Submitted to Bernoulli
arXiv: arXiv:0000.0000

Bayesian Consistency for a Nonparametric Stationary Markov Model

MINWOO CHAE[1] and STEPHEN G. WALKER[1]

[1] *Department of Mathematics, The University of Texas at Austin*
E-mail: minwooo.chae@gmail.com; s.g.walker@math.utexas.edu

We consider posterior consistency for a Markov model with a novel class of nonparametric prior. In this model, the transition density is parameterized via a mixing distribution function. Therefore, the Wasserstein distance between mixing measures can be used to construct neighborhoods of a transition density. The Wasserstein distance is sufficiently strong, *e.g.* if the mixing distributions are compactly supported, it dominates the sup–L_1 metric. We provide sufficient conditions for posterior consistency with respect to the Wasserstein metric provided that the true transition density is also parametrized via a mixing distribution. In general, when it is not be parameterized by a mixing distribution, we show the posterior distribution is consistent with respect to the average L_1 metric. Also, we provide a prior whose support is sufficiently large to contain most smooth transition densities.

Keywords: Kullback–Leibler support, mixtures, nonparametric Markov model, posterior consistency, Wasserstein metric.

1. Introduction

The Bayesian nonparametric modeling of first order time series models (i.e. Markov models) is now widely routine following the development of mixture transition densities which are tractable and describe stationary densities which are also mixture models. See for example [2] and the references within that paper. While there is now a substantial literature on Bayesian nonparametric posterior consistency for i.i.d. observations and regression models, there is a lack of decent results for Bayesian nonparametric Markov models. This is mainly due to the lack of suitable priors and metrics. The aim in this paper is to provide a novel class of nonparametric prior with techniques for demonstrating posterior consistency.

To set the scene and the notation; consider a time homogeneous $\mathcal{X}$-valued Markov chain $\mathbf{X} = (X_n)_{n\geq 0}$. The underlying probability law of the Markov chain $\mathbf{X}$ is completely determined by the initial distribution and transition kernel whose Lebesgue densities are assumed to exist, and denoted by $(y, x) \mapsto f(y|x)$. In most statistical applications, the main interest is to infer the transition kernel density f which may be considered as a finite or infinite dimensional parameter.

As is typical in i.i.d. cases, the performance of a statistical methodology, from both a frequentist and Bayesian context, may be evaluated via its large sample properties.

For example, if $\mathbf{X}$ is generated from a true transition density f_0, a reasonable estimator $\hat{f}_n$ is expected to be consistent for estimating f_0, that is, $d(\hat{f}_n, f_0) \to 0$ almost surely, or in probability, for some metric d. If f can be parametrized by a finite dimensional parameter, the Euclidean metric is a natural choice for d. In infinite dimensional cases, however, there is no natural counterpart to commonly used metrics in i.i.d. models such as the Hellinger and total variation metrics. The main difficulty arises because the distance $d(f_1(\cdot|x), f_2(\cdot|x))$ between two conditional densities depends on x, so it cannot define suitable neighborhoods on $\mathscr{F}$, where $\mathscr{F}$ is the set of every transition density of a positive Harris chain. To define a suitable topology on $\mathscr{F}$, the dependence on x should somehow be eliminated.

There is to date little work on constructing suitable neighborhoods on $\mathscr{F}$. Tang and Ghosal [26] considered three kinds of metric which can be applied for general families of transition density. The first one is a metric between stationary distributions, but this is too weak because different transition kernels can yield the same stationary distribution. The second and third types are maximized and integrated distances defined by

$$d_{\max}(f_1, f_2) = \sup_{x\in\mathcal{X}} d(f_1(\cdot|x), f_2(\cdot|x)), \quad \text{and} \quad d_{\text{avg}}(f_1, f_2) = \int d(f_1(\cdot|x), f_2(\cdot|x))d\nu(x),$$

where ν is some probability measure. The former, $d_{\max}$, is too strong and cannot be used unless $\mathcal{X}$ is compact, as mentioned in [9, 1]. The average distance d_{avg} is often useful for consistency, but the result in [26] is limited to a specific prior whose support is not sufficiently large. The same authors also considered the minimized distance $d_{\min}(f_1, f_2) = \inf_{x\in\mathcal{X}} d(f_1(\cdot|x), f_2(\cdot|x))$ in [9], but this is not sufficiently strong, and can be highly unsuitable for some important models [1]. Antoniano-Villalobos and Walker [1] extended the idea of minimized distance, and considered the metric

$$\tilde{d}_{\min}(f_1, f_2) = \inf_{x\in\mathcal{X}} d(\tilde{f}_1(\cdot, \cdot|x), \tilde{f}_2(\cdot, \cdot|x)),$$

where $\tilde{f}(y, z|x) = f(z|y)f(y|x)$ is a bivariate extension of the transition density f. Though it is not a metric, they found that under certain conditions, $\tilde{d}_{\min}$ yields a strong topology on $\mathscr{F}$. However, it still requires strong assumptions such as the compactness of $\mathcal{X}$.

In this paper, we consider posterior consistency for a Markov model with a novel class of nonparametric prior, which is an extension of [2]. In this model, the transition density is parameterized by a mixing distribution function, so a metric between mixing distributions can be used to construct neighborhoods of a transition density. If f_0 belongs to the model, *i.e.* it is also represented via a mixture, it is shown under reasonable conditions that the posterior is consistent with respect to the Wasserstein metric. The Wasserstein distance is a sufficiently strong metric in this model *e.g.* if mixing distributions are compactly supported, it dominates $\sup_{x\in\mathcal{C}} d(f_1(\cdot|x), f_2(\cdot|x))$ for any compact set $\mathcal{C}$, where d is the Hellinger or total variation distance. From one perspective the Wasserstein consistency in this paper can be seen as a time series version of the i.i.d. results for the mixing distribution obtained by [22].

For general f_0, which may not be parameterized by a mixing distribution, the Wasserstein distance cannot be used as a metric for consistency since no mixing distribution

exists. In this case, the posterior distribution is consistent in the average L_1 metric, for all f_0 in the Kullback–Leibler (KL) support of the prior. We provide a prior whose KL support is sufficiently large to contain most smooth transition densities. To the best of our knowledge, a nonparametric prior for transition densities with such a large support is not known in the literature.

The remainder of this paper is organized as follows. Section 2 presents the model and prior with three examples. The main results pertaining to posterior consistency in the Wasserstein and average L_1 metrics are given in Sections 3 and 4, respectively. All proofs are deferred to Section 5.

Before proceeding, it is useful to establish some notation. For any two densities f and g with respect to a σ-finite measure μ, d_H and d_V denote the Hellinger and total variation (L_1) metrics defined by $d_H^2(f,g) = \int(\sqrt{f}-\sqrt{g})^2 d\mu$ and $d_V(f,g) = \int |f-g| d\mu$, respectively. The Kullback–Leibler divergence is defined as $K(f,g) = \int \log(f/g) f d\mu$. Also, for two probability measures P and Q on a set $\Theta \subset \mathbb{R}^d$, define the L_1-Wesserstein metric by

$$d_W(P,Q) = \inf_{J \in \mathcal{J}(P,Q)} \left\{ \int_{\Theta\times\Theta} \|\theta_1 - \theta_2\| \, dJ(\theta_1,\theta_2) \right\}, \tag{1.1}$$

where $\|\cdot\|$ is the Euclidean norm and $\mathcal{J}(P,Q)$ is the set of all joint distributions J with marginals P and Q. For a given bivariate density $f(y,x)$, we use the same notation f for the marginal density (of the second component) $f(x) = \int f(y,x)dy$ and conditional density $f(y|x) = f(y,x)/f(x)$. Denote the standard normal density as ϕ, and let $\phi_\sigma(x) = \sigma^{-1}\phi(x/\sigma)$. For a metric space (S,d), its Borel σ-algebra is denoted as $\mathcal{B}(S)$ and $N(\epsilon,S,d)$ is the minimum ϵ-covering number. Let $a \lesssim b$ denote that a is smaller than b up to a constant multiple.

2. Model and prior

In this section we introduce the nonparametric Markov model studied in [2]. We provide three specific examples including the one considered in [2]. The other two examples possess interesting statistical properties.

Assume that $\mathcal{X} = \mathbb{R}$ and X_0 has a known distribution, so probabilistic properties of $\mathbf{X} = (X_n)_{n=0}^\infty$ are completely determined by the transition density f. The Markov chain $\mathbf{X}$ will be assumed to be stationary. To be more precise, we first introduce some notions concerning the stability of Markov chains. Readers are referred to the monograph [20] for details.

For $A \in \mathcal{B}(\mathbb{R})$, let $\tau_A = \inf\{n \geq 1 : X_n \in A\}$ be the first time a chain reaches A. For a positive measure φ on $(\mathbb{R},\mathcal{B}(\mathbb{R}))$, $\mathbf{X}$ is called *φ-irreducible* if $\mathbb{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathbb{R}$ and A with $\varphi(A) > 0$, where $\mathbb{P}_x$ denotes the probability of events conditional on $X_0 = x$. If, furthermore, $\mathbb{P}_x(X_n \in A \text{ i.o.}) = 1$ for every $x \in A$ and $A \in \mathcal{B}(\mathbb{R})$ with $\varphi(A) > 0$, it is called a *Harris recurrent* chain, where $\{X_n \in A \text{ i.o.}\} = \cap_{m=1}^\infty \cup_{n=m}^\infty \{X_n \in A\}$. If $\mathbf{X}$ is φ-irreducible and has a *stationary distribution*, that is, there exists a probability meausre ν such that $\int\int_A f(y|x)dyd\nu(x) = \nu(A)$ for every $A \in \mathcal{B}(\mathbb{R})$, it is called a *positive*

chain. If $\mathbf{X}$ is Harris recurrent and positive, then it is called a *positive Harris* chain. Let $\mathscr{F}$ be the set of every transition density $f(\cdot|\cdot)$ of a $\mathbb{R}$-valued positive Harris chain. Note that if the transition density f of the Markov chain $\mathbf{X}$ belongs to $\mathscr{F}$, then there exists a unique stationary distribution ν, and for every ν-integrable function h, $n^{-1}\sum_{i=1}^{n} h(X_i)$ converges $\mathbb{P}_x$-almost-surely to $\int h d\nu$ for every $x \in \mathbb{R}$.

