On excess mass behavior in Gaussian mixture models with Orlicz-Wasserstein distances

Aritra Guha [*1] **Nhat Ho** [2] **XuanLong Nguyen** [3]

Abstract

Dirichlet Process mixture models (DPMM) in combination with Gaussian kernels have been an important modeling tool for numerous data domains arising from biological, physical, and social sciences. However, this versatility in applications does not extend to strong theoretical guarantees for the underlying parameter estimates, for which only a logarithmic rate is achieved. In this work, we (re)introduce and investigate a metric, named Orlicz-Wasserstein distance, in the study of the Bayesian contraction behavior for the parameters. We show that despite the overall slow convergence guarantees for all the parameters, posterior contraction for parameters happens at almost polynomial rates in outlier regions of the parameter space. Our theoretical results provide new insight in understanding the convergence behavior of parameters arising from various settings of hierarchical Bayesian nonparametric models. In addition, we provide an algorithm to compute the metric by leveraging Sinkhorn divergences and validate our findings through a simulation study.

1. Introduction

From their origin in the work of Pearson (Pearson, 1894), mixture models have been widely used by statisticians (McLachlan & Basford, 1988; Lindsay, 1995; Mengersen et al., 2011) in variety of modern interdisciplinary domains such as medical science (Schlattmann, 2009), bioinformatics (Ji et al., 2005), survival analysis (Tsodikov et al., 2003), psychometry (Gu et al., 2018) and image classification (Permuter et al., 2006), to name just a few. The heterogeneity in data populations and associated quantities of interest has inspired the use of a variety of kernels, each with its own advantages and characteristics. Gaussian kernels are particularly popular in various inferential problems, especially those related to density estimation and clustering analysis (Kotz et al., 2001; Bailey et al., 1994.; Roeder & Wasserman, 1997; Robert, 1996; Banfield & Raftery, 1993). In addition to the choice of kernels, the Bayesian mixture modelers are also guided by the selection of prior distributions for the quantities of interest. In particular, Bayesian nonparametric priors (BNP) for mixture models are increasingly embraced, thanks to computational ease and the modeling flexibility that these rich priors entail (Escobar & West, 1995; MacEachern, 1999).

On the theoretical front, convergence rates for (Gaussian) mixture models received extensive treatments in the Bayesian paradigm (Ghosal et al., 2000; Barron et al., 1999; Ghosal & van der Vaart, 2007). There have been enormous recent progress on both density estimation and parameter estimation problems. The density estimation problem under Gaussian mixture models with BNP priors was extensively studied by (Ghosal & van der Vaart, 2001) who obtained attractive polynomial rates of contraction relative to the Hellinger distance metric. In the parameter estimation problem, the metric of choice is Wasserstein distance, which proved to be a natural tool to analyze the convergence of mixture parameters (Nguyen, 2013). Moreover, (Nguyen, 2013) showed that the fast rates for density estimation with BNP Gaussian mixtures do not extend themselves to parameter estimation scenarios. Meanwhile, practitioners have employed successfully BNP mixture models, which yield useful estimates for model parameters that provide meaningful information about the data population's heterogeneity. This state of affairs leaves a gap in the theoretical understanding and the practical usage of Bayesian mixture models. In this paper, we aim to bridge this gap by capturing more accurately the heterogenenous behavior in the rates of parameter estimation. We proceed to describe this in further detail.

[1]Data Science & AI Research, AT&T Chief Data Office [2]Department of Statistics and Data Sciences, University of Texas, Austin [3]Department of Statistics, University of Michigan, Ann Arbor. Correspondence to: Aritra Guha <aguha0109@gmail.com, aritra@umich.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1.1. Gaussian Mixture Models

Consider discrete *mixing (probability) measure* $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. Here, $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a vector of mixing weights, while atoms $\{\theta_i\}_{i=1}^{k}$ are elements in a given space $\Theta \subset \mathbb{R}^d$. Here k is used to denote the number of components, which can potentially be infinite. Mixing measure G is combined with a (multivariate) Gaussian kernel with known covariance matrix Σ, denoted by $f_\Sigma(\cdot|\theta)$, with respect to the Lebesgue measure μ to yield a mixture density p_G. Here, f_Σ, admits the following form:

$$f_\Sigma(x|\theta) := \frac{\exp(-(x-\theta)^\top \Sigma^{-1}(x-\theta)/2)}{|2\pi\Sigma|^{-1/2}}, \tag{1}$$

where $|\cdot|$ in the denominator is the determinant operator of a square matrix. To avoid notational cluttering, we remove Σ from notation in the remainder of the paper and denote it as $f(\cdot|\theta)$.

The mixture density p_G may be represented as follows.

$$p_G(.) := \int f(\cdot|\theta)\mathrm{d}G(\theta) = \sum_{i=1}^{k} p_i f(\cdot|\theta_i). \tag{2}$$

The atoms θ_i's are representatives of the underlying subpopulations. Let $X_1, \ldots, X_n$ be i.i.d. samples from a mixture density $p_{G_0}(x) = \int f(x|\theta)\mathrm{d}G_0(\theta)$, where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown discrete mixing measure with *unknown* number of support points $k_0 \in \mathbb{N} \cup \{\infty\}$. We assume in this work that all the masses $\{p_i^0\}_{i=1}^{k_0}$ are strictly positive and the atoms $\{\theta_i^0 : i \leq k_0\}$ are distinct.

A Bayesian mixture modeler places a prior distribution Π_n on a suitable space (specifically, $\overline{\mathcal{G}}(\Theta)$ of discrete measures on Θ). The posterior distribution corresponding to Π_n, both of which may vary with sample size, can be computed as:

$$\Pi_n(G \in B \big| X_{1:n}) = \frac{\int_B \prod_{i=1}^{n} p_G(X_i)\mathrm{d}\Pi_n(G)}{\int_{\overline{\mathcal{G}}(\Theta)} \prod_{i=1}^{n} p_G(X_i)\mathrm{d}\Pi_n(G)}. \tag{3}$$

Dirichlet process Gaussian mixture models: In the absense of the knowledge of the number of mixture components k_0, the learning of mixture models is carried out by the use of Bayesian non-parametric (BNP) priors, leading to the *infinite mixture* setting. One of the most popular such priors is the Dirichlet process prior (Antoniak, 1974), which uses sample draws from a base measure H to define the random components and weights of the mixture model, leading to the popular *Dirichlet Process Gaussian Mixture Models* (DPGMM) (Lo, 1984; Escobar & West, 1995). In essence, the Dirichlet process prior places zero probability on mixing measures with a finite number of supporting atoms and enables the addition of more atoms in the supporting set as the number of data points increase. The DPGMM is formulated as follows:

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, H), \\ \theta_1, \ldots, \theta_n &\overset{i.i.d.}{\sim} G, \\ X_i|\theta_i &\sim f(X_i|\theta_i), \quad \forall i = 1, \ldots, n, \end{aligned} \tag{4}$$

where DP stands for Dirichlet process, the base measure H is a distribution on Θ, and $\alpha > 0$ is a concentration parameter which controls the rate at which new atoms may be considered, by varying the tail-behavior of mixture weights. A parametric counterpart of DPGMM is the mixture of finite Gaussian mixtures prior (MFM) (Miller & Harrison, 2018), which places all its mass on mixing measures with finite number of supporting atoms. BNP priors other than DPGMM may have the effects of pushing the atoms away from each other (Xie & Xu, 2017) or encouraging the weights of mixture to have a polynomial tail behavior (De Blasi et al., 2015).

The popularity of BNP priors may partially have been promoted due to a misconception that it "automatically" determines the number of components in the posterior inference process. This issue was highlighted by (Miller & Harrison, 2014), who demonstrated that Dirichlet Process priors overestimate the true number of components, k_0, almost surely. Subsequent work (Guha et al., 2021) has provided post-processing techniques to determine k_0 consistently with Dirichlet Process priors. Their method depends on the knowledge of the parameter contraction rate, with respect to the *Euclidean Wasserstein metric*, i.e., Wasserstein metric with underlying distance metric ℓ_2, a rate that is extremely slow for the Gaussian kernels (Nguyen, 2013).

The inconsistency of estimating k_0 arises primarily because Dirichlet priors typically tend to create a large number of extraneous components. While some of these components may be in the neighborhood of the true supports, others may be outliers and in practice, can be easily eliminated from consideration by careful truncation techniques. However, the Euclidean Wasserstein distance treats both the scenarios similarly and in turn yields slow convergence rates for both sets of extraneous atoms. This calls for alternative metrics for investigating parameter estimation rates. In a recent work, (Manole & Ho, 2022) argued that Wasserstein metrics capture only the worst-case uniform rates of parameter estimation and therefore can yield extremely slow rates in comparison to the local rates observed in practice, which may vary drastically based on the likelihood curvature in the parameter neighborhood. Employing alternate distance metrics via the use of Voronoi tessallations, they showed that in the finite Gaussian mixture setting with overfitted components (where $\infty > k > k_0$), even though the uniform convergence rates may be slow as k increases, there may still be some atoms which enjoy much faster rates of convergence.

The infinite Gaussian mixture setting is generally more challenging to address, (a) since the "true" atoms are not guaranteed to be well-separated, (b) each true atom may be surrounded by potentially infinitely many atoms a posteriori and (c) a posteriori samples can potentially have a significant portion of atomic masses attributed to outlier regions of the parameter space. We argue in this work that in the infinite Gaussian mixture setting, the rates captured by Wasserstein distances for outlier masses are inadequately slow and will demonstrate that with the help of a new suitably defined choice of metric this difficulty can be alleviated.

1.2. Contribution

As a primary contribution of this work we study a generalized class of metrics called *Orlicz-Wasserstein* metrics, in the context of parameter estimation arising in infinite mixture models. We show that an in-depth analysis using this metric helps alleviate a number of the concerns attributable to the use of Wasserstein distances for quantifying the rates of parameter convergence arising in infinite Gaussian mixtures. This class of distance metrics generalizes the Wasserstein metric relative to the Orlicz norm using a variety of choices of convex functions. They encompass a very wide range of distances on the space of probability measures, including the Euclidean Wasserstein metrics as a special case. By making appropriate choices of convex functions we can obtain a fast, almost polynomial contraction rates for atomic masses in outlier regions of the parameter space. This is very different from the slow local contraction behavior around the true atoms under the standard Wasserstein metric. This helps us establish informative and useful finer details about the convergence behavior of parameter estimates underlying the usage of Gaussian mixture models in clustering. We believe the usage of Orlicz-Wasserstein metrics for parameter estimation in Dirichlet process Gaussian mixture models opens a new range of directions for future research that aim for developing statistically sound and computationally efficient strategies for posterior sampling with mixture models.

Organization. The remainder of the paper is organized as follows. Section 2 provides necessary backgrounds about posterior contraction of parameters in Gaussian mixture models under Wasserstein distances. Section 3.1 introduces *Orlicz-Wasserstein* distances and some of its key properties. Section 3.4 provides computational approximations to calculating Orlicz-Wasserstein metrics for two mixing measures. Section 3.2 presents exact lower bounds for the Hellinger metric with respect to Orlicz-Wasserstein distances for Gaussian kernels. Section 3.3 uses the results in Section 3.2 to provide the key results in the paper with regards to contraction behavior using Orlicz-Wasserstein metrics. Proofs of results are deferred to the Appendices.

Notation. For any function $g : \mathcal{X} \to \mathbb{R}$, we denote $\widetilde{g}(\omega)$ as the Fourier transformation of function g. Given two densities p, q (with respect to the Lebesgue measure μ), the squared Hellinger distance is given by $h^2(p,q) = (1/2)\int(\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)$. For any metric d on Θ, we define the open ball of d-radius ϵ around $\theta_0 \in \Theta$ as $B_d(\epsilon, \theta_0)$. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore, we denote A^c as the complement of set A for any set A while $B(x, r)$ denotes the ball, with respect to the l_2 norm, of radius $r > 0$ centered at $x \in \mathbb{R}^d$. The expression $D(\epsilon, \mathscr{P}, d)$ used in the paper denotes the ϵ-packing number of the space $\mathscr{P}$ relative to the metric d. d is replaced by h to denote the hellinger norm. Finally, we use $\mathrm{Diam}(\Theta) = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$ to denote the diameter of a given parameter space Θ relative to the l_2 norm, $\|\cdot\|$, for elements in $\mathbb{R}^d$. Regarding the space of mixing measures, let $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively denote the space of all mixing measures with exactly and at most k support points, all in Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. $\overline{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ. Finally, $\mathcal{M}(\Theta)$ stands for the space of all probability measures on Θ.

2. Posterior contraction under Wasserstein distance

Following the work of (Nguyen, 2013), Wasserstein distances have been used to explore parameter estimation rates of mixture models, embodied through their mixing measures. In this section, we outline the basic concepts as follows. Let $\Theta \subset \mathbb{R}^d$. Moreover, define $\mathcal{M}(\Theta) = \{P : P \text{ is a probability measure on } \Theta\}$.

