Bernoulli **21**(1), 2015, 618–646
DOI: 10.3150/13-BEJ582

Posterior contraction of the population polytope in finite admixture models

XUANLONG NGUYEN

Department of Statistics, University of Michigan, 456 West Hall, Ann Arbor, MI 48109-1107, USA.
E-mail: xuanlong@umich.edu

We study the posterior contraction behavior of the latent population structure that arises in admixture models as the amount of data increases. We adopt the geometric view of admixture models – alternatively known as topic models – as a data generating mechanism for points randomly sampled from the interior of a (convex) population polytope, whose extreme points correspond to the population structure variables of interest. Rates of posterior contraction are established with respect to Hausdorff metric and a minimum matching Euclidean metric defined on polytopes. Tools developed include posterior asymptotics of hierarchical models and arguments from convex geometry.

Keywords: Bayesian asymptotics; convex geometry; convex polytope; Hausdorff metric; latent mixing measures; population structure; rates of convergence; topic simplex

1. Introduction

We study a class of hierarchical mixture models for categorical data known as the admixtures, which were independently developed in the landmark papers by Pritchard, Stephens and Donnelly [13] and Blei, Ng and Jordan [4]. The former set of authors applied their modeling to population genetics, while the latter considered applications in text processing and computer vision, where their models are more widely known as the *latent Dirichlet allocation* model, or a topic model. Admixture modeling has been applied to and extended in a vast number of fields of engineering and sciences – in fact, the Google scholar pages for these two original papers alone combine for more than a dozen thousands of citations. In spite of their wide uses, asymptotic behavior of hierarchical models such as the admixtures remains largely unexplored, to the best of our knowledge.

A finite admixture model posits that there are k populations, each of which is characterized by a Δ^d-valued vector $\boldsymbol{\theta}_j$ of frequencies for generating a set of discrete values $\{0,1,\ldots,d\}$, for $j=1,\ldots,k$. Here, Δ^d is the d-dimensional probability simplex. A sampled individual may have mixed ancestry and as a result inherits some fraction of its

This is an electronic reprint of the original article published by the ISI/BS in *Bernoulli*, 2015, Vol. 21, No. 1, 618–646. This reprint differs from the original in pagination and typographic detail.

values from each of its ancestral populations. Thus, an individual is associated with a proportion vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k) \in \Delta^{k-1}$, where β_j denotes the proportion of the individual's data that are generated according to population j's frequency vector $\boldsymbol{\theta}_j$. This yields a vector of frequencies $\boldsymbol{\eta} = \sum_{j=1}^{k} \beta_j \boldsymbol{\theta}_j \in \Delta^d$ associated with that individual. In most applications, one does not observe $\boldsymbol{\eta}$ directly, but rather an i.i.d. sample generated from a multinomial distribution parameterized by $\boldsymbol{\eta}$. The collection of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ is referred to as the *population structure* in the admixture. In population genetics modeling, $\boldsymbol{\theta}_j$ represents the allele frequencies at each locus in an individual's genome from the jth population. In text document modeling, $\boldsymbol{\theta}_j$ represents the frequencies of words generated by the jth topic, while an individual is a document, that is, a collection of words. In computer vision, $\boldsymbol{\theta}_j$ represents the frequencies of objects generated by the jth scenary topic, while an individual is a natural image, that is, a collection of scenary objects. The primary interest is the inference of the population structure on the basis of sampled data. In a Bayesian estimation setting, the population structure is assumed random and endowed with a prior distribution – accordingly one is interested in the behavior of the posterior distribution of the population structure given the available data.

The goal of this paper is to obtain contraction rates of the posterior distribution of the latent population structure that arises in admixture models, as the amount of data increases. Admixture models present a canonical mixture model for categorical data in which the population structure provides the support for the mixing measure. Existing works on convergence behavior of mixing measures in a mixture model are quite rare, in either frequentist or Bayesian estimation literature. Chen provided the optimal convergence rate of mixing measures in several finite mixtures for univariate data [5] (see also [10]). Recent progress on multivariate mixture models include papers by Rousseau and Mengersen [14] and Nguyen [12]. In [12], posterior contraction rates of mixing measures in several finite and infinite mixture models for multivariate and continuous data were obtained. Toussile and Gassiat established consistency of a penalized MLE procedure for a finite admixture model [17]. This issue has also attracted increased attention in machine learning. Recent papers by Arora *et al.* [2] and Anandkumar *et al.* [1] study convergence properties of certain computationally efficient learning algorithms based on matrix factorization techniques.

There are a number of questions that arise in the convergence analysis of admixture models for categorical data. The first question is to find a suitable metric in order to establish rates of convergence. It would be ideal to establish convergence for each individual element $\boldsymbol{\theta}_i$, for $i = 1, \ldots, k$. This is a challenging task due to the problems of identifiability. A (relatively) minor issue is known as "label-switching" problem. That is, one can identify the collection of $\boldsymbol{\theta}_i$'s only up to a permutation. A deeper problem is that any $\boldsymbol{\theta}_j$ that can be expressed as a convex combination of the others $\boldsymbol{\theta}_{j'}$ for $j' \neq j$ may be difficult to identify, estimate, and analyze. To get around this difficulty, we propose to study the convergence of population structure variables through its convex hull $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, which shall be referred to as the *population polytope*. Convergence of convex polytopes can be evaluated in terms of Hausdorff metric $d_{\mathcal{H}}$, a metric commonly utilized in convex geometry [15]. Moreover, under some geometric identifiability conditions, it can be shown that convergence in Hausdorff metric entails convergence of all extreme points of the polytope via a minimum-matching distance metric (defined in Section 2). This is the theory we aim for in this paper. Note however that in a typical setting of topic modeling where $d \geq k$, all population structure variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ in general positions in Δ^d *are* extreme points of the population polytope. Thus, in this setting, convergence in Hausdorff metric entails convergence of the population structure variables up to a permutation of labels. Convergence behavior of (the posterior of) non-extreme points among $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$, when $k > d$, remains elusive as of this writing.

The second question in an asymptotic study of a hierarchical model is how to address multiple quantities that define the amount of empirical data. The admixture model we consider has two asymptotic quantities that play asymmetric roles – m is the number of individuals, and n is the number of data points associated with each individual. Both m and n are allowed to increase to infinity. A simple way to think about this asymptotic setting is to let m go to infinity, while $n := n(m)$ tends to infinity at a certain rate which may be constrained with respect to m. Let Π be a prior distribution on variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$. The goal is to derive a vanishing sequence of $\delta_{m,n}$, depending on both m and n, such that the posterior distribution of the $\boldsymbol{\theta}_i$'s satisfies, for some sufficiently large constant C,

$$\Pi(d_{\mathcal{H}}(G, G_0) \geq C\delta_{m,n} | \mathcal{S}_{[n]}^{[m]}) \to 0$$

in $P^m_{\mathcal{S}_{[n]}|G_0}$-probability as $m \to \infty$ and $n = n(m) \to \infty$ suitably. Here, $P^m_{\mathcal{S}_{[n]}|G_0}$ denotes the true distribution associated with population polytope G_0 that generates a given $m \times n$ data set $\mathcal{S}_{[n]}^{[m]}$. As mentioned, $\delta_{m,n}$ is also the posterior contraction rate for the extreme points among population structure variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$.

Overview of results

Suppose that $n \to \infty$ at a rate constrained by $\log m < n$ and $\log n = o(m)$. In an overfitted setting, that is, when the true population polytope may have less than k extreme points, we show that under some mild identifiability conditions the posterior contraction rate in either Hausdorff or minimum-matching distance metric is $\delta_{m,n} \asymp [\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}]^{1/(2(p+\alpha))}$, where $p = (k-1) \wedge d$ is the intrinsic dimension of the population polytope while α denotes the regularity level near boundary of the support of the density function for $\boldsymbol{\eta}$ (to be defined in sequel). On the other hand, if either the true population polytope is known to have exactly k extreme points, or if the pairwise distances among the extreme points are bounded from below by a known positive constant, then the contraction rate is improved to a parametric rate $\delta_{m,n} \asymp [\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}]^{1/(2(1+\alpha))}$.

The constraints on $n = n(m)$, and the appearance of quantity $\log n/m$ in the convergence rate are quite interesting. Both the constraints and the derived rate are rooted in a condition on the required thickness of the prior support of the marginal densities of the data and an upper bound on the entropy of the space of such densities. This suggests an interesting interaction between layers in the latent hierarchy of the admixture model worthy of further investigation. For instance, it is not clear whether posterior consistency continues to hold if n falls outside of the specified range, and what effects this has on convergence rates, with or without additional assumptions on the data. This appears quite difficult with our present set of techniques.

We also establish minimax lower bounds for both settings. In the overfitted setting, the obtained lower bound is $(mn)^{-1/(q+\alpha')}$, where $q = \lfloor k/2 \rfloor \wedge d$, and α' is a non-negative constant to be defined in the sequel that satisfies $\alpha' \leq \alpha$. This lower bound can be strengthened with additional conditions on the model. Although this lower bound does not match exactly with the posterior contraction rate, both are notably nonparametrics-like for depending on dimensionality d and on k. In particular, if $n \asymp m$, and $k \geq 2d$, the posterior contraction rate becomes $(\log m/m)^{-1/(2(d+\alpha))}$. Compare this to the lower bound $m^{-2/(d+\alpha')}$, whose exponent differs approximately by only a factor of 4 for large d.

Method of proofs and tools

The general framework of posterior asymptotics for density estimation has been well-established [7, 16] (see also [3, 8, 9, 19, 20]). This framework continues to be very useful, but the analysis of mixing measure estimation in multi-level models presents distinct new challenges. In Section 4, we shall formulate an abstract theorem (Theorem 4) on posterior contraction of latent variables of interest in an admixture model, given $m \times n$ data, by reposing on the framework of [7] (see also [12]). The main novelty here is that we work on the space of latent variables (e.g., space of latent population structures endowed with Hausdorff or a comparable metric) as opposed to the space of data densities endowed with Hellinger metric. A basic quantity is the *Hellinger information* of the Hausdorff metric for a given subset of polytopes. Indeed, the Hellinger information is a fundamental quantity running through the analysis, which ties together the amount of data m and n – key quantities that are associated with different levels in the model hierarchy.

The bulk of the paper is devoted to establishing properties of the Hellinger information, which are fed into Theorem 4 so as to obtain concrete convergence rates. This is achieved through a number of inequalities which illuminate the relationship between Hausdorff distance of a given pair of population polytopes G, G', and divergence functionals (e.g., Kullback–Leibler divergence or variational distance) of the induced marginal data densities. The technical challenges lie in the fact that in order to relate G to the marginal density of the data, one has to integrate out multiple layers of latent variables, $\boldsymbol{\eta}$ and $\boldsymbol{\beta}$. Techniques in convex geometry come in very handily in the derivation of both lower and upper bounds [15].

The remainder of the paper is organized as follows. The model and main results are described in Section 2. Section 3 describes the basic geometric assumptions and their consequences. An abstract theorem for posterior contraction for $m \times n$ data setting is formulated in Section 4, whose conditions are verified in the subsequent sections. Section 5 presents inequalities for Hausdorff distances which result in lower bounds on the Hellinger information, while Section 6 provides a lower bound on Kullback–Leibler neighborhoods of the prior support (that is, a bound the prior thickness). Proofs of main theorems and other technical lemmas are presented in Section 7 and the Appendices.

Notations

$B_p(\boldsymbol{\theta}, r)$ denotes a closed p-dimensional Euclidean ball centered at $\boldsymbol{\theta}$ and has radius r. G_ε denotes the Minkowsky sum $G_\varepsilon := G + B_{d+1}(\mathbf{0}, \varepsilon)$. $\operatorname{bd} G, \operatorname{extr} G, \operatorname{Diam} G, \operatorname{aff} G, \operatorname{vol}_p G$ denote the boundary, the set of extreme points, the diameter, the affine span, and the p-dimensional volume of set G, respectively. "Extreme points" and "vertices" are interchangeable throughout this paper. We define the dimension of a convex polytope to be the dimension of its affine hull. It is a well-known fact that if a polytope has k extreme points in general positions in Δ^d, then its dimension is $(k-1) \wedge d$. Set-theoretic difference between two sets is defined as $G \triangle G' = (G \setminus G') \cup (G' \setminus G)$. $N(\varepsilon, \mathcal{G}, d_{\mathcal{H}})$ denotes the covering number of $\mathcal{G}$ in Hausdorff metric $d_{\mathcal{H}}$. $D(\varepsilon, \mathcal{G}, d_{\mathcal{H}})$ is the packing number of $\mathcal{G}$ in Hausdorff metric. Several divergence measures for probability distributions are employed: $K(p,q), h(p,q), V(p,q)$ denote Kullback–Leibler divergence, Hellinger and total variation distance between two densities p and q defined with respect to a measure on a common space: $K(p,q) = \int p \log(p/q)$, $h^2(p,q) = \frac{1}{2}\int(\sqrt{p} - \sqrt{q})^2$ and $V(P,Q) = \frac{1}{2}\int |p-q|$. In addition, we define $K_2 = \int p[\log(p/q)]^2$. Throughout the paper, $f(m,n,\varepsilon) \lesssim g(m,n,\varepsilon)$, equivalently, $f = \mathrm{O}(g)$, means $f(m,n,\varepsilon) \le Cg(m,n,\varepsilon)$ for some constant C independent of asymptotic quantities m, n and ε – details about the dependence of C are made explicit unless obvious from the context. Similarly, $f(m,n,\varepsilon) \gtrsim g(m,n,\varepsilon)$ or $f = \Omega(g)$ means $f(m,n,\varepsilon) \ge Cg(m,n,\varepsilon)$.

2. Main results

Model description

As mentioned in the Introduction, the central objects of the admixture model are *population structure* variables $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, whose convex hull is called the *population polytope*: $G = \operatorname{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ reside in d-dimensional probability simplex Δ^d. $k < \infty$ is assumed known. Note that G has at most k vertices (i.e., extreme points) among $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$.