For a prior on (a subset of) $\mathscr{F}$, we consider a class of bivariate mixtures. Let Θ be a subset of a Euclidean space, $\mathscr{P}$ a set of Borel probability measures on Θ, and $\mathscr{K} = \{K_\theta : \theta \in \Theta\}$ be a class of probability measures on $\mathbb{R} \times \mathbb{R}$, where K_θ has the continuous and positive Lebesgue density $(y, x) \mapsto k_\theta(y, x)$. If not explicitly specified in examples, we assume that $\mathscr{P}$ is the set of every Borel probability measure. Equip $\mathscr{P}$ with the weak topology, and let Π be the prior distribution on $(\mathscr{P}, \mathcal{B}(\mathscr{P}))$. Let

$$f_P(y, x) = \int k_\theta(y, x) dP(\theta) \tag{2.1}$$

be the density of a bivariate mixture. As explained in the introduction, the notation f_P is used to denote the conditional density

$$f_P(y|x) = \frac{f_P(y, x)}{\int f_P(y, x) dy}, \tag{2.2}$$

and the marginal

$$f_P(x) = \int f_P(y, x) dy, \tag{2.3}$$

as well as the joint density (2.1).

Since the conditional density (2.2) is parameterized by a mixing distribution $P \in \mathscr{P}$, for a Bayesian analysis of the Markov chain $\mathbf{X}$, we only need to choose an appropriate parametric family $\mathscr{K}$ and put a prior on the space of mixing distribution $\mathscr{P}$. The only requirement for $\mathscr{K}$ and $\mathscr{P}$ is that $f_P(\cdot|\cdot) \in \mathscr{F}$ for every $P \in \mathscr{P}$. As mentioned in [2], $f_P(\cdot|\cdot) \in \mathscr{F}$ if the two marginals of $f_P(\cdot, \cdot)$ are identical. Based on this idea, they proposed to use bivariate normal kernels with the same mean and variance parameters; see Section 2.1. In Section 2.2, we provide a novel example such that $f_P(\cdot, \cdot)$ have different marginals but $f_P(\cdot|\cdot) \in \mathscr{F}$. As a consequence, it is possible to construct a prior on $\mathscr{F}$ which has a sufficiently large KL support; see Section 4. Furthermore, we propose a general method for $f_P(\cdot, \cdot)$ to have the identical marginals via a copula. This general approach can be applied in practice for flexible modeling.

One natural choice of Π is a Dirichlet process [8], which is commonly adopted for density estimation, again as a mixing distribution; see [15]. A Markov chain Monte Carlo algorithm and some statistical applications, with our first example, are provided in [2], using the Dirichlet process prior. This can be extended to more general $\mathscr{K}$ without much difficulty.

Note that mixtures can be used for modelling transition densities directly in a Markov model. For example, [26] and [18] considered mixture densities of the form $f(y|x) = \int k_\theta(y|x) dP(\theta)$ and $f(y|x) = \int k_\theta(y) dP(\theta|x)$, respectively. Also, [9] considered $f(y|x) = g(y - \rho x)$ for a density g and AR coefficient $\rho \in (-1, 1)$, in which g can be modelled

as a mixture. These approaches are different to (2.2) which is derived from a bivariate mixture model. In particular, it is difficult to study how flexible these mixtures are, while our proposal (2.2) can have a large KL support with an appropriate choice of $\mathscr{K}$ and $\mathscr{P}$.

2.1. Symmetric normal mixtures

As illustrated in [2], the kernel distribution K_θ can be taken as a bivariate normal distribution with mean $(\mu, \mu)^T$ and variance

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{2.4}$$

for some $\theta = (\mu, \sigma^2, \rho)^T$. Since k_θ is symmetric in the sense that $k_\theta(y, x) = k_\theta(x, y)$ for every $x, y \in \mathbb{R}$, $f_P(\cdot, \cdot)$ is also symmetric, so we call it as a *symmetric normal mixture*. Let Θ be a subset of $\mathbb{R} \times (0, \infty) \times [0, 1]$. By the symmetry, we have that $\int f_P(y, x)dy = \int f_P(x, y)dy = f_P(x)$, so

$$\int f_P(y|x) f_P(x)dx = \int f_P(y, x)dx = f_P(y)$$

for every $P \in \mathscr{P}$. That means $f_P(\cdot)$ is the density of the stationary distribution of $f_P(\cdot|\cdot)$. Therefore, both the transition and stationary densities are explicitly expressed as a function of P. Since $f_P(y|x) > 0$ for every $y, x \in \mathbb{R}$, it corresponds to an irreducible and aperiodic chain, so we conclude that $f_P(\cdot|\cdot) \in \mathscr{F}$ for every $P \in \mathscr{P}$.

2.2. Non-symmetric normal mixtures

Although symmetric normal mixtures form a large class of transition density, it may not be sufficiently flexible. Note that probabilistic properties of a positive Harris chain are completely determined by the density $f(y, x) = f(y|x)f(x)$, where $f(y|x)$ is the transition density and $f(x) = \int f(y, x)dy$ is the stationary density. Note here that $f(y, x)$ need not be symmetric, as it was in Section 2.1. Therefore, a transition density derived from a symmetric normal mixture cannot capture some important data structures such as asymmetry and skewness, as reported in [33, 34].

For a more flexible stationary Markov model, general bivariate normal kernels with different mean (or variance) parameters can be considered. However, $\mathscr{K}$ or $\mathscr{P}$ should be chosen carefully because $f_P(\cdot|\cdot)$ may not belong to $\mathscr{F}$ without further constraints on $\mathscr{P}$ or $\mathscr{K}$. To see this, assume that $k_\theta(y, x)$ is the bivariate normal density with mean $\mu = (\mu_1, \mu_2)$ and variance $\Sigma = (\Sigma_{ij})$, where $\theta = (\mu, \Sigma)$. If $\mu_1 = \mu_2 = 0$ and $\rho\Sigma_{11}/\Sigma_{22} = 1$, then the transition density $k_\theta(\cdot|\cdot)$ corresponds to the standard random walk, which is non–stationary. One may impose constraints on $\mathscr{P}$ such as symmetry, *i.e.* $P((\mu_1, \mu_2) \in B) = P((\mu_2, \mu_1) \in B)$ for every Borel set B and $P(\Sigma_{11} = \Sigma_{22}) = 1$, so that $f_P(\cdot|\cdot) \in \mathscr{F}$. However, it still results in a symmetric joint distribution which is not sufficiently flexible.

We provide simple constraints on the form of k_θ assuring $f_P(\cdot|\cdot) \in \mathscr{F}$. Let $\theta = (\mu_1, \mu_2, \sigma)$ and k_θ be the bivariate normal density of the form

$$k_\theta(y, x) = \phi_\sigma(y - \mu_1)\phi_\sigma(x - \mu_2). \tag{2.5}$$

It should be noted that $f_P(\cdot, \cdot)$ does not have the same marginals, so $f_P(\cdot)$ is not the stationary distribution of $f_P(\cdot|\cdot)$. Although the form of the stationary distribution is not explicitly given, the following theorem still assures stationarity.

***Theorem* 2.1.** Let k_θ be defined as (2.5) and $\Theta = [-M, M]^2 \times [\sigma_1, \sigma_2]$ for some positive constants $\sigma_1 < \sigma_2$ and M. Then, $f_P(\cdot|\cdot) \in \mathscr{F}$ for every $P \in \mathscr{P}$.

We call (2.5) a *non-symmetric normal mixture*. In Section 4, it will be shown that a non-symmetric normal mixture prior has large KL support.

2.3. Copula-based kernels

As noted at the beginning of this section it is sufficient for $f_P(\cdot|\cdot) \in \mathscr{F}$ so that $f_P(\cdot, \cdot)$ has the same marginals. Obviously, it is sufficient that K_θ has the same marginals for every $\theta \in \Theta$. Such a kernel can be constructed via a copula. With a slight abuse of notation, we use K_θ to denote both univariate and bivariate kernels, and their cumulative distribution functions, where it is clear from the context.

Recall that a bivariate distribution function C on $[0, 1]^2$ is called a (two-dimensional) copula if its marginal distributions are uniform. Formally, a function $C : [0, 1]^2 \to [0, 1]$ is called a *copula* if

(1) $C(0, x) = C(x, 0) = 0$ and $C(1, x) = C(x, 1) = x$ for all $x \in [0, 1]$;
(2) for every $a, b, c, d \in [0, 1]$ with $a \le b$ and $c \le d$

$$C(a, c) - C(a, d) - C(b, c) + C(b, d) \ge 0.$$

If $(y, x) \mapsto f(y, x)$ is a bivariate density with the same marginals $x \mapsto f(x)$, then it can be written as

$$f(y, x) = f(y)f(x)c(F(y), F(x)) \tag{2.6}$$

for some copula density c by the Sklar's theorem [24], where F is the cumulative distribution function of f.

The copula approach to Markov models has been considered in the frequentist literature; for some important references, Darsow et al. [7] studied the mathematical relationship between Markov processes and copulas. Joe [13] considered a class of parametric copulas and parametric stationary distribution. A semiparametric approach based on parametric copula and nonparametric stationary distributions is studied in [5, 6], and in particular, Chen et al. [6] proved that a smooth functional of a sieve MLE is asymptotically normal and efficient.

We consider mixtures of kernels of the form (2.6) in a Bayesian framework. For a given parametric family $\mathscr{S} = \{K_\alpha : \alpha \in \mathscr{A}\}$ of univariate distributions and $\mathscr{C} = \{C_\beta :$

$\beta \in \mathscr{B}\}$ of copulas, the bivariate kernel family can be constructed as $\mathscr{K} = \{K_\theta : \theta \in \Theta\}$, where $\Theta = \mathscr{A} \times \mathscr{B}$, $K_\theta \in \mathscr{K}$ is the bivariate distribution with the density $k_\theta(y,x) = k_\alpha(y)k_\alpha(x)c_\beta(K_\alpha(y), K_\alpha(x))$, and $k_\alpha(\cdot)$ and c_β are densities of $K_\alpha \in \mathscr{S}$ and C_β, respectively. Let $\mathscr{P}$ be a class of product probability measure $P = P^{\mathscr{A}} \times P^{\mathscr{B}}$ on Θ. Since the mixing measure has a product form, the marginal density $f_P(\cdot)$ is given as $\int k_\theta(x) dP^{\mathscr{A}}(\alpha)$. This implies that we can model the stationary distribution and the dependence structure of the Markov chain separately.

A class of univariate normal distributions is a good candidate for $\mathscr{S}$. The bivariate normal family described in Section 2.1 is in fact a special case that $\mathscr{S}$ is a class of univariate normal distributions and $\mathscr{C}$ is a class of Gaussian copula. Mixtures of Student t-distributions are also popularly used for robust modelling, see [23]. Wu and co-authors [34, 33] have studied some Bayesian inferential methods for copula mixtures, and these can be extended to Markov models without much technical difficulty.

3. Posterior consistency with the Wasserstein metric

For the Markov model given in Section 2, we provide sufficient conditions for the posterior consistency in the Wasserstein metric. Note that for the Wasserstein consistency to hold, it must be implicitely assumed that $f_0 = f_{P_0}$ for some $P_0 \in \mathscr{P}$. Under a certain compactness assumption on Θ, the main theorem for the posterior consistency can be applied to various families of $\mathscr{K}$ and priors on $\mathscr{P}$. We first study the property of the Wasserstein metric in the nonparametric Markov model.

