Posterior contraction of the population polytope in finite admixture models [1]

XuanLong Nguyen
xuanlong@umich.edu

Technical report 528
Department of Statistics
University of Michigan

May 24, 2012

Abstract

We study the posterior contraction behavior of the latent population structure that arises in admixture models as the amount of data increases. An admixture model — alternatively known as a topic model — specifies k populations, each of which is characterized by a Δ^d-valued vector of frequencies for generating a set of discrete values in $\{0, 1, \ldots, d\}$. The population polytope is defined as the convex hull of the k frequency vectors. Under the admixture specification, each of m individuals generates an i.i.d. frequency vector according to a probability distribution defined on the (unknown) population polytope G_0, and then generates n data points according to the sampled frequency vector. Given a prior distribution over the space of population polytopes, we establish rates at which the posterior distribution contracts to G_0, under the Hausdorff metric and a minimum matching Euclidean metric, as the amount of data $m \times n$ tends to infinity. Rates are obtained for the overfitted setting, i.e., when the number of extreme points of G_0 is bounded above by k, and for the setting in which the number of extreme points of G_0 is known. Minimax lower bounds are also established. Our analysis combines posterior asymptotics techniques for the estimation of mixing measures in hierarchical models with elementary arguments in convex geometry.

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 [Pritchard et al., 2000] and Blei, Ng and Jordan [Blei et al., 2003]. The former set of authors applied their modeling to population genetics, while the latter considered applications in text processing, where their models are more widely known as the "latent

[1] AMS 2000 subject classification. Primary 62F15, 62G05; secondary 62G20.

Key words and phrases: latent mixing measures, convex polytope, population structure, topic simplex, Bayesian estimation, posterior consistency, rates of convergence, latent Dirichlet allocation.

This research was supported in part by NSF grants CCF-1115769 and OCI-1047871. The author thank Qiaozhu Mei and Jian Tang for stimulating discussions on topic models, which helped to motivate this work.

Dirichlet allocation" topic models. 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.

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 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 refered 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 j-th population. In text document modeling, $\boldsymbol{\theta}_j$ represents the frequencies of words generated by the j-th topic, while an individual is a document, i.e., a collection of words. 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 [Chen, 1995]. This result was subsequentially extended to a Bayesian estimation setting [Ishwaran et al., 2001]. Nguyen recently obtained posterior contraction rates of mixing measures in several finite and infinite mixture models for multivariate and continuous data [Nguyen, 2012]. This issue has also attracted increased attention in machine learning. Notably, there are a couple of very recent papers that study the convergence of the population structure arising in admixture models for certain computationally efficient learning algorithms based on matrix factorization techniques [Arora et al., 2012, Anandkumar et al., 2012]. Their results will be briefly discussed in the sequel.

There are several interesting aspects that arise in the convergence analysis of admixture models for categorical data. First, it is not unreasonable to suspect that in general the population structure represented by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ may be estimated up to its convex hull $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. 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 and estimate. Throughout this paper, G will be the focus of our study, and is referred to as the *population polytope*, whose geometric

properties will be intensively exploited using ideas from convex geometry. We adopt the Hausdorff metric for evaluating the posterior contraction rates of the population polytope. Hausdorff is a natural choice for analyzing estimators of sets (e.g., [Dumbgen and Walther, 1996, Tsybakov, 1997, Singh et al., 2009]). One virtue of estimation via the admixture model is that it is possible to estimate not only the boundary of G, but also all its extreme points. Indeed, our analysis is also achieved for a "minimum-matching" metric (defined in Section 2) which measures how well each of the extreme points is estimated, under some geometric identifiability conditions.

The second aspect is concerned with the analysis of posterior contraction of latent mixing measures in a hierarchical mixture model. The general framework of posterior asymptotics for density estimation has been well-established (see, e.g., [Ghosal et al., 2000] and their list of references up to year 2000). The analysis of mixing measure estimation in multilevel models remains generally quite challenging. In the context of admixture models, suppose that there are m individuals, each of which is observed via n sampled data points, then typically both m and n are required to increase in order to achieve posterior contraction. In an overfitted setting, i.e., 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 $\left(\frac{\log m}{m} \vee \frac{\log n}{n} \vee \frac{\log n}{m}\right)^{\frac{1}{2(p+\alpha)}}$, 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 $\boldsymbol{\eta}$. However, 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 of exponent $\frac{1}{2(1+\alpha)}$. The quantity $\log n/m$ in the rate is non-standard and appears particularly interesting, which reflects the interactions between multiple levels in the latent hierarchy of the admixture model. This appears to suggest that n may not grow too fast relative to m. We also establish minimax lower bounds for both settings. In the overfitted setting the obtained lower bound is $(mn)^{-\frac{1}{q+\alpha}}$, where $q = \lfloor k/2 \rfloor \wedge d$, unless additional constraints are imposed on the prior. Moreover, if the distribution for the frequency vector $\boldsymbol{\eta}$ is uniform, we obtain a minimax lower bound of the order $m^{-1/q}$, which does not depend on n, the amount of data that provide support for the bottom level in the model hierarchy.

The main technical ingredients of our convergence analysis involve a number of inequalities which establish the relationship between Hausdorff distance (and equivalently, the minimum matching Euclidean distance) between a given pair of population polytopes G, G', and several divergences (e.g., Kullback-Leibler divergence or total variational distance) between the induced densities of the $m \times n$ data points. These bounds are derived via elementary arguments in convex geometry [Schneider, 1993]. The general posterior contraction proof strategy consists of an existence-of-tests argument, which is turned into a convergence theorem in a standard way [Ghosal et al., 2000]. Because we work in the Hausdorff metric on the space of population polytopes (as opposed to the Hellinger metric on the space of data densities), we are forced to deal with non-convex subsets in the space

of convex polytopes. As a result, the power of the tests are controlled in terms of the so-call *Hellinger information* of the Hausdorff metric for a given subset of polytopes, which appears in both the exponent and the constant of the power bound. 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.

As mentioned earlier the existing works on admixture models include the recent papers by Arora et al [Arora et al., 2012] and Anandkumar et al [Anandkumar et al., 2012]. Both sets of authors analyzed specific learning algorithms for recovering the population structure by taking the viewpoint of matrix factorization. They both work on the setting where the number of extreme points k is known, and $k \ll d$. Arora et al Arora et al. [2012] additionally required interesting but very special conditions on the nature of the extreme points, for which a polynomial time learning algorithm exists, and established an estimation error rate for the algorithm. Anandkumar et al Anandkumar et al. [2012] proposed a novel moment-based estimation method and obtained a consistency result. By contrast, we analyze general Bayesian estimation without concerning a specific inference algorithm. (This goes without saying that under general conditions the posterior contraction entails convergence of procedures such as the maximum likelihood estimation method). The posterior contraction rates and minimax results obtained in this paper appear new. The posterior asymptotics and convex geometric techniques developed here are quite distinct from the existing works.

The remainder of the paper is organized as follows. The model and the statement of main results are described in Section 2. Section 3 describes the basic geometric assumptions and their consequences. A general theorem for posterior contraction is formulated in Section 4, whose conditions are verified in the subsequent sections. Section 5 proves a contraction result which helps to establish a key lower bound on the Hellinger information, while Section 6 provides a lower bound on the Kullback-Leibler neighborhood of the prior support. Proofs of main theorems and other technical lemmas are presented in Section 7.

Notations. $B_p(\boldsymbol{\theta}, r)$ denotes a p-dimensional radius r Euclidean ball centered at $\boldsymbol{\theta}$. G_ϵ denotes the Minkowsky sum $G_\epsilon := G + B_{d+1}(\mathbf{0}, \epsilon)$. $\mathrm{bd}\, G, \mathrm{extr}\, G, \mathrm{Diam}\, G, \mathrm{aff}\, G, \mathrm{vol}_p\, G$ 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. $N(\epsilon, G, d_{\mathcal{H}})$ denotes the covering number of G in Hausdorff metric $d_{\mathcal{H}}$. $D(\epsilon, G, d_{\mathcal{H}})$ is the packing number of G in Hausdorff metric. Several divergence measures for probability distributions are employed: $K(p,q), h(p,q), V(p,q)$ denote the Kullback-Leibler divergence, Hellinger and total variational 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 $K_2 = \int p[\log(p/q)]^2$. Several probability distributions are analyzed throughout the paper: $P_{\boldsymbol{\beta}}, P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}\times\mathcal{S}|G}$ are the distribution of $\boldsymbol{\beta}$, the distribution of $\boldsymbol{\eta}$ given G, and the joint distribution of $\boldsymbol{\eta}$ and an n-sample $\mathcal{S}_{[n]}$ given G, respectively. P_G denotes the marginal density of $\mathcal{S}_{[n]}$ given G (by having $\boldsymbol{\eta}$ integrated out). The lower-case $p_{\boldsymbol{\beta}}, p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}\times\mathcal{S}|G}, p_G$

are the corresponding densities.