A random vector $\boldsymbol{\eta} \in G$ is parameterized by $\boldsymbol{\eta} = \beta_1\boldsymbol{\theta}_1 + \cdots + \beta_k\boldsymbol{\theta}_k$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k) \in \Delta^{k-1}$ is a random vector distributed according to a distribution $P_{\boldsymbol{\beta}|\gamma}$ for some parameter γ (both [13] and [4] used the Dirichlet distribution). Given $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$, this induces a probability distribution $P_{\boldsymbol{\eta}|G}$ whose support is the convex set G. Details of this distribution, suppressed for the time being, are given explicitly by Equations (14) and (15). [To be precise, $P_{\boldsymbol{\eta}|G}$ should be written as $P_{\boldsymbol{\eta}|\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k;G}$. That is, G is always attached with a specific set of $\boldsymbol{\theta}_j$'s. Throughout the paper, this specification of G is always understood but notationally suppressed to avoid cluttering.]

For each individual $i = 1, \ldots, m$, let $\boldsymbol{\eta}_i \in \Delta^d$ be an independent random vector distributed by $P_{\boldsymbol{\eta}|G}$. The observed data associated with i, $\mathcal{S}^i_{[n]} = (X_{ij})_{j=1}^n$ are assumed to be i.i.d. draws from the multinomial distribution $\operatorname{Mult}(\boldsymbol{\eta}_i)$ specified by $\boldsymbol{\eta}_i := (\eta_{i0}, \ldots, \eta_{id})$. That is, $X_{ij} \in \{0, \ldots, d\}$ such that $P(X_{ij} = l|\boldsymbol{\eta}_i) = \eta_{il}$ for $l = 0, \ldots, d$.

Admixture models are simple when specified in a hierarchical manner as given above. The relevant distributions are written down below. The joint distribution of the generic random variable $\boldsymbol{\eta}$ and n-vector $\mathcal{S}_{[n]}$ (dropping superscript i used for indexing a specific individual) is denoted by $P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}$ and its density $p_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}$. We have

$$p_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\boldsymbol{\eta}_i,\mathcal{S}^i_{[n]}) = p_{\boldsymbol{\eta}|G}(\boldsymbol{\eta}_i)\times\prod_{j=1}^{n}\prod_{l=0}^{d}\eta_{il}^{\mathbb{I}(X_{ij}=l)}. \tag{1}$$

The distribution of $\mathcal{S}_{[n]}$, denoted by $P_{\mathcal{S}_{[n]}|G}$, is obtained by integrating out $\boldsymbol{\eta}$, which yields the following density with respect to counting measure:

$$p_{\mathcal{S}_{[n]}|G}(\mathcal{S}^i_{[n]}) = \int_G \prod_{j=1}^{n}\prod_{l=0}^{d}\eta_{il}^{\mathbb{I}(X_{ij}=l)}\,\mathrm{d}P_{\boldsymbol{\eta}|G}(\boldsymbol{\eta}_i). \tag{2}$$

The joint distribution of the full data set $\mathcal{S}^{[m]}_{[n]} := (\mathcal{S}^i_{[n]})_{i=1}^m$, denoted by $P^m_{\mathcal{S}_{[n]}|G}$, is a product distribution:

$$P^m_{\mathcal{S}_{[n]}|G}(\mathcal{S}^{[m]}_{[n]}) := \prod_{i=1}^{m} P_{\mathcal{S}_{[n]}|G}(\mathcal{S}^i_{[n]}). \tag{3}$$

Admixture models are customarily introduced in an equivalent way as follows [4, 13]: For each $i = 1,\ldots,m$, draw an independent random variable $\boldsymbol{\beta}\in\Delta^{k-1}$ as $\boldsymbol{\beta}\sim P_{\boldsymbol{\beta}|\gamma}$. Given i and $\boldsymbol{\beta}$, for $j=1,\ldots,n$, draw $Z_{ij}|\boldsymbol{\beta}\overset{\text{i.i.d.}}{\sim}\mathrm{Mult}(\boldsymbol{\beta})$. Z_{ij} takes values in $\{1,\ldots,k\}$. Now, data point X_{ij} is randomly generated by $X_{ij}|Z_{ij}=l,\boldsymbol{\theta}\sim\mathrm{Mult}(\boldsymbol{\theta}_l)$. This yields the same joint distribution of $\mathcal{S}^i_{[n]} = (X_{ij})_{j=1}^n$ as the one described earlier. The use of latent variables Z_{ij} is amenable to the development of computational algorithms for inference. However, this representation bears no significance within the scope of this work.

Asymptotic setting and metrics on population polytopes

Assume the data set $\mathcal{S}^{[m]}_{[n]} = (\mathcal{S}^i_{[n]})_{i=1}^m$ of size $m\times n$ is generated according an admixture model given by "true" parameters $\boldsymbol{\theta}^*_1,\ldots,\boldsymbol{\theta}^*_k$. $G_0 = \mathrm{conv}(\boldsymbol{\theta}^*_1,\ldots,\boldsymbol{\theta}^*_k)$ is the true population polytope. Under the Bayesian estimation framework, the population structure variables $(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)$ are random and endowed with a prior distribution Π. The main question to be addressed in this paper is the contraction behavior of the posterior distribution $\Pi(G|\mathcal{S}^{[m]}_{[n]})$, as the number of data points $m\times n$ goes to infinity.

It is noted that we do not always assume that the number of extreme points of the population polytope G_0 is k. We work in a general overfitted setting where k only serves as the upper bound of the true number of extreme points for the purpose of model parameterization. The special case in which the number of extreme points of G_0 is known a priori is also interesting and will be considered.

Let $\operatorname{extr} G$ denote the set of extreme points of a given polytope G. $\mathcal{G}^k$ is the set of population polytopes in Δ^d such that $|\operatorname{extr} G| \leq k$. Let $\mathcal{G}^* = \bigcup_{2\leq k<\infty} \mathcal{G}^k$ be the set of population polytopes that have finite number of extreme points in Δ^d. A natural metric on $\mathcal{G}^*$ is the following "minimum-matching" Euclidean distance:

$$d_{\mathcal{M}}(G, G') = \max_{\boldsymbol{\theta}\in \operatorname{extr} G} \min_{\boldsymbol{\theta}'\in \operatorname{extr} G'} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \vee \max_{\boldsymbol{\theta}'\in \operatorname{extr} G'} \min_{\boldsymbol{\theta}\in \operatorname{extr} G} \|\boldsymbol{\theta}' - \boldsymbol{\theta}\|.$$

A more common metric is the Hausdorff metric:

$$d_{\mathcal{H}}(G, G') = \min\{\varepsilon \geq 0 | G \subset G'_\varepsilon; G' \subset G_\varepsilon\} = \max_{\boldsymbol{\theta}\in G} d(\boldsymbol{\theta}, G') \vee \max_{\boldsymbol{\theta}'\in G'} d(\boldsymbol{\theta}', G).$$

Here, $G_\varepsilon = G + B_{d+1}(\mathbf{0}, \varepsilon) := \{\boldsymbol{\theta} + e | \boldsymbol{\theta} \in G, e \in \mathbb{R}^{d+1}, \|e\| \leq 1\}$, and $d(\boldsymbol{\theta}, G') := \inf\{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \boldsymbol{\theta}' \in G'\}$. Observe that $d_{\mathcal{H}}$ depends on the boundary structure of sets, while $d_{\mathcal{M}}$ depends on only extreme points. In general, $d_{\mathcal{M}}$ dominates $d_{\mathcal{H}}$, but under additional mild assumptions the two metrics are equivalent (see Lemma 1).

We introduce a notion of regularity for a family probability distributions defined on convex polytopes $G \in \mathcal{G}^*$. This notion is concerned with the behavior near the boundary of the support of distributions $P_{\boldsymbol{\eta}|G}$. We say a family of distributions $\{P_{\boldsymbol{\eta}|G} | G \in \mathcal{G}^k\}$ is α-regular if for any $G \in \mathcal{G}^k$ and any $\boldsymbol{\eta}_0 \in \operatorname{bd} G$,

$$P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \varepsilon) \geq c\varepsilon^\alpha \operatorname{vol}_p(G \cap B_{d+1}(\boldsymbol{\eta}_0, \varepsilon)), \tag{4}$$

where p is the number of dimensions of the affine space $\operatorname{aff} G$ that spans G, constant $c > 0$ is independent of $G, \boldsymbol{\eta}_0$ and ε.

According to Lemma 4 (in Section 6) α-regularity holds for a range of α, when $P_{\boldsymbol{\beta}|\gamma}$ is a Dirichlet distribution, but there may be other choices. We will see that α plays an important role in characterize the rates of contraction for the posterior of the population polytope.

Assumptions. *Π is a prior distribution on $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ such that the following hold for the relevant parameters that reside in the support of Π:*

- (S0) *Geometric properties (A1) and (A2) listed in Section 3 are satisfied uniformly for all G.*
- (S1) *Each of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ is bounded away from the boundary of Δ^d. That is, if $\boldsymbol{\theta}_j = (\theta_{j,0}, \ldots, \theta_{j,d})$ then $\min_{l=0,\ldots,d} \theta_{j,l} > c_0$ for all $j = 1, \ldots, k$.*
- (S2) *For any small ε, $\Pi(\|\boldsymbol{\theta}_j - \boldsymbol{\theta}_j^*\| \leq \varepsilon \ \forall j = 1, \ldots, k) \geq c_0' \varepsilon^{kd}$, for some $c_0' > 0$.*
- (S3a) *$P_{\boldsymbol{\beta}}$ induces a family of distributions $\{P_{\boldsymbol{\eta}|G} | G \in \mathcal{G}^k\}$ that is α-regular.*
- (S3b) *$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ is distributed (a priori) according to a symmetric probability distribution $P_{\boldsymbol{\beta}}$ on Δ^{k-1}. That is, the random variables $\beta_1, \ldots, \beta_k$ are a priori exchangeable.*

Theorem 1. *Let $G_0 \in \mathcal{G}^k$ and G_0 is in the support of prior Π. Let $p = (k-1) \wedge d$. Under assumptions (S0)–(S3) of the admixture model, as $m \to \infty$ and $n \to \infty$ such as*

$\log\log m \le \log n = \mathrm{o}(m)$, *for some sufficiently large constant* C *independent of* m *and* n,

$$\Pi(d_{\mathcal{M}}(G_0, G) \ge C\delta_{m,n} | \mathcal{S}_{[n]}^{[m]}) \longrightarrow 0 \tag{5}$$

in $P^m_{\mathcal{S}_{[n]}|G_0}$*-probability. Here,*

$$\delta_{m,n} = \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}\right]^{1/(2(p+\alpha))}.$$

The same statement holds for the Hausdorff metric $d_{\mathcal{H}}$.

Remarks.

1. Geometric assumption (S0) and its consequences are presented in the next section. (S0), (S1) and (S2) are mild assumptions observed in practice (cf. [4, 13]).
2. The assumption in (S3b) that $P_{\boldsymbol{\beta}}$ is symmetric is relatively strong, but it has been widely adopted in practice (e.g., symmetric Dirichlet distributions, including the uniform distribution). This technical condition is not intrinsic to the theory, and is required only to establish an upper bound of the Kullback–Leibler distance in terms of Hausdorff distance. In fact, it may be replaced if such an upper bound can be established by some other means. See also the remark following the statement of Lemma 7.
3. In practice, $P_{\boldsymbol{\beta}}$ may be further parameterized as $P_{\boldsymbol{\beta}|\gamma}$, where γ is endowed with a prior distribution. Then, it would be of interest to also study the posterior contraction behavior for γ. In this paper, we have opted to focus only on convergence behavior of the population structure to simplify the exposition and the results.
4. The appearance of both m^{-1} and n^{-1} in the contraction rate suggests that if either m or n is small, the rate would suffer even if the total amount of data $m \times n$ increases. What is quite interesting is the appearance of $\log n/m$. This is rooted in an entropy condition (cf. Theorem 4 in Section 4), which requires an upper bound of the KL divergence in terms of Hausdorff distance. It is possible that the appearance of $\log n/m$ is due to our general proof technique of posterior contraction presented in Section 4. From a hierarchical modeling viewpoint, this result highlights an interesting interaction of sample sizes provided to different levels in the model hierarchy. This issue has not been widely discussed in the hierarchical modeling literature in a theoretical manner, to the best of our knowledge.
5. Note the constraints that $n > \log m$ and $\log n = \mathrm{o}(m)$ are required in order to obtain rates of posterior contraction. These constraints are related to the term $\log n/m$ mentioned above – they stem from the upper bound on Kullback–Leibler in Lemma 6. The remark following the statement of this lemma explains why the upper bound almost always grows with n. A very special situation is presented in Lemma 5 where an upper bound on Kullback–Leibler distance can be obtained that is independent of n. However, such a situation cannot be verified in any reasonable estimation setting. This suggests that with our proof technique, we almost always

require n to grow at a constrained rate relatively to m in order to obtain posterior contraction rates.

6. Constant α plays an important role in the rate exponent. Intuitively, the larger α is, the weaker the guaranteed probability mass accrued near the boundary of the population polytope, which implies less data observed for points located near the boundary. This entails a weaker guarantee on the rate of convergence. Indeed, a key step in the proof of the theorem is that Equation (4) enables us to transfer an upper bound on the diminishing variational distance between, say distributions $P_{\boldsymbol{\eta}|G}$ and $P_{\boldsymbol{\eta}|G'}$, to an upper bound on the Hausdorff distance between G and G', while incurring an extra term α in the exponent.
7. The exponent $\frac{1}{2(p+\alpha)}$ suggests a slow, nonparametric-like convergence rate. Moreover, later in Theorem 3 we show that this is qualitatively quite close to a minimax lower bound. On the other hand, the following theorem shows that it is possible to achieve a parametric rate if additional constraints are imposed on the true G_0 and/or the prior Π:

Theorem 2. *Let $G_0 \in \mathcal{G}^k$ and G_0 is in the support of prior* Π*. Assume* (S0)–(S3a), (S3b)*, and either one of the following two conditions hold:*

(a) $|\operatorname{extr} G_0| = k$*, or*
(b) *There is a known constant $r_0 > 0$ such that the pairwise distances of the extreme points of all G in the support of the prior are bounded from below by r_0.*

Then, as $m \to \infty$ and $n \to \infty$ such that $\log m < n$ and $\log n = o(m)$, Equation (5) holds with

$$\delta_{m,n} = \left[\frac{\log m}{m} + \frac{\log n}{n} + \frac{\log n}{m}\right]^{1/(2(1+\alpha))}.$$

The same statement holds for the Hausdorff metric $d_{\mathcal{H}}$.

