Minimax Posterior Contraction Rates for Unconstrained Distribution Estimation on $[0,1]^d$ under Wasserstein Distance

Anonymous authors
Paper under double-blind review

Abstract

We obtain asymptotic minimax optimal posterior contraction rates for estimation of probability distributions on $[0,1]^d$ under the Wasserstein-v metrics using Bayesian Histograms. To the best of our knowledge, our analysis is the first to provide minimax optimal posterior contraction rates under the Wasserstein-v metrics for every problem dimension $d \geq 1$. Our proof technique takes advantage of the conjugacy of the Bayesian Histogram.

1 Introduction

The Wasserstein metric is a popular tool for comparing two distributions μ and ν defined on a common metric space $(E^d, \|\cdot - \cdot\|_2)$ where $E \subseteq \mathbb{R}$. For $1 \leq v < \infty$, the Wasserstein distance W_v is defined as

$$W_v(\mu_1, \mu_2) := \left(\inf_{\pi \in \mathcal{M}(\mu_1,\mu_2)} \int \|x-y\|_2^v \, d\pi(x,y) \right)^{1/v}, \tag{1}$$

where $\mathcal{M}(\mu_1, \mu_2)$ is the set of couplings of μ_1 and μ_2; specifically the joint probability measures on $E \times E$ with marginals μ_1 and μ_2 respectively. Some benefits of using the Wasserstein metric include its sensitivity to distance in the underlying space, ability to compare distributions regardless of continuity level, and its 1-dimension equivalent representation as the L^v distance between quantile functions, which facilitates quantile function inference (Zhang et al., 2020).

In this paper we study the problem of non–parametrically estimating a distribution P_0 on E^d (where $E = [0,1]$) under the Wasserstein metric from n independent and identically distributed (i.i.d) random variables $Y_1, \ldots, Y_n$ drawn from P_0. Our focus is on the unconstrained problem; that is, we place no additional assumptions on P_0. From the viewpoint of analyzing only frequentist estimators, this is a well studied problem. The frequentist convergence rates of the empirical measure under the expected Wasserstein distance are studied in (Fournier & Guillin, 2015; Singh & Póczos, 2018; Bobkov & Ledoux, 2019; Weed & Bach, 2019) to varying degrees of generality. A consequence of the work of (Singh & Póczos, 2018) is that on the metric space $([0,1]^d, \|\cdot - \cdot\|_2)$, for $d \in \mathbb{N}$, for the class of Borel probability measures, the empirical measure is minimax optimal (at least up to logarithmic terms) for every $v \geq 1$. Further, the minimax rate is lower bounded by $n^{-1/2v}$ for $d \leq 2v$, and $n^{-1/d}$ for $d > 2v$.

Far less has been done in providing frequentist guarantees for Bayesian statistical procedures when the inferential goal is to estimate a non-parametric distribution underneath a Wasserstein distance. In a non-parametric Bayesian model aimed at inferring a probability distribution on E^d, for each sample size n, a prior Π_{0n} is placed on the space of Borel probability measures on E^d. We denote this space $\mathcal{P}_d(E)$. The sample size n posterior distribution, which we denote $\Pi_n(\cdot|Y_1, Y_2, \ldots, Y_n)$, is a regular conditional distribution over $\mathcal{P}_d(E)$ induced from the likelihood and the prior Π_{0n}. Given a distance function $\tilde{d}$ between probability measures on E (e.g Kullback-Leibler, Hellinger, Wasserstein–1, Total Variation, etc.) we say the sequence of posterior distributions contracts at the rate ϵ_n under P_0 if $\Pi_n(P \in \mathcal{P}_d(E) : \tilde{d}(P_0, P) \geq M_n \epsilon_n)$ converges in probability to 0 as $n \to \infty$ for *any* arbitrarily slowly increasing sequence M_n when $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} P_0$. If for every $P_0 \in \mathcal{P}_d(E)$ the Posterior Contraction Rate (PCR) ϵ_n is achieved, and ϵ_n is the frequentist minimax rate over $\mathcal{P}_d(E)$, we say the Bayesian method is agnostic to the prior choice in the presence of an infinite amount of data under class $\mathcal{P}_d(E)$.

Ghosal et al. (2000) provides a general three condition strategy for proving these PCRs, but their approach is more difficult to use when $\tilde{d}$ is a Wasserstein metric. Challenges include $W_v, v \geq 2$ not being dominated by Total Variation or Hellinger distances, causing the need for explicit test construction. Also, the Kullback-Leibler neighborhood condition, which ensures such neighborhoods of P_0 have sufficient prior mass, may make it more difficult to achieve the minimax rate under $W_v, v \geq 1$ because depending on the model under consideration, approximation of distributions under the Kullback-Leibler divergence may not be achievable at the square of minimax rate under W_v [1]. In light of these challenges, there have been far fewer theoretical advances in proving minimax optimal PCRs for distribution estimation under $W_v, v \geq 1$ than under Total Variation and Hellinger distances. Chae et al. (2021) successfully derives posterior contraction rates under Wasserstein distance when $E = \mathbb{R}$, but their results are restricted to dimension $d = 1$. Camerlenghi et al. (2022) develops a framework to study Wasserstein PCRs for models where the posterior distribution is not available through Bayes formula. But the only model in which they apply their framework to derive PCRs for each $d \in \mathbb{N}, v \geq 1$ is the one placing a Dirichlet process prior on the data generating distribution. In addition, the PCR derived for $P_0 \in \mathcal{P}_d([0,1])$ is $\gtrsim n^{-\frac{1}{2}\frac{1}{(d+v)}}$ which via the discussion earlier in this section is slower decaying than the minimax rate by a polynomial factor for every $d \in \mathbb{N}, v \geq 1$.Wasserstein distance PCRs for estimation of the mixing distribution in a convolved data generating distribution where the noise distribution is known are derived in Rousseau & Scricciolo (2023),Gao & van der Vaart (2016), and Scricciolo (2018) but these papers do not focus on directly estimating the data generating distribution under Wasserstein distance.