3.1. Wasserstein metric

As mentioned in the introduction, it is generally not easy to define a suitable metric d on a space of transition densities. Since we parameterize the transition density by $P \in \mathscr{P}$, any metric on $\mathscr{P}$ induces a semimetric on $\mathscr{F}_0 = \{f_P(\cdot|\cdot) : P \in \mathscr{P}\} \subset \mathscr{F}$. We use the Wasserstein metric d_W defined as (1.1). It is well-known that

$$d_W(P,Q) = \sup\left\{\left|\int h\, dP - \int h\, dQ\right| : h \in \mathscr{L}(\Theta)\right\}, \tag{3.1}$$

by the Kantorovich-Rubinstein theorem [14], where $\mathscr{L}(\Theta)$ is the set of all functions $h : \Theta \to \mathbb{R}$ such that $|h(\theta_1) - h(\theta_2)| \le \|\theta_1 - \theta_2\|$ for every $\theta_1, \theta_2 \in \Theta$. Note that the topology generated by d_W is stronger than the weak topology on $\mathscr{P}$, and coincides if Θ is bounded, [10].

The Wasserstein metric between mixing measures P_1 and P_2 of two mixture densities $f_1 = \int k_\theta dP_1(\theta)$ and $f_2 = \int k_\theta dP_2(\theta)$ is typically stronger than the Hellinger and total variation metric between f_1 and f_2. The reverse does not generally hold. Thus, it is regarded as a strong metric in density estimation problems. For details about this and more recent results, see [22]. The forthcoming Theorem 3.1 asserts that d_W is also strong enough as a distance in the Markov model considered in Section 2. For example, under

a certain condition it dominates $d_{\mathcal{C}}(f_{P_1}, f_{P_2}) = \sup_{x\in\mathcal{C}} d_V\big(f_{P_1}(\cdot|x), f_{P_2}(\cdot|x)\big)$ for every compact set $\mathcal{C}$. Therefore, posterior consistency with respect to d_W implies consistency with respect to $d_{\mathcal{C}}$ for any compact $\mathcal{C}$. For a simple illustration, consider two bivariate normal distributions $N((\mu_1,\mu_1)^T, \Sigma_1)$ and $N((\mu_2,\mu_2)^T, \Sigma_2)$, where the Σ_j's are defined in (2.4) with $\sigma^2 = 1$ and $\rho = \rho_j$ for $j = 1, 2$. This is the case, in our terminology, that $\mathscr{K}$ is a bivariate normal family and P_j's are Dirac measures. Since each conditional distribution is $N(\mu_j + \rho_j(x-\mu_j), 1-\rho_j^2)$, if $\rho_1 \neq \rho_2$ we have $d_{\max}(f_{P_1}, f_{P_2}) = 2$. We assume the following condition for the kernel family.

(**M**) For every compact subset $\mathcal{C}$ of $\mathbb{R}$, there exist a continuous function $g : \mathbb{R} \to [0,\infty)$ and positive constants γ_1 and γ_2 such that $\int g(y)dy < \infty$ and

$$\gamma_1 \le k_\theta(x) \le \gamma_2 \tag{3.2}$$
$$|k_\theta(y,x_1) - k_\theta(y,x_2)| \le g(y)\|x_1 - x_2\| \tag{3.3}$$
$$|k_{\theta_1}(y,x) - k_{\theta_2}(y,x)| \le g(y)\|\theta_1 - \theta_2\| \tag{3.4}$$

for every $x, x_1, x_2 \in \mathcal{C}, y \in \mathbb{R}$ and $\theta, \theta_1, \theta_2 \in \Theta$.

Note that γ_1, γ_2 and g are allowed to depend on a compact set $\mathcal{C}$. Condition (**M**) is required for technical reasons, and typically holds when Θ is bounded and the scale parameter of k_θ is bounded away from zero.

***Theorem* 3.1.** For any probability measures P_1 and P_2,

$$\int d_V\big(f_{P_1}(\cdot|x), f_{P_2}(\cdot|x)\big) f_{P_2}(x)dx \le 2d_V(f_{P_1}(\cdot,\cdot), f_{P_2}(\cdot,\cdot)). \tag{3.5}$$

Furthermore, if (**M**) holds, then for every compact $\mathcal{C} \subset \mathbb{R}$ there exists a constant $D > 0$ such that

$$\sup_{x\in\mathcal{C}} d_V\big(f_{P_1}(\cdot|x), f_{P_2}(\cdot|x)\big) \le D d_W(P_1, P_2).$$

3.2. Posterior consistency with the Wasserstein metric

To investigate asymptotic properties, we assume that there exists the *true* transition density $f_0 \in \mathscr{F}$ generating the observation $X_0, X_1, \ldots, X_n$. Also, we assume that $f_0(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$ for some $P_0 \in \mathscr{P}$. Denote the density of stationary distribution of $f_0(\cdot|\cdot)$ as $f_0(\cdot)$. The joint density $f_0(\cdot,\cdot)$ refers to $f_0(y,x) = f_0(y|x)f_0(x)$, which is consistent in our notation. Recall that $f_0(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$, but it is not necessarily required that $f_0(\cdot) = f_{P_0}(\cdot)$ and $f_0(\cdot,\cdot) = f_{P_0}(\cdot,\cdot)$.

Let Π^n be the posterior distribution given $X_0, X_1, \ldots, X_n$. Then for any measurable subset A of $\mathscr{P}$, the posterior probability is given by $\Pi^n(A) = L_{nA}/I_n$, where $L_{nA} = \int_A R_n(P)d\Pi(P)$, $I_n = \int_{\mathscr{P}} R_n(P)d\Pi(P)$ for $n \ge 0$ ($L_{0A} = \Pi(A)$ and $I_0 = 1$), and

$$R_n(P) = \prod_{i=1}^n \frac{f_P}{f_0}(X_i|X_{i-1}).$$

For $f_1, f_2 \in \mathscr{F}_0$, let $h_n(f_1, f_2) = \frac{1}{2}\, d_H^2\big(f_1(\cdot|X_n), f_2(\cdot|X_n)\big)$. Also let

$$f_{nA}(y|x) = \int f_P(y|x) d\Pi_A^n(P),$$

where Π_A^n is the posterior distribution restricted and renormalized to the set A. The posterior distribution is said to be *consistent* at f_0 with respect to a (psuedo-)metric d if

$$\Pi^n\Big(\{P \in \mathscr{P} : d(f_P, f_0) > \epsilon\}\Big) \to 0 \quad \mathbb{P}\text{-almost surely},$$

for every $\epsilon > 0$, where $\mathbb{P}$ is the underlying probability measure generating $\mathbf{X}$.

The posterior probability of $B_\epsilon^c = \{P \in \mathscr{P} : d_W(P, P_0) > \epsilon\}$ is given by $\Pi^n(B_\epsilon^c) = L_{nB_\epsilon^c}/I_n$. Typically, $\Pi^n(B_\epsilon^c)$ can be shown to converge almost surely to zero by proving, roughly speaking, that $\mathbb{P}$-almost-surely $L_{nB_\epsilon^c} < \exp(-n\delta)$ for some $\delta > 0$ and $I_n \geq \exp(-n\epsilon)$ for every $\epsilon > 0$. The lower bound of the denominator I_n can be obtained by the so-called KL support condition as in the i.i.d. cases; see [9, 1] and the forthcoming condition (**K**). In the literature on Markov models, proofs for the upper bound of the numerator $L_{nB_\epsilon^c}$ rely on the martingale approach proposed by [30, 31] for i.i.d. models. We use similar techniques to [9, 1], where B_ϵ^c is partitioned into a finite number of sets $B_j, 1 \leq j \leq N$, satisfying (5.5). Then, for each j, it is typically needed to bound $\sup_{x \in C} d_V(f_{P_1}(\cdot|x), f_{P_2}(\cdot|x))$ for some set C with $\int_C f_0(x)dx > 0$. We construct partitions via Wasserstein balls (with different radii) so this uniform bound can be achieved by Theorem 3.1. The following two conditions are required for the posterior consistency.

(**I**) For $P \in \mathscr{P}$, $f_P(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$ implies $P = P_0$.
(**K**) For every $\epsilon > 0$

$$\Pi\left(\left\{P \in \mathscr{P} : \int K(f_0(\cdot|x), f_P(\cdot|x)) f_0(x)dx < \epsilon\right\}\right) > 0.$$

If two marginals of $f_P(\cdot,\cdot)$ are identical for every $P \in \mathscr{P}$, that is,

$$\int f_P(y, x)dx = \int f_P(x, y)dx \quad \text{for all } y \in \mathbb{R},$$

then $f_P(\cdot)$ is the stationary distribution of the transition density $f_P(\cdot|\cdot)$, so $f_P(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$ implies that $f_P(\cdot,\cdot) = f_{P_0}(\cdot,\cdot)$. Therefore in this case, condition (**I**) is equivalent to the following condition:

(**I′**) For $P \in \mathscr{P}$, $f_P(\cdot,\cdot) = f_{P_0}(\cdot,\cdot)$ implies $P = P_0$.

Note that condition (**I′**) can be viewed as an identifiability condition for bivariate mixture models. In general, a family of bivariate normal mixtures does not satisfy (**I′**) because a convolution of two normal distributions is again a normal distribution. If we put a restriction on Θ or $\mathscr{P}$, normal mixtures can be shown to be identifiable. For example,

the class of all finite mixtures of normal distributions can be shown to be identifiable. For the identifiability of continuous normal mixtures, some conditions on Θ, such as compactness, are required. There is a vast amount of literature about the identifiability of mixtures including both finite and infinite mixtures; see [27, 29, 25, 28] for general conditions and [4] for infinite normal mixtures. Without the identifiability condition (**I**), the posterior distribution cannot be consistent in the Wasserstein metric, and so a weaker metric or topology should be considered.

***Theorem* 3.2** (Posterior consistency)**.** Assume that (**M**), (**I**), (**K**) hold, Θ is bounded and $\mathscr{P}$ is compact in d_W. If $f_0 = f_{P_0} \in \mathscr{F}$ for some $P_0 \in \mathscr{P}$, then

$$\Pi^n\Big(\big\{P \in \mathscr{P} : d_W(P, P_0) > \epsilon\big\}\Big) \to 0$$

$\mathbb{P}$-almost-surely for every $\epsilon > 0$.

Note that the compactness of $\mathscr{P}$ in d_W is mild because the set of every Borel probability measure on a compact set is compact in d_W; see [19, 10]. We apply Theorem 3.2 to three examples considered in Section 2. A Dirichlet process mixture prior is one of the most important priors, so we consider it in the examples. Let $\mathrm{DP}(\alpha, G_0)$ be the Dirichlet process with mean measure G_0 and precision parameter $\alpha > 0$; see for example [12]. Assume that the support of G_0 is Θ, so every Borel probability measure on Θ is contained in the weak support of $\mathrm{DP}(\alpha, G_0)$, [17].

