The Sketched Wasserstein Distance for mixture distributions

Xin Bing*   Florentina Bunea†   Jonathan Niles-Weed‡

Abstract

The Sketched Wasserstein Distance (W^S) is a new probability distance specifically tailored to finite mixture distributions. Given any metric d defined on a set $\mathcal{A}$ of probability distributions, W^S is defined to be the most discriminative convex extension of this metric to the space $\mathcal{S} = \text{conv}(\mathcal{A})$ of mixtures of elements of $\mathcal{A}$. Our representation theorem shows that the space $(\mathcal{S}, W^S)$ constructed in this way is isomorphic to a Wasserstein space over $\mathcal{X} = (\mathcal{A}, d)$. This result establishes a universality property for the Wasserstein distances, revealing them to be uniquely characterized by their discriminative power for finite mixtures. We exploit this representation theorem to propose an estimation methodology based on Kantorovich–Rubenstein duality, and prove a general theorem that shows that its estimation error can be bounded by the sum of the errors of estimating the mixture weights and the mixture components, for any estimators of these quantities. We derive sharp statistical properties for the estimated W^S in the case of p-dimensional discrete K-mixtures, which we show can be estimated at a rate proportional to $\sqrt{K/N}$, up to logarithmic factors. We complement these bounds with a minimax lower bound on the risk of estimating the Wasserstein distance between distributions on a K-point metric space, which matches our upper bound up to logarithmic factors. This result is the first nearly tight minimax lower bound for estimating the Wasserstein distance between discrete distributions. Furthermore, we construct $\sqrt{N}$ asymptotically normal estimators of the mixture weights, and derive a $\sqrt{N}$ distributional limit of our estimator of W^S as a consequence. An extensive simulation study and a data analysis provide strong support on the applicability of the new Sketched Wasserstein Distance between mixtures.

Keyword: Probability distance, mixture distribution, Wasserstein distance, optimal rates of convergence, limiting distribution, topic model.

1 Introduction

This paper proposes a new metric between probability distributions, tailored to finite mixture distributions on a Polish space $\mathcal{Y}$. This metric, which we call the *Sketched Wasserstein Metric* (W^S) for reasons we detail below, provides a generic and flexible framework for estimation and inference tasks involving finite mixtures. Our work is motivated by, and can be placed at, the intersection of the following two strands of work.

Distances between probability measures. There are many metrics between probability distributions: total variation, Hellinger distance, L^p metrics, the Kolmogorov and Lévy metrics (for distributions on $\mathbb{R}$), the Prokhorov metric, Wasserstein distances, energy distances, maximum mean discrepancy, to give just a short list of examples of classical distances. More recently, the success of "metric learning" approaches in machine learning [5, 15, 30, 62] opens the door to

---

*Department of Statistical Sciences, University of Toronto. E-mail: `xin.bing@utoronto.ca`

†Department of Statistics and Data Science, Cornell University. E-mail: `fb238@cornell.edu`.

‡Courant Institute for Mathematical Sciences and Center for Data Science, New York University. `jnw@cims.nyu.edu`

much more sophisticated metrics, including ones learned by deep neural networks, albeit via black-box type algorithms. These metrics all find use in various statistical contexts, for instance in two-sample and goodness-of-fit testing [46, 48], minimum distance estimation [58, 59], and robust statistics [42]. However, in high-dimensional contexts, these metrics all can have very different properties, in terms of their sensitivity to outliers, their ability to capture geometric features of the problem, and the structure necessary to define them [12].

In many specific applications, it is useful to tailor a specific distance, particularly one that reflects geometric features of the problem at hand; for instance, this is often cited as an important benefit of the Wasserstein distances [34]. We take this view in this work, and develop a new Wasserstein-type distance, tailored to structured distributions, in which the structure is given by a mixture model.

Mixture modeling. Mixture models are a cornerstone of statistical theory and practice. They give a principled way to capture heterogeneity in data and to identify and isolate subpopulations in varied datasets. Moreover, the mixture components in these models give rise to interpretable summaries of the data: in the analysis of text documents, mixture components can correspond to topics, which each document may be related to; in chemistry and biology, mixture components can correspond to particular compounds or biological molecules, which can arise in combination; in ecological modeling, mixture components can correspond to sub-species, different proportions of which appear in different habitats. The literature on mixture models goes back more than a century, starting with the classical example of Pearson [43], and continues to be a very active area of research. For space limitations we only refer to the monographs [37] for a general, fundamental treatment, and to [36, 54], which are more closely related to estimation of discrete mixtures, which we treat in detail in this work.

Despite the diversity of distance measures between probability distributions and the ubiquity of mixture models, most classical distances do not respect this interpretable structure. The following simple example illustrates this point. We revisit Karl Pearson's analysis [43] of Neapolitan crabs, one of the earliest examples of a mixture model in the statistical literature. By examining the distribution of body measurements of 1000 crabs obtained in Naples, Pearson identified two Gaussian mixture components, corresponding to two different species, which gave rise to the particular non-symmetric distribution observed in the data (left orange curve). Consider, however, an alternate population, which includes a third sub-species. We illustrate such an alternative composition in Figure 1. It gives rise to the distribution depicted as the green curve in the left panel, and it is almost indistinguishable from the orange curve. From an ecological standpoint, however, the two mixtures are structurally different, representing very different populations, but this difference cannot be captured by the classical distances.

The main goal of this work is to define and analyze a new, more discriminative, distance, the Sketched Wasserstein Distance W^S, dedicated to mixture distributions, that can detect such structural differences, despite potential overall similarity between the distributions that are being compared.

