Posterior asymptotics in Wasserstein metrics on the real line[*]

Minwoo Chae

Department of Industrial and Management Engineering,
Pohang University of Science and Technology, South Korea
e-mail: mchae@postech.ac.kr

Pierpaolo De Blasi

University of Torino and Collegio Carlo Alberto, Italy
e-mail: pierpaolo.deblasi@unito.it
and

Stephen G. Walker

Department of Mathematics,
University of Texas at Austin, USA
e-mail: s.g.walker@math.utexas.edu

Abstract: In this paper, we use the class of Wasserstein metrics to study asymptotic properties of posterior distributions. Our first goal is to provide sufficient conditions for posterior consistency. In addition to the well-known Schwartz's Kullback–Leibler condition on the prior, the true distribution and most probability measures in the support of the prior are required to possess moments up to an order which is determined by the order of the Wasserstein metric. We further investigate convergence rates of the posterior distributions for which we need stronger moment conditions. The required tail conditions are sharp in the sense that the posterior distribution may be inconsistent or contract slowly to the true distribution without these conditions. Our study involves techniques that build on recent advances on Wasserstein convergence of empirical measures. We apply the results to some examples including a Dirichlet process mixture prior and conduct a simulation study for further illustration.

Contents

1. Introduction

The Wasserstein distance originally arose in the problem of optimal transportation [43] and is often called the Kantorovich or transportation distance. We refer to [42] for the history about this metric. For two Borel probability measures P and Q on the real line, the Wasserstein metric of order p, $p \in [1, \infty)$, is defined as

$$W_p(P, Q) = \inf_{\pi \in \mathscr{C}(P,Q)} \left(\int_{\mathbb{R}^2} |x - y|^p d\pi(x, y) \right)^{1/p},$$

where $\mathscr{C}(P, Q)$ is the set of every coupling π of P and Q, that is, a Borel probability measure on $\mathbb{R}^2$ with marginals P and Q, respectively.

There are a wide number of applications of Wasserstein metrics, *e.g.* Wasserstein generative adversarial networks (GAN; [1, 25]), approximate Markov chain Monte Carlo (MCMC; [38]), distributionally robust optimization (DRO; [31]) and clustering ([4, 32]). However, exhaustive study on statistical properties such as the convergence behavior of the empirical measure with respect to W_p have been conducted only recently, see [7, 18, 46, 15]. In particular, the great success of Wasserstein GAN in machine learning society accelerated the study of Wasserstein metrics in statistics community as a discrepancy measure between probabilities; [41, 34, 5]. Recently, [3] proposed the use of the Wasserstein distance in the implementation of Approximate Bayesian Computation (ABC) to approximate the posterior distribution. In nonparametric Bayesian inference, [36, 12] used Wasserstein metrics to study asymptotic properties of posterior distributions, but W_p was considered as a distance between mixing distributions rather than a distance between mixture densities themselves. As a result, the Wasserstein metrics in these papers yielded a stronger topology than the total variation distance on the space of density functions. In general, W_p, $1 \le p < \infty$, metrizes the weak convergence of probability measures in a bounded metric space. Specifically, if the diameter of the underlying metric space is bounded by 1, one has the relationship $d_P^2 \le W_1 \le 2d_P \le d_V$, where d_P and d_V are Lévy-Prokhorov and total variation distances, see [23]. In an unbounded metric space, the second and third inequalities do not hold because W_1 is not a bounded metric.

In this article, we utilize the Wasserstein distances to study asymptotic behavior of posterior distributions under the assumption that data are generated

from a fixed true distribution and we focus on nonparametric Bayesian density estimation on the real line. To set the stage, let $X_1, \ldots, X_n$ be the observations which are independent and identically distributed random variables from the true distribution P_0 possessing a density p_0. Let $\mathcal{F}$ be a collection of probability densities in $\mathbb{R}$ equipped with the weak topology, and Π be a prior distribution on $\mathcal{F}$. Then the posterior probability of a measurable set $A \subset \mathcal{F}$ is given as

$$\Pi(p \in A \mid X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n p(X_i)/p_0(X_i) d\Pi(p)}{\int \prod_{i=1}^n p(X_i)/p_0(X_i) d\Pi(p)} \qquad (1.1)$$

by the Bayes formula. Throughout the paper, we allow the prior Π to depend on the sample size n, but often abbreviate this dependency in the notation of both prior and posterior distributions. If clarification is necessary, the prior and posterior will be denoted Π_n and $\Pi_n(\cdot \mid X_1, \ldots, X_n)$, respectively. The posterior distribution is said to be consistent with respect to a (pseudo-)metric d if

$$\Pi\Big(d(p, p_0) > \epsilon \mid X_1, \ldots, X_n\Big) \to 0 \text{ in probability for every } \epsilon > 0,$$

where the convergence in probability is taken with respect to the true distribution P_0. If ϵ is replaced by ϵ_n for some sequence $\epsilon_n \to 0$, the convergence rate of the posterior distributions is said to be at least ϵ_n. There is a huge amount of research articles concerning asymptotic properties of the posterior distribution. We refer to the monograph [22] for the history and details about this topic.

Of key importance is the Kullback–Leibler (KL) support condition developed by [39]. A fixed prior Π is said to satisfy the KL support condition if

$$\Pi\Big(p : K(p_0, p) < \epsilon\Big) > 0 \quad \text{for every } \epsilon > 0, \qquad (1.2)$$

where $K(p_0, p) = \int \log[p_0(x)/p(x)] dP_0(x)$ is the KL divergence. If the prior depends on the sample size, the KL condition (1.2) can be replaced by

$$\liminf_{n \to \infty} \Pi_n\Big(p : K(p_0, p) < \epsilon\Big) > 0 \quad \text{for every } \epsilon > 0. \qquad (1.3)$$

Conditions (1.2) and (1.3) became standard for proving posterior consistency. In particular, it gives a suitable lower bound of the denominator in (1.1) and it implies posterior consistency in the weak topology, that is with respect to the Lévy-Prokhorov distance, see [39] and Section 6.4 of [22]. A variation of the KL support condition to obtain a convergence rate is developed by [20]. It is formally expressed as

$$\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2} \quad \text{for all large enough } n, \qquad (1.4)$$

where

$$\mathcal{K}_n = \left\{ p \in \mathcal{F} : \int \log \frac{p_0}{p} dP_0 \leq \epsilon_n^2, \int \left(\log \frac{p_0}{p} \right)^2 dP_0 \leq \epsilon_n^2 \right\}.$$

In literature, studies on posterior asymptotics have focused on strong metrics such as the total variation, Hellinger and uniform metrics. For those purposes, some non-trivial conditions such as the bounded entropy or prior summability are assumed in addition to the KL conditions, see [19, 45, 2, 11] for example.

On the other hand, it is surprising that careful analysis of the convergence rates with respect to a weak metric such as the Lévy-Prokhorov and Kolmogorov has not been studied in literature, considering that the KL support condition is sufficient for the consistency in those metrics. [11] studied the convergence rate of the posterior distribution with respect to the Lévy-Prokhorov metric, but their rate $n^{-1/4}$ have a lot of room for improvement. Furthermore, they used the Lévy-Prokhorov rate as a tool for proving the consistency in total variation, and did not focus on the convergence rate itself.

Wasserstein metrics W_p, $1 \leq p < \infty$ metrize weak convergence in a bounded space, but it generates a stronger topology in general. Indeed, neither the KL support condition (1.2) nor (1.4) are sufficient for posterior consistency with respect to W_p. If P_0 is a standard Cauchy density, for example, $W_p(P, P_0) = \infty$ for any P and $p \geq 1$. Therefore, for any prior except the one putting all its mass on P_0, the posterior distribution is inconsistent with respect to W_p. This simple example shows that tails or moments of probability measures play an important role for handling W_p.

For a sequence P_n of probability measures, it is well-known that $W_p(P_n, P) \to 0$ if and only if P_n converges to P weakly and $M_p(P_n) \to M_p(P)$, see [43], p.212, where $M_p(P) = \int |x|^p dP(x)$. Therefore, for the Wasserstein consistency to hold, the posterior moment should converge to the true moment, see Theorem 2.1. However, while the moment consistency of frequentist's nonparametric estimators such as the the empirical distribution is straightforward, it is non-trivial to show that the posterior moment converges to the true moment even with a very popular prior such as a Dirichlet process mixture. This is mainly because tails of probability measures in the support of the prior should be considered simultaneously.

To prove posterior consistency, we will leverage on the KL condition. We provide two different approaches which are of independent interest; see the proof of Theorem 2.2. The first one targets directly posterior moment consistency and relies on a result from [45]. The second one has less stringent conditions but the proof is more complicated. Specifically, we construct uniformly consistent tests based on the empirical distribution by exploiting suitable upper bounds of Wasserstein metrics. We then show that, to achieve posterior consistency with respect to W_p, moments of densities must be suitably bounded. In particular, the posterior needs to put most of its mass on distributions that possess moments up to an order determined by that of the Wasserstein metric. In practice, the posterior moment condition can be worked out by means of exponentially small prior probability on the complement set, cf. Lemma 8.2. In Section 5.2 we provide an illustration in the specific example of Dirichlet process mixture prior.

Both approaches for posterior consistency can be extended to obtain suitable convergence rates with the KL condition (1.4). While the first approach gives the convergence rate for the moment, the second approach gives the rate with respect to W_p^p relying on slightly stronger moment conditions, see Theorems 3.1 and 3.2. For convergence rates with the second approach, we rely on new upper bounds on Wasserstein metrics that can be of independent interest, cf.

Lemma 8.7. Interestingly, the posterior moment conditions for consistency and convergence rates are nearly necessary, that is the posterior distribution may be inconsistent or contract slowly to the true distribution when they are not satisfied. Finally, we obtain convergence rates for the case $p = \infty$ in Theorem 4.1, for which we need to restrict to probability measures on a bounded space.

To the best of our knowledge, this paper is the first result on posterior asymptotics with the Wasserstein metric. The remainder of this paper is organized as follows. Results on posterior consistency and its convergence rate with respect to W_p, for $1 \leq p < \infty$, are considered in Sections 2 and 3, respectively. Posterior asymptotics with respect to W_∞ is studied in Section 4. Section 5 considers more details with specific examples. Some numerical results complementing our theory are provided in Section 6. Concluding remarks and proofs are given in Sections 7 and 8, respectively.

Notation

Before proceeding, we introduce some further notation; for two real numbers a and b, their minimum and maximum are denoted by $a \wedge b$ and $a \vee b$, respectively. Inequality $a \lesssim b$ means that a is less than a constant multiple of b, where the constant is universal unless specified. Upper cases such as P and Q refer to probability measures corresponding to the densities denoted by lower cases and vise versa. The empirical measure based on $X_1, \ldots, X_n$ is denoted $\mathbb{P}_n$. For a real-valued function f, its expectation with respect to P is denoted Pf. The expectation with respect to the true distribution is often denoted $\mathrm{E}f(X)$. The restriction of P onto a set A is denoted $P|_A$.

2. Consistency with respect to W_p

Recall that $W_p(P_n, P) \to 0$ if and only if P_n converges weakly to P and $M_p(P_n) \to M_p(P)$, see Theorem 7.12 of [43]. Also, the KL support condition (1.2), or (1.3), guarantees posterior consistency with respect to the Lévy-Prokhorov metric which induces the weak convergence. Therefore, it is natural under the KL support condition to guess that posterior consistency with respect to W_p is equivalent to the consistency of the pth moment, that is, for every $\epsilon > 0$,

$$\Pi\Big(\big|M_p(P) - M_p(P_0)\big| > \epsilon \mid X_1, \ldots, X_n\Big) \to 0 \quad \text{in probability.} \tag{2.1}$$

If (2.1) holds, we say that the posterior moment of order p is consistent. For $p = 1$, the moment consistency can be easily implied by W_1-consistency by the help of the duality theorem by [30], see also [17, 14, 44], which asserts that

$$W_1(P, Q) = \sup_{f \in \mathscr{L}} \left| \int f(x)dP(x) - \int f(x)dQ(x) \right|,$$

where $\mathscr{L}$ is the class of every Lipschitz function whose Lipschitz constant is bounded by 1. Since the map $x \mapsto |x|$ belongs to $\mathscr{L}$, we have that

$$|M_1(P) - M_1(Q)| \leq W_1(P, Q).$$

Although such an explicit bound does not exist for $p > 1$, one can show that posterior consistency with respect to the Wasserstein distance is equivalent to the moment consistency under the KL support condition.

Theorem 2.1. *For a prior* Π*, suppose that the KL condition* (1.3) *holds. Then, the consistency of the p-th moment* (2.1) *is equivalent to that*

$$\Pi\Big(W_p(P, P_0) \geq \epsilon \mid X_1, \ldots, X_n\Big) \longrightarrow 0 \quad \text{in probability for every } \epsilon > 0. \quad (2.2)$$

We provide two different approaches for proving posterior consistency with respect to W_p which are of independent interests. The first approach relies on a result from [45]; namely that if $\mathcal{C}$ is a convex set of probability measures and $\inf_{P \in \mathcal{C}} H(P_0, P) > 0$ then $\Pi(\mathcal{C}|X_1, \ldots, X_n) \to 0$ in probability, where $H(P, Q)$ denotes the Hellinger distance between P and Q. This approach directly uses Theorem 2.1 by establishing the consistency of the pth moment (2.1). The proof based on this approach is very simple as it only needs a single application of the Cauchy–Schwarz inequality. However, it requires the moment of order $2p$ to be bounded a posteriori.

The second approach constructs a uniformly consistent sequence of tests based on the convergence of empirical distribution. The uniformity does not make any problem for the compact support case, *i.e.* $P_0([-1, 1]) = 1$ and $P([-1, 1]) = 1$ for every P in the support of the prior Π. If probability measures in the support of the prior have unbounded support, however, problems may happen due to probability measures with large moments. This problem can be avoided if the moments are suitably bounded a posteriori, as expressed through condition (2.3) below. The second approach relies on a rather complicated proof, but it only needs the moment of order $p + \delta$, for some $\delta > 0$, to be bounded.

Theorem 2.2. *Assume that* Π *satisfies the KL condition* (1.3)*. Furthermore, assume that there exist positive constants* K *and* δ *such that* $M_{p+\delta}(P_0) < \infty$ *and*

$$\Pi\Big(M_{p+\delta}(P) \leq K \mid X_1, \ldots, X_n\Big) \to 1 \text{ in probability.} \quad (2.3)$$

Then for every $\epsilon > 0$*,*

$$\Pi\Big(W_p(P, P_0) > \epsilon \mid X_1, \ldots, X_n\Big) \longrightarrow 0 \text{ in probability.}$$

It should be emphasized that assumptions in Theorem 2.2 are nearly necessary. Certainly, $M_p(P_0) < \infty$ is necessary. Since the consistency with respect to W_p entails the consistency of the pth moment by Theorem 2.1, it is also necessary that for some constant K,

$$\Pi\Big(M_p(P) \leq K \mid X_1, \ldots, X_n\Big) \to 1 \quad \text{in probability.} \quad (2.4)$$

On the other hand, $M_p(P_0) < \infty$ and (2.4) are not sufficient for the posterior distribution to be consistent with respect to W_p, as shown in the following example.

Example. Let $P_0 = \delta_0$, $P_n = (1 - n^{-1})\delta_0 + n^{-1}\delta_{x_n}$ and $\Pi(\{P_0\}) = \Pi(\{P_n\}) = 1/2$, where δ_x is the Dirac measure at x and $x_n = n^{1/p}$. Obviously, the KL condition (1.3) holds. Furthermore, $W_p^p(P_0, P_n) = M_p(P_n) = 1 < \infty$ and $M_{p+\delta}(P_n) = n^{\delta/p} \to \infty$ for every $\delta > 0$. Since $P_0(X_1 = 0, \ldots, X_n = 0) = 1$ and $P_n(X_1 = 0, \ldots, X_n = 0) = (1 - n^{-1})^n \to e^{-1} > 0$, the posterior distribution is inconsistent with respect to W_p. Here, condition (2.4) holds, but (2.3) is violated for any $\delta > 0$.