1.1 Contributions

Our main contribution is Theorem 1. In it we obtain PCRs for every dimension $d \geq 1$ and for every distance $W_v, v \geq 1$ and the PCRs achieved are minimax optimal at least up to logarithmic terms. To the best of our knowledge, our result is the first to provide a minimax optimal PCR across each $(d \in \mathbb{N}, v \geq 1)$ setting for estimating an unconstrained $P_0 \in \mathcal{P}_d([0,1])$. These rates are achieved using a Bayesian Histogram model that partitions $[0,1]^d$ into b_n^d equal area squares where $b_n := 2^{\lceil \log_2(k_n) \rceil}$ for a sequence k_n growing as a function of the sample size n at the appropriate rate, uses the Multinomial likelihood to weight the constant density within each square, and places a sample size dependent Dirichlet prior distribution on the weight vector with prior concentration vector $\boldsymbol{\alpha}_{b_n}$ (of dimension b_n^d). This model induces a sequence of posterior distributions $\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ over $\mathcal{P}_d([0,1])$. In Theorem 1, we show that

$$\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}(P \in \mathcal{P}_d([0,1]) : W_v(P_0, P) \geq \epsilon_n(d,v)) \overset{\text{i.p } P_0}{\to} 0$$

provided that

1. If $d \leq 2v$ then $k_n = n^{\frac{1}{2v}}$, $\sum_{\boldsymbol{j} \in 2^{d\lceil \log_2(k_n) \rceil}} \boldsymbol{\alpha}_{\boldsymbol{j},b_n} \lesssim n^{\frac{1}{2}}$, $\epsilon_n \asymp n^{-\frac{1}{2v}}$

2. If $d > 2v$, then $k_n = n^{\frac{1}{d}}$, $\sum_{\boldsymbol{j} \in 2^{d\lceil \log_2(k_n) \rceil}} \boldsymbol{\alpha}_{\boldsymbol{j},b_n} \lesssim n^{1-\frac{v}{d}}$, $\epsilon_n \asymp n^{-\frac{1}{d}}$

where log terms in ϵ_n which are specified in the theorem are ignored here. In the problem of providing optimal PCRs for estimating distributions on $[0,1]^d$ under Wasserstein-v distance, our results close the gap between the minimax rates for this problem and the Wasserstein PCRs provided by Camerlenghi et al. (2022).

The remainder of this paper is organized as follows. In Section 2 we formally introduce the Bayesian Histogram model. In Section 3 we state the main theorem and the two fundamental lemmas upon which the main theorem depends. We then prove the main theorem. In Section 4 we provide the proofs of the lemmas and in section 5 we provide concluding remarks.

[1] Chae et al. (2021) (p.3644) already encounters Kullback-Liebler condition limitations when only estimating distributions on $\mathbb{R}$

2 Bayesian Histogram

2.1 General Notations

Since we always consider probability distributions on $[0,1)^d$, we drop the E notation of the introduction and denote

$$\mathcal{P}_d := \{\text{Borel Probability Measures on } [0,1)^d\}$$

Excluding the right end points are a notational convenience but extension of the arguments that follow to include the right endpoint is trivial.

For $b, d \in \mathbb{N}$, we denote $[b] := \{1, 2, \ldots, b\}$ and $[b]^d := \prod_{j=1}^{d}[b]$. For $B \subseteq \mathbb{R}^d$, $\mathcal{B}(B)$ denotes the Borel measurable subsets of B. For $j \in \mathbb{N}$, $\mathcal{S}^{j-1}$ refers to the $(j-1)$ dimensional probability simplex. That is $\mathcal{S}^{j-1} := \{(x_1, \ldots, x_j) \in \mathbb{R}^j : \sum_{t=1}^{j} x_t = 1, x_t \geq 0 \text{ for } t \in [j]\}$. Also note that $\mathbb{R}_+ := \{x \in \mathbb{R} : x > 0\}$ and for $z \in \mathbb{N}$ and $\boldsymbol{\alpha} \in \mathbb{R}_+^z$, the Dirichlet probability measure Dirichlet : $\mathcal{B}(\mathcal{S}^{z-1}) \to [0,1]$ is given by

$$\text{Dirichlet}(G|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \int_G \prod_{i=1}^{z} x_i^{\alpha_i - 1} \mathrm{d}\boldsymbol{x}, \tag{2}$$

where $B(\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_z)) := \frac{\prod_{j=1}^{z} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{z} \alpha_j)}$ is the z dimensional Beta function and $\Gamma(x)$ denotes the Gamma function evaluated at x and $G \in \mathcal{B}(\mathcal{S}^{z-1})$. For $b \in \mathbb{N}$ and a multi-index $\boldsymbol{i} = (i_1, i_2, \ldots, i_d) \in [b]^d$, define

$$A_{\boldsymbol{i},b} := \left[\frac{i_1 - 1}{b}, \frac{i_1}{b}\right) \times \left[\frac{i_2 - 1}{b}, \frac{i_2}{b}\right) \times \ldots, \times \left[\frac{i_d - 1}{b}, \frac{i_d}{b}\right). \tag{3}$$

Clearly, $\{A_{\boldsymbol{i},b}\}_{\boldsymbol{i} \in [b]^d}$ form a partition of $[0,1)^d$. For a vector of weights $\boldsymbol{\pi} = \{\pi_{\boldsymbol{j}}\}_{\boldsymbol{j} \in [b]^d} \in \mathcal{S}^{bd-1}$, the d dimensional Histogram probability measure Histogram : $\mathcal{B}([0,1)^d) \to [0,1]$ is a weighted mixture of uniform distributions on the partition sets $A_{\boldsymbol{i},b}$, defined by

$$\text{Histogram}(G|\boldsymbol{\pi}, b) := \int_G \sum_{\boldsymbol{i} \in [b]^d} b^d \pi_{\boldsymbol{i}} \mathbb{I}(\boldsymbol{y} \in A_{\boldsymbol{i},b}) \mathrm{d}\boldsymbol{y}, \tag{4}$$

where $G \in \mathcal{B}([0,1)^d)$.