Definition 1. Given $\mu, \nu \in \mathcal{M}(\Theta)$ and the l_2 metric $\|\cdot\|$ on $\mathbb{R}^d$, the Wasserstein distance (Villani, 2009) of order r seeks a joint measure $\pi \in \Pi$ minimizing

$$W_r(\mu, \nu) := \left(\inf_{\pi \in \Pi} \int_{\Theta \times \Theta} \|\theta_1 - \theta_2\|^r d\pi(\theta_1, \theta_2)\right)^{1/r}. \quad (5)$$

Here, Π is the set of couplings of μ and ν denoted by $\Pi = \{\pi : \gamma^1_\# \pi = \mu, \gamma^2_\# \pi = \nu\}$, where γ^1, γ^2 are functions that project onto the first and second coordinates of $\Theta \times \Theta$ respectively.

In particular, as shown by (Nguyen, 2013), given two discrete measures $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$, a coupling between $\boldsymbol{p}$ and $\boldsymbol{p'}$ is a joint distribution $\boldsymbol{q}$ on $[1 \ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix

$\boldsymbol{q} = (q_{ij})_{1\le i\le k, 1\le j\le k'} \in [0,1]^{k\times k'}$ with marginal probabilities $\sum_{i=1}^{k} q_{ij} = p'_j$ and $\sum j = 1^{k'} q_{ij} = p_i$ for any $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k'$. We use $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ to denote the space of all such couplings of $\boldsymbol{p}$ and $\boldsymbol{p'}$. For any $r \ge 1$, the r-th order Wasserstein distance between G and G' is given by

$$W_r(G, G') \quad = \quad \inf_{\boldsymbol{q}\in\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})} \left(\sum_{i,j} q_{ij} \|\theta_i - \theta'_j\|^r \right)^{1/r} . \quad (6)$$

(Heinrich & Kahn, 2018) show that with Gaussian kernels, the minimax rate for estimation is dependent on the number of extra components and goes down as the number of potential components increases, meaning it gets harder to accurately cluster the observations as we have more and more extra components. The Gaussian kernel being smooth fits in as many components as possible without changing the mixture density and therefore achieves a very slow parameter contraction rate. With potentially infinitely many extra components (while using Dirichlet Process priors), rates are even slower. In fact, (Nguyen, 2013) shows that for DPGMM with posterior distribution $\Pi_n(\cdot|X_{1:n})$, the following holds true.

$$\Pi_n \left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_0) \lesssim (\log n)^{-1/2} \middle| X_{1:n} \right) \to 1 \quad (7)$$

in p_{G_0}-probability. This bound can be shown to be tight leveraging the results of (Ded, 2013). On the other hand, it has been shown that ordinary-smooth kernels need only a power of $-\log(\epsilon)$ components to approximate an infinite component mixing density upto ϵ- approximation in $\mathbb{L}_q$ distance (Nguyen, 2013; Gao & van der Vaart, 2016). Correspondingly, Laplace kernels need a polynomial power of $(1/\epsilon)$ many components for the same degree of approximation. This combined with (7) suggests that BNP priors use a lot more extra components to fit the true mixture distribution than is necessary, especially with Gaussian kernel. The extra components can potentially arise from two different sources, (i) multiple supporting atoms in the posterior trying to approximate each true atom, (ii) or excessively many outlier atoms in the posterior sample. If condition (ii) is true, this may potentially have negative consequences for using Gaussian kernels for clustering purposes. From Eq. (7), we are only able to conclude that

$$\Pi_n \left(G = \sum p_i \delta_{\theta_i} \quad : \quad \sum_j p_j \mathbb{1}_{\{\|\theta_j - \theta_i^0\| > \eta \ \forall i\}} \right.$$
$$\left. \gtrsim \quad \log(n)^{-1}/\eta^2 \middle| X_{1:n} \right) \to 1 \quad (8)$$

which states that masses attributed to outlier atoms (those $> \eta$ distance from any "true" atom) vanish at only a slow logarithmic rate. Clearly, while standard Wasserstein distances are the popular choices of metrics, they do not help differentiate between the sources of extra atoms, and thereby are not useful while discarding outlier atoms. To facilitate this distinction of the source of excess atoms, in this paper we consider a generalisation of standard (Euclidean) Wasserstein metrics called *Orlicz-Wasserstein* distances which allow placement of higher weight penalties on outliers and thereby help to identify outlier atoms better. We proceed in the following sections to describe this in further detail.

3. A generalized metric for contraction of mixing measures

In existing literature thus far, the rates of parameter estimation have been extensively studied with respect to Euclidean Wasserstein distances, in the works of (Nguyen, 2013; Ho & Nguyen, 2016b;a; Gao & van der Vaart, 2016; Guha et al., 2021). As part of this work, we extend such results to the regime of *Orlicz-Wassertein* metrics which take a more careful consideration of the geometry of the parameter space. In that regard, for the sake of completeness, we first introduce the reader to the notion of Orlicz norms and spaces as follows.

3.1. Orlicz-Wasserstein distance

The Orlicz norm is defined as follows (Wellner, 2017).

Definition 2. Let μ be a $\sigma-$finite measure on a space $\mathcal{X}$ with metric $\|\cdot\|$. Assume that $\Phi : [0,\infty) \to [0,\infty)$ be a convex function satisfying:

(i) $\dfrac{\Phi(x)}{x} \to \infty, \quad \text{as } x \to \infty,$

(ii) $\dfrac{\Phi(x)}{x} \to 0, \quad \text{as } x \to 0.$

Then, the Orlicz space is defined as follows:

$$L_\Phi := \big\{ f : \mathcal{X} \to \mathbb{R} | \ \exists\ \lambda \in \mathbb{R}^+ \text{ s.t.} \int_X \Phi(\|f(x)\|/\lambda)\, d\mu(x) \le 1 \big\}.$$

Moreover, the Orlicz norm corresponding to $f \in L_\Phi$ is given by:

$$\|f\|_\Phi := \inf\{\lambda \in \mathbb{R}^+ : \int_X \Phi(\|f(x)\|/\lambda)\, d\mu(x) \le 1\}. \quad (9)$$

Without loss of generalisation, we will assume $\mathcal{X} = \mathbb{R}^d$, with $\|\cdot\|$ denoting the standard Euclidean metric. Notice that when $\Phi(x) = x^p$ with $p \ge 1$, the Orlicz norm, $\|f\|_\Phi$ is the same as the $\mathbb{L}_p$-norm. In this sense, the Orlicz norm generalizes the concept of $\mathbb{L}_p$-norm for $p \ge 1$. Recall that,

a coupling between two probability measures ν_1 and ν_2 on $\mathbb{R}^d$ is a joint distribution on $\mathbb{R}^d \times \mathbb{R}^d$ with corresponding marginal distributions ν_1 and ν_2. Corresponding to the Orlicz norms, we define the *Orlicz-Wasserstein metric* which generalizes the W_r-metric as follows.

Definition 3. Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$. Assume that $\Phi : [0,\infty) \to [0,\infty)$ is a convex function satisfying conditions (i) and (ii) in Definition 2. We define the *Orlicz-Wasserstein* distance between ν_1 and ν_2 as follows:

$$W_\Phi(\nu_1,\nu_2) := \inf_{\nu\in\mathcal{Q}(\nu_1,\nu_2)} \inf\{\lambda\in\mathbb{R}^+ : \int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/\lambda)\, d\nu(x,y) \le 1\}, \tag{10}$$

where $\mathcal{Q}(\nu_1,\nu_2)$ is the set of all possible couplings of ν_1 and ν_2.

Orlicz Wasserstein distances have been briefly introduced in the works of (Kell, 2017; Sturm, 2011), however, the utility of the metrics for contraction properties of parameter estimation has remained hitherto unexplored. Also, following Lemma 3.1 of (Sturm, 2011), we see under some minor regularity conditions, for every Φ, ν_1, ν_2, there exists $\lambda_{\min}$ and ν_{opt} such that $\lambda_{\min} = W_\Phi(\nu_1,\nu_2)$ and $\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/\lambda_{\min})\, d\nu_{\text{opt}}(x,y) = 1$. This combined with Fubini's theorem establishes the equivalence of the definitions in this work and those of (Sturm, 2011; Kell, 2017).

Note that when $\Phi(x) = x^r$ for $r \ge 1$, then $W_\Phi(\nu_1,\nu_2) = W_r(\nu_1,\nu_2)$, the usual Wasserstein distance of order r between ν_1 and ν_2. The following lemma demonstrates that Orlicz-Wasserstein defines a proper metric on $(\mathbb{R}^d, \|\cdot\|)$.

Lemma 1. *The Orlicz-Wasserstein W_Φ is a distance metric on the set of probability measures on $(\mathbb{R}^d, \|\cdot\|)$, namely, it is symmetric and satisfies the identity and triangle inequality properties.*

The proof of Lemma 1 is in Appendix B.1. The notion of Orlicz-Wasserstein distance may encompass a stronger notion of metrics than that of the usual Wasserstein distance to compare probability measures as evidenced by the following lemma.

Lemma 2. *Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$. Also assume Φ, Ψ are convex functions satisfying conditions (i) and (ii) in Definition 2. Suppose that for all $x > 0$, $\Phi(x) \le \Psi(x)$. Then, we have*

$$W_\Phi(\nu_1,\nu_2) \le W_\Psi(\nu_1,\nu_2).$$

The proof of Lemma 2 is in Appendix B.2. Note that the supremum of convex functions is also a convex function. Therefore, as a corollary to the above lemma we obtain the following inequality.

Corollary 1. *Let $\Phi_1(\cdot)$ be a polynomial convex function and $\Phi_2(\cdot)$ an exponential convex function. Ψ is the supremum of $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$. Then the following holds, for any G, G', $1 > \alpha > 0$.*

$$\begin{aligned} W_\Psi(G,G') &\ge W_{\alpha\Phi_1+(1-\alpha)\Phi_2}(G,G') \\ &\ge \alpha W_{\Phi_1}(G,G') + (1-\alpha) W_{\Phi_2}(G,G') \end{aligned} \tag{11}$$

An important property of the Wasserstein distances is that if one mixing measure is close to another in Wasserstein distance, it provides a way to control the corresponding contraction rates of the atoms and the masses associated with them. The following lemma provides a similar result for Orlicz-Wasserstein norms.

Lemma 3. *Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$, $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ be mixing measures such that $\theta_j, \theta_i^0 \in \mathbb{R}^d$ for all i, j. Assume that $\Phi : [0,\infty) \to [0,\infty)$ is a convex function satisfying conditions (i) and (ii) in Definition 2. Then*

$$\sum_j p_j \mathbb{1}_{\{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i\}} \le \left(\Phi\left(\frac{\eta}{W_\Phi(G,G_0)}\right)\right)^{-1}. \tag{12}$$

Here, k_0, k can also take the value ∞.

The proof of Lemma 3 is in Appendix B.5. Lemma 3 allows us to identify the amount of mass transferred over large distances, when the mass transfer occurs between two measures G and G_0. Note that the constraint on Φ is very minimal, thereby lending flexibility to the result. Since operations like supremums of convex functions or compositions of a convex function with a non-decreasing convex function (this is the outer function), also yield convex functions, Lemma 3 is a standalone result of interest as a generalisation of Bernstein/Hoeffding type inequalities for mixing measures.

3.2. Lower bound of Hellinger distance based on Orlicz-Wasserstein metric

In the previous section, we state results to control the cost of mass transfer attributable to large transportation distances using Orlicz-Wasserstein distances. This is an important result in understanding the contraction behaviors of support points in the outlier regions of parameter space. Traditionally, contraction behavior has been extensively studied (Ghosal & van der Vaart, 2001) in the regime of mixture densities p_G. The following results help us connect our understanding of posterior contraction on space of mixture densities to that of mixing measure, relative to that of Orlicz-Wasserstein distances. This is stated as follows in the next theorem.

Theorem 1. *Let Φ be a convex function satisfying conditions (i) and (ii) in Definition 2 such that $\Phi(x) \le \exp(x^\beta) - 1$ for some $16/15 > \beta > 1$. Then, as $\Theta = [-\bar\theta, \bar\theta]^d$, for*

any mixing measures G, G', with corresponding densities $p_G, p_{G'}$, we have

$$W_\Phi(G,G') \lesssim C\bigg(\frac{\bar{\theta}^{5/4}}{(\log(1/h(p_G,p_{G'})))^{1/8}} + \bigg(\frac{1}{\log(1/h(p_G,p_{G'}))}\bigg)^{11/8} + \bigg(\frac{1}{\log(c/h(p_G,p_{G'})(\log(1/h(p_G,p_{G'})))^{d/4})}\bigg)^{1/2}\bigg) \quad (13)$$

for constants C, c dependent on the dimension and known covariance matrix.

The proof of Theorem 1 is in Appendix A.1. The key technical novelty of the proof lies in the idea of convolving the mixing measures with a mollifier which is exponentially integrable while its Fourier transform is smoother than the Gaussian location kernel. This helps to smoothly transition the problem of bounding distances on mixing measures to the Fourier transform domain of corresponding mixture densities. We make a few comments about the above theorem.