By Theorem 2.2, the proof of the Wasserstein consistency boils down to

$$\Pi\Big(M_p(P) \le K \mid X_1, \ldots, X_n\Big) \longrightarrow 1 \quad \text{in probability} \tag{2.5}$$

for a constant K, a condition that seems easy to prove at first sight. However, the proof is not simple even with a well-known prior which puts all of its mass on the space of light-tailed distributions, that is, distribution with large or infinite tail index. Here, if a distribution function F satisfies $1 - F(x) = x^{-\alpha}L(x)$ for large enough x, where $L(\cdot)$ is a slowly varying function satisfying $\lim_{y\to\infty} L(xy)/L(y) = 1$ for any $x > 0$, the positive constant α is called the (right) tail index of F, see [33] for a Bayesian consistency of the tail index. It should be noted that a light-tail, *i.e.* large tail index, does not guarantee a small value of moment, which makes the proof of posterior consistency in W_p difficult. This is in stark contrast to that the moment of the empirical distribution can be trivially shown to be consistent. In Section 5.2, we are able to work out the case of Dirichlet process mixture prior by using Lemma 8.2, that is by establishing that the prior puts exponentially small mass to probability measures P with $M_p(P) > K$. See Theorem 5.1.

3. Convergence rates with respect to W_p

For a given rate sequence ϵ_n, suppose that $\Pi(\mathcal{K}_n) \ge e^{-n\epsilon_n^2}$ for every large enough n. Based on this condition, which is used to find a lower bound of the integrated likelihood, the denominator in the expression (1.1), we will extend the results of Section 2 to obtain a convergence rate. The main task in this section is to find additional assumptions required to achieve the convergence rate ϵ_n. An extension of the first proof for Theorem 2.2 requires the moment of order $2p$ as follows.

Theorem 3.1. *Assume that the prior Π satisfies the KL condition (1.4) for a sequence ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$. Furthermore, assume that there exists a constant K such that $M_{2p}(P_0) \le K$ and*

$$\Pi\Big(P: M_{2p}(P) > K \mid X_1, \ldots, X_n\Big) \to 0 \quad \text{in probability.}$$

Then

$$\Pi\Big(|M_p(P) - M_p(P_0)| > K'\epsilon_n \mid X_1, \ldots, X_n\Big) \to 0 \quad \text{in probability}$$

for some constant $K' > 0$.

Note that $M_p(P)$ is a linear functional of P for which the semi-parametric Bernstein–von Mises (BvM) theorem may hold, see [9, 37]. In this case, the convergence rate of the marginal posterior distribution of $M_p(P)$ would be the parametric rate $n^{-1/2}$ even though the global posterior convergence rate ϵ_n may be slower. However, while Theorem 3.1 is very general, the semi-parametric BvM theorem holds under rather strong conditions. For example, the above mentioned papers consider only specific priors and relied on the assumption that p_0 is compactly supported and bounded away from zero. It is sometimes possible to obtain the parametric convergence rate for the finite-dimensional parameter of interests without the semi-parametric BvM theorem. However, the proof typically relies on the LAN (locally asymptotically normal) expansion of the log-likelihood, see [6, 10].

Next, we consider an extension of the testing approach. To achieve the convergence rate ϵ_n, we will construct a sequence of consistent test

$$P_0\phi_n \to 0 \quad \text{and} \quad \sup_{P \in \mathcal{F}_n^c} P(1 - \phi_n) \leq e^{-3n\epsilon_n^2},$$

where $\mathcal{F}_n = \{P : W_p^p(P, P_0) \leq K\epsilon_n\} \cap \mathcal{F}_0$. Here, $\mathcal{F}_0$ will be defined as a collection of probability measures whose tails and moments are suitably bounded. Then, it will suffices for the desired result to show that $\Pi(\mathcal{F}_0^c \mid X_1, \ldots, X_n) \to 0$ in probability.

A consistent sequence of tests will be constructed based on the convergence of the empirical distribution to the true distribution. Note that there are well-known concentration inequalities of the form $P(W_p^p(\mathbb{P}_n, P) > \epsilon_n) \leq \delta_n$, where δ_n is a decaying sequence, and those inequalities might be directly used to define tests as

$$\phi_n = \begin{cases} 1 & \text{if } W_p^p(\mathbb{P}_n, P_0) > \epsilon_n \\ 0 & \text{otherwise.} \end{cases}$$

However, such a simple approach does not give sharp convergence rates of the posterior distribution. For example, if we apply the concentration inequality by [18], for any P with $W_p^p(P, P_0) > 2^p\epsilon_n$ and $M_{2p+\delta}(P) < \infty$, we have

$$\begin{aligned}
P(1 - \phi_n) &= P\Big(W_p^p(\mathbb{P}_n, P_0) \leq \epsilon_n\Big) \\
&\leq P\Big(W_p^p(\mathbb{P}_n, P) \geq 2^{-(p-1)}W_p^p(P, P_0) - \epsilon_n\Big) \\
&\leq P\Big(W_p^p(\mathbb{P}_n, P) \geq \epsilon_n\Big) \leq c_1\left(e^{-c_2 n\epsilon_n^2} + \frac{1}{n\epsilon_n^2}\frac{1}{(n\epsilon_n)^{\delta/2p}}\right)
\end{aligned} \tag{3.1}$$

where c_1 and c_2 are constants. Here, the constants c_1 and c_2 depends on the moments of P, so it is not easy to bound (3.1) uniformly. Furthermore, the second term in the right hand side of (3.1) is of polynomial order in $n\epsilon_n^2$ which decays too slowly compared to $e^{-n\epsilon_n^2}$. In turn, the use of ϕ_n would give a much slower convergence rate than ϵ_n.

Theorem 3.2 below is our main results concerning convergence rates of the posterior distribution. Proof of Theorem 3.2 relies on the set-up in [18]. In particular, Lemmas 8.3 and 8.5 can be easily deduced from the results in [18].

We build on these two lemmas to develop some techniques whose details are different from [18]. As mentioned above, we need to construct a sequence of tests decaying with an exponential rate. As far as we know, this is not possible with the proof technique used in [18]. Given a bounded moment condition, we achieve this by the help of Lemma 8.7. The condition $\epsilon_n \geq \sqrt{(\log n)/n}$ in Theorem 3.2 is assumed only for technical reason. Although we could not succeed to eliminate this condition, we believe the result is valid for any $\epsilon_n \downarrow 0$ with $n\epsilon_n^2 \to \infty$.

Theorem 3.2. *Assume that the prior Π satisfies the KL condition* (1.4) *for a sequence ϵ_n with $\epsilon_n \downarrow 0$ and $\epsilon_n \geq \sqrt{(\log n)/n}$. Furthermore, assume that there exist positive constants K and δ such that $M_{2p+\delta}(P_0) < \infty$ and $\Pi(M_{2p+\delta}(P) \leq K \mid X_1, \ldots, X_n) \to 1$ in probability. Then, for some constant $K' > 0$,*

$$\Pi\Big(W_p^p(P, P_0) \geq K'\big\{\epsilon_n + I(p = 1)\log \epsilon_n^{-1}\big\} \mid X_1, \ldots, X_n \Big) \longrightarrow 0$$

in probability.

Assumptions in Theorems 3.2 should be understood as sufficient conditions so that $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$ guarantees ϵ_n as the posterior convergence rate for any $\epsilon_n \gg n^{-1/2}$. For the empirical measure to achieve the rate $n^{-1/2}$ with respect to W_p^p, the same moment condition $M_{2p+\delta}(P_0) < \infty$ is considered in [18]. They provided an example showing that this moment condition cannot be weakened in general. As illustrated in the example at the end of this section, the posterior moment condition $\Pi(M_{2p+\delta}(P) \leq K \mid X_1, \ldots, X_n) \to 1$ cannot also be weakened.

When $p > 1$, Theorem 3.2 gives a rate ϵ_n with respect to W_p^p rather than W_p. This result is more relevant to [18] than [7]. In particular, condition $M_{2p+\delta}(P_0) < \infty$ is the same to Eq. (3) of [18], and much weaker than

$$\int \frac{[F_0(x)(1 - F_0(x))]^{p/2}}{p_0(x)^{p-1}} dx < \infty \tag{3.2}$$

which is a necessary and sufficient condition in [7] for that $\mathrm{E}[W_p(\mathbb{P}_n, P_0)] \asymp n^{-1/2}$. When $p = 1$, $M_{2p+\delta}(P_0) < \infty$ is only slightly stronger than (3.2), which is reduced to $\int \sqrt{F_0(x)(1 - F_0(x))}dx < \infty$, where F_0 is the cumulative distribution function of P_0. If $p > 1$, however, $M_{2p+\delta}(P_0) < \infty$ is much weaker than (3.2) which may not be satisfied even when P_0 is compactly supported. Note that if P_0 is standard normal, (3.2) is satisfied if and only if $1 \leq p < 2$. As mentioned in [7], the rate $\mathrm{E}[W_p(\mathbb{P}_n, P_0)] \asymp n^{-1/2}$ cannot be obtained under moment-type conditions considered in Theorem 3.2. Therefore, we would need a stronger assumption such as (3.2) to replace W_p^p by W_p in Theorem 3.2. Since we are focusing on the moment-type condition in the present paper, we do not address more detail about condition (3.2). Instead, we consider the metric W_∞ in Section 4 with a stronger assumption. Specifically, P_0 will be assumed to be supported on a bounded interval. This is necessary to obtain the consistency with respect to W_∞. The result in Section 4 guarantees the rate ϵ_n with respect to W_p, not W_p^p, at least when probabilities are compactly supported.

Note that our approach does not guarantee the rate $n^{-1/2}$ which is minimax optimal and achieved by the empirical measure under some general conditions, *e.g.* [47, 18, 7]. Our approach gives the rate $n^{-1/2}$ only if the prior puts sufficiently large mass around the KL neighborhood of p_0. This is mainly because our approach relies on the general approach of [20] for which the KL condition $\Pi(\mathcal{K}_n) \gtrsim e^{-n\epsilon_n^2}$ plays an important role to determine the rate. Also, note that the testing approach only gives sharp rates when the distance is compatible with the natural statistical distance of the model, the Hellinger distance in our case, see [27] for extensive discussion on this point. In this regards, it might not be possible to obtain sharp rates based on the testing approach. Hence, a different approach would be necessary to achieve the rate $n^{-1/2}$, *e.g.* the approach in [27, 48]. Another possible approach would be to utilize the functional Bernstein–von Mises theorem. Specifically, the approach given in [8], combining with the Kantorovich-Rubinstein representation, might give the rate $n^{-1/2}$ at least for W_1, and further limiting distribution of the posterior distribution. Note that the above papers are limited to specific priors and probability measures on a bounded set, while the present paper focuses on the moment condition for the posterior convergence rate. With these approaches, it would be highly interesting to investigate sufficient conditions to achieve the rate $n^{-1/2}$.

For $p = 1$, an additional logarithmic term in Theorem 3.2 can be eliminated if we assume a slightly stronger condition, which is satisfied if $p_0(x) \leq K|x|^{-(3+\delta)}$ and $\Pi(p(x) \leq K|x|^{-(3+\delta)} \ \forall x \mid X_1, \ldots, X_n) \to 1$ in probability for some positive constants K and δ, see Theorem 8.9 for details. Since Theorem 8.9 relies on some technical assumptions, we defer its statement to Section 8, and provide here a simpler statement, Theorem 3.2. Proofs of these theorems are quite similar.

Finally, we note that moment conditions in Theorem 3.2 cannot be weakened to $\delta < 0$ as shown in the following example.

Example. Let $P_0 = \delta_0$, $P_n = (1 - n^{-1})\delta_0 + n^{-1}\delta_{x_n}$ and $\Pi(\{P_0\}) = \Pi(\{P_n\}) = 1/2$. Certainly, the KL condition (1.4) holds for any $\epsilon_n \gg n^{-1/2}$. Note that the likelihood given P_n equals to $(1 - n^{-1})^n$ P_0-almost-surely, which converges to $e^{-1} > 0$. If $x_n = n^{1/(2p)+\delta}$ for small enough $\delta > 0$, then $W_p^p(P_0, P_n) = n^{-(1/2-p\delta)}$. Therefore, the posterior distribution is consistent with respect to W_p^p, but the rate of convergence is strictly slower than $\sqrt{(\log n)/n}$. In this example, note that $\int |x|^{2p} dP_n(x) = n^{2\delta p} \to \infty$, so the posterior moment condition in Theorem 3.2 is not satisfied.

4. Convergence rates with respect to W_∞

Since $W_p(P, Q)$ monotonically increases in p, one may define $W_\infty(P, Q) = \lim_{p\to\infty} W_p(P, Q)$ which, according to [24], corresponds to

$$W_\infty(P, Q) = \inf \left\{ \epsilon > 0 : P(A) \leq Q(A^\epsilon), \ \forall A \in \mathcal{R} \right\},$$

where $A^\epsilon = \{x : |x - y| < \epsilon \text{ for some } y \in A\}$ is the ϵ-enlargement of A and $\mathcal{R}$ is the set of all Borel subsets of $\mathbb{R}$. This representation of W_∞ bears similarities

with the Lévy-Prokhorov metric

$$d_P(P,Q) = \inf \left\{ \epsilon > 0 : P(A) \leq Q(A^\epsilon) + \epsilon, \; \forall A \in \mathcal{R} \right\}$$

which metrizes the weak convergence.

The metric W_∞ induces a much stronger topology than the weak topology even in a bounded metric space. In an unbounded space, if the tail index of two probability measures P and Q are different, then $W_\infty(P,Q)$ is typically infinity. For example, if P and Q are Student's t-distributions with ν_1 and ν_2 degrees of freedom with $\nu_1 \neq \nu_2$, then $W_\infty(P,Q) = \infty$. Therefore, it is meaningless to study asymptotics with W_∞ in an unbounded space.

In this section, we assume that P_0 is supported in the unit interval $[0,1]$, and so are all probability measures in the support of the prior. Our benchmarking assumption is $\inf_{x \in [0,1]} p_0(x) \geq c_0$ for some constant $c_0 > 0$, which is a necessary and sufficient condition for that $P_0[W_\infty(\mathbb{P}_n, P_0)] \asymp n^{-1/2}$, see [7].

Theorem 4.1. *Suppose that p_0 is a density on $[0,1]$ and $\inf_{x \in [0,1]} p_0(x) \geq c_0$ for some constant $c_0 > 0$. Also, assume that the prior Π satisfies the KL condition (1.4) for a sequence ϵ_n with $\epsilon_n \downarrow 0$ and $\epsilon_n \geq \sqrt{(\log n)/n}$ and $\Pi(P([0,1]) = 1) = 1$. Then, for some constant $K > 0$,*

$$\Pi\left(W_\infty(P,P_0) > K\epsilon_n \;\middle|\; X_1, \ldots, X_n\right) \longrightarrow 0 \quad \text{in probability.} \qquad (4.1)$$

5. Examples

In this section, we consider the posterior moment condition (2.5) with two examples. In the first example, we illustrate the idea of a novel approach handling the second moment condition without full technical details. The approach relies on a special property of gamma distributions. The second example considers higher order moments, and concrete posterior convergence rates are derived.

Note that (2.5) holds trivially if the prior satisfies

$$\Pi\left(p : p(x) \leq K' t_p(x) \; \forall x\right) = 1 \qquad (5.1)$$

for some K', where t_p is the density of the Student's t distribution with p degrees of freedom. Such a prior can be easily constructed by conditioning well-known priors by the event in the left hand side of (5.1). Although the prior probability for this event would be close to 1 with most priors and large enough K, this conditioning might be unnatural in practice.

5.1. Mixture of gamma distributions

Suppose it is required to establish weak consistency alongside a functional constraint; such as

$$\Pi\left(p : \int x^2 p(x)\,dx > K \mid X_1, \ldots, X_n\right) \to 0$$

in probability, for some finite $K > 0$, with the prior Π on density functions on $(0, \infty)$. This would be for establishing W_1 consistency.