2.2 Bayesian Histogram Model Definition

We suppose $Y_1, Y_2, \ldots, Y_n, \ldots \overset{iid}{\sim} P_0$ where $P_0 \in \mathcal{P}_d$. For $b \in \mathbb{N}$, let $\boldsymbol{\alpha}_b := \{\alpha_{\boldsymbol{j},b}\}_{\boldsymbol{j} \in [b]^d} \in \mathbb{R}_+^{bd}$. For an increasing sequence k_n, let $b_n := 2^{K_n}$, where $K_n := \lceil \log_2(k_n) \rceil$, $\boldsymbol{\pi}_n := \{\pi_{n,\boldsymbol{j}}\}_{\boldsymbol{j} \in [b_n]^d} \in \mathcal{S}^{b_n d - 1}$. For $n \in \mathbb{N}$, the Bayesian Histogram model likelihood and prior are given by

$$Y_1, \ldots, Y_n | \boldsymbol{\pi}_n \overset{i.i.d}{\sim} \text{Histogram}(\cdot|\boldsymbol{\pi}_n, b_n), \qquad \boldsymbol{\pi}_n | \boldsymbol{\alpha}_{b_n} \sim \text{Dirichlet}(\cdot|\boldsymbol{\alpha}_{b_n}). \tag{5}$$

Also, let $z_n^*(\cdot|Y_1, \ldots, Y_n)$ refer to the posterior probability measure over $\mathcal{S}^{b_n d - 1}$ derived from equation 5. As $\alpha_{\boldsymbol{i},b_n} > 0$ for every $\boldsymbol{i} \in [b_n]^d$ and for every $n \in \mathbb{N}$, equation 5 induces a sequence of posterior distributions over $\mathcal{P}_d$. Specifically let $\psi_b : \mathcal{S}^{bd-1} \to \mathcal{P}_d$ be the map that takes a given $\boldsymbol{\pi} = \{\pi_{\boldsymbol{j}}\}_{\boldsymbol{j} \in [b]^d}$ and produces its corresponding Histogram probability measure. That is

$$\psi_b(\boldsymbol{\pi}) = \text{Histogram}(\cdot|\boldsymbol{\pi}, b). \tag{6}$$

For a measurable set $B \subseteq \mathcal{P}_d$, the posterior measure $\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ is

$$\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}(B|Y_1, \ldots, Y_n) = z_n^*(\psi_{b_n}^{-1}(B)|Y_1, \ldots, Y_n). \tag{7}$$

Due to conjugacy, it is straightforward to show that

$$z_n^*(\cdot|Y_1, \ldots, Y_n) = \text{Dirichlet}(\cdot|\boldsymbol{\alpha}_{\boldsymbol{b_n}}^{\boldsymbol{*}}),$$

where for $\boldsymbol{i} \in [b_n]^d$

$$\alpha^*_{\boldsymbol{i},b_n} = \alpha_{\boldsymbol{i},b_n} + \sum_{j=1}^{n} \mathbb{I}(Y_j \in A_{\boldsymbol{i},b_n}). \tag{8}$$

Now allowing $\boldsymbol{\alpha}_{b_n} \in \{x \in \mathbb{R} : x \geq 0\}^{b_n d}$, we define the sequence of estimators for P_0, denoted $\bar{P}_n$, by

$$\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}} := \psi_{b_n}\left\{\left(\frac{\alpha^*_{\boldsymbol{i},b_n}}{\sum_{\boldsymbol{j}\in[b_n]^d}\alpha^*_{\boldsymbol{j},b_n}}\right)_{\boldsymbol{i}\in[b_n]^d}\right\} = \psi_{b_n}\{(E_{z_n^*}(\pi_{\boldsymbol{i}}|Y_1,\ldots,Y_n))_{\boldsymbol{i}\in[b_n]^d}\}, \tag{9}$$

where the second equality above holds if $\alpha_{\boldsymbol{i},b_n} > 0$ for $\boldsymbol{i} \in [b_n]^d$.

We note that posterior distributions derived from improper prior distributions are not considered in this work, and therefore to consider the posterior measure sequence Π_n we require that $\alpha_{\boldsymbol{i},b_n} > 0$ for $\boldsymbol{i} \in [b_n]^d$. However, we allow $\bar{P}_n$ to be defined regardless of whether or not the prior distribution over the simplex is proper. In particular, it is still defined in the event that some or all of the $\alpha_{\boldsymbol{i},b_n}$ parameters are zero. When the prior distribution is proper, $\bar{P}_n$ has an additional interpretation: it is the posterior mean Histogram. In the lemmas and theorems that follow that involve analysis of the posterior distribution sequence Π_n, we make clear that we require $\alpha_{\boldsymbol{i},b_n} > 0$ for $\boldsymbol{i} \in [b_n]^d$ and $n \in \mathbb{N}$.

$\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ (and $\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}$) are indexed by the choice of k_n (which determines the total number of bins) and $\boldsymbol{\alpha}_{b_n}$, which gives the prior concentrations on those bins. In the subsequent subsection we establish constraints on k_n and $\boldsymbol{\alpha}_{b_n}$ that ensure $\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ and $\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ are minimax statistical procedures.

3 Posterior Contraction Results

Our results utilize the following two assumptions for $d \in \mathbb{N}$ and $v \geq 1$.

Assumption 1. *For $n \in \mathbb{N}$*

$$k_n = \begin{cases} n^{1/2v} & d \leq 2v, \\ n^{1/d} & d > 2v, \end{cases}$$

and

Assumption 2.

$$\sum_{\boldsymbol{j}\in[b_n]^d} \alpha_{\boldsymbol{j},b_n} \lesssim \begin{cases} n^{1/2} & d \leq 2v. \\ n^{1-\frac{v}{d}} & d > 2v. \end{cases}$$

Our main PCR result is the following theorem.