The next theorem produces minimax lower bounds that are qualitatively quite similar to the nonparametric-like rates obtained in Theorem 1. In the following theorem, $\boldsymbol{\eta}$ is not parameterized by $\boldsymbol{\beta}$ and $\boldsymbol{\theta}_j$'s as in the admixture model. Instead, we shall simply replace assumptions (S3a) and (S3b) on $P_{\boldsymbol{\beta}|\gamma}$ by either one of the following assumptions on $P_{\boldsymbol{\eta}|G}$:

(S4) There is a non-negative constant α' such that for any pair of p-dimensional polytopes $G' \subset G$ that satisfy Property A1,

$$V(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G'}) \lesssim d_{\mathcal{H}}(G, G')^{\alpha'} \operatorname{vol}_p G \setminus G'.$$

(S4′) For any p-dimensional polytope G, $P_{\boldsymbol{\eta}|G}$ is the uniform distribution on G.

Note that the condition of α-regularity (cf. Equation (4)) implies that $\alpha \geq \alpha'$. In particular, if (S4′) is satisfied, then both (S3a)/(S3b) and (S4) hold with $\alpha = \alpha' = 0$.

Since a parameterization for $\boldsymbol{\eta}$ is not needed, the overall model can be simplified as follows: Given population polytope $G \in \Delta^d$, for each $i = 1, \ldots, m$, draw $\boldsymbol{\eta}_i \overset{\text{i.i.d.}}{\sim} P_{\boldsymbol{\eta}|G}$. For each $j = 1, \ldots, n$, draw $\mathcal{S}^i_{[n]} = (X_{ij})_{j=1}^n \overset{\text{i.i.d.}}{\sim} \text{Mult}(\boldsymbol{\eta}_i)$.

Theorem 3. *Suppose that $G_0 \in \mathcal{G}^k$ satisfies assumptions* (S0), (S1) *and* (S2). *Point estimates $\hat{G} = \hat{G}(\mathcal{S}^{[m]}_{[n]})$ take value in the set $\mathcal{G}^*$. In the following, the multiplying constants in $\gtrsim$ depend only on constants specified by these assumptions.*

(a) *Let $q = \lfloor k/2 \rfloor \wedge d$. Under assumption* (S4), *we have*

$$\inf_{\hat{G} \in \mathcal{G}^*} \sup_{G_0 \in \mathcal{G}^k} P^m_{\mathcal{S}_{[n]}|G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn}\right)^{1/(q+\alpha')}.$$

(b) *Let $q = \lfloor k/2 \rfloor \wedge d$. Under assumption* (S4′), *we have*

$$\inf_{\hat{G} \in \mathcal{G}^*} \sup_{G_0 \in \mathcal{G}^k} P^m_{\mathcal{S}_{[n]}|G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{m}\right)^{1/q}.$$

(c) *Assume* (S4), *and that either condition* (a) *or* (b) *of Theorem 2 holds, then*

$$\inf_{\hat{G} \in \mathcal{G}^*} \sup_{G_0 \in \mathcal{G}^k} P^m_{\mathcal{S}_{[n]}|G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn}\right)^{1/(1+\alpha')}.$$

Furthermore, if (S4) *is replaced by* (S4′), *the lower bound becomes $1/m$.*

Remarks.

1. There is a gap between the posterior contraction rate in Theorem 1 and the minimax lower bound in Theorem 3(a). This is expected because the infimum over point estimates $\hat{G}$ is taken over $\mathcal{G}^*$, as opposed to $\mathcal{G}^k$. Nonetheless, the lower bounds are notably dependent on d and k, thereby provide a partial justification for the nonparametics-like posterior contraction rates. It is interesting to note that if $k \geq 2d \gg \alpha$, and allowing $m \asymp n$, the rate exponents differ approximately by only a factor of 4. That is, $m^{-1/2(d+\alpha)}$ vis-à-vis $m^{-2/(d+\alpha')}$.
2. The nonparametrics-like lower bounds in part (a) and (b) in the overfitted setting are somewhat surprising even if $P_{\boldsymbol{\beta}}$ is known exactly (e.g., $P_{\boldsymbol{\beta}}$ is uniform distribution). Since we are more likely to be in the overfitted setting than knowing the exact number of extreme points, an implication of this is that it is important in practice to impose a lower bound on the pairwise distances between the extreme points of the population polytope.
3. The results in part (b) and (c) under assumption (S4′) present an interesting scenario in which the obtained lower bounds do not depend on n, which determines the amount of data at the bottom level in the model hierarchy.

3. Geometric assumptions and basic lemmas

In this section, we discuss the geometric assumptions postulated in the main theorems, and describe their consequences using elementary arguments in convex geometry of Euclidean spaces. These results relate Hausdorff metric, the minimum-matching metric, and the volume of the set-theoretic difference of polytopes. These relationships prove crucial in obtaining explicit posterior contraction rates. Here, we state the properties and prove the results for p-dimensional polytopes and convex bodies of points in Δ^d, for a given $p \leq d$. (Convex bodies are bounded convex sets that may have an unbounded number of extreme points. Within this section, the detail of the ambient space is irrelevant. For instance, Δ^d may be replaced by $\mathbb{R}^{d+1}$ or a higher dimensional Euclidean space.)

Property A1 (Property of thick body). *For some $r, R > 0$, $\boldsymbol{\theta}_c \in \Delta^d$, G contains the spherical ball $B_p(\boldsymbol{\theta}_c, r)$ and is contained in $B_p(\boldsymbol{\theta}_c, R)$.*

Property A2 (Property of non-obtute corners). *For some small $\delta > 0$, at each vertex of G there is a supporting hyperplane whose angle formed with any edges adjacent to that vertex is bounded from below by δ.*

We state key geometric lemmas that will be used throughout the paper. Bounds such as those given by Lemma 2 are probably well-known in the folklore of convex geometry (for instance, part (b) of that lemma is similar to (but not precisely the same as) Lemma 2.3.6. from [15]). Due to the absence of direct references, we include the proof of this and other lemmas in the Appendix.

Lemma 1.

(a) $d_{\mathcal{H}}(G, G') \leq d_{\mathcal{M}}(G, G')$.
(b) *If the two polytopes G, G' satisfy Property A2, then $d_{\mathcal{M}}(G, G') \leq C_0 d_{\mathcal{H}}(G, G')$, for some positive constant $C_0 > 0$ depending only on δ.*

According to part (b) of this lemma, convergence of a sequence of convex polytope $G \in \mathcal{G}^k$ to $G_0 \in \mathcal{G}^k$ in Hausdorff metric entails the convergence of the extreme points of G to those of G_0. Moreover, they share the same rate as the Hausdorff convergence.

Lemma 2. *There are positive constants C_1 and c_1 depending only on r, R, p such that for any two p-dimensional convex bodies G, G' satisfying Property A1:*

(a) $\mathrm{vol}_p\, G \triangle G' \geq c_1 d_{\mathcal{H}}(G, G')^p$.
(b) $\mathrm{vol}_p\, G \triangle G' \leq C_1 d_{\mathcal{H}}(G, G')$.

Remark. The exponents in both bounds in Lemma 2 are attainable. Indeed, for the lower bound in part (a), consider a fixed convex polytope G. For each vertex $\boldsymbol{\theta}_i \in G$, consider point x that lie on edges incident to $\boldsymbol{\theta}_i$ such that $\|x - \boldsymbol{\theta}_i\| = \varepsilon$. Let G' be the convex hull of all such x's and the remaining vertices of G. Clearly, $d_{\mathcal{H}}(G, G') = \mathrm{O}(\varepsilon)$, and $\mathrm{vol}_p\, G \setminus G' \le \mathrm{O}(\varepsilon^p)$. Thus, for the collection of convex polytopes G' constructed in this way, $\mathrm{vol}_p(G \triangle G') \asymp d_{\mathcal{H}}(G, G')^p$. The upper bound in part (b) is also tight for a broad class of convex polytopes, as exemplified by the following lemma.

Lemma 3. *Let G be a fixed polytope and $|\mathrm{extr}\, G| = k < \infty$. G' an arbitrary polytope in $\mathcal{G}^*$. Moreover, either one of the following conditions holds:*

(a) $|\mathrm{extr}\, G'| = k$, *or*
(b) *The pairwise distances between the extreme points of G' is bounded away from a constant $r_0 > 0$.*

Then, there is a positive constant $\varepsilon_0 = \varepsilon_0(G)$ depending only on G, a positive constant $c_2 = c_2(G)$ in case (a) *and $c_2 = c_2(G, r_0)$ in case* (b)*, such that*

$$\mathrm{vol}_p\, G \triangle G' \ge c_2 d_{\mathcal{H}}(G, G')$$

as soon as $d_{\mathcal{H}}(G, G') \le \varepsilon_0(G)$.

Remark. We note that the bound obtained in this lemma is substantially stronger than the one obtained by Lemma 2 part (a). This is due to the asymmetric roles of G, which is held fixed, and G', which can vary. As a result, constant c_2 as stated in the present lemma is independent of G' but allowed to be dependent on G. By contrast, constant c_1 in Lemma 2 part (a) is independent of both G and G'.

4. An abstract posterior contraction theorem

In this section, we state an abstract posterior contraction theorem for hierarchical models, whose proof is given in the Appendix. The setting of this theorem is a general hierarchical model defined as follows

$$G \sim \Pi, \qquad \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_m | G \sim P_{\boldsymbol{\eta}|G},$$
$$\mathcal{S}^i_{[n]} | \boldsymbol{\eta}_i \sim P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_i} \qquad \text{for } i = 1, \ldots, m.$$

The detail of conditional distributions in above specifications is actually irrelevant. Thus, results in this section may be of general interest for hierarchical models with $m \times n$ data.

As before $p_{\mathcal{S}_{[n]}|G}$ is marginal density of the generic $\mathcal{S}_{[n]}$ which is obtained by integrating out the generic random vector $\boldsymbol{\eta}$ (e.g., see Equation (2)). We need several key notions. Define the Hausdorff ball as:

$$B_{d_{\mathcal{H}}}(G_1, \delta) := \{G \in \Delta^d \colon\ d_{\mathcal{H}}(G_1, G) \le \delta\}.$$

A useful quantity for proving posterior concentration theorems is the Hellinger information of Hausdorff metric for a given set:

Definition 1. *Fix $G_0 \in \mathcal{G}^*$. For a fixed n, the sample size of $\mathcal{S}_{[n]}$, define the Hellinger information of $d_{\mathcal{H}}$ metric for set $\mathcal{G} \subset \mathcal{G}^*$ as a real-valued function on the positive reals $\Psi_{\mathcal{G},n} : \mathbb{R}_+ \to \mathbb{R}$:*

$$\Psi_{\mathcal{G},n}(\delta) := \inf_{G \in \mathcal{G}; d_{\mathcal{H}}(G_0,G) \geq \delta/2} h^2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G}). \tag{6}$$

We also define $\Phi_{\mathcal{G},n} : \mathbb{R}_+ \to \mathbb{R}$ to be an arbitrary non-negative valued function on the positive reals such that for any $\delta > 0$,

$$\sup_{G,G' \in \mathcal{G}; d_{\mathcal{H}}(G,G') \leq \Phi_{\mathcal{G},n}(\delta)} h^2(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \Psi_{\mathcal{G},n}(\delta)/4.$$

In both definitions of Φ and Ψ, we suppress the dependence on (the fixed) G_0 to simplify notations. Note that if $G_0 \in \mathcal{G}$, it follows from the definition that $\Phi_{\mathcal{G},n}(\delta) < \delta/2$.

Remark. Suppose that conditions of Lemma 7(b) hold, so that

$$h^2(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \frac{n}{c_0} C_0 d_{\mathcal{H}}(G, G').$$

Then it suffices to choose $\Phi_{\mathcal{G},n}(\delta) = \frac{c_0}{4nC_0} \Psi_{\mathcal{G},n}(\delta)$.

Define the neighborhood of the prior support around G_0 in terms of Kullback–Leibler distance of the marginal densities $p_{\mathcal{S}_{[n]}|G}$:

$$B_K(G_0, \delta) = \{G \in \mathcal{G}^* | K(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G}) \leq \delta^2; K_2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G}) \leq \delta^2\}. \tag{7}$$

Theorem 4. *Let $\mathcal{G}$ denote the support of the prior Π. Fix $G_0 \in \mathcal{G}$ and suppose that*

(a) *$m \to \infty$ and $n \to \infty$ at a certain rate relative to m.*
(b) *There is a large constant C, a sequence of scalars $\varepsilon_{m,n} \to 0$ defined in terms of m and n such that $m\varepsilon_{m,n}^2$ tends to infinity, such that*

$$\sup_{G_1 \in \mathcal{G}} \log D(\Phi_{\mathcal{G},n}(\varepsilon), \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \varepsilon/2), d_{\mathcal{H}})$$
$$+ \log D(\varepsilon/2, \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_0, 2\varepsilon) \setminus B_{d_{\mathcal{H}}}(G_0, \varepsilon), d_{\mathcal{H}}) \leq m\varepsilon_{m,n}^2 \tag{8}$$
$$\forall \varepsilon \geq \varepsilon_{m,n},$$
$$\Pi(B_K(G_0, \varepsilon_{m,n})) \geq \exp[-m\varepsilon_{m,n}^2 C]. \tag{9}$$

(c) *There is a sequence of positive scalars M_m such that*

$$\Psi_{\mathcal{G},n}(M_m \varepsilon_{m,n}) \geq 8\varepsilon_{m,n}^2 (C+4), \tag{10}$$

$$\exp(2m\varepsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-m\Psi_{\mathcal{G},n}(j\varepsilon_{m,n})/8] \to 0. \tag{11}$$

Then, $\Pi(G\colon\ d_{\mathcal{H}}(G_0, G) \geq M_m \varepsilon_{m,n} | \mathcal{S}_{[n]}^{[m]}) \to 0$ in $P^m_{\mathcal{S}_{[n]}|G_0}$-probability as m and $n \to \infty$.

Condition (8) is referred to as entropy condition for certain sets in the support of the prior. Condition (9) is concerned with the "thickness" of the prior as measured by the Kullback–Leibler distance (see also [7]). Conditions (10) and (11) are related to the Hellinger information function (see also [12]). The proof of this theorem is deferred to the Appendix. As noted above, this result is applicable to any hierarchical models for $m \times n$ data. The choice of Hausdorff metric $d_{\mathcal{H}}$ is arbitrary here, and can be replaced by any other valid metric (e.g., $d_{\mathcal{M}}$). The remainder of the paper is devoted to verifying the conditions of this theorem so it can be applied. These conditions hinge on our having established a lower bound for the Hellinger information function $\Psi_{\mathcal{G},n}(\cdot)$ (via Theorem 5), and a lower bound for the prior probability defined on Kullback–Leibler balls $B_K(G_0,\cdot)$ (via Theorem 6). Both types of results are obtained by utilizing the convex geometry lemmas described in the previous section.