With the usual Kullback–Leibler support condition, we write the posterior as

$$d\Pi(p \mid X_1, \ldots, X_n) = \frac{\prod_{i=1}^{n} \alpha(x_i) \frac{p(x_i)}{p_0(x_i)} d\Pi(p)}{\int \prod_{i=1}^{n} \alpha(x_i) \frac{p(x_i)}{p_0(x_i)} d\Pi(p)},$$

for some function α. Using standard arguments, the P_0-expectation of the numerator over the set A is

$$\int_A \left(\int \alpha \, dP \right)^n d\Pi(p)$$

and the reciprocal of the denominator is upper bounded, i.e.

$$e^{nd} \, e^{n \int \log(1/\alpha) \, dP_0} \quad \text{a.s.}$$

for all large n, for any $d > 0$; see [39] for details about this argument. Also note

$$\exp \left\{ n \int \log(1/\alpha) \, dP_0 \right\} \le \left(\int \alpha^{-1} \, dP_0 \right)^n.$$

Hence, if $A = \{p : \int \alpha^{-1} \, dP > K\}$ and we construct the prior Π so that

$$\int \alpha^{-1} \, dP > K \implies \int \alpha \, dP < \epsilon < \left(\int \alpha^{-1} \, dP_0 \right)^{-1},$$

then, a.s. for all large n, using the Markov inequality and the Borel–Cantelli lemma,

$$\Pi(A \mid X_1, \ldots, X_n) \le e^{nd} \, \epsilon^n \left(\int \alpha^{-1} \, dP_0 \right)^n \to 0 \quad \text{a.s.}$$

We obtain the second moment result by taking $\alpha(x) = 1/x^2$ and so we need to ensure for the prior, for any $\epsilon > 0$, there exists a $K < \infty$ such that $\int x^2 \, dP > K$ implies $\int x^{-2} \, dP < \epsilon$.

Let κ_0 be the true second moment and assume we can construct the model $p(x)$ such that for any $\epsilon > 0$ there exists a $K > 0$ such that

$$\int x^2 p(x) \, dx > K \implies \int x^{-2} p(x) \, dx < \epsilon < 1/\kappa_0.$$

This also implies $K > \kappa_0$. Such an example arises with the gamma distribution, so consider $p(x) = \Gamma(a)^{-1} x^{a-1} e^{-x}$, where we have $\mathrm{E}\, X^2 = a(a+1)$ and $\mathrm{E}\, X^{-2} = [(a-1)(a-2)]^{-1}$. Hence, $\mathrm{E}\, X^2 > K$ implies $\mathrm{E}\, X^{-2} < \epsilon$ for some suitably large K, for $a > 2$.

For a more general nonparametric model, consider

$$p(x) = \sum_{j=1}^{M} w_j \Gamma(x \mid a_j, b),$$

a mixture of gamma distributions. We can assign priors to M and w but to describe the prior for $a = (a_1, \ldots, a_M)$ and b there is no loss in generality in

fixing them. Now

$$\mathrm{E}\, X^2 = \sum_{j=1}^{M} w_j \frac{a_j(1+a_j)}{b^2}$$

and, if, as assumed, $a_j > 2 + \delta$, then

$$\mathrm{E}\, X^{-2} = \sum_{j=1}^{M} w_j \frac{b^2}{(a_j - 1)(a_j - 2)}.$$

To obtain our required condition, we take the prior so that if a single $a_j(1 + a_j)/b^2 > K$ then it is true for all j. This ensures that if $\mathrm{E}\, X^2 > K$ then $a_j(1 + a_j)/b^2 > K$ for all j and then it is also true that $b^2/((a_j - 1)(a_j - 2)) < \xi/K$, for all j, for some $\xi < \infty$, which is fixed. Indeed, $\xi = (2+\delta)(3+\delta)/(\delta(1+\delta))$, so $\mathrm{E}\, X^{-2} < \xi/K$.

Hence, we take the prior for (a, b) as

$$\pi(a, b) = \pi(b) \left[(1 - q) \prod_{j=1}^{M} g_{c-}(a_j|b) + q \prod_{j=1}^{M} g_{c+}(a_j|b) \right],$$

where g_{c-} is a density on $(0, c)$ and g_{c+} a density on (c, ∞) for some c. Here, $g_{c-}(a_j|b)$ puts all the mass on $a_j(1 + a_j) < cb^2$ and $g_{c+}(a_j|b)$ puts all the mass on $a_j(1 + a_j) > cb^2$. In practice, we can take c so large that the part of the prior which contributes to the posterior will only be the g_{c-} component.

5.2. Dirichlet process mixture

Consider a Dirichlet process mixture prior

$$p(x) = \int \phi_\sigma(x - z)dG(z), \quad G \sim \mathrm{DP}(\alpha H), \tag{5.2}$$

where $\mathrm{DP}(\alpha H)$ denotes the Dirichlet process with base measure αH, $\phi_\sigma(x) = \sigma^{-1}\phi(x/\sigma)$ and ϕ is the standard normal density. In practice, an inverse gamma prior is usually imposed for σ^2, but we consider a fixed sequence $\sigma = \sigma_n \to 0$ for technical convenience. Note that the sequence σ_n controls the convergence rate. Specifically, with a suitable sequence σ_n, one can prove that $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$, see [21, 40]. While these papers extensively studied posterior convergence rates with respect to the Hellinger metric, posterior moments have not been studied thoroughly. Only Section 8 of [21] slightly touched the tail mass of the posterior distribution. However, their result relies on the assumption that P_0 is compactly supported, and cannot be directly used to bound the posterior moments.

Note that the posterior moment condition (2.5) is similar to

$$\Pi\Big(P(B_m) \leq K2^{-pm} \ \forall m \geq 0 \mid X_1, \ldots, X_n\Big) \longrightarrow 1 \text{ in probability}, \tag{5.3}$$

where $B_0 = (-1, 1]$ and $B_m = (-2^m, 2^m]\setminus(-2^{m-1}, 2^{m-1}]$ for $m \geq 1$. Since

$$\frac{1}{2^p} \sum_{m \geq 1} 2^{pm} P(B_m) \leq M_p(P) \leq \sum_{m \geq 0} 2^{pm} P(B_m) \tag{5.4}$$

for any probability measure P and $p \geq 1$, (2.5) is implied by

$$\Pi\Big(P(B_m) \leq K' 2^{-(p+\delta)m} \ \forall m \geq 0 \mid X_1, \ldots, X_n\Big) \longrightarrow 1 \text{ in probability}$$

for some positive constants K' and δ.

Suppose that $P_0(B_m) \lesssim 2^{-pm}$ for every $m \geq 0$. Under the assumption that $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$, it is not difficult to show that

$$\Pi\Big(|P(B_m) - P_0(B_m)| \gtrsim \epsilon_n \mid X_1, \ldots, X_n\Big) \to 0 \text{ in probability}$$

for every $m \geq 0$. If $2^{-pm} \geq \epsilon_n$, or equivalently $m \leq p^{-1} \log_2 \epsilon_n^{-1}$, then the posterior probability

$$\Pi\Big(P(B_m) \lesssim 2^{-pm} \mid X_1, \ldots, X_n\Big)$$

will be close to 1 for large enough n with high P_0-probability. More generally, one can show that

$$\Pi\Big(P(B_m) \lesssim 2^{-pm} \ \forall m \leq p^{-1} \log_2 \epsilon_n^{-1} \mid X_1, \ldots, X_n\Big) \to 1 \text{ in probability.}$$

If $m > p^{-1} \log_2 \epsilon_n^{-1}$, however, one cannot bound $P(B_m)$ by 2^{-pm} because the convergence rate ϵ_n is larger than 2^{-pm}. In this case, the prior must play a role, that is, the prior probability that $P(B_m) \gtrsim 2^{-pm}$ should be small. In fact, this prior probability should be exponentially small, with an order $e^{-cn\epsilon_n^2}$ for some constant $c > 0$, to guarantee that the posterior probability also decays, cf. Lemma 8.2. To this aim we will make use of

$$G(B_m) \sim \text{Beta}\Big(\alpha H(B_m), \alpha\big(1 - H(B_m)\big)\Big)$$

for every $m \geq 0$, which in particular implies that the prior expectation of $G(B_m)$ equals $H(B_m)$. If H is a normal distribution (any H with sub-Gaussian tail would actually work), the prior expectation of $G(B_m)$ is much smaller than 2^{-pm} for every large enough m.

Theorem 5.1. *Let H be the normal distribution with mean μ_H and variance σ_H^2. Let Π be a Dirichlet process mixture prior (5.2) with $\alpha > 1$ and $\sigma = \sigma_n = n^{-1/5}$. Also, suppose that p_0 is twice continuously differentiable with a sub-Gaussian tail, and satisfies $\int (p_0''/p_0)^2 + (p_0'/p_0)^4 dP_0 < \infty$. Then, for any $p < 4$, there exists a constant K such that*

$$\Pi\Big(M_p(P) \leq K \mid X_1, \ldots, X_n\Big) \to 1 \quad \text{in probability.}$$

If the prior and P_0 satisfy conditions in Theorem 5.1, the posterior distribution is consistent with the rate $n^{-2/5}$ with respect to W_p^p for $p < 2$, up to a

logarithmic factor. Once P_0 possesses a smoother density, it is possible to prove the consistency of higher order moments, see Lemma 8.13 for more details.

If we impose a prior on σ^2, it can be deduced from the proof that the assertion of Lemma 8.13 is still valid provided that σ^2 is bounded a posteriori, that is, $\Pi(\sigma^2 > K \mid X_1, \ldots, X_n) \to 0$ for some constant $K > 0$. Under mild assumptions, the posterior distribution of σ^2 will be concentrated around 0 unless P_0 itself is a location mixture of normal distributions. If P_0 is a location mixture of normal distributions, the posterior probability that $\sigma^2 > \sigma_0^2 + \epsilon$ vanishes, where σ_0^2 is the true parameter.

6. Numerical study

Although theoretical results given in previous sections provide reasonable sufficient conditions for the Wasserstein consistency, those conditions are not easy to verify in practice. With a DP mixture prior, for example, the rate ϵ_n determined by $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$ plays an important role for the consistency with respect to W_p. However, it is very difficult to find exact rate ϵ_n satisfying $\Pi(\mathcal{K}_n) \asymp e^{-n\epsilon_n^2}$. Note also that if P_0 has an unbounded support, the posterior distribution is typically inconsistent with respect to W_∞. Since $W_p \uparrow W_\infty$ as $p \uparrow \infty$, the posterior distribution will be consistent with respect to W_p only for small values of p, where the threshold value depends on ϵ_n. Perhaps the most interesting cases would be $p = 1$ or $p = 2$, so in this section, we empirically show that the posterior distribution tends to be consistent with respect to W_1 and W_2 with popularly used priors.

We consider DP mixtures of Gaussian priors described in Section 5.2. Instead of a decaying sequence σ_n, we put an inverse gamma prior on σ^2 as usual in practice. Specifically, we used $H = N(\mu_H, \sigma_H)$, $\sigma^2 \sim \Gamma^{-1}(\beta, \lambda)$ and $\alpha \sim \Gamma(\beta_\alpha, \lambda_\alpha)$ with $\sigma_H = \beta = \lambda = \beta_\alpha = \lambda_\alpha = 1$ and $\mu_H = 0$, where β and λ denotes the shape and rate parameters of the gamma distribution. In addition to the location mixture, we also consider a location-scale mixture

$$p(x) = \int \phi_\sigma(x - z) dG(z, \sigma), \quad G \sim \mathrm{DP}(\alpha H),$$

where H is the normal-inverse gamma distribution. In this case, we used $H = \mathrm{N\text{-}\Gamma}^{-1}(\mu_H, \sigma_H, \beta, \lambda)$ and $\alpha \sim \Gamma(\beta_\alpha, \lambda_\alpha)$ with $\sigma_H = \beta = \lambda = \beta_\alpha = \lambda_\alpha = 1$ and $\mu_H = 0$, where $(X, Y) \sim \mathrm{N\text{-}\Gamma}^{-1}(\mu, \sigma, \beta, \lambda)$ means that $X \mid Y \sim N(\mu, Y/\sigma)$ and $Y \sim \Gamma^{-1}(\beta, \lambda)$. Note that an inverse gamma distribution has a tail of polynomial order, so with a location-scale mixture, the prior probability that $P(B_m) \geq 2^{-pm}$ may not be too small.

There are several computational algorithms sampling from a posterior distribution based on a Dirichlet process mixture prior, see [35, 29] and references therein. Unfortunately, given a posterior sample P, it is very difficult to compute the Wasserstein distance $W_p(P, P_0)$, see Theorem 3 of [31]. Instead of directly calculating $W_p(P, P_0)$, we can easily generate a Markov chain sample $Y_1, \ldots, Y_N$ from the posterior predictive distribution $\int p(x) d\Pi(p \mid X_1, \ldots, X_n)$. Then, the corresponding empirical distribution $\widetilde{\mathbb{P}}_N$ can be used as a proxy of the posterior

predictive distribution. Note that the empirical distribution from an ergodic Markov chain, as well as the one from an iid sample, contracts to the stationary distribution with respect to the Wasserstein metrics, see [18]. However, it is still not easy to compute $W_p(\widetilde{\mathbb{P}}_N, P_0)$. To evaluate $W_p(\widetilde{\mathbb{P}}_N, P_0)$, we first approximate P_0 by a discrete measure $\mathbb{Q}_M$ and find $W_p(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$. If M is a multiple of N, one can easily find exact value of $W_p(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$ based on the following lemma taken from [7].

Lemma 6.1. *For given two collections of real numbers $x_1 \leq \cdots \leq x_N$ and $y_1 \leq \cdots \leq y_N$, let $\mathbb{P}$ and $\mathbb{Q}$ be the corresponding empirical measures. Then, for any $p \geq 1$,*

$$W_p^p(\mathbb{P}, \mathbb{Q}) = \frac{1}{N} \sum_{k=1}^{N} |x_k - y_k|^p.$$

To approximate P_0 by $\mathbb{Q}_M$, assume for a moment that P_0 is symmetric about the origin. For an even integer M, let $x_k = q(1/2 + k/M)$ for $k = 0, \ldots, M/2 - 1$ and $\mathbb{Q}_M$ be the probability measure such that $\mathbb{Q}_M(\{x_0\}) = 2/M$, $\mathbb{Q}_M(\{x_k\}) = 1/M$ and $\mathbb{Q}_M(\{-x_k\}) = 1/M$ for $k \geq 1$, where $q : (0,1) \to \mathbb{R}$ is the quantile function of P_0. Then,

$$W_p^p(P_0, \mathbb{Q}_M) \leq 2 \int_{x_{M/2-1}}^{\infty} |x - x_{M/2-1}|^p dP_0(x) + \frac{2}{M/2 - 1} \sum_{k=1}^{M/2-1} |x_k - x_{k-1}|^p.$$
(6.1)

Since $W_p(P_0, \mathbb{Q}_M) \to 0$ as $M \to \infty$, one can approximate $W_p(\widetilde{\mathbb{P}}_N, P_0)$ by $W_p(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$. For a non-symmetric P_0, a similar approximation $\mathbb{Q}_M$ can be obtained after replacing the origin by the median. For various true distributions– standard uniform, standard normal, Laplace, Student's t with 20, 10, 5 degrees of freedoms–the approximation error, the upper bound of $W_p(P_0, \mathbb{Q}_M)$, is depicted in Figure 1. When $p = 1$ and $p = 2$, the approximation of P_0 by $\mathbb{Q}_M$ is quite accurate for all cases. On the other hand, for $p = 4$ and $p = 8$, the approximation is not reliable unless the support of the true distribution is bounded.