Theorem 1. *Let $Y_1,\ldots,Y_n \overset{iid}{\sim} P_0 \in \mathcal{P}_d$. Suppose $\gamma > 1$ and k_n satisfies Assumption 1 and*

$$\epsilon_n(d,v) := C_0(d,v)\begin{cases} n^{-\frac{1}{2v}}\log^{\frac{\gamma}{v}}(n) & d < 2v, \\ n^{-\frac{1}{2v}}\log^{\frac{1+\gamma}{v}}(n) & d = 2v, \\ n^{-\frac{1}{d}}\log^{\frac{\gamma}{v}}(n) & d > 2v, \end{cases} \tag{10}$$

Now assuming that for each $n \in \mathbb{N}$ and $\boldsymbol{j} \in [b_n]^d$, $\alpha_{\boldsymbol{j},b_n} > 0$, and that $\boldsymbol{\alpha}_{b_n}$ satisfies Assumption 2, we have that for $1 \leq v < \infty$ and $d \in \mathbb{N}$ and $C_0(d,v)$ sufficiently large,

$$\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}(P \in \mathcal{P}_d : W_v(P_0,P) \geq \epsilon_n(d,v)) \overset{i.p\ P_0}{\to} 0,$$

where $b_n = 2^{\lceil \log_2(k_n)\rceil}$.

According to Singh & Póczos (2018),

$$\inf_{\tilde{P}} \sup_{P_0\in\mathcal{P}_d} \mathbb{E}_{P_0} W_v(\tilde{P},P_0) \gtrsim \begin{cases} n^{-\frac{1}{2v}} & d \leq 2v, \\ n^{-\frac{1}{d}} & d > 2v, \end{cases} \tag{11}$$

where the inf is taken over all estimators $\tilde{P}$ from n observations. Thus the PCRs of Theorem 1 are up to logarithmic terms attaining the minimax rates. The assumption on the prior concentrations, Assumption 2, is flexible enough to support a vague prior. Specifically, the mean of a Dirichlet distribution with common concentration on all categories is a discrete uniform distribution, so the practitioner wishing to encode vagueness by asserting that under the prior on average all bin probabilities are equal will want to set all prior bin concentrations to a common value. When $d \leq 2v$, by assumption 1, the number of bins is $\asymp n^{\frac{d}{2v}}$, thus Assumption 2 is satisfied when each concentration is set to $Cn^{-(\frac{d}{2v}-\frac{1}{2})}$ for some $C > 0$. Likewise when $d > 2v$, by assumption 1, there are $\asymp n$ bins and Assumption 2 is satisfied when all concentrations are $Cn^{-\frac{v}{d}}$. Also note that while Assumption 2 places an upper bound on the total volume of the prior concentrations to ensure the prior does not overwhelm the empirical Histogram at large sample sizes, it in general does not place any shape restrictions on the prior; in particular other prior shapes besides the uniform can be constructed.

The proof of Theorem 1 is composed from the following two auxiliary lemmas. The first auxiliary lemma upper bounds the rate of convergence of the posterior mean histogram, $\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}$, towards P_0 in mean W_v distance. The second lemma establishes a PCR *around* $\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}$, rather than P_0. It is the second lemma that leverages the conjugacy of this model.

Lemma 1. *Let $Y_1, \ldots, Y_n \overset{iid}{\sim} P_0 \in \mathcal{P}_d$. Suppose k_n satisfies Assumption 1, $\boldsymbol{\alpha}_{b_n}$ satisfies Assumption 2 and that for $n \in \mathbb{N}$ and $\boldsymbol{j} \in [b_n]^d$, $\alpha_{\boldsymbol{j},b_n} \geq 0$. Then*

$$\mathbb{E}_{P_0} W_v(P_0, \bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}) \lesssim \begin{cases} n^{-\frac{1}{2v}} & d < 2v. \\ n^{-\frac{1}{2v}} \log^{\frac{1}{v}}(n) & d = 2v. \\ n^{-\frac{1}{d}} & d > 2v. \end{cases}$$

Lemma 2. *Let $Y_1, \ldots, Y_n \overset{iid}{\sim} P_0 \in \mathcal{P}_d$. Suppose k_n satisfies Assumption 1. Let $\gamma > 1$, and let $\{\tau_n(d,v)\}_{n=1}^{\infty}$ be a sequence satisfying*

$$\tau_n(d,v) = C_1(d,v) \begin{cases} n^{-\frac{1}{2v}} \log^{\frac{\gamma}{v}}(n) & d < 2v, \\ n^{-\frac{1}{2v}} \log^{\frac{1+\gamma}{v}}(n) & d = 2v, \\ n^{-\frac{1}{d}} \log^{\frac{\gamma}{v}}(n) & d > 2v, \end{cases} \tag{12}$$

Then, provided that $\alpha_{\boldsymbol{j},b_n} > 0$ for each $n \in \mathbb{N}$ and $\boldsymbol{j} \in [b_n]^d$, we have that for $1 \leq v < \infty$ and $d \in \mathbb{N}$ and $C_1(d,v)$ sufficiently large

$$\mathbb{E}_{p_0} \Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}(P \in \mathcal{P}_d : W_v(P, \bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}) \geq \tau_n(d,v)) \to 0 \text{ as } n \to \infty.$$

The main technical challenges appear in proving the auxiliary lemmas. Given Lemmas 1 and 2, Theorem 1 follows easily and we show this now. For ease in notation, through the remainder of the paper we drop the k_n and $\boldsymbol{\alpha}_{b_n}$ subscripts from the notation for the posterior, thus $\Pi_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ is referred to as Π_n (and $\bar{P}_{n,k_n,\boldsymbol{\alpha}_{b_n}}$ is referred to as $\bar{P}_n$). This does not cause ambiguity in what follows because the values of k_n and $\boldsymbol{\alpha}_{b_n}$ are given in assumptions 1 and 2.

