entropy

MDPI

Article

PAC-Bayes Bounds on Variational Tempered Posteriors for Markov Models

Imon Banerjee [1], Vinayak A. Rao [1,*,†] and Harsha Honnappa [2,†]

[1] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA; ibanerj@purdue.edu
[2] School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA; honnappa@purdue.edu
* Correspondence: varao@purdue.edu
† These authors contributed equally to this work.

Abstract: Datasets displaying temporal dependencies abound in science and engineering applications, with Markov models representing a simplified and popular view of the temporal dependence structure. In this paper, we consider Bayesian settings that place prior distributions over the parameters of the transition kernel of a Markov model, and seek to characterize the resulting, typically intractable, posterior distributions. We present a Probably Approximately Correct (PAC)-Bayesian analysis of variational Bayes (VB) approximations to tempered Bayesian posterior distributions, bounding the model risk of the VB approximations. Tempered posteriors are known to be robust to model misspecification, and their variational approximations do not suffer the usual problems of over confident approximations. Our results tie the risk bounds to the mixing and ergodic properties of the Markov data generating model. We illustrate the PAC-Bayes bounds through a number of example Markov models, and also consider the situation where the Markov model is misspecified.

Keywords: ergodicity; Markov chain; probably approximately correct; variational Bayes

Citation: Banerjee, I.; Rao, V. A.; Honnappa, H. PAC-Bayes Bounds on Variational Tempered Posteriors for Markov Models. *Entropy* **2021**, *23*, 313. https://doi.org/10.3390/e23030313

Academic Editor: Pierre Alquier

Received: 8 February 2021
Accepted: 4 March 2021
Published: 6 March 2021

1. Introduction

This paper presents probably approximately correct (PAC)-Bayesian bounds on variational Bayesian (VB) approximations of fractional or tempered posterior distributions for Markov data generation models. Exact computation of either standard or tempered posterior distributions is a hard problem that has, broadly speaking, spawned two classes of computational methods. The first, Markov chain Monte Carlo (MCMC), constructs ergodic Markov chains to approximately sample from the posterior distribution. MCMC is known to suffer from high variance and complex diagnostics, leading to the development of variational Bayesian (VB) [1] methods as an alternative in recent years. VB methods pose posterior computation as a variational optimization problem, approximating the posterior distribution of interest by the 'closest' element of an appropriately defined class of 'simple' probability measures. Typically, the measure of closeness used by VB methods is the Kullback–Leibler (KL) divergence. Excellent introductions to this so-called *KL-VB* method can be found in [2–4]. More recently, there has also been interest in alternative divergence measures, particularly the α-Rényi divergence [5–7], though in this paper, we focus on the KL-VB setting.

Theoretical properties of VB approximations, and in particular asymptotic frequentist consistency, have been studied extensively under the assumption of an independent and identically distributed (i.i.d.) data generation model [4,8,9]. On the other hand, the common setting where data sets display temporal dependencies presents unique challenges. In this paper, we focus on homogeneous Markov chains with parameterized transition kernels, representing a parsimonious class of data generation models with a wide range of applications. We work in the Bayesian framework, focusing on the posterior distribution over the unknown parameters of the transition kernel. Our theory develops PAC bounds

that link the ergodic and mixing properties of the data generating Markov chain to the Bayes risk associated with approximate posterior distributions.

Frequentist consistency of Bayesian methods, in the sense of concentration of the posterior distribution around neighborhoods of the 'true' data generating distribution, have been established in significant generality, in both the i.i.d. [10–12] and in the non-i.i.d. data generation setting [13,14]. More recent work [14–16] has studied fractional or tempered posteriors, a class of generalized Bayesian posteriors obtained by combining the likelihood function raised to a fractional power with an appropriate prior distribution using Bayes' theorem. Tempered posteriors are known to be robust against model misspecification: in the Markov setting we consider, the associated stationary distribution as well as mixing properties are sensitive to model parameterization. Further, tempered posteriors are known to be much simpler to analyze theoretically [14,16]. Therefore, following [14–16] we focus on tempered posterior distributions on the transition kernel parameters, and study the rate of concentration of variational approximations to the tempered posterior. Equivalently, as shown in [16] and discussed in Section 1.1, our results also apply to so-called α-variational approximations to standard posterior distributions over kernel parameters. The latter are modifications of the standard KL-VB algorithm to address the well-known problem of overconfident posterior approximations.

While there have been a number of recent papers studying the consistency of approximate variational posteriors [5,8,15] in the large sample limit, rates of convergence have received less attention. Exceptions include [9,15,17], where an i.i.d. data generation model is assumed. [15] establishes PAC-Bayes bounds on the convergence of a variational tempered posterior with fractional powers in the range $[0, 1)$, while [9] considers the standard variational posterior case (where the fractional power equals 1). [17], on the other hand, establishes PAC-Bayes bounds for risk-sensitive Bayesian decision making problems in the standard variational posterior setting. The setting in [15] allows for model misspecification and the analysis is generally more straightforward than that in [9,17]. Our work extends [15] to the setting of a discrete-time Markov data generation model.

Our first results in Theorem 1 and Corollary 1 of Section 2 establish PAC-Bayes bounds for sequences with arbitrary temporal dependence. Our resultsgeneralize [15], [Theorem 2.4] to the non-i.i.d. data setting in a straightforward manner. Note that Theorem 1 also recovers ([16], [Theorem 3.3]), which is established under different 'existence of test' conditions. Our objective in this paper is to explicate how the ergodic and mixing properties of the Markov data generating process influences the PAC-Bayes bound. The sufficient conditions of our theorem, bounding the mean and variance of the log-likelihood ratio of the data, allows for developing this understanding, without the technicalities of proving the existence of test conditions intruding on the insights.

In Section 3, we study the setting where the data generating model is a stationary α-mixing Markov chain. Stationarity means that the Markov chain is initialized with the invariant distribution corresponding to the parameterized transition kernel, implying all subsequent states also follow this marginal distribution. The α-mixing condition ensures that the variance of the likelihood ratio of the Markov data does not grow faster than linear in the sample size. Our main results in this setting are applicable when the state space of the Markov chain is either continuous or discrete. The primary requirement on the class of data generating Markov models is for the log-likelihood ratio of the parameterized transition kernel and invariant distribution to satisfy a Lipschitz property. This condition implies a decoupling between the model parameters and the random samples, affording a straightforward verification of the mean and variance bounds. We highlight this main result by demonstrating that it is satisfied by a finite state Markov chain, a birth-death Markov chain on the positive integers, and a one-dimensional Gaussian linear model.

In practice, the assumption that the data generating model is stationary is unlikely to be satisfied. Typically, the initial distribution is arbitrary, with the state distribution of the Markov sequence converging weakly to the stationary distribution. In this setting, we must further assume that the class of data generating Markov chains are geometrically ergodic.

We show that this implies the boundedness of the mean and variance of the log-likelihood ratio of the data generating Markov chain. Alternatively, in Theorem 4 we directly impose a drift condition on random variables that bound the log-likelihood ratio. Again, in this more general nonstationary setting, we illustrate the main results by showing that the PAC-Bayes bound is satisfied by a finite state Markov chain, a birth-death Markov chain on the positive integers, and a one-dimensional Gaussian linear model.

In preparation for our main technical results starting in Section 2 we first note relevant notations and definitions in the next section.

1.1. Notations and Definitions

We broadly adopt the notation in [15]. Let the sequence of random variables $X^n = (X_0, \ldots, X_n) \subset \mathbb{R}^{m\times(n+1)}$ represent a dataset of $n+1$ observations drawn from a joint distribution $P_{\theta_0}^{(n)}$, where $\theta_0 \in \Theta \subseteq \mathbb{R}^d$ is the 'true' parameter underlying the data generation process. We assume the state space $S \subseteq \mathbb{R}^m$ of the random variables X_i is either discrete-valued or continuous, and write $\{x_0, \ldots, x_n\}$ for a realization of the dataset. We also adopt the convention that $0\log(0/0) = 0$.

For each $\theta \in \Theta$, we will write $p_\theta^{(n)}$ as the probability density of $P_\theta^{(n)}$ with respect to some measure $Q^{(n)}$, i.e., $p_\theta^{(n)} := \frac{dP_\theta^{(n)}}{dQ^{(n)}}$, where $Q^{(n)}$ is either Lebesgue measure or the counting measure. Unless stated otherwise, all probabilities, expectations and variances, which we represent as P, $\mathrm{E}[X]$ and $\mathrm{Var}[X]$, are with respect to the true distribution $P_{\theta_0}^{(n)}$.

Let $\pi(\theta)$ be a *prior* distribution with support Θ. The α^{te}*-fractional posterior* is defined as

$$\pi_{n,\alpha^{te}|X^n}(d\theta) := \frac{e^{-\alpha^{te} r_n(\theta,\theta_0)(X^n)}\pi(d\theta)}{\int e^{-\alpha^{te} r_n(\theta,\theta_0)(X^n)}\pi(d\theta)}, \tag{1}$$

where, for $\theta_0, \theta \in \Theta$, $r_n(\theta,\theta_0)(\cdot) := \log\left(\frac{p_{\theta_0}^{(n)}(\cdot)}{p_\theta^{(n)}(\cdot)}\right)$, is the log-likelihood ratio of the corresponding density functions, and $\alpha^{te} \in (0,\infty)$ is a tempering coefficient. Setting $\alpha^{te} = 1$ recovers the standard Bayesian posterior. Note that we will use superscripts to distinguish different quantities that are referred to just as α in the literature.

The *Kullback–Leibler* (KL) divergence between distributions P, Q is defined as

$$\mathcal{K}(P,Q) := \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) p(x)dx,$$

where p, q are the densities corresponding to P, Q on some sample space $\mathcal{X}$. In particular, the KL divergence between the distributions parameterized by θ_0 and θ is

$$\begin{aligned}\mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)}) &:= \int \log\left(\frac{p_{\theta_0}^{(n)}(x_0,\ldots,x_n)}{p_\theta^{(n)}(x_0,\ldots,x_n)}\right) p_{\theta_0}^{(n)}(x_0,\ldots,x_n)dx_0\cdots dx_n \\ &= \int r_n(\theta,\theta_0)(x_0,\ldots,x_n)p_{\theta_0}^n(x_0,\ldots,x_n)dx_0\cdots dx_n. \end{aligned} \tag{2}$$

The α^{re}*-Rényi divergence* $D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})$ is defined as

$$D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)}) := \frac{1}{\alpha^{re}-1}\log\int \exp(-\alpha^{re} r_n(\theta,\theta_0)(x_0,\ldots,x_n))p_{\theta_0}^{(n)}(x_0,\ldots,x_n)dx_0\cdots dx_n, \tag{3}$$

where $\alpha^{re} \in (0,1)$. As $\alpha^{re} \to 1$, the α^{re}-Rényi divergence recovers the KL divergence.

Let $\mathcal{F}$ be some class of distributions with support in $\mathbb{R}^d$ and such that any distribution P in $\mathcal{F}$ is absolutely continuous with respect to the tempered posterior: $P \ll \pi_{n,\alpha^{te}|X^n}$.

Many choices of $\mathcal{F}$ exist; for instance (see also [15]), $\mathcal{F}$ can be the set of Gaussian measures, denoted $\mathcal{F}_{id}^{\Phi}$:

$$\mathcal{F}_{id}^{\Phi} = \{\Phi(d\theta; \mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma_{d\times d} \in \text{P.D.}\}, \tag{4}$$

where P.D. references the class of positive definite matrices. Alternately, $\mathcal{F}$ can be the family of *mean-field* or factored distributions where the components θ_i of θ are independent of each other. Let $\tilde{\pi}_{n,\alpha^{te}|X^n}$ be the variational approximation to the tempered posterior, defined as

$$\tilde{\pi}_{n,\alpha^{te}|X^n} := \underset{\rho\in\mathcal{F}}{\arg\min}\ \mathcal{K}(\rho, \pi_{n,\alpha^{te}|X^n}) \tag{5}$$

It is easy to see that finding $\tilde{\pi}_{n,\alpha^{te}|X^n}$ in Equation (5) is equivalent to the following optimization problem:

$$\tilde{\pi}_{n,\alpha^{te}|X^n} := \underset{\rho\in\mathcal{F}}{\arg\max}\left[\int r_n(\theta,\theta_0)(x_0,\ldots,x_n)\rho(d\theta) - \left(\alpha^{te}\right)^{-1}\mathcal{K}(\rho,\pi)\right]. \tag{6}$$

Setting $\alpha^{te} = 1$ again recovers the usual variational solution that seeks to approximate the posterior distribution with the closest element of $\mathcal{F}$ (the right-hand side above is called the evidence lower bound (ELBO)). Other settings of α^{te} constitute α^{te}-variational inference [16], which seeks to regularize the 'overconfident' approximate posteriors that standard variational methods tend to produce.

Our results in this paper focus on parametrized Markov chains. We term a Markov chain as 'parameterized' if the transition kernel $p_\theta(\cdot|\cdot)$ is parametrized by some $\theta \in \Theta \subseteq \mathbb{R}^d$. Let $q^{(0)}(\cdot)$ be the initial density (defined with respect to the Lebesgue measure over $\mathbb{R}^m$) or initial probability mass function. Then, the joint density is $p_\theta^{(n)}(x_0,\ldots,x_n) = q^{(0)}(x_0)\prod_{i=0}^{n-1} p_\theta(x_{i+1}|x_i)$; recall, this joint density $p_\theta^{(n)}(x_0,\ldots,x_n)$ corresponds to the walk probability of a time-homogeneous Markov chain. We assume that corresponding to each transition kernel p_θ, $\theta \in \Theta$, there exists an invariant distribution $q_\theta^{(\infty)} \equiv q_\theta$ that satisfies

$$q_\theta(x) = \int p_\theta(x|y)q_\theta(dy) \quad \forall x \in \mathbb{R}^m, \theta \in \Theta.$$

We will also use q_θ to designate the density of the invariant measure (as before, this is with respect to the Lebesgue or counting measure for continuous or discrete state spaces, respectively). A Markov chain is stationary if its initial distribution is the invariant probability distribution, that is, $X_0 \sim q_\theta$.

Our results in the ensuing sections will be established under strong mixing conditions [18] on the Markov chain. Specifically, recall the definition of the α-mixing coefficients of a Markov chain $\{X_n\}$:

Definition 1 (α-mixing coefficient)**.** *Let $\mathcal{M}_i^j$ denote the σ-field generated by the Markov chain $\{X_k : i \le k \le j\}$ parameterized by $\theta \in \Theta$. Then, the α-mixing coefficient is defined as*

$$\alpha_k = \sup_{t>0}\ \sup_{(A,B)\in\mathcal{M}_{-\infty}^t\times\mathcal{M}_{t+k}^{\infty}} |P_\theta(A\cap B) - P_\theta(A)P_\theta(B)|. \tag{7}$$

Informally speaking, the α-mixing coefficients $\{\alpha_k\}$ measure the dependence between any two events A (in the 'history' σ-algebra) and B (in the 'future' σ-algebra) with a time lag k. We note that we do not use superscripts to identify these α parameters, since they are the only ones with subscripts, and can be identified through this.

2. A Concentration Bound for the α^{re}-Rényi Divergence

The object of analysis in what follows is the probability measure $\tilde{\pi}_{n,\alpha^{te}|X^n}(\theta)$, the variational approximation to the tempered posterior. Our main result establishes a bound on the Bayes risk of this distribution; in particular, given a sequence of loss functions $\ell_n(\theta,\theta_0)$, we bound $\int \ell_n(\theta,\theta_0)\tilde{\pi}_{n,\alpha^{te}|X^n}(\theta)d\theta$. Following recent work in both the i.i.d. and dependent sequence settings [14–16], we will use $\ell_n(\theta,\theta_0) = D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})$, the α^{re}-Rényi divergence between $P_\theta^{(n)}$ and $P_{\theta_0}^{(n)}$ as our loss function. Unlike loss functions like Euclidean distance, Rényi divergence compares θ and θ_0 through their effect on observed sequences, so that issues like parameter identifiability no longer arise. Our first result generalizes [15], [Theorem 2.1] to a general non-i.i.d. data setting.

Proposition 1. *Let $\mathcal{F}$ be a subset of all probability distributions on Θ. For any $\alpha^{re} \in (0,1)$, $\epsilon \in (0,1)$ and $n \geq 1$, the following probabilistic uniform upper bound on the expected α^{re}-Rényi divergence holds:*

$$P\left[\sup_{\rho\in\mathcal{F}} \int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\rho(d\theta) \leq \frac{\alpha^{re}}{1-\alpha^{re}} \int r_n(\theta,\theta_0)\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi) + \log(\frac{1}{\epsilon})}{1-\alpha^{re}}\right] \geq 1-\epsilon. \tag{8}$$

The proof of Proposition 1 follows easily from [15], and we include it in Appendix B.1.1 for completeness. Mirroring the comments in [15], when $\rho = \tilde{\pi}_{n,\alpha^{te}}$ this result is precisely [14, Theorem 3.4]. We also note from [14] that $\forall\ \alpha^{re}, \beta \in (0,1]$ α^{re}-Rényi divergences are all equivalent through the following inequality $\frac{\alpha^{re}(1-\beta)}{\beta(1-\alpha^{re})} D_\beta \leq D_{\alpha^{re}} \leq D_\beta\ \forall\ \alpha^{re} \leq \beta$. Hence, for the subsequent results, we simplify by assuming that $\alpha^{te} = \alpha^{re}$. This probabilistic bound implies the following PAC-Bayesian concentration bound on the model risk computed with respect to the fractional variational posterior:

Theorem 1. *Let $\mathcal{F}$ be a subset of all probability distributions parameterized by Θ, and assume there exist $\epsilon_n > 0$ and $\rho_n \in \mathcal{F}$ such that*

i. $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) = \int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq n\epsilon_n$,
ii. $\int \mathrm{Var}(r_n(\theta,\theta_0))\rho_n(d\theta) \leq n\epsilon_n$, *and*
iii. $\mathcal{K}(\rho_n,\pi) \leq n\epsilon_n$.

Then, for any $\alpha^{re} \in (0,1)$ and $(\epsilon,\eta) \in (0,1)\times(0,1)$,

$$P\left[\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}}(d\theta|X^{(n)}) \leq \frac{(\alpha^{re}+1)n\epsilon_n + \alpha^{re}\sqrt{\frac{n\epsilon_n}{\eta}} - \log(\epsilon)}{1-\alpha^{re}}\right] \geq 1-\epsilon-\eta. \tag{9}$$

The proof of Theorem 1 is a generalization of [15] (Theorem 2.4) to the non-i.i.d. setting, and a special case of [16] (Theorem 3.1), where the problem setting includes latent variables. We include a proof for completeness. As noted in [15], the sufficient conditions follow closely from [13] and we will show that they hold for a variety of Markov chain models.

A direct corollary of Theorem 1 follows by setting $\eta = \frac{1}{n\epsilon_n}$, $\epsilon = e^{-n\epsilon_n}$ and using the fact that $e^{-n\epsilon_n} \geq \frac{1}{n\epsilon_n}$. Note that Equation (9) is vacuous if $\eta + \epsilon > 1$. Therefore, without loss of generality, we restrict ourselves to the condition $\frac{2}{n\epsilon_n} < 1$.

Corollary 1. *Assume $\exists\ \epsilon_n > 0$, $\rho_n \in \mathcal{F}$ such that the following conditions hold:*

i. $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) = \int E[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq n\epsilon_n$,
ii. $\int \mathrm{Var}(r_n(\theta,\theta_0))\rho_n(d\theta) \leq n\epsilon_n$, *and*
iii. $\mathcal{K}(\rho_n,\pi) \leq n\epsilon_n$.

Then, for any $\alpha^{re} \in (0,1)$,

$$P\left[\int D_{\alpha^{re}}(P_{\theta}^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}}(d\theta|X^{(n)}) \leq \frac{2(\alpha^{re}+1)\epsilon_n}{1-\alpha^{re}}\right] \geq 1 - \frac{2}{n\epsilon_n}. \tag{10}$$

We observe that Theorem 1 and Corollary 1 place no assumptions on the nature of the statistical dependence between data points. However, verification of the sufficient conditions is quite hard, in general. One of our key contributions is to verify that under reasonable assumptions on the smoothness of the transition kernel, the sufficient conditions of Theorem 1 and Corollary 1 are satisfied by ergodic Markov chains.