3.2.1. Symmetric normal mixtures

Let

$$\Theta = \Big\{(\mu, \sigma^2, \rho) : \mu \in [-M, M], \sigma_1^2 \le \sigma^2 \le \sigma_2^2, |\rho| \le 1 - \delta\Big\}$$

for some positive constants $M, \delta, \sigma_1, \sigma_2$. For $\theta = (\mu, \sigma^2, \rho)$, let $k_\theta(y, x)$ be the bivariate normal densities with mean vector $(\mu, \mu)^T$ and variance Σ, where Σ is defined in (2.4). Note that

$$k_\theta(y, x) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(\frac{z}{2(1-\rho^2)}\right),$$

where

$$z = \frac{1}{\sigma^2}\Big\{(y-\mu)^2 - 2\rho(y-\mu)(x-\mu) + (x-\mu)^2\Big\}.$$

Condition (**I′**), which is equivalent to (**I**), is well studied in [4] under the compactness of Θ. What remains to prove is (**M**) and (**K**).

***Corollary* 3.1.** Let $\mathscr{K}$ be the class of symmetric normal distributions described above. If Π is the $\mathrm{DP}(\alpha, G_0)$ prior on $\mathscr{P}$, where G_0 has full support on Θ, the posterior distribution is consistent in d_W at every $P_0 \in \mathscr{P}$.

3.2.2. Non–symmetric normal mixtures

Let

$$\Theta = \left\{(\mu_1, \mu_2, \sigma) : \mu_1, \mu_2 \in [-M, M], \sigma_1^2 \leq \sigma^2 \leq \sigma_2^2\right\}$$

for some positive constants $\sigma_1 < \sigma_2$ and M. For $\theta = (\mu_1, \mu_2, \sigma)$, let K_θ be the bivariate normal distribution with mean (μ_1, μ_2) and variance $\sigma^2 I$. Conditions (**M**) and (**K**) can be proved in the same way as the symmetric normal mixture.

Although the identifiability condition (**I**′) holds, (**I**) is not assured for every $P_0 \in \mathscr{P}$. For example, if P_0 is the Dirac measure at $(\mu_{01}, \mu_{02}, \sigma_0) \in \Theta$, then $f_P(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$ for every product measure of the form $P = \delta_{\mu_{01},\sigma_0} \times Q$, where $\delta_{\mu_{01},\sigma_{01}}$ is the Dirac measure at (μ_{01}, σ_0) and Q is a probability measure on $[-M, M]$. Thus, we consider a slightly weaker topology than the one induced by d_W. Let $\mathscr{P}_0 \subset \mathscr{P}$ be the collection of every P such that $f_P(\cdot|\cdot) = f_{P_0}(\cdot|\cdot)$. If a sequence (P_n) in $\mathscr{P}_0$ converges to P_∞ in d_W, then $\lim_n f_{P_n}(y|x) = f_{P_\infty}(y|x)$ for every y and x, so, $\mathscr{P}_0$ is d_W-closed. Therefore, we can define the distance between P and $\mathscr{P}_0$ as $d_0(P, \mathscr{P}_0) = \inf_{Q \in \mathscr{P}_0} d_W(P, Q)$. A neighborhood of $f_{P_0}(\cdot|\cdot)$ can be defined using the pseudo-metric d_0 as $A_\epsilon = \{f_P \in \mathscr{F} : P \in B_\epsilon\}$, where $B_\epsilon = \{P \in \mathscr{P} : d_0(P, \mathscr{P}_0) < \epsilon\}$. The result of Theorem 3.1 still holds in the sense that, for every compact set $\mathcal{C} \subset \mathbb{R}$, there exists a constant $C > 0$ such that $\sup_{x \in \mathcal{C}} d_V(f_P(\cdot|x), f_{P_0}(\cdot|x)) \leq C\epsilon$ for every $\epsilon > 0$ and $P \in B_\epsilon$. Thus, A_ϵ can be regarded as a strong and suitable neighborhood of $f_{P_0}(\cdot|\cdot)$.

***Corollary* 3.2.** Let $\mathscr{K}$ be the class of non-symmetric normal distributions described above. If Π is the $\mathrm{DP}(\alpha, G_0)$ prior on $\mathscr{P}$, where G_0 has full support on Θ, $\Pi^n(B_\epsilon) \to 1$ $\mathbb{P}$-almost-surely for every $\epsilon > 0$.

3.2.3. Copula–based kernels

As particular examples of copula based kernels, we consider semiparametric models that can effectively model the dependence structure of the Markov chain, as studied in [5, 6]. We parameterize the class of stationary distributions and of copulas separately, and denote them as $\mathscr{S} = \{K_\alpha : \alpha \in \mathscr{A}\}$ and $\mathscr{C} = \{C_\beta : \beta \in \mathscr{B}\}$, respectively, where $\mathscr{A}$ and $\mathscr{B}$ are subsets of Euclidean spaces. Let $\theta = (\alpha, \beta)$, $\Theta = \mathscr{A} \times \mathscr{B}$, and $k_\theta(y, x) = k_\alpha(y)k_\alpha(x)c_\beta(y, x)$. Let $\mathscr{P}$ be the set of every product probability probability measure on Θ of the form $P \times \delta_\beta$, where P is a probability measure on $\mathscr{A}$ and δ_β is the Dirac measure at $\beta \in \mathscr{B}$. This is a semiparametric model considered in [6].

Condition (**I**) is equivalent to (**I**′) and can be handled with general approaches introduced in Section 3. In particular, once $\mathscr{C}$ is identifiable, *i.e.*, $C_{\beta_1} = C_{\beta_2}$ implies $\beta_1 = \beta_2$, it sufficies to check that $\int k_\alpha(\cdot)dP_1(\alpha) = \int k_\alpha(\cdot)dP_2(\alpha)$ implies $P_1 = P_2$. Condition (**K**) can be proved in the same way to symmetric normal mixtures. More specifically, if

$$\left\{\frac{f_{P_0}(y, x)}{f_P(y, x)}\right\}^\delta \tag{3.6}$$

is bounded by an integrable function for some $\delta > 0$ and every $P \in \mathscr{P}$, (3.4) implies condition (**K**); see the proof of Corollary 3.1.

The technical condition (**M**) can be satisfied under mild integrability and smoothness conditions on k_α and c_β. For this, let $k'_\alpha(x)$, $\dot{k}_\alpha(x)$, $\dot{K}_\alpha(x)$, $c'_\beta(u,v)$ and $\dot{c}_\beta(u,v)$ be partial derivatives of maps $x \mapsto k_\alpha(x)$, $\alpha \mapsto k_\alpha(x)$, $\alpha \mapsto K_\alpha(x)$, $v \mapsto c_\beta(u,v)$ and $\beta \mapsto c_\beta(u,v)$, respectively. For simplicity, assume that $c_\beta(u,v) = c_\beta(v,u)$ for every $u, v \in [0,1]$ and β, then $\partial c_\beta(u,v)/\partial u = c'_\beta(v,u)$. We also assume that the marginal density k_α satisfies (3.2) for every compact $\mathcal{C} \subset \mathbb{R}$. If $k'_\alpha(x)$, $\dot{k}_\alpha(x)$ and $\dot{K}_\alpha(x)$ are bounded uniformly in α and $x \in \mathbb{R}$, and

$$\begin{aligned} \int \sup_{x\in\mathcal{C}} \sup_{\theta} k_\alpha(y) c_\beta(K_\alpha(y), K_\alpha(x)) dy < \infty \\ \int \sup_{x\in\mathcal{C}} \sup_{\theta} k_\alpha(y) \|c'_\beta(K_\alpha(y), K_\alpha(x))\| dy < \infty \end{aligned} \tag{3.7}$$

for every compact $\mathcal{C} \subset \mathbb{R}$, then (3.3) holds. Also, (3.4) holds provided that

$$\begin{aligned} &\int \sup_{x\in\mathcal{C}} \sup_{\theta} \|\dot{k}_\alpha(y)\| c_\beta(K_\alpha(y), K_\alpha(x)) dy < \infty \\ &\int \sup_{x\in\mathcal{C}} \sup_{\theta} k_\alpha(y) \|c'_\beta(K_\alpha(x), K_\alpha(y))\| dy < \infty \\ &\int \sup_{x\in\mathcal{C}} \sup_{\theta} k_\alpha(y) \|\dot{c}_\beta(K_\alpha(x), K_\alpha(y))\| dy < \infty. \end{aligned} \tag{3.8}$$

As a concrete example, we consider the location family of Student t-distributions and the Farlie-Gumbel-Morgenstern copula, see [21]. That is, k_α is the density of the Student t-distribution with the median $\alpha \in [-M, M]$, scale parameter σ, and $d \geq 1$ degrees of freedom, and the copula is given as

$$C_\beta(u,v) = uv + \beta uv(1-u)(1-v), \quad |\beta| \leq \beta_{\max}$$

for some constant $\beta_{\max} < 1$, and $\Theta = \{(\alpha, \beta) \in \mathbb{R}^2 : |\alpha| < M, |\beta| \leq \beta_{\max}\}$. Note that $c_\beta(u,v) = 1 + \beta(2u-1)(2v-1)$, and integrability conditions (3.7) and (3.8) are easily satisfied. Since $k'_\alpha(x)$, $\dot{k}_\alpha(x)$ and $\dot{K}_\alpha(x)$ are uniformly bounded, condition (**M**) is satisfied as described above. Condition (**K**) can be proved by (3.6), which holds for small enough $\delta > 0$ because the tail of $x \mapsto k_\alpha(x)$ has a polynomial order and c_β is bounded away from zero. The class of location mixtures of Student t-distributions is identifiable, see [29], so the monotonicity of $\beta \mapsto C_\beta(u,v)$ implies the identifiability (**I′**).

4. Posterior consistency with the average L_1 metric

In this section, we consider the posterior consistency for a general true transition density f_0. As mentioned earlier, the Wasserstein distance cannot be used as a metric for the posterior consistency because f_0 may not be presented as a mixture, *i.e.* no P_0 exists. As done in [26], we consider the average metric d_{avg} defined as

$$d_{\text{avg}}(f_1, f_2) = \int d_V(f_1(\cdot|x), f_2(\cdot|x)) f_0(x) dx.$$

For the posterior consistency to hold at f_0, it is standard that the prior puts sufficient mass around f_0 in the sense of the KL support condition (**K**). Note that

$$\int K\big(f_0(\cdot|x), f_P(\cdot|x)\big) f_0(x)dx = K\big(f_0(\cdot,\cdot), f_P(\cdot,\cdot)\big) - K\big(f_0(\cdot), f_P(\cdot)\big),$$

so condition (**K**) can be implied by the following condition:

(**K′**) for every $\epsilon > 0$

$$\Pi\Big(\big\{P \in \mathscr{P} : K(f_0(\cdot,\cdot), f_P(\cdot,\cdot)) < \epsilon\big\}\Big) > 0.$$

Condition (**K′**) is the KL support condition for bivariate densities required for i.i.d. models. In particular with the non-symmetric normal mixtures, condition (**K′**) is very mild and satisfied for most bivariate density f_0 provided that Π has full weak support; see for example Theorem 2 of [32]. Recall that most important priors on $\mathscr{P}$, including the Dirichlet process, has full weak support, [3].