1.1 A new distance for mixture distributions

A class of mixture models based on a set of mixing distributions $\mathcal{A} = \{A_1, \ldots, A_K\}$ on a Polish space $\mathcal{Y}$ is a $(K-1)$-dimensional subfamily of the space $\mathscr{P}(\mathcal{Y})$ of probability distributions on $\mathcal{Y}$ consisting of convex combinations of the components of $\mathcal{A}$, given by

$$\mathcal{S} := \{r \in \mathscr{P}(\mathcal{Y}) : r = \sum_{i=1}^{K} \alpha_k A_k, \alpha := (\alpha_1, \ldots, \alpha_K) \in \Delta_K\},$$
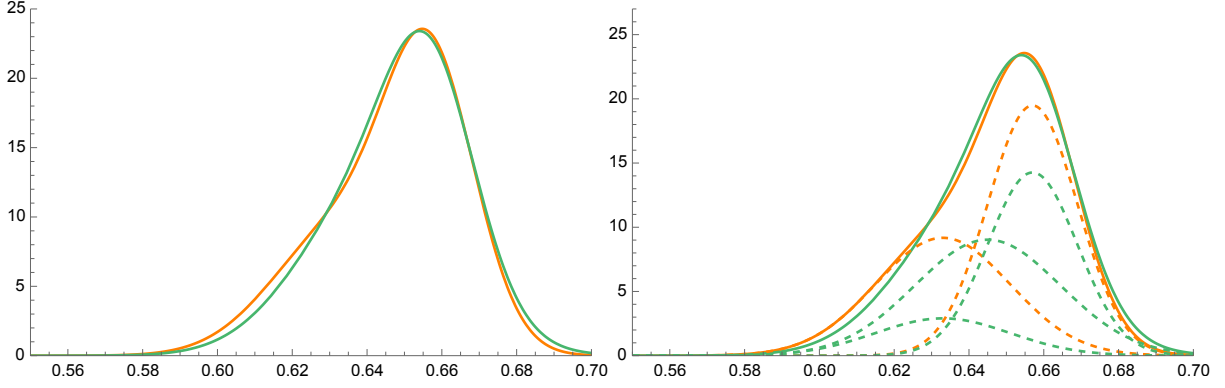
2

Figure 1: Pearson's crab mixture [43], in orange, and an alternative mixture, in green. Though these mixtures are close in total variation (left plot), they arise from very different sub-populations (right plot). On this example, $TV(r, s) \approx 0.062$, but $W^S(r, s) \approx 0.254$.

where Δ_K is the probability simplex in $\mathbb{R}^K$ and α is the vector of mixture weights. Throughout, we make a natural identifiability assumption that requires that none of the mixing components is itself a linear combination of other elements of $\mathcal{A}$:

Assumption 1. For each $k \in [K] := \{1, \ldots, K\}$, $A_k \notin \text{span}(\mathcal{A} \setminus A_k)$.

As described above, in many contexts, the mixing components $\mathcal{A}$ are interpretable units of analysis, for instance, semantic topics, interacting groups of proteins, or sub-populations of a community. Based on the application, we assume that the statistician has defined a metric d on the set $\mathcal{A}$. The metric may be chosen to reflect meaningful features of the problem at hand, and we do not make any assumptions on how d arises or is computed—it can be one of the standard distances on probability distributions, such as the total variation or Wasserstein distance, or a more exotic distance, for instance, one computed via a machine learning procedure or tailored to the particular data analysis task. In Section 5, we give an analysis of text data from more than 5000 NeurIPS publications from 1987 to 2015, where we equip the set of mixing components with the Jaccard distance [28]. Though this metric is common in topic modeling, it is defined only at the topic level (i.e., the mixture components), and does not readily apply to documents (i.e., mixtures).

Our first main contribution is to propose a new distance, W^S, which lifts the metric d on $\mathcal{A}$ to the set $\mathcal{S}$. In particular, W^S is uniquely characterized as the largest—that is, most discriminative—jointly convex metric on $\mathcal{S}$ for which the inclusion $\iota : (\mathcal{A}, d) \to (\mathcal{S}, W^S)$ is an isometry, therefore guaranteeing that W^S reproduces geometric features of d while being as useful as possible for downstream estimation and testing tasks. Despite the abstractness of this definition, we show that computing $W^S(r, s)$ for $r, s \in \mathcal{S}$ reduces to computing a Wasserstein distance between low-dimensional representations of r and s as elements of Δ_K. This fact justifies our calling W^S the *Sketched Wasserstein Distance*.

Our second main contribution is to study both estimation and inference for W^S, when the mixing components $\mathcal{A}$ are potentially unknown. Though W^S is well defined for both discrete and continuous mixtures, estimation of the mixing components comes with challenges that differ substantially between the discrete and continuous case. Our general bounds apply to both cases, but our main statistical contribution is to give a complete treatment of estimation and inference for W^S between potentially high-dimensional discrete mixtures, where both $|\mathcal{Y}|$ and K are allowed to potentially grow with the sample size. Our results will be developed for generic estimators of the discrete mixing components, and we will then further specialize to the

setting of topic-model estimation as our chief example. Estimation and inference for W^S for continuous measures will be treated in follow-up work.

In the following section, we provide a more detailed summary of our main results.

1.2 Our contributions

We consider throughout a finite set $\mathcal{A}$ of probability distributions on $\mathcal{Y}$ satisfying Assumption 1, equipped with a metric d, i.e., a nonnegative function satisfying the standard axioms:

- $d(A_k, A_\ell) = d(A_\ell, A_k)$ for all $k, \ell \in [K]$,

- $d(A_k, A_\ell) = 0$ if and only if $k = \ell$,

- $d(A_k, A_\ell) \leq d(A_k, A_m) + d(A_m, A_\ell)$ for all $k, \ell, m \in [K]$.

We denote by $\mathcal{S}$ the set of mixtures based on $\mathcal{A}$, that is, the convex hull of $\mathcal{A}$ in $\mathscr{P}(\mathcal{Y})$.