Observe that the first two conditions in Corollary 1 ensure that the distribution ρ_n concentrates on parameters $\theta \in \Theta$ around the true parameter θ_0, while the third condition requires that ρ_n not diverge from the prior π rapidly as a function of the sample size n. In general, verifying the first and third conditions is relatively straightforward. The second condition, on the other hand, is significantly more complicated in the current setting of dependent data, as the variance of $r_n(\theta, \theta_0)$ includes correlations between the observations $\{X_0, \ldots, X_n\}$. In the next section, we will make assumptions on the transition kernels (and corresponding invariant densities) that 'decouple' the temporal correlations and the model parameters in the setting of strongly mixing and ergodic Markov chain models, and allow for the verification of the conditions in Corollary 1. Towards this, Propositions 2 and 3 below characterize the expectation and variance of the log-likelihood ratio $r_n(\cdot,\cdot)$ in terms of the one-step transition kernels of the Markov chain. First, consider the expectation of $r_n(\cdot,\cdot)$ in condition (i).

Proposition 2. *Fix* $\theta_1, \theta_2 \in \Theta$ *and consider the parameterized Markov transition kernels* p_{θ_1} *and* p_{θ_2}*, and initial distributions* $q_{\theta_1}^{(0)}$ *and* $q_{\theta_2}^{(0)}$*. Let* $p_{\theta_1}^{(n)}$ *and* $p_{\theta_2}^{(n)}$ *be the corresponding joint probability densities; that is,*

$$p_{\theta_j}^{(n)}(x_0, \ldots, x_n) = q_{\theta_j}^{(0)}(x_0)\prod_{i=1}^{n} p_{\theta_i}(x_i|x_{i-1}) \tag{11}$$

for $j \in \{1,2\}$*. Then, for any* $n \geq 1$*, the log-likelihood ratio* $r_n(\theta_2, \theta_1)$ *satisfies*

$$\mathrm{E}_{\theta_1}[r_n(\theta_2,\theta_1)] = \sum_{i=1}^{n} \mathrm{E}_{\theta_1}\left[\log\left(\frac{p_{\theta_1}(X_i|X_{i-1})}{p_{\theta_2}(X_i|X_{i-1})}\right)\right] + \mathrm{E}_{\theta_1}[Z_0], \tag{12}$$

where $Z_0 := \log\left(\frac{q_{\theta_1}^{(0)}(X_0)}{q_{\theta_2}^{(0)}(X_0)}\right)$*. The expectation in the first term is with respect to the joint density function* $p_{\theta_1}(y,x) = p_{\theta_1}(y|x)q_{\theta_1}^{(i-1)}(x)$ *where the marginal density satisfies*

$$q_{\theta_1}^{(i-1)}(x) = \begin{cases} \int p_{\theta_1}^{(i-1)}(x_0, \ldots, x_{i-2}, x)dx_0 \cdots dx_{i-2} & \text{for } i > 1, \text{and} \\ q_{\theta_1}^{(0)}(x) & \text{for } i = 1. \end{cases}$$

If the Markov chain is also stationary under θ_1*, then Equation (12) simplifies to*

$$\mathrm{E}_{\theta_1}[r_n(\theta_2,\theta_1)] = n\mathrm{E}_{\theta_1}\left[\log\left(\frac{p_{\theta_1}(X_1|X_0)}{p_{\theta_2}(X_1|X_0)}\right)\right] + \mathrm{E}_{\theta_1}[Z_0]. \tag{13}$$

Notice that $\mathrm{E}_{\theta_1}[r_n(\theta_2,\theta_1)]$ is precisely the KL divergence, $\mathcal{K}(P_{\theta_1}^{(n)}, P_{\theta_2}^{(n)})$. Next, the following proposition uses [19] (Lemma 1.3) to upper bound the variance of the log-likelihood ratio.

Proposition 3. *Fix* $\theta_1, \theta_2 \in \Theta$ *and consider parameterized Markov transition kernels* p_{θ_1} *and* p_{θ_2}*, with initial distributions* $q_{\theta_1}^{(0)}$ *and* $q_{\theta_2}^{(0)}$*. Let* $p_{\theta_1}^{(n)}$ *and* $p_{\theta_2}^{(n)}$ *be the corresponding joint*

probability densities of the sequence $(x_0, \ldots, x_n)$, *and* $q_{\theta_j}^{(i)}$ *the marginal density for* $i \in \{1, \ldots, n\}$ *and* $j \in \{1, 2\}$. *Fix* $\delta > 0$ *and, for each* $i \in \{1, \ldots, n\}$, *define*

$$C_{\theta_1,\theta_2}^{(i)} := \int \left| \log\left(\frac{p_{\theta_1}(x_i|x_{i-1})}{p_{\theta_2}(x_i|x_{i-1})} \right) \right|^{2+\delta} p_{\theta_1}(x_i|x_{i-1}) q_{\theta_1}^{(i-1)}(x_{i-1}) dx_i dx_{i-1}.$$

Similarly, define $Z_0 := \log\left(\frac{q_{\theta_1}^{(0)}(X_0)}{q_{\theta_2}^{(0)}(X_0)} \right)$, *and* $D_{1,2} := \mathrm{E}_{\theta_1}|Z_0|^{2+\delta}$. *Suppose the Markov chain corresponding to* θ_1 *is* α*-mixing with coefficients* $\{\alpha_k\}$. *Then,*

$$\begin{aligned} \mathrm{Var}(r_n(\theta_1, \theta_2)) < & \sum_{i,j=1}^{n} \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + C_{\theta_1,\theta_2}^{(j)} + \sqrt{C_{\theta_1,\theta_2}^{(i)} C_{\theta_1,\theta_2}^{(j)}}) \right) \left(\alpha_{|i-j|-1}^{\delta/(2+\delta)} \right) \\ & + \sum_{i=1}^{n} \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + D_{1,2} + \sqrt{C_{\theta_1,\theta_2}^{(i)} D_{1,2}}) \right) \left(\alpha_{i-1}^{\delta/(2+\delta)} \right) \qquad (14) \\ & + \mathrm{Cov}(Z_0, Z_0). \qquad (15) \end{aligned}$$

Note that this result holds for any parameterized Markov chain. In particular, when the Markov chain is stationary, $C_{\theta_1,\theta_2}^{(i)} = C_{\theta_1,\theta_2}^{(1)} \ \forall \ i$ and $\forall \theta \in \Theta$, and Equation (14) simplifies to

$$\begin{aligned} \mathrm{Var}(r_n(\theta_1, \theta_2)) < & \, n \left(\frac{4}{n} + 6n^{\delta/2} C_{\theta_1,\theta_2}^{(1)} \right) \left(\sum_{k \geq 0} \alpha_k^{\delta/(2+\delta)} \right) \\ & + \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(1)} + D_{1,2} + \sqrt{C_{\theta_1,\theta_2}^{(1)} D_{1,2}}) \right) \left(\sum_{k \geq 1} \alpha_k^{\delta/(2+\delta)} \right) \\ & + \mathrm{Cov}(Z_0, Z_0). \qquad (16) \end{aligned}$$

If the sum $\sum_{k \geq 0} \alpha_k^{\delta/(2+\delta)}$ is infinite, the bound is trivially true. For it to be finite, of course, the coefficients α_k must decay to zero sufficiently quickly. For instance, Theorem A.1.2 shows that if the Markov chain is geometrically ergodic, then the α-mixing coefficients are geometrically decreasing. We will use this fact when the Markov chain is non-stationary, as in Section 4. In the next section, however, we first consider the simpler stationary Markov chain setting where geometric ergodic conditions are not explicitly imposed. We also note that unless only a finite number of α_k are nonzero, the sum $\sum_{k \geq 0} \alpha_k^{\delta/(2+\delta)}$ is infinite when $\delta = 0$, and our results will typically require $\delta > 0$.

3. Stationary Markov Data-Generating Models

Observe that the PAC-Bayesian concentration bound in Corollary 1 specifically requires bounding the mean and variance of the log-likelihood ratio $r_n(\theta, \theta_0)$. We ensure this by imposing regularity conditions on the log-ratio of the one-step transition kernels and the corresponding invariant densities. Specifically, we assume the following conditions that decouple the model parameters from the random samples, allowing us to verify the bounds in Corollary 1.

Assumption 1. *There exist positive functions* $M_k^{(1)}(\cdot, \cdot)$ *and* $M_k^{(2)}(\cdot)$, $k \in \{1, 2, \ldots, m\}$ *such that for any parameters* $\theta_1, \theta_2 \in \Theta$, *the log of the ratio of one-step transition kernels and the log of the ratio of the invariant distributions satisfy, respectively,*

$$|\log p_{\theta_1}(x_1|x_0) - \log p_{\theta_2}(x_1|x_0)| \leq \sum_{k=1}^{m} M_k^{(1)}(x_1, x_0) |f_k^{(1)}(\theta_2, \theta_1)| \ \forall \ (x_0, x_1), \text{ and} \qquad (17)$$

$$|\log q_{\theta_1}(x) - \log q_{\theta_2}(x)| \leq \sum_{k=1}^{m} M_k^{(2)}(x) |f_k^{(2)}(\theta_2, \theta_1)| \ \ \forall \ \ x. \tag{18}$$

We further assume that for some $\delta > 0$, the functions $f_k^{(1)}$, $f_k^{(2)}$ and $M_k^{(1)}$ satisfy the following:

i. there exist constants $C_k^{(t)}$ and measures $\rho_n \in \mathcal{F}$ such that $\int |f_k^{(t)}(\theta, \theta_0)|^{2+\delta} \rho_n(d\theta) < \frac{C_k^{(t)}}{n}$ for $t \in \{1, 2\}$, $n \geq 1$ and $k \in \{1, 2, \ldots, m\}$, and

ii. there exists a constant B such that $\int M_k^{(1)}(x_1, x_0)^{2+\delta} p_{\theta_j}(x_1|x_0) q_{\theta_j}^{(0)}(x_0) dx_1 dx_0 < B$, $k \in \{1, \ldots, m\}$ and $j \in \{1, 2\}$.

The following examples illustrate Equations (17) and (18) for discrete and continuous state Markov chains.

Example 1. *Suppose $\{X_0, \ldots, X_n\}$ is generated by the birth-death chain with parameterized transition probability mass function,*

$$p_\theta(j|i) = \begin{cases} \theta & \text{if } j = i - 1, \\ 1 - \theta & \text{if } j = i + 1. \end{cases}$$

In this example, the parameter θ denotes the probability of birth. We shall see that, $m = 3$: $M_1^{(1)}(X_1, X_0) = I_{[X_1 = X_0 + 1]}$, $M_2^{(1)}(X_1, X_0) = I_{[X_1 = X_0 - 1]}$, and $M_3^{(1)}(X_1, X_0) = 1$. We also define $M_1^{(2)}(X_0) = 1$, and set $M_2^{(2)}(X_0)$ and $M_3^{(2)}(X_0)$ both to $X_0 - 1$. Let $f_1^{(1)}(\theta, \theta_0) = \log\left[\frac{\theta_0}{\theta}\right]$, $f_2^{(1)}(\theta, \theta_0) = \log\left[\frac{1-\theta_0}{1-\theta}\right]$, $f_3^{(1)}(\theta, \theta_0) = 0$, $f_1^{(2)}(\theta, \theta_0) = -f_3^{(2)}(\theta, \theta_0) = \log\left[\frac{1-\theta_0}{1-\theta}\right]$, and $f_2^{(2)}(\theta, \theta_0) = \log\left[\frac{\theta_0}{\theta}\right]$. The derivation of these terms and that they satisfy the conditions of Assumption 1 is provided in the proof of Proposition 6.

Example 2. *Suppose $\{X_0, \ldots, X_n\}$ is generated by the 'simple linear' Gauss–Markov model*

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

where $\{W_n\}$ is a sequence of i.i.d. standard Gaussian random variables. Then, $m = 2$, with $M_1^{(1)}(X_n, X_{n-1}) = |X_n X_{n-1}|$, $M_2^{(1)}(X_n, X_{n-1}) = X_n^2$, $M_1^{(2)}(x) = \frac{x^2}{2}$ and $M_2^{(2)}(X) = 0$. Corresponding to these, we have $f_1^{(1)}(\theta, \theta_0) = (\theta - \theta_0)$, $f_2^{(1)}(\theta, \theta_0) = (\theta_0^2 - \theta^2)$, $f_1^{(2)}(\theta_0, \theta_0) = (\theta_0^2 - \theta^2)$ and $f_2^{(2)}(\theta_0, \theta_0) = 0$. The derivation of these quantities and that these satisfy the conditions of Assumption 1 under appropriate choice of ρ_n is shown in the proof of Proposition 10.

Note that assuming the same number m of $M_k^{(1)}$ and $M_k^{(2)}$ involves no loss of generality, since these functions can be set to 0. Both Equations (17) and (18) can be viewed as generalized Lipschitz-smoothness conditions, recovering the usual Lipschitz-smoothness when $m = 1$ and when $f_k^{(t)}$ is Euclidean distance. Our generalized conditions are useful for distributions like the Gaussian, where Lipschitz smoothness does not apply. From Jensen's inequality we have $\int |f_k^{(t)}(\theta, \theta_0)| \rho_n(d\theta)| \leq \left[\int |f_k^{(t)}(\theta, \theta_0)|^{2+\delta} \rho_n(d\theta)\right]^{\frac{1}{2+\delta}}$, and Assumption 1(i) above implies that for some constant $C > 0$ and $k \in \{1, 2, \ldots, m\}$, $t \in \{1, 2\}$,

$$\int |f_k^{(t)}(\theta, \theta_0)| \rho_n(d\theta) \leq \frac{C}{n^{1/(2+\delta)}} < \frac{C}{\sqrt{n}}. \tag{19}$$

Assumption 1(i) is satisfied in a variety of scenarios, for example, under mild assumptions on the partial derivatives of the functions $f_k^{(t)}$. To illustrate this, we present the following proposition.

Proposition 4. *Let $f(\theta, \theta_0)$ be a function on a bounded domain with bounded partial derivatives with $f(\theta_0, \theta_0) = 0$. Let $\{\rho_n(\cdot)\}$ be a sequence of probability densities on θ such that $\mathrm{E}_{\rho_n}[\theta] = \theta_0$ and $\mathrm{Var}_{\rho_n}[\theta] = \frac{\sigma^2}{n}$ for some $\sigma > 0$. Then, for some $C > 0$,*

$$\int |f(\theta, \theta_0)|^{2+\delta} \rho_n(d\theta) < \frac{C}{n}. \tag{20}$$

Proof. Define $\partial_\theta f(\theta, \theta_0) := \frac{\partial f(\theta, \theta_0)}{\partial \theta}$ as the partial derivative of the function f. By the mean value theorem, $|f(\theta, \theta_0)| = |\theta - \theta_0||\partial_\theta f(\theta^*, \theta_0)|$, for some $\theta^* \in [\min\{\theta, \theta_0\}, \max\{\theta, \theta_0\}]$. Since the partial derivatives are bounded, there exists $L \in \mathbb{R}$ such that $\partial_\theta f(\theta^*, \theta_0) < L$, and $\int |f(\theta, \theta_0)|^{2+\delta} \rho_n(d\theta) < L^{2+\delta} \int |\theta - \theta_0|^{2+\delta} \rho_n(d\theta)$. Choose $G > 0$ be such that $|\theta| < G$, then $\left|\frac{\theta - \theta_0}{2G}\right|^{2+\delta} < \left|\frac{\theta - \theta_0}{2G}\right|^2$. Therefore, $\int |\theta - \theta_0|^{2+\delta} \rho_n(d\theta) < (2G)^{2+\delta} \mathrm{Var}\left[\frac{\theta}{2G}\right] < (2G)^\delta \frac{\sigma^2}{n}$. Now choosing $(2G)^\delta \sigma^2$ as C completes the proof. □

If $\partial_\theta f_k^{(t)}$ is continuous and Θ is compact, then $\partial_\theta f_k^{(t)}$ is always bounded. Furthermore, observe that if $\mathrm{E}\left[M_k^{(1)}(X_1, X_0)^{2+\delta}\right] < B$, without loss of generality we can use Jensen's inequality to conclude that, for all $0 < a < 2 + \delta$, $\mathrm{E}\left[M_k^{(1)}(X_1, X_0)^a\right] < B^{\frac{a}{2+\delta}} < B$.

We can now state the main theorem of this section.

Theorem 2. *Let $\{X_0, \ldots, X_n\}$ be generated by a stationary, α-mixing Markov chain parametrized by $\theta_0 \in \Theta$. Suppose that Assumption 1 holds and that the α-mixing coefficients satisfy $\sum_{k \geq 1} \alpha_k^{\delta/(2+\delta)} < +\infty$. Furthermore, assume that $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n} C$ for some constant $C > 0$. Then, the conditions of Corollary 1 are satisfied with $\epsilon_n = \mathrm{O}\left(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\right)$.*

Theorem 2 is satisfied by a large class of Markov chains, including chains with countable and continuous state spaces. In particular, if the Markov chain is geometrically ergodic, then it follows from Equation (A4) (in the appendix) that $\sum_{k \geq 1} \alpha_k^{\delta/(2+\delta)} < +\infty$. Observe that in order to achieve $O(\frac{1}{\sqrt{n}})$ convergence, we need $\delta \leq 1$. Key to the proof of Theorem 2 is the fact that the variance of the log-likelihood ratio can be controlled via the application of Assumption 1 and Proposition 3. Note also that as δ decreases, satisfying the condition $\sum_{k \geq 1} \alpha_k^{\delta/(2+\delta)}$ requires the Markov chain to be faster mixing.

We now illustrate Theorem 2 for a number of Markov chain models. First, consider a birth-death Markov chain on a finite state space.

Proposition 5. *Suppose the data-generating process is a birth-death Markov chain, with one-step transition kernel parametrized by the birth probability $\theta_0 \in \Theta$. Let $\mathcal{F}$ be the set of all Beta distributions. We choose the prior to be a Beta distribution. Then, the conditions of Theorem 2 are satisfied and $\epsilon_n = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$.*

Proof. The proof of Proposition 5 follows from the more general Proposition 8, by fixing the initial distribution to the invariant distribution under θ_0. Therefore it has been omitted. We simply refer to the proof of Proposition 8 under a more general setup in Appendix B.3. □

The birth-death chain on the finite state space is, of course, geometrically ergodic and the α-mixing coefficients α_k decay geometrically. Note that the invariant distribution of this Markov chain is uniform over the state space, and consequently this is a particularly simple example. A more complicated and more realistic example is a birth-death Markov chain on the nonnegative integers. We note that if the probability of birth θ in a birth-death Markov chain on positive integers is greater than 0.5, then the Markov chain is transient, and consequently, not ergodic. Hence, our prior should be chosen to have support within $(0, 0.5)$. For that purpose, we define the class of scaled beta distributions.

Definition 2 (Scaled Beta). *If X is a beta distribution on with parameters a and b, then Y is said to be a scaled beta distribution with same parameters on the interval $(c, m+c)$ if*

$$Y = mx + c \;\; ; \;\; (m, c) \in \mathbb{R}^2$$

and in that case, the pdf of Y is obtained as

$$f(y) = \begin{cases} \frac{1}{m\text{Beta}(a,b)} \left(\frac{y-c}{m}\right)^{a-1} \left(1 - \frac{y-c}{m}\right)^{b-1} & \text{if } y \in (c, m+c), \\ 0 & \text{otherwise.} \end{cases}.$$

Here, $\mathrm{E}[Y] = m\frac{a}{a+b} + c$ and $\mathrm{Var}[Y] = m^2 \frac{ab}{(a+b)^2(a+b+1)}$. For the birth-death chain, we set $m = 0.5$ and $c = 0$ giving it support on $(0, \frac{1}{2})$. Setting $m = 2$ and $c = -1$ gives a beta distribution rescaled to have support on $(-1, 1)$.

Proposition 6. *Suppose the data-generating process is a positive recurrent birth-death Markov chain on the positive integers parameterized by the birth probability $\theta_0 \in (0, \frac{1}{2})$. Further let $\mathcal{F}$ be the set of all Beta distributions rescaled to have support $(0, \frac{1}{2})$. We choose the prior to be a scaled Beta distribution on $(0, 1/2)$ with parameters a and b. Then, the conditions of Theorem 2 are satisfied with $\epsilon_n = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$.*

Proof. The proof of Proposition 6 (for the stationary case) follows from the more general Proposition 9 (the nonstationary case) by fixing the initial distribution to the invariant distribution under θ_0. We omit the proof and simply refer to the proof of Proposition 9 under a more general setup in Appendix B.3. □

Unlike with the finite state-space, the invariant distribution now depends on the parameter $\theta \in \Theta$, and verification of the conditions of the proposition is more involved. In Appendix A.2, we prove that the class of scaled beta distributions satisfy the condition $\mathcal{K}(\rho_n, \pi) \leq n\epsilon_n$ when the prior π is a beta or an uniform distribution. This fact will allow us to prove the above propositions.