2 Statement of main results

Model description. As mentioned in the Introduction, a central object of the admixture model is a *population polytope* represented by $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ are k points in the 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 $\boldsymbol{\eta} = \beta_1\boldsymbol{\theta}_1 + \ldots, \beta_k\boldsymbol{\theta}_k$, 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 Pritchard et al. [2000] and Blei et al. [2003] used the Dirichlet distribution). This induces a probability distribution $P_{\boldsymbol{\eta}|G}$ whose support is the convex set G.

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 $\mathrm{Mult}(\boldsymbol{\eta}_i)$ 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 $l = 0, \ldots, d$. The joint distribution of $\boldsymbol{\eta}$ and $\mathcal{S}_{[n]}$ (without using the superscript i for indexing a specific individual) is denoted by $P_{\boldsymbol{\eta}\times\mathcal{S}|G}$ and its density $p_{\boldsymbol{\eta}\times\mathcal{S}|G}$. The marginal distribution of $\mathcal{S}_{[n]}$ and its density are denoted by P_G and p_G, respectively.

Admixture models are customarily introduced in an equivalent way as follows [Blei et al., 2003, Pritchard et al., 2000]: 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 $Z_{ij}|\boldsymbol{\beta} \overset{iid}{\sim} \mathrm{Mult}(\boldsymbol{\beta})$. 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 that a 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 a "true" population polytope $G_0 = \mathrm{conv}(\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_k)$. Under the Bayesian estimation framework, G is 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 will also 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 $\mathcal{G}^* = \cup_{2 \leq k < \infty} \mathcal{G}^k$ 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:

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

Here, $G_\epsilon = G + B_{d+1}(\mathbf{0}, \epsilon) := \{\boldsymbol{\theta} + e | \boldsymbol{\theta} \in G, e \in \mathbb{R}^{d+1}, \|e\| \leq 1\}$, 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 $\boldsymbol{\eta}_0 \in \operatorname{bd} G$,

$$P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \epsilon) \gtrsim \epsilon^\alpha \operatorname{vol}_p(G \cap B_{d+1}(\boldsymbol{\eta}_0, \epsilon)).$$

where p is the number of dimensions of the affine space $\operatorname{aff} G$ that spans G.

Assumptions.

(S0) Geometric properties (A1) and (A2) listed in Section 3 are satisfied uniformly for all G in the support of the prior Π.

(S1) The prior support for 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 $\min_{l=0,\ldots,d} \boldsymbol{\theta}_{j,l} > c_0$ for all $j = 1, \ldots, k$.

(S2) Each $\boldsymbol{\theta}_j$ has a Lesbegue density function on its support that is bounded away from 0.

(S3) $\boldsymbol{\beta}$ is distributed (a priori) according to a symmetric probability distribution $P_{\boldsymbol{\beta}}$ on Δ^{k-1}.

(S4) $P_{\boldsymbol{\beta}}$ induces a family of distributions $\{P_{\boldsymbol{\eta}|G} | G \in \mathcal{G}^k\}$ that is α-regular.

Theorem 1. *Fix $G_0 \in \mathcal{G}^k$. Let $p = (k-1) \wedge d$. Under Assumptions (S0–S4) of the admixture model, we have:*

$$\Pi(d_{\mathcal{M}}(G_0, G) \geq \delta_{m,n} | \mathcal{S}^{[m]}_{[n]}) \to 0 \tag{1}$$

in P_{G_0}-probability as both m and n tend to infinity. Here,

$$\delta_{m,n} = \left[\frac{C_1 \log m}{m} \vee \frac{C_2 \log n}{n} \vee \frac{C_3 \log n}{m} \right]^{\frac{1}{2(p+\alpha)}},$$

for some positive constants C_1, C_2, C_3 that are independent of m and n. The same statement holds for the Hausdorff metric $d_{\mathcal{H}}$.

Remarks. (i) The geometric assumptions (S0) and their consequences are presented in the next section. (S0)(S1) and (S2) are very mild assumptions often observed in practice. (S4) is a standard assumption that holds for a range of α, when $P_{\boldsymbol{\beta}|\gamma}$ is a Dirichlet distribution (see Lemma 6), but there may be other choices. The assumption in (S3) that $P_{\boldsymbol{\beta}}$ is symmetric is relatively strong, but it has been commonly used in practice (e.g., symmetric Dirichlet distributions, including the uniform distribution). It may be difficult to try to relax this assumption if one insists on using Hausdorff metric, see the remark following the statement of Lemma 9.

(ii) 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 polytope to simplify the exposition and the results.

(iii) 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 $\log n/m$, which suggests that n may not grow too fast compared to m. This can be explained by the observation that as n increases, the space of the data vectors $\mathcal{S}_n$ increases in dimensions. Consequentially, the prior support gets "thinner" in probability mass, which in turn affects the posterior contraction rate. From a hierarchical modeling viewpoint, this provides a cautionary tale about balancing between 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.

(iv) The exponent $\frac{1}{2(p+\alpha)}$ appears quite weak. The following theorem shows that it is possible to achieve a parametric rate if additional constraints are imposed either on the true G_0, or the prior Π:

Theorem 2. *Fix $G_0 \in \mathcal{G}^k$. Assume (S0–S4), 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 (including G_0), are bounded from below by r_0.*

Then, the posterior contraction given in Eq. (1) *holds with*

$$\delta_{m,n} = \left[\frac{C_1 \log m}{m} \vee \frac{C_2 \log n}{n} \vee \frac{C_3 \log n}{m} \right]^{\frac{1}{2(1+\alpha)}},$$

for some positive constants C_1, C_2, C_3 that are independent of m and n. The same statement holds for the Hausdorff metric $d_{\mathcal{H}}$.

The next result shows that the nonparametrics-like rates obtained in Theorem 1 may be not too far off from a minimax optimal rate. 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 (S3) and (S4) on $P_{\boldsymbol{\beta}|\gamma}$ by either one of the following assumptions on $P_{\boldsymbol{\eta}|G}$:

(S5) For any pair of p-dimensional polytopes $G' \subset G$ that satisfy property A1,

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

(S5') For any p-dimensional polytope G, $P_{\boldsymbol{\eta}|G}$ is the uniform distribution on G.[2]

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 $\boldsymbol{\eta}_i \overset{iid}{\sim} P_{\boldsymbol{\eta}|G}$. For each $j = 1, \ldots, n$, draw $\mathcal{S}^i_{[n]} = (X_{ij})_{j=1}^n \overset{iid}{\sim} \text{Mult}(\boldsymbol{\eta}_i)$.

Theorem 3. *(a) Let $q = \lfloor k/2 \rfloor \wedge d$. Under Assumption (S5), we have*

$$\inf_{\hat{G}} \sup_{G_0 \in \mathcal{G}^k} P_{G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn} \right)^{\frac{1}{q+\alpha}}.$$

The infimum is taken over all estimators $\hat{G} = \hat{G}(\mathcal{S}^{[m]}_{[n]})$ of G, on the basis of the $m \times n$-data set $\mathcal{S}^{[m]}_{[n]}$.

(b) Let $q = \lfloor k/2 \rfloor \wedge d$. Under Assumption (S5'), we have

$$\inf_{\hat{G}} \sup_{G_0 \in \mathcal{G}^k} P_{G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{m} \right)^{\frac{1}{q}}.$$

(c) Assume (S5'), and that either condition (a) or (b) of Theorem 2 holds, then

$$\inf_{\hat{G}} \sup_{G_0} P_{G_0} d_{\mathcal{H}}(G_0, \hat{G}) \gtrsim \left(\frac{1}{mn} \right)^{\frac{1}{1+\alpha}}.$$

Furthermore, if (S5) is replaced by (S5'), the lower bound becomes $1/m$.

[2] It is straightforward to show that (S5') entails (S5) for $\alpha = 0$, by invoking Lemma 2 (b).

Remarks. (i) There remain some gaps between the posterior contraction rates in Theorem 1 and Theorem 2 and the minimax lower bounds in Theorem 3, especially in the rate exponent (by a factor of 2 or 4 if we allow $m \asymp n$). This may be partly attributable to the slightly enlarged models considered in Theorem 3, due to the relaxed parameterization. We do not know if the gaps are also due to our proof techniques, or the nature of the Bayesian estimation in the studied models.

(ii) 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). In practice, we are more likely to be in the overfitted setting than knowing the extract number of extreme points. Thus, it is important to impose a lower bound in the prior on the pairwise distances between the extreme points of the population polytope.

(iii) The results in part (b) and (c) under assumption (S5') 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.

(iv) It is worth mentioning that the exponent for m in the lower bounds $(1/mn)^{1/(d+\alpha)}$ of part (a) (when $k \geq 2d$) and $(1/m)^{1/d}$ of part (b) (when $k \geq 2d$ and (S5') holds) appear compatible to a general minimax optimal rate $(\log m/m)^{1/(d+\alpha)}$ for estimating the support of the density function $p_{\boldsymbol{\eta}|G}$, assuming that an iid m-sample of $\boldsymbol{\eta}$ is *directly* observed [Tsybakov, 1997, Singh et al., 2009]. A word of caution about making this comparison is that while the latter problem is easier due to the direct observations of $\boldsymbol{\eta}$, the density support for $\boldsymbol{\eta}$ is not required to be convex as is the case with admixture models.