With the above six true distributions, we generated $n = 50, 100, 200, \ldots, 6400$ samples and obtained $N = 10^4$ MCMC samples from the posterior predictive distributions after 1000 burn-in periods. Then, we evaluated the Wasserstein distance $W_p(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$ between the empirical distribution $\widetilde{\mathbb{P}}_N$ of MCMC sample and the discrete approximation $\mathbb{Q}_M$ of P_0 with $M = 2 \times 10^5$. We considered $p = 1$ and $p = 2$ only because because the approximation by $\mathbb{Q}_M$ is not reliable for large p. We repeated the above procedure for 100 times and the median among 100 repetitions are depicted in Figures 2 and 3. As can be seen, the posterior predictive distributions become closer to the approximation $\mathbb{Q}_M$ of the true distribution as the sample size increases. Interestingly, it seems that the location-scale mixture prior also gives consistent posterior distributions with respect to both W_1 and W_2 for all cases. Figure 4 shows similar results with a location mixture prior with different hyperparameter $H = N(0, 10^4)$. Note that a normal distribution with large variance is a natural choice for H in
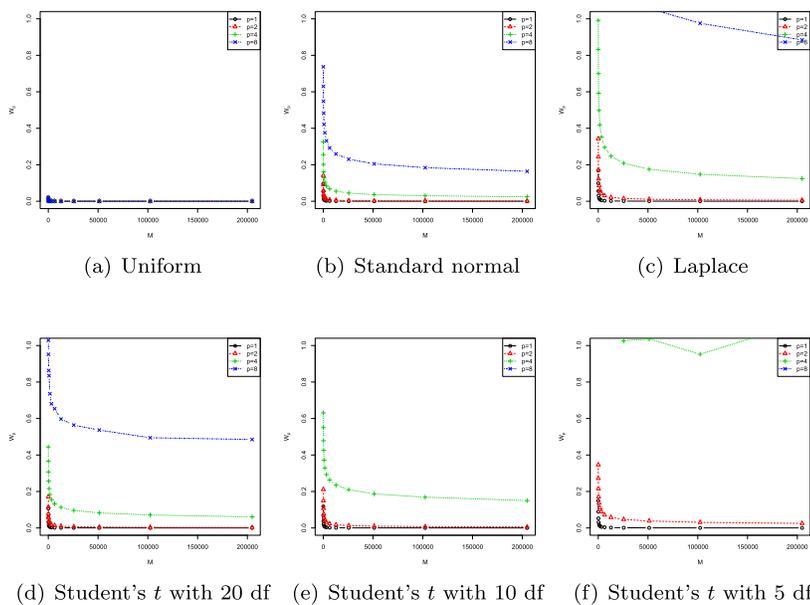
(a) Uniform (b) Standard normal (c) Laplace

(d) Student's t with 20 df (e) Student's t with 10 df (f) Student's t with 5 df

Fig 1. *The upper bound of $W_p(P_0, \mathbb{Q}_M)$ for various true distributions.*

practice. The results in Figure 4 shows that the posterior distribution seems to be consistent with respect to W_2, but more samples are needed to dominate prior probabilities on the tail. This is because some posterior predictive samples might be very large when the number of observation is small, and $W_2(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$ is more sensitive to these large samples than $W_1(\widetilde{\mathbb{P}}_N, \mathbb{Q}_M)$.

7. Discussion

In this paper, we provided sufficient conditions for posterior consistency with respect to the Wasserstein metrics and the convergence rate to be ϵ_n in addition to the well-known KL conditions. Based on our main theorem, the posterior probability that $W_p^p(P, P_0) \gtrsim \epsilon_n$ vanishes if $M_{2p+\delta}(P)$ is bounded by a constant for some $\delta > 0$ with high posterior probability. A similar moment condition has been used in [18] to show that $W_p^p(\mathbb{P}_n, P_0) \asymp n^{-1/2}$ with high probability. The moment condition cannot be weakened in general as illustrated in our examples. Under a stronger condition (3.2), which is a necessary and sufficient condition for $W_p(\mathbb{P}_n, P_0) \asymp n^{-1/2}$, we conjecture that the posterior probability that $W_p(P, P_0) \gtrsim \epsilon_n$ would vanish.

We note that asymptotic results given in this paper might be utilized to obtain posterior consistency and its convergence rate with respect to strong metrics such as the total variation. For this, the key is to obtain posterior convergence rate in the Wasserstein metric and bound the total variation between smooth
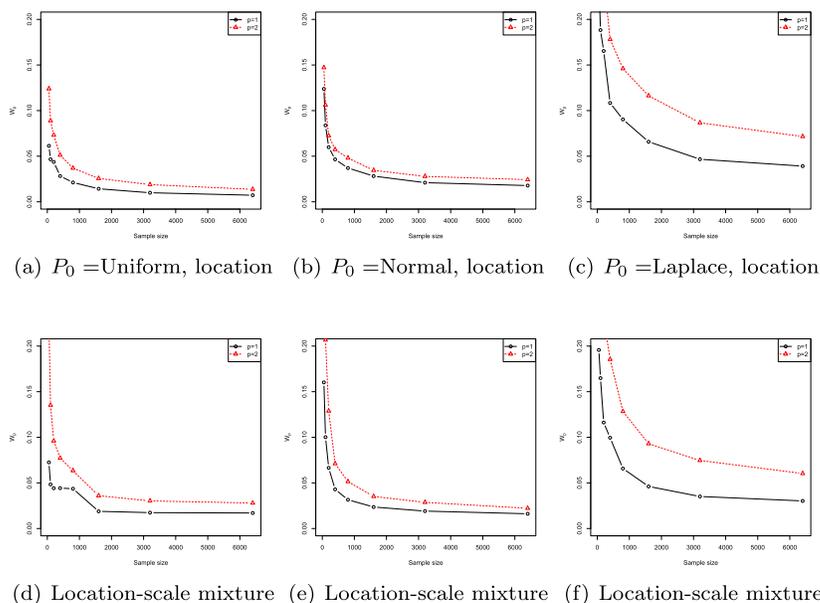
(a) P_0 =Uniform, location   (b) P_0 =Normal, location   (c) P_0 =Laplace, location

(d) Location-scale mixture (e) Location-scale mixture (f) Location-scale mixture

Fig 2. *Wasserstein distances between the true distribution–uniform (left), normal (middle), and Laplace (right)–and the posterior predictive distributions based on location (upper) and location-scale mixtures (lower) of Gaussians.*

densities by a power of the Wasserstein metric. More precisely, if P and Q possess smooth Lebesgue densities p and q, one can prove that $\|p-q\|_1 \lesssim W_p^\alpha(P,Q)$ for some $\alpha > 0$, see [13] for a sharp inequality. This is a certain reverse inequality because the total variation generates a stronger topology than W_p in the space of all probability measures on a bounded metric space. This kind of reverse inequality and related theory for posterior consistency are the main motivation of the present paper, which was firstly considered in [11].

We conclude by discussing an example where total variation consistency and a mild condition implies Wasserstein consistency. This is a non–trivial finding. For a given kernel density function k on $\mathbb{R}$, consider a location mixture

$$p(x) = \int k(x - z)dG(z), \tag{7.1}$$

which is often called a convolution. A prior Π on p can be induced from a prior on the mixing distribution G. With slight abuse of notation, we use the same notation Π for the prior of G. Suppose that the true distribution is also of the form (7.1), that is $p_0(x) = \int k(x-z)dG_0(z)$ for some probability measure G_0. In this case, posterior consistency with respect to the total variation automatically implies the consistency in W_2. Suppose that k is symmetric about the origin, $\int x^2 k(x)dx < \infty$, and that $\tilde{k}(t) \neq 0$ for every $t \in \mathbb{R}$, where $\tilde{k}$ is the Fourier transform of f defined as $\tilde{k}(t) = \int e^{-itx}k(x)dx$.
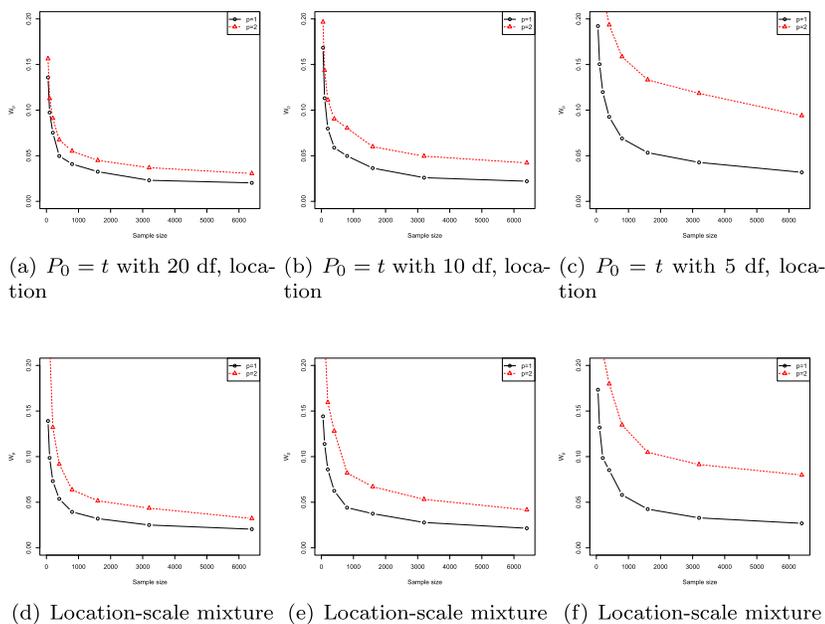
(a) $P_0 = t$ with 20 df, location

(b) $P_0 = t$ with 10 df, location

(c) $P_0 = t$ with 5 df, location

(d) Location-scale mixture (e) Location-scale mixture (f) Location-scale mixture

FIG 3. *Wasserstein distances between the true distribution–Student's t distribution with 20 (left), 10 (middle), and 5 (right) degrees of freedom–and the posterior predictive distributions based on location (upper) and location-scale mixtures (lower) of Gaussians.*

Theorem 7.1. *Suppose that the kernel k satisfies the assumption described above. Assume also that $M_2(G_0) < \infty$ and $\Pi(M_2(G) < \infty) = 1$. Then, $E\Pi(\|p - p_0\|_1 > \epsilon \mid X_1, \ldots, X_n) \to 0$ for every $\epsilon > 0$ implies that*

$$E\Pi\Big(W_2(P, P_0) > \epsilon \mid X_1, \ldots, X_n\Big) \to 0$$

for every $\epsilon > 0$.

Note that the condition $\Pi(M_2(G) < \infty) = 1$ is easily satisfied for well-known priors. For example, if we put a Dirichlet process prior for G, the tail of G is much lighter than that of its mean, see [16]. Posterior consistency with respect to the total variation can also be easily established using a standard technique.

8. Proofs

Firstly, we introduce a basic set-up which is taken from [18] with slight modification, see also [15, 46]. For nonnegative integers l, let $\mathcal{P}_l$ be the natural partition of $(-1, 1]$ into 2^l translations of $(-2^{-l}, 2^{-l}]$, that is,

$$\mathcal{P}_l = \Big\{(-1 + k2^{1-l}, -1 + (k+1)2^{1-l}] : k = 0, 1, \ldots, 2^l - 1\Big\}.$$

(a) $P_0 = $ Uniform   (b) $P_0 = $ Normal   (c) P_0 Laplace

(d) $P_0 = t$ with 20 df   (e) $P_0 = t$ with 10 df   (f) $P_0 = t$ with 5 df
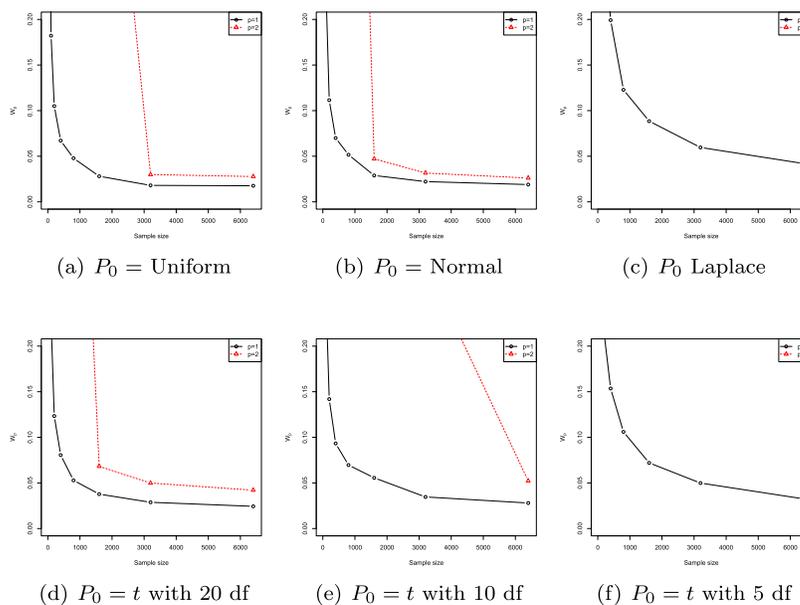
FIG 4. *Wasserstein distances between the true distributions and the posterior predictive distributions based on the location mixture with $H = N(0, 10^4)$.*

Let $B_0 = (-1, 1]$ and $B_m = (-2^m, 2^m] \setminus (-2^{m-1}, 2^{m-1}]$ for $m \geq 1$. Let $\pi_m : \mathbb{R} \to \mathbb{R}$ be the function defined as $\pi_m(x) = x/2^m$, and $\mathcal{R}_{B_m} P$ be the probability measure on $(-1, 1]$ defined as the π_m-image of $P|_{B_m}/P(B_m)$, that is, for any Borel set $F \subset (-1, 1]$,

$$\mathcal{R}_{B_m} P(F) = \frac{P(\pi_m^{-1}(F) \cap B_m)}{P(B_m)}$$

or $\mathcal{R}_{B_m} P(F) = 0$ according as $P(B_m) > 0$ or $P(B_m) = 0$.

To get some insight of overall proofs, we next address how one can obtain the consistency of the empirical distribution with respect to W_p. Suppose for a moment that P_0 is supported on $[-1, 1]$. Then, Lemma 8.3 implies that if $|\mathbb{P}_n(F) - P_0(F)|$ is sufficiently small for every $F \in \mathcal{P}_l$ and $l \leq L$, where L is a large enough constant, then $W_p(\mathbb{P}_n, P_0)$ will also be small. Since there are various tools to bound the deviation $|\mathbb{P}_n(F) - P_0(F)|$, *e.g.* the inequality by [26], it is not difficult to prove that the empirical distribution converges to P_0 in probability with respect to W_p, $1 \leq p < \infty$, with the help of Lemma 8.3.

In case that P_0 has an unbounded support, Lemma 8.5 can be applied for the Wasserstein consistency of $\mathbb{P}_n$. Indeed, if $|\mathbb{P}_n(\pi_m^{-1}(F)) - P_0(\pi_m^{-1}(F))|$ is sufficiently small for every $F \in \mathcal{P}_l$, $l \leq L$ and $m \leq M$, where L and M are large constants, then $W_p(\mathbb{P}_n, P_0)$ will be small. Note that L and M can be chosen as large but fixed constants, so the consistency of $\mathbb{P}_n$ can be similarly proven using a large deviation inequality such as the Hoeffding's inequality. Here, it plays an

important role that $M_p(\mathbb{P}_n)$ converges to $M_p(P_0)$ by the law of large numbers, because once the pth moment of $\mathbb{P}_n$ and P_0 is bounded, it is relatively easy to prove the Wasserstein consistency, see the proof of Theorem 2.2 and Lemma 8.6 for details.

8.1. Frequently used results from literature

The KL condition (1.4) gives a suitable lower bound of the integrated likelihood, that is, the denominator in (1.1). Once this condition holds, the posterior probability of a sequence of subsets $\mathcal{F}_n$ of $\mathcal{F}$ can be shown to converge to 1 if the prior probability of $\mathcal{F}_n^c$ or likelihood is sufficiently small. The latter can often be expressed through the existence of a certain sequence of uniformly consistent tests. Lemmas 8.1 and 8.2 are taken from [20] with slight modification for the simplicity. The rate sequence ϵ_n is assumed that $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$.

Lemma 8.1. *Suppose that* $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$ *and assume that there exists a sequence of tests* ϕ_n *such that*

$$P_0\phi_n \to 0 \quad \text{and} \quad \sup_{P \in \mathcal{F}_n^c} P(1 - \phi_n) \leq e^{-3n\epsilon_n^2}$$

for $\mathcal{F}_n \subset \mathcal{F}$. *Then,* $\Pi\left(\mathcal{F}_n^c \mid X_1, \ldots, X_n\right) \to 0$ *in probability.*

Lemma 8.2. *Suppose that* $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$ *and* $\Pi(\mathcal{F}_n^c) \leq e^{-3n\epsilon_n^2}$ *for* $\mathcal{F}_n \subset \mathcal{F}$. *Then,* $\Pi\left(\mathcal{F}_n^c \mid X_1, \ldots, X_n\right) \to 0$ *in probability.*

The following lemmas are taken from [18] with slight modification, see also [15, 46]. Since the statement of Lemma 8.3 is slightly different from these papers, we provide a detailed proof for the reader's convenience.