Both Proposition 5 and Proposition 6 assume a discrete state space. The next example considers a strictly stationary simple linear model (as defined in Example 2), which has a continuous, unbounded state space.

Proposition 7. *Suppose the data-generating model is a stationary simple linear model:*

$$X_n = \theta_0 X_{n-1} + W_n, \tag{21}$$

where $\{W_n\}$ are i.i.d. standard Gaussian random variables and $|\theta_0| < 1$. Suppose that $\mathcal{F}$ is the class of all beta distributions rescaled to have the support $(-1, 1)$. Then, the conditions of Theorem 2 are satisfied with $\epsilon_n = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$.

Proof. This is a special case of the more general non-stationary simple linear model which is detailed in Proposition 10. Therefore, the proof of the fact that the simple linear model satisfies Assumption 1 when starting from stationarity is deferred to the proof of Proposition 10. The simple linear model with $|\theta_0| < 1$ has geometrically decreasing (and therefore summable) α-mixing coefficients as a consequence of [20] (eq. (15.49)) and Theorem A.1.2. Combining these two facts, it follows that the conditions of Theorem 2 are satisfied. □

Observe that Theorem 1 (and Corollary 1) are general, and hold for *any* dependent data-generating process. Therefore, there can be Markov chains that satisfy these, but do not satisfy Assumption 1 which entails some loss of generality. However, as our examples demonstrate, common Markov chain models do indeed satisfy the latter assumption.

4. Non-Stationary, Ergodic Markov Data-Generating Models

We call a time-homogeneous Markov chain *non-stationary* if the initial distribution $q^{(0)}$ is not the invariant distribution. There are two sets of results in this setting: in Theorem 3 and Theorem 4 we explicitly impose the α-mixing condition, while in Theorem 5 we impose a f-geometric ergodicity condition (Definition A.1.2 in the appendix). As seen in Equation (A4) (in the appendix) if the Markov chain is also geometrically ergodic, then $\forall\ \delta > 0, \sum \alpha_k^{\delta/(2+\delta)} < \infty$. This condition can be relaxed, albeit at the risk of more complicated calculations that, nonetheless, mirror those in the geometrically ergodic setting. A common thread through these results is that we must impose some integrability or regularity conditions on the functions $M_k^{(1)}$.

First, in Theorem 3 we assume that the $M_k^{(1)}$ functions in Assumption 1 are uniformly bounded and that the α-mixing condition is satisfied. This result holds for both discrete and continuous state space settings.

Theorem 3. *Let $\{X_0, \ldots, X_n\}$ be generated by an α-mixing Markov chain parametrized by $\theta_0 \in \Theta$ with transition probabilities satisfying Assumption 1 and with known initial distribution $q^{(0)}$. Let $\{\alpha_k\}$ be the α-mixing coefficients under θ_0, and assume that $\sum_{k\geq 1} \alpha_k^{\delta/(2+\delta)} < +\infty$. Suppose that there exists $B \in \mathbb{R}$ such that $\sup_{x,y} |M_k^{(1)}(x,y)| < B$ for all $k \in \{1,2,\ldots,m\}$ in Assumption 1. Furthermore, assume that there exists $\rho_n \in \mathcal{F}$ such that $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n}C$ for some constant $C > 0$. If the initial distribution $q^{(0)}$ satisfies $E_{q^{(0)}} |M_k^{(2)}(X_0)|^2 < +\infty$ for all $k \in \{1,2,\ldots,m\}$, then the conditions of Corollary 1 are satisfied with $\epsilon_n = \mathrm{O}\Big(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\Big)$.*

The following result in Proposition 8 illustrates Theorem 3 in the setting of a finite state birth-death Markov chain.

Proposition 8. *Suppose the data-generating process is a finite state birth-death Markov chain, with one-step transition kernel parametrized by the birth probability θ_0. Let $\mathcal{F}$ be the set of all Beta distributions. We choose the prior on θ_0 to be a Beta distribution. Then, the conditions of Theorem 3 are satisfied with $\epsilon_n = \mathrm{O}\Big(\frac{1}{\sqrt{n}}\Big)$ for any initial distribution $q^{(0)}$.*

Theorem 3 also applies to data generated by Markov chains with countably infinite state spaces, so long as the class of data-generating Markov chains is strongly ergodic and the initial distribution has finite second moments. The following example demonstrates this in the setting of a birth-death Markov chain on the positive integers, where the initial distribution is assumed to have finite second moments.

Proposition 9. *Suppose the data-generating process is a birth-death Markov chain on the non-negative integers, parameterized by the probability of birth $\theta_0 \in (0, \frac{1}{2})$. Further let $\mathcal{F}$ be the set of all Beta distributions rescaled upon the support $(0, \frac{1}{2})$. Let $q^{(0)}$ be a probability mass function on non-negative integers such that $\sum_{i=1}^{\infty} i^2 q^{(0)}(i) < +\infty$. We choose the prior to be a scaled Beta distribution on $(0, 1/2)$ with parameters a and b. Then, the conditions of Theorem 3 are satisfied with $\epsilon_n = \mathrm{O}\Big(\frac{1}{\sqrt{n}}\Big)$.*

Since continuous functions on a compact domain are bounded, we have the following (easy) corollary (stated without proof).

Corollary 2. *Let $\{X_0, \ldots, X_n\}$ be generated by an α-mixing Markov chain parametrized by $\theta_0 \in \Theta$ on a compact state space, and with initial distribution $q^{(0)}$. Suppose the α-mixing coefficients satisfy $\sum_{k\geq 1} \alpha_k^{\delta/(2+\delta)} < +\infty$, and that Assumption 1 holds with continuous functions $M_k^{(1)}(\cdot,\cdot)$, $k \in \{1,2,\ldots,m\}$. Furthermore, assume that there exists ρ_n such that $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n}C$ for some constant C. Then, Theorem 3 is satisfied with $\epsilon_n = \mathrm{O}\Big(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\Big)$.*

In general, the $M_k^{(1)}$ functions will not be uniformly bounded (consider the case of the Gauss–Markov simple linear model in Example 2), and stronger conditions must be imposed on the data-generating Markov chain itself. The following assumption imposes a 'drift' condition from [21]. Specifically, [21] (Theorem 2.3) shows that under the conditions of Assumption 2, the moment generating function of an aperiodic Markov chain $\{X_n\}$ can be upper bounded by a function of the moment generating function of X_0. Together with the α-mixing condition, Assumption 2 implies that this Markov data generating process satisfies Corollary 1.

Assumption 2. *Consider a Markov chain $\{X_n\}$ parameterized by $\theta_0 \in \Theta$. Let $\mathcal{M}_{-\infty}^n$ denote the σ-field generated by $\{X_{-\infty}, \ldots, X_{n-1}, X_n\}$. Denote the stochastic process $\{M_n^k\} := \{M_k^{(1)}(X_n, X_{n-1})\}$; recall $M_k^{(1)}$, for each $k = 1, \ldots, m_1$, are defined in Assumption 1. For each $k = 1, \ldots, m$, assume the process $\{M_n^k\}$ satisfies the following conditions:*

- *The drift condition holds for $\{M_n^k\}$, i.e., $\mathrm{E}\left[M_n^k - M_{n-1}^k | \mathcal{M}_{-\infty}^{n-1}, M_{n-1}^k > a\right] \leq -\epsilon$ for some $\epsilon, a > 0$.*
- *For some $\lambda > 0$ and $\mathcal{D} > 0$, $\mathrm{E}\left[e^{\lambda(M_n^k - M_{n-1}^k)} | \mathcal{M}_{-\infty}^{n-1}\right] \leq \mathcal{D}$.*

Under this drift condition, the next theorem shows that Corollary 1 is satisfied.

Theorem 4. *Let $\{X_0, \ldots, X_n\}$ be generated by an aperiodic α-mixing Markov chain parametrized by $\theta_0 \in \Theta$ and initial distribution $q^{(0)}$. Suppose that Assumption 1 and Assumption 2 hold, and that the α-mixing coefficients satisfy $\sum_{k \geq 1} \alpha_k^{\delta/(2+\delta)} < +\infty$. Furthermore, assume $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n}C$ for some constant $C > 0$. If $\int e^{\lambda M_k^{(1)}(y,x)} p_{\theta_0}(y|x) q_1^{(0)}(x) dx < +\infty$ for all $k = 1, \ldots, m_1$, then the conditions of Corollary 1 are satisfied with $\epsilon_n = \mathrm{O}\left(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\right)$.*

Verifying the conditions in Theorem 4 can be quite challenging. Instead, we suggest a different approach that requires f-geometric ergodicity. Unlike the drift condition in Assumption 2, f-geometric ergodicity additionally requires the existence of a petite set. As noted before, geometric ergodicity implies α-mixing with geometrically decaying mixing coefficients. As with Theorem 4, we assume for simplicity that the Markov chain is aperiodic.

Theorem 5. *Let $\{X_0, \ldots, X_n\}$ be generated by an aperiodic Markov chain parametrized by $\theta_0 \in \Theta$ with known initial distribution $q^{(0)}$, and assumed to be V-geometrically ergodic for some $V : \mathbb{R}^m \to [1, \infty)$. Suppose that Assumption 1 holds and $\int M_k^{(1)}(y, x)^{2+\delta} p_{\theta_0}(y|x) dy < V(x) \ \forall \ k, x$ and some $\delta > 0$. Furthermore, assume that $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n}C$ for some constant $C > 0$. If the initial distribution $q^{(0)}$ satisfies $E_{q^{(0)}}[V(X_0)] < +\infty$, then the conditions of Corollary 1 are satisfied with $\epsilon_n = \mathrm{O}\left(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\right)$.*

The following Proposition 10 shows, the simple linear model satisfies Theorem 5 when the parameter θ_0 is suitably restricted.

Proposition 10. *Consider the simple linear model satisfying the equation*

$$X_n = \theta_0 X_{n-1} + W_n, \tag{22}$$

where $\{W_n\}$ are i.i.d. standard Gaussian random variables and $|\theta_0| < 2^{\frac{1}{4+2\delta}-1}$ for $\delta > 0$. Let $\mathcal{F}$ be the space of all scaled Beta distributions on $(-1, 1)$ and suppose the prior π is a uniform distribution on $(-1, 1)$. Then, the conditions of Theorem 5 are satisfied with $\epsilon_n = \mathrm{O}\left(\max(\frac{1}{\sqrt{n}}, \frac{n^{\delta/2}}{n})\right)$, if the initial distribution $q^{(0)}$ satisfies $E_{q^{(0)}}[X_0^{4+2\delta}] < +\infty$.

5. Misspecified Models

We show next how our results can be extended to the misspecified model setting. Assume that the true data generating distribution is parametrized by $\theta_0 \notin \Theta$. Let $\theta_n^* := \arg\min_{\theta\in\Theta} \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})$ represent the closest parametrized distribution in the variational family to the data-generating distribution. Further, assume our usual conditions:

i. $\int \mathrm{E}[r_n(\theta, \theta_n^*)]\rho_n(d\theta) \leq n\epsilon_n$,
ii. $\int \mathrm{Var}(r_n(\theta, \theta_n^*))\rho_n(d\theta) \leq n\epsilon_n$.

Now, since $r_n(\theta, \theta_0) = r_n(\theta, \theta_n^*) + r_n(\theta_n^*, \theta_0)$, we have

$$\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) \leq \mathrm{E}[r_n(\theta_0, \theta_n^*)] + n\epsilon_n. \tag{23}$$

Similarly, decomposing the variance it follows that

$$\mathrm{Var}[r_n(\theta, \theta_0)] = \mathrm{Var}[r_n(\theta, \theta_n^*)] + \mathrm{Var}[r_n(\theta_n^*, \theta_0)] + 2\mathrm{Cov}[r_n(\theta, \theta_n^*), r_n(\theta_n^*, \theta_0)]. \tag{24}$$

Using the fact that $2ab \leq a^2 + b^2$ on the covariance term $2\mathrm{Cov}[r_n(\theta, \theta_n^*), r_n(\theta_n^*, \theta_0)] = 2\mathrm{E}[(r_n(\theta, \theta_n^*) - \mathrm{E}[r_n(\theta, \theta_n^*)])(r_n(\theta_n^*, \theta_0) - \mathrm{E}[r_n(\theta_n^*, \theta_0)])]$, we have

$$\mathrm{Var}[r_n(\theta, \theta_0)] \leq 2\mathrm{Var}[r_n(\theta, \theta_n^*)] + 2\mathrm{Var}[r_n(\theta_n^*, \theta_0)]. \tag{25}$$

Integrating both sides with respect to $\rho_n(d\theta)$ we get

$$\begin{aligned}\int \mathrm{Var}[r_n(\theta, \theta_0)]\rho_n(d\theta) &\leq 2\int \mathrm{Var}[r_n(\theta, \theta_n^*)]\rho_n(d\theta) + 2\int \mathrm{Var}[r_n(\theta_n^*, \theta_0)]\rho_n(d\theta) \\ &\leq 2n\epsilon_n + 2\mathrm{Var}[r_n(\theta_n^*, \theta_0)].\end{aligned} \tag{26}$$

Consequently, we arrive at the following result:

Theorem 6. *Let $\mathcal{F}$ be a subset of all probability distributions parameterized by Θ. Let $\theta_n^* = \arg\min_{\theta\in\Theta} \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})$ and assume there exist $\epsilon_n > 0$ and $\rho_n \in \mathcal{F}$ such that*

i. $\int \mathrm{E}[r_n(\theta, \theta_n^*)]\rho_n(d\theta) \leq n\epsilon_n$,
ii. $\int \mathrm{Var}(r_n(\theta, \theta_n^*))\rho_n(d\theta) \leq n\epsilon_n$, *and*
iii. $\mathcal{K}(\rho_n, \pi) \leq n\epsilon_n$.

Then, for any $\alpha^{re} \in (0,1)$ and $(\epsilon, \eta) \in (0,1)\times(0,1)$,

$$P\left[\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}}(d\theta|X^{(n)}) \leq \frac{(\alpha^{re}+1)n\epsilon_n + \mathrm{E}[r_n(\theta_0, \theta_n^*)] + \alpha^{re}\sqrt{\frac{2n\epsilon_n + 2\mathrm{Var}[r_n(\theta_n^*, \theta_0)]}{\eta}} - \log(\epsilon)}{1-\alpha^{re}}\right] \geq 1 - \epsilon - \eta. \tag{27}$$

The proof of this theorem is straightforward and follows from the proof of Theorem 1 by plugging in the upper bounds for KL-divergence from Equation (23), and variance from Equation (26) to Equation (A13). A sketch of the proof is presented in the appendix.

6. Conclusions

Concentration of the KL-VB model risk, in terms of the expected α^{re}-Rényi divergence, is well established under the i.i.d. data generating model assumption. Here, we extended this to the setting of Markov data generating models, linking the concentration rate to the mixing and ergodic properties of the Markov model. Our results apply to both stationary and non-stationary Markov chains, as well as to the situation with misspecified models. There remain a number of open questions. An immediate one is to extend the current

analysis to continuous-time Markov chains and Markov jump processes, possibly using uniformization of the continuous time model. Another direction is to extend this to the setting of non-homogeneous Markov chains, where analogues of notions such as stationarity are less straightforward. Further, as noted in the introduction, [14] establish PAC-Bayes bounds under slightly weaker 'existence of test functions' conditions, while our results are established under the stronger conditions used by [15] for the i.i.d. setting. Weakening the conditions in our analysis is important, but complicated. A possible path is to build on results from [22], who provides conditions form the existence of exponentially powerful test functions exist for distinguishing between two Markov chains. It is also known that there exists a likelihood ratio test separating any two ergodic measures [23]. However, leveraging these to establish the PAC-Bayes bounds for the KL-VB posterior is a challenging effort that we leave to future papers. Finally it is of interest to generalize our PAC-bounds to posterior approximations beyond KL-variational inference, such as α^{re}-Rényi posterior approximations [6], and loss-calibrated posterior approximations [24,25].

Author Contributions: Formal analysis, I.B.; Investigation, I.B.; Methodology, I.B., V.A.R. and H.H.; Resources, V.A.R. and H.H.; Validation, V.A.R. and H.H. All authors have read and agreed to the published version of the manuscript.

Funding: National Science Foundation : IIS-1816499; DMS-1812197.

Acknowledgments: Rao and Honnappa acknowledge support from NSF DMS-1812197. In addition, Rao acknowledges NSF IIS-1816499 for supporting this project.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Technical Desiderata

Appendix A.1. Definitions Related to Markov Chains

As noted before, ergodicity plays an acute role in establishing our results. We consolidate various definitions used throughout the paper in this appendix. Recall that we assume the parameterized Markov chain possesses an invariant probability density or mass function q_θ under parameter $\theta \in \Theta$. Our results in Section 4 also rely on the ergodic properties of the Markov chain, and we assume that the Markov chain is f-geometrically ergodic [20] (Chapter 15). First, refer to the definition of the functional norm $\|\cdot\|_f$, from Definition A.1.1,

Definition A.1.1 (f-norm). *The functional norm in f-metric of a measure v, or the f-norm of v is*

$$\|v\|_f = \sup_{g:|g|<f} \left| \int g dv \right|, \tag{A1}$$

where f and g are any two functions.

An immediate consequence of this definition is that if f_1, f_2 are two functions such that $f_1 < f_2$ (i.e., for all points in the support of the functions), then

$$\|v\|_{f_1} \leq \|v\|_{f_2}. \tag{A2}$$

Now that we have defined the $\|\cdot\|_f$ norm, we can now define f-geometric ergodicity. In the following, we assume the Markov chain is positive Harris; see [20] for a definition. This is a mild and fairly standard assumption in Markov chain theory.

Definition A.1.2 (f-geometric ergodicity). *For any function f, Markov chain $\{X_n\}$ parameterized by $\theta \in \Theta$ is said to be f-geometrically ergodic if it is positive Harris and there exists a constant $r_f > 1$, that depends on f, such that for any $A \in \mathcal{B}(X)$,*

$$\sum_{n=1}^{n} r_f^n \left\| P_\theta(X_n \in A | X_0 = x) - \int_A q_\theta(y) dy \right\|_f < \infty. \tag{A3}$$

It is straightforward to see that this is equivalent to

$$\left\| P_\theta(X_n \in A | X_0 = x) - \int q_\theta(y) dy \right\|_f \leq C r_f^{-n}$$

for an appropriate constant C (which may depend on the state x), that is, the Markov chain approaches steady state at a geometrically fast rate. If a Markov chain is f-geometrically ergodic for $f \equiv 1$, then, it is simply termed as *geometrically ergodic*. It is straightforward to see (via Theorem A.1.2 in the Appendix) that a geometrically ergodic Markov chain is also α-mixing, with mixing coefficients satisfying

$$\sum_{k \geq 0} \alpha_k^v < \infty \ \forall \ v > 0, \tag{A4}$$

showing that, under geometric ergodicity, the α-mixing coefficients raised to any positive power v are finitely summable. We note here that the most standard procedure to establish f-geometric ergodicity for any Markov chain is through the verification of the drift condition. The drift condition is a sufficient condition for a Markov chain to be f-geometrically ergodic, as long as there exists a set (called petite set) towards which the Markov chain drifts to (see Assumption A.1.1 in the appendix). If a Markov chain is f-geometrically ergodic with $f \equiv V$, for some particular function V, then we call it V-geometrically ergodic.

We defined V-geometric ergodicity in the previous sections. In this section, we provide a sufficient condition for a Markov chain to be V-geometrically ergodic. First, we recall the definition of resolvent from [20] (Chapter 5).

Definition A.1.3 (Resolvent). *Let $n \in \{0, 1, 2, \dots\}$ and q_n be such that $q_n \geq 0 \ \forall \ n$ and $\sum_{n=1}^{\infty} q_n = 1$. Note that q_n can be thought of being a probability mass function for a random variable "q" taking values on non-negative integers. Then, the resolvent of a Markov chain with respect to q is given by $K_q(x, A)$ where,*

$$K_q(x, A) = \sum_{n=0}^{\infty} q_n P(X_n \in A | X_0 = x). \tag{A5}$$

Then, the definition of petite sets follows (see, for Reference, [20] (Chapter 5)).