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-obstute corners): For some small $\delta > 0$, the angle between every pair of adjacent edges of G is less than $\pi - \delta$.

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, p.

Proof. (a) Let $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ and $G' = \mathrm{conv}(\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_{k'})$. 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 $d_{\mathcal{H}}(G, G') = \epsilon$ for some small $\epsilon > 0$. Take an extreme point of G, say $\boldsymbol{\theta}_1$. Due to A2, there is a small constant $\delta' > 0$ depending only on δ, p, such that there is a ray emanating from $\boldsymbol{\theta}_1$ that intersects with 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 $B_p(\boldsymbol{\theta}_1, \epsilon)$.

Let H be a $p-1$-dimensional hyperplane in $\mathbb{R}^p$ that touches $B_p(\boldsymbol{\theta}_1, \epsilon)$ at x. Define $C(x)$, resp. $C_\epsilon(x)$, 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, the line segment $[x, x'] \in H$ and $\|x - x'\| \leq \epsilon \cot \delta'$. Suppose that the ray emanating from x through x' intersects with $\mathrm{bd}\, G_\epsilon$ at x''. Then, $\|x' - x''\| \leq \epsilon / \sin \delta'$, which implies that $\|x - x''\| \leq O(\epsilon)$. This entails that $\mathrm{Diam}\, C_\epsilon(x) \leq O(\epsilon)$.

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

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) $\mathrm{vol}_p\, G \triangle G' \leq C_1 d_{\mathcal{H}}(G, G')$.

Both bounds in this lemma are probably well-known in the folklore of convex geometry. For instance, part (b) is similar to (but not precisely the same as) Lemma 2.3.6. from Schneider [1993]. We include a proof below due to the absence of a more direct reference.

Proof. (a) Let $d_{\mathcal{H}}(G, G') = \epsilon$. There exists either a point $x \in \mathrm{bd}\, G$ such that $G' \cap B_p(x, \epsilon/2) = \emptyset$, or a point $x' \in \mathrm{bd}\, G'$ such that $G \cap B_p(x, \epsilon/2) = \emptyset$. Without loss of generality, assume the former. Thus, $\mathrm{vol}_p\, G \triangle G' \geq \mathrm{vol}_p\, B_p(x, \epsilon/2) \cap G$. 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 Condition 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, $\mathrm{vol}_p\, B_d(x, \epsilon/2) \cap G \gtrsim \epsilon^p$.

(b) Let $d_{\mathcal{H}}(G, G') = \epsilon$. Then $G' \subset G_\epsilon$ and $G \subset G'_\epsilon$. Take any point $x \in \operatorname{bd} G$, let x' be the intersection between $\operatorname{bd} G_\epsilon$ 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 $\|x' - x\| \leq \epsilon/\cos\varphi \leq \epsilon R/r$, so $\|x' - \boldsymbol{\theta}_c\|/\|x - \boldsymbol{\theta}_c\| \leq 1 + \epsilon R/r^2$. In other words, $G_\epsilon - \boldsymbol{\theta}_c \subset (1 + \epsilon R/r^2)(G - \boldsymbol{\theta}_c)$. So, $\operatorname{vol}_p G' \setminus G \leq \operatorname{vol}_p G_\epsilon \setminus G \leq [(1 + \epsilon R/r^2)^p - 1] \operatorname{vol}_p G \lesssim \epsilon$. We obtain a similar bound for $\operatorname{vol}_p G \setminus G'$, which concludes the proof. □

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 $\|x - \boldsymbol{\theta}_i\| = \epsilon$. Let G' be the convex hull of all such x's and the remaining vertices of G. Clearly, $d_{\mathcal{H}}(G, G') = O(\epsilon)$, and $\operatorname{vol}_p G \setminus G' \leq O(\epsilon^p)$. Thus, for the collection of convex polytopes G' constructed in this way, $\operatorname{vol}_p(G \triangle G') \asymp d_{\mathcal{H}}(G, G')^p$. The upper bound in part (b) is also tight for a broad class of convex polytopes, as exemplified by the following lemma.

Lemma 3. *Fix a polytope G (i.e., $|\operatorname{extr} G| = k < \infty$). G' is an arbitrary polytope that satisfies properties A1and A2, and suppose that either one of the following conditions holds:*

- *(a) $|\operatorname{extr} G'| = k$, 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 c_2 such that

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

if $d_{\mathcal{H}}(G, G')$ is sufficiently small. c_2 depends only on G in case (a) and G, r_0 in case (b).

Proof. We provide a proof for case (a). Let $G = \operatorname{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ and $G' = \operatorname{conv}(\boldsymbol{\theta}'_1, \ldots, \boldsymbol{\theta}'_k)$. Since G is fixed, both G and G' satisfies A2and A1(for some fixed $\boldsymbol{\theta}_c$, radii r, R such that $0 < r < R$). Let $d_{\mathcal{H}}(G, G') = \epsilon$ for a small $\epsilon > 0$. Due to property A2and Lemma 1 (b) for each vertex of G, say $\boldsymbol{\theta}_i$, there is a vertice of G', say $\boldsymbol{\theta}'_i$, such that $\boldsymbol{\theta}'_i \in B_p(\boldsymbol{\theta}_i, C_0\epsilon)$ for some constant $C_0 > 1$. Moreover, there is at least one vertice of G, say $\boldsymbol{\theta}_1$, for which $\|\boldsymbol{\theta}'_1 - \boldsymbol{\theta}_1\| \geq \epsilon$.

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 at least one half-space that contains G, and one that does not contain G.

If (i) is true, by property A1, there is 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 $\epsilon r/R$. Let $B \subset S$ be a homothetic transformation of S with respect to center $\boldsymbol{\theta}_1$ that maps $x \in S$ to $x' \in B$ such that $\|\boldsymbol{\theta}_1 - x'\|/\|\boldsymbol{\theta}_1 - x\| = \frac{r/R}{2(r/R+C_0)}$ (constant 2 in the denominator can be replaced by any other constant greater than 1). It is simple to verify that for sufficiently small ϵ, $B \cap G' = \emptyset$. Moreover, $\mathrm{vol}_{p-1} B$ is a multiple of $\mathrm{vol}_{p-1} S$ (independent of ϵ), so it is bounded from below by a constant. Let Q be a p-pyramid which has apex $\boldsymbol{\theta}_1'$ and base B. It follows that $\mathrm{relint}\, Q \cap \mathrm{relint}\, G' = \emptyset$, which implies that $\mathrm{relint}\, Q \subset G \setminus G'$. ($\mathrm{relint}$ stands for the relative interior of a set). Hence, $\mathrm{vol}_p\, G \setminus G' \geq \mathrm{vol}_p\, Q \geq \frac{1}{p}\epsilon r/R\, \mathrm{vol}_{p-1}\, B \gtrsim \epsilon$.

If (ii) is true, the same argument can be applied to show that one can construct a p-pyramid contained in $G' \setminus G$ such that whose volume is bounded from below by a multiple of ϵ. If (iii) is true, a similar argument continues to apply, but we may have either $\mathrm{vol}_p\, G' \setminus G$ or $\mathrm{vol}_p\, G \setminus G' \gtrsim \epsilon$, depending on the relative distance of $\boldsymbol{\theta}_1'$ to the hyperplanes that provide the support for the $p-1$ dimensional faces adjacent to $\boldsymbol{\theta}_1$. In particular, if there is a face (supported by hyperplane H) such that the distance from $\boldsymbol{\theta}_1'$ to H is $\Omega(\epsilon)$, but the half-space supported by H that contains $\boldsymbol{\theta}_1'$ but does not contain G, then $\mathrm{vol}_p\, G' \setminus G \gtrsim \epsilon$. If, on the other hand, the associated half-space does contain G, then $\mathrm{vol}_p\, G \setminus G' \gtrsim \epsilon$.

The proof for case (b) is similar and is omitted. □

4 A general 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. Define the Hausdorff ball as:

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

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

$$C_{k,n}(\mathcal{G}, r) := \inf_{G \in \mathcal{G};\, d_{\mathcal{H}}(G_0, G) \geq r/2} h^2(p_{G_0}, p_G). \tag{2}$$

The following two lemmas on the existence of tests highlight the fundamental role of the Hellinger information of $d_{\mathcal{H}}$ metric for a given set. Both lemmas require the following condition:

Condition C. $\mathcal{G}$ is a subset of $\mathcal{G}^*$. All $G \in \mathcal{G}$ are bounded away from the boundary from Δ^d. That is, there is a constant $c_0 > 0$ such that $\min_{l=0,\ldots,d} \eta_l \geq c_0$ for all $\boldsymbol{\eta} =$

$(\eta_0, \ldots, \eta_d) \in \mathcal{G}$. Moreover, all $G \in \mathcal{G}$ satisfy geometric A2, so that Lemma 1 (b) holds for some constant C_0.