Lemma 8.3. *Assume that two probability measures P and Q are supported on* $(-1, 1]$. *Then,*

$$W_p^p(P, Q) \leq \kappa_p \left(\sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} |P(F) - Q(F)| + 2^{-Lp} \right)$$

for every $L \geq 1$, where κ_p is a constant depending only on $p \geq 1$.

Proof. For a Borel partition $\{A_k : k \geq 1\}$ of a Borel set $A \subset \mathbb{R}$ and two finite measures P and Q on A with equal mass, define the finite measure $\overline{P}$ as

$$\overline{P}|_{A_k} = \frac{Q(A_k)}{P(A_k)} P|_{A_k}$$

if it is well-defined, that is, $P(A_k) = 0$ implies $Q(A_k) = 0$. Here, $P|_{A_k}$ and $\overline{P}|_{A_k}$ denote the restrictions of P and $\overline{P}$, respectively, onto A_k. We say $\overline{P}$ is the $\{A_k : k \geq 1\}$-approximation of P to Q. Then, we have the following lemma whose proof is explicitly given in [15] (pp. 1189–1190).

Lemma 8.4. *Suppose that the $\{A_k : k \geq 1\}$-approximation $\overline{P}$ of P to Q is well-defined. Then, there exists a coupling ξ of P and $\overline{P}$ such that*

$$\xi\Big(\{(x,y) : x \neq y\}\Big) = \frac{1}{2}\sum_{k \geq 1}|P(A_k) - Q(A_k)|.$$

For $l \geq 0$, let P_l be the $\mathcal{P}_l$-approximation of P to Q. We only consider the case that P_l is well-defined for all $l \geq 0$. The other case can be handled with further details, see Proposition 1 of [46].

Since $P_l(F) = Q(F)$ for $F \in \mathcal{P}_l$, we have $W_p(P_l, Q) \leq 2^{-(l-1)}$ for every $l \geq 0$. Furthermore, it is easy to check that, for $F \in \mathcal{P}_l$, $P_l(F) = P_{l+1}(F)$ and $P_{l+1}|_F$ is the $\{C \in \mathcal{P}_{l+1} : C \subset F\}$-approximation of $P_l|_F$ to $Q|_F$. Therefore, by Lemma 8.4, there exists a coupling ξ_{l+1} of P_l and P_{l+1} such that

$$\xi_{l+1}\Big(\{(x,y) : x \neq y\}\Big) = \frac{1}{2}\sum_{F \in \mathcal{P}_l}\sum_{\substack{C:C \subset F \\ C \in \mathcal{P}_{l+1}}}\Big|P_l|_F(C) - Q|_F(C)\Big|$$

$$= \frac{1}{2}\sum_{F \in \mathcal{P}_l}\sum_{\substack{C:C \subset F \\ C \in \mathcal{P}_{l+1}}}\Big|P_l(C) - Q(C)\Big| = \frac{1}{2}\sum_{F \in \mathcal{P}_l}\sum_{\substack{C:C \subset F \\ C \in \mathcal{P}_{l+1}}}\Big|Q(C) - \frac{Q(F)}{P(F)}P(C)\Big|.$$

It follows that there exist random variables $Z_0, Z_1, Z_2, \ldots$ in a same probability space, say $(S, \mathcal{S}, \mu)$, such that

$$\mu\Big(|Z_{l+1} - Z_l| \leq 2^{-(l-1)}\Big) = 1,$$

$$\mu\Big(Z_{l+1} \neq Z_l\Big) = \frac{1}{2}\sum_{F \in \mathcal{P}_l}\sum_{\substack{C:C \subset F \\ C \in \mathcal{P}_{l+1}}}\Big|Q(C) - \frac{Q(F)}{P(F)}P(C)\Big|$$

and Z_l is marginally distributed as P_l. Let $N = \inf\{l : Z_{l+1} \neq Z_l\}$, where the infimum of the empty set is set to be infinity. Then, conditional on the event $\{N = l\}$ with $l < L$, where L is a fixed positive integer, we have

$$|Z_0 - Z_L| \leq \sum_{l'=l}^{L-1}|Z_{l'} - Z_{l'+1}| \leq 2^{-(l-2)}$$

with probability one. It follows that

$$\mathbb{E}|Z_0 - Z_L|^p \leq \sum_{l=0}^{L-1}\mathbb{E}\Big[|Z_0 - Z_L|^p \ \Big| \ N = l\Big]\mu\Big(N = l\Big)$$

$$\leq \sum_{l=0}^{L-1}2^{-(l-2)p}\mu\Big(Z_{l+1} \neq Z_l\Big)$$

$$= \frac{1}{2}\sum_{l=0}^{L-1}2^{-(l-2)p}\sum_{F \in \mathcal{P}_l}\sum_{\substack{C:C \subset F \\ C \in \mathcal{P}_{l+1}}}\Big|Q(C) - \frac{Q(F)}{P(F)}P(C)\Big|.$$

Therefore,

$$W_p^p(P,Q) \leq 2^{p-1}\Big(W_p^p(P,P_L) + W_p^p(P_L,Q)\Big)$$

$$\leq 2^{p-1}\Big(\frac{1}{2}\sum_{l=0}^{L-1}2^{-(l-2)p}\sum_{F\in\mathcal{P}_l}\sum_{\substack{C:C\subset F\\C\in\mathcal{P}_{l+1}}}\Big|Q(C) - \frac{Q(F)}{P(F)}P(C)\Big| + 2^{-p(L-1)}\Big)$$

Since

$$\Big|Q(C) - \frac{Q(F)}{P(F)}P(C)\Big| = \frac{1}{P(F)}\Big|Q(C)P(F) - Q(F)P(C)\Big|$$

$$\leq \frac{1}{P(F)}\Big[P(F)|Q(C) - P(C)| + P(C)|P(F) - Q(F)|\Big]$$

$$= \frac{P(C)}{P(F)}|P(F) - Q(F)| + |P(C) - Q(C)|,$$

we have

$$\frac{W_p^p(P,Q)}{2^{p-1}}$$

$$\leq \frac{1}{2}\sum_{l=0}^{L-1}2^{-(l-2)p}\sum_{F\in\mathcal{P}_l}\sum_{\substack{C:C\subset F\\C\in\mathcal{P}_{l+1}}}\Big[\frac{P(C)}{P(F)}|P(F) - Q(F)| + |P(C) - Q(C)|\Big]$$

$$+2^{-p(L-1)}$$

$$= \frac{1}{2}\sum_{l=0}^{L-1}2^{-(l-2)p}\Big[\sum_{F\in\mathcal{P}_l}|P(F) - Q(F)| + \sum_{F\in\mathcal{P}_{l+1}}|P(F) - Q(F)|\Big] + 2^{-p(L-1)}$$

$$= \frac{1}{2}\sum_{l=1}^{L-1}2^{-(l-2)p}\sum_{F\in\mathcal{P}_l}|P(F) - Q(F)|$$

$$+\frac{1}{2}\sum_{l=1}^{L}2^{-(l-3)p}\sum_{F\in\mathcal{P}_l}|P(F) - Q(F)| + 2^{-p(L-1)}$$

$$\leq \frac{1+2^p}{2}\sum_{l=1}^{L}2^{-(l-2)p}\sum_{F\in\mathcal{P}_l}|P(F) - Q(F)| + 2^{-p(L-1)},$$

where the second equality holds because $\sum_{F\in\mathcal{P}_0}|P(F) - Q(F)| = 0$. □

Lemma 8.5. *For two probability measures P and Q on $\mathbb{R}$,*

$$W_p^p(P,Q) \leq \sum_{m\geq 0}2^{mp}\Big[2^{p-1}|P(B_m) - Q(B_m)|$$

$$+ \Big(P(B_m)\wedge Q(B_m)\Big)W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q)\Big].$$

(8.1)

Proof. The proof is explicitly given in [18] (pp. 714–715). □

8.2. Proof of Theorem 2.1

Since the KL condition (1.3) holds, the posterior distribution is consistent with respect to the Lévy-Prokhorov metric d_P, see Theorem 6.25 of [22]. Therefore, there exists a real sequence $\epsilon_{1n} \downarrow 0$ such that

$$\Pi\Big(d_P(P, P_0) > \epsilon_{1n} \mid X_1, \ldots, X_n\Big) \longrightarrow 0 \quad \text{in probability.}$$

To see this, let $N_0 = 1$, and for every $m \geq 1$, choose $N_m > N_{m-1}$ such that

$$\mathrm{E}\left[\Pi\Big(d_P(P, P_0) > \frac{1}{m+1} \mid X_1, \ldots, X_n\Big)\right] \leq \frac{1}{m+1} \quad \text{for every } n \geq N_m.$$

Define $\epsilon_{1n} = (m+1)^{-1}$ if $N_m \leq n < N_{m+1}$. Then, $\epsilon_{1n} \to 0$ and for $N_m \leq n < N_{m+1}$, we have

$$\mathrm{E}\left[\Pi\Big(d_P(P, P_0) > \epsilon_{1n} \mid X_1, \ldots, X_n\Big)\right] \leq \frac{1}{m+1} \longrightarrow 0$$

as $n \to \infty$.

Now, suppose that (2.1) holds. Then, in a similar way, we can construct a sequence $\epsilon_{2n} \downarrow 0$ such that

$$\Pi\Big(|M_p(P) - M_p(P_0)| > \epsilon_{2n} \mid X_1, \ldots, X_n\Big) \longrightarrow 0 \quad \text{in probability.}$$

Let

$$\mathcal{F}_n = \Big\{P \in \mathcal{F} : d_P(P, P_0) \leq \epsilon_{1n}, \ |M_p(P) - M_p(P_0)| \leq \epsilon_{2n}\Big\}$$

and $P_n \in \mathcal{F}_n$ such that

$$W_p(P_n, P_0) \geq \sup_{P \in \mathcal{F}_n} W_p(P, P_0) - \epsilon_{1n}.$$

Note that (P_n) is a non-random sequence of probability measures such that $d_P(P_n, P_0) \to 0$ and $M_p(P_n) \to M_p(P_0)$. It follows that $W_p(P_n, P_0) \to 0$. Since $\Pi(\mathcal{F}_n \mid X_1, \ldots, X_n) \to 1$ in probability, we conclude that (2.2) holds.

Conversely, suppose that (2.2) holds. Then, similarly as before, we can construct a sequence $\epsilon_{3n} \downarrow 0$ such that

$$\mathrm{E}\left[\Pi\Big(W_p(P, P_0) > \epsilon_{3n} \mid X_1, \ldots, X_n\Big)\right] \longrightarrow 0.$$

Let

$$\mathcal{F}'_n = \Big\{P \in \mathcal{F} : W_p(P, P_0) \leq \epsilon_{3n}\Big\}.$$

and $P'_n \in \mathcal{F}'_n$ such that

$$|M_p(P'_n) - M_p(P_0)| \geq \sup_{P \in \mathcal{F}'_n} |M_p(P) - M_p(P_0)| - \epsilon_{3n}.$$

Again, (P'_n) is a non-random sequence with $W_p(P'_n, P_0) \to 0$, so we have $|M_p(P'_n) - M_p(P_0)| \to 0$. Since $\Pi(\mathcal{F}'_n \mid X_1, \ldots, X_n) \to 1$ in probability, we conclude that (2.1) holds. $\square$

8.3. Proof of Theorem 2.2

We first provide a simple proof relying on a stronger moment condition than the one in the statement of Theorem 2.2. For this, we assume that $M_{2p}(P_0) \leq K$ and

$$\Pi\Big(P:\ M_{2p}(P) \leq K \mid X_1, \ldots, X_n\Big) \to 1 \quad \text{in probability.}$$

In view of the characterization of posterior consistency in the Wasserstein distance of Theorem 2.1, we will establish that $\Pi\Big(\mathcal{C}_j \mid X_1, \ldots, X_n\Big) \to 0$ in probability for $j = 1, 2$, where $\mathcal{C}_1$ and $\mathcal{C}_2$ are the following two convex sets

$$\mathcal{C}_1 = \{P:\ M_p(P) - M_p(P_0) > \epsilon, \quad M_{2p}(P) \leq K\}$$
$$\mathcal{C}_2 = \{P:\ M_p(P_0) - M_p(P) > \epsilon, \quad M_{2p}(P) \leq K\}$$

To this aim, it suffices to show that $\inf_{P \in \mathcal{C}_1} H(P_0, P) > 0$ and $\inf_{P \in \mathcal{C}_2} H(P_0, P) > 0$. For $P \in \mathcal{C}_1 \cup \mathcal{C}_2$, we have

$$|M_p(P) - M_p(P_0)|^2 = \left| \int |x|^p \left(\sqrt{p(x)/p_0(x)} + 1 \right) \left(1 - \sqrt{p(x)/p_0(x)} \right) p_0(x) dx \right|^2$$

$$\leq H^2(P_0, P) \int |x|^{2p} \left(1 + \sqrt{p(x)/p_0(x)} \right)^2 p_0(x) dx$$

by the Cauchy–Schwarz inequality. The integral of the right term is itself upper bounded by

$$K + 2 \int |x|^{2p} \sqrt{p(x)\, p_0(x)} dx + K \leq 4K$$

by virtue of $\sqrt{p\, p_0} \leq \frac{1}{2}(p + p_0)$. Hence, we get $H(P_0, P) \geq \epsilon/(2\sqrt{K})$ for $P \in \mathcal{C}_1 \cup \mathcal{C}_2$. $\qquad\square$

Now, we will prove Theorem 2.2 without the moment condition of order $2p$.

Lemma 8.6. *For positive constants ϵ, δ and K assume that*

$$P(B_m) + Q(B_m) \leq K 2^{-(p+\delta)m} \quad \text{for } m \geq 0, \tag{8.2}$$

and

$$\left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right| \leq \epsilon \quad \text{for } m \leq M, F \in \mathcal{P}_l, l \leq L, \tag{8.3}$$

where M and L are positive integers. Then,

$$W_p^p(P, Q) \leq K' \left[2^{-\delta M} + 2^{-Lp} + 2^{Mp} L\epsilon \right],$$

where K' is a constant depending only on δ, K and p.