Definition A.1.4 (Petite Sets). *Let $X_0, \dots, X_n$ be n samples from a Markov chain taking values on the state space $\mathcal{X}$. Let C be a set. We shall call C to be v_q petite if*

$$K_q(x, B) \geq v_q(B)$$

for all $x \in C$ and $B \in \mathcal{B}(\mathcal{X})$, and a non-trivial measure v_q on $\mathcal{B}(\mathcal{X})$, and a probability mass function q on $\{1, 2, 3, \dots\}$

Now, let $\Delta V(x) := E[V(X_n) | X_{n-1} = x] - V(x)$ for $V : S \to [1, \infty)$.

Assumption A.1.1 (Drift condition). *[20] (Chapter 5) Suppose the chain $\{X_n\}$ is, aperiodic and ψ-irreducible . Let there exists a petite set C, constants $b < \infty, \beta > 0$, and a non-trivial function $V : S \to [1, \infty)$ satisfying*

$$\Delta V(x) \leq -\beta V(x) + bI_{x \in C} \ \ \forall x \in S. \tag{A6}$$

If a Markov chain drifts towards a petite set then it is V-geometrically ergodic. Suppose, for simplicity, that $V(x) = |X|$. Then, the drift condition becomes $E[|X_n \| X_{x-1}] - |X_{n-1}| = -\beta |X_n| + bI_{X_n \in C}$. The left hand side of this equation represents the change in the state of the Markov chain in one time epoch. Thus, the condition in Assumption A.1.1 essentially states that the Markov chain drifts towards a petite set C and then, once it reaches that set, moves to any point in the state space with at least some probability independent of C.

Theorem A.1.1 (Geometrically ergodic theorem). *Suppose that $\{X_n\}$ is satisfies Assumption A.1.1. Then, the set $S_V = \{x : V(x) < \infty\}$ is absorbing, i.e., $P_\theta(X_1 \in S_V | X_0 = x) = 1 \ \forall x \in S_V$, and full, i.e., $\psi(S_V^c) = 0$. Furthermore, $\exists$ constants $r > 1, R < \infty$ such that, for any $A \in \mathcal{B}(S)$,*

$$\left\| P_\theta(X_n \in A | X_0 = x) - \int_A q_\theta(y) dy \right\|_V \leq R r^{-n} V(x). \tag{A7}$$

Any aperiodic and ψ-irreducible Markov chain satisfying the drift condition is geometrically ergodic. A consequence of Equation (A2) is that if, $\{X_n\}$ is V-geometrically ergodic, then for any other function U, such that $|U| < V$, it is also U-geometrically ergodic. In essence, a geometrically ergodic Markov chain is asymptotically uncorrelated in a precise sense. Recall ρ-mixing coefficients defined as follows. Let $\mathcal{A}$ be a sigma field and $\mathcal{L}^2(A)$ be the set of square integrable, real valued, $\mathcal{A}$ measurable functions.

Definition A.1.5 (ρ-mixing coefficient). *Let $\mathcal{M}_i^j$ denote the sigma field generated by the measures X_k, where $i \leq k \leq j$. Then,*

$$\rho_k = \sup_{t > 0} \sup_{(f,g) \in \mathcal{L}^2(\mathcal{M}_{-\infty}^t) \times \mathcal{L}^2(\mathcal{M}_{t+k}^\infty)} |\mathrm{Corr}(f, g)|, \tag{A8}$$

where Corr *is the correlation function.*

Theorem A.1.2. *If X_n is geometrically ergodic, then it is α-mixing. That is, there exists a constant $c > 0$ such that $\alpha_k = \mathrm{O}(e^{-ck})$.*

Proof. By [26] (Theorem 2) it follows that a geometrically ergodic Markov chain is asymptotically uncorrelated with ρ-mixing coefficients (see Definition A.1.5) that satisfy $\rho_k = \mathrm{O}(e^{-ck})$. Furthermore, it is well known that [18,26] $\alpha_k \leq \frac{1}{4}\rho_k$, implying $\alpha_k = \mathrm{O}(e^{-ck})$. □

Appendix A.2. Bounding the KL-Divergence between Beta Distributions

The following results will be utilized in the proofs of Propositions 8–10.

Lemma A.2.1. *Let $\theta_0 \in (0, 1)$. Let, ρ_n be a sequence of Beta distributions with parameters $a_n = n\theta_0$ and $b_n = n(1 - \theta_0)$. Let π denote an uniform distribution, $U(0, 1)$. Then, $\mathcal{K}(\rho_n, \pi) < C + \frac{1}{2}\log(n)$, for some constant $C > 0$.*

Proof. Without loss of generality, we can assume $a_n > 1$ and $b_n > 1$. The same form of the result can be obtained in all the other cases, by appropriate use of the bounds presented in the proof. We write the KL divergence $\mathcal{K}(\rho_n, \pi)$ as $\int \log\left(\frac{\rho_n}{\pi}\right) \rho_n(d\theta)$. Since π is uniform,

$\pi(\theta) = 1$ whenever $\theta \in (0,1)$. Hence, the KL-divergence can be written as the negative of the entropy of ρ_n $\int_0^1 \log(\rho_n(\theta))\rho_n(d\theta)$, which can be written as

$$\mathcal{K}(\rho_n, \pi) = (a_n - 1)\psi(a_n) + (b_n - 1)\psi(b_n) - (a_n + b_n - 2)\psi(a_n + b_n) - \log \text{Beta}(a_n, b_n), \tag{A9}$$

where ψ is the digamma function. Using Stirling's approximation on $\text{Beta}(a_n, b_n)$ yields,

$$\text{Beta}(a_n, b_n) = \sqrt{2\pi}\frac{a_n^{a_n - 1/2} b_n^{b_n - 1/2}}{(a_n + b_n)^{a_n + b_n - 1/2}}(1 + o(1)).$$

Hence, setting $C_1 = \log(2\sqrt{\pi})$, we can write $-\log \text{Beta}(a_n, b_n)$ as,

$$-\log \text{Beta}(a_n, b_n) = C_1 - (a_n - \frac{1}{2})\log(a_n) - (b_n - \frac{1}{2})\log(b_n) + (a_n + b_n - \frac{1}{2})\log(a_n + b_n) + \log(1 + o(1)).$$

From [27] we have that $\log(x) - \frac{1}{x} < \psi(x) < \log(x) - \frac{1}{2x}$ $\forall$ $x > 0$. Since we assumed $a_n > 1$ and $b_n > 1$, the fact that $\psi(x) < \log(x) - \frac{1}{2x}$ implies

$$(a_n - 1)\psi(a_n) < (a_n - 1)\log(a_n) - \frac{a_n - 1}{2a_n} \text{ and,}$$
$$(b_n - 1)\psi(b_n) < (b_n - 1)\log(b_n) - \frac{b_n - 1}{2b_n}.$$

Finally, using the fact that $\log(x) - \frac{1}{x} < \psi(x)$, we get,

$$-(a_n + b_n - 2)\psi(a_n + b_n) < -(a_n + b_n - 2)\log(a_n + b_n) + \frac{a_n + b_n - 2}{a_n + b_n}.$$

Therefore, after much cancellation, the KL-divergence

$$(a_n - 1)\psi(a_n) + (b_n - 1)\psi(b_n) - (a_n + b_n - 2)\psi(a_n + b_n) - \log \text{Beta}(a_n, b_n)$$

can be upper bounded by

$$-\frac{1}{2}\log(a_n) - \frac{1}{2}\log(b_n) + \frac{3}{2}\log(a_n + b_n) + \frac{a_n + b_n - 2}{a_n + b_n} - \frac{a_n - 1}{2a_n} - \frac{b_n - 1}{2b_n}.$$

Now, plugging in the values of a_n and b_n, we get Plugging in the values of a_n and b_n, we get as upper bound for the KL-divergence as,

$$\begin{aligned}\mathcal{K}(\rho_n, \pi) &< -\frac{1}{2}\log(n\theta_0) - \frac{1}{2}\log(n(1-\theta_0)) + \frac{3}{2}\log(n) + \frac{n-2}{n} - \frac{n\theta_0 - 1}{2n\theta_0} - \frac{n(1-\theta_0) - 1}{2n(1-\theta_0)} \\ &= \frac{1}{2}\log(n) - \frac{1}{2}(\log(\theta_0) + \log(1-\theta_0)) + 3 - \frac{2}{n} - \frac{1}{2n\theta_0} - \frac{1}{2n(1-\theta_0)} \\ &< C + \frac{1}{2}\log(n),\end{aligned}$$

for some large enough positive constant C. This completes our proof. □

Proposition A.2.1. *Let $\theta_0 \in (0,1)$. Let, ρ_n be a sequence of Beta distributions with parameters $a_n = n\theta_0$ and $b_n = n(1-\theta_0)$. Let π denote an Beta distribution, with parameters (a, b). Then, $\mathcal{K}(\rho_n, \pi) < C + \frac{1}{2}\log(n)$, for some constant $C > 0$.*

Proof. Without loss of generality, we assume $a > 1$ and $b > 1$. As mentioned in the proof of Lemma A.2.1, the other cases follows similarly. We write the KL-divergence between ρ_n and π as,

$$\mathcal{K}(\rho_n, \pi) = \int \log\Big(\frac{\rho_n}{\pi}\Big)\rho_n(d\theta) = \int \log\Big(\frac{\rho_n}{U}\Big)\rho_n(d\theta) + \int \log\Big(\frac{U}{\pi}\Big)\rho_n(d\theta),$$

where, U is an uniform distribution on $(0,1)$. We analyze the second term in the above expression. The second term can be written as,

$$\begin{aligned}\int \log\Big(\frac{U}{\pi}\Big)\rho_n(d\theta) &= \int \log\left(\frac{1}{\frac{1}{\text{Beta}(a,b)}\theta^{a-1}(1-\theta)^{b-1}}\right)\rho_n(d\theta) \\ &= C_1 - (a-1)\int \log(\theta)\rho_n(d\theta) - (b-1)\int \log(1-\theta)\rho_n(d\theta),\end{aligned}$$

where C_1 is $\log(\text{Beta}(a,b))$. Since, ρ_n follows a Beta distribution with parameters $a_n = n\theta_0$ and $b_n = n(1-\theta_0)$, we get that,

$$\int \log\Big(\frac{U}{\pi}\Big)\rho_n(d\theta) = C_1 - (a-1)[\psi(a_n) - \psi(a_n+b_n)] - (b-1)[\psi(b_n) - \psi(a_n+b_n)]$$

Since, $\log(x) - \frac{1}{x} < \psi(x) < \log(x) - \frac{1}{2x}$, looking at the term $[\psi(a_n) - \psi(a_n+b_n)]$, we get that,

$$\begin{aligned}-[\psi(a_n) - \psi(a_n+b_n)] &= -[\psi(n\theta_0) - \psi(n\theta_0 + n(1-\theta_0))] \\ &= -[\psi(n\theta_0) - \psi(n)].\end{aligned}$$

Using the lower bound on $\psi(n\theta_0)$ and the upper bound on $\psi(n)$, we get

$$\begin{aligned}-[\psi(a_n) - \psi(a_n+b_n)] &< -\log(n\theta_0) + \frac{1}{n\theta_0} + \log(n) - \frac{1}{2n} \\ &= -\log(\theta_0) + \frac{2-\theta_0}{2n\theta_0}.\end{aligned}$$

Furthermore, similarly, we get that,

$$-[\psi(b_n) - \psi(a_n+b_n)] < -\log(1-\theta_0) + \frac{2-(1-\theta_0)}{2n(1-\theta_0)}.$$

Therefore it follows that

$$\begin{aligned}&\max\{-(a-1)[\psi(a_n) - \psi(a_n+b_n)], -(b-1)[\psi(b_n) - \psi(a_n+b_n)]\} \\ &\quad< \max\left\{(a-1)\left[-\log(\theta_0) + \frac{2-\theta_0}{2n\theta_0}\right], (b-1)\left[-\log(1-\theta_0) + \frac{2-(1-\theta_0)}{2n(1-\theta_0)}\right]\right\} \\ &\quad< C,\end{aligned}$$

for a large positive constant C. Using the above bounds, we finally show that,

$$\begin{aligned}C_1 - (a-1)[\psi(a_n) - \psi(a_n+b_n)] - (b-1)[\psi(b_n) - \psi(a_n+b_n)] \\ < C_1 + 2C,\end{aligned}$$

which can be upper bounded by C' for some large constant C'. Finally, we upper bound $\int \log(\frac{\rho_n}{U})\rho_n(d\theta)$ by Lemma A.2.1 thereby completing the proof. □

Appendix B. Proofs of Main Results

Appendix B.1. Proofs for A Concentration Bound for the α^{re}-Rényi Divergence

Appendix B.1.1. Proof of Proposition 1

We start by recalling the variational formula of Donsker and Varadhan [28].

Lemma B.1.1 (Donsker-Varadhan). *For any probability distribution function π on Θ, and for any measurable function $h : \Theta \to \mathbb{R}$, if $\int e^h d\pi < \infty$, then*

$$\log \int e^h d\pi = \sup_{\rho \in \mathcal{M}^+(\Theta)} \left\{ \int h d\rho - \mathcal{K}(\rho, \pi) \right\} \tag{A10}$$

Now, fix $\alpha^{re} \in (0,1)$, and $\theta \in \Theta$. First, observe that by the definition of the α^{re}-Rényi divergence we have

$$\mathrm{E}_{\theta_0}^{(n)}[\exp(-\alpha^{re} r_n(\theta, \theta_0))] = \exp[-(1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})]$$

Multiplying both sides of the equation by $\exp[(1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})$ and integrating with respect to (w.r.t.) $\pi(\theta)$ it follows that

$$\begin{aligned} &\int \mathrm{E}_{\theta_0}^{(n)} \left[\exp\Big(-\alpha^{re} r_n(\theta, \theta_0) + (1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \Big) \right] \pi(d\theta) = 1, \text{or} \\ &\mathrm{E}_{\theta_0}^{(n)} \left[\int \exp\Big(-\alpha^{re} r_n(\theta, \theta_0) + (1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \Big) \pi(d\theta) \right] = 1. \end{aligned}$$

Define $h(\theta) := -\alpha^{re} r_n(\theta, \theta_0) + (1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})$. Then, applying Lemma B.1.1 to the integrand on the left hand side (l.h.s.) above, it follows that

$$\mathrm{E}_{\theta_0}^{(n)} \left[\exp\left(\sup_{\rho \in \mathcal{M}^+(\Theta)} \left[\int h(\theta) \rho(d\theta) - \mathcal{K}(\rho, \pi) \right] \right) \right] = 1.$$

Multiply both sides of this equation by $\epsilon > 0$ to obtain

$$\mathrm{E}_{\theta_0}^{(n)} \left[\exp\left(\sup_{\rho \in \mathcal{M}^+(\Theta)} \left[\int h(\theta) \rho(d\theta) - \mathcal{K}(\rho, \pi) + \log(\epsilon) \right] \right) \right] = \epsilon.$$

Now, by Markov's inequality, we have

$$P_{\theta_0}^{(n)} \left[\sup_{\rho \in \mathcal{M}^+(\Theta)} \int (-\alpha^{re} r_n(\theta, \theta_0) + (1-\alpha^{re}) D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})) \rho(d\theta) - \mathcal{K}(\rho, \pi) + \log(\epsilon) \geq 0 \right] \leq \epsilon. \tag{A11}$$

Thus, it follows via complementation that

$$P_{\theta_0}^{(n)} \left[\forall \rho \in \mathcal{F}(\Theta) \int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)}) \rho(d\theta) \leq \frac{\alpha^{re}}{(1-\alpha^{re})} \int r_n(\theta, \theta_0) \rho(d\theta) + \frac{\mathcal{K}(\rho, \pi) - \log(\epsilon)}{1-\alpha^{re}} \right] \geq 1 - \epsilon,$$

thereby completing the proof. □

Appendix B.1.2. Proof of Theorem 1

Recall the definition of the fractional posterior and the VB approximation,

$$\pi_{n,\alpha^{re}|X^n} = \frac{\exp^{-\alpha^{re} r_n(\theta,\theta_0)(X^n)} \pi(d\theta)}{\int \exp^{-\alpha^{re} r_n(\gamma,\theta_0)(X^n)} \pi(d\gamma)}, \ \tilde{\pi}_{n,\alpha^{re}|X^n} = \arg\min_{\rho \in \mathcal{F}} \mathcal{K}(\rho, \pi_{n,\alpha^{re}|X^{(n)}}).$$

It follows by definition of the KL divergence that

$$\tilde{\pi}_{n,\alpha^{re}|X^n} = \arg\min_{\rho\in\mathcal{F}}\left\{-\alpha^{re}\int r_n(\theta,\theta_0)\rho(d\theta) + \mathcal{K}(\rho,\pi)\right\}, \tag{A12}$$

where π is the prior distribution. Following Proposition 1 it follows that for any $\epsilon > 0$

$$\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}(d\theta|X^n) \leq \frac{\alpha^{re}}{(1-\alpha^{re})}\int r_n(\theta,\theta_0)\rho(d\theta) + \frac{\mathcal{K}(\rho,\pi) - \log(\epsilon)}{1-\alpha^{re}},$$

with probability $1-\epsilon$. We fix an $\eta \in (0,1)$. Using Chebychev's inequality, we have

$$\begin{aligned}
&P_{\theta_0}^{(n)}\left[\frac{\alpha^{re}}{1-\alpha^{re}}\int r_n(\theta,\theta_0)\rho_n(d\theta) \geq \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta)\right.\\
&\qquad\left. + \frac{\alpha^{re}}{1-\alpha^{re}}\sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho_n(d\theta)]}{\eta}} + \frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha^{re}}\right]\\
&= P_{\theta_0}^{(n)}\left[\frac{\alpha^{re}}{1-\alpha^{re}}\int r_n(\theta,\theta_0)\rho_n(d\theta) - \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta) - \frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha^{re}}\right.\\
&\qquad\left. \geq \frac{\alpha^{re}}{1-\alpha^{re}}\sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho_n(d\theta)]}{\eta}}\right]\\
&\leq \frac{\mathrm{Var}\left[\frac{\alpha^{re}}{1-\alpha^{re}}\int r_n(\theta,\theta_0)\rho_n(d\theta) - \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta) - \frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha^{re}}\right]}{\frac{(\alpha^{re})^2}{(1-\alpha^{re})^2}\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho_n(d\theta)]}{\eta}}.
\end{aligned}$$

Note that $\frac{\alpha^{re}}{1-\alpha^{re}}\int E(r_n(\theta,\theta_0))\rho_n(d\theta)$ and $\frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha^{re}}$ are constants with respect to the data, implying

$$\begin{aligned}
&\mathrm{Var}\left[\frac{\alpha^{re}}{1-\alpha^{re}}\int r_n(\theta,\theta_0)\rho_n(d\theta) - \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta) - \frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha^{re}}\right]\\
&\qquad = \frac{(\alpha^{re})^2}{(1-\alpha^{re})^2}\mathrm{Var}\left[\int r_n(\theta,\theta_0)\rho_n(d\theta)\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
P_{\theta_0}^{(n)}\left[\frac{\alpha^{re}}{1-\alpha^{re}}\int r_n(\theta,\theta_0)\rho_n(d\theta) \geq \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathrm{E}[r_n(\theta,\theta_0)]\rho_n(d\theta)\right.\\
\left. + \frac{\alpha^{re}}{1-\alpha^{re}}\sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho_n(d\theta)]}{\eta}} + \frac{\mathcal{K}(\rho_n,\pi)}{1-\alpha}\right] \leq \eta.
\end{aligned}$$

From Proposition 1, with probability $1-\epsilon$ the following holds

$$\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}|X^n}(d\theta) \leq \frac{\alpha^{re}\int r_n(\theta,\theta_0)\rho_n(d\theta) + \mathcal{K}(\rho_n,\pi) - \log(\epsilon)}{1-\alpha^{re}}.$$

Therefore, with probability $1-\eta-\epsilon$ the following statement holds

$$\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}|X^n}(d\theta) \leq \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) + \frac{\alpha^{re}}{1-\alpha^{re}}\sqrt{\frac{\mathrm{Var}[\int r_n(\theta,\theta_0)\rho_n(d\theta)]}{\eta}} + \frac{\mathcal{K}(\rho_n,\pi)-\log(\epsilon)}{1-\alpha^{re}}. \tag{A13}$$

Next, we observe that

$$\mathrm{Var}\left[\int r_n(\theta,\theta_0)\rho_n(d\theta)\right] = E_{\theta_0}^{(n)}\left[\left|\int r_n(\theta,\theta_0)\rho_n(d\theta) - E\left[\int r_n(\theta,\theta_0)\rho_n(d\theta)\right]\right|^2\right] \leq \int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta),$$

by a straightforward application of Jensen's inequality to the inner integral on the left hand side. Finally, following the hypotheses (i), (ii) and (iii), we have,

$$\begin{aligned}\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}|X^n}(d\theta) &\leq \frac{\alpha^{re}}{1-\alpha^{re}}\int\left(\mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)}) + \sqrt{\frac{\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta)}{\eta}}\right)\rho_n(d\theta) \\ &\quad + \frac{1}{\alpha^{re}}(\mathcal{K}(\rho_n,\pi)-\log(\epsilon)) \\ &\leq \frac{\alpha^{re}(\epsilon_n + \sqrt{\frac{n\epsilon_n}{\eta}})}{1-\alpha^{re}} + \frac{n\epsilon_n - \log(\epsilon)}{1-\alpha^{re}},\end{aligned}$$

thereby concluding the proof. □

Appendix B.1.3. Proof of Proposition 2

We define $Y_i := \log\left(\frac{p_{\theta_1}(X_i|X_{i-1})}{p_{\theta_2}(X_i|X_{i-1})}\right)$ for $i = 1,\ldots,n$, and $Z_0 = \log\left(\frac{q_1^{(0)}(X_0)}{q_2^{(0)}(X_0)}\right)$. Then, using the Markov property we can see that the Kullback–Leibler divergence between the joint distributions $P_{\theta_1}^{(n)}$ and $P_{\theta_2}^{(n)}$ satisfies $\mathcal{K}\left(P_{\theta_1}^{(n)}, P_{\theta_2}^{(n)}\right) = \sum_{i=1}^n \mathrm{E}_{\theta_1}[Y_i] + \mathrm{E}_{\theta_1}[Z_0]$. If the Markov chain $\{X_i\}$ is stationary under θ_1, so is $\{Y_i\}$. Hence $Y_i \overset{d}{=} Y_1$ and the above equation reduces to,

$$\mathcal{K}\left(P_{\theta_1}^{(n)}, P_{\theta_2}^{(n)}\right) = n\mathrm{E}_{\theta_1}[Y_1] + \mathrm{E}_{\theta_1}[Z_0]. \tag{A14}$$

□

Appendix B.1.4. Proof of Proposition 3

First, recall the following result from [19].