(i) The upper bound on the RHS of equation (13) depends on a power of log-Hellinger distance between the corresponding mixture densities. This strengthens the result in Theorem 2 of (Nguyen, 2013), who obtained a $(\log(1/h))^{-1/2}$ upper bound for $W_2(G,G')$. The result in Theorem 1 is obtained in terms of Orlicz-Wasserstein distances relative to an exponential convex function, thus lending it more flexibility.

(ii) The key object to obtaining this result is to find a suitable mollifier Z_δ, which we choose as $c\frac{1}{\delta}(\int \exp(-itx/\delta)\exp(-t^4)\mathrm{d}t)^2$ with c being the constant of proportionality for the proof of Theorem 1. However, we believe a more refined choice of mollifier can yield sharper estimates on the RHS of equation (13).

(iii) The result is obtained with exact computation of the involvement of $\bar{\theta}$. Therefore, it can also be used for posterior contraction rates with sieve priors, although for this work we study only compactly supported priors.

Outline of proof of Theorem 1: Here, we provide a proof strategy for Theorem 1, which relies on the following triangle inequality with Orlicz-Wasserstein distance between G and G':

$$W_\Phi(G,G') \le W_\Phi(G,G*Z_{\delta,d}) + W_\Phi(G',G'*Z_{\delta,d}) + W_\Phi(G*Z_{\delta,d},G'*Z_{\delta,d}), \quad (14)$$

where $Z_{\delta,d}(x_1,\ldots,x_d) := \prod_{i=1}^d \zeta_\delta(x_i)$ and $\zeta_\delta(x) := c\frac{1}{\delta}(\int \exp(-itx/\delta)\exp(-t^4)\mathrm{d}t)^2$, with c being the constant of proportionality. To control both $W_\Phi(G,G*Z_{\delta,d})$ and $W_\Phi(G',G'*Z_{\delta,d})$, we use the following lemma:

Lemma 4. *Assume that $\nu_2 = \nu_1 * Z_{\delta,d}$ where ν_1 is a given probability measure on $(\mathbb{R}^d, \|\cdot\|)$. Furthermore, suppose that $\Phi(x) \le \exp(x^\alpha) - 1$ for some $1 < \alpha < 4/3$. Then, there exists universal constant C_α depending only on α such that*

$$W_\Phi(\nu_1,\nu_2) \le C_\alpha \delta.$$

The proof of Lemma 4 is in Appendix B.3. For the final term $W_\Phi(G*Z_{\delta,d}, G'*Z_{\delta,d})$, we can upper bound it using the following result:

Lemma 5. *Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$ and let Φ be a convex function satisfying conditions (i) and (ii) in Definition 2. Then, we obtain that*

$$W_\Phi(\nu_1,\nu_2) \le 2\inf\{\lambda \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/\lambda)\, d|\nu_1(x) - \nu_2(x)| \le 1\}.$$

The proof of Lemma 5 is in Appendix B.4. Using triangle inequality and Lemmas 4 and 5, we obtain

$$W_\Phi(G,G') \precsim \delta + \inf\{\lambda \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/\lambda)\cdot|(G-G')*Z_{\delta,d}(x)|\mathrm{d}x \le 1\}.$$

We then decompose the integral with respect to $\mathbb{R}^d$ into two integrals: one with respect to $\|x\| \le M$ and one with respect to $\|x\| > M$, and after some algebraic manipulations, we have

$$\begin{aligned}\inf\bigg\{&\lambda \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/\lambda)\cdot|(G-G')*Z_{\delta,d}(x)|\mathrm{d}x \le 1\bigg\} \\ &\precsim \frac{M}{\log(C/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2}))} + \frac{(d\bar{\theta})^{5/4}}{\log(3/2)M^{1/4}} + \frac{\delta^{5/4}}{M^{1/4}},\end{aligned}$$

for any $M > 0$ where C is some universal constant. Collecting these results leads to

$$W_\Phi(G,G') \precsim \inf_{\delta,M}\bigg\{\delta + \frac{M}{\log(C/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2}))} + \frac{(d\bar{\theta})^{5/4}}{\log(3/2)M^{1/4}} + \frac{\delta^{5/4}}{M^{1/4}}\bigg\}.$$

Solving the minimization problem, we obtain the conclusion of Theorem 1.

In the next section, we use Theorem 1 to establish posterior contraction bounds of parameter estimating in Dirichet Process Gaussian mixtures.

3.3. Posterior contraction with Orlicz Wasserstein distances

On the parameter estimation front, (Nguyen, 2013; Guha et al., 2021; Ohn & Lin, 2020) establish logarithmic rates for estimating mixing measures in Dirichlet Process Gaussian mixtures. While (Nguyen, 2013) establishes an approximately $\log(n)^{-1/2}$ rate of contraction relative to the W_2 metric, more recently, (Ohn & Lin, 2020) establish minimax type $\approx \log(n)$ rates relative to the W_1 metric. Putting the results in context with Lemma 3, both those results imply, $\sum_j p_j \mathbb{1}_{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i} \approx \log(n)$, meaning the mass of posterior sample atoms in the region of parameter space not populated by atoms of the true (data-generating) mixing measure decays logarithmically. This puts the use of DPGMMs for clustering in a negative light.

In this section, we show that a much stronger almost polynomial rate can be established for this objective, facilitated by the use of Orlicz-Wasserstein metrics. To facilitate our presentation, we consider the following notation.

$$\mathcal{EX}_\eta(\Theta, r) := \Big\{ G = \sum p_i \delta_{\theta_i} \in \overline{\mathcal{G}}(\Theta_{n1}) : \sum_j p_j \mathbb{1}_{\{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i\}} \geq r \Big\}. \quad (15)$$

$\mathcal{EX}_\eta(\Theta, r)$ here denotes the set of mixing measures which devote at least r probability mass to atoms which are away from the atoms of G_0 by distance η. To study the contraction of mixing measure of DPGMMs, we impose the following assumption on the base distribution H.

(P.1) The base distribution H is supported on $\Theta = [-\bar{\theta}, \bar{\theta}]^d$, and absolutely continuous with respect to the Lebesgue measure μ on Θ and admits a density function $g(\cdot)$. Also, H is approximately uniform, i.e., $\min_{\theta \in \Theta} g(\theta) > \dfrac{c_0}{\mu(\Theta)} > 0$.

Let $f_1(n, d) := (\log(n)/(d+2) - \log(\log n))^{-1/8}$.

Theorem 2. *Given the Dirichlet Process Gaussian mixture models* (4)*, if Φ satisfies the assumptions in Theorem 1, then for any $\eta > 0$ the following holds:*

$$\Pi_n \Big(G \in \overline{\mathcal{G}}(\Theta) : W_\Phi(G, G_0) \geq f_1(n, d) \,\Big|\, X_{1:n} \Big) \overset{P^n_{G_0}}{\to} 0.$$

The proof of Theorem 2 is in Appendix A.2. The following result is a simple corollary of Theorem 2.

Corollary 2. *Given all the assumptions in Theorem 2,*

$$\Pi_n \Big(G \in \mathcal{EX}_\eta \Big(\Theta, 2 \exp\Big(\frac{-\eta \log(n)^{1/8}}{(d+2)} \Big) \Big) \,\Big|\, X_{1:n} \Big) \overset{P_{G_0}}{\to} 0. \quad (16)$$

The proof of Corollary 2 is in Appendix A.3.

Remarks: (i) Corollary 2 suggests that if η can be chosen sufficiently small so that each η-neighborhood contains at most one true atom, Gaussian mixture models can be useful choices in clustering as well since outlier atoms vanish at almost polynomial rates.

(ii) We believe the rate of contraction can be optimized further with a more refined choice of $\Phi(\cdot)$, however, we make no such attempts in this work. In particular, given existence of a mollifier integrable relative to $\exp(x^\beta) - 1 (\beta \approx 2)$ with a strictly sub-Gaussian Fourier transform, the same proof technique can be used to show that the Orlicz-Wasserstein rate (relative to W_Φ, with $\Phi(x) := \exp(x^2)$) of $\log(n)^{-1/2}$ (possibly ignoring $\log(\log(n))$ terms) can be achieved, in which case the excess mass would contract polynomially $\approx \exp(-c(\log(n)^{1/2})^2)$. Corollary 2 reveals the power of Orlicz-Wasserstein distances for Gaussian mixture models. On the other hand, this exponential choice of Φ does not improve on the bound for heavy tailed kernels such as Laplace location mixtures.

We show in this section that Orlicz-Wasserstein metrics provide strong theoretical guarantees for mixing measures. This raises the natural question as to how such a metric can be computed for arbitrary choices of Φ. We provide some guidance in that regard in the following section.

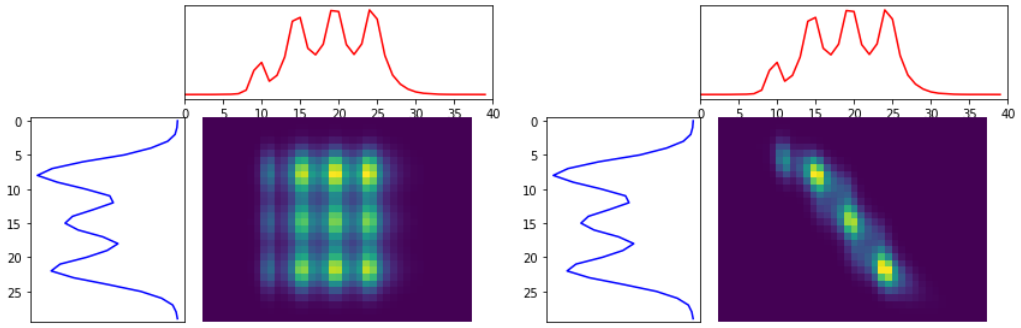
3.4. Computation of the Orlicz-Wasserstein

In practice, the Euclidean Wasserstein distance is computed for samples of the respective distributions. The exact computation turns out to be a linear programming problem which scales to the order of $O(n^3 \log(n))$, where n is the combined sample size of the two sampling distributions for which the distance is being calculated. (Cuturi, 2013) shows that using entropic regularization this can be drastically improved to $O(n^2)$ (Altschuler et al., 2017; Lin et al., 2019; 2022). Further speed-ups and easiness of computation via the use of dual formulation of the entropic regularization has been explored by the works of (Seguy et al., 2017; Genevay et al., 2016; Genevay, 2019). Here we consider the entropic regularized version of the Orlicz-Wasserstein metrics.

Computational procedure: In that respect, we consider solving the following problem as a surrogate to equation (10).

$$\begin{aligned} W^\lambda_\Phi(\nu_1, \nu_2) &:= \inf_{\nu \in \mathcal{Q}(\nu_1, \nu_2)} A_\Phi(\nu_1, \nu_2), \\ P^\lambda_\Phi(\nu_1, \nu_2) &:= \arg \inf_{\nu \in \mathcal{Q}(\nu_1, \nu_2)} A_\Phi(\nu_1, \nu_2), \end{aligned} \quad (17)$$

where $A_\Phi(\nu_1, \nu_2) := \inf\{\eta \in \mathbb{R}^+ : \int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x - y\|/\eta)\, d\nu(x, y) - (1/\lambda)(H(\nu)) \leq 1\}$ with $H(\mu)$ used to denote the Shannon entropy of distribution μ. To obtain solutions for equation (17), we resort to using outputs from



(a) Standard Wasserstein W_1 (b) Orlicz-Wasserstein with exponential Φ

Figure 1. Transportation plans. (a) Entropic OT produces more global plans and is unable to capture local structure of mass transfers. (b) Entropic Orlicz-Wasserstein penalizes mass transfers over large distances

Sinkhorn divergence computations.

Algorithm 1 Computing Orlicz Wasserstein distances between two discrete probability measures

Input M, $\lambda, \boldsymbol{r}, \boldsymbol{c}, \epsilon$.
Output $W_\Phi^\lambda(\nu_1, \nu_2)$.
I = $(\boldsymbol{r} > 0)$; $\boldsymbol{r} = \boldsymbol{r}(I)$; $M = M(I, :)$;
$x_{\text{upp}} = \max(M)/\Phi^{-1}(1)$,
$x_{\text{low}} = [S(M, \lambda, \boldsymbol{r}, \boldsymbol{c}) + \frac{1}{2\lambda}(H(\boldsymbol{r}) + H(\boldsymbol{c}))]/\Phi^{-1}(1 + \frac{1}{\lambda}(H(\boldsymbol{r}) + H(\boldsymbol{c}))$
$fx_{\text{upp}} = S(\Phi(M/x_{\text{upp}}), \lambda, \boldsymbol{r}, \boldsymbol{c}), fx_{\text{low}} = S(\Phi(M/x_{\text{low}}), \lambda, \boldsymbol{r}, \boldsymbol{c})$.
while $|x_{\text{low}} - x_{\text{upp}}| < \epsilon$ not converged do
$x_{\text{new}} = (x_{\text{low}} * fx_{\text{upp}} - x_{\text{upp}} * fx_{\text{low}})/(fx_{\text{upp}} - fx_{\text{low}}))$.
if $x_{\text{new}} < x_{\text{upp}}$ and $x_{\text{new}} > x_{\text{low}}$ do
$fx_{\text{new}} = S(\Phi(M/x_{\text{upp}}), \lambda, \boldsymbol{r}, \boldsymbol{c})$
if $fx_{\text{new}} < 1$, $x_{\text{upp}} = x_{\text{new}}$, $fx_{\text{upp}} = fx_{\text{new}}$.
else: $x_{\text{low}} = x_{\text{new}}$, $fx_{\text{low}} = fx_{\text{new}}$
end if
else $x_{\text{new}} = (x_{\text{low}} + x_{\text{upp}})/2$. repeat Step 9-12.
end if
end while
return $W_\Phi^\lambda(\nu_1, \nu_2) := x_{\text{upp}}$.