Proof. Since $W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q) \le 2^p$ and (8.2) holds, the summation in the right hand side of (8.1) over $m > M$ is bounded by $c_1 2^{-\delta M}$, where c_1 is a constant depending only on δ, K and p. Therefore, $W_p^p(P, Q)$ is bounded by

$$\sum_{m=0}^{M} 2^{mp} \left[2^{p-1} |P(B_m) - Q(B_m)| \right.$$
$$\left. + \left(P(B_m) \wedge Q(B_m)\right) W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q) \right] + c_1 2^{-\delta M}$$

by Lemma 8.5. Note that

$$\left| \mathcal{R}_{B_m}P(F) - \mathcal{R}_{B_m}Q(F) \right| = \left| \frac{P(\pi_m^{-1}(F) \cap B_m)}{P(B_m)} - \frac{Q(\pi_m^{-1}(F) \cap B_m)}{Q(B_m)} \right|$$
$$= \frac{|P(\pi_m^{-1}(F) \cap B_m)Q(B_m) - P(B_m)Q(\pi_m^{-1}(F) \cap B_m)|}{P(B_m)Q(B_m)}$$
$$\le \frac{1}{P(B_m)Q(B_m)} \left[P(\pi_m^{-1}(F) \cap B_m) \Big| P(B_m) - Q(B_m) \Big| \right.$$
$$\left. + P(B_m) \Big| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \Big| \right].$$

By Lemma 8.3 and the last display, we have

$$\left(P(B_m) \wedge Q(B_m)\right) W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q)$$
$$\le \kappa_p \left(P(B_m) \wedge Q(B_m)\right)$$
$$\quad \times \left[\sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} |\mathcal{R}_{B_m}P(F) - \mathcal{R}_{B_m}Q(F)| + 2^{-Lp} \right]$$
$$\le \kappa_p \left[\sum_{l=1}^{L} 2^{-lp} \left\{ \Big| P(B_m) - Q(B_m) \Big| \right.\right.$$
$$\left. + \sum_{F \in \mathcal{P}_l} \Big| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \Big| \right\} \quad (8.4)$$
$$\left. + 2^{-Lp} \left(P(B_m) \wedge Q(B_m)\right) \right]$$
$$\le \kappa_p \left[\Big| P(B_m) - Q(B_m) \Big| + 2^{-Lp} \left(P(B_m) \wedge Q(B_m)\right) \right.$$
$$\left. + \sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} \Big| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \Big| \right].$$

It follows that

$$W_p^p(P, Q)$$

$$\leq c_1 2^{-\delta M} + \sum_{m=0}^{M} 2^{mp} \left\{ (2^{p-1} + \kappa_p) \Big| P(B_m) - Q(B_m) \Big| \right.$$

$$\left. + \kappa_p 2^{-Lp} \Big(P(B_m) \wedge Q(B_m) \Big) \right\}$$

$$+ \kappa_p \sum_{m=0}^{M} 2^{mp} \sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} \Big| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \Big|$$

$$\leq c_1 2^{-\delta M} + \epsilon(2^{p-1} + \kappa_p) \sum_{m=0}^{M} 2^{mp} + K\kappa_p 2^{-Lp} \sum_{m=0}^{M} 2^{-\delta m}$$

$$+ \kappa_p \epsilon \sum_{m=0}^{M} 2^{mp} \sum_{l=1}^{L} 2^{-(p-1)l}$$

$$\leq c_2 \Big(2^{-\delta M} + 2^{Mp}\epsilon + 2^{-Lp} + 2^{Mp} L\epsilon \Big),$$

where the second inequality holds by (8.2), (8.3) and that the cardinality of $\mathcal{P}_l$ is 2^l. Here, c_2 is a constant depending only on δ, K and p. □

By (5.4), we have that

$$P_0(B_m) \leq 2^{p+\delta} K 2^{-(p+\delta)m} \quad \text{for } m \geq 0.$$

and $\Pi(\mathcal{F}_0 \mid X_1, \ldots, X_n) \to 1$ in probability, where

$$\mathcal{F}_0 = \Big\{ P : P(B_m) \leq 2^{p+\delta} K 2^{-(p+\delta)m} \text{ for all } m \geq 0 \Big\}.$$

Suppose that a sufficiently small $\epsilon > 0$ is given. We will prove that for some function $g : (0, \infty) \to (0, \infty)$, with $g(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$,

$$\Pi \Big(W_p^p(P, P_0) \geq g(\epsilon) \ \Big| \ X_1, \ldots, X_n \Big) \longrightarrow 0 \quad \text{in probability.} \tag{8.5}$$

Let M and L be the largest integer less than or equal to $(\log_2 \epsilon^{-1})/(2p)$ and $\log_2 \epsilon^{-1}$, respectively. Then,

$$2^{-\delta M} + 2^{-Lp} + 2^{Mp} L\epsilon \leq 2^\delta \epsilon^{\delta/(2p)} + 2^p \epsilon^p + \sqrt{\epsilon} \log_2 \epsilon^{-1}. \tag{8.6}$$

Let

$$\mathcal{F}_{m,F,+} = \Big\{ P : P(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) < \epsilon \Big\}$$

$$\mathcal{F}_{m,F,-} = \Big\{ P : P_0(\pi_m^{-1}(F) \cap B_m) - P(\pi_m^{-1}(F) \cap B_m) < \epsilon \Big\}.$$

Then, by Lemma 8.6 and (8.6), there exists a constant c_1, depending only on δ, K and p, such that

$$P \in \mathcal{F}_0 \quad \text{and} \quad P \in \bigcap_{m \leq M} \bigcap_{l \leq L} \bigcap_{F \in \mathcal{P}_l} (\mathcal{F}_{m,F,+} \cap \mathcal{F}_{m,F,-}) \equiv \mathcal{F}_\epsilon$$

implies that

$$W_p^p(P, P_0) \leq c_1 \left(2^\delta \epsilon^{\delta/(2p)} + 2^p \epsilon^p + \sqrt{\epsilon} \log_2 \epsilon^{-1} \right) \equiv g(\epsilon).$$

Certainly, $g(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$. Since $\Pi(\mathcal{F}_0^c \mid X_1, \ldots, X_n) \to 0$ in probability and the KL condition (1.3) holds, by Schwartz's theorem (see Theorem 6.25 of [22] if Π depends on n), it is sufficient for (8.5) to construct a sequence ϕ_n of tests such that

$$P_0\phi_n + \sup_{P \in \mathcal{F}_\epsilon^c} P(1 - \phi_n) \leq e^{-cn} \tag{8.7}$$

for some constant $c > 0$ and every large enough n.

Let

$$\phi_{m,F,+} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) > \epsilon/2 \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{m,F,-} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) < -\epsilon/2 \\ 0 & \text{otherwise.} \end{cases}$$

Then, by the Hoeffding's inequality,

$$(P_0\phi_{m,F,+}) \vee (P_0\phi_{m,F,-}) \leq e^{-n\epsilon^2/2}.$$

Also, for $P \in \mathcal{F}_{m,F,+}^c$,

$$P(1 - \phi_{m,F,+}) \leq P\left(\mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P(\pi_m^{-1}(F) \cap B_m) \leq -\epsilon/2 \right) \leq e^{-n\epsilon^2/2}$$

by the Hoeffding's inequality. Similarly, for $P \in \mathcal{F}_{m,F,-}^c$,

$$P(1 - \phi_{m,F,-}) \leq e^{-n\epsilon^2/2}.$$

Therefore, if we define

$$\phi_n = \max_{m \leq M} \max_{l \leq L} \max_{F \in \mathcal{P}_l} (\phi_{m,F,+} \vee \phi_{m,F,-}),$$

then,

$$P_0\phi_n \leq \sum_{m \leq M} \sum_{l \leq L} \sum_{F \in \mathcal{P}_l} P_0(\phi_{m,F,+} + \phi_{m,F,-}) \leq 2^{L+1}(L+1)(M+1)e^{-n\epsilon^2/2}.$$

Since L, M and ϵ does not depend on n, ϕ_n satisfies (8.7) for some $c > 0$ and large enough n, which completes the proof. □

8.4. Proof of Theorem 3.1

For a given sequence δ_n, let

$$\mathcal{C}_{n,1} = \{p : M_p(P) - M_p(P_0) > \delta_n, M_{2p}(P) \leq K\}$$

$$\mathcal{C}_{n,2} = \{p : M_p(P_0) - M_p(P) > \delta_n, M_{2p}(P) \le K\}$$

and $\mathcal{C}_n = \mathcal{C}_{n,1} \cup \mathcal{C}_{n,2}$. Then, it can be shown that

$$\inf_{p \in \mathcal{C}_n} H^2(p, p_0) \ge \frac{\delta_n^2}{4K}$$

as in the first proof of Theorem 2.2. For any measurable set $\mathcal{C}$, let $\Pi_n^{\mathcal{C}}$ be the posterior distribution restricted and renormalized onto $\mathcal{C}$, that is,

$$\Pi_n^{\mathcal{C}}(A) = \frac{1}{L_n(\mathcal{C})} \int_A \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p) \quad \text{for all measurable } A \subset \mathcal{C}$$

and let $\bar{p}_n^{\mathcal{C}}(x) = \int_{\mathcal{C}} p(x) d\Pi_n^{\mathcal{C}}(p)$, where

$$L_n(\mathcal{C}) = \int_{\mathcal{C}} \prod_{i=1}^n \frac{p}{p_0}(X_i) d\Pi(p).$$

Since

$$\frac{L_n(\mathcal{C})}{L_{n-1}(\mathcal{C})} = \frac{\bar{p}_{n-1}^{\mathcal{C}}}{p_0}(X_n),$$

we have

$$\mathrm{E}\left[L_n^{1/2}(\mathcal{C}) \mid \mathcal{G}_{n-1}\right] = L_{n-1}^{1/2}(\mathcal{C})\left(1 - \tfrac{1}{2}H^2(p_0, \bar{p}_{n-1}^{\mathcal{C}})\right),$$

where $\mathcal{G}_{n-1}$ is the σ-algebra generated by $X_1, \dots, X_{n-1}$. Since $\mathcal{C}_{n,j}$ is convex, we have $H^2(p_0, \bar{p}_{n-1}^{\mathcal{C}_{n,j}}) \ge \delta_n^2/(4K)$ for $j = 1, 2$. Therefore,

$$\mathrm{E}L_n^{1/2}(\mathcal{C}_{n,j}) \le \left(1 - \frac{\delta_n^2}{8K}\right)^n \le e^{-c_1 n \delta_n^2}$$

for all large enough n, where $c_1 > 0$ is a constant depending only on K. It follows that $L_n(\mathcal{C}_{n,j})$ is upper bounded by $e^{-c_2 n \delta_n^2}$ with probability tending to 1 for some constant c_2. Thus, if we take $\delta_n = K' \epsilon_n$ for large enough K', the proof is complete.

8.5. Proof of Theorem 3.2

Lemma 8.7. *For positive constants α, δ, ϵ and K, suppose that*

$$P(B_m) + Q(B_m) \le K 2^{-(2p+\delta)m} \quad \text{for } m \ge 0, \tag{8.8}$$

and

$$\left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right| \le \frac{K(2+\alpha)^{-mp}\epsilon}{(l+1)^2} \tag{8.9}$$

$$\text{for } m \le M, \ F \in \mathcal{P}_l, \ l \le L.$$

Then,

$$W_p^p(P,Q) \leq K' \left[2^{-Lp} + \epsilon + 2^{-(p+\delta)M} \right],$$

where K' is a constant depending only on α, K and p. If $p > 1$, condition (8.9) can be replaced by a slightly weaker condition that

$$\left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right| \leq K(2+\alpha)^{-mp}\epsilon \tag{8.10}$$
$$\text{for } m \leq M, \ F \in \mathcal{P}_l, \ l \leq L.$$

Proof. By (8.8) and that $W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q) \leq 2^p$, the summation in the right hand side of (8.1) over $m > M$ is bounded by a constant multiple of $2^{-(p+\delta)M}$. Since $B_0 \in \mathcal{P}_0$,

$$\sum_{m=0}^{M} 2^{mp} \big| P(B_m) - Q(B_m) \big|$$
$$= \sum_{m=0}^{M} 2^{mp} \big| P(\pi_m^{-1}(B_0) \cap B_m) - Q(P(\pi_m^{-1}(B_0) \cap B_m) \big|$$
$$\leq K\epsilon \sum_{m=0}^{M} \left(\frac{2}{2+\alpha} \right)^{mp} = K\epsilon \frac{1 - (1+\alpha/2)^{-(M+1)p}}{1 - (1+\alpha/2)^{-p}},$$

where the inequality holds by (8.9) with $l = 0$. Therefore,

$$W_p^p(P,Q) \leq \sum_{m=0}^{M} 2^{mp} \Big(P(B_m) \wedge Q(B_m) \Big) W_p^p(\mathcal{R}_{B_m}P, \mathcal{R}_{B_m}Q)$$
$$+ K' \Big(\epsilon + 2^{-(p+\delta)M} \Big)$$

by Lemma 8.5, where K' is a constant depending only on α, K and p. By (8.4), the summation in the last display is bounded by

$$\kappa_p \sum_{m=0}^{M} 2^{mp} \left[\left| P(B_m) - Q(B_m) \right| + 2^{-Lp} \Big(P(B_m) \wedge Q(B_m) \Big) \right.$$
$$\left. + \sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} \left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right| \right] \tag{8.11}$$

Since the cardinality of $\mathcal{P}_l$ is 2^l and $\sum_{l=1}^{\infty}(l+1)^{-2} < \infty$, the first assertion follows from (8.8) and (8.9).

If $p > 1$ and (8.9) is replaced by (8.10), we have

$$\sum_{l=1}^{L} 2^{-lp} \sum_{F \in \mathcal{P}_l} \left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right|$$

$$\leq \sum_{l=1}^{L} K 2^{-l(p-1)} (2+\alpha)^{-mp} \epsilon \leq \frac{K}{2^{p-1}-1} (2+\alpha)^{-mp} \epsilon.$$

Therefore, we have the same conclusion with a different constant K'. $\qquad\square$

Lemma 8.8. *If $X_1, \ldots, X_n \overset{\mathrm{iid}}{\sim} P$, then*

$$P(\mathbb{P}_n(B) \leq P(B) - \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2P(B)}\right)$$

$$P(\mathbb{P}_n(B) \geq P(B) + \epsilon) \leq \exp\left(-\frac{n\epsilon^2}{2\{P(B)+\epsilon/3\}}\right)$$

for every $n \geq 1$ and $\epsilon \geq 0$.

Proof. See Theorem 1 of [28]. $\qquad\square$

Before proving Theorem 3.2, we state and prove a similar one. Theorem 8.9 below is devised for eliminating the logarithmic term $\log \epsilon_n^{-1}$ in Theorem 3.2. Proofs of Theorem 3.2 and 8.9 are quite similar, so we do not provide all details to avoid the repetition. We provide a detailed proof only for Theorem 8.9 because this contains the most technical part caused by the factors $(l+1)^{-2}$ and $(l+1)^{-4}$. These factors appear for handling the last term in (8.11). If $p > 1$, we need not consider these factors by the second assertion of Lemma 8.7. For $p = 1$, we can avoid the technical factors $(l+1)^{-2}$ and $(l+1)^{-4}$, with an additional logarithmic factor in the rate. If we want to eliminate the term $\log \epsilon_n^{-1}$, the statement would become more complicated as Theorem 8.9. For conciseness we decided to include Theorem 3.2 in the main texts rather than Theorem 8.9.

Theorem 8.9. *Assume that the prior Π satisfies the KL condition (1.4) for a sequence ϵ_n with $\epsilon_n \downarrow 0$ and $\epsilon_n \geq \sqrt{(\log n)/n}$. Furthermore, assume that there exist positive constants K and δ such that*

$$P_0(\pi_m^{-1}(F) \cap B_m) \leq \frac{K}{(l+1)^4} 2^{-(2p+\delta)m} \quad \text{for } m \geq 0, F \in \mathcal{P}_l, l \geq 0$$

and $\Pi(\mathcal{F}_n^c \mid X_1, \ldots, X_n) \to 0$ in probability, where

$$\mathcal{F}_n = \bigcap_{m \geq 0} \bigcap_{l \leq L} \bigcap_{F \in \mathcal{P}_l} \left\{ P : P(\pi_m^{-1}(F) \cap B_m) \leq \frac{K}{(l+1)^4} 2^{-(2p+\delta)m} \right\}$$

and L is the largest integer less than or equal to $(\log_2 \epsilon_n^{-1})/p$. Then, for some constant $K' > 0$,

$$\Pi\left(W_p^p(P, P_0) \geq K'\epsilon_n \mid X_1, \ldots, X_n \right) \longrightarrow 0 \quad \text{in probability.} \tag{8.12}$$

Proof. Let M be the largest integer less than or equal to $(p+\delta)^{-1} \log_2 \epsilon_n^{-1}$. Let $\alpha > 0$ be a sufficiently small constant such that $(1 + \alpha/2)^{2p} < 2^\delta$. For $m \le M$ and $F \in \mathcal{P}_l$ with $l \le L$, let

$$\mathcal{F}_{m,F,+} = \left\{ P \in \mathcal{F}_n : P(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) \right.$$

$$\left. \le \frac{2K_1}{(l+1)^2}(2+\alpha)^{-mp}\epsilon_n \right\}$$

$$\mathcal{F}_{m,F,-} = \left\{ P \in \mathcal{F}_n : P_0(\pi_m^{-1}(F) \cap B_m) - P(\pi_m^{-1}(F) \cap B_m) \right.$$

$$\left. \le \frac{2K_1}{(l+1)^2}(2+\alpha)^{-mp}\epsilon_n \right\},$$

where $K_1 > 0$ is a large constant described below. Then, by Lemma 8.7,

$$P \in \bigcap_{m \le M} \bigcap_{l \le L} \bigcap_{F \in \mathcal{P}_l} (\mathcal{F}_{m,F,+} \cap \mathcal{F}_{m,F,-}) \equiv \mathcal{F}_n'$$

implies that $W_p^p(P, P_0) \le K_2 \epsilon_n$ for some constant K_2. Since $\Pi(\mathcal{K}_n) \ge e^{-n\epsilon_n^2}$, by Lemma 8.1, it is sufficient for (8.12) to construct a sequence ϕ_n of tests such that

$$P_0 \phi_n \to 0 \quad \text{and} \quad \sup_{P \in (\mathcal{F}_n')^c} P(1 - \phi_n) \le e^{-3n\epsilon_n^2} \tag{8.13}$$

for every large enough n.