5. Inequalities for the Hausdorff distance

The following results guarantee that as marginal densities of $\mathcal{S}_{[n]}$ get closer in total variation distance metric (or Hellinger metric), so do the corresponding population polytopes in Hausdorff metric (or minimum matching metric). This gives a lower bound for the Hellinger information defined by Equation (6), because h is related to V via inequality $h \geq V$.

Theorem 5.

(a) *Let G, G' be two convex bodies in Δ^d. G is a p-dimensional body containing spherical ball $B_p(\boldsymbol{\theta}_c, r)$, while G' is p'-dimensional body containing $B_{p'}(\boldsymbol{\theta}_c, r)$ for some $p, p' \leq d, r > 0, \boldsymbol{\theta}_c \in \Delta^d$. In addition, assume that both $p_{\boldsymbol{\eta}|G}$ and $p_{\boldsymbol{\eta}|G'}$ are α-regular densities on G and G', respectively. Then, there is $c_1 > 0$ independent of G, G' such that*

$$c_1 d_{\mathcal{H}}(G,G')^{(p\vee p')+\alpha} \leq V(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) + 6(d+1)\exp\left[-\frac{n}{8(d+1)} d_{\mathcal{H}}(G,G')^2\right].$$

(b) *Assume further that G is fixed convex polytope, G' an arbitrary polytope, $p' = p$, and that either $|\operatorname{extr} G'| = |\operatorname{extr} G|$ or the pairwise distances of extreme points of G' is bounded from below by a constant $r_0 > 0$. Then, there are constants $c_2, C_3 > 0$ depending only on G and r_0 (and independent of G') such that*

$$c_2 d_{\mathcal{H}}(G,G')^{1+\alpha} \leq V(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) + 6(d+1)\exp\left[-\frac{n}{C_3(d+1)} d_{\mathcal{H}}(G,G')^2\right].$$

Remark. Part (a) holds for varying pairs of G, G' satisfying certain conditions. It is consequence of Lemma 2(a). Part (b) produces a tighter bound, but it holds only for a fixed G, while G' is allowed to vary while satisfying certain conditions. This is a consequence of Lemma 3. Constants c_1, c_2 are the same as those from Lemmas 2(a) and 3, respectively.

Proof of Theorem 5. (a) The main idea of the proof is the construction of a suitable test set in order to distinguish $p_{\mathcal{S}_{[n]}|G'}$ from $p_{\mathcal{S}_{[n]}|G}$. The proof is organized as a sequence of steps.

Step 1. Given a data vector $\mathcal{S}_{[n]} = (X_1, \ldots, X_n)$, define $\hat{\boldsymbol{\eta}}(\mathcal{S}) \in \Delta^d$ such that the i-element of $\hat{\boldsymbol{\eta}}(\mathcal{S})$ is $\frac{1}{n}\sum_{j=1}^n \mathbb{I}(X_j = i)$ for each $i = 0, \ldots, d$. In the following, we simply use $\hat{\boldsymbol{\eta}}$ to ease the notations. By the definition of the variational distance,

$$V(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) = \sup_A |P_{\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A) - P_{\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A)|, \tag{12}$$

where the supremum is taken over all measurable subsets of Δ^d.

Step 2. Fix a constant $\varepsilon > 0$. By Hoeffding's inequality and the union bound, under the conditional distribution $P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}$,

$$P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}\Big(\max_{i=0,\ldots,d} |\hat{\eta}_i - \eta_i| \geq \varepsilon\Big) \leq 2(d+1)\exp(-2n\varepsilon^2)$$

with probability one (as $\boldsymbol{\eta}$ is random). It follows that

$$\begin{aligned} P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\| \geq \varepsilon) &\leq P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}\Big(\max_{i=0,\ldots,d} |\hat{\eta}_i - \eta_i| \geq \varepsilon(d+1)^{-1/2}\Big) \\ &\leq 2(d+1)\exp[-2n\varepsilon^2/(d+1)]. \end{aligned}$$

The same bound holds under $P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}$.

Step 3. Define event $B = \{\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\| < \varepsilon\}$. Take any (measurable) set $A \subset \Delta^d$,

$$\begin{aligned} &|P_{\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A) - P_{\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A)| \\ &\quad = |P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A; B) + P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A; B^C) \\ &\qquad - P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A; B) - P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A; B^C)| \\ &\quad \geq |P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A; B) - P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A; B)| \\ &\qquad - 4(d+1)\exp[-2n\varepsilon^2/(d+1)]. \end{aligned} \tag{13}$$

Step 4. Let $\varepsilon_1 = d_{\mathcal{H}}(G, G')/4$. For any $\varepsilon \leq \varepsilon_1$, recall the outer ε-parallel set $G_\varepsilon = (G + B_{d+1}(\mathbf{0}, \varepsilon))$, which is full-dimensional $(d+1)$ even though G may not be. By triangular inequality, $d_{\mathcal{H}}(G_\varepsilon, G'_\varepsilon) \geq d_{\mathcal{H}}(G, G')/2$. We shall argue that for any $\varepsilon \leq \varepsilon_1$, there is a constant $c_1 > 0$ independent of G, G', ε and ε_1 such that either one of the two scenarios holds:

(i) There is a set $A^* \subset G \setminus G'$ such that $A^*_\varepsilon \cap G'_\varepsilon = \varnothing$ and $\mathrm{vol}_p(A^*) \geq c_1\varepsilon_1^p$, or
(ii) There is a set $A^* \subset G' \setminus G$ such that $A^*_\varepsilon \cap G_\varepsilon = \varnothing$ and $\mathrm{vol}_{p'}(A^*) \geq c_1\varepsilon_1^{p'}$.

Indeed, since $\varepsilon \leq d_{\mathcal{H}}(G, G')/4$, either one of the following two inequalities holds: $d_{\mathcal{H}}(G \setminus G'_{3\varepsilon}, G') \geq d_{\mathcal{H}}(G, G')/4$ or $d_{\mathcal{H}}(G' \setminus G_{3\varepsilon}, G) \geq d_{\mathcal{H}}(G, G')/4$. If the former inequality holds, let $A^* = G \setminus G'_{3\varepsilon}$. Then, $A^* \subset G \setminus G'$ and $A^*_\varepsilon \cap G'_\varepsilon = \varnothing$. Moreover, by Lemma 2(a),

$\mathrm{vol}_p(A^*) \geq c_1\varepsilon_1^p$, for some constant $c_1 > 0$ independent of $\varepsilon, \varepsilon_1, G, G'$, so A^* satisfies (i). In fact, using the same argument as in the proof of Lemma 2(a) there is a point $x \in \mathrm{bd}\, G$ such that $G' \cap B_p(x, \varepsilon_1) = \varnothing$. Combined with the α-regularity of $P_{\boldsymbol{\eta}|G}$, we have $P_{\boldsymbol{\eta}|G}(A^*) \geq \varepsilon^\alpha \,\mathrm{vol}_p(G \cap B_p(x, \varepsilon_1)) \geq c_1 \varepsilon^{p+\alpha}$ for some constant $c_1 > 0$. If the latter inequality holds, the same argument applies by defining $A^* = G' \setminus G_{3\varepsilon}$ so that (ii) holds.

Step 5. Suppose that (i) holds for the chosen A^*. This means that $P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A^*_\varepsilon; B) \leq P_{\boldsymbol{\eta}|G'}(\boldsymbol{\eta} \in A^*_{2\varepsilon}) = 0$, since $A^*_{2\varepsilon} \cap G' = \varnothing$, which is a consequence of $A^*_\varepsilon \cap G'_\varepsilon = \varnothing$. In addition,

$$
\begin{aligned}
P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A^*_\varepsilon; B) &\geq P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(\boldsymbol{\eta} \in A^*; B) \\
&\geq P_{\boldsymbol{\eta}|G}(A^*) - P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}(B^C) \\
&\geq P_{\boldsymbol{\eta}|G}(A^*) - 2(d+1)\exp(-2n\varepsilon^2/(d+1)) \\
&\geq c_1\varepsilon_1^{p+\alpha} - 2(d+1)\exp(-2n\varepsilon^2/(d+1)).
\end{aligned}
$$

Hence, by Equation (13) $|P_{\mathcal{S}_{[n]}|G}(\hat{\boldsymbol{\eta}} \in A^*_\varepsilon) - P_{\mathcal{S}_{[n]}|G'}(\hat{\boldsymbol{\eta}} \in A^*_\varepsilon)| \geq c_1\varepsilon_1^{p+\alpha} - 6(d+1) \times \exp(-2n\varepsilon^2)$. Set $\varepsilon = \varepsilon_1$, the conclusion then follows by invoking Equation (12). The scenario of (ii) proceeds in the same way.

(b) Under the condition that the pairwise distances of extreme points of G' are bounded from below by $r_0 > 0$, the proof is very similar to part (a), by invoking Lemma 3. Under the condition that $|\,\mathrm{extr}\, G'| = k$, the proof is also similar, but it requires a suitable modification for the existence of set A^*. For any small ε, let $\tilde{G}_\varepsilon$ be the minimum-volume homethetic transformation of G, with respect to center $\boldsymbol{\theta}_c$, such that $\tilde{G}_\varepsilon$ contains G_ε. Since $B_p(\boldsymbol{\theta}_c, r) \subset G \subset B_p(\boldsymbol{\theta}_c, R)$ for $R = 1$, it is simple to see that $d_{\mathcal{H}}(G, \tilde{G}_\varepsilon) \leq \varepsilon R/r = \varepsilon/r$.

Set $\varepsilon_1 = d_{\mathcal{H}}(G, G')r/4$. We shall argue that for any $\varepsilon \leq \varepsilon_1$, there is a constant $c_0 > 0$ independent of G', ε and ε_1 such that either one of the following two scenarios hold:

(iii) There is a set $A^* \subset G \setminus G'$ such that $A^*_\varepsilon \cap G'_\varepsilon = \varnothing$ and $\mathrm{vol}_p(A^*) \geq c_2\varepsilon_1$, or
(iv) There is a set $A^* \subset G' \setminus G$ such that $A^*_\varepsilon \cap G_\varepsilon = \varnothing$ and $\mathrm{vol}_p(A^*) \geq c_2\varepsilon_1$.

Indeed, note that either one of the following two inequalities holds: $d_{\mathcal{H}}(G \setminus \tilde{G}'_{3\varepsilon}, G') \geq d_{\mathcal{H}}(G, G')/4$ or $d_{\mathcal{H}}(G' \setminus \tilde{G}_{3\varepsilon}, G) \geq d_{\mathcal{H}}(G, G')/4$. If the former inequality holds, let $A^* = G \setminus \tilde{G}'_{3\varepsilon}$. Then, $A^* \subset G \setminus G'$ and $A^*_\varepsilon \cap \tilde{G}'_\varepsilon = \varnothing$. Observe that both G and $\tilde{G}'_{3\varepsilon}$ have the same number of extreme points by the construction. Moreover, G is fixed so that all geometric Properties A2, A1 are satisfied for both G and $\tilde{G}'_{3\varepsilon}$ for sufficiently small $d_{\mathcal{H}}(G, G')$. By Lemma 3, $\mathrm{vol}_p(A^*) \geq c_2\varepsilon_1$. Hence, (iii) holds. If the latter inequality holds, the same argument applies by defining $A^* = G' \setminus \tilde{G}_{3\varepsilon}$ so that (iv) holds.

Now the proof of the theorem proceeds in the same manner as in part (a). □

6. Concentration properties of the prior support

In this section, we study properties of the support of the prior probabilities as specified by the admixture model, including bounds for the support of the prior as defined by Kullback–Leibler neighborhoods.

α-regularity

Let $\boldsymbol{\beta}$ be a random variable taking values in Δ^{k-1} that has a density $p_{\boldsymbol{\beta}}$ (with respect to the $k-1$-dimensional Hausdorff measure $\mathcal{H}^{k-1}$ on $\mathbb{R}^k$). For a definition of the Hausdorff measure, see [6], which in our case reduces to the Lebesgue measure defined on simplex Δ^{k-1}. Define random variable $\boldsymbol{\eta} = \beta_1\boldsymbol{\theta}_1 + \cdots + \beta_k\boldsymbol{\theta}_k$, which takes values in $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. Write $\boldsymbol{\eta} = L\boldsymbol{\beta}$, where $L = [\boldsymbol{\theta}_1 \quad \cdots \quad \boldsymbol{\theta}_k]$ is a $(d+1)\times k$ matrix. If $k \leq d+1$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ are generally linearly independent, in which case matrix L has rank $k-1$. By the change of variable formula [6] (Chapter 3), $P_{\boldsymbol{\beta}}$ induces a distribution $P_{\boldsymbol{\eta}|G}$ on $G \subset \Delta^d$, which admits the following density with respect to the $k-1$ dimensional Hausdorff measure $\mathcal{H}^{k-1}$ on Δ^d:

$$p_{\boldsymbol{\eta}}(\boldsymbol{\eta}|G) = p_{\boldsymbol{\beta}}(L^{-1}(\boldsymbol{\eta}))J(L)^{-1}. \tag{14}$$

Here, $J(L)$ denotes the Jacobian of the linear map. On the other hand, if $k \geq d+1$, then L is generally d-ranked. The induced distribution for $\boldsymbol{\eta}$ admits the following density with respect to the $k-(d+1)$-dimensional Hausdorff measure on $\mathbb{R}^{d+1}$:

$$p_{\boldsymbol{\eta}}(\boldsymbol{\eta}|G) = \int_{L^{-1}\{\boldsymbol{\eta}\}} p_{\boldsymbol{\beta}}(\boldsymbol{\beta})J(L)^{-1}\mathcal{H}^{k-(d+1)}(\mathrm{d}\boldsymbol{\beta}). \tag{15}$$

A common choice for $P_{\boldsymbol{\beta}}$ is the Dirichlet distribution, as adopted by [4, 13]: given parameter $\gamma \in \mathbb{R}^k_+$, for any $A \subset \Delta^{k-1}$,

$$P_{\boldsymbol{\beta}}(\boldsymbol{\beta} \in A|\gamma) = \int_A \frac{\Gamma(\sum \gamma_j)}{\prod_{j=1}^k \Gamma(\gamma_j)} \prod_{j=1}^k \beta_j^{\gamma_j - 1}\mathcal{H}^{k-1}(\mathrm{d}\boldsymbol{\beta}).$$

Lemma 4. *Let $\boldsymbol{\eta} = \sum_{j=1}^k \beta_j\boldsymbol{\theta}_j$, where $\boldsymbol{\beta}$ is distributed according to a $k-1$-dimensional Dirichlet distribution with parameters $\gamma_j \in (0,1]$ for $j = 1, \ldots, k$.*

(a) *If $k \leq d+1$, there is constant $\varepsilon_0 = \varepsilon_0(k) > 0$, and constant $c_6 = c_6(\gamma, k, d) > 0$ dependent on γ, k and d such that for any $\varepsilon < \varepsilon_0$,*

$$\inf_{G \subset \Delta^d} \inf_{\boldsymbol{\eta}^* \in G} P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \leq \varepsilon) \geq c_6\varepsilon^{k-1}.$$

(b) *If $k > d+1$, the statement holds with a lower bound $c_6\varepsilon^{d+\sum_{i=1}^k \gamma_i}$.*

A consequence of this lemma is that if $\gamma_j \leq 1$ for all $j = 1, \ldots, k$, $k \leq d+1$ and G is $k-1$-dimensional, then the induced $P_{\boldsymbol{\eta}|G}$ has a Hausdorff density that is bounded away from 0 on the entire its support Δ^{k-1}, which implies 0-regularity. On the other hand, if $\gamma_j \leq 1$ for all j, $k > d+1$, and G is d-dimensional, the $P_{\boldsymbol{\eta}|G}$ is at least $\sum_{j=1}^k \gamma_j$-regularity. Note that the α-regularity condition is concerned with the density behavior near the boundary of its support, and thus is weaker than what is guaranteed here.