Lemma 4. *Suppose that $\mathcal{G}$ satisfies Condition C. Fix a pair of $(G_0, G_1) \in (\mathcal{G}^k \times \mathcal{G})$ and let $r = d_{\mathcal{H}}(G_0, G_1)$. Then, there exist tests $\{\varphi_{m,n}\}$ that have the following properties:*

$$P_{G_0}\ \varphi_{m,n} \leq D \exp[-mC_{k,n}(\mathcal{G}, r)] \tag{3}$$

$$\sup_{G \in \mathcal{G} \cap B_{\mathcal{H}}(G_1, r/2)} P_G\ (1 - \varphi_{m,n}) \leq \exp[-mC_{k,n}(\mathcal{G}, r)]. \tag{4}$$

Here $D = D\left(\frac{c_0 C_{k,n}(\mathcal{G},r)}{4nC_0}, \mathcal{G} \cap B_{\mathcal{H}}(G_1, r/2), d_{\mathcal{H}}\right)$ denotes a packing number, i.e., the maximal number of elements in $\mathcal{G} \cap B_{\mathcal{H}}(G_1, r/2)$ that are mutually separated by at least $c_0 C_{k,n}(\mathcal{G}, r)/(4nC_0)$ in Hausdorff distance.

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

Lemma 5. *Suppose that $\mathcal{G}$ satisfies condition C. Fix $G_0 \in \mathcal{G}^k$. Suppose that for some non-increasing function $D(\epsilon)$, some $\epsilon_{m,n} \geq 0$ and every $\epsilon > \epsilon_{m,n}$,*

$$\begin{aligned} &\sup_{G_1 \in \mathcal{G}} D(c_0 C_{k,n}(\mathcal{G}, \epsilon)/(4nC_0), \mathcal{G} \cap B_{\mathcal{H}}(G_1, \epsilon/2), d_{\mathcal{H}}) \\ &\times D(\epsilon/2, \mathcal{G} \cap B_{\mathcal{H}}(G_0, 2\epsilon) \setminus B_W(G_0, \epsilon), d_{\mathcal{H}}) \leq D(\epsilon). \end{aligned} \tag{5}$$

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

$$P_{G_0}\ \varphi_{m,n} \leq D(\epsilon) \sum_{t=t_0}^{\lceil 1/\epsilon \rceil} \exp[-mC_{k,n}(\mathcal{G}, t\epsilon)/8] \tag{6}$$

$$\sup_{G \in \mathcal{G}: d_{\mathcal{H}}(G_0, G) > t_0 \epsilon} P_G\ (1 - \varphi_{m,n}) \leq \exp[-mC_{k,n}(\mathcal{G}, t_0\epsilon)/8]. \tag{7}$$

Remarks. (i) We note the appearance of two packing numbers in the upper bound for the test power. The first quantity is the packing number of the thin Hausdorff layer, i.e., the set $\mathcal{G} \cap B_{\mathcal{H}}(G_0, 2\epsilon) \setminus B_{\mathcal{H}}(G_0, \epsilon)$. This is similar to a quantity that arises the analysis of Ghosal et al. [2000]. The second quantity is the packing number for the small ball, i.e., $\mathcal{G} \cap B_{\mathcal{H}}(G_1, \epsilon/2)$ in terms of smaller balls in Hausdorff metric. This extra term appears to be intrinsic to the analysis of the latent polytope $G \in \mathcal{G}$, as opposed to the data density p_G, and is attributed to the non-convexity of the Hausdorff balls when restricted to a subset $\mathcal{G}$ (for instance, when $\mathcal{G} = \mathcal{G}^k$). See the proof of Lemma 4 for more details. (ii) Note also the roles of the Hellinger information function: it appears as the exponent in the powers of the tests, but it also provides the radius for the smaller balls which define the second packing number. This feature was also observed in the posterior asymptotics using Wasserstein metric [Nguyen, 2012].

The two aforementioned lemmas form the core argument for establishing the following general theorem for posterior contraction of latent mixing measures in the admixture model for discrete data. Define the Kullback-Leibler neighborhood of G_0 under the admixture model and the prior distribution Π on population polytope G as:

$$B_K(G_0, \delta) = \{G \in \mathcal{G}^* : K(p_{G_0}, p_G) \leq \delta; K_2(p_{G_0}, p_G) \leq \delta\}. \tag{8}$$

Theorem 4. *Let $G_0 \in \mathcal{G}^k$ for some $k < \infty$. Assume the following:*

(a) Π is a prior distribution on $\mathcal{G}^*$ such that the support of the prior is a subset $\mathcal{G} \subset \mathcal{G}^*$ which satisfies condition C.

(b) There is a sequence of subsets $\mathcal{G}_m \subset \mathcal{G}^*$.

(c) There is a sequence $\epsilon_{m,n} \to 0$ such that $m\epsilon_{m,n}^2$ is bounded away from 0 or tending to infinity, and a sequence M_m such that

$$\log D(\epsilon/2, \mathcal{G}_m \cap B_{\mathcal{H}}(G_0, 2\epsilon) \setminus B_{\mathcal{H}}(G_0, \epsilon), d_{\mathcal{H}}) + \sup_{G_1 \in \mathcal{G}_m} D(c_0 C_{k,n}(\mathcal{G}_m, \epsilon)/(4nC_0), \mathcal{G}_m \cap B_{\mathcal{H}}(G_1, \epsilon/2), d_{\mathcal{H}}) \leq m\epsilon_{m,n}^2$$

$$\text{for all } \epsilon \geq \epsilon_{m,n}, \tag{9}$$

$$\frac{\Pi(\mathcal{G}^* \setminus \mathcal{G}_m)}{\Pi(B_K(G_0, \epsilon_{m,n}))} = o(\exp(-2m\epsilon_{m,n}^2)), \tag{10}$$

$$\frac{\Pi(\mathcal{G}_m \cap B_{\mathcal{H}}(G_0, 2j\epsilon_{m,n}) \setminus B_{\mathcal{H}}(G_0, j\epsilon_{m,n}))}{\Pi(B_K(\epsilon_{m,n}))} \leq \exp[mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/16]$$

$$\text{for all } j \geq M_m, \tag{11}$$

$$\exp(2m\epsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/16] \to 0. \tag{12}$$

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

The proof of this theorem follows the method of Ghosal, Ghosh and van der Vaart [Ghosal et al., 2000], and is deferred to the Appendix. 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 $C_{k,n}(\mathcal{G}_m, \cdot)$ (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 Contraction properties

The following contraction result guarantees that as the data densities get closer, so do the population polytopes. This gives a lower bound for the Hellinger information defined by Eq. (2), since the Hellinger distance h can be lower bounded by the variational distance 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 $B_{d+1}(\boldsymbol{\theta}_c, r) \cap G$, while G' is p'-dimensional body containing $B_{d+1}(\boldsymbol{\theta}_c, r) \cap G'$ 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 a constant $c_1 > 0$ independent of G, G' such that*

$$V(p_G, p_{G'}) \geq c_1 d_{\mathcal{H}}(G, G')^{(p \vee p')+\alpha} - 6(d+1)\exp\bigg[- \frac{n}{8(d+1)} d_{\mathcal{H}}(G, G')^2 \bigg].$$

(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 such that

$$V(p_G, p_{G'}) \geq c_2 d_{\mathcal{H}}(G, G')^{1+\alpha} - 6(d+1)\exp\bigg[- \frac{n}{C_3(d+1)} d_{\mathcal{H}}(G, G')^2 \bigg].$$

Proof. (a) 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,

$$V(p_G, p_{G'}) = \sup_A |P_G(\hat{\boldsymbol{\eta}} \in A) - P_{G'}(\hat{\boldsymbol{\eta}} \in A)|, \tag{13}$$

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

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

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

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

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

The same bound holds under $P_{\boldsymbol{\eta}\times\mathcal{S}|G'}$. Now, define event $B = \{\|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\| < \epsilon\}$. Take any (measurable) set $A \subset \Delta^d$,

$$\begin{aligned} & |P_G(\hat{\boldsymbol{\eta}} \in A) - P_{G'}(\hat{\boldsymbol{\eta}} \in A)| \\ = \; & |P_{\boldsymbol{\eta}\times\mathcal{S}|G}(\hat{\boldsymbol{\eta}} \in A; B) + P_{\boldsymbol{\eta}\times\mathcal{S}|G}(\hat{\boldsymbol{\eta}} \in A; B^C) \\ & - P_{\boldsymbol{\eta}\times\mathcal{S}|G'}(\hat{\boldsymbol{\eta}} \in A; B) - P_{\boldsymbol{\eta}\times\mathcal{S}|G'}(\hat{\boldsymbol{\eta}} \in A; B^C)| \\ \geq \; & |P_{\boldsymbol{\eta}\times\mathcal{S}|G}(\hat{\boldsymbol{\eta}} \in A; B) - P_{\boldsymbol{\eta}\times\mathcal{S}|G'}(\hat{\boldsymbol{\eta}} \in A; B)| - 4(d+1)\exp[-2n\epsilon^2/(d+1)] \end{aligned} \tag{14}$$

Let $\epsilon_1 = d_{\mathcal{H}}(G, G')/4$. For any $\epsilon \leq \epsilon_1$, recall the outer ϵ-parallel set $G_\epsilon = (G + B_{d+1}(\mathbf{0}, \epsilon))$, which is full-dimensional $(d+1)$ eventhough G may not be. By triangular

inequality, $d_{\mathcal{H}}(G_\epsilon, G'_\epsilon) \geq d_{\mathcal{H}}(G, G')/2$. We shall argue that for any $\epsilon \leq \epsilon_1$, 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 $A^*_\epsilon \cap G'_\epsilon = \emptyset$ and $\mathrm{vol}_p(A^*) \geq c_1 \epsilon_1^p$, or

(ii) There is a set $A^* \subset G' \setminus G$ such that $A^*_\epsilon \cap G_\epsilon = \emptyset$ and $\mathrm{vol}_{p'}(A^*) \geq c_1 \epsilon_1^{p'}$.

