. However, this would require to estimate the φ-mixing coefficients, which is still an open question.