Consider two discrete probability measures, $\boldsymbol{r}$ (with m atoms, $\{x_i\}_{i=1}^m$) and $\boldsymbol{c}$ (with n atoms, $\{y_i\}_{i=1}^m$). Let $M_{n\times m}$ be a distance matrix such that $M_{ij} = c(x_i, y_j)$ for some cost function $c(\cdot, \cdot)$. Let $S(M, \lambda, r, c)$ be used to denote the Sinkhorn divergence optimized objective function for cost matrix M, regularization parameter λ and $d(M, \lambda, r, c) = \langle S(M, \lambda, r, c), M\rangle$ be used to denote the transport cost. Algorithm 1 defines a procedure to obtain a regularised Orlicz-Wasserstein distance between $\nu_1 = \sum_i r_i \delta_{x_i}$ and $\nu_2 = \sum_i c_i \delta_{y_i}$ in such a scenario by iteratively updating the value of Orlicz-Wasserstein distance until convergence. The crucial intuition behind Algorithm 1 is that $\inf_{\nu\in\mathcal{Q}(\nu_1,\nu_2)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x - y\|/\eta)\, d\nu(x, y) - (1/\lambda)(H(\nu))$ is a monotonically non-increasing function of η. Therefore the solution to the Orlicz Wasserstein distance can be obtained by a binary search once upper and lower limits are known. This is rigorously explained in Proposition 1 in Appendix C.

Simulations settings: We provide a demonstration of the utility of using Orlicz-Wassestein distances in Figure 1. We consider two mixing densities, ν_1 on the y-axis is a 3-mixture of univariate normal distributions with means at $[3, 4, 5]$, common $\sigma = 0.3$ and mixture weights $[0.37, 0.3, 0.33]$. On the other hand ν_2 represented in the x-axis is a 4-mixture of univariate Laplace kernels with means at $[7, 8, 9, 6]$, scale parameters $[0.3, 0.3, 0.3, 0.1]]$ and mixture weights $[0.30, 0.32, 0.32, 0.06]$. The left plot of Figure 1 shows the transportation plan for output of Sinkhorn mechanism with regularisation parameter 0.01, while the right plot shows the same for transportation plan obtained via Algorithm 1 with $\lambda = 0.01$ ($\Phi(\cdot) = \exp(\cdot/\beta) - 1$, $\beta = 1.1$). We have the following remarks.

Remark: The entropic Orlicz-Wasserstein procedure produces sharper transport plans. This indicates that it performs a shrinkage procedure on the space of transportation plans. This can have potential benefits towards obtaining robust plans and provide a promising direction of future research. Additionally, while entropic Euclidean Wasserstein transport plans distribute the mass of the outlier atom of ν_2 (mean=6, weight= 0.06), its Orlicz-Wasserstein counterpart manages to avoid it entirely. By penalizing mass transfers over large distances, Orlicz-Wasserstein distances are able to restrict attention to localised transportation plans. This in turn helps capture the small outlier mass associated with aposteriori DPGMM samples, as seen in Section 3.3.

4. Conclusion

In this work, we discuss the shortcomings of traditional Wasserstein metrics to perform clustering with Gaussian mixture models. We re-introduce a metric, called Orlicz-Wasserstein distances, with novel application to the context of estimating parameter convergence rates of hierarchical and mixture models and provide sound theoretical justifications of its ability to address the concerns associated with traditional Wasserstein distances. We also provide a theoretically sound approximate algorithm to compute the distance metric, and also show that convergence rates of Orlicz-Wasserstein distances carry over to the approximate distance. Lastly, we provide a preliminary simulation study to initiate a discussion on future research with Orlicz-Wasserstein distances. Since they allow low/high penalty on mass transfers over large distances, depending on the choice of function Φ, this lends flexibility to extending mass transfers over local/global regions and consequentially may be used as a device for smoothing/sharpening standard OT plans. Combined with dimension reduction techniques this can lend usage to a number of application domains such as anomaly detection and robust optimal transport.

Acknowledgements

This research is supported in part by grants NSF CAREER DMS-1351362, NSF CNS-1409303, a research gift from Adobe Research and a Margaret and Herman Sokol Faculty Award. Nhat Ho acknowledges support from the NSF IFML 2019844 and the NSF AI Institute for Foundations of Machine Learning.

References

Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 2013.

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Antoniak, C. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

Athey, T. L. and Vogelstein, J. T. Autogmm: Automatic gaussian mixture modeling in python. *CoRR*, abs/1909.02688, 2019. URL `http://arxiv.org/abs/1909.02688`.

Bailey, T. L., Elkan, C., et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

Banfield, J. and Raftery, A. Model based gaussian and non-gaussian clustering. *Biometrics*, 49:803821, 1993.

Barron, A., Schervish, M., and Wasserman, L. The consistency of posterior distributions in nonparametric problems. *Ann. Statist*, 27:536–561, 1999.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Pruenster, I., and Ruggiero, M. Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 212–229, 2015.

Escobar, M. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

Gao, F. and van der Vaart, A. W. Posterior contraction rates for deconvolution of dirichlet-laplace mixtures. *Electronic Journal of Statistics*, 10:608–627, 2016.

Genevay, A. *Entropy-Regularized Optimal Transport for Machine Learning. (Régularisation Entropique du Transport Optimal pour le Machine Learning)*. PhD thesis, PSL Research University, Paris, France, 2019. URL `https://tel.archives-ouvertes.fr/tel-02319318`.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic Optimization for Large-scale Optimal Transport. In NIPS (ed.), *NIPS 2016 - Thirtieth Annual Conference on Neural Information Processing System*, Proc. NIPS 2016, Barcelona, Spain, December 2016. URL `https://hal.archives-ouvertes.fr/hal-01321664`.

Ghosal, S. and van der Vaart, A. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29: 1233–1263, 2001.

Ghosal, S. and van der Vaart, A. Convergence rates of posterior distributions for noniid observations. *Ann. Statist.*, 35(1):192–223, 2007.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

Green, P. and Richardson, S. Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics*, 28:355–377, 2001.

Gu, Y., Liu, J., Xu, G., and Ying, Z. Hypothesis testing of the q-matrix. *Psychometrika*, pp. 515537, 2018.

Guha, A., Ho, N., and Nguyen., X. On posterior contraction of parameters and interpretability in bayesian mixture modeling. *Bernoulli*, pp. 2159–2188, 2021.

Heinrich, P. and Kahn, J. Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics*, 46(6A):2844 – 2870, 2018.

Ho, N. and Nguyen, X. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016a.

Ho, N. and Nguyen, X. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10 (1):271–307, 2016b.

Ji, Y., Wu, C., Liu, P., Wang, J., and Coombes, K. R. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, 02 2005.

Jiao, L. et al. Egmm: An evidential version of the gaussian mixture model for clustering. *Applied Soft Computing*, 129, 2022.

Kell, M. On interpolation and curvature via wasserstein geodesics. *Advances in Calculus of Variations*, 10(2):125–167, 2017. URL https://doi.org/10.1515/acv-2014-0040.

Kotz, S., Kozubowski, T. J., and Podgorski, K. *The Laplace distribution and generalizations*. Birkhauser Boston, Inc., Boston, MA., 2001.

Lin, T., Ho, N., and Jordan, M. On efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms. In *International Conference on Machine Learning*, pp. 3982–3991, 2019.

Lin, T., Ho, N., and Jordan, M. I. On the efficiency of entropic regularized algorithms for optimal transport. *Journal of Machine Learning Research (JMLR)*, 23:1–42, 2022.

Lindsay, B. *Mixture models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA, 1995.

Lo, A. On a class of bayesian nonparametric estimates : I. density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

MacEachern, S. Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science, American Statistical Association*, 1999.

MacEachern, S. and Muller, P. Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7:223–238, 1998.

Manole, T. and Ho, N. Refined convergence rates for maximum likelihood estimation under finite mixture models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14979–15006. PMLR, 17–23 Jul 2022.

Manole, T. and Khalili, A. Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The Annals of Statistics*, 49(6):3043 – 3069, 2021.

McLachlan, G. and Basford, K. *Mixture models: Inference and Applications to Clustering*. Marcel-Dekker, New York, 1988.

Mengersen, K. L., Robert, C., and Titterington, M. *Mixtures: Estimation and Applications*. Wiley, 2011.

Miller, J. and Harrison, M. Inconsistency of pitman-yor process mixtures for the number of components. *Journal of Machine Learning Research*, 15:3333–3370, 2014.

Miller, J. W. and Harrison, M. T. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113, 2018.

Nguyen, X. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1): 370–400, 2013.

Ohn, I. and Lin, L. Optimal bayesian estimation of gaussian mixtures with growing number of components. 2020. URL https://arxiv.org/abs/2007.09284.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 185:71–110, 1894.

Permuter, H., Francos, J., and Jermyn, I. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition*, 39(4):695–706, 2006. ISSN 0031-3203. Graph-based Representations.

Robert, C. *Mixtures of distributions: inference and estimation. In Markov Chain Monte Carlo in Practice (W. Gilks, S. Richardson and D. Spiegelhalter, eds.)*. Chapman and Hall, London., 1996.

Roeder, K. and Wasserman, L. Practical bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.*, 92:894–902., 1997.

Schlattmann, P. *Medical Applications of Finite Mixture Models*. Springer, 2009.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. 2017. URL https://arxiv.org/abs/1711.02283.

Stein, E. and Shakarchi, R. *Complex Analysis*. Princeton University Press, 2010.

Sturm, K.-T. Generalized orlicz spaces and wasserstein distances for convex-concave scale functions. 2011. URL https://arxiv.org/abs/1104.4223.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. Estimating cure rates from survival data. *Journal of the American Statistical Association*, 98(464):1063–1078, 2003. doi: 10.1198/016214503000000001007. PMID: 21151838.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Villani, C. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathemtical Sciences]*. Springer, Berlin, 2009.

Wellner, J. A. The bennett-orlicz norm. 2017. URL https://arxiv.org/abs/1703.01721.

Wong, W. H. and Shen, X. Probability inequalities for likelihood ratios and convergences of sieves mles. *Annals of Statistics*, 23:339–362, 1995.

Xie, F. and Xu, Y. Bayesian repulsive Gaussian mixture model. *arXiv:1703.09061v2 [stat.ME]*, 2017.

Supplement to "On Excess Mass Behavior in Gaussian Mixture Models with Orlicz-Wasserstein Distances"

In this supplementary material, we present proofs of key results in Appendix A and proofs of lemmas in Appendix B. We then provide theoretical guarantee for the algorithm to compute the entropic regularized Orlicz-Wasserstein in Appendix C.

A. Proofs of key results

Notation revisited For any function $g : \mathcal{X} \to \mathbb{R}$, we denote $\widetilde{g}(\omega)$ as the Fourier transformation of function g. Given two densities p, q (with respect to the Lebesgue measure μ), the squared Hellinger distance is given by $h^2(p,q) = (1/2)\int(\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)$. For any metric d on Θ, we define the open ball of d-radius ϵ around $\theta_0 \in \Theta$ as $B_d(\epsilon, \theta_0)$. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore, we denote A^c as the complement of set A for any set A while $B(x, r)$ denotes the ball, with respect to the l_2 norm, of radius $r > 0$ centered at $x \in \mathbb{R}^d$. The expression $D(\epsilon, \mathscr{P}, d)$ used in the paper denotes the ϵ-packing number of the space $\mathscr{P}$ relative to the metric d. d is replaced by h to denote the hellinger norm. Finally, we use $\mathrm{Diam}(\Theta) = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$ to denote the diameter of a given parameter space Θ relative to the l_2 norm, $\|\cdot\|$, for elements in $\mathbb{R}^d$. Regarding the space of mixing measures, let $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively denote the space of all mixing measures with exactly and at most k support points, all in Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. Moreover, $\overline{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ. Finally, $\mathcal{M}(\Theta)$ stands for the space of all probability measures on Θ.

A.1. Proof of Theorem 1

We present the proof of Theorem 1 for the lower bound of Hellinger distance between mixing density functions based on Orlicz-Wasserstein metric between their corresponding mixing measures.