For $m \le M$ and $F \in \mathcal{P}_l$ with $l \le L$, let

$$\phi_{m,F,+} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) > \frac{K_1}{(l+1)^2}(2+\alpha)^{-mp}\epsilon_n \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{m,F,-} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) < \frac{-K_1}{(l+1)^2}(2+\alpha)^{-mp}\epsilon_n \\ 0 & \text{otherwise.} \end{cases}$$

Then, by Lemma 8.8,

$$P_0 \phi_{m,F,+} \le \exp\left[\frac{-K_1^2(l+1)^{-4}(2+\alpha)^{-2mp}n\epsilon_n^2}{2\{P_0(\pi_m^{-1}(F) \cap B_m) + K_1(l+1)^{-2}(2+\alpha)^{-mp}\epsilon_n/3\}}\right].$$

If $P_0(\pi_m^{-1}(F) \cap B_m) > K_1(l+1)^{-2}(2+\alpha)^{-mp}\epsilon_n/3$, then

$$P_0 \phi_{m,F,+} \le \exp\left[-\frac{K_1^2(l+1)^{-4}(2+\alpha)^{-2mp}n\epsilon_n^2}{4P_0(\pi_m^{-1}(F) \cap B_m)}\right]$$

$$\le \exp\left[-\frac{K_1^2(2+\alpha)^{-2mp}n\epsilon_n^2}{4K2^{-(2p+\delta)m}}\right] = \exp\left[-\frac{K_1^2\beta^m n\epsilon_n^2}{4K}\right],$$

where $\beta = 2^\delta(1+\alpha/2)^{-2p} > 1$. Otherwise,

$$P_0 \phi_{m,F,+} \le \exp\left[-\frac{3}{4}K_1(l+1)^{-2}(2+\alpha)^{-mp}n\epsilon_n\right]$$

$$\leq \exp\left[-\frac{3}{4}K_1 n\left(\frac{1}{p}\log_2 \epsilon_n^{-1} + 1\right)^{-2}\epsilon_n^{1+\frac{p\log_2(2+\alpha)}{p+\delta}}\right] \leq e^{-K_1 n\epsilon_n^2}$$

for large enough n, where the second inequality holds because $m \leq M \leq (p+\delta)^{-1}\log_2 \epsilon_n^{-1}$ and $l \leq L \leq (\log_2 \epsilon_n^{-1})/p$, and the third inequality holds because

$$\frac{p\log_2(2+\alpha)}{p+\delta} = \frac{\log_2(2+\alpha)^p}{\log_2 2^{p+\delta}} \quad \text{and} \quad \frac{(2+\alpha)^p}{2^{p+\delta}} = \frac{(1+\alpha/2)^p}{2^\delta} < 1.$$

Since $\epsilon_n \geq \sqrt{(\log n)/n}$, we have

$$\sum_{m\leq M}\sum_{l\leq L}\sum_{F\in\mathcal{P}_l} P_0\phi_{m,F,+} \leq ML2^L \max_{m\leq M}\max_{l\leq L}\max_{F\in\mathcal{P}_l} P_0\phi_{m,F,+}$$

$$\leq \frac{1}{p(p+\delta)}\left(\log_2 \epsilon_n^{-1}\right)^2\left(\frac{1}{\epsilon_n}\right)^{1/p} \max_{m\leq M}\max_{l\leq L}\max_{F\in\mathcal{P}_l} P_0\phi_{m,F,+} \longrightarrow 0$$

as $n \to \infty$ provided that K_1 is large enough. Also, if K_1 is sufficiently large, for $P \in \mathcal{F}_{m,F,+}^c$ with $F \in \mathcal{P}_l$,

$$P(1 - \phi_{m,F,+})$$
$$= P\left(\mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P(\pi_m^{-1}(F) \cap B_m) \leq -\frac{K_1}{(l+1)^2}(2+\alpha)^{-mp}\epsilon_n\right)$$
$$\leq \exp\left[-\frac{K_1^2(l+1)^{-4}(2+\alpha)^{-2mp}n\epsilon_n^2}{2P(\pi_m^{-1}(F) \cap B_m)}\right] \leq \exp\left[-\frac{K_1^2\beta^m}{2K}n\epsilon_n^2\right] \leq e^{-3n\epsilon_n^2}$$

for large enough n, where the first inequality holds by Lemma 8.8. A similar inequalities for $P_0\phi_{m,F,-}$ and $P(1 - \phi_{m,F,-})$ can also be obtained. Therefore, if we define

$$\phi_n = \max_{m\leq M}\max_{l\leq L}\max_{F\in\mathcal{P}_l}(\phi_{m,F,+} \vee \phi_{m,F,-})$$

and K_1 is sufficiently large, then ϕ_n satisfies (8.13) for all large enough n. This completes the proof. □

Proof of Theorem 3.2 for $p > 1$. We first claim that if

$$P_0(B_m) \leq K2^{-(2p+\delta)m} \quad \text{for } m \geq 0$$

and $\Pi(\mathcal{F}_0^c \mid X_1, \ldots, X_n) \to 0$ in probability, where

$$\mathcal{F}_0 = \bigcap_{m\geq 0}\left\{P : P(B_m) \leq K2^{-(2p+\delta)m}\right\},$$

then (8.12) holds for some constant K'. The proof of this claim is the same to that of Theorem 8.9 if we replace $\mathcal{F}_n$ by $\mathcal{F}_0$ and eliminate the factors $(l+1)^{-2}$ and $(l+1)^{-4}$ in all equations, which is possible due to the second assertion of Lemma 8.7.

Once we adjust the constant K, two conditions of the claim is satisfied by (5.4). Hence the proof is complete. □

Proof of Theorem 3.2 for $p = 1$. If (8.8) and (8.10) hold with $p = 1$, then it holds that

$$W_1(P, Q) \leq K' \left[2^{-L} + L\epsilon + 2^{-(1+\delta)M} \right]. \tag{8.14}$$

This can be proved as in Lemma 8.7. The only difference is that the last term in (8.11) is bounded as

$$\kappa_p \sum_{m=0}^{M} 2^m \sum_{l=1}^{L} 2^{-l} \sum_{F \in \mathcal{P}_l} \left| P(\pi_m^{-1}(F) \cap B_m) - Q(\pi_m^{-1}(F) \cap B_m) \right|$$

$$\leq K\kappa_p \epsilon \sum_{m=0}^{M} \sum_{l=1}^{L} \left(\frac{2}{2+\alpha} \right)^m \leq K' L\epsilon,$$

where K' is a constant depending only on α, K and p.

As in the proof of Theorem 3.2, we next claim that if

$$P_0(B_m) \leq K 2^{-(2+\delta)m} \quad \text{for } m \geq 0$$

and $\Pi(\mathcal{F}_0^c \mid X_1, \ldots, X_n) \to 0$ in probability, where

$$\mathcal{F}_0 = \bigcap_{m \geq 0} \left\{ P : P(B_m) \leq K 2^{-(2+\delta)m} \right\},$$

then

$$\Pi \left(W_p^p(P, P_0) \geq K' \epsilon_n \log \epsilon_n^{-1} \mid X_1, \ldots, X_n \right) \longrightarrow 0 \quad \text{in probability}$$

for some constant K'. To prove this, define L, M and α as in Theorem 8.9 with $p = 1$. Also, for $m \leq M$ and $F \in \mathcal{P}_l$ with $l \leq L$, let

$$\mathcal{F}_{m,F,+} = \left\{ P \in \mathcal{F}_0 : P(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) \leq 2K_1(2+\alpha)^{-m} \epsilon_n \right\}$$

$$\mathcal{F}_{m,F,-} = \left\{ P \in \mathcal{F}_0 : P_0(\pi_m^{-1}(F) \cap B_m) - P(\pi_m^{-1}(F) \cap B_m) \leq 2K_1(2+\alpha)^{-m} \epsilon_n \right\},$$

where $K_1 > 0$ is a large constant as in the proof of Theorem 8.9. Then, by (8.14),

$$P \in \bigcap_{m \leq M} \bigcap_{l \leq L} \bigcap_{F \in \mathcal{P}_l} (\mathcal{F}_{m,F,+} \cap \mathcal{F}_{m,F,-}) \equiv \mathcal{F}_n'$$

implies that $W_p^p(P, P_0) \leq K_2 \epsilon_n \log \epsilon_n^{-1}$ for some constant K_2. Once we change the definition of ϕ_n as

$$\phi_n = \max_{m \leq M} \max_{l \leq L} \max_{F \in \mathcal{P}_l} (\phi_{m,F,+} \vee \phi_{m,F,-}),$$

where

$$\phi_{m,F,+} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) > K_1(2+\alpha)^{-m}\epsilon_n \\ 0 & \text{otherwise} \end{cases}$$

$$\phi_{m,F,-} = \begin{cases} 1 & \text{if } \mathbb{P}_n(\pi_m^{-1}(F) \cap B_m) - P_0(\pi_m^{-1}(F) \cap B_m) < -K_1(2+\alpha)^{-m}\epsilon_n \\ 0 & \text{otherwise,} \end{cases}$$

the remaining proof of the claim is the same to that of Theorem 8.9.

Once we adjust the constant K, two conditions of the claim is satisfied by (5.4). Hence the proof is complete. □

8.6. Proof of Theorem 4.1

Let $\mathcal{F}_\epsilon = \{P \in \mathcal{F}_0 : W_\infty(P, P_0) \le \epsilon\}$. We will show that for every small enough $\epsilon \ge K_1\sqrt{(\log n)/n}$ and $n \ge n_0$, there exists a test ϕ such that

$$P_0\phi \le e^{-K_2 n\epsilon^2} \quad \text{and} \quad \sup_{P \in \mathcal{F}_{2\epsilon}^c} P(1-\phi) \le e^{-K_2 n\epsilon^2}, \tag{8.15}$$

where K_1, K_2 and n_0 are constants depending only on c_0. Since $\Pi(\mathcal{K}_n) \ge e^{-n\epsilon_n^2}$, (8.15) and Lemma 8.1 guarantees (4.1) for large enough constant $K > 0$.

Let $\epsilon > 0$ be given. Let N be the smallest integer greater than or equal to ϵ^{-1}. Let $I_j = [(j-1)\epsilon, j\epsilon)$ for $j = 1, \ldots, N-1$ and $I_N = [(N-1)\epsilon, 1]$. Let $I_{jk} = \cup_{l=j}^{j+k-1} I_l$ for $j = 1, \ldots, N$ and $k = 1, \ldots, N-j+1$. Let $\mathscr{I}$ and $\mathscr{B}$ be the collections of every interval I_{jk} and every finite union of I_{jk}, respectively. Note that the cardinalities of $\mathscr{I}$ and $\mathscr{B}$ are $N(N+1)/2$ and $2^N - 1$, respectively.

We first claim that for $P \in \mathcal{F}_0$,

$$P(I_{jk}) - P_0(I_{jk}) \le \frac{c_0\epsilon}{2} \quad \forall j, k \text{ implies that } W_\infty(P_0, P) \le 2\epsilon. \tag{8.16}$$

If B is either $[0,1]$ or $[0,(N-1)\epsilon)$, it is obvious that $P(B) \le P_0(B^\epsilon)$. Also, for $B = I_{jk}$ for some (j,k), with $B \ne [0,1]$ and $B \ne [0,(N-1)\epsilon)$,

$$P(B) \le P_0(B) + \frac{c_0\epsilon}{2} < P_0(B^\epsilon).$$

Thus, $P(B) \le P_0(B^\epsilon)$ for every $B \in \mathscr{I}$. For $B \in \mathscr{B} - \mathscr{I}$, we have $B = \cup_{l=1}^L B_l$ for some $L \ge 2$, where $B_l \in \mathscr{I}$ and B_l's are ϵ-separated. Thus,

$$P_0(B^\epsilon) \ge \sum_{l=1}^L P_0(B_l) + (L-1)c_0\epsilon.$$

It follows by (8.16) that

$$P(B) = \sum_{l=1}^L P(B_l) \le \sum_{l=1}^L P_0(B_l) + \frac{Lc_0\epsilon}{2} \le \sum_{l=1}^L P_0(B_l) + (L-1)c_0\epsilon \le P_0(B^\epsilon).$$

Thus, we have $P(B) \le P_0(B^\epsilon)$ for every $B \in \mathscr{B}$. Next, for any Borel subset A of $[0,1]$, let $J = \{j : A \cap I_j \ne \emptyset\}$ and $C = \cup_{j \in J} I_j$. Then, we have $C \in \mathscr{B}$ and $A \subset C \subset A^\epsilon$. Therefore,

$$P(A) \le P(C) \le P_0(C^\epsilon) \le P_0(A^{2\epsilon}).$$

This proves (8.16).

By (8.16), Then, $\mathcal{F}_{2\epsilon} \supset \cap_{j,k} \mathcal{F}_{jk}$, where

$$\mathcal{F}_{jk} = \{P : P(I_{jk}) - P_0(I_{jk}) \le c_0\epsilon/2\}.$$

Define test functions ϕ_{jk} as $\phi_{jk} = 1$ if

$$\mathbb{P}_n(I_{jk}) > P_0(I_{jk}) + \frac{c_0\epsilon}{4}$$

and $\phi_{jk} = 0$ otherwise. Then, for every $P \in \mathcal{F}_{jk}^c$, we have

$$\begin{aligned}
P(1 - \phi_{jk}) &= P\left\{\mathbb{P}_n(I_{jk}) \le P_0(I_{jk}) + \frac{c_0\epsilon}{4}\right\} \\
&= P\left\{\mathbb{P}_n(I_{jk}) \le P(I_{jk}) + P_0(I_{jk}) - P(I_{jk}) + \frac{c_0\epsilon}{4}\right\} \\
&\le P\left\{\mathbb{P}_n(I_{jk}) < P(I_{jk}) - \frac{c_0\epsilon}{4}\right\} \le \exp\left[-\frac{nc_0^2\epsilon^2}{8}\right],
\end{aligned}$$

where the last inequality holds by the Hoeffding's inequality. Let $\phi = \max_{j,k} \phi_{jk}$. Applying the Hoeffding's inequality again, we have

$$\begin{aligned}
P_0\phi \le \sum_{j,k} P_0\phi_{jk} &\le \frac{N(N+1)}{2} \exp\left[-\frac{nc_0^2\epsilon^2}{8}\right] \\
&\le \exp\left[2\log(\epsilon^{-1}+1) - \frac{nc_0^2\epsilon^2}{8}\right].
\end{aligned} \tag{8.17}$$

Therefore, we can choose constants $K_1, K_2 > 0$ and n_0 such that if

$$\epsilon \ge K_1\sqrt{(\log n)/n},$$

then the right hand side of (8.17) is bounded by $\exp(-K_2 n\epsilon^2)$ for every $n \ge n_0$. This completes the proof of (8.15). $\square$

8.7. Proof of Theorem 5.1

Lemma 8.10. *There exist universal constants $c_1, c_2 > 0$ such that*

$$\frac{c_1}{\epsilon} \le \Gamma(\epsilon) \le \frac{c_2}{\epsilon}$$

for every $\epsilon \in (0,1]$.

Proof. Note that

$$\Gamma(\epsilon) = \int_0^\infty x^{\epsilon-1} e^{-x} dx = \int_0^1 x^{\epsilon-1} e^{-x} dx + \int_1^\infty x^{\epsilon-1} e^{-x} dx.$$

It is easy to show that there exist constants a_1 and a_2 such that

$$a_1 \le \int_1^\infty x^{\epsilon-1} e^{-x} dx \le a_2$$

for every $\epsilon \in (0,1]$. The assertion follows because $e^{-1} \le e^{-x} \le 1$ for every $x \in (0,1]$ and $\int_0^1 x^{\epsilon-1} dx = \epsilon^{-1}$. □

Lemma 8.11. *Suppose that $X \sim \text{Beta}(\alpha\epsilon, \alpha(1-\epsilon))$, $\alpha\epsilon \le 1$ and $\alpha(1-\epsilon) \ge 1$. Then,*

$$P(X > t) \le C_\alpha(1 - t^{\alpha\epsilon}),$$

where C_α is a constant depending only on α.