Bounds on KL divergences

Suppose that the population polytope G is endowed with a prior distribution on $\mathcal{G}^k$ (via prior on the population structures $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$). Given G, the marginal density $p_{\mathcal{S}_{[n]}|G}$ of n-vector $\mathcal{S}_{[n]}$ is obtained via Equation (2). To establish the concentration properties of Kullback–Leibler neighborhood B_K as induced by the prior, we need to obtain an upper bound on the KL divergences for the marginal densities in terms of Hausdorff metric on population polytopes. First, consider a very special case.

Lemma 5. *Let $G, G' \subset \Delta^d$ be closed convex sets satisfying Property A1. Moreover, assume that*

(a) $G \subset G'$, $\operatorname{aff} G = \operatorname{aff} G'$ *is p-dimensional, for $p \leq d$.*
(b) $P_{\boldsymbol{\eta}|G}$ *(resp. $P_{\boldsymbol{\eta}|G'}$) are uniform distributions on G (resp. G').*

Then, there is a constant $C_1 = C_1(r,p) > 0$ such that $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq C_1 d_{\mathcal{H}}(G, G')$.

Proof. First, we note a well-known fact of KL divergences: the divergence between marginal distributions (e.g., on $\mathcal{S}_{[n]}$) is bounded from above by the divergence between joint distributions (e.g., on $\boldsymbol{\eta}$ and $\mathcal{S}_{[n]}$ via Equation (1)):

$$K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq K(P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}, P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}).$$

Due to the hierarchical specification, $p_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G} = p_{\boldsymbol{\eta}|G} \times p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}$ and $p_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'} = p_{\boldsymbol{\eta}|G'} \times p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}$, so $K(P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G}, P_{\boldsymbol{\eta}\times\mathcal{S}_{[n]}|G'}) = K(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'})$. The assumption $\operatorname{aff} G = \operatorname{aff} G'$ and moreover $G \subset G'$ implies that $K(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}) < \infty$. In addition, $P_{\boldsymbol{\eta}|G}$ and $P_{\boldsymbol{\eta}|G'}$ are assumed to be uniform distributions on G and G', respectively, so

$$K(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}) = \int \log \frac{1/\operatorname{vol}_p G}{1/\operatorname{vol}_p G'} \, \mathrm{d}P_{\boldsymbol{\eta}|G}.$$

By Lemma 2(b), $\log[\operatorname{vol}_p G' / \operatorname{vol}_p G] \leq \log(1 + C_1 d_{\mathcal{H}}(G, G')) \leq C_1 d_{\mathcal{H}}(G, G')$ for some constant $C_1 = C_1(r,p) > 0$. This completes the proof. □

Remark. The previous lemma requires a particularly stringent condition, $\operatorname{aff} G = \operatorname{aff} G'$, and moreover $G \subset G'$, which is usually violated when $k < d+1$. However, the conclusion is worth noting in that the upper bound does not depend on the sample size n (for $\mathcal{S}_{[n]}$). The next lemma removes this condition and the condition that both $p_{\boldsymbol{\eta}|G}$ and $p_{\boldsymbol{\eta}|G'}$ be uniform. As a result the upper bound obtained is weaker, in the sense that the bound is not in terms of a Hausdorff distance, but in terms of a Wasserstein distance.

Let $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ denote a coupling of $P(\boldsymbol{\eta}|G)$ and $P(\boldsymbol{\eta}|G')$, that is, a joint distribution on $G \times G'$ whose induced marginal distributions of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are equal to $P(\boldsymbol{\eta}|G)$ and $P(\boldsymbol{\eta}|G')$, respectively. Let $\mathcal{Q}$ be the set of all such couplings. The Wasserstein distance between $p_{\boldsymbol{\eta}|G}$ and $p_{\boldsymbol{\eta}|G'}$ is defined as

$$W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}) = \inf_{Q \in \mathcal{Q}} \int \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| \, \mathrm{d}Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2).$$

Lemma 6. *Let $G, G' \subset \Delta^d$ be closed convex subsets such that any $\boldsymbol{\eta} = (\eta_0, \dots, \eta_d) \in G \cup G'$ satisfies $\min_{l=0,\dots,d} \eta_l > c_0$ for some constant $c_0 > 0$. Then*

$$K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \frac{n}{c_0} W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}).$$

Remark. As $n \to \infty$, the upper bound tends to infinity. This is expected, because the marginal distribution $P_{\mathcal{S}_{[n]}|G}$ should degenerate. Since typically $\operatorname{aff} G \neq \operatorname{aff} G'$, Kullback–Leibler distances between $P_{\mathcal{S}_{[n]}|G}$ and $P_{\mathcal{S}_{[n]}|G'}$ should typically tend to infinity.

Proof of Lemma 6. Associating each sample $\mathcal{S}_{[n]} = (X_1, \dots, X_n)$ with a $d+1$-dimensional vector $\boldsymbol{\eta}(\mathcal{S}) \in \Delta^d$, where $\boldsymbol{\eta}(\mathcal{S})_i = \frac{1}{n} \sum_{j=1}^n \mathbb{I}(X_j = i)$ for each $i = 0, \dots, d$. The density of $\mathcal{S}_{[n]}$ given G (with respect to the counting measure) takes the form:

$$p_{\mathcal{S}_{[n]}|G}(\mathcal{S}_{[n]}) = \int_G p(\mathcal{S}_{[n]}|\boldsymbol{\eta}) \, \mathrm{d}P(\boldsymbol{\eta}|G) = \int_G \exp\left(n \sum_{i=0}^d \boldsymbol{\eta}(\mathcal{S})_i \log \boldsymbol{\eta}_i \right) \mathrm{d}P(\boldsymbol{\eta}|G).$$

Due to the convexity of Kullback–Leibler divergence, by Jensen inequality, for any coupling $Q \in \mathcal{Q}$:

$$\begin{aligned} K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) &= K\left(\int p(\mathcal{S}_{[n]}|\boldsymbol{\eta}_1) \, \mathrm{d}Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \int p(\mathcal{S}_{[n]}|\boldsymbol{\eta}_2) \, \mathrm{d}Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \right) \\ &\leq \int K(p(\mathcal{S}_{[n]}|\boldsymbol{\eta}_1), p(\mathcal{S}_{[n]}|\boldsymbol{\eta}_2)) \, \mathrm{d}Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2). \end{aligned}$$

It follows that $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \inf_Q \int K(p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_1}, p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_2}) \, \mathrm{d}Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$.

Note that $K(P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_1}, P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_2}) = \sum_{\mathcal{S}_{[n]}} n(K(\boldsymbol{\eta}(\mathcal{S}), \boldsymbol{\eta}_2) - K(\boldsymbol{\eta}(\mathcal{S}), \boldsymbol{\eta}_1)) p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_1}$, where the summation is taken over all realizations of $\mathcal{S}_{[n]} \in \{0, \dots, d\}^n$. For any $\boldsymbol{\eta}(\mathcal{S}) \in \Delta^d$, $\boldsymbol{\eta}_1 \in G$ and $\boldsymbol{\eta}_2 \in G'$,

$$\begin{aligned} |K(\boldsymbol{\eta}(\mathcal{S}), \boldsymbol{\eta}_1) - K(\boldsymbol{\eta}(\mathcal{S}), \boldsymbol{\eta}_2)| &= \left| \sum_{i=0}^d \boldsymbol{\eta}(\mathcal{S})_i \log(\eta_{1,i}/\eta_{2,i}) \right| \\ &\leq \sum_i \boldsymbol{\eta}(\mathcal{S})_i |\eta_{1,i} - \eta_{2,i}| / c_0 \\ &\leq \left(\sum_i \boldsymbol{\eta}(\mathcal{S})_i^2 \right)^{1/2} \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| / c_0 \\ &\leq \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| / c_0. \end{aligned}$$

Here, the first inequality is due the assumption, the second due to Cauchy–Schwarz. It follows that $K(P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_1}, P_{\mathcal{S}_{[n]}|\boldsymbol{\eta}_2}) \leq n \|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| / c_0$, so $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \leq \frac{n}{c_0} W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'})$. $\square$

Lemma 7. *Let $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ and $G' = \mathrm{conv}(\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_k)$ (same k). A random variable $\boldsymbol{\eta} \sim P_{\boldsymbol{\eta}|G}$ is parameterized by $\boldsymbol{\eta} = \sum_j \beta_j \boldsymbol{\eta}_j$, while a random variable $\boldsymbol{\eta} \sim P_{\boldsymbol{\eta}|G'}$ is parameterized by $\boldsymbol{\eta} = \sum_j \beta'_j \boldsymbol{\eta}'_j$, where $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ are both distributed according to a symmetric probability density $p_{\boldsymbol{\beta}}$.*

(a) *Assume that both G, G' satisfy Property A2. Then, for small $d_{\mathcal{H}}(G, G')$, $W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}) \le C_0 d_{\mathcal{H}}(G, G')$ for some constant C_0 specified by Lemma 1.*

(b) *Assume further that assumptions in Lemma 6 hold, then $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) \le \frac{n}{c_0} C_0 d_{\mathcal{H}}(G, G')$.*

Remark. In order to obtain an upper bound for $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'})$ in terms of $d_{\mathcal{H}}(G, G')$, the assumption that $p_{\boldsymbol{\beta}}$ is symmetric appears essential. That is, random variables $\beta_1, \ldots, \beta_k$ are exchangeable under $p_{\boldsymbol{\beta}}$. Without this assumption, it is possible to have $d_{\mathcal{H}}(G, G') = 0$, but $K(p_{\mathcal{S}_{[n]}|G}, p_{\mathcal{S}_{[n]}|G'}) > 0$.

Proof of Lemma 7. By Lemma 1 under Property A2, $d_{\mathcal{M}}(G, G') \le C_0 d_{\mathcal{H}}(G, G')$ for some constant C_0. Let $d_{\mathcal{H}}(G, G') \le \varepsilon$ for some small $\varepsilon > 0$. Assume without loss of generality that $|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j| \le C_0 \varepsilon$ for all $j = 1, \ldots, k$ (otherwise, simply relabel the subscripts for $\boldsymbol{\theta}'_j$'s).

Let $Q(\boldsymbol{\eta}, \boldsymbol{\eta}')$ be a coupling of $P_{\boldsymbol{\eta}|G}$ and $P_{\boldsymbol{\eta}|G'}$ such that under Q, $\boldsymbol{\eta} = \sum_{j=1}^k \beta_j \boldsymbol{\theta}_j$ and $\boldsymbol{\eta}' = \sum_{j=1}^k \beta_j \boldsymbol{\theta}'_j$, that is, $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$ share the *same* $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a random variable with density $p_{\boldsymbol{\beta}}$. This is a valid coupling, since $p_{\boldsymbol{\beta}}$ is assumed to be symmetric.

Under distribution Q, $\mathbb{E}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \le \mathbb{E} \sum_{j=1}^k \beta_j \|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\| \le C_0 \varepsilon \mathbb{E} \sum_{j=1}^k \beta_j = C_0 \varepsilon$. Hence, $W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G'}) \le C_0 \varepsilon$. Part (b) is an immediate consequence. □

Recall the definition of Kullback–Leibler neighborhood given by Equation (7). We are now ready to prove the main result of this section.

Theorem 6. *Under assumptions* (S1) *and* (S2)*, for any G_0 in the support of prior Π, for any $\delta > 0$ and $n > \log(1/\delta)$*

$$\Pi(G \in B_K(G_0, \delta)) \ge c(\delta^2/n^3)^{kd},$$

where constant $c = c(c_0, c'_0)$ depends only on c_0, c'_0.

Proof. We shall invoke a bound of [21] (Theorem 5) on the KL divergence. This bound says that if p and q are two densities on a common space such that $\int p^2/q < M$, then for some universal constant $\varepsilon_0 > 0$, as long as $h(p,q) \le \varepsilon < \varepsilon_0$, there holds: $K(p,q) = \mathrm{O}(\varepsilon^2 \log(M/\varepsilon))$, and $K_2(p,q) := \int p(\log(p/q))^2 = \mathrm{O}(\varepsilon^2 [\log(M/\varepsilon)]^2)$, where the big O constants are universal.