Lemma B.1.2. *[19] (Lemma 1.2) Let $X_{-\infty},\ldots,X_1,X_2,\ldots$ be an α-mixing Markov chain with α-mixing coefficients given by α_k. Let $\mathcal{M}_a^b$ be the sigma-field generated by the subsequence $(X_a,X_{a+1},\ldots,X_b)$. Let $\eta_t \in \mathcal{M}_{-\infty}^t$ and $\tau_t \in \mathcal{M}_{t+k}^\infty$ be adapted random variables such that $|\eta_t| \leq 1, |\tau_t| \leq 1$. Then,*

$$\sup_t \sup_{\eta_t,\tau_t} |\mathrm{E}[\eta_t\tau_t] - \mathrm{E}[\eta_t]\mathrm{E}[\tau_t]| \leq 4\alpha_k. \tag{A15}$$

This lemma provides an upper bound on the covariance of events η and τ, as shown next.

Lemma B.1.3. *Let $\eta \in \mathcal{M}_{-\infty}^{t}$ $\tau \in \mathcal{M}_{t+k}^{\infty}$ be such that, $E|\eta|^{2+\delta} \leq C_1, E|\tau|^{2+\delta} \leq C_2$ for some $\delta > 0$. Then, for a fixed $n < +\infty$, we have*

$$|\mathrm{E}\eta\tau - \mathrm{E}\eta\mathrm{E}\tau| \leq \left(\frac{4}{n} + 2n^{\delta/2}(C_1 + C_2) + 2n^{\delta/2}\sqrt{C_1C_2}\right)\alpha_k^{2\delta/(2+\delta)}. \tag{A16}$$

Proof. Let $N < +\infty$ be a fixed number. We get from the triangle inequality that

$$\begin{aligned}|\mathrm{E}\eta\tau - \mathrm{E}\eta\mathrm{E}\tau| \leq\ & |\mathrm{E}\eta\tau I_{[|\eta|\leq N,|\tau|\leq N]} - \mathrm{E}\eta I_{[|\eta|\leq N]}\mathrm{E}\tau I_{[|\tau|\leq N]}| \\ &+ |\mathrm{E}\eta\tau I_{[|\eta|\geq N,|\tau|\leq N]} - \mathrm{E}\eta I_{[|\eta|\geq N]}\mathrm{E}\tau I_{[|\tau|\leq N]}| \\ &+ |\mathrm{E}\eta\tau I_{[|\eta|\leq N,|\tau|\geq N]} - \mathrm{E}\eta I_{[|\eta|\leq N]}\mathrm{E}\tau I_{[|\tau|\geq N]}| \\ &+ |\mathrm{E}\eta\tau I_{[|\eta|\geq N,|\tau|\geq N]} - \mathrm{E}\eta I_{[|\eta|\geq N]}\mathrm{E}\tau I_{[|\tau|\geq N]}|.\end{aligned} \tag{A17}$$

Multiplying and dividing the first term by N^2 and applying Lemma B.1.2, we get $|\mathrm{E}\eta\tau I_{[|\eta|\leq N,|\tau|\leq N]} - \mathrm{E}\eta I_{[|\eta|\leq N]}\mathrm{E}\tau I_{[|\tau|\leq N]}| \leq 4N^2\alpha_k$. For the second term, if $|\tau| \leq N$, then $\tau \leq N$ and $\tau \geq -N$. Plugging this in the second term we get,

$$|\mathrm{E}\eta\tau I_{[|\eta|\geq N,|\tau|\leq N]} - \mathrm{E}\eta I_{[|\eta|\geq N]}\mathrm{E}\tau I_{[|\tau|\leq N]}| \leq \left|N\mathrm{E}\eta I_{[|\eta|\geq N} + N\left[\mathrm{E}\eta I_{[|\eta|\geq N]}\right]\right| \tag{A18}$$

$$= 2N|\mathrm{E}\eta I_{[|\eta|\geq N]}|. \tag{A19}$$

Since $|\eta| \geq N$, we have $1 \leq \frac{|\eta|^{1+\delta}}{N^{1+\delta}}$. Following this,

$$|2N\mathrm{E}\eta I_{[|\eta|\geq N]}| \leq 2N\left|\mathrm{E}\left[\frac{|\eta|^{2+\delta}}{N^{1+\delta}} I_{[|\eta|\geq N]}\right]\right| \tag{A20}$$

$$\leq 2N\frac{1}{N^{1+\delta}}|\mathrm{E}\eta^{2+\delta}| \leq 2\frac{C_1}{N^{\delta}}. \tag{A21}$$

Similarly, we can also write for the third term, $|\mathrm{E}\eta\tau I_{[|\eta|\leq N,|\tau|\geq N]} - \mathrm{E}\eta I_{[|\eta|\leq N]}\mathrm{E}\tau I_{[|\tau|\geq N]}| \leq 2\frac{C_2}{N^{\delta}}$. Finally, for the last term we get that by Cauchy-Schwarz inequality,

$$|\mathrm{E}\eta\tau I_{[|\eta|\geq N,|\tau|\geq N]} - \mathrm{E}\eta I_{[|\eta|\geq N]}\mathrm{E}\tau I_{[|\tau|\geq N]}| \leq \sqrt{\mathrm{Var}\left[\eta I_{[|\eta|\geq N]}\right]\mathrm{Var}\left[\tau I_{[|\tau|\geq N]}\right]} \tag{A22}$$

$$< 2\sqrt{\mathrm{Var}\left[\eta I_{[|\eta|\geq N]}\right]\mathrm{Var}\left[\tau I_{[|\tau|\geq N]}\right]} \tag{A23}$$

$$\leq 2\sqrt{\mathrm{E}\left[\eta^2 I_{[|\eta|\geq N]}\right]\mathrm{E}\left[\tau^2 I_{[|\tau|\geq N]}\right]}. \tag{A24}$$

Since $|\eta| > N$, $1 < \frac{|\eta|^{\delta}}{N^{\delta}}$. Similarly, $1 < \frac{|\tau|^{\delta}}{N^{\delta}}$. Plugging these in the previous equation, we get,

$$\sqrt{\mathrm{E}\left[\eta^2 I_{[|\eta|\geq N]}\right]\mathrm{E}\left[\tau^2 I_{[|\tau|\geq N]}\right]} \leq \sqrt{\frac{1}{N^{2\delta}}\mathrm{E}\left[|\eta|^{2+\delta} I_{[|\eta|\geq N]}\right]\mathrm{E}\left[|\tau|^{2+\delta} I_{[|\tau|\geq N]}\right]} \tag{A25}$$

$$\leq \frac{1}{N^{\delta}}\sqrt{C_1C_2}. \tag{A26}$$

Combining the four upper bounds above, we get,

$$|\mathrm{E}\eta\tau - \mathrm{E}\eta\mathrm{E}\tau| \leq 4N^2\alpha_k + \frac{2}{N^{\delta}}(C_1 + C_2) + \frac{2}{N^{\delta}}\sqrt{C_1C_2}. \tag{A27}$$

Now, in particular, setting $N = n^{-1/2}\alpha_k^{-1/(2+\delta)}$ it follows that

$$|\mathrm{E}\eta\tau - \mathrm{E}\eta\mathrm{E}\tau| \leq \frac{4}{n}\alpha_k^{\delta/(2+\delta)} + 2n^{\delta/2}\alpha_k^{\delta/(2+\delta)}(C_1 + C_2) + 2n^{\delta/2}\alpha_k^{\delta/(2+\delta)}\sqrt{C_1C_2} \tag{A28}$$

$$= \left(\frac{4}{n} + 2n^{\delta/2}(C_1 + C_2) + 2n^{\delta/2}\sqrt{C_1C_2}\right)\alpha_k^{\delta/(2+\delta)}. \tag{A29}$$

□

Lemma B.1.4. *Let $\{X_t\}$ be an α-mixing Markov chain with mixing coefficient α_k. Further assume that $\mathrm{E}|X_t|^{2+\delta} \leq C_1$ and $\mathrm{E}|X_{t+k}|^{2+\delta} \leq C_2$ for some $\delta > 0$. Then, for any t and any $n > 0$*

$$|\mathrm{Cov}(X_t, X_{t+k})| \leq \left(\frac{4}{n} + 2n^{\delta/2}(C_1 + C_2) + 2n^{\delta/2}\sqrt{C_1C_2}\right)\alpha_k^{\delta/(2+\delta)}. \tag{A30}$$

Proof. Set $\eta = X_t, \tau = X_{t+k}$ in Lemma B.1.3. □

We also need to establish the following technical lemma.

Lemma B.1.5. *Let $\{X_t\}$ be an α-mixing Markov Chain with mixing coefficients $\{\alpha_t\}$. Then the process $\{Y_t\}$ where $Y_t := \log\left(\frac{p_{\theta_0}(X_t|X_{t-1})}{p_\theta(X_t|X_{t-1})}\right)$ is also α-mixing with mixing coefficients $\{\tilde{\alpha}_t\}$ where $\tilde{\alpha}_t = \alpha_{t-1}$.*

Proof. By Z_i denote the paired random measure (X_i, X_{i-1}). Let $\mathcal{M}_i^j$ denote the sigma field generated by the measures X_k, where $i \leq k \leq j$. By $\mathcal{G}_i^j$ denote the sigma field generated by the measures Z_k, where $i \leq k \leq j$. Let $C \in \mathcal{M}_{i-1}^j$. Then, C can be expressed as $(C_{i-1} \times C_i \times \cdots \times C_j)$. for $C_{i-1} \in \mathcal{M}_{i-1}^{i-1}, C_i \in \mathcal{M}_i^i \ldots$ and so on. Now, consider a map. $T_i^j : (C_{i-1} \times C_i \times \cdots \times C_j) \to (C_{i-1} \times C_i \times C_i \times \cdots \times C_{j-1} \times C_{j-1} \times C_j)$. Note that, $T_i^j(C) \in \mathcal{G}_i^j$. It is easy to see that $\mathcal{G}_i^j = T_i^j(\mathcal{M}_{i-1}^j) \cup \mathcal{M}_{i-1}^{*j}$, where $T_i^j(\mathcal{M}_{i-1}^j)$ is obtained by applying the map T_i^j to each element of $\mathcal{M}_{i-1}^j$. If we assume this latter set to be the range and $\mathcal{M}_{i-1}^j$ to be the domain, then, by construction, T_i^j is a bijection. Furthermore, the two classes are made of disjoint sets, i.e., if $A \in T_i^j(\mathcal{M}_{i-1}^j)$ and $A^* \in \mathcal{M}_{i-1}^{*j}$, then $A \cap A^* = \phi$. Furthermore, note that $\mathcal{M}_{i-1}^{j*}$ is made of impossible sets. i.e., $P(A^*) = 0 \ \forall \ A^* \in \mathcal{M}_{i-1}^{j*}$. Now consider the α-mixing coefficients for Z_i. By definition, it is given by

$$\begin{aligned}\alpha_k^z &= \sup_i \sup_{A \in \mathcal{G}_{-\infty}^i, B \in \mathcal{G}_{i+k}^\infty} |P(A \cap B) - P(A)P(B)| \\ &= \sup_i \sup_{A \in \mathcal{G}_{-\infty}^i, B \in \mathcal{G}_{i+k}^\infty} |P((A^o \cup A^*) \cap (B^o \cup B^*)) - P((A^o \cup A^*))P((B^o \cup B^*))|.\end{aligned}$$

where,

$$\begin{array}{ll} A = (A^o \cup A^*) & B = (B^o \cup B^*) \\ A^o \in \mathcal{T}_{-\infty}^i(M_{-\infty}^i) & A^* \in \mathcal{M}_{-\infty}^{*i} \\ B^o \in T_{i+k-1}^\infty(\mathcal{M}_{j+k-1}^\infty) & B^* \in \mathcal{M}_{j+k-1}^{*\infty}. \end{array}$$

Then, the expression for the α-mixing coefficient can be reduced into

$$\alpha_k^z = \sup_i \sup_{A^o \in T_{-\infty}^i(\mathcal{M}_{-\infty}^i), B^o \in T_{i+k-1}^\infty(\mathcal{M}_{i+k-1}^\infty)} |P(A^o \cap B^o) - P(A^o)P(B^o)|.$$

Note that, by bijection property of T_i^j, we can find $A' \in \mathcal{M}_{-\infty}^i$ and $B' \in \mathcal{M}_{i+k-1}^\infty$ such that

$$\begin{aligned}\alpha_k^z &= \sup_i \sup_{A' \in \mathcal{M}_{-\infty}^i, B' \in \mathcal{M}_{i+k-1}^\infty} |P(T_{-\infty}^i(A') \cap T_{i+k-1}^\infty(B')) - P(T_{-\infty}^i(A'))P(T_{i+k-1}^\infty(B'))|. \\ &= \alpha_{k-1}.\end{aligned}$$

Now, $\log\left(\frac{p_{\theta_0}(X_n|X_{n-1})}{p_\theta(X_n|X_{n-1})}\right)$ is just a function of the paired Markov chain Z_i, therefore it has α-mixing coefficient α_{k-1}. □

We now proceed to the proof of Proposition 3. Let $\{X_k\}$ be a stationary α-mixing Markov chain under θ_1 with mixing coefficients $\{\alpha_k\}$. Observe that the log-likelihood can be expressed as

$$\begin{aligned}r_n(\theta_2, \theta_1) &= \sum_{i=1}^n \log\left(\frac{p_{\theta_1}(X_i|X_{i-1})}{p_{\theta_2}(X_i|X_{i-1})}\right) + \log\left(\frac{q_1^{(0)}(X_0)}{q_2^{(0)}(X_0)}\right) \\ &\equiv \sum_{i=1}^n Y_i + Z_0.\end{aligned}$$

Therefore, the variance of the log-likelihood ratio is simply

$$\begin{aligned}\mathrm{Var}_{\theta_1}[r_n(\theta_2, \theta_1)] &= \mathrm{Var}_{\theta_1}\left[\sum_{i=1}^n Y_i + Z_0\right] \\ &= \sum_{i,j=1}^n \mathrm{Cov}_{\theta_1}(Y_i, Y_j) + \sum_{i=1}^n \mathrm{Cov}_{\theta_1}(Y_i, Z_0) + \mathrm{Cov}_{\theta_1}(Z_0, Z_0).\end{aligned}$$

It follows from Lemma B.1.5 that $\{Y_k\}$ is a stochastic process with α-mixing coefficients α_{k-1}. Therefore, using Lemma B.1.4 we have

$$\begin{aligned}|\mathrm{Cov}_{\theta_1}(Y_i, Y_j)| &= |\mathrm{E}_{\theta_1} Y_i Y_j - \mathrm{E}_{\theta_1} Y_i \mathrm{E}_{\theta_1} Y_j| \\ &< \left(\frac{4}{n} + 2n^{\delta/2}(\mathrm{E}_{\theta_1}|Y_i|^{2+\delta} + \mathrm{E}_{\theta_1}|Y_j|^{2+\delta}\right. \\ &\quad \left. + \sqrt{\mathrm{E}_{\theta_1}|Y_i|^{2+\delta}\mathrm{E}_{\theta_1}|Y_j|^{2+\delta}})\right)\alpha_{|j-i|-1}^{\delta/(2+\delta)} \\ &= \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + C_{\theta_1,\theta_2}^{(j)} + \sqrt{C_{\theta_1,\theta_2}^{(i)} C_{\theta_1,\theta_2}^{(j)}})\right)\alpha_{|j-i|-1}^{\delta/(2+\delta)}.\end{aligned}$$

Similarly, as above we can also say

$$|\mathrm{Cov}_{\theta_1}(Y_i, Z_0)| < \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + D_{1,2} + \sqrt{C_{\theta_1,\theta_2}^{(i)} D_{1,2}})\right)\left(\alpha_{i-1}^{\delta/(2+\delta)}\right)$$

Combining, the two upper bounds above, we get the first result:

$$\begin{aligned}\mathrm{Var}_{\theta_1}\left[r_n(\theta_2, \theta_1)\right] &< \sum_{i,j=1}^n \left(\frac{4}{n} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + C_{\theta_1,\theta_2}^{(j)} + \sqrt{C_{\theta_1,\theta_2}^{(i)} C_{\theta_1,\theta_2}^{(j)}})\right)\left(\alpha_{|i-j|-1}^{\delta/(2+\delta)}\right) \\ &\quad + \sum_{i=1}^n \left(\frac{4}{n^2} + 2n^{\delta/2}(C_{\theta_1,\theta_2}^{(i)} + D_{1,2} + \sqrt{C_{\theta_1,\theta_2}^{(i)} D_{1,2}})\right)\left(\alpha_{i-1}^{\delta/(2+\delta)}\right) \\ &\quad + \mathrm{Var}[Z_0, Z_0].\end{aligned}$$

If $\{X_i\}$ is stationary under θ_1, so is $\{Y_i\}$. Therefore, $\mathrm{E}_{\theta_1}|Y_i|^{2+\delta} = \mathrm{E}_{\theta_1}|Y_1|^{2+\delta} = C^{(1)}_{\theta_1,\theta_2}\ \forall\ i$, and

$$\begin{aligned}\sum_{i,j=1}^{n}\mathrm{Cov}_{\theta_1}(Y_i,Y_j) &\leq \sum_{i,j=1}^{n}\left(\frac{4}{n}+6n^{\delta/2}C^{(1)}_{\theta_1,\theta_2}\right)\alpha^{\delta/(2+\delta)}_{|j-i|-1}\\ &\leq n\left(\frac{4}{n}+6n^{\delta/2}C^{(1)}_{\theta_1,\theta_2}\right)\left(\sum_{h\geq 1}\alpha^{\delta/(2+\delta)}_{h-1}\right).\end{aligned} \tag{A31}$$

Again, using Lemma B.1.4 on $\mathrm{Cov}_{\theta_1}(Y_i,Z_0)$, yields

$$\sum_{i=1}^{n}\mathrm{Cov}_{\theta_1}(Y_i,Z_0) \leq \left(\frac{4}{n}+2n^{\delta/2}(C_\theta+D_{1,2}+\sqrt{C_\theta D_{1,2}})\right)\left(\sum_{h\geq 1}\alpha^{\delta/(2+\delta)}_{h}\right). \tag{A32}$$

Finally, using Equations (A31) and (A32) we have

$$\begin{aligned}\mathrm{Var}_{\theta_1}[r_n(\theta_2,\theta_1)] \leq\ & n\left(\frac{4}{n}+6n^{\delta/2}C^{(1)}_{\theta_1,\theta_2}\right)\left(\sum_{h\geq 1}\alpha^{\delta/(2+\delta)}_{h-1}\right)+\\ & \left(\frac{4}{n}+2n^{\delta/2}(C^{(1)}_{\theta_1,\theta_2}+D_{1,2}+\sqrt{C^{(1)}_{\theta_1,\theta_2}D_{1,2}})\right)\left(\sum_{h\geq 1}\alpha^{\delta/(2+\delta)}_{h}\right)\\ & +\mathrm{Cov}_{\theta_1}(Z_0,Z_0).\end{aligned}$$

□

Appendix B.2. Proofs for Stationary Markov Data-Generating Models

Proof of Theorem 2

Part 1: Verifying condition (i) of Corollary 1.