Indeed, observe that either one of the following two inequalities holds: $d_{\mathcal{H}}(G \backslash G'_{3\epsilon}, G') \geq d_{\mathcal{H}}(G, G')/4$ or $d_{\mathcal{H}}(G' \setminus G_{3\epsilon}, G) \geq d_{\mathcal{H}}(G, G')/4$. If the former inequality holds, let $A^* = G \setminus G'_{3\epsilon}$. Then, $A^* \subset G \setminus G'$ and $A^*_\epsilon \cap G'_\epsilon = \emptyset$. Moreover, by Lemma 2 (a), $\mathrm{vol}_p(A^*) \geq c_1 \epsilon_1^p$, for some constant $c_1 > 0$ independent of $\epsilon, \epsilon_1, G, G'$, so A^* satisfies (i). Combined with the α-regularity of $P_{\boldsymbol{\eta}|G}$, we have $P_{\boldsymbol{\eta}|G}(A^*) \geq c_1 \epsilon^{p+\alpha}$ for some constant $c_1 > 0$. If the latter inequality holds, the same argument applies by defining $A^* = G' \setminus G_{3\epsilon}$ so that (ii) holds.

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

$$\begin{aligned} P_{\boldsymbol{\eta}\times\mathcal{S}|G}(\hat{\boldsymbol{\eta}} \in A^*_\epsilon; B) &\geq P_{\boldsymbol{\eta}\times\mathcal{S}|G}(\boldsymbol{\eta} \in A^*; B) \\ &\geq P_{\boldsymbol{\eta}|G}(A^*) - P_{\boldsymbol{\eta}\times\mathcal{S}|G}(B^C) \\ &\geq P_{\boldsymbol{\eta}|G}(A^*) - 2(d+1)\exp(-2n\epsilon^2/(d+1)) \\ &\geq c_1 \epsilon_1^{p+\alpha} - 2(d+1)\exp(-2n\epsilon^2/(d+1)), \end{aligned}$$

for some constant $c_1 > 0$. Hence, by Eq. (14) $|P_G(\hat{\boldsymbol{\eta}} \in A^*_\epsilon) - P_{G'}(\hat{\boldsymbol{\eta}} \in A^*_\epsilon)| \geq c_1 \epsilon_1^{p+\alpha} - 6(d+1)\exp(-2n\epsilon^2)$. Set $\epsilon = \epsilon_1$, the conclusion then follows by invoking Eq. (13). 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 involking Lemma 3. Under the condition that $|\operatorname{extr} G'| = k$, the proof is also similar, but it requires a suitable modification for the existence of set A^*. For any small ϵ, let $\tilde{G}_\epsilon$ be the minimum-volume homethetic transformation of G, with respect to center $\boldsymbol{\theta}_c$, such that $\tilde{G}_\epsilon$ 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 $d_{\mathcal{H}}(G, \tilde{G}_\epsilon) \leq \epsilon R/r = \epsilon/r$.

Set $\epsilon_1 = d_{\mathcal{H}}(G, G')r/4$. We shall argue that for any $\epsilon \leq \epsilon_1$, 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 $A^*_\epsilon \cap G'_\epsilon = \emptyset$ and $\mathrm{vol}_p(A^*) \geq c_2 \epsilon_1$, or

(iv) There is a set $A^* \subset G' \setminus G$ such that $A^*_\epsilon \cap G_\epsilon = \emptyset$ and $\mathrm{vol}_p(A^*) \geq c_2 \epsilon_1$.

Indeed, note that either one of the following two inequalities holds: $d_{\mathcal{H}}(G \setminus \tilde{G}'_{3\epsilon}, G') \geq d_{\mathcal{H}}(G, G')/4$ or $d_{\mathcal{H}}(G' \setminus \tilde{G}_{3\epsilon}, G) \geq d_{\mathcal{H}}(G, G')/4$. If the former inequality holds, let $A^* = G \backslash \tilde{G}'_{3\epsilon}$. Then, $A^* \subset G \backslash G'$ and $A^*_\epsilon \cap \tilde{G}'_\epsilon = \emptyset$. Observe that both G and $\tilde{G}'_{3\epsilon}$ have the same number of extreme points by the construction. Moreover, G is fixed so that all geometric

properties A2, A1are satisfied for both G and $\tilde{G'}_{3\epsilon}$ for sufficiently small $d_{\mathcal{H}}(G, G')$. By Lemma 3, $\mathrm{vol}_p(A^*) \geq c_2\epsilon_1$. Hence, (iii) holds. If the latter inequality holds, the same argument applies by defining $A^* = G' \setminus \tilde{G}_{3\epsilon}$ 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 Kullback-Leibler balls.

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 on $\mathbb{R}^k$). Define random variable $\boldsymbol{\eta} = \beta_1\boldsymbol{\theta}_1 + \ldots + \beta_k\boldsymbol{\theta}_k$, which takes values in $G = \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. Write $\boldsymbol{\eta} = L\boldsymbol{\beta}$, where $L = [\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_k]$ 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 [Evans and Gariepy, 1992] (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 on Δ^d:

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

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 d-dimensional Hausdorff measure on $\mathbb{R}^{d+1}$:

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

A common choice for $P_{\boldsymbol{\beta}}$ is the Dirichlet distribution, as adopted by Pritchard et al. [2000], Blei et al. [2003]: given parameter $\gamma \in \mathbb{R}^k_+$, for any $A \subset \Delta^{k-1}$,

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

Lemma 6. *Let $\boldsymbol{\eta} = \sum_{j=1}^k \beta_j\boldsymbol{\theta}_k$, 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 $\epsilon_0 = \epsilon_0(k) > 0$, and constant $c_6 = c_6(\gamma, k, d) > 0$ dependent on γ, k and d such that for any $\epsilon < \epsilon_0$,

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

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

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 $\mathcal{G}^k$. Under the admixture model specification, this induces a (prior) distribution on the space of marginal densities p_G of the data vector $\mathcal{S}_{[n]}$. To establish the concentration properties of the Kullback-Leibler neighborhood B_K as induced by the prior distribution, 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 7. *Let $G, G' \in \Delta^d$ 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 $K(p_G, p_{G'}) \leq C_1 d_{\mathcal{H}}(G, G')$ whenever $d_{\mathcal{H}}(G, G')$ is sufficiently small.

Proof. First, we note a well-known fact of KL divergences: the divergence between marginal distributions are bounded from above by the divergence between joint distributions:

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

Due to conditional independence, $p_{\boldsymbol{\eta} \times \mathcal{S}|G} = p_{\boldsymbol{\eta}|G} \times p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}$ and $p_{\boldsymbol{\eta} \times \mathcal{S}|G'} = p_{\boldsymbol{\eta}|G'} \times p_{\mathcal{S}_{[n]}|\boldsymbol{\eta}}$, so $K(P_{\boldsymbol{\eta} \times \mathcal{S}|G}, P_{\boldsymbol{\eta} \times \mathcal{S}|G'}) = K(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'})$. Since $P_{\boldsymbol{\eta}|G}$ and $P_{\boldsymbol{\eta}|G'}$ are assumed to be uniform distributions on G and G', respectively, and $G \subset G'$, we obtain that

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

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 particular stringent condition, $\operatorname{aff} G = \operatorname{aff} G'$, 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. Moreover, sample size n now appears as a linear term in the upper bound, as one would expect.

Let $Q(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ denote a coupling of $P(\boldsymbol{\eta}|G)$ and $P(\boldsymbol{\eta}|G')$, i.e., 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

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

Lemma 8. *Let $G, G' \subset \Delta^d$ be closed convex subsets such that any $\boldsymbol{\eta} = (\eta_0, \ldots, \eta_d) \in G \cup G'$ satisfies $\min_{l=0,\ldots,d} \eta_l > c_0$ for some constant $c_0 > 0$. Then*

$$K(p_G, p_{G'}) \leq \frac{n}{c_0} W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}).$$

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

$$p_G(\mathcal{S}_{[n]}) = \int_{\boldsymbol{\eta} \in G} p(\mathcal{S}_{[n]}|\boldsymbol{\eta}) dP(\boldsymbol{\eta}|G) = \int_{\boldsymbol{\eta} \in G} \exp\left(n \sum_{i=0}^d \boldsymbol{\eta}(\mathcal{S})_i \log \boldsymbol{\eta}_i \right) dP(\boldsymbol{\eta}|G).$$

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

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

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

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 $\mathcal{S}_{[n]} \in \{0, \ldots, d\}^n$. For any $\boldsymbol{\eta}(\mathcal{S}) \in \Delta^d$, $\boldsymbol{\eta}_1 \in G$ and $\boldsymbol{\eta}_2 \in G'$,

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

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 $K(p_G, p_{G'}) \leq \frac{n}{c_0} W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'})$. □

Lemma 9. *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 sufficiently small $d_{\mathcal{H}}(G, G')$, $W_1(p_{\boldsymbol{\eta}|G}, p_{\boldsymbol{\eta}|G'}) \leq C_0 d_{\mathcal{H}}(G, G')$ for some constant C_0 specified by Lemma 1.