1.2.1 Construction of W^S and a deterministic error bound

In Section 2, we define our main object of interest, the Sketched Wasserstein Distance. As described above, we define the function W^S on $\mathcal{S} \times \mathcal{S}$ to be the largest jointly convex function agreeing with d on $\mathcal{A} \times \mathcal{A}$. The main result of Section 2 is that W^S is a metric on $\mathcal{S}$, and moreover that it is in fact closely connected to a Wasserstein distance on a *finite* metric space. Concretely, let $\mathcal{X} = (\mathcal{A}, d)$ be the metric space corresponding to the mixture components. We can identify the set $\mathscr{P}(\mathcal{X})$ of probability distributions on $\mathcal{X}$ with the simplex Δ_K, and given $\alpha, \beta \in \Delta_K$, we write $W(\alpha, \beta; d)$ for the Wasserstein distance

$$W(\alpha, \beta; d) = \inf_{\pi \in \Pi(\alpha, \beta)} \sum_{k,\ell=1}^{K} d(A_k, A_\ell) \pi_{k\ell},$$

where

$$\Pi(\alpha, \beta) = \left\{ \pi \in \Delta_{K \times K} : \begin{array}{ll} \sum_{\ell=1}^{K} \pi_{k\ell} = \alpha_k & \forall k \in [K] \\ \sum_{k=1}^{K} \pi_{k\ell} = \beta_\ell & \forall \ell \in [K] \end{array} \right\}$$

denotes the set of couplings between α and β.

Given $\alpha \in \Delta_K$ and $r \in \mathcal{S}$, we write, by convention, $r = A\alpha$ if $r = \sum_{k=1}^{K} \alpha_k A_k$, with $A_k \in \mathcal{A}$. Under Assumption 1, for each $r \in \mathcal{S}$, there exists a unique $\alpha \in \Delta_K$ for which $r = A\alpha$. Theorem 1 shows that if $r, s \in \mathcal{S}$ satisfy $r = A\alpha$ and $s = A\beta$, then W^S given by the abstract definition (1) satisfies

$$W^S(r, s) = W(\alpha, \beta; d).$$

In other words, W^S between r and s reduces to a Wasserstein distance between the mixture weights α and β, regarded as *sketches* of r and s in the space $\mathcal{S}$. We exploit this result to show that $(\mathcal{S}, W^S)$ is indeed a (complete) metric space, and that W^S possesses a dual formulation in terms of Lipschitz functions on $\mathcal{X}$.

The idea to use a Wasserstein distance to compare the mixtures $r = A\alpha$ and $s = A\beta$ is not new, and has appeared explicitly or implicitly in prior work on mixture models [6, 13, 16, 25, 40, 61, 63]. In these works, $W(\alpha, \beta; d)$ arose naturally as a quantity of interest for practical tasks involving mixtures. It was also prominently advocated as an analytical tool in [39], in which α and β are are called "latent mixing measures," and where the Wasserstein distance is identified as a convenient metric for proving minimax estimation rates for mixture estimation problems.

4

Our investigation takes a different perspective from these prior works. In the statistics literature, [25, 39, 40] use $W(\alpha, \beta; d)$ as a mathematical tool when analyzing estimators of finite mixtures (with d taken to be the Euclidean metric on the parameter space or an f-divergence between mixture components). However, these works do not establish $W(\alpha, \beta; d)$ as a bona fide metric on the space of mixtures, nor identify it as a canonical lifting of a distance on the mixture components. Our work is the first to establish this link. The papers that do view $W(\alpha, \beta; d)$ as a metric on mixtures themselves [13, 16, 63] consider the special case where d is itself a Wasserstein distance and focus on particular mixture types ([63] on topic models, [13, 16] on Gaussian mixtures). These works do not explore the possibility of lifting more general metrics, nor analyze the statistical properties of the resulting distance. By contrast, we develop our theory for arbitrary mixtures and arbitrary metrics, give general estimation bounds, and conduct a complete statistical analysis for the case of discrete mixtures.

We advocate for W^S as an object of interest in its own right as a tool for statistical data analysis, where d can be chosen by the statistician for the application in question. Moreover, we derive W^S from first principles as a canonical way to lift the distance d to the whole mixture model S. Our mathematical derivation of W^S in Section 2 gives additional justification for its use in the statistical literature.

Towards obtaining statistical results, in Section 2.1 we prove an error bound on W^S when neither α and β nor the mixing distributions A are known exactly. Concretely, letting $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^K$ and $\bar{d}$ be any metric on $[K] \times [K]$, we define a generic plug-in approximation $\bar{W}^S(r, s)$ for W^S and prove a deterministic error bound

$$|W^S(r, s) - \bar{W}^S(r, s)| \lesssim \operatorname{diam}(\mathcal{X})(\|\alpha - \bar{\alpha}\|_1 + \|\beta - \bar{\beta}\|_1) + \|d - \bar{d}\|_\infty \,.$$

Crucially, we do not require $\bar{\alpha}$, $\bar{\beta}$, and $\bar{d}$ to be bona fide estimators: $\bar{\alpha}$ and $\bar{\beta}$ do not need to lie in Δ_K, and $\bar{d}$ need not be of the "plug-in" form $\bar{d}(k, k') = d(\widehat{A}_k, \widehat{A}_{k'})$ for estimators $\widehat{A}_k, \widehat{A}_{k'} \in \mathscr{P}(\mathcal{Y})$, but can instead be any estimator of the metric. For instance, when d is an optimal transport distance, $\bar{d}$ could be defined using an approximation obtained via entropic regularization [14] or Gaussian smoothing [22], which have been shown to possess better statistical or computational properties [1, 20, 21, 38]. We exploit this flexibility in what follows to obtain estimators of W^S with tractable distributional limits allowing for valid asymptotic inference.