Thus, we focus on the non-symmetric normal mixtures. For condition (**K′**) to hold for a large class of f_0, small values of σ should be considered. It should be noted that by considering small σ, the support of the prior contains much more transition densities than that considered in Section 3.2.2. For technical convenience and notational simplicity, we restrict the parameter set to $\Theta = [-M, M]^2 \times (0, M]$, where $M > 0$ is an arbitrary constant. In this case, (**K′**) is easily satisfied provided that $f_0(\cdot,\cdot)$ is supported on $[-M, M]^2$. We provide a DP prior yielding consistent posterior at f_0 satisfying (**K′**). Note that our construction is based on a strong tail assumption on the prior. We leave more delicate construction of priors, incorporating nearly optimal convergence rate and $M \to \infty$, as our future work.

Consider Gaussian mixtures of the form

$$f_{P,\sigma}(y, x) = \int \phi_\sigma(y - z_1)\phi_\sigma(x - z_2)dP(z),$$

where $z = (z_1, z_2)$. Let $\mathscr{P}$ be the set of probability measures on $[-M, M]^2$. For a given probability measure G_0 whose support is $[-M, M]^2$, let Π_1 be $\mathrm{DP}(\alpha, G_0)$. Let Π_2 be a prior on $(0, M]$ satisfying $\Pi_2((0, \eta_n]) \le e^{-cn}$ for some constant $c > 0$ and sequence $\eta_n \downarrow 0$. We assume that $e^{a\eta_n^{-2}} = o(n)$ as $n \to \infty$ for any constant $a > 0$, which holds if $G_0([x, M])$ increases sufficiently slowly as $x \to 0$. Let $\Pi = \Pi_1 \times \Pi_2$ be the product prior on $\mathscr{P} \times (0, M]$. We further assume the following:

(**U**) The true joint density $f_0(\cdot,\cdot)$ is supported on $[-M, M]^2$ and there exists a probability measure φ on $\mathbb{R}$, $\lambda > 0$, and an integer $k \ge 1$ such that

$$\mathbb{P}_x(X_k \in B) \ge \lambda\varphi(B)$$

for every initial x and $B \in \mathcal{B}(\mathbb{R})$.

Condition (**U**) is closely related to the assumptions of uniform ergodicity and Doeblin recurrence; see [20]. Technically, it is required for Hoeffding's inequality to hold for Markov chains [11].

***Theorem* 4.1.** With the prior described above, assume that conditions (**K**′) (f_P replaced by $f_{P,\sigma}$) and (**U**) hold. Then, the posterior is consistent in d_{avg}, *i.e.* for every $\epsilon > 0$

$$\Pi^n\Big(\big\{(P,\sigma) \in \mathscr{P} \times (0,M] : d_{\text{avg}}(f_{P,\sigma}, f_0) > \epsilon\big\}\Big) \to 0$$

$\mathbb{P}$-almost-surely.

5. Proofs

5.1. Proofs for Section 2

5.1.1. Proof of Theorem 2.1.

Assume that $\mathbf{X}$ is a Markov chain with transition $f_P(\cdot|\cdot)$ and let $A \in \mathcal{B}(\mathbb{R})$ be a given set with positive Lebesgue measure. Since $f_P(y|x) > 0$ for every $y, x \in \mathbb{R}$, $\mathbf{X}$ is aperiodic in the sense of [20]. Since $f_P(y|x) = \int_\Theta k_\theta(y,x)dP(\theta) / \int_\Theta k_\theta(x)dP(\theta)$, we have that

$$\inf_{\theta\in\Theta} \phi_\sigma(y-\mu_1) = \inf_{\theta\in\Theta} k_\theta(y|x) \le f_P(y|x) \le \sup_{\theta\in\Theta} k_\theta(y|x) = \sup_{\theta\in\Theta} \phi_\sigma(y-\mu_1) \tag{5.1}$$

for every $y, x \in \mathbb{R}$. If the Lebesgue measure of A^c is equal to zero, it is obvious that $\mathbb{P}_x(X_n \in A \text{ i.o.}) = 1$. Otherwise,

$$\mathbb{P}_x(X_n \in A \text{ i.o.}) = \lim_{m\to\infty} \mathbb{P}_x(\cup_{n=m}^\infty \{X_n \in A\}) = 1 - \lim_{m\to\infty} \mathbb{P}_x(\cap_{n=m}^\infty \{X_n \in A^c\})$$

and, by (5.1),

$$\mathbb{P}_x(\cap_{n=m}^\infty \{X_n \in A^c\}) = \lim_{N\to\infty} \mathbb{P}_x(\cap_{n=m}^N \{X_n \in A^c\}) \le \limsup_{N\to\infty} (1-\alpha)^{N-m+1} = 0,$$

where $\alpha = \int_A \inf_{\theta\in\Theta} \phi_\sigma(y-\mu_1)dy > 0$. Therefore, $\mathbf{X}$ is Harris recurrent. Also, by (5.1), $\lim_{C\to\infty} \sup_{n\ge1} \mathbb{P}_x(|X_n| > C) = 0$ for every $x \in \mathbb{R}$, which implies that $\mathbf{X}$ is bounded in probability on average in the sense of [20]. Thus, there exists at least one invariant measure of $f_P(\cdot|\cdot)$ by Theorem 12.0.1 of [20]. By the Ergodic theorem (Theorem 13.0.1 in [20]), it suffices to show that $\sup_{x\in\mathbb{R}} \mathbb{E}_x(\tau_A) < \infty$, where $\mathbb{E}_x$ denotes the expectation under $\mathbb{P}_x$. If the Lebesgue measure of A^c is equal to zero, this is trivial because $P_x(\tau_A = 1) = 1$. Otherwise, by (5.1),

$$\mathbb{P}_x(\tau_A = t) = \mathbb{P}_x(X_t \in A | X_s \in A^c, s < t) \le (1-\alpha)^{t-1},$$

so we have that

$$\sup_{x\in\mathbb{R}} \mathbb{E}_x(\tau_A) = \sum_{t=1}^\infty t(1-\alpha)^{t-1} < \infty.$$

This completes the proof. □

5.2. Proofs for Section 3

5.2.1. Proof of Theorem 3.1

Note that

$$f_{P_1}(y|x) - f_{P_2}(y|x) = \frac{f_{P_1}(y,x)}{f_{P_1}(x)} - \frac{f_{P_2}(y,x)}{f_{P_2}(x)}$$

for every $x, y \in \mathbb{R}$. Since

$$\frac{b}{a} - \frac{d}{c} = \frac{b(c-a)}{ac} + \frac{b-d}{c} \tag{5.2}$$

for every positive numbers a, b, c and d, $|f_{P_1}(y|x) - f_{P_2}(y|x)|$ is bounded by

$$\frac{f_{P_1}(y,x)}{f_{P_1}(x)f_{P_2}(x)}\big|f_{P_1}(x) - f_{P_2}(x)\big| + \frac{1}{f_{P_2}(x)}\big|f_{P_1}(y,x) - f_{P_2}(y,x)\big|. \tag{5.3}$$

Thus, the left hand side of (3.5) is bounded by

$$\begin{aligned}&\iint \Big[f_{P_1}(y|x)\big|f_{P_1}(x) - f_{P_2}(x)\big| + \big|f_{P_1}(y,x) - f_{P_2}(y,x)\big|\Big]dydx\\ &\le d_V(f_{P_1}(\cdot), f_{P_2}(\cdot)), + d_V(f_{P_1}(\cdot,\cdot), f_{P_2}(\cdot,\cdot)) \le 2d_V(f_{P_1}(\cdot,\cdot), f_{P_2}(\cdot,\cdot)),\end{aligned}$$

which completes the proof of (3.5).

Note that $d_V(f_{P_1}(\cdot|x), f_{P_2}(\cdot|x)) = 2\sup_{B\in\mathcal{B}(\mathbb{R})}|f_{P_1}(B|x) - f_{P_2}(B|x)|$, where $f_P(B|x) = \int_B f_P(y|x)dy$. Let a compact set $\mathcal{C} \subset \mathbb{R}$ be given. Then, for every $x \in \mathcal{C}$ and $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}f_{P_1}(B|x) - f_{P_2}(B|x) &= \int_B \left[\frac{\int k_\theta(y,x)dP_1(\theta)}{f_{P_1}(x)} - \frac{\int k_\theta(y,x)dP_2(\theta)}{f_{P_2}(x)}\right] dy\\ &\le C_1 \int_B \left\{ f_{P_1}(y,x)\left|\int k_\theta(x)dP_2(\theta) - \int k_\theta(x)dP_1(\theta)\right| \right.\\ &\quad \left. + \left|\int k_\theta(y,x)dP_1(\theta) - \int k_\theta(y,x)dP_2(\theta)\right|\right\} dy\end{aligned} \tag{5.4}$$

by (3.2) and (5.2), where $C_1 > 0$ is a constant depending only on γ_1. By (3.4), two terms

$$\left|\int k_\theta(x)dP_2(\theta) - \int k_\theta(x)dP_1(\theta)\right|$$

and

$$\left|\int k_\theta(y,x)dP_1(\theta) - \int k_\theta(y,x)dP_2(\theta)\right|$$

are bounded by a constant multiple of $d_W(P_1, P_2)$, and $g(y)d_W(P_1, P_2)$, respectively. Since $\int_B f_{P_1}(y,x)dy \le f_{P_1}(x) \le \gamma_2$ by (3.2), and $\int g(y)dy < \infty$, (5.4) is bounded by a constant multiple of $d_W(P_1, P_2)$. □

5.2.2. Proof of Theorem 3.2

We first provide three lemmas. In particular, the proof of Lemma 5.1 is motivated from the martingale approach [30, 31] for i.i.d. models. Similar techniques can be found in [2, 9].

***Lemma* 5.1.** For any measurable $A \subset \mathscr{P}$, if

$$\liminf_m \frac{1}{m} \sum_{n=0}^{m-1} h_{n-1}\big(f_{n-1A}, f_0\big) > \epsilon \quad \mathbb{P}\text{-almost-surely} \tag{5.5}$$

for some constant $\epsilon > 0$, then there exists $\delta > 0$, depending only on ϵ, such that $\limsup_n e^{n\delta} L_{nA} \leq 1$ $\mathbb{P}$-almost-surely.