Let $G_0 = \mathrm{conv}(\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_k)$. Consider a random set $G \in \mathcal{G}^k$ represented by $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, and the event $\mathcal{E}$ that $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*_j\| \le \varepsilon$ for all $j = 1, \ldots, k$. For the pair of G_0 and G, consider a coupling Q for $P_{\boldsymbol{\eta}|G}$ and $P_{\boldsymbol{\eta}|G_0}$ such that any $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ distributed by Q is parameterized by $\boldsymbol{\eta}_1 = \beta_1 \boldsymbol{\theta}_1 + \cdots + \beta_k \boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_2 = \beta_1 \boldsymbol{\theta}^*_1 + \cdots + \beta_k \boldsymbol{\theta}^*_k$

(that is, under the coupling $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ share the same vector $\boldsymbol{\beta}$). Then, under Q, $\mathbb{E}\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| \le \varepsilon$. This entails that $W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G_0}) \le \varepsilon$. (We note here that the argument appears similar to the one from Lemma 7, but we do not need to assume that $p_{\boldsymbol{\beta}}$ be symmetric in this theorem.) If G is randomly distributed according to prior Π, under assumption (S2), the probability of event $\mathcal{E}$ is lower bounded by $c_0'\varepsilon^{kd}$. By Lemma 6, $h^2(p_{G_0}, p_G) \le K(p_{G_0}, p_G)/2 \le (n/c_0)W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G_0}) \le n\varepsilon/(2c_0)$. Note that the density ratio $p_{\mathcal{S}_{[n]}|G}/p_{\mathcal{S}_{[n]}|G_0} \le (1/c_0)^n$, which implies that $\sum_{\mathcal{S}_{[n]}} p_{\mathcal{S}_{[n]}|G_0}{}^2/p_{\mathcal{S}_{[n]}|G} \le (1/c_0)^n$. We can apply the upper bound described in the previous paragraph to obtain:

$$K_2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G}) = \mathrm{O}\left(\frac{n\varepsilon}{2c_0}\left[\frac{1}{2}\log\frac{2c_0}{n\varepsilon} + n\log\frac{1}{c_0}\right]^2\right).$$

Here, the big O constant is universal. If we set $\varepsilon = \delta^2/n^3$, then the quantity in the right hand side of the previous display is bounded by $\mathrm{O}(\delta^2)$ as long as $n > \log(1/\delta)$. Combining with the probability bound $c_0'\varepsilon^{kd}$ derived above, we obtain the desired result. □

7. Proofs of main theorems and auxiliary lemmas

Proof of Theorem 1 (Overfitted setting). The proof proceeds by verifying conditions of Theorem 4. Let $\varepsilon_{m,n} = (\log m/m)^{1/2} + (\log n/m)^{1/2} + (\log n/n)^{1/2}$. Let $\mathcal{G} := \operatorname{supp}\Pi \subset \mathcal{G}^k$. Starting with the entropy condition (8), we note that

$$\log D(\varepsilon/2, \mathcal{G} \cap B_{\mathcal{H}}(G_0, 2\varepsilon), d_{\mathcal{H}}) \le \log N(\varepsilon/4, \mathcal{G} \cap B_{\mathcal{H}}(G_0, 2\varepsilon), d_{\mathcal{H}}) = \mathrm{O}(1).$$

By Theorem 5(a), assumption (S3a) and the general inequality that $h \ge V$, we have:

$$\Psi_{\mathcal{G},n}(\varepsilon) \ge [c_1(\varepsilon/2)^{p+\alpha} - 6(d+1)\mathrm{e}^{-n\varepsilon^2/32(d+1)}]^2,$$

where $p = \min(k-1, d)$. So $\Psi_{\mathcal{G},n}(\varepsilon) \ge c\varepsilon^{2(p+\alpha)}$ as long as $c_1(\varepsilon/2)^{p+\alpha} \ge 12(d+1)\exp[-n\varepsilon^2/32(d+1)]$. Here, c is a constant depending on c_1, p, d. This is satisfied if ε is bounded from below by a large multiple of $\varepsilon_{m,n} > (\log n/n)^{1/2}$. Using $\Phi_{\mathcal{G},n}(\delta) := \frac{c_0}{4nC_0}\Psi_{\mathcal{G},n}(\delta)$ (cf. Remark following Definition 1) it follows that

$$\begin{aligned}
&\log D(c_0\Psi_{\mathcal{G},n}(\varepsilon)/(4nC_0), \mathcal{G} \cap B_{\mathcal{H}}(G_1, \varepsilon/2), d_{\mathcal{H}}) \\
&\quad \le \log N(c_0 c\varepsilon^{2(p+\alpha)}/(4nC_0), \mathcal{G} \cap B_{\mathcal{H}}(G_1, \varepsilon/2), d_{\mathcal{H}}) \\
&\quad \lesssim \log(n^{kd}\varepsilon^{-(2p+2\alpha-1)kd}) \le m\varepsilon^2,
\end{aligned}$$

where the last inequality holds since ε is bounded from below by a large multiple of $\varepsilon_{m,n} > (\log n/m)^{1/2} + (\log m/m)^{1/2}$. Thus, the entropy condition (8) is established.

To verify condition Equation (11), we note that for some constant $c > 0$,

$$\exp(2m\varepsilon_{m,n}^2)\sum_{j \ge M_m}\exp[-m\Psi_{\mathcal{G},n}(j\varepsilon_{m,n})/8]$$

$$\leq \exp(2m\varepsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-cm(j\varepsilon_{m,n})^{2(p+\alpha)}/8]$$
$$\lesssim \exp(2m\varepsilon_{m,n}^2) \exp[-cm(M_m\varepsilon_{m,n})^{2(p+\alpha)}/8],$$

where the right side of the above display vanishes if $(M_m\varepsilon_{m,n})^{p+\alpha}$ is a sufficiently large multiple of $\varepsilon_{m,n}$. This holds if we choose $M_m = M\varepsilon_{m,n}^{-(p+\alpha-1)/(p+\alpha)}$ for a large constant M. Equation (10) also holds.

It remains to verify Equation (9). By Theorem 6, as long as $n \gtrsim \log(1/\varepsilon_{m,n})$,

$$\begin{aligned}\log \Pi(G \in B_K(G_0, \varepsilon_{m,n})) &\geq c(c_0) \log(\varepsilon_{m,n}^2/n^3)^{kd} \\ &= c(c_0)kd(2\log \varepsilon_{m,n} - 3\log n).\end{aligned}$$

Equation (9) holds for a sufficiently large constant C because $\varepsilon_{m,n} > (\log n/m)^{1/2} + (\log m/m)^{1/2}$, and the constant that $n > \log m$.

Now, we can apply Theorem 4 to obtain a posterior contraction rate $M_m\varepsilon_{m,n} \asymp \varepsilon_{m,n}^{1/(p+\alpha)}$. □

Proof of Theorem 2. The proof proceeds in exactly the same way as Theorem 1, except that part (b) of Theorem 5 is applied instead of part (a). Accordingly, p is replaced by 1 in the rate exponent. □

Proof of Theorem 3 (Minimax lower bounds). (a) The proof involves the construction of a pair of polytopes in $\mathcal{G}^k$ whose set difference has small volume for a given Hausdorff distance. We consider two separate cases: (i) $k/2 \leq d$ and (ii) $k > 2d$.

If $k/2 \leq d$, consider a $q = \lfloor k/2 \rfloor$-simplex G_0 that is spanned by $q+1$ vertices in general positions. Take a vertex of G_0, say $\boldsymbol{\theta}_0$. Construct G_0' by chopping G_0 off by an ε-cap that is obtained by the convex hull of $\boldsymbol{\theta}_0$ and q other points which lie on the edges adjacent to $\boldsymbol{\theta}_0$, and of distance ε from $\boldsymbol{\theta}_0$. Clearly, G_0' has $2q \leq k$ vertices, so both G_0 and G_0' are in $\mathcal{G}^k$. We have $d_{\mathcal{H}}(G_0, G_0') \asymp \varepsilon$, and $\mathrm{vol}_q(G_0 \setminus G_0') \asymp \varepsilon^q$. Due to assumption (S4), $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \varepsilon^{q+\alpha'}$. We note here and for the rest of the proof, the multiplying constants in asymptotic inequalities depend only on r, R, δ of Properties A1 and A2.

If $k > 2d$, consider a d-dimensional polytope G_0 which has $k-d+1$ vertices in general positions. Construct G_0' in the same way as above (by chopping G_0 off by an ε-cap that contains a vertex $\boldsymbol{\theta}_0$ which has d adjacent vertices). Then, G_0' has $(k-d+1)-1+d = k$ vertices. Thus, both G_0' and G_0 are in $\mathcal{G}^k$. We have $d_{\mathcal{H}}(G_0, G_0') \asymp \varepsilon$, and $\mathrm{vol}_d(G_0 \setminus G_0') \asymp \varepsilon^d$. Due to assumption (S4), $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \varepsilon^{d+\alpha'}$.

To combine the two cases, let $q = \min(\lfloor k/2 \rfloor, d)$. We have constructed a pair of $G_0, G_0' \in \mathcal{G}^k$ such that $d_{\mathcal{H}}(G_0, G_0') \asymp \varepsilon$, and $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \varepsilon^{q+\alpha'}$. By Lemma 6, $K(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G_0'}) \lesssim nW_1(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim nV(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \leq Cn\varepsilon^{q+\alpha'}$ for some constant $C > 0$ independent of ε and n. Note that the second inequality in the above display is due to Theorem 6.15 of [18].

Applying the method due to Le Cam (cf. [22], Lemma 1), for any estimator $\hat{G} \in \mathcal{G}^*$,

$$\max_{G \in \{G_0, G'_0\}} P_{\mathcal{S}_{[n]}|G_0} d_{\mathcal{H}}(G, \hat{G}) \gtrsim \varepsilon \left(1 - \frac{1}{2} V(P^m_{\mathcal{S}_{[n]}|G_0}, P^m_{\mathcal{S}_{[n]}|G'_0})\right).$$

Here, $P^m_{\mathcal{S}_{[n]}|G_0}$ denotes the (product) distribution of the m-sample $\mathcal{S}^1_{[n]}, \ldots, \mathcal{S}^m_{[n]}$. Thus,

$$\begin{aligned} V^2(P^m_{\mathcal{S}_{[n]}|G_0}, P^m_{\mathcal{S}_{[n]}|G'_0}) &\le h^2(P^m_{\mathcal{S}_{[n]}|G_0}, P^m_{\mathcal{S}_{[n]}|G'_0}) \\ &= 1 - \int [P^m_{\mathcal{S}_{[n]}|G_0} P^m_{\mathcal{S}_{[n]}|G'_0}]^{1/2} \\ &= 1 - [1 - h^2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G'_0})]^m \\ &\le 1 - (1 - Cn\varepsilon^{q+\alpha'})^m. \end{aligned}$$

The last inequality is due to $h^2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G'_0}) \le K(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G'_0}) \le Cn\varepsilon^{q+\alpha'}$. Thus,

$$\max_{G \in \{G_0, G'_0\}} P_{\mathcal{S}_{[n]}|G_0} d_{\mathcal{H}}(G, \hat{G}) \gtrsim \varepsilon \left(1 - \frac{1}{2} [1 - (1 - Cn\varepsilon^{q+\alpha'})^m]^{1/2}\right).$$

Letting $\varepsilon^{q+\alpha'} = \frac{1}{Cmn}$, the right side of the previous display is bounded from below by $\varepsilon(1 - \frac{1}{2}(1 - 1/2)^{1/2})$.

(b) We employ the same construction of G_0 and G'_0 as in part (a). Using the argument used in the proof of Lemma 5, $K(p_{\mathcal{S}_{[n]}|G'_0}, p_{\mathcal{S}_{[n]}|G_0}) = \int \log[\operatorname{vol}_q G_0 / \operatorname{vol}_q G'_0] \, \mathrm{d}P_{\boldsymbol{\eta}|G_0} \le \int \log(1 + C\varepsilon^q) P_{\boldsymbol{\eta}|G_0} \lesssim \varepsilon^q$. So, $h^2(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G'_0}) \le K(p_{\mathcal{S}_{[n]}|G'_0}, p_{\mathcal{S}_{[n]}|G_0}) \lesssim \varepsilon^q$. Then, the proof proceeds as in part (a).

(c) Let G'_0 be a polytope such that $|\operatorname{extr} G'_0| = |\operatorname{extr} G_0| = k$ and $d_{\mathcal{H}}(G'_0, G_0) = \varepsilon$. By Lemma 2, $\operatorname{vol}_p(G_0 \triangle G'_0) = \mathrm{O}(\varepsilon)$, where $p = (k-1) \wedge d$. The proof proceeds as in part (a) to obtain $(1/mn)^{1/(1+\alpha')}$ rate for the lower bound under assumption (S4). Under assumption (S4′), as in part (b), the dependence on n can be removed to obtain $1/m$ rate. □

Proof of α-regularity of the Dirichlet-induced densities in Lemma 4. First, consider the case $k \le d+1$. For $\boldsymbol{\eta}^* \in G$, write $\boldsymbol{\eta}^* = \beta^*_1 \boldsymbol{\theta}_1 + \cdots + \beta^*_k \boldsymbol{\theta}_k$. For $\boldsymbol{\beta} \in \Delta^{k-1}$ such that $|\beta_i - \beta^*_i| \le \varepsilon/k$ for all $i = 1, \ldots, k-1$, we have $\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| = \|\sum_{i=1}^k (\beta_i - \beta^*_i)\boldsymbol{\theta}_i\| \le \sum_{i=1}^k |\beta_i - \beta^*_i| \le 2\sum_{i=1}^{k-1} |\beta_i - \beta^*_i| \le 2\varepsilon$. Here, we used the fact that $\|\boldsymbol{\theta}_i\| \le 1$ for any $\boldsymbol{\theta}_i \in \Delta^d$. Without loss of generality, assume that $\beta^*_k \ge 1/k$. Then, for any $\varepsilon < 1/k$

$$\begin{aligned} &P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \le 2\varepsilon) \\ &\quad \ge P_{\boldsymbol{\beta}}(|\beta_i - \beta^*_i| \le \varepsilon/k; i = 1, \ldots, k-1) \\ &\quad = \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \int_{\beta_i \in [0,1]; |\beta_i - \beta^*_i| \le \varepsilon/k; i = 1, \ldots, k-1} \prod_{i=1}^{k-1} \beta_i^{\gamma_i - 1} \left(1 - \sum_{i=1}^{k-1} \beta_i\right)^{\gamma_k - 1} \mathrm{d}\beta_1 \cdots \mathrm{d}\beta_{k-1} \end{aligned}$$

$$\geq \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \prod_{i=1}^{k-1} \int_{\max(\gamma_i^*-\varepsilon/k,0)}^{\min(\gamma_i^*+\varepsilon/k,1)} \beta_i^{\gamma_i-1}\,\mathrm{d}\beta_i \geq \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} (\varepsilon/k)^{k-1}.$$

Both the second and the third inequality in the previous display exploits the fact that since $\gamma_i \leq 1$, $x^{\gamma_i-1} \geq 1$ for any $x \leq 1$.