1.2.2 Estimation and inference of W^S between discrete mixtures

In Section 3, we give our main statistical results, which focus on estimation and inference of W^S for mixtures of p-dimensional discrete distributions $A_k \in \Delta_p$ for each $k \in [K]$, which we collect into a matrix $A \in \mathbb{R}^{p \times K}$. Given access to N i.i.d. samples from r and s, we define an estimator $\widehat{W}^S(r, s)$ and establish i) its finite sample behavior, including minimax rate optimality and ii) its asymptotic distribution. In both (i) and (ii), we consider the high-dimensional setting where $p = p(N)$ may depend on the sample size, and in (i) we allow $K = K(N)$ to grow with N as well. We do *not* assume that A is known and conduct our analysis under the assumption that we have access to an estimator $\widehat{A} \in \mathbb{R}^{p \times K}$, which we do *not* assume is independent of the samples from r and s. Beyond mathematical completeness, this setting is motivated by the concrete application to topic models, where often r, s, and A are estimated simultaneously on the same data. We further obtain stronger results under the more restrictive settings where $\widehat{A}$ is independent of the samples and where A is assumed to be known.

The main estimators we consider are of weighted least squares type: letting $\widehat{r}$ and $\widehat{s}$ denote empirical frequency estimates of r and s based on the observed samples, we estimate the mixture proportions by $\widehat{\alpha} = \widehat{A}^+ \widehat{r}$ and $\widehat{\beta} = \widehat{A}^+ \widehat{s}$, where $\widehat{A}^+$ is a carefully chosen left inverse of $\widehat{A}$. The

vectors $\widehat{\alpha}$ and $\widehat{\beta}$ are not, properly speaking, estimators, since they do not necessarily lie in Δ_K. We construct our estimator $\widehat{W}^S(r, s)$ by applying the plug-in approach described above with $\bar{\alpha} = \widehat{\alpha}$, $\bar{\beta} = \widehat{\beta}$, and $\bar{d}(k, k') = d(\widehat{A}_k, \widehat{A}_{k'})$.

In Section 3.1, we prove finite sample upper bounds on $\widehat{W}^S(r, s)$, which we obtain by combining an analysis of the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ with the deterministic error bound of Section 2.1. We show that with high probability,

$$\|\widehat{\alpha} - \alpha\|_1 \lesssim \sqrt{\frac{K \log K}{N}} + \epsilon_{\widehat{A}},$$

where $\epsilon_{\widehat{A}}$ denotes an error term that depends on how well $\widehat{A}$ approximates A. We recall that this error bound holds *without* assuming that $\widehat{A}$ is independent of $\widehat{r}$ and $\widehat{s}$, and we moreover prove that this term can be made smaller if $\widehat{A}$ is independent of $\widehat{r}$ and $\widehat{s}$, or removed entirely when A is known. By combining this bound with the results of Section 2.1, we establish in Theorem 4 that $\widehat{W}^S(r, s)$ enjoys a similar finite sample guarantee.

In Section 3.2, we complement these results with lower bounds, which establish that $\widehat{W}^S(r, s)$ is minimax rate optimal. Our lower bounds in fact hold for a more general setting: we prove a lower bound on Wasserstein distance estimation for any finite metric space, and show that these distances cannot be estimated on metric spaces of size K at a rate faster than $\sqrt{K/(N \log^2 K)}$, thus matching our upper bound up to logarithmic factors when A is known. Perhaps surprisingly, this is the first lower bound for Wasserstein estimation on finite spaces with nearly optimal dependence on the size of the space.

In Section 3.3, we treat the inference problem for W^S. Our main result in Section 3.3.1 shows that $\widehat{\alpha}$ is asymptotically normal, at a $\sqrt{N}$, rate when K is fixed, even if $p = p(N)$ and A is unknown. We then obtain the asymptotic distribution of $\widehat{W}^S(r, s)$ by an application of the δ-method in Theorem 7. These results justify our choice of the least squares approach in estimating α and β. Indeed, in the definition of $\widehat{W}^S(r, s)$ it may be more intuitive to employ bona fide estimators $\bar{\alpha}, \bar{\beta} \in \Delta_K$ in place of the pseudo-estimators $\widehat{\alpha}, \widehat{\beta} \in \mathbb{R}^K$, for instance by employing restricted weighted least squares or maximum likelihood estimators, which are guaranteed to lie in Δ_K, and these choices also lead to point estimates of W^S that have rate comparable, in some regimes, to that of the proposed $\widehat{W}^S(r, s)$. However, a rate analysis alone does not offer full insight into the study of estimation of W^S, since $\widehat{\alpha}$ and $\widehat{\beta}$ have net advantages over other restricted estimators from the perspective of their asymptotic distribution. Our empirical results indicate that the distribution of $\widehat{\alpha}$ and $\widehat{\beta}$ converges rapidly to normality, a property not shared by the restricted least squares or maximum likelihood estimators. In the case of the restricted least squares estimator $\widehat{\alpha}_{rls}$, this fact can be explained by the presence of an additional term in the expression for $\widehat{\alpha}_{rls} - \alpha$ corresponding to the Lagrange multipliers. Removing these terms gives rise exactly to our proposed un-restricted least squares estimator $\widehat{\alpha}$.

In Section 3.4, we specialize to the setting of topic models to discuss explicitly the estimation of A.

In Section 4, we illustrate via simulation our distributional convergence results and empirically evaluate the performance of the m-out-of-N bootstrap for estimating the limiting distribution of our estimator, for various choices of $m = N^\gamma$, $\gamma \in (0, 1)$.

Finally, Section 5 contains an application of W^S to the analysis of a real data set based on a corpus 5811 NeurIPS articles published between 1987 and 2015 [44].

Notation. The following notation is used throughout the paper. For any positive integer d, we write $[d] = \{1, \dots, d\}$. For any vector $v \in \mathbb{R}^d$, we write its ℓ_q norm as $\|v\|_q$ for all $0 \le q \le \infty$.