Proof. Note that for a sequence of real random variables (Y_n) and real sequence (b_n) with $b_n \uparrow \infty$, if $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n)/b_n^2 < \infty$ then

$$\frac{1}{b_n} \sum_{k=1}^{n} \Big(Y_k - \mathbb{E}(Y_k | Y_1, \ldots, Y_{k-1}) \Big) \to 0 \tag{5.6}$$

almost surely; see [16].

Since

$$f_{nA}(X_{n+1}|X_n) = \int f_P(X_{n+1}|X_n) d\Pi_A^n(P) = \frac{\int_A f_P(X_{n+1}|X_n) R_n(P) d\Pi(P)}{\int_A R_n(P) d\Pi(P)},$$

we have

$$\frac{L_{n+1A}}{L_{nA}} = \frac{f_{nA}}{f_0}(X_{n+1}|X_n) \tag{5.7}$$

for every measurable $A \subset \mathscr{P}$. Thus,

$$\begin{aligned} \mathbb{E}\Big(\sqrt{L_{n+1A}} | \mathcal{F}_n \Big) &= \sqrt{L_{nA}} \Big(1 - \tfrac{1}{2} d_H^2\big(f_{nA}(\cdot|X_n), f_0(\cdot|X_n)\big) \Big) \\ &= \sqrt{L_{nA}} \Big(1 - h_n\big(f_{nA}, f_0\big) \Big), \end{aligned}$$

where $\mathbb{E}$ is the expectation under the true distribution and $\mathcal{F}_n = \sigma(X_i : i \leq n)$ is the σ-algebra generated by $X_0, \ldots, X_n$. Therefore, the sequence $(M_m, \mathcal{F}_m)$ forms a martingale, where

$$M_m = \sum_{n=1}^{m} \Big\{ \sqrt{L_{nA}/L_{n-1A}} - 1 + h_{n-1}\big(f_{n-1A}, f_0\big) \Big\}.$$

Since $\mathbb{E}(L_{n+1A}/L_{nA}) = 1$ by (5.7) and the Hellinger distance is bounded by $\sqrt{2}$, the variance of each summand in the last display is bounded by a constant. Thus, $M_m/m \to 0$ $\mathbb{P}$-almost-surely by (5.6). Condition (5.5) implies that

$$\limsup_{m \to \infty} \frac{1}{m} \sum_{n=1}^{m} \sqrt{\frac{L_{nA}}{L_{n-1A}}} < 1 - \epsilon$$

$\mathbb{P}$-almost-surely. Since

$$\sqrt{L_{mA}} = \prod_{n=1}^{m} \sqrt{\frac{L_{nA}}{L_{n-1A}}} \leq \left(\frac{1}{m} \sum_{n=1}^{m} \sqrt{\frac{L_{nA}}{L_{n-1A}}} \right)^m,$$

we have

$$\limsup_{m\to\infty} (L_{mA})^{1/2m} \leq 1 - \epsilon$$

$\mathbb{P}$-almost-surely. Thus, for every small enough $\gamma > 0$, with $\delta = -2\log(1 - \epsilon + \gamma) > 0$, we have $\limsup_m e^{m\delta} L_{mA} \leq 1$ $\mathbb{P}$-almost-surely. □

***Lemma* 5.2.** If (**M**) holds and Θ is bounded, then for every fixed $z \in \mathbb{R}$ and $Q \in \mathscr{P}$, the real-valued function $(x, P) \mapsto d_V\big(f_P(\cdot|x), f_Q(\cdot|z)\big)$ defined on $\mathbb{R} \times \mathscr{P}$ is continuous, where $\mathbb{R} \times \mathscr{P}$ is equipped with the product metric.

Proof. By (3.3) and (3.4), for a given compact set $\mathcal{C} \subset \mathbb{R}$, there exists a constant $C > 0$ such that

$$|k_\theta(x_1) - k_\theta(x_2)| \leq C\|x_1 - x_2\| \quad \text{and} \quad |k_{\theta_1}(x) - k_{\theta_2}(x)| \leq C\|\theta_1 - \theta_2\| \tag{5.8}$$

for all $x, x_1, x_2 \in \mathcal{C}$ and $\theta, \theta_1, \theta_2 \in \Theta$. We first claim that maps $(x, P) \mapsto f_P(x)$ and $(x, P) \mapsto f_P(y, x)$, for every $y \in \mathbb{R}$, from $\mathcal{C} \times \mathscr{P}$ to $\mathbb{R}$ are continuous with respect to the product topology.

Let (x_n, P_n) be a sequence in $\mathcal{C} \times \mathscr{P}$ converging to (x_∞, P_∞) and $y \in \mathbb{R}$ be given. Then,

$$\begin{aligned}
&\left| \int k_\theta(x_n) dP_n(\theta) - \int k_\theta(x_\infty) dP_\infty(\theta) \right| \\
&\leq \left| \int k_\theta(x_n) dP_n(\theta) - \int k_\theta(x_n) dP_\infty(\theta) \right| + \left| \int k_\theta(x_n) dP_\infty(\theta) - \int k_\theta(x_\infty) dP_\infty(\theta) \right| \\
&\leq C_1\big(d_W(P_n, P_\infty) + \|x_n - x_\infty\|\big)
\end{aligned}$$

for some constant $C_1 > 0$, where the last inequality holds by (5.8). Also,

$$\begin{aligned}
&\left| \int k_\theta(y, x_n) dP_n(\theta) - \int k_\theta(y, x_\infty) dP_\infty(\theta) \right| \\
&\leq \left| \int k_\theta(y, x_n) dP_n(\theta) - \int k_\theta(y, x_\infty) dP_n(\theta) \right| + \left| \int k_\theta(y, x_\infty) dP_n(\theta) - \int k_\theta(y, x_\infty) dP_\infty(\theta) \right| \\
&\leq g(y)\big(\|x_n - x_\infty\| + d_W(P_n, P_\infty)\big),
\end{aligned}$$

where g is the function in condition (**M**). This proves the claim.

Since for some fixed x_0 and θ_0

$$k_\theta(y, x) \leq k_{\theta_0}(y, x_0) + g(y)\big(\|x - x_0\| + \|\theta - \theta_0\|\big)$$

by (3.3) and (3.4), there exists a constant $M > 0$ such that $f_P(y,x) \le k_{\theta_0}(y,x_0) + Mg(y)$ for every $x \in \mathcal{C}$, $y \in \mathbb{R}$ and $P \in \mathscr{P}$ due to the boundedness of $\mathcal{C}$ and Θ. Now, by (3.2) and the dominated convergence theorem

$$d_V\big(f_{P_n}(\cdot|x_n), f_Q(\cdot|z)\big) = \int \left| \frac{f_{P_n}(y,x_n)}{f_{P_n}(x_n)} - \frac{f_Q(y,z)}{f_Q(z)} \right| dy,$$

converges to $d_V\big(f_{P_\infty}(\cdot|x_\infty), f_Q(\cdot|z)\big)$. Since $\mathcal{C}$ can be arbitrarily large, this completes the proof. □

***Lemma* 5.3.** If (**I**) holds and $f_0 = f_{P_0} \in \mathscr{F}$, then for every $P_1 \in \mathscr{P} - \{P_0\}$,

$$\liminf_m \frac{1}{m} \sum_{n=1}^{m} h_n\big(f_{P_1}, f_{P_0}\big) > 0 \tag{5.9}$$

$\mathbb{P}$-almost-surely.

Proof. By the strong law of large number for positive Harris chain, for every $P \in \mathscr{P}$,

$$\frac{1}{m} \sum_{n=1}^{m} h_n(f_P, f_{P_0}) \to \tfrac{1}{2} \int d_H^2\big(f_P(\cdot|x), f_{P_0}(\cdot|x)\big) f_0(x) dx$$

$\mathbb{P}$-almost-surely. Therefore, if the left hand side of (5.9) is equal to zero with positive $\mathbb{P}$-probability, then $d_H\big(f_{P_1}(\cdot|x), f_{P_0}(\cdot|x)\big) = 0$ for f_0-almost-surely. Since $f_{P_0}(y|x) > 0$ for every $y, x \in \mathbb{R}$, f_0 has full support on $\mathbb{R}$. By the continuity of $f_{P_j}(\cdot|\cdot)$, $j = 0, 1$, this implies that $f_{P_1}(y|x) = f_{P_0}(y|x)$ for every $x, y \in \mathbb{R}$. By the identifiability condition (**I**), this contradicts that $P_1 \neq P_0$. □

***Lemma* 5.4.** Assume that (**I**), (**M**) hold, Θ is compact and $f_0 = f_{P_0} \in \mathscr{F}$. Then, for every $P_1 \in \mathscr{P} - \{P_0\}$, there exists a $\delta > 0$ such that (5.5) hold with $A = \{P \in \mathscr{P} : d_W(P, P_1) < \delta\}$.

Proof. By Lemma 5.3, there exists $x_1 \in \mathbb{R}$ such that $\epsilon \equiv d_V\big(f_{P_1}(\cdot|x_1), f_{P_0}(\cdot|x_1)\big) > 0$. Also, by Lemma 5.2 and the equivalence of d_H and d_V, there exists a $\delta > 0$ such that $\max\{\|x - x_1\|, d_W(P, P_1)\} < \delta$ implies that

$$\max\Big\{ d_H\big(f_P(\cdot|x), f_{P_1}(\cdot|x_1)\big), d_H\big(f_{P_0}(\cdot|x), f_{P_0}(\cdot|x_1)\big) \Big\} < \epsilon/5.$$

If $\max\{\|x - x_1\|, d_W(P, P_1)\} < \delta$, then we have

$$\begin{aligned} d_H\big(f_{P_0}(\cdot|x), f_P(\cdot|x)\big) &\ge d_H\big(f_{P_1}(\cdot|x_1), f_{P_0}(\cdot|x_1)\big) \\ &\quad - d_H\big(f_{P_0}(\cdot|x_1), f_{P_0}(\cdot|x)\big) - d_H\big(f_P(\cdot|x), f_{P_1}(\cdot|x_1)\big) \\ &\ge \epsilon/2 - \epsilon/5 - \epsilon/5 = \epsilon/10, \end{aligned}$$

so

$$\begin{aligned} d_H\big(f_{nA}(\cdot|x), f_{P_0}(\cdot|x)\big) &\geq d_H\big(f_{P_1}(\cdot|x), f_{P_0}(\cdot|x)\big) - d_H\big(f_{P_1}(\cdot|x), f_{nA}(\cdot|x)\big) \\ &\geq \epsilon/10 - d_H\big(f_{P_1}(\cdot|x), f_{nA}(\cdot|x)\big), \end{aligned}$$

where $A = \{P \in \mathscr{P} : d_W(P, P_1) < \delta\}$. By Theorem 3.1 and the convexity of Hellinger balls, we can choose $\delta > 0$ sufficiently small, so that $d_H\big(f_{P_1}(\cdot|x), f_{nA}(\cdot|x)\big) < \epsilon/20$ for $\|x - x_1\| < \delta$. Finally, by ergodicity, the cardinality of the set $\{n \leq m : \|X_n - x_1\| < \delta\}$ divided by m converges almost surely to a positive constant γ, and we conclude that the left hand side of (5.5) is greater than or equal to $\gamma\epsilon^2/800$. □

Proof of Theorem 3.2. Condition (**K**) implies that for every $c > 0$

$$\limsup_{n\to\infty} e^{nc} I_n = \infty$$

$\mathbb{P}$-almost-surely; by Lemma 3.3 of [1].