(b) Assume further that assumptions in Lemma 8 hold, then $K(p_G, p_{G'}) \leq \frac{n}{c_0} C_0 d_{\mathcal{H}}(G, G')$.

Remark. In order to obtain an upper bound for $K(p_G, p_{G'})$ in terms of $d_{\mathcal{H}}(G, G')$, the assumption that $p_{\boldsymbol{\beta}}$ is symmetric appears essential. Without this assumption, it is possible to have $d_{\mathcal{H}}(G, G') = 0$, but $K(p_G, p_{G'}) > 0$.

Proof. By Lemma 1 under property A2, $d_{\mathcal{M}}(G, G') \leq C_0 d_{\mathcal{H}}(G, G')$ for some constant C_0. Let $d_{\mathcal{H}}(G, G') \leq \epsilon$ for some small $\epsilon > 0$. Assume without loss of generality that $|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j| \leq C_0 \epsilon$ 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 $\boldsymbol{\eta}' = \sum_{j=1}^k \beta_j \boldsymbol{\theta}'_j$, i.e., $\boldsymbol{\eta}$ 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, $\mathbb{E}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \leq \mathbb{E} \sum_{j=1}^k \beta_j \|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\| \leq C_0 \epsilon \mathbb{E} \sum_{j=1}^k \beta_j = C_0 \epsilon$. Hence $W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G'}) \leq C_0 \epsilon$. Part (b) is an immediate consequence. □

Recall the definition of the Kullback-Leibler neighborhood given by Eq. (8). The main result of this section is the following:

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

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

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

Proof. We shall invoke a bound of Wong and Shen [1995] (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 $\epsilon_0 > 0$, as long as $h(p, q) \leq \epsilon < \epsilon_0$, there holds: $K(p, q) = O(\epsilon^2 \log(M/\epsilon))$, and $K_2(p, q) := \int p(\log(p/q))^2 = O(\epsilon^2 [\log(M/\epsilon)]^2)$, 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 $\|\boldsymbol{\theta}_j - \boldsymbol{\theta}^*_j\| \leq \epsilon$ 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 $\boldsymbol{\eta}_1 = \beta_1 \boldsymbol{\theta}_1 + \ldots \beta_k \boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_2 = \beta_1 \boldsymbol{\theta}^*_1 + \ldots \beta_k \boldsymbol{\theta}^*_k$ (that is, under

the coupling $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ share the same vector $\boldsymbol{\beta}$). Then, under Q, $\mathbb{E}\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\| \leq \epsilon$. This entails that $W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G_0}) \leq \epsilon$. (We note here that the argument appears similar to the one from Lemma 9, 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 $\Omega(\epsilon^{kd})$. By Lemma 8, $h^2(p_{G_0}, p_G) \leq K(p_{G_0}, p_G)/2 \leq (n/c_0)W_1(P_{\boldsymbol{\eta}|G}, P_{\boldsymbol{\eta}|G_0}) \leq n\epsilon/(2c_0)$. Note that the density ratio $p_G/p_{G_0} \leq (1/c_0)^n$, which implies that $\sum_{\mathcal{S}_{[n]}} p_{G_0}^2/p_G \leq (1/c_0)^n$. We can apply the upper bound described in the previous paragraph to obtain:

$$K_2(p_{G_0}, p_G) = O\bigg(\frac{n\epsilon}{2c_0}\bigg[\frac{1}{2}\log\frac{2c_0}{n\epsilon} + n\log\frac{1}{c_0}\bigg]^2\bigg).$$

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

7 Proofs of main theorems and auxiliary lemmas

Proof of Theorem 1 (Overfitted setting).

Proof. The proof proceeds by verifying conditions of Theorem 4. Let $\epsilon_{m,n}$ be a large multiple of $(\log m/m)^{1/2} \vee (\log n/m)^{1/2} \vee (\log n/n)^{1/2}$.

Choose the sequence of subsets $\mathcal{G}_m$ simply as $\mathcal{G}_m = \mathcal{G}^k$, so that $\Pi(\mathcal{G}^* \setminus \mathcal{G}_m) = 0$. Condition (10) trivially holds. Turning to the entropy conditions, we note that

$$\log D(\epsilon/2, \mathcal{G}^k \cap B_{\mathcal{H}}(G_0, 2\epsilon), d_{\mathcal{H}}) \quad \leq \quad \log N(\epsilon/4, \mathcal{G}^k \cap B_{\mathcal{H}}(G_0, 2\epsilon), d_{\mathcal{H}}) = O(1).$$

By Theorem 5 (a), assumption (S4) and the general inequality that $h \geq V$, we have: $C_{k,n}(\mathcal{G}^k, \epsilon) \gtrsim [\epsilon^{p+\alpha} - 6(d+1)e^{-n\epsilon^2/8(d+1)}]^2$, where p is defined as $p = \min(k-1, d)$. So $C_{k,n}(\mathcal{G}^k, \epsilon) \gtrsim \epsilon^{2(p+\alpha)}$ as long as $\epsilon^{p+\alpha} \geq 12(d+1)\exp[-n\epsilon^2/8(d+1)]$. This is satisfied for any $\epsilon \geq \epsilon_{m,n} \gtrsim (\frac{p+\alpha}{2}\log n/n)^{1/2}$. It follows that

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

where the last inequality holds since $\epsilon \geq \epsilon_{m,n} \gtrsim (\log n/m)^{1/2} \vee (\log m/m)^{1/2}$. Thus, the entropy condition (9)) is established.

To verify condition Eq. (12), we note that for some constant $c > 0$,

$$\begin{aligned}
&\exp(2m\epsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/16] \\
&\lesssim \exp(2m\epsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-cm(j\epsilon_{m,n})^{2(p+\alpha)}] \\
&\lesssim \exp(2m\epsilon_{m,n}^2)\exp[-cm(M_m\epsilon_{m,n})^{2(p+\alpha)}],
\end{aligned}$$

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

It remains to verify Eq. (11). By Theorem 6, $\log \Pi(G \in B_K(G_0, \epsilon_{m,n})) \gtrsim \log(\epsilon_{m,n}/n^3)^{kd} = kd(\log \epsilon_{m,n} - 3\log n)$.

Moreover, $\Pi(B_{\mathcal{H}}(G_0, 2j\epsilon_{m,n}) \setminus B_{\mathcal{H}}(G_0, j\epsilon_{m,n})) \leq \Pi(B_{\mathcal{H}}(G_0, 2j\epsilon_{m,n}))$. Take any $j \geq M_m$, if $d_{\mathcal{H}}(G, G_0) \leq j\epsilon_{m,n}$, then at least one of G's extreme points is within $O(j\epsilon_{m,n})$ distance from G_0's extreme points, by Lemma 1 (b). By an union bound, and the assumption that the prior densities for $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ are bounded away from 0, $\Pi(B_{\mathcal{H}}(G_0, 2j\epsilon_{m,n})) \lesssim k^2(2j\epsilon_{m,n})^d$. As the result, the logarithm of the left side of Eq. (11) is upper bounded by

$$\begin{aligned}\log[k^2(2j\epsilon_{m,n})^d(n^3/\epsilon_{m,n})^{kd}] &\leq \log(k^2 2^d) + d\log j + kd\log(1/\epsilon_{m,n}) + 3kd\log n \\ &\lesssim m(j\epsilon_{m,n})^{2(p+\alpha)} \lesssim mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/16\end{aligned}$$

The last inequality of the previous display is due to Theorem 5 (a). The next to the last inequality holds because for any $j \geq M_m$, $m(j\epsilon_{m,n})^{2(p+\alpha)} \gtrsim m\epsilon_{m,n}^2 \gtrsim \log n \vee \log(1/\epsilon_{m,n})$, and that $m(j\epsilon_{m,n})^{2(p+\alpha)} \gtrsim \log j$.

Now, we can apply Theorem 4 to obtain a posterior contraction rate $M_m\epsilon_{m,n} \asymp \epsilon_{m,n}^{1/(p+\alpha)} \asymp [\frac{\log m}{m} \vee \frac{\log n}{m} \vee \frac{\log n}{n}]^{\frac{1}{2(p+\alpha)}}$. □

Proof of Theorem 2. The proof proceeds in exactly the same way as in 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 $d_{\mathcal{H}}(G_0, G_0') \asymp \epsilon$, and $\mathrm{vol}_q(G_0 \setminus G_0') \asymp \epsilon^q$. Due to Assumption (S5), $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \epsilon^{q+\alpha}$.