In this proof, we denote $a \lesssim b$ to imply that $a \leq C \cdot b$ for a universal constant C dependent on α, d, and $\bar{\theta}$. Also, $f * g$ will denote the outcome of convolution operation on functions f and g. Now, we consider the following density function in $\mathbb{R}$:

$$K(x) := c\left(\int_{-\infty}^{\infty} \exp(-itx)\exp(-t^4)\mathrm{d}t\right)^2, \tag{18}$$

where c is a proportionality constant so that $\int_{-\infty}^{\infty} K(x)\mathrm{d}x = 1$. Lemma 6 shows that $K(\cdot)$ is integrable.

Moreover, Lemma 7 shows that the characteristic function $\hat{K}(\cdot)$, corresponding to $K(\cdot)$ satisfies,

$$|\hat{K}(x)| \lesssim \exp(-(x/2)^4).$$

The strategy to obtain upper bounds for $W_\Phi(G, G')$ is to convolve G with mollifiers, $Z_{\delta,d}(\cdot)$, of the form $Z_{\delta,d}(x) = \prod_{i=1}^d \frac{1}{\delta}K(x_i/\delta)$ for $\delta > 0$, where $x = (x_1, \ldots, x_d)$. In particular, by triangle inequality and following Lemma 1 we can write:

$$W_\Phi(G, G') \leq W_1(G, G * Z_{\delta,d}) + W_\Phi(G', G' * Z_{\delta,d}) + W_\Phi(G * Z_{\delta,d}, G' * Z_{\delta,d}).$$

For $\Phi(x) = \exp((7/32)x) - 1$, following Lemma 4 we find that

$$W_\Phi(G, G * Z_{\delta,d}) \leq C_\alpha \delta.$$

Therefore, we can write

$$W_\Phi(G, G') \leq 2C_\alpha\delta + W_\Phi(G * Z_{\delta,d}, G' * Z_{\delta,d}).$$

For every $M > 0$, we have

$$\begin{aligned} W_\Phi(G * Z_{\delta,d}, G' * Z_{\delta,d}) &\leq 2\inf\{\lambda \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/\lambda) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x \leq 1\} \\ &\leq 2\inf\{\lambda \in \mathbb{R}^+ : \ s_1 \leq 1/2 \text{ and } \ s_2 \leq 1/2\}, \\ &\leq 2\max\{\inf\{\lambda \in \mathbb{R}^+ : \ s_1 \leq 1/2\}, \inf\{\lambda \in \mathbb{R}^+ : \ s_2 \leq 1/2\}\}, \end{aligned} \tag{19}$$

with the first inequality following from Lemma 5 and the third inequality comes from the monotonicity of function Φ. Here, we denote

$$\begin{aligned} s_1 &= \int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/\lambda) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x, \\ s_2 &= \int\limits_{\|x\|_2 > M} \Phi(\|x\|/\lambda) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x. \end{aligned}$$

We now proceed to bound $T_1 = \inf\{\lambda \in \mathbb{R}^+ : \ s_1 \leq 1/2\}$ and $T_2 = \inf\{\lambda \in \mathbb{R}^+ : \ s_2 \leq 1/2\}$.

Bounding for T_1: Using Holder's inequality, we obtain

$$\begin{aligned} &\inf\{\lambda \in \mathbb{R}^+ : \int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/\lambda) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x \leq 1/2\} \\ \leq \ &\inf\{\lambda > 0 : \int\limits_{\|x\| \leq M} \exp((\|x\|/\lambda)^\beta) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x \leq 3/2\} \\ \leq \ &\inf\Big\{\lambda > 0 : \Big(\int\limits_{\|x\| \leq M} \exp((M/\lambda)^\beta)\mathrm{d}x\Big)^{1/2} \Big(\int\limits_{\|x\| \leq M} |(G - G') * Z_{\delta,d}(x)|^2\mathrm{d}x\Big)^{1/2} \leq 3/2\Big\} \\ \leq \ &\inf\Big\{\lambda > 0 : \frac{\pi^{d/4}}{\sqrt{(\frac{d}{2}+1)\Gamma(d/2)}} M^{d/2} \exp((M/\lambda)^\beta)\|(G - G') * Z_{\delta,d}(x)\|_2 \leq 3/2\Big\} \\ = \ &\frac{M}{(\log(c_d/(\|(G - G') * \zeta_{\delta,d}\|_2 M^{d/2})))^{1/\beta}}. \end{aligned} \tag{20}$$

Since f is Gaussian distribution, we have $\tilde{f}(\omega) \geq c_f \exp(-\alpha \sum_{i=1}^d \omega_i^2)$ for some $c_f, \alpha > 0$. Given that inequality, we find that

$$\begin{aligned} \|(G - G') * Z_{\delta,d}\|_2^2 &= \int |\widetilde{G} - \widetilde{G'}|^2(\omega)|\widetilde{K}_{\delta,d}(\omega)|^2 d\omega = \int |\widetilde{f}(\widetilde{G} - \widetilde{G'})|^2(\omega)\frac{|\widetilde{K}_{\delta,d}(\omega)|^2}{|\widetilde{f}(\omega)|^2} d\omega \\ &\leq \|p_G - p_{G'}\|_2^2 \sup_{\omega \in \mathbb{R}^d} \frac{|\widetilde{K}_{\delta,d}(\omega)|^2}{|\widetilde{f}(\omega)|^2} \\ &\leq 4\|f\|_\infty h^2(p_G, p_{G_0}) \sup_{\omega \in \mathbb{R}^d} \Big\{\frac{1}{c_f^2} \cdot \prod_{i=1}^d \exp(-\delta^4|\omega_i|^4)\exp(2\alpha|\omega_i|^2)\Big\}. \end{aligned}$$

By taking derivatives, we obtain the maximum as

$$\sup_{\omega_i \in \mathbb{R}} \Big\{\exp(-\delta^4|\omega_i|^4)\exp(2\alpha|\omega_i|^2)\Big\} = \exp(\alpha^2/\delta^4).$$

Plugging these results into equation (20) leads to

$$\begin{aligned} &\inf\{\lambda \in \mathbb{R}^+ : \int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/\lambda) \cdot |(G - G') * Z_{\delta,d}(x)|\mathrm{d}x \leq 1/2\} \\ &\qquad \leq \frac{M}{(\log(c/(h(p_G, p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2})))^{1/\beta}} \end{aligned} \tag{21}$$

for some universal constant c.

Bounding for T_2: For any $M > 0$, we denote

$$\begin{aligned} k' &= \inf\{\lambda \in \mathbb{R}^+ : \mathbb{E}_{X\sim(G-G')}(\Phi(\|X\|^{5/4}/\lambda M^{1/4}) \leq 1/2\}, \\ k'' &= \inf\{\lambda \in \mathbb{R}^+ : \mathbb{E}_{Y\sim Z_{\delta,d}}(\Phi(\|Y\|^{5/4}/\lambda M^{1/4}) \leq 1/2\}. \end{aligned} \tag{22}$$

Then, by the convexity of Φ we have

$$\inf\{\lambda \in \mathbb{R}^+ : \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}(\Phi(\|X+Y\|^{5/4}/\lambda M^{(1/4)}) \leq 1/2\} \leq 2^{1/4}(k'+k").$$

The above inequality is because of the following inequalities:

$$\begin{aligned} & \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}(\Phi(\|X+Y\|^{5/4}/2^{1/4}(k'+k")M^{(1/4)}) \\ \leq\ & \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}(\Phi(2^{1/4}(\|X\|^{5/4}+\|Y\|^{5/4})/2^{1/4}(k'+k")M^{(1/4)})) \\ =\ & \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}\left(\Phi\left(\left(k'\|X\|^{5/4}+\|Y\|^{5/4}\right)/(k'+k")M^{(1/4)}\right)\right) \\ \leq\ & \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}\left(\Phi\left(\frac{k'}{k'+k"}\left(\frac{\|X\|^{5/4}}{k'M^{(1/4)}}\right)+\frac{k"}{k'+k"}\left(\frac{\|Y\|^{5/4}}{k"M^{(1/4)}}\right)\right)\right) \\ \leq\ & \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}\frac{k'}{k'+k"}\Phi\left(\frac{\|X\|^{5/4}}{k'M^{(1/4)}}\right)+\frac{k"}{k'+k"}\Phi\left(\frac{\|Y\|^{5/4}}{k"M^{(1/4)}}\right) \leq \frac{1}{2}. \end{aligned} \tag{23}$$

The first inequality follows from $\|a+b\|^p \leq 2^{p-1}(\|a\|^p+\|b\|^p)$. The second last inequality follows from convexity of Φ and the final inequality follows from equation (22). Therefore, we obtain that

$$\begin{aligned} \inf\{\lambda \in \mathbb{R}^+ :\ & \int_{\|x\|_2>M} \Phi(\|x\|/\lambda)\cdot|(G-G')*Z_{\delta,d}(x)|\mathrm{d}x \leq 1/2\} \\ \leq\ & \inf\{\lambda \in \mathbb{R}^+ : \int_{\|x\|_2>M} \Phi(\|x\|^{5/4}/\lambda M^{(1/4)})\cdot|(G-G')*Z_{\delta,d}(x)|\mathrm{d}x \leq 1/2\} \\ \leq\ & \inf\{\lambda \in \mathbb{R}^+ : \mathbb{E}_{X\sim G-G', Y\sim Z_{\delta,d}}(\Phi(\|X+Y\|^{5/4}/\lambda M^{(1/4)}) \leq 1/2\} \\ \lesssim\ & \inf\{\lambda > 0 : \int_{\mathbb{R}^d} \exp((\|x\|^{5/4}/\lambda M^{1/4})^\beta)\cdot|(G-G')(x)|\mathrm{d}x \leq 3/2\} \\ +\ & \inf\{\lambda > 0 : \int_{\mathbb{R}^d} \exp((\|x\|^{5/4}/\lambda M^{1/4})^\beta)\cdot|Z_{\delta,d}(x)|\mathrm{d}x \leq 3/2\} \\ \lesssim\ & \frac{(d\bar{\theta})^{5/4}}{M^{1/4}} + C\delta^{5/4}/M^{1/4}, \end{aligned} \tag{24}$$

where $C = \inf\{\lambda > 0 : \int_{\mathbb{R}^d} \exp((\|x\|^{5/4}/\lambda)^\beta)\cdot|K_{1,d}(x)|\mathrm{d}x < \infty$ as $K_{1,d}(x) \sim O(\exp(-|x|^{4/3}))$ for large $|x|$, by Lemma 6. Hence, using these results we get

$$\begin{aligned} W_\Phi(G,G') &\lesssim \delta + \max\left\{\frac{(d\bar{\theta})^{5/4}}{M^{1/4}} + C\delta^{5/4}/M^{1/4}, \frac{M}{(\log(c/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2})))^{1/\beta}}\right\} \\ &\leq \delta + \frac{(d\bar{\theta})^{5/4}}{M^{1/4}} + C\delta^{5/4}/M^{1/4} + \frac{M}{(\log(c/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2})))^{1/\beta}}. \end{aligned} \tag{25}$$

Choosing $M = (\log(1/h(p_G,p_{G_0})))^{1/2}$ and $\delta = \dfrac{2\alpha^2}{\log(1/h(p_G,p_{G_0}))}$ in equation (25) we obtain,

$$\begin{aligned} W_\Phi(G,G') \lesssim (\log(1/h(p_G,p_{G_0})))^{-1} &+ \frac{(d\bar{\theta})^{5/4}}{(\log(1/h(p_G,p_{G_0})))^{1/8}} + \left(\frac{1}{\log(1/h(p_G,p_{G_0}))}\right)^{11/8} \\ &+ \left(\frac{1}{\log(c/h(p_G,p_{G_0})(\log(1/h(p_G,p_{G_0})))^{d/4})}\right)^{(1/\beta)-(1/2)} \end{aligned} \tag{26}$$

As a consequence, we obtain the conclusion of the theorem.

A.2. Proof of Theorem 2

The proof of this result follows by an application of Lemma 8, 9 and 10 in combination with Theorem 2.1 in (Ghosal et al., 2000). To facilitate the presentation, we break the proof into several steps.