Proof. Let p be the pdf of X, that is,

$$p(x) = \frac{\Gamma(\alpha)}{\Gamma(\alpha\epsilon)\Gamma(\alpha(1-\epsilon))} x^{\alpha\epsilon-1}(1-x)^{\alpha(1-\epsilon)-1}.$$

By Lemma 8.10,

$$P(X > t) = \int_t^1 p(x) dx \le c_\alpha \epsilon \int_t^1 x^{\alpha\epsilon-1} dx = \frac{c_\alpha}{\alpha}(1 - t^{\alpha\epsilon}),$$

where c_α is a constant depending only on α. □

Lemma 8.12. *Suppose that $\Pi(\mathcal{K}_n) \ge e^{-n\epsilon_n^2}$ and $\epsilon_n \ge \sqrt{(\log n)/n}$. Then, there exists a universal constant $K > 0$ such that*

$$\Pi\Big(\big|P(B_m) - P_0(B_m)\big| \le K\epsilon_n \ \forall m \le C \log \epsilon_n^{-1} \mid X_1, \ldots, X_n\Big) \to 1 \quad \text{in probability}$$

for every $C > 0$.

Proof. Let $C > 0$ be given. For eacn n and $m \le C \log \epsilon_n^{-1}$, let

$$\psi_m = I\Big(\big|\mathbb{P}_n(B_m) - P_0(B_m)\big| > K\epsilon_n/2\Big),$$

where K is a universal constant described below. Using the Hoeffding's inequality, it is not difficult to prove that

$$P_0\psi_m \lesssim e^{-K^2 n\epsilon_n^2/2} \quad \text{and} \quad \sup_{P \in \mathcal{G}_m^c} P(1 - \psi_n) \lesssim e^{-K^2 n\epsilon_n^2/2},$$

where $\mathcal{G}_m = \{P : |P(B_m) - P_0(B_m)| \le K\epsilon_n\}$. Let

$$\phi_n = \max_{m \le C \log \epsilon_n^{-1}} \psi_m.$$

Then, we have

$$P_0\phi_n \lesssim C \log \epsilon_n^{-1} e^{-K^2 n\epsilon_n^2/2} \to 0 \quad \text{and} \quad \sup_{P \in \mathcal{F}_n^c} P(1 - \phi_n) \lesssim e^{-K^2 n\epsilon_n^2/2},$$

where

$$\mathcal{F}_n = \bigcap_{m \leq C \log \epsilon_n^{-1}} \mathcal{G}_m.$$

Thus, the proof is complete by Lemma 8.1 provided that $K^2/2 \geq 3$. □

Lemma 8.13. *Let ϵ_n be a sequence such that $\epsilon_n \to 0$ and $\epsilon_n \geq \sqrt{\log n/n}$. Let H be the normal distribution with mean μ_H and variance σ_H^2. For a Dirichlet process mixture prior (5.2) with $\alpha > 1$ and $\sigma = \sigma_n \to 0$, suppose that $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$. Also, for some $p \in [1, \infty)$, assume that $P_0(B_m) \leq K2^{-pm}$ for every $m \geq 0$, and that $\epsilon_n \leq An^{-p/(2+2p)}$ for every n, where K and A are constants. Then,*

$$\Pi\Big(P(B_m) \leq K'2^{-pm} \ \forall m \geq 0 \mid X_1, \ldots, X_n \Big) \to 1 \quad \text{in probability,}$$

where K' is a large enough constant.

Proof. Let $\widetilde{\epsilon}_n = L\epsilon_n$, and define $\widetilde{\mathcal{K}}_n$ as $\mathcal{K}_n$ after replacing ϵ_n by $\widetilde{\epsilon}_n$, where L is a large constant described below. Then, $\Pi(\widetilde{\mathcal{K}}_n) \geq \Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$. Note that

$$G(B) \sim \text{Beta}\Big(\alpha H(B), \alpha\big(1 - H(B)\big)\Big)$$

for any Borel set B. Also, $\alpha H(\overline{B}_m) \leq 1$ and $\alpha(1 - H(\overline{B}_m)) \geq 1$ for every large enough m, where

$$\overline{B}_m = (-\infty, -2^{m-1}] \cup (2^{m-1}, \infty).$$

Thus, by Lemma 8.11, for any $K' \geq 2^p$,

$$\begin{aligned} \Pi\big(G(B_m) > K'2^{-pm}\big) &\leq \Pi\big(G(\overline{B}_m) > K'2^{-pm}\big) \\ &\leq \Pi\big(G(\overline{B}_m) > 2^{-pm}\big) \leq C_\alpha\big(1 - 2^{-pm\alpha H(\overline{B}_m)}\big) \end{aligned} \tag{8.18}$$

for every large enough m, where C_α is a constant depending only on α. Note that $1 - \Phi_\sigma(x) \leq e^{-x^2/(2\sigma^2)}/2$, where Φ_σ is the cdf of ϕ_σ, so we have

$$H(\overline{B}_m) \leq \frac{1}{2}\left[\exp\left\{-\frac{1}{2}\left(\frac{2^{m-1} - \mu_H}{\sigma_H}\right)^2\right\} + \exp\left\{-\frac{1}{2}\left(\frac{2^{m-1} + \mu_H}{\sigma_H}\right)^2\right\}\right].$$

Since $1 - e^{-x} \leq x$, the right hand side of (8.18) is bounded by

$$\frac{\log 2}{2} p\alpha C_\alpha m \left[\exp\left\{-\frac{1}{2}\left(\frac{2^{m-1} - \mu_H}{\sigma_H}\right)^2\right\} + \exp\left\{-\frac{1}{2}\left(\frac{2^{m-1} + \mu_H}{\sigma_H}\right)^2\right\}\right]$$

$$\lesssim p\alpha C_\alpha \exp\{-C_{\sigma_H} 2^{2m}\}$$

for every large enough m, where C_{σ_H} is a constant depending only on σ_H.

Note that P is the convolution of G and $N(0, \sigma_n^2)$. If $Y_1 = Y_2 + Y_3$, where Y_2 and Y_3 are independent random variables following G and $N(0, \sigma_n^2)$, respectively, then

$$P(\overline{B}_m) = \Pr(|Y_1| > 2^{m-1}) \leq \Pr(|Y_2| > 2^{m-2}) + \Pr(|Y_3| > 2^{m-2})$$
$$\leq G(\overline{B}_{m-1}) + 2\big(1 - \Phi_{\sigma_n}(2^{m-2})\big).$$

Hence,

$$P(B_m) \leq P(\overline{B}_m) \leq G(\overline{B}_{m-1}) + 2\big(1 - \Phi_{\sigma_n}(2^{m-2})\big)$$
$$\leq G(\overline{B}_{m-1}) + \exp\left(-\frac{2^{2(m-2)}}{2\sigma_n^2}\right) \leq G(\overline{B}_{m-1}) + 2^{-pm},$$

where the last inequality holds for every large enough m. It follows for any constants $C > 0$ and large enough n that

$$\Pi\big(P(B_m) > (K'+1)2^{-pm} \text{ for some } m \geq C\log\widetilde{\epsilon}_n^{-1}\big)$$
$$\leq \Pi\big(G(\overline{B}_{m-1}) > K'2^{-pm} \text{ for some } m \geq C\log\widetilde{\epsilon}_n^{-1}\big)$$
$$= \Pi\big(G(\overline{B}_m) > K'2^{-p(m+1)} \text{ for some } m \geq C\log\widetilde{\epsilon}_n^{-1} - 1\big)$$
$$\leq \sum_{m \geq C\log\widetilde{\epsilon}_n^{-1} - 1} \Pi\big(G(\overline{B}_m) > 2^{-pm}\big)$$
$$\lesssim p\alpha C_\alpha \sum_{m \geq C\log\widetilde{\epsilon}_n^{-1} - 1} \exp\left(-C_{\sigma_H} 2^{2m}\right) \lesssim p\alpha C_\alpha \exp\left(-\frac{C_{\sigma_H}}{4}\widetilde{\epsilon}_n^{-2C\log 2}\right).$$

If we take $C = (p\log 2)^{-1}$, the right hand side of the last display is bounded by

$$p\alpha C_\alpha \exp\left(-\frac{C_{\sigma_H}}{4}\widetilde{\epsilon}_n^{-2/p}\right).$$

Since $\epsilon_n \leq An^{-p/(2+2p)}$, $n\epsilon_n^2$ is bounded by a constant multiple of $\epsilon_n^{-2/p}$ for every n. Hence, if L is large enough, we have that

$$\Pi\Big(P(B_m) > (K'+1)2^{-pm} \text{ for some } m \geq C\log\widetilde{\epsilon}_n^{-1} \mid X_1, \ldots, X_n\Big) \to 0$$

in probability by Lemma 8.2.

Note that by Lemma 8.12,

$$\Pi\Big(\big|P(B_m) - P_0(B_m)\big| \leq K''\widetilde{\epsilon}_n \ \forall m \leq C\log\widetilde{\epsilon}_n^{-1} \mid X_1, \ldots, X_n\Big) \to 1$$

in probability, where K'' is a constant. Since

$$P_0(B_m) + K''\widetilde{\epsilon}_n \leq (K + K'')2^{-pm}$$

for every $m \leq C\log\widetilde{\epsilon}_n^{-1}$, the proof is complete. $\qquad \square$

It is shown in the proof of Theorem 2 in [21] that $\Pi(\mathcal{K}_n) \geq e^{-n\epsilon_n^2}$ with $\epsilon_n = cn^{-2/5}(\log n)^2$ for some constant $c > 0$. For any $p < 4$, note that $\epsilon_n \leq n^{-p/(2+2p)}$ for all large enough n. Hence, the proof of Theorem 5.1 is complete by (5.4) and Lemma 8.13. □

8.8. Proof of Theorem 7.1

Denote $p_G(x) = \int k_\sigma(x - z)dG(z)$. We use the result of [36]. It is shown in the proof of Theorem 2 in [36] that

$$
\begin{aligned}
&W_2^2(G, G_0) \\
&\leq C \inf_{\delta \in (0,1)} \left[\delta^2 + \|p_G - p_{G_0}\|_1^{2(s-2)/(1+2s)} \left\{ \int_{-1/\delta}^{1/\delta} \tilde{k}(t)^{-2}dt \right\}^{(s-2)/(1+2s)} \right]
\end{aligned}
$$

for any $s > 2$ and G with $M_2(G) < \infty$, where C is a constant depending only on s. Note that Theorem 2 of [36] assumed that G and G_0 are discrete probability measures with bounded supports, but finiteness of the second moment suffices as discussed therein. The right hand side of the last display tends to zero as $\|p_G - p_{G_0}\|_1 \to 0$. It follows that for every $\epsilon > 0$,

$$
\Pi\Big(W_2^2(G, G_0) > \epsilon \mid X_1, \ldots, X_n \Big) \to 0.
$$

in probability. Since $W_2^2(P_G, P_{G_0}) \leq W_2^2(G, G_0)$, the proof is complete. □

Acknowledgments

References

[1] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning* 214–223.

[2] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[3] BERNTON, E., JACOB, P. E., GERBER, M. and ROBERT, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 235–269. MR3928142

[4] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory* **54** 781–790. MR2444554

[5] Biau, G., Cadre, B., Sangnier, M. and Tanielian, U. (2018). Some theoretical properties of GANs. *arXiv:1803.07819.* MR4124334

[6] Bickel, P. and Kleijn, B. (2012). The semiparametric Bernstein–von Mises theorem. *Ann. Statist.* **40** 206–237. MR3013185

[7] Bobkov, S. and Ledoux, M. (2019). *One-dimensional empirical measures, order statistics, and Kantorovich transport distances* **261**. Memoirs of the American Mathematical Society. MR4028181

[8] Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473

[9] Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597

[10] Chae, M., Kim, Y. and Kleijn, B. (2019). The semi-parametric Bernstein–von Mises theorem for regression models with symmetric errors. *Statist. Sinica* **29**. MR3932526

[11] Chae, M. and Walker, S. G. (2017). A novel approach to Bayesian consistency. *Electron. J. Stat.* **11** 4723–4745. MR3729657

[12] Chae, M. and Walker, S. G. (2019). Bayesian consistency for a nonparametric stationary Markov model. *Bernoulli* **25** 877–901. MR3920360

[13] Chae, M. and Walker, S. G. (2020). Wasserstein upper bounds of the total variation for smooth densities. *Statist. Probab. Lett.* **163** 1–6. MR4083852

[14] De Acosta, A. (1982). Invariance principles in probability for triangular arrays of *B*-valued random vectors and some applications. *Ann. Probab.* **10** 346–373. MR0647509

[15] Dereich, S., Scheutzow, M. and Schottstedt, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 1183–1203. MR3127919

[16] Doss, H. and Sellke, T. (1982). The tails of probabilities chosen from a Dirichlet prior. *Ann. Statist.* **10** 1302–1305. MR0673666

[17] Dudley, R. M. (1989). *Real Analysis and Probability.* Chapman and Hall/CRC. MR0982264

[18] Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341

[19] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105

[20] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[21] Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[22] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonpara-*

metric Bayesian Inference. Cambridge University Press. MR3587782

[23] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

[24] GIVENS, C. R. and SHORTT, R. M. (1984). A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal* **31** 231–240. MR0752258

[25] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V. and COURVILLE, A. C. (2017). Improved training of Wasserstein GANs. In *Proc. Neural Information Processing Systems* 5767–5777.

[26] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[27] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985

[28] JANSON, S. (2016). Large deviation inequalities for sums of indicator variables. *arXiv:1609.00533*.

[29] KALLI, M., GRIFFIN, J. E. and WALKER, S. G. (2011). Slice sampling mixture models. *Stat. Comput.* **21** 93–105. MR2746606

[30] KANTOROVICH, L. V. and RUBINSTEIN, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ.* **13** 52–59. MR0102006

[31] KUHN, D., ESFAHANI, P. M., NGUYEN, V. A. and SHAFIEEZADEH-ABADEH, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics* 130–166. INFORMS.

[32] LALOË, T. (2010). L_1-Quantization and clustering in Banach spaces. *Math. Methods Statist.* **19** 136–150. MR2723146

[33] LI, C., LIN, L. and DUNSON, D. B. (2019). On posterior consistency of tail index for Bayesian kernel mixture models. *Bernoulli* **25** 1999–2028. MR3961238

[34] LIANG, T. (2018). On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv:1811.03179*.

[35] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

[36] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422

[37] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[38] RUDOLF, D. and SCHWEIZER, N. (2018). Perturbation theory for Markov chains via Wasserstein distance. *Bernoulli* **24** 2610–2639. MR3779696

[39] SCHWARTZ, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **4** 10–26. MR0184378

[40] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[41] SINGH, S., UPPAL, A., LI, B., LI, C.-L., ZAHEER, M. and PÓCZOS, B.

(2018). Nonparametric density estimation with adversarial losses. In *Proc. Neural Information Processing Systems* 10246–10257. Curran Associates Inc.

[42] VERSHIK, A. M. (2013). Long history of the Monge–Kantorovich transportation problem. *Math. Intelligencer* **35** 1–9. MR3133757

[43] VILLANI, C. (2003). *Topics in Optimal Transportation*. Amer. Math. Soc. MR1964483

[44] VILLANI, C. (2008). *Optimal Transport: Old and New*. Springer Science & Business Media. MR2459454

[45] WALKER, S. G. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32** 2028–2043. MR2102501

[46] WEED, J. and BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560

[47] WEED, J. and BERTHET, Q. (2019). Estimation of smooth densities in Wasserstein distance. **99** 3118–3119.

[48] YOO, W. W., RIVOIRARD, V. and ROUSSEAU, J. (2017). Adaptive supremum norm posterior contraction: Wavelet spike-and-slab and anisotropic Besov spaces. *arXiv:1708.01909*.