Let $A = \{P \in \mathscr{P} : d_W(P, P_0) \geq \epsilon\}$. Then by Lemma 5.4, for every $P \in A$, we can choose a $\delta_P > 0$ such that $\liminf m^{-1} \sum_{n=1}^{m} h_{n-1}(f_{n-1A_P}, f_{P_0}) > 0$, where $A_P = \{\widetilde{P} \in \mathscr{P} : d_W(\widetilde{P}, P) < \delta_P\}$. Since A is compact in d_W, we can choose a finite collection $\{P_1, \ldots, P_M\}$ such that $A \subset \cup_{m=1}^{M} A_{P_m}$. Since $\Pi^n(A_{P_m}) \to 0$ $\mathbb{P}$-almost-surely for every m by Lemma 5.1, the proof is complete. □

5.2.3. Proof of Corollary 3.1

Since $|\mu|$ is bounded above and σ^2 is bounded away from zero and infinity, (3.2) is easily satisfied. To prove (3.3) and (3.4), it suffices to check that partial derivatives of $(\theta, x) \mapsto k_\theta(y, x)$, viewed as a map from $\Theta \times \mathcal{C}$, are bounded by an integrable function, where $\mathcal{C}$ is a compact set. Note that partial derivatives of $(\theta, x) \mapsto k_\theta(y, x)$ is equal to $k_\theta(y, x)$ times partial derivatives of $(\theta, x) \mapsto \log k_\theta(y, x)$. Let $\ell(y, x, \theta) = \log k_\theta(y, x)$, then its partial derivatives are given as

$$\begin{aligned} \frac{\partial}{\partial x}\ell(y, x, \theta) &= \frac{(x - \mu) - \rho(y - \mu)}{\sigma^2(1 - \rho^2)} \\ \frac{\partial}{\partial \mu}\ell(y, x, \theta) &= \frac{(1 - \rho)\{(y - \mu) + (x - \mu)\}}{\sigma^2(1 - \rho^2)} \\ \frac{\partial}{\partial \sigma^2}\ell(y, x, \theta) &= \frac{1}{\sigma^2}\left(\frac{z}{2(1 - \rho^2)} - 1\right) \\ \frac{\partial}{\partial \rho}\ell(y, x, \theta) &= \frac{\rho}{1 - \rho^2} + \frac{(y - \mu)(x - \mu)}{\sigma^2(1 - \rho^2)} \\ &\quad - \frac{\rho\{(y - \mu)^2 - 2\rho(y - \mu)(x - \mu) + (x - \mu)^2\}}{\sigma^2(1 - \rho^2)^2}. \end{aligned}$$

The map $\tilde{g}$ defined by

$$\tilde{g}(y) = \sup_{x\in\mathcal{C}} \sup_{\theta\in\Theta} (y - \mu)^2 k_\theta(y, x)$$

is of order $O(\exp\{-y^2/(2\sigma_2^2)\})$ as $|y| \to \infty$, so it is Lebesgue integrable. Thus if we let $g(y) = C(1 + \tilde{g}(y))$ for a sufficiently large constant $C > 0$, then both (3.3) and (3.4) are satisfied.

To prove (**K**), note that

$$\inf_{P\in\mathscr{P}} f_P(y,x) \geq \inf_{\theta\in\Theta} k_\theta(y,x) \gtrsim \exp\{-C_1(y^2+x^2)\}$$
$$\sup_{P\in\mathscr{P}} f_P(y,x) \leq \sup_{\theta\in\Theta} k_\theta(y,x) \lesssim \exp\{-C_2(y^2+x^2)\}$$

as $\|(y,x)\| \to \infty$ for some constants $C_1, C_2 > 0$. Thus, there exists a small enough $\delta > 0$ and a function h such that

$$\int\int h(y,x) f_{P_0}(y,x) dy dx < \infty$$

and

$$\left\{\frac{f_{P_0}(y,x)}{f_P(y,x)}\right\}^\delta < h(y,x)$$

for every $P \in \mathscr{P}$. Since

$$\int K\big(f_{P_0}(\cdot|x), f_P(\cdot|x)\big) f_0(x) dx = K(f_{P_0}(\cdot,\cdot), f_P(\cdot,\cdot)) - K(f_{P_0}(\cdot), f_P(\cdot))$$
$$\leq K(f_{P_0}(\cdot,\cdot), f_P(\cdot,\cdot)) \leq \frac{1}{\delta}\log\int\int\left\{\frac{f_{P_0}(y,x)}{f_P(y,x)}\right\}^\delta f_{P_0}(y,x) dy dx$$

and $f_P(y,x) \to f_{P_0}(y,x)$ as $d_W(P,P_0) \to 0$, where the convergence of $f_P(y,x)$ holds by (3.1) and (3.4), it holds by the dominated convergence theorem that

$$\int K\big(f_{P_0}(\cdot|x), f_P(\cdot|x)\big) f_0(x) dx \to 0$$

as $d_W(P,P_0) \to 0$. Thus (**K**) holds because Π has the full weak support. □

5.2.4. Proof of Corollary 3.2

It is easy to check that the result of Lemmas 5.3 and 5.4 holds for every $P_1 \in \mathscr{P} - \mathscr{P}_0$. Since B_ϵ is open, $\mathscr{P} - B_\epsilon$ is compact, so we can follow the same line of the proof of Theorem 3.2 by replacing the set A with $\mathscr{P} - B_\epsilon$. □

5.3. Proofs for Section 4

***Lemma* 5.5.** For every $\epsilon > 0$, there exists a constant $C > 0$ (depending only on M) and an integer N (depending only on M and ϵ) such that

$$\sup_{|x|\leq M} d_V(f_{P_1,\sigma_1}(\cdot|x), f_{P_2,\sigma_2}(\cdot|x)) \leq e^{C\eta_n^{-2}}\Big(d_W(P_1,P_2) + |\sigma_1 - \sigma_2|\Big) + \epsilon$$

for every $P_j \in \mathscr{P}$, $\sigma_j \geq \eta_n$, for $j = 1, 2$ and $n \geq N$.

Proof. Let ϵ be given. Assume that $P_1, P_2 \in \mathscr{P}$, $\sigma_1, \sigma_2 > \eta_n$ and $|x| \le M$. Note that there exists a constant $D > 0$, depending only on M and ϵ, such that

$$\int_{\{|y|>D\}} f_{P,\sigma}(y|x)dy \le \int_{\{|y|>D\}} \sup_{\sigma\le M} \sup_{|z|\le M} \phi_\sigma(y-z)dy < \frac{\epsilon}{2}$$

for every $P \in \mathscr{P}$ and $\sigma \le M$. Therefore, by (5.3), $|f_{P_1,\sigma_1}(B|x) - f_{P_2,\sigma_2}(B|x)|$ is bounded by

$$\frac{1}{f_{P_2,\sigma_2}(x)}\left[\left|f_{P_1,\sigma_1}(x) - f_{P_2,\sigma_2}(x)\right| + \int_{\{|y|\le D\}} \left|f_{P_1,\sigma_1}(y,x) - f_{P_2,\sigma_2}(y,x)\right|dy\right] + \epsilon \quad (5.10)$$

for every Borel set B, where $f_{P,\sigma}(B|x) = \int_B f_{P,\sigma}(y|x)dy$. Note that

$$\begin{aligned}\left|\frac{\partial}{\partial x}\phi_\sigma(x)\right| &= \frac{|x|}{\sigma^2}\phi_\sigma(x) \lesssim \frac{|x|}{\sigma^3}\\ \left|\frac{\partial}{\partial \sigma}\phi_\sigma(x)\right| &\le \left(\frac{1}{\sigma} + \frac{x^2}{\sigma^3}\right)\phi_\sigma(x) \lesssim \frac{1+x^2}{\sigma^4}.\end{aligned}$$

Thus, we have

$$\max\left\{|f_{P_1,\sigma_1}(x) - f_{P_2,\sigma_1}(x)|, |f_{P_2,\sigma_1}(x) - f_{P_2,\sigma_2}(x)|\right\} \lesssim \frac{d_W(P_1,P_2) + |\sigma_1 - \sigma_2|}{\eta_n^4}$$

$$\max\left\{|f_{P_1,\sigma_1}(y,x) - f_{P_2,\sigma_1}(y,x)|, |f_{P_2,\sigma_1}(y,x) - f_{P_2,\sigma_2}(y,x)|\right\} \lesssim \frac{d_W(P_1,P_2) + |\sigma_1 - \sigma_2|}{\eta_n^5}$$

by the Kantorovich-Rubinstein representation (3.1). Since

$$f_{P_2,\sigma_2}(x) \gtrsim \inf_{|z|\le M} \exp\left(-\frac{(x-z)^2}{2\sigma_2^2}\right) \ge \exp\left(-\frac{2M^2}{\eta_n^2}\right),$$

the proof is complete by (5.10). □

5.3.1. Proof of Theorem 4.1

Let $\epsilon > 0$ be given. Note that (**K**) implies that $\liminf_n e^{n\delta} I_n = \infty$ for every $\delta > 0$. It follows that $\Pi(\sigma < \eta_n | X_1, \ldots, X_n) \to 0$ $\mathbb{P}$-almost-surely. Thus, it suffices to prove that $\limsup_n e^{cn} L_{nA} = 0$ $\mathbb{P}$-almost-surely for some constant $c > 0$, where

$$A = \{(P,\sigma) \in \mathscr{P} \times [\eta_n, M] : d_{\mathrm{avg}}(f_P, f_0) > \epsilon\}.$$

Here, the dependence of A on n is abbreviated for notational convenience.