For any matrix $M \in \mathbb{R}^{d_1 \times d_2}$ and any sets $S_1 \subseteq [d_1]$ and $S_2 \subseteq [d_2]$, we write $M_{S_1}.$ and M_{S_2} the submatrices of M corresponding to rows in S_1 and columns in S_2, respectively. We further write $\|M\|_{\infty,1} = \max_i \|M_i.\|_1$ and $\|M\|_{1,\infty} = \max_j \|M_j\|_1$. For any symmetric matrix $Q \in \mathbb{R}^{d \times d}$, we write $\lambda_k(Q)$ the kth largest eigenvalue of Q for $k \in [d]$. The symbol Δ_d is reserved for the set of all vectors in the d-dimensional simplex $\{v \in \mathbb{R}_+^d : \|v\|_1 = 1\}$.

2 A new distance on the space of mixtures $\mathcal{S}$

We recall that $\mathcal{A} = \{A_1, \ldots, A_K\}$ is a set of probability distributions on a Polish space $\mathcal{Y}$, satisfying Assumption 1. We assume that $\mathcal{A}$ is equipped with a metric, d, and we denote the resulting metric space by $\mathcal{X} := (\mathcal{A}, d)$. Our goal is to lift the metric d to be a metric on the whole space $\mathcal{S} = \operatorname{conv}(\mathcal{A})$, the set of mixtures arising from the mixing components $A_1, \ldots, A_K$. Specifically, to define a new distance $W^S(\cdot, \cdot)$ between elements of $\mathcal{S}$, we propose that it satisfies two desiderata:

R1 : The function W^S should be a jointly convex function of its two arguments.

R2 : The function W^S should agree with the original distance d on the mixture components i.e., $W^S(A_k, A_\ell) = d(A_k, A_\ell)$ for $k, \ell \in [K]$.

We do *not* specifically impose the requirement that W^S be a metric, since many useful measure of distance between probability distributions, such as the Kullback–Leibler divergence, do not possess this property; nevertheless, in Corollary 1, we show that our construction does indeed give rise to a metric on $\mathcal{S}$.

R1 is motivated by both mathematical and practical considerations. Mathematically speaking, this property is enjoyed by a wide class of well behaved metrics, such as those arising from norms on vector spaces, and, in the specific context of probability distributions, holds for any f-divergence. Metric spaces with jointly convex metrics are known as Busemann spaces, and possess many useful analytical and geometrical properties [32, 50]. Practically speaking, convexity is crucial for both computational and statistical purposes, as it can imply, for instance, that optimization problems involving W^S are computationally tractable and that minimum-distance estimators using W^S exist almost surely. **R2** says that the distance should accurately reproduce the original distance d when restricted to the original mixture components, and therefore that W^S should capture the geometric structure induced by d.

To identify a unique function satisfying **R1** and **R2**, note that the set of such functions is closed under taking pointwise suprema. Indeed, the supremum of convex functions is convex [see, e.g., 26, Proposition IV.2.1.2], and given any set of functions satisfying **R2**, their supremum clearly satisfies **R2** as well.

For $r, s \in \mathcal{S}$, we therefore define $W^S(r, s)$ by

$$W^S(r, s) := \sup_{\phi(\cdot, \cdot): \phi \text{ satisfies } \mathbf{R1}\&\mathbf{R2}} \phi(r, s). \tag{1}$$

This function is the largest—most discriminative—quantity satisfying both **R1** and **R2**.

To give an example of the gap in the literature filled by our construction, consider the case where the elements of $\mathcal{A}$ are multivariate Gaussian distributions, each with covariance matrix $\Sigma \succ 0$. A natural metric on $\mathcal{A}$ is the *Mahalanobis distance*,

$$d_M(\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)) := \|\Sigma^{-1/2}(\mu_1 - \mu_2)\|_2 = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)}. \tag{2}$$

The metric d_M measures the distance between the means of the elements of $\mathcal{A}$ after applying the whitening transformation $\Sigma^{-1/2}$. A naïve lifting of d_M to the set $\mathcal{S}$ of mixtures of elements of $\mathcal{A}$ can be obtained by defining

$$\tilde{d}_M(r,s) := \|\Sigma^{-1/2}(\mu(r) - \mu(s))\|_2 \quad r, s \in \mathcal{S}, \tag{3}$$

where $\mu(r)$ and $\mu(s)$ denote the mean of r and s, respectively. The function $\tilde{d}_M$ satisfies **R1** and **R2**, but it does not effectively discriminate between elements of $\mathcal{S}$ and fails to be a metric: $\tilde{d}_M(r,s) = 0$ if and only if $\mu(r) = \mu(s)$, and not (as required for a metric) iff $r = s$. Thus, $\tilde{d}_M$ fails to distinguish *different* mixtures that may have the same means. On the other hand, by Eq. (1), $W^S(r,s) \geq \tilde{d}_M(r,s)$, and Theorem 1, below, shows that $W^S(r,s)$ *is* a metric on $\mathcal{S}$, and effectively discriminates between mixtures of elements of $\mathcal{A}$.

Our first main result characterizes W^S as an optimal transport distance. Recall that given $\alpha \in \Delta_K$ and $r \in \mathcal{S}$, we write, by convention, $r = A\alpha$ if $r = \sum_{k=1}^{K} \alpha_k A_k$, with $A_k \in \mathcal{A}$.

Theorem 1. *Under Assumption 1, for all $r, s \in \mathcal{S}$, with $\alpha, \beta \in \Delta_K$ such that $r = A\alpha$ and $s = A\beta$,*

$$W^S(r,s) = W(\alpha, \beta; d)$$
$$= \sup_{f \in \mathcal{F}} f^\top(\alpha - \beta). \tag{4}$$

where $W(\alpha, \beta; d)$ denotes the Wasserstein distance on $\mathcal{X}$ and where

$$\mathcal{F} = \{f \in \mathbb{R}^K : f_k - f_\ell \leq d(A_k, A_\ell), \ \forall k, \ell \in [K], \ f_1 = 0\}. \tag{5}$$

Proof. The proof can be found in Appendix A.1. □

Theorem 1 reveals a surprising universality result for the Wasserstein distances. On a convex set of measures, the Wasserstein distance is uniquely specified as the largest jointly convex function taking prescribed values at the extreme points. This characterization gives an axiomatic justification of the Wasserstein distance for statistical applications involving mixtures.