Proof of Theorem 1. By the triangle inequality and the union bound

$$\begin{aligned} \mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P_0, P) \geq \epsilon_n(d,v)\right)\right] &\leq \mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P_0, \bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right)\right] \\ &\quad + \mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P, \bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right)\right] \\ &= P_0\left[W_v(P_0, \bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right] \\ &\quad + \mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P, \bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right)\right], \end{aligned} \tag{13}$$

Using Markov's inequality and Lemma 1

$$P_0\left[W_v(P_0,\bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right] \leq 2\frac{\mathbb{E}_{p_0}W_v(P_0,\bar{P}_n)}{\epsilon_n(d,v)} \lesssim \left\{\begin{array}{ll} \dfrac{n^{-\frac{1}{2v}}}{n^{-\frac{1}{2v}}\log^{\frac{\gamma}{v}}(n)} & d<2v \\ \dfrac{n^{-\frac{1}{2v}}\log^{\frac{1}{v}}(n)}{n^{-\frac{1}{2v}}\log^{\frac{1+\gamma}{v}}(n)} & d=2v \\ \dfrac{n^{-\frac{1}{d}}}{n^{-\frac{1}{d}}\log^{\frac{\gamma}{v}}(n)} & d>2v \end{array}\right\} \to 0 \text{ as } n\to\infty. \tag{14}$$

Setting $C_0(d,v) \geq 2C_1(d,v)$ we have that $\tau_n(d,v) \leq \frac{\epsilon_n(d,v)}{2}$ for every $v \geq 1, d \in \mathbb{N}$ where $\tau_n(d,v)$ is as defined in Lemma 2. Using this and Lemma 2, we have that for every $v \geq 1, d \in \mathbb{N}$,

$$\mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P,\bar{P}_n) \geq \frac{\epsilon_n(d,v)}{2}\right)\right] \leq \mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P,\bar{P}_n) \geq \tau_n(d,v)\right)\right] \to 0 \tag{15}$$

as $n \to \infty$. By equations 13, 14, and 15, we conclude that for all $d \in \mathbb{N}, v \geq 1$,

$$\mathbb{E}_{p_0}\left[\Pi_n\left(P \in \mathcal{P}_d : W_v(P_0,P) \geq \epsilon_n(d,v)\right)\right] \to 0 \tag{16}$$

as $n \to \infty$. By Markov the theorem statement follows. □

4 Proofs of Auxiliary Lemmas

In this section we prove Lemmas 1 and 2. First we need to state a couple of technical tools.

4.1 Technical Tools

The first tool is the multiresolution upper bound on the Wasserstein distance. See Weed & Bach (2019) section 3 or Singh & Póczos (2018) appendix section A for a good review. Here we use an application of this general result for the metric space $([0,1)^d, \|\cdot\|_2)$.

Lemma 3. *(Wasserstein Multiresolution Upper Bound) Let $\mathcal{S}_0 = [0,1)^d$ and for $k \in \mathbb{N}$,*

$$\mathcal{S}_k := \left\{\left[\frac{i_1-1}{2^k}, \frac{i_1}{2^k}\right) \times \left[\frac{i_2-1}{2^k}, \frac{i_2}{2^k}\right) \times \cdots \times \left[\frac{i_d-1}{2^k}, \frac{i_d}{2^k}\right) \text{ for } (i_1,i_2,\ldots,i_d) \in [2^k]^d\right\},$$

If μ, ν are probability measures on $[0,1)^d$, then for $v \geq 1$

$$W_v^v(\mu_1,\mu_2) \leq d^{v/2}\left(\left(\frac{1}{2}\right)^{Kv} + \sum_{k=1}^{K}\left(\frac{1}{2}\right)^{(k-1)v}\sum_{S\in\mathcal{S}_k}|\mu_1(S)-\mu_2(S)|\right).$$

Proof. This is a straightforward application of proposition 1 of Weed & Bach (2019). □

The next technical tool is an upper bound on the L_1 concentration of a Multinomial distribution around its mean.

Lemma 4 (Multinomial concentration)**.** *If $(X_1,\ldots,X_k) \sim$ Multinomial$(n,p_1,\ldots,p_k)$ and $Z := \sum_{j=1}^k |X_j - np_j|$, then*

$$\mathbb{E}(Z/n) \leq \sqrt{\frac{k-1}{n}}.$$

Proof. Applying Jensen's inequality and then Cauchy-Schwarz

$$\mathbb{E}(\frac{Z}{n}) \leq \sum_{j=1}^k \sqrt{\mathrm{Var}(\frac{X_j}{n})} = \frac{1}{\sqrt{n}}\sum_{j=1}^k\sqrt{p_j(1-p_j)} \leq \frac{1}{\sqrt{n}}\sqrt{\sum_{j=1}^k p_j \sum_{j=1}^k (1-p_j)} = \sqrt{\frac{k-1}{n}}$$

□

The last tool is the concentration of the Dirichlet distribution around its mean in the L_1 norm.

Lemma 5. *(Dirichlet Concentration) Let $k \in \mathbb{N}$ and $(\pi_1, \pi_2, \ldots, \pi_k) \sim$ Dirichlet$(\alpha_1, \alpha_2, \ldots, \alpha_k)$. Then for $\delta > 0$*

$$\mathbb{P}\left(\sum_{j=1}^{k} |\pi_j - \mathbb{E}(\pi_j)| \geq \frac{(\bar{\alpha})^{-\frac{1}{2}}\sqrt{k}}{\delta}\right) \leq \delta,$$

where $\bar{\alpha} := \sum_{j=1}^{k} \alpha_j$.