Step 1: First we compute the contraction rate relative to the Hellinger metric, i.e., assume that

$$\frac{\bar{\theta}^d}{\epsilon_n^{d+2}} \log\left(\frac{\bar{\theta}}{\epsilon_n}\right) = o(n) \text{ and } n\epsilon_n^2 \to \infty.$$

Then we show that

$$\Pi_n(G \in \overline{\mathcal{G}}(\Theta) : h(p_G, p_{G_0}) \geq L\epsilon_n | X_1, \ldots, X_n) \overset{P_{G_0}}{\to} 0. \tag{27}$$

We apply Theorem 7.1 in (Ghosal et al., 2000), with $\epsilon = L\epsilon_n$ and $D(\epsilon) = \exp\left(c_1 \left(\frac{\bar{\theta}}{\sqrt{\lambda_{\min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}^2}{\lambda_{\min}\epsilon_n^2}\right)\right)$, where $L \geq 2$ is a large constant to be chosen later and c_1 is the constant in equation (43). Lemma 9 shows the validity of this choice of $D(\epsilon)$. Then there exists a test function ϕ_n that satisfies

$$\begin{aligned} P_{G_0}^n \phi_n &\leq \exp\left(c_1 \left(\frac{\bar{\theta}}{\sqrt{\lambda_{\min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}^2}{\lambda_{\min}\epsilon_n^2}\right)\right) \\ &\quad \times \exp(-KnL^2\epsilon_n^2) \frac{1}{1 - \exp(-KnL^2\epsilon_n^2)}, \\ \sup_{G \in \overline{\mathcal{G}}(\Theta): h(p_G, p_{G_0}) \geq L\epsilon_n} P_G^n(1 - \phi_n) &\leq \exp(-KnL^2\epsilon_n^2). \end{aligned} \tag{28}$$

Now, we have

$$\begin{aligned} &\mathbb{E}_{P_{G_0}} \Pi_n(G \in \overline{\mathcal{G}}(\Theta) : h(p_G, p_{G_0}) \geq L\epsilon_n | X_1, \ldots, X_n)\phi_n \\ &\quad \leq P_{G_0}^n \phi_n \leq 2\exp\left(c_1 \left(\frac{\bar{\theta}}{\sqrt{\lambda_{min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}^2}{\lambda_{min}\epsilon_n^2}\right) - KnL^2\epsilon_n^2\right). \end{aligned} \tag{29}$$

Based on computation with the posterior,

$$\begin{aligned} \Pi_n(G : h(p_G, p_{G_0} \geq \epsilon_n) | X_1, \ldots, X_n)(1 - \phi_n) &= \frac{\int_{G \in \overline{\mathcal{G}}(\Theta): h(p_G, p_{G_0}) \geq \epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1 - \phi_n)}{\int_{G \in \overline{\mathcal{G}}(\Theta)} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)} \\ &\leq \frac{\int_{G \in \overline{\mathcal{G}}(\Theta): h(p_G, p_{G_0}) \geq \epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1 - \phi_n)}{\int_{G \in \overline{\mathcal{G}}(\Theta): K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2 (\log(M/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)}, \end{aligned}$$

where $M = \exp(d\lambda_{min}^{-1}(5\bar{\theta}_0^2 + 4\bar{\theta}^2))$, with λ_{min} being the minimum eigenvalue of Σ.

Step 1.1: In this step we show that

$$\begin{aligned} &\int_{G \in \overline{\mathcal{G}}(\Theta): K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2 (\log(M/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G) \\ &\quad \gtrsim \exp(-(1 + C)n\lambda_{min}\epsilon_n^2) \frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^D}{(2\Gamma(d/2 + 1))^D (2D)^{D-1}} \left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}}\right)^{2(D-1)+dD} \\ &\qquad \text{with } p_{G_0}^n \text{ probability} \to 1, \end{aligned} \tag{30}$$

for all $C > 0$ and $\epsilon_n > 0$ is sufficiently small, where $D = D(\sqrt{\lambda_{min}}\epsilon_n, \Theta, \|.\|) \approx \left(\frac{\bar{\theta}}{\epsilon_n}\right)^d$ stands for the maximal $\sqrt{\lambda_{min}}\epsilon_n$-packing number for Θ under $\|.\|$ norm, and $\Gamma(\cdot)$ is the gamma function. First we show that

$$\{G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_0) \lesssim \sqrt{\lambda_{min}}\epsilon_n\} \\ \subset \{G \in \overline{\mathcal{G}}(\Theta) : K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n))^2\}, \tag{31}$$

for ϵ_n sufficiently small.

Since $\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x) \leq M$ by Lemma 10, it follows by an application of Theorem 5 in (Wong & Shen, 1995) that for $\epsilon_n < 1/2(1-e^{-1})^2$,

$$h(p_G, p_{G_0}) \lesssim \epsilon_n^2 \implies K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n))^2.$$

Following Example 1 in (Nguyen, 2013), $h^2(p_G, p_{G_0}) \leq \frac{W_2^2(G, G_0)}{8\lambda_{min}}$ for Gaussian location mixtures.

Similarly, from (Nguyen, 2013) it also follows that $K(p_G, p_{G_0}) \leq \frac{W_2^2(G, G_0)}{2\lambda_{min}}$. Combining the above displays, equation (31) follows.

Following Lemma 8.1 in (Ghosal et al., 2000), for every $C, \epsilon, M > 0$ and any measure Π on the set $\{G \in \overline{\mathcal{G}}(\Theta) : K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n))^2\}$, we have,

$$P_{G_0}^n\left(\int \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)}\mathrm{d}\Pi_n(G) \leq \exp(-(1+C)n\epsilon^2)\right) \leq \frac{1}{C^2 n\epsilon^2(\log(M/\epsilon))^2}. \tag{32}$$

The result in equation (30) now follows by an application of Lemma 8 in combination with equations (31) and (32) using the fact that $n\epsilon_n^2 \to \infty$.

Step 1.2: Let the event in (30) be denoted as T_n. Then

$$\begin{aligned} & \mathbb{E}_{P_{G_0}}\left[\Pi_n(G : h(p_G, p_{G_0}) \geq L\epsilon_n)|X_1, \ldots, X_n)(1-\phi_n)\right] \leq P_{G_0}(T_n^C) \\ + \quad & P_{G_0}(T_n)\frac{\exp((1+C)n\lambda_{min}\epsilon_n^2)}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^D}{(2\Gamma(d/2+1))^D(2D)^{D-1}}\left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}}\right)^{2(D-1)+dD}} \sup_{G \in \overline{\mathcal{G}}(\Theta): h(p_G, p_{G_0}) \geq L\epsilon_n} P_G^n(1-\phi_n) \\ \lesssim \quad & \frac{\exp((1+C)n\lambda_{min}\epsilon_n^2)}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^D}{(2\Gamma(d/2+1))^D(2D)^{D-1}}\left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}}\right)^{2(D-1)+dD}}\exp(-KnL^2\epsilon_n^2) + o(1). \end{aligned} \tag{33}$$

The final step follows from simple computation similar to that of the Proof of Theorem 2.1 in (Ghosal et al., 2000) and using the fact that $\frac{\bar{\theta}^d}{\epsilon_n^{d+2}}\log\left(\frac{\bar{\theta}}{\epsilon_n}\right) = o(n)$. Combining equations (29) and (33) and using the condition $\frac{\bar{\theta}^d}{\epsilon_n^{d+2}}\log\left(\frac{\bar{\theta}}{\epsilon_n}\right) = o(n)$, it follows that for L large enough

$$\Pi_n(G \in \overline{\mathcal{G}}(\Theta) : h(p_G, p_{G_0}) \geq L\epsilon_n|X_1, \ldots, X_n) \overset{P_{G_0}}{\to} 0. \tag{34}$$

Step 2: For some sufficiently large L with $\epsilon_n = L(\log n)n^{-1/(d+2)}$ satisfies $\frac{\bar{\theta}^d}{\epsilon_n^{d+2}}\log\left(\frac{\bar{\theta}}{\epsilon_n}\right) = o(n)$. Therefore we get, from the result in Step 1of this proof

$$\Pi_n\left(G \in \overline{\mathcal{G}}(\Theta) : h(p_G, p_{G_0}) \geq \frac{L(\log n)}{n^{1/(d+2)}} \,\middle|\, X_{1:n}\right) \overset{P_{G_0}^n}{\to} 0.$$

Now, from Theorem 1, we have

$$\Pi_n\left(G \in \overline{\mathcal{G}}(\Theta) : W_\Phi(G, G_0) \geq f_1(n, d) \,\middle|\, X_{1:n}\right) \overset{P_{G_0}^n}{\to} 0,$$

where $f_1(n, d) := (\log(n)/(d+2) - \log(\log n))^{-1/8}$.

A.3. Proof of Corollary 2

Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$, $G = \sum_{j=1}^{k} p_i \delta_{\theta_i}$. Suppose $\boldsymbol{q} = (q_{ij})_{1\leq i\leq k_0, 1\leq j\leq k} \in [0,1]^{k_0\times k}$ is a coupling between $\boldsymbol{p_0} = (p_1^0, \ldots, p_{k_0}^0)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$, with $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ represents the space of all such couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p}$. Using the proof technique similar to Lemma 3, we get

$$\begin{aligned}
&\sum q_{ij} \exp((\|\theta_i^0 - \theta_j\|/k)^\beta) \\
&\geq \sum q_{ij} \mathbb{1}_{\{\|\theta_i^0 - \theta_j\|\geq\eta\}} \exp((\eta/k)^\beta) \\
&\geq \sum p_j \mathbb{1}_{\{\|\theta_i^0 - \theta_j\|\geq\eta \text{ for all } i\}} \exp((\eta/k)^\beta),
\end{aligned}$$

for all $1 < \beta < 16/15$.

We denote $K = \inf\{\lambda \geq 0 : \sum p_j \mathbb{1}_{\{\|\theta_i^0 - \theta_j\|\geq\eta \text{ for all } i\}} \exp((\eta/\lambda)^\beta) \leq 2\}$. Then, we find that

$$\begin{aligned}
&K \geq \eta \left(\log\left(\frac{1}{\sum p_j \mathbb{1}_{\{\|\theta_i^0 - \theta_j\|\geq\eta \text{ for all } i\}}}\right)\right)^{-1/\beta}, \quad \text{and} \\
&\sum_j p_j \mathbb{1}_{\{\|\theta_j - \theta_i^0\|>\eta \text{ for all } i\}} \leq 2\exp\left(\frac{-\eta}{W_\Phi(G, G_0)}\right).
\end{aligned}$$

Putting these results together with Theorem 2 leads to

$$\Pi_n\left(G \in \mathcal{EX}_\eta\left(\Theta, 2\exp\left(-\left(\frac{\eta\log(n)^{1/8}}{(d+2)}\right)^\beta\right)\right) \,\middle|\, X_{1:n}\right) \overset{P_{G_0}}{\to} 0$$

in $P_{G_0}^n$ probability. Since this result holds for all $1 < \beta < 16/15$, we obtain the conclusion.

B. Proofs for Lemmas

We now present the proofs for all lemmas in Section 3.

B.1. Proof of Lemma 1

We need to show the following properties of Orlicz-Wasserstein:

(i) $W_\Phi(\nu_1, \nu_2) = W_\Phi(\nu_2, \nu_1)$ for any probability measures ν_1, ν_2 on $(\mathbb{R}^d, \|\cdot\|)$.

(ii) $W_\Phi(\mu, \mu) = 0$ for any probability measure μ on $(\mathbb{R}^d, \|\cdot\|)$.

(iii) $W_\Phi(\nu_1, \nu_2) \leq W_\Phi(\nu_1, \nu_3) + W_\Phi(\nu_3, \nu_2)$ for any probability measures ν_1, ν_2, ν_3 on $(\mathbb{R}^d, \|\cdot\|)$.

(i) follows easily from the fact $\|x - y\| =$ is symmetric with respect to $x, y \in \mathbb{R}^d$.

For (ii) consider the coupling, $\nu(x, y) = \mu(x)\mathbb{1}_{x=y}$, then it is clear to see that for any $k > 0$, $\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x - y\|/k)\, d\nu(x, y) = 0$ and therefore $W_\Phi(\mu, \mu) = 0$.

For part (iii), assume that $W_\Phi(\nu_1, \nu_3) = k_1, W_\Phi(\nu_3, \nu_2) = k_2$. Then, it is enough to show that there exists a coupling ν of ν_1 and ν_2 such that $\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x - y\|/(k_1 + k_2))\, \mathrm{d}\nu(x, y) \leq 1$.

By results from (Villani, 2003; 2009), there exists a coupling μ_1 of ν_1 and ν_3 and a coupling μ_2 of ν_2 and ν_3 such that,

$$\begin{aligned}
\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x - z\|/k_1)\, d\mu_1(x, z) &\leq 1 \\
\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|z - y\|/k_2)\, d\mu_2(y, z) &\leq 1. \qquad (35)
\end{aligned}$$

Then, by a result in probability theory there exists a probability measure μ on $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ such that

$$\begin{aligned} \int_{x\in\mathbb{R}^d} \mu(\mathrm{d}x, y, z) &= \mu_2(y, z) \\ \int_{x\in\mathbb{R}^d} \mu(x, \mathrm{d}y, z) &= \mu_1(x, z) \end{aligned} \tag{36}$$

Define $\nu(x, y) := \int_{z\in\mathbb{R}^d} \mu(x, y, \mathrm{d}z)$. Then, we obtain that

$$\begin{aligned} & \int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/(k_1+k_2))\,\mathrm{d}\nu(x,y) \\ = & \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/(k_1+k_2))\,\mathrm{d}\mu(x,y,z) \\ \leq & \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Phi((\|x-z\|+\|y-z\|)/(k_1+k_2))\,\mathrm{d}\mu(x,y,z) \\ \leq & \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Phi\left(\frac{k_1}{k_1+k_2}\frac{\|x-z\|}{k_1}+\frac{k_2}{k_1+k_2}\frac{\|y-z\|}{k_2}\right)\mathrm{d}\mu(x,y,z) \\ \leq & \frac{k_1}{k_1+k_2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi\left(\frac{\|x-z\|}{k_1}\right)\mathrm{d}\mu_1(x,z) \\ + & \frac{k_2}{k_1+k_2}\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi\left(\frac{\|y-z\|}{k_2}\right)\mathrm{d}\mu_2(y,z) \leq 1. \end{aligned}$$

The first inequality follows from the triangle inequality property of $\|\cdot\|$, while the last inequality follows from the convexity of Φ.