Our Theorem 1 can also be viewed as an extension of a similar result in the nonlinear Banach space literature: Weaver's theorem [56, Theorem 3.3 and Lemma 3.5] identifies the norm defining the *Arens–Eells space*, a Banach space into which the Wasserstein space isometrically embeds, as the largest seminorm on the space of finitely supported signed measures which reproduces the Wasserstein distance on pairs of Dirac measures.

Corollary 1. *Under Assumption 1, $(\mathcal{S}, W^S)$ is a complete metric space.*

The proof appears in Appendix A.2. Corollary 1 shows that $W^S(r,s)$ is the Wasserstein distance between the mixture weights α, β, regarded as *sketches (S)* of r and s in the space $\mathcal{X}$. This motivates its notation and the name of the proposed new distance, the *Sketched Wasserstein Distance*. Whether $\mathcal{Y}$ is finite or infinite, $\mathcal{X}$ is always a small, discrete metric space.

Remark 1. If we work with models that are not identifiable, we pay a price: the quantity defined by (1) is no longer a metric. This happens even if Assumption 1 is only slightly relaxed to

Assumption 2. For each $k \in [K]$, $A_k \notin \text{conv}(\mathcal{A} \setminus A_k)$.

The proof of Theorem 1 shows that under Assumption 2 we have, for all $r, s \in \mathcal{S}$

$$
\begin{aligned}
W^S(r,s) &= \inf_{\alpha,\beta \in \Delta_K : r = A\alpha, s = A\beta} W(\alpha, \beta; d) \qquad\qquad\qquad (6)\\
&= \inf_{\substack{\alpha,\beta \in \Delta_K \\ r = A\alpha, s = A\beta}} \sup_{f \in \mathcal{F}} f^\top (\alpha - \beta) \\
&= \sup_{f \in \mathcal{F}} \inf_{\substack{\alpha,\beta \in \Delta_K \\ r = A\alpha, s = A\beta}} f^\top (\alpha - \beta).
\end{aligned}
$$

However, W^S is only a semi-metric in this case: it is symmetric, and $W^S(r,s) = 0$ iff $r = s$, but it no longer satisfies the triangle inequality in general, as we show by example in Appendix A.2.1.

2.1 A deterministic error bound

In Theorem 1 we established that

$$
W^S(r,s) = \sup_{f \in \mathcal{F}} f^\top (\alpha - \beta), \qquad\qquad\qquad (7)
$$

for any $r, s \in \mathcal{S}$ with $r = A\alpha$ and $s = A\beta$. Then, if $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^K$ are any generic (pseudo) estimators of α and β, and if $\bar{d}$ is any metric on $[K]$, we propose to construct the generic estimator $\bar{W}^S(r,s)$ as a plug-in estimator of the right-hand side of (7), defined as

$$
\bar{W}^S(r,s) := \sup_{f \in \bar{\mathcal{F}}} f^\top (\bar{\alpha} - \bar{\beta}), \qquad\qquad\qquad (8)
$$

with

$$
\bar{\mathcal{F}} = \{ f \in \mathbb{R}^K : f_k - f_\ell \le \bar{d}(k, \ell), \ \forall k, \ell \in [K], \ f_1 = 0 \}. \qquad\qquad\qquad (9)
$$

We begin by stating a deterministic result that shows how the errors in estimating α, β and d combine into the estimation error of W^S. Let $\bar{\alpha}, \bar{\beta} \in \mathbb{R}^K$ be any generic estimators of α and β such that

$$
\|\bar{\alpha} - \alpha\|_1 \le \epsilon_\alpha, \qquad \|\bar{\beta} - \beta\|_1 \le \epsilon_\beta, \qquad\qquad\qquad (10)
$$

for some deterministic sequences $\epsilon_\alpha, \epsilon_\beta \ge 0$. Let d be such that:

$$
\max_{k,k' \in [K]} \left| \bar{d}(k, k') - d(A_k, A_{k'}) \right| \le \epsilon_d, \qquad\qquad\qquad (11)
$$

holds up to a label permutation, for some deterministic sequence $\epsilon_d \ge 0$.

Theorem 2. *Under Assumption 1, let r, s be any mixture distributions in $\mathcal{S}$ with $r = A\alpha$ and $s = A\beta$. For any $\bar{\alpha}, \bar{\beta}$ and $\bar{A}$ as in Eq. (10) – Eq. (11), the estimator $\bar{W}^S(r,s)$ given by Eq. (8) satisfies the following inequality*

$$
\left| \bar{W}^S(r,s) - W^S(r,s) \right| \le (\mathrm{diam}(\mathcal{X}) + \epsilon_d)(\epsilon_\alpha + \epsilon_\beta) + 2\,\epsilon_d,
$$

where the diameter of $\mathcal{X} = (\mathcal{A}, d)$ is $\mathrm{diam}(\mathcal{X}) := \max_{k,k'} d(A_k, A_{k'})$.

Proof. The proof can be found in Appendix A.3. $\qquad\qquad\qquad\square$

The proof of Theorem 2 is obtained by a stability analysis of the linear programs Eq. (7) and 8, the crucial step of which is to bound the Hausdorff distance between the polytopes $\mathcal{F}$ and $\bar{\mathcal{F}}$.