Proof. Basic properties of the Dirichlet distribution give that for $j \in \{1, 2, \ldots, k\}$, $\pi_j \sim \text{Beta}(\alpha_j, \bar{\alpha} - \alpha_j)$. Also, if $X \sim \text{Beta}(\alpha, \beta)$ then $\text{Var}(X) = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$. Using these properties, in addition to Jensen's inequality and Cauchy–Schwarz inequality, we have that

$$\begin{aligned}
\mathbb{E}\left(\sum_{j=1}^{k} |\pi_j - \mathbb{E}(\pi_j)|\right) &\leq \sum_{j=1}^{k} \sqrt{\text{Var}(\pi_j)} \\
&= \sum_{j=1}^{k} \sqrt{\frac{\alpha_j(\bar{\alpha} - \alpha_j)}{\bar{\alpha}^2(\bar{\alpha}+1)}} \\
&\leq (\bar{\alpha})^{-\frac{3}{2}} \sum_{j=1}^{k} \sqrt{\alpha_j(\bar{\alpha} - \alpha_j)} \\
&\leq (\bar{\alpha})^{-\frac{3}{2}} \sqrt{\left(\sum_{j=1}^{k} \alpha_j\right)\left(\sum_{j=1}^{k} \bar{\alpha} - \alpha_j\right)} \\
&= (\bar{\alpha})^{-\frac{3}{2}} \sqrt{\bar{\alpha}(\bar{\alpha}k - \bar{\alpha})} \\
&\leq (\bar{\alpha})^{-\frac{1}{2}} \sqrt{k}. \qquad (17)
\end{aligned}$$

By Markov the result follows. □

4.2 Proof of Lemma 1

We now prove Lemma 1.

Proof of lemma 1. Due to the nesting of the dyadic models and since $b_n = 2^{K_n}$, we have for $k \in \{1, 2, \ldots, K_n\}$ and $S \in \mathcal{S}_k$ a set $I_{s,k,n} \subseteq [b_n]^d$ such that

$$S = \bigcup_{\boldsymbol{j} \in I_{s,k,n}} A_{\boldsymbol{j},b_n}. \qquad (18)$$

Moreover $\{\cup_{\boldsymbol{j} \in I_{S,k,n}} A_{\boldsymbol{j},b_n}\}_{S \in \mathcal{S}_k}$ partitions $[0,1)^d$ and $\{I_{s,k,n}\}_{S \in \mathcal{S}_k}$ partitions $[b_n]^d$. Using this and lemma 3, we have that

$$\mathbb{E}_{p_0} W_v^v(P_0, \bar{P}_n) \lesssim \left(\frac{1}{2}\right)^{K_n v} + \sum_{k=1}^{K_n} \left(\frac{1}{2}\right)^{(k-1)v} \mathbb{E}_{p_0} \sum_{S \in \mathcal{S}_k} \left| \bar{P}_n\Big(\bigcup_{\boldsymbol{j} \in I_{S,k,n}} A_{\boldsymbol{j},b_n}\Big) - P_0(S) \right|. \qquad (19)$$

By definition of $\bar{P}_n$ and since the $A_{\boldsymbol{j},b_n}$ are disjoint,

$$\begin{aligned}
\left| \bar{P}_n\Big(\bigcup_{\boldsymbol{j} \in I_{S,k,n}} A_{\boldsymbol{j},b_n}\Big) - P_0(S) \right| &= \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \frac{\alpha_{\boldsymbol{j},b_n} + \sum_{t=1}^{n} \mathbb{I}(Y_t \in A_{\boldsymbol{j},b_n})}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i},b_n}} - P_0(S) \right| \\
&= \left| \frac{\sum_{\boldsymbol{j} \in I_{S,k,n}} \alpha_{\boldsymbol{j},b_n} + \sum_{t=1}^{n} \mathbb{I}(Y_t \in \bigcup_{\boldsymbol{j} \in I_{S,k,n}} A_{\boldsymbol{j},b_n})}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i},b_n}} - P_0(S) \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left| \frac{n}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} \frac{\sum_{t=1}^n \mathbb{I}(Y_t \in S)}{n} - P_0(S) \right| + \frac{\sum_{\boldsymbol{j} \in I_{S,k,n}} \alpha_{\boldsymbol{j}, b_n}}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} \\
&\leq \left| \frac{n}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} - 1 \right| \frac{\sum_{t=1}^n \mathbb{I}(Y_t \in S)}{n} + \left| \frac{\sum_{t=1}^n \mathbb{I}(Y_t \in S)}{n} - P_0(S) \right| \\
&\quad + \frac{\sum_{\boldsymbol{j} \in I_{S,k,n}} \alpha_{\boldsymbol{j}, b_n}}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}}.
\end{aligned} \tag{20}
$$

Using this and that $\mathcal{S}_k$ partitions $[0,1)^d$ and Lemma 4 yields

$$
\mathbb{E}_{p_0} \sum_{S \in \mathcal{S}_k} \left| \bar{P}_n(\bigcup_{\boldsymbol{j} \in I_{S,k,n}} A_{\boldsymbol{j}, b_n}) - P_0(S) \right| \lesssim \frac{\sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j}, b_n}}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} + n^{-\frac{1}{2}} \sqrt{|\mathcal{S}_k|}
$$

Using this and equation 19 and that $|\mathcal{S}_k| = 2^{dk}$

$$
\begin{aligned}
\mathbb{E}_{p_0} W_v^v(P_0, \bar{P}_n) &\lesssim \left(\frac{1}{2}\right)^{K_n v} + \frac{\sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} \sum_{k=1}^{K_n} \left(\frac{1}{2}\right)^{(k-1)v} + n^{-1/2} \sum_{k=1}^{K_n} 2^{-k(v - \frac{d}{2})} \\
&\lesssim k_n^{-v} + \frac{\sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}}{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i}, b_n}} + n^{-\frac{1}{2}} \left(\max(1, 2^{K_n(\frac{d}{2} - v)}) \mathbb{I}(d \neq 2v) + K_n \mathbb{I}(d = 2v) \right)
\end{aligned} \tag{21}
$$