We substitute the true parameter θ_0 for θ_1 and θ for θ_2. We also set $q^{(0)}_1$ to be the invariant distribution of the Markov chain under θ_0, q_0, and $q^{(0)}_2$ as the invariant distribution of the Markov chain under θ, q_θ. Applying the fact that these Markov chains are stationary to Proposition 2, we have

$$\begin{aligned}\mathcal{K}(P^{(n)}_{\theta_0},P^{(n)}_{\theta}) &= n\mathrm{E}\left[\log\left(\frac{p_{\theta_0}(X_1|X_0)}{p_\theta(X_1|X_0)}\right)\right]+\mathrm{E}[Z_0],\\ &\leq n\sum_{j=1}^{m}\mathrm{E}\left[M^{(1)}_j(X_1,X_0)\right]|f^{(1)}_j(\theta,\theta_0)|+\sum_{k=1}^{m}\mathrm{E}[M^{(2)}_k(X_0)]|f^{(2)}_k(\theta,\theta_0)|,\end{aligned} \tag{A33}$$

where the inequality follows from Assumption 1. Therefore, it follows that

$$\begin{aligned}\int\mathcal{K}(P^{(n)}_{\theta_0},P^{(n)}_{\theta})\rho_n(d\theta) \leq\ & n\sum_{j=1}^{m}\mathrm{E}\left[M^{(1)}_j(X_1,X_0)\right]\int|f^{(1)}_j(\theta,\theta_0)|\rho_n(d\theta)\\ & +\sum_{k=1}^{m}\mathrm{E}[M^{(2)}_k(X_0)]|\int f^{(2)}_k(\theta,\theta_0)|\rho_n(d\theta).\end{aligned}$$

By Assumption 1(i), it follows that

$$\int\mathcal{K}(P^{(n)}_{\theta_0},P^{(n)}_{\theta})\rho_n(d\theta) \leq n\sum_{j=1}^{m}\mathrm{E}\left[M^{(1)}_j(X_1,X_0)\right]\frac{C}{\sqrt{n}}+\sum_{k=1}^{m}\mathrm{E}[M^{(2)}_k(X_0)]\frac{C}{\sqrt{n}} \leq n\epsilon^{(1)}_n,$$

where $\epsilon_n^{(1)} = \mathrm{O}\left(\frac{1}{\sqrt{n}}\right)$.

Part 2: Verifying condition (ii) of Corollary 1. Again, using Proposition 3 along with the fact that the Markov chain is stationary we have

$$\begin{aligned}\mathrm{Var}[r_n(\theta,\theta_0)] \leq n\left(\frac{4}{n} + 6n^{\delta/2}C_{\theta_0,\theta}^{(1)}\right)\left(\sum_{k\geq 0}\alpha_k^{\delta/(2+\delta)}\right) \\ + \left(\frac{4}{n^2} + 2n^{\delta/2}(C_{\theta_0,\theta}^{(1)} + D_{\theta_0,\theta} + \sqrt{C_{\theta_0,\theta}^{(1)}D_{\theta_0,\theta}})\right)\left(\sum_{k\geq 1}\alpha_k^{\delta/(2+\delta)}\right) \\ + \mathrm{Var}[Z_0].\end{aligned}$$

It then follows that

$$\begin{aligned}\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq n\left(\frac{4}{n} + 6n^{\delta/2}\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta)\right)\left(\sum_{k\geq 1}\alpha_{k-1}^{\delta/(2+\delta)}\right) + \int \mathrm{Var}[Z_0]\rho_n(d\theta) \\ + \left(\frac{4}{n^2} + 2n^{\delta/2}(\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta) \right. \\ \left. + \int D_{\theta_0,\theta}\rho_n(d\theta) + \int \sqrt{C_{\theta_0,\theta}^{(1)}D_{\theta_0,\theta}}\rho_n(d\theta))\right)\left(\sum_{k\geq 1}\alpha_k^{\delta/(2+\delta)}\right).\end{aligned}$$

First, consider the term $\int C_{\theta_0,\theta}^{(1)}\rho_n(\theta)$, and observe that

$$\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta) = \int \mathrm{E}\log\left|\frac{p_{\theta_0}(X_1|X_0)}{p_\theta(X_1|X_0)}\right|^{2+\delta}\rho_n(d\theta).$$

By Assumption 1, we have

$$\int \mathrm{E}\log\left|\frac{p_{\theta_0}(X_1|X_0)}{p_\theta(X_1|X_0)}\right|^{2+\delta}\rho_n(d\theta) \leq \int \mathrm{E}\left[\sum_{j=1}^m M_j^{(1)}(X_1,X_0)|f_k^{(1)}(\theta,\theta_0)|\right]^{2+\delta}\rho_n(d\theta).$$

Since the function $x \mapsto x^{2+\delta}$ is convex, we can apply Jensen's inequality to obtain,

$$\left(\sum_{j=1}^m M_j^{(1)}(X_1,X_0)|f_k^{(1)}(\theta,\theta_0)|\right)^{2+\delta} \leq m^{1+\delta}\sum_{k=1}^m M_j^{(1)}(X_1,X_0)^{2+\delta}|f_k^{(1)}(\theta,\theta_0)|^{2+\delta}.$$

Therefore, it follows that

$$\begin{aligned}\int \mathrm{E}\log\left|\frac{p_{\theta_0}(X_1|X_0)}{p_\theta(X_1|X_0)}\right|^{2+\delta}\rho_n(d\theta) \leq m^{1+\delta}\sum_{k=1}^m \mathrm{E}[M_k^{(1)}(X_1,X_0)^{2+\delta}] \\ \times \int |f_k^{(1)}(\theta,\theta_0)|^{2+\delta}\rho_n(d\theta).\end{aligned}$$

By Assumption 1, $\int |f_k(\theta,\theta_0)|^{2+\delta}\rho_n(d\theta) < \frac{C}{n}$ and $\mathrm{E}[M_k^{(1)}(X_1,X_0)^{2+\delta}] < B$, implying that

$$\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta) \leq m^{1+\delta}\sum_{k=1}^m B\frac{C}{n} = m^{2+\delta}\frac{BC}{n}.$$

Since $\left(\sum_{k\geq 0}\alpha_k^{\delta/(2+\delta)}\right) < \infty$, it follows that $\left(\frac{4}{n} + 6n^{\delta/2}\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta)\right)\left(\sum_{k\geq 1}\alpha_{k-1}^{\delta/(2+\delta)}\right) = \mathrm{O}(\frac{n^{\delta/2}}{n})$. Similarly, we can show that $\int D_{\theta_0,\theta}\rho_n(d\theta) = \mathrm{O}(\frac{1}{n})$, and $\int \mathrm{Var}[Z_0]\rho_n(d\theta) = \mathrm{O}(\frac{1}{n})$.

For the final term $\int\sqrt{C_{\theta_0,\theta}^{(1)}D_{\theta_0,\theta}}\rho_n(d\theta)$, use the Cauchy-Schwarz inequality to obtain the upper bound $\left(\int C_{\theta_0,\theta}^{(1)}\rho_n(d\theta)\int D_{\theta_0,\theta}\rho_n(d\theta)\right)^{1/2}$ which is also of order $O(\frac{1}{n})$. Combining all of these together we have

$$\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq n\epsilon_n^{(2)},$$

for some $\epsilon_n^{(2)} = \mathrm{O}(\frac{n^{\delta/2}}{n})$.

Since $\mathcal{K}(\rho_n,\pi) < \sqrt{n}C = n\frac{C}{\sqrt{n}}$, it follows that $\mathcal{K}(\rho_n,\pi) < n\epsilon_n^{(3)}$, where $\epsilon_n^{(3)} = \mathrm{O}(1/\sqrt{n})$ as before. Finally, by choosing $\epsilon_n = \max(\epsilon_n^{(1)},\epsilon_n^{(2)},\epsilon_n^{(3)})$, our theorem is proved. □

Appendix B.3. Proofs for Non-Stationary, Ergodic Markov Data-Generating Models

Appendix B.3.1. Proof of Theorem 3

Part 1: Verifying condition (i) of Corollary 1: As in the proof of Theorem 2 substitute the true parameter θ_0 for θ_1 and θ for θ_2 in . We also set $q_1^{(0)}$ and $q_2^{(0)}$ to the distribution $q^{(0)}$. Applying Proposition 2 to the corresponding transition kernels and initial distribution we have,

$$\begin{aligned}\mathcal{K}(P_{\theta_0}^{(n)},P_\theta^{(n)}) &= \sum_{i=1}^n \mathrm{E}\left[\log\left(\frac{p_{\theta_0}(X_i|X_{i-1})}{p_\theta(X_i|X_{i-1})}\right)\right] + \mathrm{E}\left[\log\left(\frac{D(X_0)}{D(X_0)}\right)\right] \\ &= \sum_{i=1}^n \mathrm{E}\left[\log\left(\frac{p_{\theta_0}(X_i|X_{i-1})}{p_\theta(X_i|X_{i-1})}\right)\right].\end{aligned} \tag{A34}$$

Now, applying Assumption 1, we can bound the previous equation as follows,

$$\begin{aligned}\mathcal{K}(P_{\theta_0}^{(n)},P_\theta^{(n)}) &\leq \sum_{i=1}^n \mathrm{E}\left[\sum_{k=1}^m M_k^{(1)}(X_i,X_{i-1})|f_k^{(1)}(\theta,\theta_0)|\right] \\ &= \sum_{i=1}^n\sum_{k=1}^m \mathrm{E}\left[M_k^{(1)}(X_i,X_{i-1})\right]|f_k^{(1)}(\theta,\theta_0)|.\end{aligned} \tag{A35}$$

Since $M_k^{(1)}$'s are bounded there exists a constant Q so that,

$$\begin{aligned}\int \mathcal{K}(P_{\theta_0}^{(n)},P_\theta^{(n)})\rho_n(d\theta) &\leq Q\int\sum_{i=1}^n\sum_{k=1}^m |f_k^{(1)}(\theta,\theta_0)|\rho_n(d\theta) \\ &= Qn\sum_{k=1}^m\int |f_k^{(1)}(\theta,\theta_0)|\rho_n(d\theta).\end{aligned}$$

By Assumption 19 in Assumption 1, it follows that

$$\int \mathcal{K}(P_{\theta_0}^{(n)},P_\theta^{(n)})\rho_n(d\theta) \leq Qn\sum_{k=1}^m\frac{C}{\sqrt{n}} = nmQ\frac{C}{\sqrt{n}} = n\epsilon_n^{(1)},$$

for some $\epsilon_n^{(1)} = \mathrm{O}(\frac{1}{\sqrt{n}})$.

Part 2: Verifying condition (ii) of Corollary 1: As in the previous part, $Z_0 = 0$, implying that D_{θ,θ_0}. Applying Proposition 3 and integrating with respect to ρ_n, we obtain

$$\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) \le \sum_{i=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)\right)\left(\alpha_{i-1}^{\delta/(2+\delta)}\right) + \sum_{i,j=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\Big(\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)+\int C^{(j)}_{\theta_0,\theta}\rho_n(d\theta)+\int\sqrt{C^{(i)}_{\theta_0,\theta}C^{(j)}_{\theta_0,\theta}}\rho_n(d\theta)\Big)\right) \times\left(\alpha_{|i-j|-1}^{\delta/(2+\delta)}\right). \tag{A36}$$

First, consider the term $\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)$. Using Assumption 1, we can upper bound $C^{(i)}_{\theta_0,\theta}$ as,

$$\begin{aligned} C^{(i)}_{\theta_0,\theta} &\le \mathrm{E}\left[\sum_{k=1}^{m} M_k^{(1)}(X_i,X_{i-1})|f_k^{(1)}(\theta,\theta_0)|\right]^{2+\delta} \\ &\le \sum_{k=1}^{m} m^{1+\delta}\mathrm{E}\left[\left(M_k^{(1)}(X_i,X_{i-1})|f_k^{(1)}(\theta,\theta_0)|\right)^{2+\delta}\right] \text{(by Jensen's inequality)} \\ &= \sum_{k=1}^{m} m^{1+\delta}\mathrm{E}\left[M_k^{(1)}(X_i,X_{i-1})^{2+\delta}\right]|f_k^{(1)}(\theta,\theta_0)|^{2+\delta}. \end{aligned}$$

Since $M_k^{(1)}$'s are upper bounded by Q, it follows from the previous expression that, $C^{(i)}_{\theta_0,\theta} \le \sum_{k=1}^{m} m^{1+\delta}Q^{2+\delta}|f_k^{(1)}(\theta,\theta_0)|^{2+\delta}$.

Hence, from Assumption 1, we get,

$$\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta) \le \sum_{k=1}^{m} m^{1+\delta}Q^{2+\delta}\int |f_k^{(1)}(\theta,\theta_0)|^{2+\delta}\rho_n(d\theta) \le (mQ)^{2+\delta}\frac{C}{n}.$$

Using the upper bound above, we can say for an L large enough, $\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta) \le \frac{L}{n}$. Next, by the Cauchy-Schwarz inequality, we have that $\int\sqrt{C^{(i)}_{\theta_0,\theta}C^{(j)}_{\theta_0,\theta}\rho_n(d\theta))} < \sqrt{\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)\int C^{(j)}_{\theta_0,\theta}\rho_n(d\theta))} \le \frac{L}{n}$. Thus, we have the following upper bound.

$$\begin{aligned} \int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) &\le \sum_{i=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\frac{L}{n}\right)\left(\alpha_{i-1}^{\delta/(2+\delta)}\right) + \sum_{i,j=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\Big(\frac{L}{n}+\frac{L}{n}+\frac{L}{n}\Big)\right)\left(\alpha_{|i-j|-1}^{\delta/(2+\delta)}\right) \\ &= \left(\frac{4}{n}+2n^{\delta/2}\frac{L}{n}\right)\left(\sum_{i=1}^{n}\alpha_{i-1}^{\delta/(2+\delta)}\right) + \left(\frac{4}{n}+6n^{\delta/2}\frac{L}{n}\right)\left(\sum_{i,j=1}^{n}\alpha_{|i-j|-1}^{\delta/(2+\delta)}\right). \end{aligned}$$

Since $\sum_{i,j=1}^{n}\alpha_{|i-j|-1}^{\delta/(2+\delta)} < n\sum_{k\ge1}\alpha_{k-1}^{\delta/(2+\delta)} < \infty$, we have that for some $\epsilon_n^{(2)} = \mathrm{O}(\frac{n^{\delta/2}}{n})$,

$$\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) < n\epsilon_n^{(2)}.$$

Since $\mathcal{K}(\rho_n,\pi) \le \sqrt{n}C$, following the concluding argument in Theorem 2 completes the proof. □

Appendix B.3.2. Proof of Proposition 8

We verify Assumption 1 and the proof follows from Theorem 3. For $i \in \{1, 2, \ldots, K-1\}$,

$$p_\theta(j|i) = \begin{cases} \theta & \text{if } j = i-1, \\ 1-\theta & \text{if } j = i+1. \end{cases}$$

If $i = 0$ or $i = K$, then the Markov chain goes back to 1 or $K-1$, respectively, with probability 1. With the convention $\log \frac{0}{0} = 0$, the log ratio of the transition probabilities becomes,

$$|\log p_{\theta_0}(X_1|X_0) - \log p_\theta(X_1|X_0)| = I_{[X_1 = X_0+1]} \log\left(\frac{\theta_0}{\theta}\right) + I_{[X_1 = X_0-1]} \log\left(\frac{1-\theta_0}{1-\theta}\right).$$

In this case, $m = 2$. $M_1^{(1)}(X_1, X_0) = I_{[X_1 = X_0+1]}$ and $M_2^{(1)}(X_1, X_0) = I_{[X_1 = X_0-1]}$, both of which are bounded. Let $f_1^{(1)}(\theta, \theta_0) := \log\left(\frac{\theta_0}{\theta}\right)$ suppose $f_2^{(1)}(\theta, \theta_0) := \log\left(\frac{1-\theta_0}{1-\theta}\right)$.

The stationary distribution $q_\theta(i) = \frac{1}{K} \ \forall \ i \in 1, 2, \ldots, K$. Hence the log of the ratio of the invariant distribution becomes

$$\log q_0(x) - \log q_\theta(x) = 0, \tag{A37}$$

and we can set $M_i^{(2)}(\cdot) := 1$ and $f_i^{(2)}(\cdot, \cdot) := 0$ for $i \in \{1, 2\}$. Thus, to prove the concentration bound for this Markov chain it is enough to assume that $\delta = 1$ and show that $\int [f_1^{(1)}(\theta, \theta_0)]^3 \rho_n(d\theta) < \frac{C}{n}$ and $\int [f_2^{(1)}(\theta, \theta_0)]^3 \rho_n(d\theta) < \frac{C}{n}$ for some constant $C > 0$.