If $k \geq 2d$, 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 $d_{\mathcal{H}}(G_0, G_0') \asymp \epsilon$, and $\mathrm{vol}_d(G_0 \setminus G_0') \asymp \epsilon^d$. Due to Assumption (S5), $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \epsilon^{d+\alpha}$.

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 $d_{\mathcal{H}}(G_0, G_0') \asymp \epsilon$, and $V(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim \epsilon^{q+\alpha}$. By Lemma 8, $K(p_{G_0}, p_{G_0'}) \lesssim nW_1(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \lesssim nV(p_{\boldsymbol{\eta}|G_0}, p_{\boldsymbol{\eta}|G_0'}) \leq Cn\epsilon^{q+\alpha}$ for some constant $C > 0$.

Applying the method due to Le Cam (cf. [Yu, 1997], Lemma 1), for any estimator $\hat{G}$,

$$\max_{G\in\{G_0,G_0'\}} P_{G_0} d_{\mathcal{H}}(G,\hat{G}) \gtrsim \epsilon(1-\frac{1}{2}V(p_{G_0}^{[m]},p_{G_0'}^{[m]})).$$

Here, $p_{G_0}^{[m]}$ denotes the marginal density of the m-sample $\mathcal{S}_{[n]}^1,\ldots,\mathcal{S}_{[n]}^m$. Thus, $V^2(p_{G_0}^{[m]},p_{G_0'}^{[m]}) \leq h^2(p_{G_0}^{[m]},p_{G_0'}^{[m]}) = 1-\int[p_{G_0}^{[m]}p_{G_0'}^{[m]}]^{1/2} = 1-(1-h^2(p_{G_0},p_{G_0'}))^m \leq 1-(1-Cn\epsilon^{q+\alpha})^m$. The last inequality is due to $h^2(p_{G_0},p_{G_0'}) \leq K(p_{G_0},p_{G_0'}) \leq Cn\epsilon^{q+\alpha}$. Thus,

$$\max_{G\in\{G_0,G_0'\}} P_{G_0} d_{\mathcal{H}}(G,\hat{G}) \gtrsim \epsilon(1-\frac{1}{2}[1-(1-Cn\epsilon^{q+\alpha})^m]^{1/2}).$$

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

(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 7 $K(p_{G_0'},p_{G_0}) = \int \log[\mathrm{vol}_q\, G_0/\,\mathrm{vol}_q\, G_0']dP_{\boldsymbol{\eta}|G_0} \leq \int \log(1+C\epsilon^q)P_{\boldsymbol{\eta}|G_0} \lesssim \epsilon^q$. So, $h^2(p_{G_0},p_{G_0'}) \leq K(p_{G_0'},p_{G_0}) \lesssim \epsilon^q$. 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 $d_{\mathcal{H}}(G_0',G_0)=\epsilon$. By Lemma 2, $\mathrm{vol}_p(G_0 \triangle G_0') = O(\epsilon)$, 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 (S5). Under assumption (S5'), as in part (b), the dependence on n can be removed to obtain $1/m$ rate.

Proof of the existence of tests in Lemma 4.

Proof. Define $\mathcal{P}_1 = \{p_G | G \in \mathcal{G}\cap B_{\mathcal{H}}(G_1,r/2)\}$. We note in passing that that this is generally not a convex set of densities for $\mathcal{S}_{[n]}$. For instance, if $\mathcal{G}=\mathcal{G}_k$, which is a non-convex set, then $\mathcal{P}_1$ is non-convex. Thus, a straightforward application of standard results on existence of tests (cf. [Cam, 1986], Chapter 4) is not possible. Consider a maximal $c_1 r$-packing in $d_{\mathcal{H}}$ metric for the set $G\in\mathcal{G}\cap B_{\mathcal{H}}(G_1,r/2)$, where c_1 is a positive constant to be determined. This yields a set of $D=D(c_1 r,\mathcal{G}\cap B_{\mathcal{H}}(G_1,r/2),d_{\mathcal{H}})$ elements $\tilde{G}_1,\ldots,\tilde{G}_D \in \mathcal{G}\cap B_{\mathcal{H}}(G_1,r/2)$.

Next, we note the following fact: for any $t=1,\ldots,D$, if $G\in\mathcal{G}\cap B_{\mathcal{H}}(G_1,r/2)$ and $d_{\mathcal{H}}(G,\tilde{G}_t)\leq c_1 r$, then by Lemma 9

$$h^2(p_G,p_{\tilde{G}_t}) \leq K(p_G,p_{\tilde{G}_t}) \leq \frac{n}{c_0}C_0 d_{\mathcal{H}}(G,\tilde{G}_t) \leq \frac{n}{c_0}C_0 c_1 r$$

Choose $c_1 = \frac{c_0}{4nrC_0}C_{k,n}(\mathcal{G},r)$, so that $h^2(p_G,p_{\tilde{G}_t}) \leq \frac{1}{4}C_{k,n}(\mathcal{G},r)$. By definition, $h^2(p_{G_0},p_{\tilde{G}_t}) \geq C_{k,n}(\mathcal{G},r)$. Thus, by triangle inequality, $h(p_{G_0},p_G) \geq \frac{1}{2}C_{k,n}(\mathcal{G},r)^{1/2}$.

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

$$P_{G_0}\ \omega_{m,n}^{(t)} \leq \exp[-mh^2(p_{G_0},p_{\tilde{G}_t})/8],$$
$$\sup_{P_2\in\mathcal{P}_2(t)} P_2\ (1-\omega_{m,n}^{(t)}) \leq \exp[-mh^2(p_{G_0},p_{\tilde{G}_t})/8].$$

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

$$\begin{aligned} P_{G_0}\ \varphi_{m,n} &\le D \times \exp[-mC_{k,n}(\mathcal{G}, r)/8], \\ \sup_{G\in\mathcal{G}\cap B_{\mathcal{H}}(G_1, r/2)} P_G\ (1-\varphi_{m,n}) &\le \exp[-mC_{k,n}(\mathcal{G}, r)/8]. \end{aligned}$$

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 $G \in \mathcal{G} \cap B_{\mathcal{H}}(G_1, r/2)$ there is some $d = 1, \ldots, D$ such that $d_{\mathcal{H}}(G, \tilde{G}_t) \le c_1 r$, so that $p_G \in \mathcal{P}_2(t)$. □

Proof of the existence of tests in Lemma 5.

Proof. The proof utilizes a peeling idea of Ghosal et al. [2000], and then apply a packing argument as in the previous proof. For a given $t \in \mathbb{N}$ choose a maximal $t\epsilon/2$-packing for set $S_t = \{G : t\epsilon < d_{\mathcal{H}}(G_0, G) \le (t+1)\epsilon\}$. This yields a set S'_t of at most $D(t\epsilon/2, S_t, d_{\mathcal{H}})$ points. Moreover, every $G \in S_t$ is within distance $t\epsilon/2$ 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 Eqs. (3) and (4), where r is taken to be $r = t\epsilon$. 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 $t \le \lceil 1/\epsilon \rceil$. Then, by union bound, and the condition that $D(\epsilon)$ is non-increasing,

$$\begin{aligned} &P_{G_0}\ \varphi_{m,n} \\ &\quad\le \sum_{t=t_0}^{\lceil 1/\epsilon\rceil} \sum_{G_1\in S'_t} D\Big(\frac{c_0 C_{k,n}(\mathcal{G}, t\epsilon)}{4nC_0}, \mathcal{G}\cap B_{\mathcal{H}}(G_1, t\epsilon/2), d_{\mathcal{H}}\Big) \exp[-mC_{k,n}(\mathcal{G}, t\epsilon)/8] \\ &\quad\le D(\epsilon) \sum_{t\ge t_0} \exp[-mC_{k,n}(\mathcal{G}, t\epsilon)/8] \end{aligned}$$

$$\sup_{G\in\cup_{u\ge t_0} S_u} P_G\ (1-\varphi_n) \le \sup_{u\ge t_0}\ \exp[-mC_{k,n}(\mathcal{G}, u\epsilon)/8] \le \exp[-mC_{k,n}(\mathcal{G}, t_0\epsilon)/8],$$

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

Proof of α-regularity of the Dirichlet-induced densities in Lemma 6.