B.2. Proof of Lemma 2

Fix a coupling ν of ν_1 and ν_2. Consider λ satisfying

$$\begin{aligned} &\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/\lambda)\,d\nu(x,y) < \infty, \\ &\int_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/\lambda)\,d\nu(x,y) < \infty, \\ &\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/\lambda)\,d\nu(x,y) \leq \int_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/\lambda)\,d\nu(x,y), \end{aligned}$$

and thus, we find that

$$\begin{aligned} &\left\{\lambda : \int_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/\lambda)\,d\nu(x,y) \leq 1\right\} \\ &\quad\subset \left\{\lambda : \int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/\lambda)\,d\nu(x,y) \leq 1\right\}. \end{aligned}$$

As a consequence, we obtain the conclusion of Lemma 2 since infimum of a set is smaller than the infimum of its subset.

B.3. Proof of Lemma 4

Consider $X \sim \nu_1$ and $Y \sim Z_{\delta,d}$. Let K be such that

$$\int_{\mathbb{R}} \exp((7/32)|y_i/K|^{\alpha} - (7/16)|y_i/\delta|^{4/3})\mathrm{d}y_i < \infty.$$

Then, we find that

$$\begin{aligned}
&\inf_{\mu}\left\{\int_{\mathbb{R}^d\times\mathbb{R}^d}\Phi(\|x-y\|/\lambda)\,d\mu(x,y):\mu\in\mathcal{Q}(\nu_1,\nu_2)\right\}\\
\leq\;&\left(\frac{1}{\delta}\right)^d\int_{\mathbb{R}^d}\exp((7/32)\|y\|^\alpha/\lambda^\alpha)\prod_{i=1}^d K_1(y_i/\delta)\prod_{i=1}^d \mathrm{d}y_i-1\\
\leq\;&\prod_{i=1}^d\left(\frac{1}{\delta}\right)\int_{\mathbb{R}}\exp((7/32)|y_i|^\alpha/\lambda^\alpha)K_1(y_i/\delta)\mathrm{d}y_i-1\\
=\;&\prod_{i=1}^d\left(\frac{1}{\delta}\right)\int_{\mathbb{R}}\phi(y_i)^2\exp((7/32)|y_i/\lambda|^\alpha-(7/16)|y_i/\delta|^{4/3})\mathrm{d}y_i-1,
\end{aligned}$$

where $\phi(\cdot)$ is the function in Lemma 6. The second inequality follows from the fact that $\|x\|_p\leq\|x\|_q$ when $p\geq q$, where $\|\cdot\|_p$ is the L_p norm. The final equality follows from Lemma 6. Now, as $|\phi(x)|\leq C_\phi$ for some constant $C_\phi<\infty$, we have following the result in Lemma 2,

$$W_\Phi(\nu_1,\nu_2)\leq C_\alpha\delta$$

where

$$C_\alpha=\inf\left\{k>0:\int_{\mathbb{R}}\exp(|y/k|^\alpha-|y|^{4/3})\mathrm{d}y-1\leq\frac{1}{C_\phi^2}\right\}.$$

Note that, C_α as defined above exists because $\alpha\leq 4/3$. As a consequence, we obtain the conclusion of the lemma.

B.4. Proof of Lemma 5

Consider a coupling, ν between ν_1 and ν_2 that keeps fixed all the mass shared between ν_1 and ν_2, and redistributes the remaining mass independently, i.e.,

$$\nu(x,y)=(\nu_1(x)\bigwedge\nu_2(y))\mathbb{1}_{x=y}+\frac{1}{(\nu_1-\nu_2)_+(\mathbb{R}^d)}(\nu_1(x)-\nu_2(x))_+(\nu_2(y)-\nu_1(y))_+ \tag{37}$$

Let k_0 be defined as

$$k_0:=\inf\{k\in\mathbb{R}^+:\int_{\mathbb{R}^d}\Phi(\|x\|/k)\,d|\nu_1(x)-\nu_2(x)|\leq 1\}. \tag{38}$$

Then, using ν as defined in the above display we get

$$\begin{aligned}
\int_{\mathbb{R}^d\times\mathbb{R}^d}\Phi(\|x-y\|/2k_0)\,d\nu(x,y)&=\int_{\mathbb{R}^d\times\mathbb{R}^d}\Phi(\|x-y\|/2k_0)\cdot\frac{1}{(\nu_1-\nu_2)_+(\mathbb{R}^d)}(\nu_1(x)-\nu_2(x))_+(\nu_2(y)-\nu_1(y))_+\\
&\leq\int_{\mathbb{R}^d\times\mathbb{R}^d}\Phi(\|x\|/k_0)(\nu_1(x)-\nu_2(x))_+\leq 1
\end{aligned}$$

Therefore,

$$W_\Phi(\nu_1,\nu_2)\leq 2\inf\{k\in\mathbb{R}^+:\int_{\mathbb{R}^d}\Phi(\|x\|/k)\,d|\nu_1(x)-\nu_2(x)|\leq 1\}.$$

As a consequence, we reach the conclusion of the lemma.

Lemma 6. *Let $f(x)=\exp(-x^4)$, and $\tilde{f}(t)=(1/2\pi)\int_{-\infty}^{\infty}\exp(-itx)f(x)\mathrm{d}x$. Then,*

$$|\tilde{f}(t)|\leq\phi(t)\exp(-7/32|t|^{4/3}), \tag{39}$$

where $\phi(t)$ is an absolutely bounded real-valued function.

Proof. Consider a rectangle on the complex plane, with vertices at $R, -R, R+i\zeta, -R+i\zeta$ respectively. Following Goursat's Theorem (Stein & Shakarchi, 2010) for integration along rectangular contours on the complex plane, the contour integral along a closed rectangle is 0.

Therefore,

$$\int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x + \int_{R}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x + \int_{-R+i\zeta}^{-R} \exp(-itx)f(x)\mathrm{d}x + \int_{R+i\zeta}^{-R+i\zeta} \exp(-itx)f(x)\mathrm{d}x = 0.$$

Now,

$$|\int_{R}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x| = |\int_{0}^{\zeta} \exp(itR - tx)f(R+ix)i\mathrm{d}x| \leq C\exp(-R^4) \to 0,$$

as $R \to \infty$. Similarly,

$$|\int_{-R+i\zeta}^{-R} \exp(-itx)f(x)\mathrm{d}x| \to 0,$$

as $R \to \infty$.

Therefore,

$$\lim_{R\to\infty} \int_{-R+i\zeta}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x = \lim_{R\to\infty} \int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x = 2\pi\tilde{f}(t).$$

Now,

$$\begin{aligned}
\lim_{R\to\infty} \int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x = 2\pi\tilde{f}(t) &= \lim_{R\to\infty} \int_{-R+i\zeta}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x \\
&= \lim_{R\to\infty} \int_{-R}^{R} \exp(it(x+i\zeta))f(x+i\zeta)\mathrm{d}x. \\
&= \lim_{R\to\infty} \int_{-R}^{R} \exp(-itx - t\zeta))\exp(-(x+i\zeta)^4)\mathrm{d}x.
\end{aligned}$$

Expanding the above expression,

$$\tilde{f}(t) = (1/2\pi)\lim_{R\to\infty} \textstyle\int_{-R}^{R} \exp(-itx - 4ix^3\zeta + 4ix\zeta^3 - t\zeta - (x^2 - 3\zeta^2)^2 + 8\zeta^4)\mathrm{d}x.$$

Substituting $\zeta = \dfrac{1}{4}\,\mathrm{sign}(t)|t|^{1/3}$ in the above equationa,

$$|\tilde{f}(t)| \leq (1/2\pi)\exp(-(7/32)|t|^{4/3}) \cdot \int_{-\infty}^{\infty} \exp(-(x^2 - (1/3)|t|^{1/2})^2)\mathrm{d}x. \tag{40}$$

The proof is complete when we note that $\phi(t) := (1/2\pi)\int_{-\infty}^{\infty} \exp(-(x^2 - (1/3)|t|^{1/2})^2)\mathrm{d}x$ is an absolutely bounded function. □

Lemma 7. *Let $k(t) = c\tilde{f}(t)^2$, where $\tilde{f}(t) = (1/2\pi)\int_{-\infty}^{\infty} \exp(-itx)\exp(-x^4)\mathrm{d}x$ and c is a constant of proportionality so that $\int_{-\infty}^{\infty} k(t)\mathrm{d}t = 1$. Then,*

$$|\int_{-\infty}^{\infty} \exp(itx)k(t)\mathrm{d}t| \lesssim \exp(-(x/2)^4) \tag{41}$$

Proof. Define $f(x) = \exp(-x^4)$. Then, by a version of the Fourier inversion theorem,

$$\int_{-\infty}^{\infty} \exp(itx)k(t)\mathrm{d}t = f * f(x),$$

where $*$ is the convolution operator. Since convolution of even functions is even, it is enough to show the result for $x > 0$. Then,

$$\begin{aligned} f * f(x) &= \int_{-\infty}^{\infty} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y \\ &= \int_{x/2}^{\infty} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y + \int_{-\infty}^{x/2} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y \\ &\leq \exp(-(x/2)^4)\int_{x/2}^{\infty} \exp(-(y-x)^4)\mathrm{d}y + \exp(-(x/2)^4)\int_{-\infty}^{x/2} \exp(-y^4)\mathrm{d}y \\ &\leq 2\exp(-(x/2)^4)\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y. \end{aligned}$$

The result holds with $C = 2\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y$ since $\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y < \infty$. □

B.5. Proof of Lemma 3

Suppose $\boldsymbol{q} = (q_{ij})_{1\leq i\leq k_0, 1\leq j\leq k} \in [0,1]^{k_0\times k}$ is a coupling between $\boldsymbol{p_0} = (p_1^0, \ldots, p_{k_0}^0)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$, with $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ representing the space of all such couplings of $\boldsymbol{p}$ and $\boldsymbol{p'}$. Then, for fixed k we have

$$\begin{aligned} \sum q_{ij}\Phi(\|\theta_i^0 - \theta_j\|/k) &\geq \sum q_{ij}\mathbb{1}_{\{\|\theta_i^0-\theta_j\|\geq\eta\}}\Phi(\eta/k) \\ &\geq \sum p_j\mathbb{1}_{\{\|\theta_i^0-\theta_j\|\geq\eta \text{ for all } i\}}\Phi(\eta/k). \end{aligned}$$

Let $K = \inf\{k \geq 0 : \sum p_j\mathbb{1}_{\{\|\theta_i^0-\theta_j\|\geq\eta \text{ for all } i\}}\Phi(\eta/k) \leq 1\}$. Then,

$$K \geq \eta\left(\Phi^{-1}\left(\frac{1}{\sum p_j\mathbb{1}_{\{\|\theta_i^0-\theta_j\|\geq\eta \text{ for all } i\}}}\right)\right)^{-1}, \tag{42}$$

where Φ^{-1} is the inverse function of the function Φ. Note that, this function exists and is concave as Φ is monotonic increasing and convex. Moreover, by Lemma 2(i), we would have that $W_\Phi(G, G_0) \geq K$, where,

$$W_\Phi(G, G_0) := \inf_{q\in\mathcal{Q}(\boldsymbol{p},\boldsymbol{p'})}\{\inf\{k \geq 0 : \sum q_{ij}\Phi(\|\theta_i^0 - \theta_j\|/k) \leq 1\}\}$$

Combining the results from equations (42) and (43) we obtain the conclusion of the lemma.

B.6. Prior mass on Wasserstein ball

Lemma 8. *Let $G \sim DP(\gamma, H_n)$. Fix $r \geq 1$. Assume $G_0 \in \mathcal{M}(\Theta)$, where $\Theta = [-\bar{\theta}, \bar{\theta}]^d$. If H_n admits condition (P.1), then the following holds*

$$\Pi\left(W_r^r(G, G_0) \leq (2^r+1)\epsilon^r\right) \geq \frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^D}{(2\Gamma(d/2+1))^D(2D)^{D-1}}\left(\frac{\epsilon}{2\bar{\theta}}\right)^{r(D-1)+dD}$$

for all ϵ sufficiently small so that $D(\epsilon, \Theta, \|.\|) > \gamma$.

Here, $D = D(\epsilon, \Theta, \|.\|)$ stands for the maximal ϵ-packing number for Θ under $\|.\|$ norm, and $\Gamma(\cdot)$ is the gamma function.