Now, consider the case $k > d+1$. The proof in the previous case applies, but we can achieve a better lower bound because the intrinsic dimensionality of G is d, not $k-1$. Since $\boldsymbol{\eta}^* \in \operatorname{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \subset \Delta^d$, by Carathéodory's theorem, $\boldsymbol{\eta}^*$ is the convex combination of $d+1$ or fewer extreme points among $\boldsymbol{\theta}_i$'s. Without loss of generality, let $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d+1}$ be such points, and write $\boldsymbol{\eta}^* = \beta_1^* \boldsymbol{\theta}_1 + \cdots + \beta_{d+1}^* \boldsymbol{\theta}_{d+1}$. Consider $\boldsymbol{\eta} = \beta_1 \boldsymbol{\theta}_1 + \cdots + \beta_k \boldsymbol{\theta}_k$, where $\|\beta_i - \beta_i^*| \leq \varepsilon/k$, for $i = 1, \ldots, d$, while $0 \leq \beta_i \leq \varepsilon/k$ for $i = d+2, \ldots, k$. Then, $\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \leq 2\varepsilon$. This implies that

$$\begin{aligned}
&P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \leq 2\varepsilon) \\
&\quad \geq P_{\boldsymbol{\beta}}(|\beta_i - \beta_i^*| \leq \varepsilon/k, i = 1, \ldots, d+1; |\beta_j| \leq \varepsilon/k, j > d+1) \\
&\quad \geq \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \prod_{i=1}^{d} \int_{\max(\gamma_i^*-\varepsilon/k,0)}^{\min(\gamma_i^*+\varepsilon/k,1)} \beta_i^{\gamma_i-1}\,\mathrm{d}\beta_i \prod_{i=d+2}^{k} \int_0^{\varepsilon/k} \beta_i^{\gamma_i-1}\,\mathrm{d}\beta_i \\
&\quad \geq \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} (\varepsilon/k)^{d+\sum_{i=d+2}^k \gamma_i} \Big/ \prod_{i=d+2}^{n} \gamma_i \gtrsim \varepsilon^{d+\sum_{i=1}^k \gamma_i}.
\end{aligned}$$

This concludes the proof. □

Appendix A: Proofs of geometric lemmas

Proof of Lemma 1. (a) Let $G = \operatorname{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ and $G' = \operatorname{conv}(\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_{k'})$. This part of the lemma is immediate from the definition by noting that for any $x \in G$, $d(x, G') \leq \min_j \|x - \boldsymbol{\theta}'_j\|$, while the maximum of $d(x, G')$ is attained at some extreme point of G.

(b) Let $d_{\mathcal{H}}(G, G') = \varepsilon$ for some small $\varepsilon > 0$. Take an extreme point of G, say $\boldsymbol{\theta}_1$. Due to A2, there is a ray emanating from $\boldsymbol{\theta}_1$ that intersects with the interior of G and the angles formed by the ray and all (exposed) edges incident to $\boldsymbol{\theta}_1$ are bounded from above by $\pi/2 - \delta$. Let x be the intersection between the ray and the boundary of $B_p(\boldsymbol{\theta}_1, \varepsilon)$.

Let H be a $p-1$-dimensional hyperplane in $\mathbb{R}^p$ that touches (intersects with) $B_p(\boldsymbol{\theta}_1, \varepsilon)$ at only x. Define $C(x)$, resp. $C_\varepsilon(x)$, to be the p-dimensional caps obtained by the intersection between G, resp. G_ε, with the half-space which contains $\boldsymbol{\theta}_1$ and which is supported by H. For any x' that lies in the intersection of H and a line segment $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_i]$, where $\boldsymbol{\theta}_i$ is another vertex of G, the line segment $[x, x'] \in H$ and $\|x - x'\| \leq \varepsilon \cot \delta$. Suppose that the ray emanating from x through x' intersects with $\operatorname{bd} G_\varepsilon$ at x''. Then, $\|x' - x''\| \leq \varepsilon / \sin \delta$, which implies that $\|x - x''\| \leq \varepsilon(\cot \delta + 1/\sin \delta)$ by triangle inequality. This entails that $\operatorname{Diam} C_\varepsilon(x) \leq C\varepsilon$, where $C = (1 + (\cot \delta + 1/\sin \delta)^2)^{1/2}$.

Now, $d_{\mathcal{H}}(G,G') = \varepsilon$ implies that $G' \cap B_p(\boldsymbol{\theta}_1, \varepsilon) \neq \varnothing$. There is an extreme point of G' in the half-space which contains $B(\boldsymbol{\theta}_1,\varepsilon)$ and is supported by H. But $G' \subset G_\varepsilon$, so there is an extreme point of G' in $C_\varepsilon(x)$. Hence, there is $\boldsymbol{\theta}'_j \in G'$ such that $\|\boldsymbol{\theta}'_j - \boldsymbol{\theta}_1\| \leq \mathrm{Diam}(C_\varepsilon(x)) \leq C\varepsilon$. Repeat this argument for all other extreme points of G to conclude that $d_{\mathcal{M}}(G,G') \leq C\varepsilon$. $\square$

Proof of Lemma 2. (a) Let $d_{\mathcal{H}}(G,G') = \varepsilon$. There exists either a point $x \in \mathrm{bd}\, G$ such that $G' \cap B_p(x, \varepsilon/2) = \varnothing$, or a point $x' \in \mathrm{bd}\, G'$ such that $G \cap B_p(x,\varepsilon/2) = \varnothing$. Without loss of generality, assume the former. Thus, $\mathrm{vol}_p\, G \triangle G' \geq \mathrm{vol}_p\, B_p(x,\varepsilon/2) \cap G$. Consider the convex cone emanating from x that circumscribes the p-dimensional spherical ball $B_p(\boldsymbol{\theta}_c, r)$ (whose existence is given by Property A1). Since $\|x - \boldsymbol{\theta}_c\| \leq R$, the angle between the line segment $[x, \boldsymbol{\theta}_c]$ and the cone's rays is bounded from below by $\sin\varphi \geq r/R$. So, $\mathrm{vol}_p\, B_d(x,\varepsilon/2) \cap G \geq c_1 \varepsilon^p$, where c_1 depends only on r, R, p.

(b) Let $d_{\mathcal{H}}(G,G') = \varepsilon$. Then $G' \subset G_\varepsilon$ and $G \subset G'_\varepsilon$. Take any point $x \in \mathrm{bd}\, G$, let x' be the intersection between $\mathrm{bd}\, G_\varepsilon$ and the ray emanating from $\boldsymbol{\theta}_c$ and passing through x. Let H_1 be a $p-1$ dimensional supporting hyperplane for G at x. There is also a supporting hyperplane H_2 of G' that is parallel to H_1 and of at most ε distance away from H_1. Since $\|\boldsymbol{\theta}_c - x\| \leq R$, while the distance from $\boldsymbol{\theta}_c$ to H_1 is lower bounded by r, the angle φ between vector $\boldsymbol{\theta}_c - x$ and the vector normal to H_1 satisfies $\cos\varphi \geq r/R$. This implies that $\|x' - x\| \leq \varepsilon/\cos\varphi \leq \varepsilon R/r$, so $\|x' - \boldsymbol{\theta}_c\|/\|x - \boldsymbol{\theta}_c\| \leq 1 + \varepsilon R/r^2$. In other words, $G_\varepsilon - \boldsymbol{\theta}_c \subset (1 + \varepsilon R/r^2)(G - \boldsymbol{\theta}_c)$. So, $\mathrm{vol}_p\, G' \setminus G \leq \mathrm{vol}_p\, G_\varepsilon \setminus G \leq [(1+\varepsilon R/r^2)^p - 1]\, \mathrm{vol}_p\, G \leq C_1 \varepsilon$, where C_1 depends only on r, R, p. We obtain a similar bound for $\mathrm{vol}_p\, G \setminus G'$, which concludes the proof. $\square$

Proof of Lemma 3. We provide a proof for case (a). Let $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ and $G' = \mathrm{conv}(\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_k)$, where G is fixed but G' is allowed to vary. Since G is fixed, it satisfies A1 and A2 for some constants r, R and δ (depending on G). Moreover, there is some $\varepsilon_0 = \varepsilon_0(G)$ depending only G such that as soon as $d_{\mathcal{H}}(G,G') \leq \varepsilon_0$, G' also satisfies A1 and A2 for constants $\delta' = \delta/2, r' = r/2, R' = 2R$.

Suppose that $d_{\mathcal{H}}(G,G') = \varepsilon$ such that $\varepsilon < \varepsilon_0$. By Lemma 1(b) for each vertex of G, say $\boldsymbol{\theta}_i$, there is a vertice of G', say $\boldsymbol{\theta}'_i$, such that $\boldsymbol{\theta}'_i \in B_p(\boldsymbol{\theta}_i, C_0\varepsilon)$ with $C_0 = C_0(G)$ depending only on δ. Moreover, there is at least one vertice of G, say $\boldsymbol{\theta}_1$, for which $\|\boldsymbol{\theta}'_1 - \boldsymbol{\theta}_1\| \geq \varepsilon$.

There are only three possible general positions for $\boldsymbol{\theta}'_1$ relatively to G. Either

(i) $\boldsymbol{\theta}'_1 \in G$, or
(ii) $\boldsymbol{\theta}'_1 \in 2\boldsymbol{\theta}_1 - G$, or
(iii) $\boldsymbol{\theta}'_1$ lies in a cone formed by all half-spaces supported by the $p-1$ dimensional faces adjacent to $\boldsymbol{\theta}_1$. Among these there is one half-space that contains G, and one that does not contain G.

If (i) is true, by Property A1, G has at least one face $S \supset \boldsymbol{\theta}_1$ such that the distance from $\boldsymbol{\theta}'_1$ to the hyperplane that provides support for S is bounded from below by $\varepsilon r/R$. Let $B \subset S$ be a homothetic transformation of S with respect to center $\boldsymbol{\theta}_1$ that maps $x \in S$ to $\tilde{x} \in B$ such that the ratio $\eta := \|\boldsymbol{\theta}_1 - \tilde{x}\|/\|\boldsymbol{\theta}_1 - x\|$ satisfies $1 - \eta = 2C_0\varepsilon/\min_{i\neq j}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| \in (0, 1/2)$. This is possible as soon as $\varepsilon < \min_{i \neq j} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\|/4C_0$. Then, for any $\boldsymbol{\theta}_j \in S$,

$j \neq 1$, under this transformation $\boldsymbol{\theta}_j \mapsto \tilde{\boldsymbol{\theta}}_j \in S$ for which $\|\tilde{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\| = (1-\eta)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_j\| \geq 2C_0\varepsilon$. Since $\|\boldsymbol{\theta}'_j - \boldsymbol{\theta}_j\| \leq C_0\varepsilon$, the construction of B implies that $\boldsymbol{\theta}'_j \notin B$. As a result, $B \cap G' = \varnothing$. Moreover, $\mathrm{vol}_{p-1} B = \eta^{p-1} \mathrm{vol}_{p-1} S \geq (1/2)^{p-1} \mathrm{vol}_{p-1} S \geq c_0(G)$, a constant depending only on G. Let Q be a p-pyramid which has apex $\boldsymbol{\theta}'_1$ and base B. It follows that $\mathrm{relint}\, Q \cap \mathrm{relint}\, G' = \varnothing$, which implies that $\mathrm{relint}\, Q \subset G \setminus G'$ (relint stands for the relative interior of a set). Hence, $\mathrm{vol}_p G \setminus G' \geq \mathrm{vol}_p Q \geq \frac{1}{p}\varepsilon r/R\, \mathrm{vol}_{p-1} B \geq \frac{1}{p}\varepsilon c_0(G) r/R$.

If (ii) is true, the same argument can be applied to show that $\mathrm{vol}_p(G' \setminus G) = \Omega(\varepsilon)$. If (iii) is true, a similar argument continues to apply: we obtain a lower bound for either $\mathrm{vol}_p G' \setminus G$ or $\mathrm{vol}_p G \setminus G'$. G has a face (supported by a hyperplane, say, H) such that the distance from $\boldsymbol{\theta}'_1$ to H is $\Omega(\varepsilon)$. If the half-space supported by H that contains $\boldsymbol{\theta}'_1$ but does not contain G, then $\mathrm{vol}_p G' \setminus G = \Omega(\varepsilon)$. If, on the other hand, the associated half-space does contain G, then $\mathrm{vol}_p G \setminus G' = \Omega(\varepsilon)$. The proof for case (b) is similar and is omitted. $\square$

Appendix B: Proof of abstract posterior contraction theorem

A key ingredient in the general analysis of convergence of posterior distributions is through establishing the existence of tests for subsets of parameters of interest. A test $\varphi_{m,n}$ is a measurable indicator function of the $m \times n$-sample $\mathcal{S}_{[n]}^{[m]} = (\mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)$ from an admixture model. For a fixed pair of convex polytopes $G_0, G_1 \in \mathcal{G}$, where $\mathcal{G}$ is a given subset of Δ^d, consider tests for discriminating G_0 against a closed Hausdorff ball centered at G_1. The following two lemmas on the existence of tests highlight the fundamental role of the Hellinger information:

Lemma 8. *Fix a pair of $(G_0, G_1) \in (\mathcal{G}^* \times \mathcal{G})$ and let $\delta = d_{\mathcal{H}}(G_0, G_1)$. Then, there exist tests $\{\varphi_{m,n}\}$ that have the following properties:*

$$P^m_{\mathcal{S}_{[n]}|G_0} \varphi_{m,n} \leq D \exp[-m\Psi_{\mathcal{G},n}(\delta)/8], \tag{16}$$

$$\sup_{G \in \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)} P^m_{\mathcal{S}_{[n]}|G}(1 - \varphi_{m,n}) \leq \exp[-m\Psi_{\mathcal{G},n}(\delta)/8]. \tag{17}$$

Here, $D := D(\Phi_{\mathcal{G},n}(\delta), \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2), d_{\mathcal{H}})$, i.e., the maximal number of elements in $\mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)$ that are mutually separated by at least $\Phi_{\mathcal{G},n}(\delta)$ in Hausdorff metric $d_{\mathcal{H}}$.