Applying Assumptions 1 and 2 now allows us to conclude that $\mathbb{E}_{p_0} W_v^v(P_0, \bar{P}_n) \lesssim n^{-\frac{1}{2}}$ when $d < 2v$, $\lesssim n^{-\frac{v}{d}} \log(n)$ when $d = 2v$ and $\lesssim n^{-\frac{v}{d}}$ when $d > 2v$. Applying Jensen's inequality to upper bound $\left(\mathbb{E}_{p_0} W_v(P_0, \bar{P}_n)\right)^v$ by $\mathbb{E}_{p_0} W_v^v(P_0, \bar{P}_n)$ completes the proof. □

4.3 Proof of Lemma 2

We now prove Lemma 2

Proof of Lemma 2. Again using Lemma 3 and the sets $I_{s,k,n}$ from equation 18, we have for $n \in \mathbb{N}, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \mathcal{S}^{b_n d - 1}$

$$
\begin{aligned}
W_v^v(\psi_{b_n}(\boldsymbol{\pi}_1), \psi_{b_n}(\boldsymbol{\pi}_2)) &\leq d^{v/2} \left[\left(\frac{1}{2}\right)^{K_n v} + \sum_{k=1}^{K_n} \left(\frac{1}{2}\right)^{(k-1)v} \sum_{S \in \mathcal{S}_k} |\psi_{b_n}(\boldsymbol{\pi}_1)(S) - \psi_{b_n}(\boldsymbol{\pi}_2)(S)| \right] \\
&= d^{v/2} \left[\left(\frac{1}{2}\right)^{K_n v} + \sum_{k=1}^{K_n} \left(\frac{1}{2}\right)^{(k-1)v} \sum_{S \in \mathcal{S}_k} |\psi_{b_n}(\boldsymbol{\pi}_1)(\bigcup_{\boldsymbol{j} \in I_{s,k,n}} A_{\boldsymbol{j}, b_n}) - \psi_{b_n}(\boldsymbol{\pi}_2)(\bigcup_{\boldsymbol{j} \in I_{s,k,n}} A_{\boldsymbol{j}, b_n})| \right] \\
&= d^{v/2} \left[\left(\frac{1}{2}\right)^{K_n v} + \sum_{k=1}^{K_n} \left(\frac{1}{2}\right)^{(k-1)v} \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{2\boldsymbol{j}} \right| \right]
\end{aligned} \tag{22}
$$

Using the above equation, the preimage form of Π_n (equation 7), the definition of $\bar{P}_n$ (equation 9), the definition of z_n^* (the posterior measure over the simplex $\mathcal{S}^{b_n d - 1}$), and that by assumption 1, $2^{-K_n v} =$

$o(\tau_n^v(d,v))$ (as $n \to \infty$) we have that almost surely under P_0 and eventually in n and for each $d \in \mathbb{N}, v \geq 1$

$$
\begin{aligned}
&\Pi_n(P \in \mathcal{P}_d : W_v(P, \bar{P}_n) \geq \tau_n(d,v)) \\
&= z_n^*(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : W_v^v(\psi_{b_n}(\boldsymbol{\pi}_1), \psi_{b_n}(\mathbb{E}_{z_n^*}(\boldsymbol{\pi}|Y_1,\ldots,Y_n))) \geq \tau_n^v(d,v)) \\
&\leq z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \frac{1}{2^{K_n v}} + \sum_{k=1}^{K_n} \frac{1}{2^{(k-1)v}} \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| \geq d^{-v/2}\tau_n^v(d,v)\right) \\
&\leq z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \sum_{k=1}^{K_n} \frac{1}{2^{(k-1)v}} \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| \geq \frac{1}{2} d^{-v/2}\tau_n^v(d,v)\right) \\
&\leq z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \sum_{k=1}^{K_n} \frac{1}{2^{kv}} \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| \geq 2^{v-1} d^{-v/2}\tau_n^v(d,v)\right),
\end{aligned}
\tag{23}
$$

Now note that

$$
\sum_{k=1}^{K_n} 2^{-kv}\left(\frac{\log^\gamma(n) 2^{\frac{dk}{2}}}{\sqrt{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i},b_n}}}\right) = \frac{\log^\gamma(n)}{\sqrt{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i},b_n}}} \sum_{k=1}^{K_n} 2^{-k(v-\frac{d}{2})} \lesssim \tau_n^v(d,v) \tag{24}
$$

To see the $\lesssim$ in equation 24, observe that by Assumption 2, the total prior concentration is dominated by n and therefore the term in front of the summand on LHS of $\lesssim$ is $\asymp \frac{\log^\gamma(n)}{\sqrt{n}}$. Thus by definition of $\tau_n^v(d,v)$, this is sufficient to conclude $LHS \lesssim \tau_n^v(d,v)$ in the $d < 2v$ case. In the $d = 2v$ case the sum contributes a factor $\log(n)$ to LHS and so again $LHS \lesssim \tau_n^v(d,v)$. In the $d > 2v$ case, the sum contributes an asymptotic factor $2^{K_n(\frac{d}{2}-v)} \asymp k_n^{\frac{d}{2}-v} = n^{\frac{1}{2}-\frac{v}{d}}$ to LHS so that $LHS \asymp \log^\gamma(n) n^{-\frac{v}{d}}$ and so again $LHS \lesssim \tau_n^v(d,v)$. By equation 24, for each $d \in \mathbb{N}, v \geq 1$, we set $C_1(d,v)$ sufficiently large so that eventually in n

$$
\sum_{k=1}^{K_n} 2^{-kv}\left(\frac{\log^\gamma(n) 2^{\frac{dk}{2}}}{\sqrt{n + \sum_{\boldsymbol{i} \in [b_n]^d} \alpha_{\boldsymbol{i},b_n}}}\right) < 2^{v-1} d^{-v/2}\tau_n^v(d,v) \tag{25}
$$