Proof. From Lemma 5 in (Nguyen, 2013),

$$\Pi\left(W_r^r(G, G_0) \leq (2^r+1)\epsilon^r\right) \geq \frac{\Gamma(\gamma)\gamma^D}{(2D)^{D-1}} \left(\frac{\epsilon}{\text{Diam}(\Theta)}\right)^{r(D-1)} \sup_S \prod_{i=1}^{D} H_n(S_i),$$

where, $S := (S_1, ..., S_D)$ denotes the D disjoint $\epsilon/2$-balls that form a maximal ϵ-packing of Θ. The supremum is taken over all such packings.

Now, $H_n(A) \geq \left(\frac{c_0}{\mu(\Theta)}\right)\mu(A)$. Moreover, $\prod_{i=1}^{D}\mu(S_i) \geq \left(\frac{(\sqrt{\pi}\epsilon)^d}{2\Gamma(d/2+1)}\right)^D$. Using this, we arrive at the result.

□

B.7. Metric entropy with Hellinger distance

Lemma 9. *Let G_0 be a discrete mixing measure with all its atoms in $\Theta = [-\tilde{\theta}, \tilde{\theta}]^d \subset \mathbb{R}^d$. Let $\mathscr{P}_{\overline{\mathcal{G}}(\Theta)} := \{p_G : G \in \overline{\mathcal{G}}(\Theta)\}$. Then, if the kernel f is multivariate Gaussian with covariance matrix Σ,*

$$\log D(\epsilon/2, \{p_G \in \mathscr{P}_{\overline{\mathcal{G}}(\Theta)} : \epsilon < h(p_G, p_{G_0}) \leq 2\epsilon\}, h) \leq c_1 \left(\frac{\tilde{\theta}}{\sqrt{\lambda_{min}}\epsilon}\right)^d \log\left(e + \frac{32e\tilde{\theta}^2}{\lambda_{min}\epsilon^2}\right) \tag{43}$$

for some universal constant c_1.

Proof. Let $N(\epsilon, \mathscr{P}, d)$ denote the ϵ-covering number of the space $\mathscr{P}$ relative to the metric d. It is related to the packing number by the following identity:

$$N(\epsilon, \mathscr{P}, h) \leq D(\epsilon, \mathscr{P}, d) \leq N(\epsilon/2, \mathscr{P}, h). \tag{44}$$

Using the result in Example 1 of (Nguyen, 2013), when $f_\Sigma(\cdot|\theta) \sim \mathcal{N}_d(\theta, \Sigma)$,

$$h^2(f_\Sigma(\cdot|\theta_i), f_\Sigma(\cdot|\theta'_j)) = 1 - \exp\left(-\frac{1}{8}\|\theta_i - \theta'_j\|^2_{\Sigma^{-1}}\right) \leq \frac{\|\theta_i - \theta'_j\|^2}{8\lambda_{min}}, \tag{45}$$

where $\|z\|_{\Sigma^{-1}} := \sqrt{z'\Sigma^{-1}z}$.

Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ and $G = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$ be mixing measures in $\overline{\mathcal{G}}(\Theta)$, with $k_0, k' \in [1, \infty]$. Let $\boldsymbol{q} = (q_{ij})_{1\leq i\leq k_0, 1\leq j\leq k'} \in [0,1]^{k_0\times k'}$ denote a coupling of $\boldsymbol{p^0}$ and $\boldsymbol{p'}$.

Using Lemma 2 of (Nguyen, 2013) with $\phi(x) = \frac{1}{2}(\sqrt{x}-1)^2$, gives us:

$$h^2(p_G, p_{G_0}) \leq \inf_{\boldsymbol{q}\in Q(\boldsymbol{p_0}, \boldsymbol{p'})} \sum_{i,j} q_{ij} \frac{\|\theta_i - \theta'_j\|^2}{8\lambda_{min}} = \frac{W_2(G, G_0)^2}{8\lambda_{min}}, \tag{46}$$

where $Q(\boldsymbol{p_0}, \boldsymbol{p'})$ is the set of all couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p'}$. Therefore, it immediately follows that:

$$\begin{aligned} & \log D(\epsilon/2, \{p_G \in \mathscr{P}_{\overline{\mathcal{G}}(\Theta)} : \epsilon < h(p_G, p_{G_0}) \leq 2\epsilon\}, h) \\ \leq\ & \log D(\sqrt{2\lambda_{min}}\epsilon, \{G : G \in \overline{\mathcal{G}}(\Theta)\}, W_2) \leq N\left(\sqrt{\frac{\lambda_{min}}{8}}\epsilon, \Theta, \|\cdot\|\right) \log\left(e + \frac{32e\tilde{\theta}^2}{\lambda_{min}\epsilon^2}\right). \end{aligned}$$

The last inequality follows by applying Eq. (44) followed by Lemma 4 part (b) of (Nguyen, 2013). The result then follows immediately.

□

B.8. Computation of M corresponding to KL ball

Lemma 10. *Let G be a discrete mixing measure with all its atoms in $\left[-\tilde{\theta}, \tilde{\theta}\right]^d$ for some $\tilde{\theta} > 0$. Furthermore, assume the atoms of G_0 lie in $\left[-\bar{\theta}, \bar{\theta}\right]^d$ where $\bar{\theta} > 0$ is given. Then, the following holds if the kernel f is multivariate Gaussian,*

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)} \mu(\mathrm{d}x) \leq \exp(d\lambda_{min}^{-1}(5\bar{\theta}^2 + 4\tilde{\theta}^2)). \tag{47}$$

Here μ is the Lebesgue measure on $\mathbb{R}^d$.

Proof. For the multivariate Gaussian kernel with covariance matrix Σ, similar to the multivariate Laplace case, using lemma 2 of (Nguyen, 2013) with $\phi(x) = \dfrac{1}{x}$, gives us:

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)} \mu(\mathrm{d}x) \leq \inf_{\boldsymbol{q} \in Q(\boldsymbol{p_0}, \boldsymbol{p'})} \sum_{i,j} q_{ij} \int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta_j')} \mu(\mathrm{d}x), \tag{48}$$

where $Q(\boldsymbol{p_0}, \boldsymbol{p'})$ is the set of all couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p'}$, and $f_\Sigma(\cdot|\theta)$ is the multivariate Gaussian kernel with covariance parameter Σ and mean parameter θ.

$$\begin{aligned} \int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta_j')} \mu(\mathrm{d}x) &= \int f_\Sigma(x|\theta_i^0) \exp\left(\frac{-\|x-\theta_i^0\|_{\Sigma^{-1}}^2 + \|x-\theta_j'\|_{\Sigma^{-1}}^2}{2}\right) \mu(\mathrm{d}x) \\ &= \int f_\Sigma(x|\theta_i^0) \exp\left(\frac{-\|\theta_j'-\theta_i^0\|_{\Sigma^{-1}}^2}{2} + \langle x-\theta_j', \Sigma^{-1}\theta_j' - \theta_i^0\rangle\right) \mu(\mathrm{d}x), \end{aligned} \tag{49}$$

where the second equality follows by simple calculation using $x - \theta_i^0 = (x - \theta_j') + (\theta_j' - \theta_i^0)$.

If $M_\Sigma(t|\theta)$ is the moment generating function of the Gaussian distribution with mean θ and covariance Σ, then

$$M_\Sigma(t|\theta) = \exp(\langle \theta, t\rangle + \frac{1}{2}\langle t, \Sigma t\rangle).$$

Using this result , we can rewrite Eq. (49) as

$$\int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta_j')} \mu(\mathrm{d}x) = \exp(\langle \theta_j' - \theta_i^0, \Sigma^{-1}\theta_i^0 + \theta_j'\rangle) \leq \exp(2d\lambda_{min}^{-1}(\tilde{\theta}+\bar{\theta})^2 + d\lambda_{min}^{-1}\bar{\theta}^2),$$

The bound on $\int (p_{G_0}(x))^2/p_G(x)\mu(\mathrm{d}x)$ then follows immediately. □

C. Theoretical guarantee of Algorithm 1

We show in this section that the output of Algorithm 1 converges to the Entropy regularised version of the Orlicz-Wasserstein distance in equation (17).

Proposition 1. *Let $\hat{W}_\Phi^\lambda(\nu_1, \nu_2)$ be the output of Algorithm 1 and $W_\Phi^\lambda(\nu_1, \nu_2)$ be as in equation* (17)*. Then*

$$|\hat{W}_\Phi^\lambda(\nu_1, \nu_2) - W_\Phi^\lambda(\nu_1, \nu_2)| < \epsilon. \tag{50}$$

Proof. Here M is the cost matrix such that $M_{ij} = \|x_i - y_j\|$. Note that $S(\Phi(M/W_\Phi^\lambda(\nu_1, \nu_2)), \lambda, r, c) < 1$ and if $S(\Phi(M/\eta), \lambda, r, c) < 1$, then $\eta < W_\Phi^\lambda(\nu_1, \nu_2))$.

Therefore, if $\hat{x}_{upp} = max(M)/\Phi^{-1}(1)$, $\hat{x}_{low} = d(M, \lambda, r, c)/\Phi^{-1}(1 + d(M, \lambda, r, c) - S(M, \lambda, r, c))$ it is enough to show that $fx_{upp} = S(\Phi(M/\hat{x}_{upp}), \lambda, r, c) < 1$,$fx_{low} = S(\Phi(M/\hat{x}_{low}), \lambda, r, c) > 1$, since it would imply $x_{upp} := \hat{W}_\Phi^\lambda(\nu_1, \nu_2) < W_\Phi^\lambda(\nu_1, \nu_2) < x_{low}$ and therefore if $|x_{upp} - x_{low}| < \epsilon$, the result holds directly.

We need to show

(i) $S(\Phi(M/\hat{x}_{upp}), \lambda, r, c) < 1$.

(ii) $S(\Phi(M/\hat{x}_{low}), \lambda, r, c) > 1$.

For (i), observe that

$$S(\Phi(M/\hat{x}_{upp}), \lambda, r, c) = \inf_{\nu \in \mathcal{Q}(\nu_1, \nu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x-y\|/\hat{x}_{upp})\, d\nu(x,y) - (1/\lambda)(H(\nu)) \tag{51}$$

$$= \inf_{\nu \in \mathcal{Q}(\nu_1, \nu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\Phi^{-1}(1)\|x-y\|/max(M))\, d\nu(x,y) - (1/\lambda)(H(\nu)) \leq 1 \tag{52}$$

The last inequality holds by monotonicity of Φ combined with $\|x-y\|/max(M) < 1$ with ν-probability 1, and the fact that $H(\nu) > 0$.

For (ii),note that for any $\nu \in \mathcal{Q}(\nu_1, \nu_2)$, it holds that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x-y\|/\eta)\, d\nu(x,y) - H(\nu)/\lambda \geq \Phi\left(\int_{\mathbb{R}^d \times \mathbb{R}^d} (\|x-y\|/\eta)\, d\nu(x,y)\right) - (H(r)+H(c))/\lambda \tag{53}$$

$$\geq \Phi((S(M,\lambda,r,c) + (H(r)+H(c))/2\lambda)/\eta)) - (H(r)+H(c))/\lambda. \tag{54}$$

Both the inequalities hold by monotonicty and convexity of Φ combined with the fact that $\forall \nu \in \mathcal{Q}(\nu_1, \nu_2)$, it holds that $H(r)+H(c) \geq H(\nu) \geq (H(r)+H(c))/2$.

Now $\Phi((S(M,\lambda,r,c)+(H(r)+H(c))/2\lambda)/\eta)) - (H(r)+H(c))/\lambda \geq 1$, for any $\eta \leq \hat{x}_{upp}$, this completes the proof. □

D. Estimation of number of components for mixing measures

In this section, we consider how Orlicz-Wasserstein distances could be used to improved estimation of the number of components with Gaussian mixture models. Gaussian Mixture models have been used for the purpose of clustering both historically (**?**) as well as in modern applications (Athey & Vogelstein, 2019; **?**; Jiao et al., 2022). From the Bayesian perspective, often used BNP priors for mixture models tend to overestimate the number of components drastically by producing multiple extraneous components around the "true" components (Miller & Harrison, 2014). This makes it difficult to estimate the number of components, where it may of interest (MacEachern & Muller, 1998; Green & Richardson, 2001).

Several recent works have explored the consistent estimation of the number of components with mixture models, both with in-processing (Manole & Khalili, 2021) and post-processing (Guha et al., 2021) techniques. However, while (Manole & Khalili, 2021) restricts attention to the overfit setting only, (Guha et al., 2021) requires the knowledge of explicit contraction rates of respective parameters in Wasserstein distances. As parameter convergence rates of Dirichlet Process Gaussian Mixture models are extremely slow (Nguyen, 2013), this would also affect the estimation of the number of components negatively. The procedures in both the works (Guha et al., 2021; Manole & Khalili, 2021) consist of two smaller steps, truncation of extraneous outlier atoms and merging of atoms which are close to the "true" atoms. The results in this work specifically, Theorem 2 provide a low threshold for truncating outlier atoms thereby eliminating outlier atoms more efficiently. Combined with an understanding of convergence behavior around the "true" atoms would allow efficient estimation of the number of components with Dirichlet Process Gaussian Mixture models.