Proof. First, consider the case $k \le d+1$. For $\boldsymbol{\eta}^* \in G$, write $\boldsymbol{\eta}^* = \beta^*_1\boldsymbol{\theta}_1 + \ldots + \beta^*_k\boldsymbol{\theta}_k$. For $\boldsymbol{\beta} \in \Delta^{k-1}$ such that $|\beta_i - \beta^*_i| \le \epsilon/k$ for all $i = 1, \ldots, k-1$, we have $\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| = \|\sum_{i=1}^{k}(\beta_i - \beta^*_{,i})\boldsymbol{\theta}_i\| \le \sum_{i=1}^{k}|\beta_i - \beta^*_i| \le 2\sum_{i=1}^{k-1}|\beta_i - \beta^*_i| \le 2\epsilon$. 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 $\epsilon < 1/k$

$$P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\| \le 2\epsilon) \ge P_{\boldsymbol{\beta}}(|\beta_i - \beta_i^*| \le \epsilon/k; i = 1, \ldots, k-1)$$
$$= \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \int_{\beta_i \in [0,1]; |\beta_i - \beta_i^*| \le \epsilon/k; i=1,\ldots,k-1} \prod_{i=1}^{k-1} \beta_i^{\gamma_i - 1} (1 - \sum_{i=1}^{k-1} \beta_i)^{\gamma_k - 1} d\beta_1 \ldots d\beta_{k-1}$$
$$\ge \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \prod_{i=1}^{k-1} \int_{\max(\gamma_i^* - \epsilon/k, 0)}^{\min(\gamma_i^* + \epsilon/k, 1)} \beta_i^{\gamma_i - 1} d\beta_i \ge \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} (\epsilon/k)^{k-1}.$$

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

Now, consider the case $k > d+1$. Since $\boldsymbol{\eta}^* \in \mathrm{conv}(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k) \subset \Delta^d$, 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 $\boldsymbol{\eta}^* = \beta_1^* \boldsymbol{\theta}_1 + \ldots \beta_{d+1}^* \boldsymbol{\theta}_{d+1}$. Consider $\boldsymbol{\eta} = \beta_1 \boldsymbol{\theta}_1 + \ldots + \beta_k \boldsymbol{\theta}_k$, where $\|\beta_i - \beta_i^*| \le \epsilon/k$, for $i = 1, \ldots, d$, while $0 \le \beta_i \le \epsilon/k$ for $i = d+2, \ldots, k$. Then, $\|\boldsymbol{\eta} - \boldsymbol{\eta}^*\| \le 2\epsilon$. This implies that

$$P_{\boldsymbol{\eta}|G}(\|\boldsymbol{\eta}-\boldsymbol{\eta}^*\| \le 2\epsilon) \ge P_{\boldsymbol{\beta}}(|\beta_i - \beta_i^*| \le \epsilon/k, i = 1, \ldots, d+1;\ |\beta_j| \le \epsilon/k, j > d+1)$$
$$\ge \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} \prod_{i=1}^{d} \int_{\max(\gamma_i^* - \epsilon/k, 0)}^{\min(\gamma_i^* + \epsilon/k, 1)} \beta_i^{\gamma_i - 1} d\beta_i \prod_{i=d+2}^{k} \int_0^{\epsilon/k} \beta_i^{\gamma_i - 1} d\beta_i$$
$$\ge \frac{\Gamma(\sum \gamma_i)}{\prod_i \Gamma(\gamma_i)} (\epsilon/k)^{d + \sum_{i=d+2}^k \gamma_i} / \prod_{i=d+2}^{n} \gamma_i \gtrsim \epsilon^{d + \sum_{i=1}^k \gamma_i}.$$

□

8 Appendix

Proof of the general posterior contraction in Theorem 4.

Proof. By a result of Ghosal et al [Ghosal et al., 2000] (Lemma 8.1, pg. 524), for every $\epsilon > 0, C > 0$ and every probability measure Π_0 supported on the set $B_K(G_0, \epsilon)$ defined by Eq. (8), we have,

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

This entails that, for a fixed $C \ge 1$, there is an event A_m with P_{G_0}-probability at least $1 - (Cm\epsilon_{m,n}^2)^{-1}$, for which there holds:

$$\int \prod_{i=1}^{n} p_G(\mathcal{S}^i_{[n]}) / p_{G_0}(\mathcal{S}^i_{[n]}) d\Pi(G) \ge \exp(-2m\epsilon_{m,n}^2) \Pi(B_K(G_0, \epsilon_{m,n})). \qquad (15)$$

Let $\mathcal{O}_m = \{G \in \mathcal{G}^* : d_{\mathcal{H}}(G_0, G) \geq M_m \epsilon_{m,n}\}$, $S_{n,j} = \{G \in \mathcal{G}_m : d_{\mathcal{H}}(G_0, G) \in [j\epsilon_{m,n}, (j+1)\epsilon_{m,n})\}$ for each $j \geq 1$. Due to Eq.(9), the condition specified by Lemma 5 is satisfied by setting $D(\epsilon) = \exp(m\epsilon_{m,n}^2)$ (constant in ϵ). Thus there exist tests $\varphi_{m,n}$ for which Eq. (6) and (7) hold. Then,

$$
\begin{aligned}
& P_{G_0}\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m) \\
= \; & P_{G_0}[\varphi_{m,n}\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)] + P_{G_0}[(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)] \\
\leq \; & P_{G_0}[\varphi_{m,n}\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)] + P_{G_0}\mathbb{I}(A_m^c) \\
& + P_{G_0}\left[(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)\mathbb{I}(A_m)\right].
\end{aligned}
$$

Applying Lemma 5, the first term in the preceeding display is bounded above by $P_{G_0}\varphi_{m,n} \leq D(\epsilon_{m,n}) \sum_{j \geq M_m} \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/8] \to 0$, thanks to Eq. (12). The second term in the above display is bounded by $(Cm\epsilon_{m,n}^2)^{-1}$ by the definition of A_m. Since $m\epsilon_{m,n}^2$ is bounded away from 0, C can be chosen arbitrarily large so that the second term can be made arbitrarily small. It remains to show that third term in the display also vanishes as $m \to \infty$. We exploit the following expression:

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

and then obtain a lower bound for the denominator by Eq. (15). For the nominator, by Fubini's theorem:

$$
\begin{aligned}
& P_{G_0}\int_{\mathcal{O}_m \cap \mathcal{G}_m} (1-\varphi_{m,n}) \prod_{i=1}^m p_G(\mathcal{S}_{[n]}^i)/p_{G_0}(\mathcal{S}_{[n]}^i)\Pi(G) \\
= \; & P_{G_0}\sum_{j \geq M_m} \int_{S_{m,j}} (1-\varphi_{m,n}) \prod_{i=1}^m p_G(\mathcal{S}_{[n]}^i)/p_{G_0}(\mathcal{S}_{[n]}^i)\Pi(G) \\
= \; & \sum_{j \geq M_m} \int_{S_{m,j}} P_G(1-\varphi_{m,n})\Pi(G) \\
\leq \; & \sum_{j \geq M_m} \Pi(S_{m,j}) \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/8], \qquad (16)
\end{aligned}
$$

where the last inequality is due to Eq. (7). In addition, by (10),

$$
\begin{aligned}
& P_{G_0}\int_{\mathcal{O}_m \setminus \mathcal{G}_m} (1-\varphi_{m,n}) \prod_{i=1}^m p_G(\mathcal{S}_{[n]}^i)/p_{G_0}(\mathcal{S}_{[n]}^i)\Pi(G) \\
= \; & \int_{\mathcal{O}_m \setminus \mathcal{G}_m} P_G(1-\varphi_{m,n})\Pi(G) \\
\leq \; & \Pi(\mathcal{G}^* \setminus \mathcal{G}_m) = o(\exp(-2m\epsilon_{m,n}^2)\Pi(B_K(G_0, \epsilon_{m,n}))). \qquad (17)
\end{aligned}
$$

Now, combining bounds (16) and (17) with condition (11), we obtain:

$$\begin{aligned} & P_{G_0}(1-\varphi_{m,n})\Pi(G \in \mathcal{O}_m | \mathcal{S}_{[n]}^1, \ldots, \mathcal{S}_{[n]}^m)\mathbb{I}(A_m) \\ \leq & \frac{o(\exp(-2m\epsilon_{m,n}^2)\Pi(B_K(G_0, \epsilon_{m,n}))) + \sum_{j \geq M_m} \Pi(S_{m,j}) \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/8]}{\exp(-2m\epsilon_{m,n}^2)\Pi(B_K(\epsilon_{m,n}))} \\ \leq & o(1) + \exp(2m\epsilon_{m,n}^2) \sum_{j \geq M_m} \exp[-mC_{k,n}(\mathcal{G}_m, j\epsilon_{m,n})/16] \end{aligned}$$

The upper bound in the preceeding display converges to 0 by Eq. (12), thereby concluding the proof.

□

References

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

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. *arXiv:1204.1956*, 2012.

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

L. Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, 1986.

J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1): 221–233, 1995.

L. Dumbgen and G. Walther. Rates of convergence for random approximations of convex sets. *Advances in Applied Probability*, 28(2):384–393, 1996.

L. Evans and R. Gariepy. *Measure theory and fine properties of functions*. CRC Press, 1992.

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

H. Ishwaran, L. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of American Statistical Association*, 96(456):1316–1332, 2001.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, revision submitted, 2012.

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

R. Schneider. *Convex bodies: Brunn-Minkowsky theory*. Cambridge University Press, 1993.

A. Singh, C. Scott, and R. Nowak. Adaptive Hausdorff estimation of density level sets. *Annals of Statistics*, 37(5B):2760–2782, 2009.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Annals of Statistics*, 25: 948–969, 1997.

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

B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.