With $C_1(d,v)$ this large, we therefore have that eventually in n

$$
\begin{aligned}
&z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \sum_{k=1}^{K_n} \frac{1}{2^{kv}} \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| \geq 2^{v-1} d^{-v/p}\tau_n^v(d,v)\right) \\
&\leq z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \exists k \in \{1,2,\ldots,K_n\} \text{ s.t } \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| \right. \\
&\qquad \left. > \log^\gamma(n)\sqrt{\frac{2^{dk}}{n + \sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j},b_n}}}\right) \\
&\leq \sum_{k=1}^{K_n} z_n^*\left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}}|Y_1,\ldots,Y_n) \right| > \log^\gamma(n)\sqrt{\frac{2^{dk}}{n + \sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j},b_n}}}\right),
\end{aligned}
\tag{26}
$$

where in the last line we have used the union bound. Now recall $\{I_{S,k,n}\}_{S \in \mathcal{S}_k}$ partitions $[b_n]^d$ for $k \in \{1,2,\ldots,K_n\}$. In particular, since $z_n^* = \text{Dirichlet}(\cdot|\{\alpha_{\boldsymbol{j},b_n}^*\}_{\boldsymbol{j} \in [b_n]^d})$, under z_n^*, $\{\sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{\boldsymbol{j}}\}_{S \in \mathcal{S}_k} \sim \text{Dirichlet}(\{\sum_{\boldsymbol{j} \in I_{S,k,n}} \alpha_{\boldsymbol{j},b_n}^*\}_{S \in \mathcal{S}_k})$. Moreover, $\sum_{S \in \mathcal{S}_k} \sum_{\boldsymbol{j} \in I_{S,k,n}} \alpha_{\boldsymbol{j},b_n}^* = \sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j},b_n}^* \overset{a.s}{=} n + \sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j},b_n}$. Finally note that by definition of $\mathcal{S}_k$, $|\mathcal{S}_k| = 2^{dk}$. So for $n \in \mathbb{N}$ and $k \in \{1,2,\ldots,K_n\}$ applying Dirichlet concentration of measure Lemma 5 with $\delta := \log^{-\gamma}(n)$, we have that for $C_1(d,v)$ sufficiently large, eventually

in n,

$$\sum_{k=1}^{K_n} z_n^* \left(\boldsymbol{\pi}_1 \in \mathcal{S}^{b_n d-1} : \sum_{S \in \mathcal{S}_k} \left| \sum_{\boldsymbol{j} \in I_{S,k,n}} \pi_{1\boldsymbol{j}} - \sum_{\boldsymbol{j} \in I_{S,k,n}} \mathbb{E}_{z_n^*}(\pi_{\boldsymbol{j}} | Y_1, \ldots, Y_n) \right| > \log^{\gamma}(n) \sqrt{\frac{2^{dk}}{n + \sum_{\boldsymbol{j} \in [b_n]^d} \alpha_{\boldsymbol{j}, b_n}}} \right) \leq K_n \log^{-\gamma}(n) \tag{27}$$

Finally note $K_n \log^{-\gamma}(n) \to 0$ since $K_n \lesssim \log(n)$ and $\gamma > 1$. By equations 23, 26 and 27, we have that for each $d \in \mathbb{N}$ and $v \geq 1$ and $C_1(d, v)$ sufficiently large,

$$\Pi_n(P \in \mathcal{P}_d : W_v(P, \bar{P}_n) \geq \tau_n(d, v)) \to 0 \text{ as } n \to \infty. \tag{28}$$

almost surely under P_0. By dominated convergence the conclusion of the lemma follows. □

5 Conclusions

In this work we obtained minimax optimal PCRs for unconstrained distribution estimation on $[0, 1]^d$ underneath the Wasserstein-v distances for every data dimension d. To the best of our knowledge these are the first PCRs achieving minimaxity for every problem dimension d under $W_v, v \geq 1$ distance. Our proof technique avoids verifying a Kullback-Liebler prior support condition by using conjugacy and a direct analysis of the posterior distribution.

These results may be useful to practicioners needing to estimate a distribution underneath a Wasserstein distance when they have some knowledge prior to data collection about the shape of the distribution they are estimating, intend to encode this through a prior distribution to potentially achieve increased accuracy at low sample sizes, and yet simultaneously require a guarantee of precision at large sample sizes that is robust to inaccurate prior selection.

An important area for future work is to determine whether for high dimensional data, Bayesian models can adaptively achieve minimax optimal PCRs underneath Wasserstein-v distances in constrained distribution estimation settings where it is safe to assume that the distribution to be estimated is of low entropy or has a smooth density.

References

Sergey Bobkov and Michel Ledoux. *One-dimensional empirical measures, order statistics, and Kantorovich transport distances*, volume 261. American Mathematical Society, 2019.

Federico Camerlenghi, Emanuele Dolera, Stefano Favaro, and Edoardo Mainini. Wasserstein posterior contraction rates in non-dominated bayesian nonparametric models. *arXiv preprint arXiv:2201.12225*, 2022.

Minwoo Chae, Pierpaolo De Blasi, and Stephen G Walker. Posterior asymptotics in wasserstein metrics on the real line. *Electronic Journal of Statistics*, 15(2):3635–3677, 2021.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, 2015.

Fengnan Gao and Aad van der Vaart. Posterior contraction rates for deconvolution of dirichlet-laplace mixtures. 2016.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pp. 500–531, 2000.

Judith Rousseau and Catia Scricciolo. Wasserstein convergence in bayesian and frequentist deconvolution models. *arXiv preprint arXiv:2309.15300*, 2023.

Catia Scricciolo. Bayes and maximum likelihood for l 1-wasserstein deconvolution of laplace mixtures. *Statistical Methods & Applications*, 27(2):333–362, 2018.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

Rui Zhang, Christian Walder, Edwin V Bonilla, Marian-Andrei Rizoiu, and Lexing Xie. Quantile propagation for wasserstein-approximate gaussian processes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21566–21578. Curran Associates, Inc., 2020.