3 Estimation and Inference for the Sketched Wasserstein Distance between discrete distributions

In Section 2 we introduced the distance W^S, valid for either continuous or discrete mixtures. In this section, and the remainder of the paper, we focus on the estimation of W^S for mixtures of p-dimensional discrete distributions $A_k \in \Delta_p$ for each $k \in [K]$. We collect the mixture components in the matrix $A \in \mathbb{R}^{p \times K}$, and the formal notation $r = A\alpha$ introduced in Section 2 above is henceforth standard matrix-vector multiplication.

Consider any estimator $\widehat{A}$ of A, assume access to N i.i.d. samples from r and s, and let

$$N\widehat{r} \sim \mathrm{Multinomial}_p(N; r), \qquad N\widehat{s} \sim \mathrm{Multinomial}_p(N; s). \tag{12}$$

We estimate the mixture proportions by weighted least squares estimators

$$\widehat{\alpha} = \widehat{A}^+ \widehat{r}, \qquad \widehat{\beta} = \widehat{A}^+ \widehat{s}, \tag{13}$$

with

$$\widehat{A}^+ := (\widehat{A}^\top \widehat{D}^{-1} \widehat{A})^{-1} \widehat{A}^\top \widehat{D}^{-1} \quad \text{and} \quad \widehat{D} := \mathrm{diag}(\|\widehat{A}_{1\cdot}\|_1, \dots, \|\widehat{A}_{p\cdot}\|_1). \tag{14}$$

The use of the pre-conditioner $\widehat{D}$ in the definition of $\widehat{A}^+$ is reminiscent of the definition of the normalized Laplacian in graph theory: it moderates the size of the j-th entry of each mixture estimate, across the K mixtures, for each $j \in [p]$. As a consequence, $\widehat{A}^\top \widehat{D}^{-1} \widehat{A}$ is a doubly stochastic matrix, and its largest eigenvalue is 1 (see also Lemma 1, below).

Using the strategy outlined in the previous section, we then estimate the Sketched Wasserstein Distance by

$$\widehat{W}^S(r, s) := \sup_{f \in \widehat{\mathcal{F}}} f^\top (\widehat{\alpha} - \widehat{\beta}), \tag{15}$$

where $\widehat{\mathcal{F}}$ is defined as in Eq. (9), with $\bar{d}$ replaced by $\widehat{d}$, where $\widehat{d}(k, k')$ is an estimator of $d(A_k, A_{k'})$. While $\widehat{d}$ can be a plug-in estimator of the form $d(\widehat{A}_k, \widehat{A}_{k'})$, it is sometimes statistically beneficial to use another estimator with better properties: for example, in the high-dimensional discrete setting, when d is the total variation or Hellinger distance, plug-in estimators are suboptimal in a minimax sense, and more sophisticated estimators based on polynomial approximation achieve faster rates [24, 29]. The choice of estimator $\widehat{d}$ only affects our final bounds via the error term ϵ_d defined in Eq. (11).

The following sub-sections are devoted to the study, under multinomial sampling schemes, of (i) The finite sample behavior of $\widehat{W}^S(r, s)$, including minimax rate optimality and of (ii) The asymptotic distribution of $\widehat{W}^S(r, s)$.

3.1 Finite sample error bounds in W^S estimation

In the following, we apply the general Theorem 2 to obtain the rate of convergence of our proposed estimator $\widehat{W}^S(r, s)$ given by Eq. (15). To this end, for any $r, s \in \mathcal{S}$ with $r = A\alpha$ and $s = A\beta$, we first establish ℓ_1 error bounds for mixture weight estimators, then proceed to evaluate the induced error bounds in W^S estimation.

3.1.1 Finite sample ℓ_1 error bounds for the weighted least squares estimator of the mixture weights

From now on, we assume $\min_{j \in [p]} \|A_{j\cdot}\|_1 > 0$ for simplicity. Otherwise, a pre-screening procedure could be used to reduce the dimension p such that the aforementioned condition holds.

Recall that both $\widehat{\alpha}$ and $\widehat{\beta}$ depend on the estimation of A. In the following we provide a general rate-analysis of $\|\widehat{\alpha} - \alpha\|_1$ that is valid for any estimator $\widehat{A} \in \mathbb{R}^{p \times K}$ of A which satisfies:

$$\widehat{A}_k \in \Delta_p, \text{ for } k \in [K], \qquad\qquad \text{rank}(\widehat{A}) = K; \qquad\qquad (16)$$

$$\max_{k \in [K]} \|\widehat{A}_k - A_k\|_1 \le \epsilon_{1,\infty}, \qquad\qquad \max_{j \in [p]} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_1}{\|A_{j\cdot}\|_1} \le \epsilon_{\infty,1}. \qquad (17)$$

Here $\epsilon_{1,\infty}$ and $\epsilon_{\infty,1}$ are some positive deterministic sequences and Eq. (17) only needs to hold up to a label permutation among columns of A.

Since $\widehat{\alpha}$ is a weighted least squares estimator, the following matrix naturally appears in our analysis

$$M := A^\top D^{-1} A, \qquad \text{with} \quad D = \text{diag}(\|A_{1\cdot}\|_1, \cdots, \|A_{p\cdot}\|_1). \qquad (18)$$

In our analysis, the matrix M is analogous to the design matrix in a linear regression problem. The properties of M relevant to our analysis are collected in Lemma 1. See Appendix A.4 for its proof.

Lemma 1. *Under Assumption 1, the following statements hold.*

(1) *M is strictly positive definite.*

(2) *$\lambda_1(M) = 1$ and $M\mathbb{1}_K = \mathbb{1}_K$.*

(3) *$\lambda_K(M) \ge \frac{\lambda_K(A^\top A)}{\|A\|_{\infty,1}}$,*

(4) *$\frac{1}{\lambda_K(M)} \le \|M^{-1}\|_{\infty,1} \le \frac{\sqrt{K}}{\lambda_K(M)}$.*

The following theorem provides a finite sample bound of $\|\widehat{\alpha} - \alpha\|_1$ with $\widehat{\alpha}$ defined in Eq. (13) for any estimator $\widehat{A}$ satisfying Eq. (16) and Eq. (17).