Note that $N(\gamma, (0,M], |\cdot|) \lesssim \gamma^{-1}$ and

$$\log N(\gamma, \mathscr{P}, d_W) \lesssim \gamma^{-2} \log \gamma^{-1},$$

where the second inequality holds by Lemma 4 of [22]. Note that implicit constants in the notation $\lesssim$ depend only on M. Let $\delta_n = e^{-C\eta_n^{-2}}\epsilon_n$, where $C > 0$ is a constant in Lemma 5.5 and $\epsilon_n \downarrow 0$ is an arbitrary sequence. By Lemma 5.5, if ϵ_n decreases sufficiently slowly, for every large enough n, we can pick $(P_{n,j}, \sigma_{n,j}) \in \mathscr{P} \times [\eta_n, M]$, for $j = 1, \ldots, N_n$, such that $\mathscr{P} = \cup_{j=1}^{N_n} A_{n,j}$, $\log N_n = o(n)$ and

$$\sup_{|x|\le M} d_V(f_{P_1,\sigma_1}(\cdot|x), f_{P_2,\sigma_2}(\cdot|x)) \le \epsilon^2/32 \tag{5.11}$$

for any pairs (P_1, σ_1) and (P_2, σ_2) in the same partition, where

$$A_{n,j} \subset \{(P,\sigma) \in \mathscr{P} \times [\eta_n, M] : d_W(P, P_{n,j}) + |\sigma - \sigma_{n,j}| < \delta_n\}.$$

Let $f_{n,j} = f_{P_{n,j},\sigma_{n,j}}$ and

$$d'_{\text{avg}}(f_1, f_2) = \int d_H^2(f_1(\cdot|x), f_2(\cdot|x)) f_0(x) dx.$$

Then

$$d'_{\text{avg}}(f_1, f_2) \ge \frac{1}{4}\int d_V^2(f_1(\cdot|x), f_2(\cdot|x)) f_0(x) dx \ge \frac{1}{4}\{d_{\text{avg}}(f_1, f_2)\}^2,$$

so $d'_{\text{avg}}(f_{P,\sigma}, f_0) > \epsilon^2/4$ for every $(P,\sigma) \in A$. By the Hoeffding's inequality for Markov chain [11], there exist constants $d, n_0 > 0$ depending only on k, λ (see condition (**U**)) and ϵ such that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{m=1}^{n} h_m(f_{P,\sigma}, f_0) - d'_{\text{avg}}(f_{P,\sigma}, f_0)\right| > \frac{\epsilon^2}{8}\right) \le e^{-dn}$$

for every $(P,\sigma) \in A$ and $n \ge n_0$. Thus,

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le j\le N_n}\left|\frac{1}{n}\sum_{m=1}^{n} h_m(f_{n,j}, f_0) - d'_{\text{avg}}(f_{n,j}, f_0)\right| > \frac{\epsilon^2}{8}\right) \\
&\le \sum_{j=1}^{N_n}\mathbb{P}\left(\left|\frac{1}{n}\sum_{m=1}^{n} h_m(f_{n,j}, f_0) - d'_{\text{avg}}(f_{n,j}, f_0)\right| > \frac{\epsilon^2}{8}\right) \\
&\le e^{-dn/2}
\end{aligned} \tag{5.12}$$

for large enough n.

Let

$$\Omega_n = \left\{\min_{1\le j\le N_n}\frac{1}{n}\sum_{m=1}^{n} h_m(f_{mA_{n,j}}, f_0) > \frac{\epsilon^2}{32}\right\}.$$

Note that

$$h_m(f_0, f_{n,j}) \le 2\Big(h_m(f_{mA_{n,j}}, f_0) + h_m(f_{n,j}, f_{mA_{n,j}})\Big)$$

and

$$h_m^2(f_{n,j}, f_{mA_{n,j}}) = 2\left(1 - \int \sqrt{f_{n,j}(y|X_m) f_{mA_{n,j}}(y|X_m)}dy\right)$$
$$\leq \int 2\left(1 - \int \sqrt{f_{n,j}(y|X_m) f_{P,\sigma}(y|X_m)}dy\right) d\Pi^m_{A_{n,j}}(P)$$
$$\leq \sup_{(P,\sigma)\in A_{n,j}} h_m^2(f_{n,j}, f_{P,\sigma}),$$

where the first inequality holds by Cauchy-Schwartz. Thus,

$$\frac{1}{n}\sum_{m=1}^{n} h_m(f_0, f_{n,j}) \leq \frac{\epsilon^2}{8}$$

for some j on Ω_n^c. It follows that

$$d'_{\text{avg}}(f_0, f_{n,j}) - \frac{1}{n}\sum_{m=1}^{n} h_m(f_0, f_{n,j}) \geq \frac{\epsilon^2}{8}$$

for some j on Ω_n^c, so $\mathbb{P}(\Omega_n^c) \leq e^{-dn/2}$ by (5.12).

Since h_m is bounded by 2, there exists a constant $\alpha \in (0,1)$, depending only on ϵ, such that $K_{n,j}/n > \alpha$ for every $1 \leq j \leq N_n$ on Ω_n, where $K_{n,j}$ is the cardinality of

$$\Big\{m \leq n : h_m(f_{mA_{n,j}}, f_0) > \epsilon^2/64\Big\}.$$

As in the proof of Lemma 5.1, we have

$$\mathbb{E}\Big(\sqrt{L_{m+1A_{n,j}}}|\mathcal{F}_m\Big) = \sqrt{L_{mA_{n,j}}}\Big(1 - h_m(f_{mA_{n,j}}, f_0)\Big)$$

for every $m, n \geq 1$. Thus, on Ω_n

$$\mathbb{E}\Big(\sqrt{L_{nA_{n,j}}}|\mathcal{F}_1\Big) \leq \left(1 - \frac{\epsilon^2}{64}\right)^{\lfloor \alpha n\rfloor - 1} \sqrt{L_{1A_{n,j}}}$$

for every j, where $\lfloor a\rfloor$ is the largest integer less than or equal to a. It follows that

$$\mathbb{E}\Big(\sqrt{L_{nA}}|\mathcal{F}_1\Big) \leq \sum_{j=1}^{N_n} \mathbb{E}\Big(\sqrt{L_{nA_{n,j}}}|\mathcal{F}_1\Big) \leq \left(1 - \frac{\epsilon^2}{64}\right)^{\lfloor \alpha n\rfloor - 1} \sum_{j=1}^{N_n} \sqrt{L_{1A_{n,j}}}$$

on Ω_n. Thus,

$$e^{n\beta}\mathbb{E}(\sqrt{L_{nA}}) = e^{n\beta}\mathbb{E}\left[1_{\Omega_n}\mathbb{E}\Big(\sqrt{L_{nA}}|\mathcal{F}_1\Big) + 1_{\Omega_n^c}\mathbb{E}\Big(\sqrt{L_{nA}}|\mathcal{F}_1\Big)\right]$$
$$\leq e^{n\beta}\left(1 - \frac{\epsilon^2}{64}\right)^{\lfloor \alpha n\rfloor - 1} \sum_{j=1}^{N_n} \mathbb{E}\sqrt{L_{1,A_{n,j}}} + e^{n\beta}\sqrt{\mathbb{P}(\Omega_n^c)\mathbb{E}\left\{\mathbb{E}\Big(\sqrt{L_{nA}}|\mathcal{F}_1\Big)\right\}^2}$$
$$\leq e^{n\beta}N_n\left(1 - \frac{\epsilon^2}{64}\right)^{\lfloor \alpha n\rfloor - 1} + e^{n\beta}\sqrt{\mathbb{P}(\Omega_n^c)}$$

for every $\beta > 0$, where the last inequality holds because

$$\mathbb{E}\sqrt{L_{nB}} \leq \sqrt{\mathbb{E}L_{nB}} \leq 1$$

for every B. Therefore, $\limsup_n e^{n\beta} L_{nA} = 0$ for sufficiently small enough $\beta > 0$ by the Borel-Cantelli lemma. □

Acknowledgements

The authors are grateful to the comments and suggestions from an Associate Editor and two referees on an earlier version of the paper.

References

[1] Antoniano-Villalobos, I. and Walker, S. G. (2015). Bayesian Consistency for Markov Models. *Sankhya A* **77** 106–125.
[2] Antoniano-Villalobos, I. and Walker, S. G. (2016). A nonparametric model for stationary time series. *J. Time Series Anal.* **37** 126–142.
[3] Bissiri, P. G. and Ongaro, A. (2014). On the topological support of species sampling priors. *Electron. J. Stat.* **8** 861–882.
[4] Bruni, C. and Koch, G. (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.* **13** 1341–1357.
[5] Chen, X. and Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *J. Econometrics* **130** 307–335.
[6] Chen, X., Wu, W. B. and Yi, Y. (2009). Efficient estimation of copula-based semiparametric Markov models. *Ann. Statist.* **37** 4214–4253.
[7] Darsow, W. F., Nguyen, B., Olsen, E. T. et al. (1992). Copulas and Markov processes. *Illinois J. Math.* **36** 600–642.
[8] Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.
[9] Ghosal, S. and Tang, Y. (2006). Bayesian consistency for Markov processes. *Sankhya* **68** 227–239.
[10] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.
[11] Glynn, P. W. and Ormoneit, D. (2002). Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics & Probability Letters* **56** 143–146.
[12] Hjort, N. L., Holmes, C., Müller, P. and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.
[13] Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.
[14] Kantorovich, L. V. and Rubinstein, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ.* **13** 52–59.

[15] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12** 351–357.
[16] Loève, M. (1963). *Probability Theory*, 3rd ed. Van Nostrand, Princeton, NJ.
[17] Majumdar, S. (1992). On topological support of Dirichlet prior. *Statist. Probab. Lett.* **15** 385–388.
[18] Mena, R. H. and Walker, S. G. (2005). Stationary autoregressive models via a Bayesian nonparametric approach. *Journal of Time Series Analysis* **26** 789–805.
[19] Merkle, M. (2000). Topics in weak convergence of probability measures. *Zb. Rad.(Beogr.)* **9** 235–274.
[20] Meyn, S. P. and Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability*. Springer Science & Business Media.
[21] Nelsen, R. B. (2003). Properties and applications of copulas: A brief survey. In *Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance* (J. Dhaene, N. Kolev and P. Morettin, eds.) 10–28. University Press USP: Sao Paulo.
[22] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400.
[23] Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statist. Comput.* **10** 339–348.
[24] Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris* **8** 229–231.
[25] Tallis, G. and Chesson, P. (1982). Identifiability of mixtures. *J. Aust. Math. Soc.* **32** 339–348.
[26] Tang, Y. and Ghosal, S. (2007). Posterior consistency of Dirichlet mixtures for estimating a transition density. *J. Statist. Plann. Inference* **137** 1711–1726.
[27] Teicher, H. (1960). On the mixture of distributions. *Ann. Math. Statist.* **31** 55–73.
[28] Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* **34** 1265–1269.
[29] Teicher, H. et al. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.
[30] Walker, S. (2003). On sufficient conditions for Bayesian consistency. *Biometrika* **90** 482–488.
[31] Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043.
[32] Wu, Y. and Ghosal, S. (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *J. Multivariate Anal.* **101** 2411–2419.
[33] Wu, J., Wang, X. and Walker, S. G. (2014). Bayesian Nonparametric Inference for a Multivariate Copula Function. *Methodol. Comput. Appl. Probab.* **16** 747–763.
[34] Wu, J., Wang, X. and Walker, S. G. (2015). Bayesian nonparametric estimation of a copula. *J. Stat. Comput. Simul.* **85** 103–116.