Proof. We begin the proof by noting that a direct application of standard results on existence of tests (cf. [11], Chapter 4) is not possible, due to the lack of convexity of the space of densities of $\mathcal{S}_{[n]}$ as G varies in some subset $\mathcal{G} \subset \mathcal{G}^*$, even if $\mathcal{G}$ is convex. This difficulty is overcome by appealing to a packing argument.

Consider a maximal $\Phi_{\mathcal{G},n}(\delta)$-packing in $d_{\mathcal{H}}$ metric for the set $\mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)$. This yields a set of $D = D(\Phi_{\mathcal{G},n}(\delta), \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2), d_{\mathcal{H}})$ elements $\tilde{G}_1, \ldots, \tilde{G}_D \in \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)$.

Next, we note the following fact: for any $t = 1, \dots, D$, if $G \in \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)$ and $d_{\mathcal{H}}(G, \tilde{G}_t) \le \Phi_{\mathcal{G},n}(\delta)$, then by the definition of Φ, $h^2(p_{\mathcal{S}_{[n]}|G}, p_{S_{[n]}|\tilde{G}_t}) \le \frac{1}{4}\Psi_{\mathcal{G},n}(\delta)$. By the definition of Hellinger information, $h^2(p_{\mathcal{S}_{[n]}|G_0}, p_{S_{[n]}|\tilde{G}_t}) \ge \Psi_{\mathcal{G},n}(\delta)$. Thus, by triangle inequality, $h(p_{\mathcal{S}_{[n]}|G_0}, p_{\mathcal{S}_{[n]}|G}) \ge \frac{1}{2}\Psi_{\mathcal{G},n}(\delta)^{1/2}$.

For each pair of $G_0, \tilde{G}_t$ there exist tests $\omega_n^{(t)}$ of $p_{\mathcal{S}_{[n]}|G_0}$ versus the Hellinger ball $\mathcal{P}_2(t) := \{p_{\mathcal{S}_{[n]}|G} | G \in \mathcal{G}^*; h(p_{\mathcal{S}_{[n]}|G}, p_{S_{[n]}|\tilde{G}_t}) \le \frac{1}{2} h(p_{\mathcal{S}_{[n]}|G_0}, p_{S_{[n]}|\tilde{G}_t})\}$ such that,

$$P^m_{\mathcal{S}_{[n]}|G_0} \omega^{(t)}_{m,n} \le \exp[-mh^2(p_{\mathcal{S}_{[n]}|G_0}, p_{S_{[n]}|\tilde{G}_t})/8],$$
$$\sup_{P_2 \in \mathcal{P}_2(t)} P_2^m(1 - \omega^{(t)}_{m,n}) \le \exp[-mh^2(p_{\mathcal{S}_{[n]}|G_0}, p_{S_{[n]}|\tilde{G}_t})/8].$$

Consider the test $\varphi_{m,n} = \max_{1 \le t \le D} \omega^{(t)}_{m,n}$, then

$$P^m_{\mathcal{S}_{[n]}|G_0} \varphi_{m,n} \le D \times \exp[-m\Psi_{\mathcal{G},n}(\delta)/8],$$
$$\sup_{G \in \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)} P^m_{\mathcal{S}_{[n]}|G}(1 - \varphi_{m,n}) \le \exp[-m\Psi_{\mathcal{G},n}(\delta)/8].$$

The first inequality is due to $\varphi_{m,n} \le \sum_{t=1}^D \omega^{(t)}_{m,n}$, and the second is due to the fact that for any $G \in \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \delta/2)$ there is some $d = 1, \dots, D$ such that $d_{\mathcal{H}}(G, \tilde{G}_t) \le \Phi_{\mathcal{G},n}(\delta)$, so that $p_{\mathcal{S}_{[n]}|G} \in \mathcal{P}_2(t)$. $\square$

Next, the existence of tests can be shown for discriminating G_0 against the complement of a closed Hausdorff ball:

Lemma 9. *Let $G_0 \in \mathcal{G}^*$ and $\mathcal{G} \subset \mathcal{G}^*$. Suppose that for some non-increasing function $D(\varepsilon)$, some $\varepsilon_{m,n} \ge 0$ and every $\varepsilon > \varepsilon_{m,n}$,*

$$\begin{aligned} &\sup_{G_1 \in \mathcal{G}} D(\Phi_{\mathcal{G},n}(\varepsilon), \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, \varepsilon/2), d_{\mathcal{H}}) \\ &\quad \times D(\varepsilon/2, \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_0, 2\varepsilon) \setminus B_W(G_0, \varepsilon), d_{\mathcal{H}}) \le D(\varepsilon). \end{aligned} \tag{18}$$

Then, for every $\varepsilon > \varepsilon_{m,n}$, and any $t_0 \in \mathbb{N}$, there exist tests $\varphi_{m,n}$ (depending on $\varepsilon > 0$) such that

$$P_{G_0} \varphi_{m,n} \le D(\varepsilon) \sum_{t=t_0}^{\lceil \mathrm{Diam}(\mathcal{G})/\varepsilon \rceil} \exp[-m\Psi_{\mathcal{G},n}(t\varepsilon)/8], \tag{19}$$

$$\sup_{G \in \mathcal{G}: d_{\mathcal{H}}(G_0, G) > t_0 \varepsilon} P_G(1 - \varphi_{m,n}) \le \exp[-m\Psi_{\mathcal{G},n}(t_0\varepsilon)/8]. \tag{20}$$

Proof. The proof consists of a standard peeling device (e.g., [7]) and a packing argument as in the previous proof. For a given $t \in \mathbb{N}$ choose a maximal $t\varepsilon/2$-packing for set $S_t =$

$\{G\colon\ t\varepsilon < d_{\mathcal{H}}(G_0, G) \le (t+1)\varepsilon\}$. This yields a set S'_t of at most $D(t\varepsilon/2, S_t, d_{\mathcal{H}})$ points. Moreover, every $G \in S_t$ is within distance $t\varepsilon/2$ of at least one of the points in S'_t. For every such point $G_1 \in S'_t$, there exists a test $\omega_{m,n}$ satisfying Equations (16) and (17), where δ is taken to be $\delta = t\varepsilon$. Take $\varphi_{m,n}$ to be the maximum of all tests attached this way to some point $G_1 \in S'_t$ for some $t \ge t_0$. Note that $G \in \mathcal{G} \subset \Delta^d$, so $t \le \lceil \mathrm{Diam}(\mathcal{G})/\varepsilon \rceil$. Then, by union bound, and the condition that $D(\varepsilon)$ is non-increasing,

$$\begin{aligned} P^m_{\mathcal{S}_{[n]}|G_0}\varphi_{m,n} &\le \sum_{t=t_0}^{\lceil \mathrm{Diam}(\mathcal{G})/\varepsilon \rceil} \sum_{G_1 \in S'_t} D(\Phi_{\mathcal{G},n}(t\varepsilon), \mathcal{G} \cap B_{d_{\mathcal{H}}}(G_1, t\varepsilon/2), d_{\mathcal{H}}) \\ &\qquad \times \exp[-m\Psi_{\mathcal{G},n}(t\varepsilon)/8] \\ &\le D(\varepsilon) \sum_{t \ge t_0} \exp[-m\Psi_{\mathcal{G},n}(t\varepsilon)/8], \end{aligned}$$

and

$$\begin{aligned} \sup_{G \in \bigcup_{u \ge t_0} S_u} P^m_{\mathcal{S}_{[n]}|G}(1-\varphi_n) &\le \sup_{u \ge t_0} \exp[-m\Psi_{\mathcal{G},n}(u\varepsilon)/8] \\ &\le \exp[-m\Psi_{\mathcal{G},n}(t_0\varepsilon)/8], \end{aligned}$$

where the last inequality is due the monotonicity of $\Psi_{\mathcal{G},n}(\cdot)$. □

Proof of the abstract posterior contraction theorem (Theorem 4). In this proof, to simplify notations denote $P_G := P_{\mathcal{S}_{[n]}|G}$. By a result of Ghosal *et al.* [7] (Lemma 8.1, page 524), for every $\varepsilon > 0, C > 0$ and every probability measure Π_0 supported on the set $B_K(G_0, \varepsilon)$ defined by Equation (7), we have,

$$P_{G_0}\left(\int \prod_{i=1}^m \frac{p_G(\mathcal{S}^i_{[n]})}{p_{G_0}(\mathcal{S}^i_{[n]})} \,\mathrm{d}\Pi_0(G) \le \exp(-(1+C)m\varepsilon^2) \right) \le \frac{1}{C^2 m\varepsilon^2}.$$

This entails that, by fixing $C = 1$, there is an event A_m with $P^m_{G_0}$-probability at least $1 - (m\varepsilon^2_{m,n})^{-1}$, for which there holds:

$$\int \prod_{i=1}^n p_G(\mathcal{S}^i_{[n]})/p_{G_0}(\mathcal{S}^i_{[n]}) \,\mathrm{d}\Pi(G) \ge \exp(-2m\varepsilon^2_{m,n})\Pi(B_K(G_0, \varepsilon_{m,n})). \tag{21}$$

Let $\mathcal{O}_m = \{G \in \mathcal{G}^*\colon\ d_{\mathcal{H}}(G_0, G) \ge M_m\varepsilon_{m,n}\}$. Due to Equation (8), the condition specified by Lemma 9 is satisfied by setting $D(\varepsilon) = \exp(m\varepsilon^2_{m,n})$ (constant in ε). Thus, there exist tests $\varphi_{m,n}$ for which Equations (19) and (20) hold with respect to $\mathcal{G} = \operatorname{supp}\Pi$ and the given G_0. Then,

$$\begin{aligned} P_{G_0}\Pi(G \in \mathcal{O}_m|\mathcal{S}^{[m]}_{[n]}) &= P_{G_0}[\varphi_{m,n}\Pi(G \in \mathcal{O}_m|\mathcal{S}^{[m]}_{[n]})] + P_{G_0}[(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m|\mathcal{S}^{[m]}_{[n]})] \\ &\le P_{G_0}[\varphi_{m,n}\Pi(G \in \mathcal{O}_m|\mathcal{S}^{[m]}_{[n]})] + P_{G_0}\mathbb{I}(A^c_m) \end{aligned}$$

$$+ P_{G_0}[(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m | \mathcal{S}^{[m]}_{[n]})\mathbb{I}(A_m)].$$

Applying Lemma 9, the first term in the preceding display is bounded above by

$$P_{G_0}\varphi_{m,n} \le D(\varepsilon_{m,n}) \sum_{j \ge M_m} \exp[-m\Psi_{\mathcal{G},n}(j\varepsilon_{m,n})/8] \to 0,$$

thanks to Equation (11). The second term in the above display is bounded by $(m\varepsilon^2_{m,n})^{-1}$ by the definition of A_m, so this term vanishes. It remains to show that third term in the display also vanishes as $m \to \infty$. By Bayes' rule,

$$\Pi(G \in \mathcal{O}_m | \mathcal{S}^{[m]}_{[n]}) = \frac{\int_{\mathcal{O}_m} \prod_{i=1}^m p_G(\mathcal{S}^i_{[n]})/p_{G_0}(\mathcal{S}^i_{[n]}) \, d\Pi(G)}{\int \prod_{i=1}^m p_G(\mathcal{S}^i_{[n]})/p_{G_0}(\mathcal{S}^i_{[n]}) \, d\Pi(G)},$$

and then obtain a lower bound for the denominator by Equation (21). For the nominator, by Fubini's theorem:

$$\begin{aligned} &P_{G_0} \int_{\mathcal{O}_m \cap \mathcal{G}} (1-\varphi_{m,n}) \prod_{i=1}^m p_G(\mathcal{S}^i_{[n]})/p_{G_0}(\mathcal{S}^i_{[n]}) \, d\Pi(G) \\ &= \int_{\mathcal{O}_m \cap \mathcal{G}} P_G(1-\varphi_{m,n}) \, d\Pi(G) \le \exp[-m\Psi_{\mathcal{G},n}(M_m\varepsilon_{m,n})/8], \end{aligned} \tag{22}$$

where the last inequality is due to Equation (20).

Now, combining bounds (22) and (21) with condition (10), we obtain:

$$\begin{aligned} &P_{G_0}(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m | \mathcal{S}^{[m]}_{[n]})\mathbb{I}(A_m) \\ &\le \frac{\exp[-m\Psi_{G_0,n}(\mathcal{G}_m, M_m\varepsilon_{m,n})/8]}{\exp(-2m\varepsilon^2_{m,n})\Pi(B_K(G_0,\varepsilon_{m,n}))}. \end{aligned}$$

The upper bound in the preceeding display converges to 0 by Equation (10), thereby concluding the proof. □

Acknowledgements

This research was supported in part by NSF Grants CCF-1115769 and OCI-1047871. The author wish to thank the anonymous referees for helpful comments, and Qiaozhu Mei and Jian Tang for stimulating discussions which helped to motivate this work.

References

[1] Anandkumar, A., Foster, D., Hsu, D., Kakade, S. and Liu, Y.K. (2012). Two SVDs suffice: Spectral decompositions for probabilistic topic modeling and latent Dirichlet allocation. arXiv:1204.6703.

[2] Arora, S., Ge, R. and Moitra, A. (2012). Learning topic models – going beyond SVD. arXiv:1204.1956.

[3] Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[4] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[5] Chen, J.H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665

[6] Evans, L.C. and Gariepy, R.F. (1992). *Measure Theory and Fine Properties of Functions. Studies in Advanced Mathematics.* Boca Raton, FL: CRC Press. MR1158660

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

[8] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[9] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics. Springer Series in Statistics.* New York: Springer. MR1992245

[10] Ishwaran, H., James, L.F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332. MR1946579

[11] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics.* New York: Springer. MR0856411

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

[13] Pritchard, J., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

[14] Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 689–710. MR2867454

[15] Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory. Encyclopedia of Mathematics and Its Applications* **44**. Cambridge: Cambridge Univ. Press. MR1216521

[16] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337

[17] Toussile, W. and Gassiat, E. (2009). Variable selection in model-based clustering using multilocus genotype data. *Adv. Data Anal. Classif.* **3** 109–134. MR2551051

[18] Villani, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **338**. Berlin: Springer. MR2459454

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

[20] Walker, S.G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866

[21] Wong, W.H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570

[22] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. New York: Springer. MR1462963

Received December 2012 and revised September 2013