As given, $\{\rho_n\}$ is a sequence of beta probability distribution functions, with parameters a_n, b_n that satisfy the constraint $\frac{a_n}{a_n + b_n} = \theta_0$. Specifically, we choose $a_n = n\theta_0$ and (therefore) $b_n = n(1-\theta_0)$. Thus, we get the following,

$$\begin{aligned} \int |f_1^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) &= \int \left|\log\left(\frac{\theta_0}{\theta}\right)\right|^3 \rho_n(d\theta) \\ &< \int \left|\frac{\theta_0}{\theta} - 1\right|^3 \rho_n(d\theta) \\ &= \frac{1}{\text{Beta}(a_n, b_n)} \int_0^1 \left|\frac{\theta_0 - \theta}{\theta}\right|^3 \theta^{a_n - 1}(1-\theta)^{b_n - 1} d\theta. \end{aligned}$$

Since $\theta_0, \theta \in (0, 1)$, so is $\frac{|\theta_0 - \theta|}{2}$, giving $|\theta_0 - \theta|^3 < 2(\theta_0 - \theta)^2$. We use that fact to arrive at

$$\begin{aligned} \int |f_1^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) &\leq \frac{2}{\text{Beta}(a_n, b_n)} \int_0^1 (\theta_0 - \theta)^2 \theta^{a_n - 4}(1-\theta)^{b_n - 1} d\theta \\ &= \frac{2\text{Beta}(a_n - 3, b_n)}{\text{Beta}(a_n, b_n)} \frac{(a_n - 3)(b_n)}{(a_n + b_n - 3)^2 (a_n + b_n - 2)}. \end{aligned}$$

From our choice of a_n and b_n, $\frac{2\text{Beta}(a_n - 3, b_n)}{\text{Beta}(a_n, b_n)} = O(1)$, and plugging the values of a_n and b_n into $\frac{(a_n - 3)(b_n)}{(a_n + b_n - 3)^2 (a_n + b_n - 2)}$, we get $\frac{(a_n - 3)(b_n)}{(a_n + b_n - 3)^2 (a_n + b_n - 2)} = \frac{1}{n} \frac{(\theta_0 - \frac{3}{n})(1 - \theta_0)}{(1 - \frac{3}{n})^2 (1 - \frac{2}{n})}$, which is upper bounded by $\frac{C_1}{n}$ for some constant $C_1 > 0$. Hence,

$$\int |f_1^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) < \frac{C_1}{n}.$$

Similarly, we can also show that,

$$\int |f_2^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) < \frac{C_2}{n}.$$

Finally, from Proposition A.2.1, we get that $\mathcal{K}(\rho_n, \pi) < C + \frac{1}{2}\log(n)$ for some large constant C. Hence, $\mathcal{K}(\rho_n, \pi) < C_3\sqrt{n}$ for some constant $C_3 > 0$. Choosing $C = \max(C_1, C_2, C_3)$, we satisfy all the conditions of Assumption 1 and Theorem 3. □

Appendix B.3.3. Proof of Proposition 9

For the purpose of this proof, we choose ρ_n's with scaled Beta distribution with parameters $a_n = n(\theta_0/2)$ and $b_n = n(1-\theta_0/2)$. Since, ρ_n is a scaled Beta distribution with the scaling factors $m = 0.5$ and $c = 0$, the pdf of ρ_n is given by

$$\rho_n(\theta) = \frac{2}{\text{Beta}(a_n, b_n)}(2\theta)^{a_n}(1-2\theta)^{b_n}$$

Since this is a scaled distribution, $E_{\rho_n}[\theta] = 2\frac{a_n}{a_n+b_n} = \theta_0$ and there exists a constant $\sigma > 0$, $\text{Var}_{\rho_n}[\theta] = \frac{\sigma^2}{n}$. Now, we analyse the transition probabilities. For $i \in \{1, 2, \dots\}$, the Birth-Death process has transition probabilities

$$p_\theta(j|i) = \begin{cases} \theta & \text{if } j = i-1, \\ 1-\theta & \text{if } j = i+1. \end{cases}$$

If $i = 0$, then the Markov chain goes to 1 with probability 1. Hence with the convention $\log\frac{0}{0} = 0$ the ratio of the log of the transition probabilities becomes,

$$|\log p_{\theta_0}(X_1|X_0) - \log p_\theta(X_1|X_0)| = I_{[X_1=X_0+1]}\log\left[\frac{\theta_0}{\theta}\right] + I_{[X_1=X_0-1]}\log\left[\frac{1-\theta_0}{1-\theta}\right].$$

In this case, $m = 3$. $M_1^{(1)}(X_1, X_0) = I_{[X_1=X_0+1]}$ and $M_2^{(1)}(X_1, X_0) = I_{[X_1=X_0-1]}$. Define $M_3^{(1)}(X_1, X_0) := 1$. All these random variables are bounded. Define $f_1^{(1)}(\theta, \theta_0) := \log\left[\frac{\theta_0}{\theta}\right]$, $f_2^{(1)}(\theta, \theta_0) := \log\left[\frac{1-\theta_0}{1-\theta}\right]$ and $f_3^{(1)}(\theta, \theta_0) := 0$. Similarly as in the proof on Proposition 8,

$$\int [f_1^{(1)}(\theta, \theta_0)]^3 \rho_n(d\theta) < \frac{C_1}{n}, \text{ and}$$
$$\int [f_2^{(1)}(\theta, \theta_0)]^3 \rho_n(d\theta) < \frac{C_2}{n}.$$

The stationary distribution is given by $q_\theta(i) = (\frac{\theta}{1-\theta})^{i-1} q_\theta(1) \ \forall \ i \in 1, 2, \dots$, so that $q_\theta(i) = (1-\theta)(\frac{\theta}{1-\theta})^{i-1}$ Hence the log of the ratio of the invariant distribution becomes

$$\log q_0(i) - \log q_\theta(i) = \log\left[\frac{1-\theta_0}{1-\theta}\right] + (i-1)\log\left[\frac{\theta_0}{\theta}\right] - (i-1)\log\left[\frac{1-\theta_0}{1-\theta}\right] \tag{A38}$$

We define $M_1^{(2)}(X_0) := 1$, and $M_2^{(2)}(X_0) = M_3^{(2)}(X_0) := X_0 - 1$. We can write $\text{E}_{q^{(0)}}[M_2^{(2)}(X_0)]^2 = \sum_{i=1}^{\infty}(i-1)^2 q^{(0)}(i) < \sum_{i=1}^{\infty} i^2 q^{(0)}(i)$. We have chosen $q^{(0)}$ such that $\sum_{i=1}^{\infty} i^2 q^{(0)}(i)$ is bounded. Hence, $\text{E}_{q^{(0)}}[M_2^{(2)}(X_0)]^2 < \infty$. To verify Assumption i define, $f_1^{(2)}(\theta, \theta_0) = -f_3^{(2)}(\theta, \theta_0) := \log\left[\frac{1-\theta_0}{1-\theta}\right]$, and define $f_2^{(2)}(\theta, \theta_0) := \log\left[\frac{\theta_0}{\theta}\right]$. Therefore following the proof of Proposition 8,

$$\int |f_1^{(2)}(\theta, \theta_0)|^3 \rho_n(d\theta) = \int |f_3^{(2)}(\theta, \theta_0)|^3 \rho_n(d\theta) = \int |f_2^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) < \frac{C_2}{n}, \text{ and ,}$$
$$\int |f_2^{(2)}(\theta, \theta_0)|^3 \rho_n(d\theta) = \int |f_1^{(1)}(\theta, \theta_0)|^3 \rho_n(d\theta) < \frac{C_1}{n}.$$

Finally, we take the KL-divergence $\mathcal{K}(\rho_n, \pi)$. ρ_n follows a scaled Beta distribution on $(0, 1/2)$ with parameters $a_n = n(\theta_0/2)$ and $b_n = n(1-\theta_0/2)$, while π follows a scaled Beta distribution on $(0, 1/2)$ with parameters a and b. Thus,

$$\mathcal{K}(\rho_n, \pi) = \int_0^{\frac{1}{2}} \log\left(\frac{\rho_n(\theta)}{\pi(\theta)}\right)\rho_n(d\theta),$$

which, by substituting $t = 2\theta$, we get,

$$\mathcal{K}(\rho_n, \pi) = 2\int_0^{1} \log\left(\frac{\rho_n(t)}{\pi(t)}\right)\rho_n(dt).$$

$\int_0^1 \log\left(\frac{\rho_n(t)}{\pi(t)}\right)\rho_n(dt)$ is the KL-divergence between a Beta distribution with parameters a_n and b_n and a Beta distribution with parameters a and b. An application of Proposition A.2.1 gives us for a constant $C_1 > 0$,

$$\int_0^{1} \log\left(\frac{\rho_n(t)}{\pi(t)}\right)\rho_n(dt) < C_1 + \frac{1}{2}\log(n).$$

Hence we can say, $\mathcal{K}(\rho_n, \pi) < 2\left[C_1 + \frac{1}{2}\log(n)\right]$. Thus, we now get that for some constant $C_3 > 0$,

$$\mathcal{K}(\rho_n, \pi) < C_3\sqrt{n}.$$

Choosing $C = \max(C_1, C_2, C_3)$ we satisfy all of the conditions of Assumption 1 and thus by Theorem 3, we are complete the proof. □

Appendix B.3.4. Proof of Theorem 4

Part 1: Verifying condition (i) of Corollary 1 As in the proof of Theorem 2 substitute the true parameter θ_0 for θ_1 and θ for θ_2. We also set our initial distributions $q_1^{(0)}$ and $q_2^{(0)}$ to the known initial distribution $q^{(0)}$. A method similar to Equation (A35), yields

$$\mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \leq \sum_{i=1}^{n}\sum_{k=1}^{m} \mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})\right]|f_k^{(1)}(\theta, \theta_0)|.$$

Because $M_k^{(1)}$s satisfy Assumption 2, it follows by the application of Theorem 2.3, [21] that $\exists\ \lambda > 0$ such that for any $0 < \kappa \leq \lambda$, and for some $\zeta \in (0, 1)$ possibly depending upon λ,

$$\mathrm{E}\left[e^{\kappa M_k^{(1)}(X_i, X_{i-1})}\middle| X_1, X_0\right] \leq \zeta^{i-1} e^{\kappa M_k^{(1)}(X_1, X_0)} + \frac{1-\zeta^i}{1-\zeta}\mathcal{D}e^{\kappa a} \quad \text{for all } i > 1.$$

We rewrite $\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_1, X_0\right]$ as follows:

$$\begin{aligned}\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_1, X_0\right] &= \frac{\mathrm{E}[\kappa M_k^{(1)}(X_i, X_{i-1})|X_1, X_0]}{\kappa} \\ &\leq \frac{\mathrm{E}[e^{\kappa M_k^{(1)}(X_i, X_{i-1})}|X_1, X_0]}{\kappa}.\end{aligned}$$

Therefore, $\sum_{i=1}^{n} \mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})\right]$ can be upper bounded as,

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})\right] &= \sum_{i=1}^{n} \mathrm{E}\left[\kappa M_k^{(1)}(X_i, X_{i-1})|X_1, X_0\right]\kappa^{-1} \\ &\leq \sum_{i=1}^{n}\left[\zeta^{i-1}\mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{1-\zeta^i}{1-\zeta}\mathcal{D}e^{\kappa a}\right]\kappa^{-1}.\end{aligned}$$

Since, $\zeta \in (0,1)$, $\zeta^i < 1$. Hence, we can write that,

$$\begin{aligned}\sum_{i=1}^{n}\left[\zeta^{i-1}\mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{1-\zeta^i}{1-\zeta}\mathcal{D}e^{\kappa a}\right]\kappa^{-1} &\leq \sum_{i=1}^{n}\left[\zeta^{i-1}\mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{1}{1-\zeta}\mathcal{D}e^{\kappa a}\right]\kappa^{-1} \\ &= \left[\frac{1-\zeta^n}{1-\zeta}\mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{n}{1-\zeta}\mathcal{D}e^{\kappa a}\right]\kappa^{-1} \\ &\leq nL,\end{aligned}$$

for a large constant L. Therefore $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)})\rho_n(d\theta)$ can be upper bounded as follows,

$$\begin{aligned}\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)})\rho_n(d\theta) &\leq \int \sum_{k=1}^{m} nL|f_k^{(1)}(\theta,\theta_0)|\rho_n(d\theta) \\ &= \sum_{k=1}^{m} nL\int |f_k^{(1)}(\theta,\theta_0)|\rho_n(d\theta).\end{aligned}$$

By Assumption 1, $\int |f_k^{(1)}(\theta,\theta_0)|\rho_n(d\theta) < \frac{C}{n}$, hence,

$$\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)})\rho_n(d\theta) \leq nL\frac{C}{\sqrt{n}}.$$

Hence, for some $\epsilon_n^{(1)} = \mathrm{O}(\frac{1}{\sqrt{n}})$, we have obtained that, $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)})\rho_n(d\theta) \leq n\epsilon_n^{(1)}$.

Part 2: Verifying condition (ii) of Corollary 1: Similar to as in the proof of Theorem 3, we upper bound $\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta)$ by

$$\int \mathrm{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq \sum_{i,j=1}^{n}\left(\frac{4}{n} + 2n^{\delta/2}\left(\int C_{\theta_0,\theta}^{(i)}\rho_n(d\theta) + \int C_{\theta_0,\theta}^{(j)}\rho_n(d\theta)\right.\right. \tag{A39}$$

$$\left.\left. + \int \sqrt{C_{\theta_0,\theta}^{(i)}C_{\theta_0,\theta}^{(j)}}\rho_n(d\theta)\right)\right)\left(\alpha_{|i-j|-1}^{\delta/(2+\delta)}\right) \tag{A40}$$

$$+ \sum_{i=1}^{n}\left(\frac{4}{n} + 2n^{\delta/2}\int C_{\theta_0,\theta}^{(i)}\rho_n(d\theta)\right)\left(\alpha_{i-1}^{\delta/(2+\delta)}\right),$$

where $C_{\theta_0,\theta}$ is upper bounded as

$$C_{\theta_0,\theta}^{(i)} \leq \sum_{k=1}^{m} m^{1+\delta}\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})\right]^{2+\delta}|f_k^{(1)}(\theta,\theta_0)|^{2+\delta}.$$

There exists a constant C_δ depending upon δ such that,

$$\begin{aligned}[M_k^{(1)}]^{2+\delta}(X_i, X_{i-1}) &= \frac{\kappa^{2+\delta}[M_k^{(1)}]^{2+\delta}(X_i, X_{i-1})^{2+\delta}}{\kappa^{2+\delta}} \\ &\leq \frac{e^{\kappa M_k^{(1)}(X_i,X_{i-1})} + C_\delta}{\kappa^{2+\delta}}.\end{aligned}$$

By expressing $\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}\Big] = \mathrm{E}\Big[\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}|X_1, X_0\Big]\Big]$ and following a method similar to the previous part, we get,

$$\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}\Big] \leq \frac{\Big[\zeta^i \mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{1-\zeta^i}{1-\zeta}\mathcal{D}e^{\kappa a}\Big] + C_\delta}{\kappa^{2+\delta}}.$$

The fact that $0 < \zeta < 1$ implies that $0 < \zeta^i < \zeta$. This gives us the following,

$$\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}\Big] \leq \frac{\Big[\zeta \mathrm{E}e^{\kappa M_k^{(1)}(X_1,X_0)} + \frac{1}{1-\zeta}\mathcal{D}e^{\kappa a}\Big] + C_\delta}{\kappa^{2+\delta}}.$$

Since $\kappa < \lambda$, by the application of Jensen's inequality, we get

$$\begin{aligned}\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}\Big] &\leq \frac{\Big[\zeta \mathrm{E}e^{\lambda M_k^{(1)}(X_1,X_0)} + \frac{1}{1-\zeta}\mathcal{D}e^{\kappa a}\Big] + C_\delta}{\kappa^{2+\delta}} \\ &= \frac{\Big[\zeta \int e^{\lambda M_k^{(1)}(x_1,x_0)} p_{\theta_0}(x_1|x_0)D(x_0)dx_1dx_0 + \frac{1}{1-\zeta}\mathcal{D}e^{\kappa a}\Big] + C_\delta}{\kappa^{2+\delta}}.\end{aligned}$$

We know that $\int |f_k^{(1)}(\theta, \theta_0)|^{2+\delta}\rho_n(d\theta) < \frac{C}{n}$. Thus, following Assumption 1 we can say that, for a large constant L, $\int C_{\theta_0,\theta}^{(i)}\rho_n(d\theta) \leq \frac{L}{n}$. The rest of the proof follows similarly as in the proof of Theorem 3, and we obtain an $\epsilon_n^{(2)} = \mathrm{O}(\frac{n^{\delta/2}}{n})$, such that,

$$\int \mathrm{Var}[r_n(\theta, \theta_0)]\rho_n(d\theta) < n\epsilon_n^{(2)}.$$

Since, $\mathcal{K}(\rho_n, \pi) \leq \sqrt{n}C$, similar arguments as in the proof of Theorem 2 holds. The theorem is thus proved.

Appendix B.3.5. Proof of Theorem 5

Part 1: Verifying condition (i) of Corollary 1 As in the proof of Theorem 2 substitute the true parameter θ_0 for θ_1 and θ for θ_2. We also set $q_1^{(0)}$ and $q_2^{(0)}$ to the known initial distribution $q^{(0)}$. Similar to the steps leading to Equation (A35), we get

$$\mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)}) \leq \sum_{i=1}^{n}\sum_{k=1}^{m} \mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})\Big]|f_k^{(1)}(\theta, \theta_0)|.$$

Consider the term $\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})\Big]$. With $q_{\theta_0}^{(i-1)}$ the marginal distribution of X_{i-1}, we have

$$\begin{aligned}\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})\Big] &= \int M_k^{(1)}(x_i, x_{i-1})p_{\theta_0}(x_i|x_{i-1})q_{\theta_0}^{(i-1)}(x_{i-1})dx_idx_{i-1}. \\ \mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})\Big] &= \int M_k^{(1)}(x_i, x_{i-1})p_{\theta_0}(x_i|x_{i-1})p_{\theta_0}^{i-1}(x_{i-1}|x_0)q_{\theta_0}^{(0)}(x_0)dx_0dx_idx_{i-1}\end{aligned}$$

Recall that the marginal density satisfies $q_{\theta_0}^{(i-1)}(x_{i-1}) = \int p_{\theta_0}^{i-1}(x_{i-1}|x_0)q_{\theta_0}^{(0)}(x_0)d(x_0)$, where $p_{\theta_0}^i(\cdot|x_0)$ is the i-step transition probability. Then

$$\mathrm{E}\Big[M_k^{(1)}(X_i, X_{i-1})\Big] = \int \mathrm{E}\Big[M_k^{(1)}(X_i, x_{i-1})|x_{i-1}\Big]p_{\theta_0}^{i-1}(x_{i-1}|x_0)q_{\theta_0}^{(0)}(x_0)dx_0dx_{i-1}.$$

Since the Markov chain $\{X_n\}$ satisfies Assumption A.1.1, we know by the application of Theorem A.1.1 that $\{X_n\}$ is V-geometrically ergodic. Hence, $\exists\ \tau < 1$, $R < \infty$ such that $\forall\ |f| < V$

$$|\int f(x_{i-1}) p_{\theta_0}^{i-1}(x_{i-1}|x_0) dx_{i-1} - \int f(x_{i-1}) q_{\theta_0}(x_{i-1}) dx_{i-1}| < RV(x_0)\tau^{i-1},$$

where q_{θ_0} is the stationary distribution, implying that

$$\int f(x_{i-1}) p_{\theta_0}^{i-1}(x_{i-1}|x_0) dx_{i-1} < \int f(x_{i-1}) q_{\theta_0}(x_{i-1}) dx_{i-1} + RV(x_0)\tau^{i-1}.$$

By the application of Jensen's inequality we get $\left(\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_{i-1}\right]\right)^{2+\delta} \leq \mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})^{2+\delta}|X_{i-1}\right] < V(X_{i-1})$. Since $V(\cdot) \geq 1$, it follows from the previous expression that $\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_{i-1}\right] < V(X_{i-1})^{1/(2+\delta)} \leq V(X_{i-1})$. Thus, setting $f(x) = \mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_{i-1} = x\right]$, we obtain

$$\begin{aligned}\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})\right] &< \int \left[\mathrm{E}\left[M_k^{(1)}(X_i, X_{i-1})|X_{i-1}\right] q_{\theta_0}(x_i) dx_{i-1} + RV(x_0)\tau^{i-1}\right] q^{(0)}(x_0) dx_0 \\ &= \mathrm{E}[M_k^{(1)}(X_1, X_0)] + \tau^{i-1} \int RV(x_0) q^{(0)}(x_0) dx_0.\end{aligned}$$

Summing from $i = 1$ to n, we get

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{E}\left[M_k^{(1)}(x_i, x_{i-1})\right] &< n\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \sum_{i=1}^{n} \tau^{i-1} \int RV(x_0) q^{(0)}(x_0) dx_0 \\ &= n\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \frac{1-\tau^n}{1-\tau} \int RV(x_0) q^{(0)}(x_0) dx_0.\end{aligned}$$

This gives us the following bound on $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \rho_n(d\theta)$:

$$\begin{aligned}\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \rho_n(d\theta) \leq \sum_{k=1}^{m} \left[n\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \frac{1-\tau^n}{1-\tau} \int RV(x_0) D(x_0) dx_0\right] \\ \times \int |f_k^{(1)}(\theta, \theta_0)| \rho_n(d\theta).\end{aligned}$$

By Assumption 1, $\int |f_k^{(1)}(\theta, \theta_0)| \rho_n(d\theta) < \frac{C}{\sqrt{n}}$. Hence, we can rewrite the previous expression as

$$\begin{aligned}\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \rho_n(d\theta) &\leq \sum_{k=1}^{m} \left[n\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \frac{1-\tau^n}{1-\tau} \int RV(x_1) D(x_1) dx_1\right] \frac{C}{\sqrt{n}} \\ &= nm \left[\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \frac{1-\tau^n}{n(1-\tau)} \int RV(x_0) D(x_0) dx_0\right] \frac{C}{\sqrt{n}}.\end{aligned}$$

Since, $\tau < 1$, $0 < 1 - \tau^n < 1$, and we rewrite the previous equation as,

$$\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \rho_n(d\theta) \leq nm \left[\mathrm{E}[M_k^{(1)}(X_1, X_0)] + \frac{1}{n(1-\tau)} \int RV(x_0) D(x_0) dx_0\right] \frac{C}{\sqrt{n}}.$$

Hence, there exists an $\epsilon_n^{(1)} = \mathrm{O}(\frac{1}{\sqrt{n}})$ such that $\int \mathcal{K}(P_{\theta_0}^{(n)}, P_{\theta}^{(n)}) \rho_n(d\theta) \leq n\epsilon_n^{(1)}$.

Part 2: Verifying condition (ii) of Corollary 1: Similar to as in the proof of Theorem 3, we upper bound $\int \text{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta)$ by

$$\int \text{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) \leq \sum_{i,j=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\left(\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)+\int C^{(j)}_{\theta_0,\theta}\rho_n(d\theta)\right.\right. \tag{A41}$$

$$\left.\left.+\int\sqrt{C^{(i)}_{\theta_0,\theta}C^{(j)}_{\theta_0,\theta}}\rho_n(d\theta)\right)\right)\left(\alpha^{\delta/(2+\delta)}_{|i-j|-1}\right) \tag{A42}$$

$$+\sum_{i=1}^{n}\left(\frac{4}{n}+2n^{\delta/2}\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta)\right)\left(\alpha^{\delta/(2+\delta)}_{i-1}\right),$$

where $C_{\theta_0,\theta}$ is upper bounded as

$$C^{(i)}_{\theta_0,\theta} \leq \sum_{k=1}^{m} m^{1+\delta}\text{E}\left[M^{(1)}_k(X_i,X_{i-1})\right]^{2+\delta}|f^{(1)}_k(\theta,\theta_0)|^{2+\delta}.$$

Since $\text{E}\left[M^{(1)}_k(X_i,X_{i-1})^{2+\delta}|X_{i-1}\right] < V(X_{i-1})$, by a similar application of V-geometric ergodicity, we can say that, $\exists\ 0<\tau<1$, such that

$$\text{E}\left[M^{(1)}_k(X_i,X_{i-1})\right]^{2+\delta} \leq n\text{E}[M^{(1)}_k(X_1,X_0)]^{2+\delta}+\tau^{i-1}\int RV(x_0)D(x_0)dx_0,$$

which, by the fact that $\tau^{i-1}<\tau$, gives us,

$$\text{E}\left[M^{(1)}_k(X_i,X_{i-1})\right]^{2+\delta} \leq \text{E}[M^{(1)}_k(X_1,X_0)]^{2+\delta}+\tau\int RV(x_0)D(x_0)dx_0.$$

By Assumption 1, we know that, $\int |f^{(1)}_k(\theta,\theta_0)|^{2+\delta}\rho_n(d\theta) < \frac{C}{n}$. Hence, for a large constant L, $\int C^{(i)}_{\theta_0,\theta}\rho_n(d\theta) \leq \frac{L}{n}$. We also see that since the chain is geometrically ergodic, by the application of Equation (A4), $\sum_{k\geq 1}\alpha_k^{\delta/(2+\delta)} < +\infty$. The rest of the proof follows similarly as in the proof of Theorem 3, and we obtain an $\epsilon^{(2)}_n = \text{O}(\frac{n^{\delta/2}}{n})$, such that,

$$\int \text{Var}[r_n(\theta,\theta_0)]\rho_n(d\theta) < n\epsilon^{(2)}_n.$$

Since, $\mathcal{K}(\rho_n,\pi) \leq \sqrt{n}C$, similar arguments as in the proof of Theorem 2 holds. The theorem is thus proved. □

Appendix B.3.6. Proof of Proposition 10

For the purpose of the proof, we choose ρ_n's with scaled Beta distribution with parameters $a_n = n\frac{1+\theta_0}{2}$ and $b_n = n\frac{1-\theta_0}{2}$. Since, ρ_n is a scaled Beta distribution with the scaling factors $m=2$ and $c=-1$, the pdf of ρ_n is given by

$$\rho_n(\theta) = \frac{1}{2\text{Beta}(a_n,b_n)}\left(\frac{1+\theta}{2}\right)^{a_n}\left(\frac{1-\theta}{2}\right)^{b_n}$$

Since this is a scaled distribution, $E_{\rho_n}[\theta] = 2\frac{a_n}{a_n+b_n}-1 = \theta_0$ and there exists a constant $\sigma>0$, $\text{Var}_{\rho_n}[\theta] = \frac{\sigma^2}{n}$. We now analyse the log-ratio of the transition probabilities for the Markov chain,

$$\log p_{\theta_0}(X_n|X_{n-1}) - \log p_\theta(X_n|X_{n-1}) = 2X_nX_{n-1}(\theta-\theta_0)+X^2_{n-1}(\theta_0^2-\theta^2).$$

Observe that in this setting, $M_1^{(1)}(X_n, X_{n-1}) = |X_n X_{n-1}|$ and $M_2^{(1)}(X_n, X_{n-1}) = X_n^2$. Next, using the fact that

$$\mathrm{E}[|X_n|^{2+\delta}|X_{n-1}] = \mathrm{E}[|X_n - \theta_0 X_{n-1} + \theta_0 X_{n-1}|^{2+\delta}|X_{n-1}],$$

and by an application of triangle inequality, we obtain

$$\begin{aligned}
\mathrm{E}[|X_n|^{2+\delta}|X_{n-1}] &\leq \mathrm{E}\Big[(|X_n - \theta_0 X_{n-1}| + |\theta_0 X_{n-1}|)^{2+\delta}|X_{n-1}\Big] \\
&= \mathrm{E}\left[\left(2\frac{|X_n - \theta_0 X_{n-1}| + |\theta_0 X_{n-1}|}{2}\right)^{2+\delta}|X_{n-1}\right] \\
&= \mathrm{E}\left[2^{2+\delta}\left(\frac{|X_n - \theta_0 X_{n-1}| + |\theta_0 X_{n-1}|}{2}\right)^{2+\delta}|X_{n-1}\right].
\end{aligned}$$

Now by using Jensen's inequality we get,

$$\begin{aligned}
\mathrm{E}[|X_n|^{2+\delta}|X_{n-1}] &\leq \mathrm{E}\left[2^{2+\delta}\left(\frac{|X_n - \theta_0 X_{n-1}|^{2+\delta} + |\theta_0 X_{n-1}|^{2+\delta}}{2}\right)|X_{n-1}\right] \\
&= 2^{1+\delta}\mathrm{E}\Big[|X_n - \theta_0 X_{n-1}|^{2+\delta}|X_{n-1}\Big] + 2^{1+\delta}|\theta_0 X_{n-1}|.
\end{aligned}$$

We know if $Y \sim N(\mu, \sigma^2)$, then $\mathrm{E}|Y-\mu|^p = \sigma^p \frac{2^{\frac{p}{2}}\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}$. Consequently,

$$\mathrm{E}[|X_n|^{2+\delta}|X_{n-1}] \leq 2^{1+\delta}\left[\frac{2^{\frac{2+\delta}{2}}\Gamma(\frac{3+\delta}{2})}{\sqrt{\pi}}\right] + 2^{1+\delta}|\theta_0 X_{n-1}|^{2+\delta}. \tag{A43}$$

It follows that,

$$\begin{aligned}
\mathrm{E}[M_1^{(1)}(X_n, X_{n-1})^{2+\delta}|X_{n-1}] &\leq 2^{1+\delta}\left[\frac{2^{\frac{2+\delta}{2}}\Gamma(\frac{3+\delta}{2})}{\sqrt{\pi}}\right]|X_{n-1}|^{2+\delta} + 2^{1+\delta}|\theta_0|^{2+\delta}|X_{n-1}|^{4+2\delta} \\
&\leq \left(2^{1+\delta}\left[\frac{2^{\frac{2+\delta}{2}}\Gamma(\frac{3+\delta}{2})}{\sqrt{\pi}}\right] + 2^{1+\delta}|\theta_0|^{2+\delta}\right)(|X_{n-1}|^{4+2\delta} + 1).
\end{aligned}$$

Since $\theta_0 < 1$, we can say,

$$\mathrm{E}[M_1^{(1)}(X_n, X_{n-1})^{2+\delta}|X_{n-1}] \leq \left(2^{1+\delta}\left[\frac{2^{\frac{2+\delta}{2}}\Gamma(\frac{3+\delta}{2})}{\sqrt{\pi}}\right] + 2^{1+\delta}\right)(|X_{n-1}|^{4+2\delta} + 1).$$

Define a constant $C_\delta := \left(2^{1+\delta}\left[\frac{2^{\frac{2+\delta}{2}}\Gamma(\frac{3+\delta}{2})}{\sqrt{\pi}}\right] + 2^{1+\delta}\right)$. The above term then becomes,

$$\mathrm{E}[M_1^{(1)}(X_n, X_{n-1})^{2+\delta}|X_{n-1}] \leq C_\delta(|X_{n-1}|^{4+2\delta} + 1).$$

Next we analyse the term $M_2^{(1)}(X_n, X_{n-1})$.

$$\begin{aligned}
\mathrm{E}\Big[M_2^{(1)}(X_n, X_{n-1})^{2+\delta}|X_{n-1}\Big] &= \mathrm{E}[X_{n-1}^{4+2\delta}|X_{n-1}] \\
&= X_{n-1}^{4+2\delta} \\
&\leq C_\delta(X_{n-1}^{4+2\delta} + 1).
\end{aligned}$$

Then, defining $V(x) := C_\delta(x^{4+2\delta}+1)$ it follows that,

$$\mathrm{E}[V(X_n)|X_{n-1}] = \mathrm{E}\Big[C_\delta(X_n^{4+2\delta}+1)|X_{n-1}\Big].$$

Using a technique similar to Equation (A43) we get,

$$\mathrm{E}\Big[C_\delta(X_n^{4+2\delta}+1)|X_{n-1}\Big] \leq \left[C_\delta(2^{3+2\delta}\left[\frac{2^{\frac{4+2\delta}{2}}\Gamma(\frac{5+2\delta}{2})}{\sqrt{\pi}}\right] + 2^{3+2\delta}|\theta_0 X_{n-1}|^{4+2\delta}+1)\right].$$

Define another constant $C'_\delta := C_\delta\left(2^{3+2\delta}\left[\frac{2^{\frac{4+2\delta}{2}}\Gamma(\frac{5+2\delta}{2})}{\sqrt{\pi}}\right] - 2^{3+2\delta}|\theta_0|^{4+2\delta}+1\right)$. Since $\delta > 0$, $\frac{2^{\frac{4+2\delta}{2}}\Gamma(\frac{5+2\delta}{2})}{\sqrt{\pi}} > 1$. Furthermore, since $|\theta_0| < 1$, so is $|\theta_0|^{4+2\delta}$. Hence,

$$2^{3+2\delta}\left[\frac{2^{\frac{4+2\delta}{2}}\Gamma(\frac{5+2\delta}{2})}{\sqrt{\pi}}\right] - 2^{3+2\delta}|\theta_0|^{4+2\delta} > 0.$$

Hence, we have shown that,

$$\mathrm{E}[V(X_n)|X_{n-1}] \leq (2^{3+2\delta}|\theta_0|^{4+2\delta})C_\delta(X_{n-1}^{4+2\delta}+1) + C'_\delta.$$

Since $|\theta_0| < 2^{\frac{1}{4+2\delta}-1}$, $2^{3+2\delta}|\theta_0|^{4+2\delta} < 1$, and we can express the above equation as,

$$\mathrm{E}[V(X_n)|X_{n-1}] \leq V(X_{n-1}) + C'_\delta.$$

Define the set $C(m) := \{x : |x|^{4+2\delta}+1 \leq m\}$. From Proposition 11.4.2, [20], for a large enough m, $C(m)$ forms a petite set. Thus, we have proved that $V(x)$ as defined in this example satisfies Assumption A.1.1, and $\{X_n\}$ is V-geometrically ergodic. The $f_j^{(1)}$'s corresponding to Assumption 1 are given by $f_1^{(1)}(\theta,\theta_0) = (\theta-\theta_0)$ and $f_2^{(1)}(\theta,\theta_0) = (\theta_0^2-\theta^2)$. Therefore, it follows that,

$$\begin{aligned}
&\partial_\theta f_1^{(1)} = 1,\\
&\partial_\theta f_2^{(1)} = -2\theta \text{ and}\\
&\quad -2 < -2\theta < 2.
\end{aligned}$$

Since $f_1^{(1)}(\theta_0,\theta_0) = f_2^{(1)}(\theta_0,\theta_0) = 0$, We just showed that they also have bounded partial derivatives. We also know that $|\theta| < 1$. Hence, by Proposition 4 $f_j^{(1)}$'s satisfy the conditions of Assumption 1.

The invariant distribution for the simple linear model Markov-chain under parameter θ is given by a gaussian distribution with mean 0 and variance $\frac{1}{1-\theta^2}$. In other words,

$$q_\theta(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1-\theta^2}{2}x^2}.$$

Analyzing the log likelihood yields,

$$\begin{aligned}
\log q_0(x) - \log q_\theta(x) &= -\frac{x^2}{2}(1-\theta_0^2) + \frac{x^2}{2}(1-\theta^2)\\
&= \frac{x^2}{2}(\theta_0^2-\theta^2).
\end{aligned}$$

Let $f_1^{(2)}(\theta_0, \theta_0) = (\theta_0^2 - \theta^2)$ and $f_1^{(2)}(\theta_0, \theta_0) = 0$. Since $f_1^{(2)}(\theta_0, \theta_0) = f_2^{(1)}(\theta_0, \theta_0)$, by following arguments similar as before, can conclude that $f_1^{(2)}(\theta_0, \theta_0)$ also satisfies the requirements of Assumption 1. Let $M_1^{(2)}(x) = \frac{x^2}{2}$ and define $M_2^{(2)}(x) := 1$. Let $X_0 \sim q_1^{(0)}$. As long as $\int x^{4+2\delta} q_1^{(0)}(x)dx < \infty$, we satisfy all the conditions required for Theorem 5. Finally we need to verify the condition that $\mathcal{K}(\rho_n, \pi) < C\sqrt{n}$ for some constant $C > 0$. The KL-divergence $\int \log\left(\frac{\rho_n(\theta)}{\pi(\theta)}\right)\rho_n(d\theta)$ becomes,

$$\mathcal{K}(\rho_n, \pi) = \int_{-1}^{1} \log\left(\frac{1}{2\text{Beta}(a_n, b_n)}\left(\frac{1+\theta}{2}\right)^{a_n}\left(\frac{1-\theta}{2}\right)^{b_n}\right) \times \frac{1}{2\text{Beta}(a_n, b_n)}\left(\frac{1+\theta}{2}\right)^{a_n}\left(\frac{1-\theta}{2}\right)^{b_n} d\theta.$$

Substituting, $y = \frac{1+\theta}{2}$, we get,

$$\begin{aligned}\mathcal{K}(\rho_n, \pi) &= \int_0^1 \log\left(\frac{1}{2\text{Beta}(a_n, b_n)}(y)^{a_n}(1-y)^{b_n}\right)\frac{1}{2\text{Beta}(a_n, b_n)}(y)^{a_n}(1-y)^{b_n}dy \\ &= \int_0^1 \log\left(\frac{1}{2}\right)\frac{1}{\text{Beta}(a_n, b_n)}(y)^{a_n}(1-y)^{b_n}dy \\ &\quad + \int_0^1 \log\left(\frac{1}{\text{Beta}(a_n, b_n)}(y)^{a_n}(1-y)^{b_n}\right)\frac{1}{\text{Beta}(a_n, b_n)}(y)^{a_n}(1-y)^{b_n}.\end{aligned}$$

The first term integrates up to $\log(1/2)$. The second term is the KL-divergence between a Uniform and Beta distribution with parameters $a_n = n\frac{1+\theta_0}{2}$ and $b_n = n(1 - \frac{1+\theta_0}{2})$ and support $[0, 1]$. Following Lemma A.2.1 it follows that $\mathcal{K}(\rho_n, \pi)$ is upper bounded by,

$$\mathcal{K}(\rho_n, \pi) < \log(1/2) + C_1 + \frac{1}{2}\log(n) < C\sqrt{n},$$

for some large constant C. This completes the proof. □

Appendix B.4. Proofs for Misspecified Models

Proof of Theorem 6

As in the proof of Theorem 1, following Equation (A13), we note that,

$$\begin{aligned}\int D_{\alpha^{re}}(P_\theta^{(n)}, P_{\theta_0}^{(n)})\tilde{\pi}_{n,\alpha^{re}|X^n}(d\theta) \leq \frac{\alpha^{re}}{1-\alpha^{re}}\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) \\ + \frac{\alpha^{re}}{1-\alpha^{re}}\sqrt{\frac{\text{Var}[\int r_n(\theta, \theta_0)\rho_n(d\theta)]}{\eta}} + \frac{\mathcal{K}(\rho_n, \pi) - \log(\epsilon)}{1-\alpha^{re}}.\end{aligned} \tag{A44}$$

Following from Equations (23) and (26), we get that,

$$\int \mathcal{K}(P_{\theta_0}^{(n)}, P_\theta^{(n)})\rho_n(d\theta) \leq \text{E}[r_n(\theta_0, \theta_n^*)] + n\epsilon_n,$$

and

$$\int \text{Var}[r_n(\theta, \theta_0)]\rho_n(d\theta) \leq 2n\epsilon_n + 2\text{Var}[r_n(\theta_n^*, \theta_0)].$$

Plugging these into Equation (A44), we are done. □

References

1. Wainwright, M.J.; Jordan, M.I. Introduction to Variational Methods for Graphical Models. *Found. Trends Mach. Learn.* **2008**, *1*, 1–103. [CrossRef]
2. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin, Germany, 2006.
3. Ormerod, J.T.; Wand, M.P. Explaining variational approximations. *Am. Stat.* **2010**, *64*, 140–153. [CrossRef]
4. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
5. Jaiswal, P.; Rao, V.; Honnappa, H. Asymptotic Consistency of α-Rényi-Approximate Posteriors. *J. Mach. Learn. Res.* **2020**, *21*, 1–42.
6. Li, Y.; Turner, R.E. Rényi divergence variational inference. In Proceedings of the 30th Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29, pp. 1073–1081.
7. Dieng, A.B.; Tran, D.; Ranganath, R.; Paisley, J.; Blei, D. Variational Inference via χ Upper Bound Minimization. In Proceedings of the 31th Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
8. Wang, Y.; Blei, D.M. Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.* **2019**, *114*, 1147–1161. [CrossRef]
9. Zhang, F.; Gao, C. Convergence rates of variational posterior distributions. *Ann. Stat.* **2020**, *48*, 2180–2207. [CrossRef]
10. Ghosal, S.; Ghosh, J.K.; Van Der Vaart, A.W. Convergence rates of posterior distributions. *Ann. Stat.* **2000**, *28*, 500–531. [CrossRef]
11. Shen, X.; Wasserman, L. Rates of convergence of posterior distributions. *Ann. Stat.* **2001**, *29*, 687–714.
12. Rousseau, J. On the frequentist properties of Bayesian nonparametric methods. *Annu. Rev. Stat. Its Appl.* **2016**, *3*, 211–231. [CrossRef]
13. Ghosal, S.; Van Der Vaart, A.W. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of Normal densities. *Ann. Stat.* **2001**, 1233–1263.
14. Bhattacharya, A.; Pati, D.; Yang, Y. Bayesian fractional posteriors. *Ann. Stat.* **2019**, *47*, 39–66. [CrossRef]
15. Alquier, P.; Ridgway, J. Concentration of tempered posteriors and of their variational approximations. *Ann. Stat.* **2020**, *48*, 1475–1497. [CrossRef]
16. Yang, Y.; Pati, D.; Bhattacharya, A. α-variational inference with statistical guarantees. *Ann. Stat.* **2020**, *48*, 886–905. [CrossRef]
17. Jaiswal, P.; Honnappa, H.; Rao, V.A. Risk-sensitive variational Bayes: Formulations and bounds. *arXiv* **2019**, arXiv:1903.05220.
18. Bradley, R.C. Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions. *Probab. Surv.* **2005**, *2*, 107–144. [CrossRef]
19. Ibragimov, I.A. Some limit theorems for stationary processes. *Theory Probab Appl.* **1962**, *7*, 349–382. [CrossRef]
20. Meyn, S.P.; Tweedie, R.L. *Markov Chains and Stochastic Stability*; Springer: Berlin, Germany, 2012.
21. Hajek, B. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Adv. Appl. Probab.* **1982**, 502–525. [CrossRef]
22. Birgé, L. Robust testing for independent non identically distributed variables and Markov chains. In *Specifying Statistical Models*; Springer: Berlin, Germany, 1983; pp. 134–162.
23. Ryabko, D. Testing statistical hypotheses about ergodic processes. In Proceedings of the IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering, Novosibirsk, Russia, 21–25 July 2008.
24. Lacoste-Julien, S.; Huszár, F.; Ghahramani, Z. Approximate inference for the loss-calibrated Bayesian. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011.
25. Jaiswal, P.; Honnappa, H.; Rao, V.A. Asymptotic consistency of loss-calibrated variational Bayes. *Stat* **2020**, *9*, e258. [CrossRef]
26. Jones, G.L. On the Markov chain central limit theorem. *Probab. Survey.* **2004**, *1*, 299–320. [CrossRef]
27. Alzer, H. On some inequalities for the gamma and psi functions. *Math. Comput.* **1997**, *66*, 373–389. [CrossRef]
28. Donsker, M.D.; Varadhan, S.S. Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.* **1975**, *28*, 1–47. [CrossRef]