Theorem 3. *Grant Assumption 1 and Eq. (12). Let $\widehat{A}$ be as stated above and assume*

$$\|M^{-1}\|_{\infty,1} \left(\epsilon_{1,\infty} + 2\epsilon_{\infty,1}\right) \le 1/2. \qquad (19)$$

Fix any $t \in (0, 1)$ and assume

$$2K \log(2K/t) \le N. \qquad (20)$$

Then, for any $r \in \mathcal{S}$ with $r = A\alpha$, with probability at least $1 - t$, the estimator $\widehat{\alpha}$ defined in Eq. (13) satisfies

$$\|\widehat{\alpha} - \alpha\|_1 \le \frac{4}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}} + 2\|M^{-1}\|_{\infty,1} \left(\epsilon_{1,\infty} + 2\epsilon_{\infty,1}\|\widehat{r} - r\|_1\right). \qquad (21)$$

Proof. The proof can be found in Appendix A.5. $\qquad\qquad\square$

Remark 2 (Discussion of the bound in Theorem 3)**.** The error bound in Eq. (21) contains two terms: the first term is the minimum error of estimating the mixture weight α even if A were known, while the second term originates from the error of estimating A, more specifically, of estimating $A^+ = M^{-1}A^\top D$.

The error of estimating A^+ (the second term in Eq. (21)) itself has two terms that are given in terms of the scaled row-wise error ($\epsilon_{\infty,1}$) and the column-wise error ($\epsilon_{1,\infty}$) of estimating A. The former is multiplied by $\|\widehat{r} - r\|_1$, which is always no greater than 2, but can also converge to zero at a fast rate. For instance, one has $\|\widehat{r} - r\|_1 = \mathcal{O}_{\mathbb{P}}(\sqrt{1/N})$ under the summability

condition $\sum_{j=1}^{p} \sqrt{r_j} = \mathcal{O}(1)$. Furthermore, the entire error term containing $\epsilon_{\infty,1}$ becomes zero when $\widehat{A}$ is estimated independently of $\widehat{r}$. This fact is also given in Corollary 2 below.

Quantities involving the "design" matrix M also affect the magnitude of the bound. In the most favorable situation where the columns of A have disjoint support, $M = \mathbf{I}_K$ and $\|M^{-1}\|_{\infty,1} = \lambda_K^{-1}(M) = 1$. More generally, in light of point (3) of Lemma 1, if $\lambda_K(A^\top A) > c\|A\|_{\infty,1}$ for some constant $c > 0$, then $\lambda_K^{-1}(M) = O(1)$.

Remark 3. Condition (19) states the requirement on how well $\widehat{A}$ estimates A in terms of both the row-wise error and the column-wise error to ensure good conditions of $\widehat{M} = \widehat{A}^\top \widehat{D}^{-1} \widehat{A}$. Indeed, Lemma 6 in Appendix A.5 shows that condition (19) ensures $\lambda_K(\widehat{M}) \geq \lambda_K(M)/2$ and $\|\widehat{M}^{-1}\|_{1,\infty} \leq 2\|M^{-1}\|_{1,\infty}$.

Condition (20) on the other hand is assumed only to simplify the bound. Otherwise, the first term in Eq. (21) gets replaced by

$$\frac{2}{\lambda_K(M)} \left(\sqrt{\frac{2K \log(2K/t)}{N}} + \frac{4K \log(2K/t)}{N} \right).$$

Theorem 3 is valid for both $\widehat{A}$ that is independent of $\widehat{r}$ and $\widehat{A}$ that depends on $\widehat{r}$. In the following corollary, we show that the rate of $\|\widehat{\alpha} - \alpha\|_1$ could be improved when using either an independent $\widehat{A}$ or one has direct access to the true A.

Corollary 2. *Grant Assumption 1 and Eq. (12). Fix any $t \in (0,1)$ and assume (20) holds.*

(a) *Assume $\widehat{A}$ is independent of $\widehat{r}$ and condition (19) holds. Then, with probability at least $1 - t$,*

$$\|\widehat{\alpha} - \alpha\|_1 \leq \frac{4}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}} + 2\|M^{-1}\|_{1,\infty}\, \epsilon_{1,\infty}.$$

(b) *Assume $\widehat{A} = A$. Then, with probability at least $1 - t$,*

$$\|\widehat{\alpha} - \alpha\|_1 \leq \frac{4}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}}.$$

Proof. The proof can be found in Appendix A.5. □

Comparing to Theorem 4, the rate in part (a) of Corollary 2 does not contain the term $4\|M^{-1}\|_{\infty,1}\epsilon_{\infty,1}\|\widehat{r} - r\|_1$. We view this as the advantage of using an independent $\widehat{A}$ relative to $\widehat{r}$. It is worth mentioning that independence between $\widehat{A}$ and $\widehat{r}$ is natural in several applications where one has a large set of pre-seen data to estimate the mixture components $\mathcal{A}$ and then wants to quantify the distances for unseen new data.

3.1.2 Finite sample ℓ_1 bounds for alternative mixture weight estimators

As mentioned before, other estimators (restricted to Δ_K) of the mixture weights α, β can be used in (8) to estimate $W^S(r, s)$ in lieu of the pseudo-estimators $\widehat{\alpha}, \widehat{\beta}$. Our general Theorem 2 ensures that finite sample bounds of the induced estimator of the W^S distance can be easily derived from their ℓ_1-norm rates of convergence. In this section, we discuss two alternative estimators of the mixture weights.

The closest restricted estimator to $\widehat{\alpha}$ in (13) is the following restricted (weighted) least squares estimator

$$\widehat{\alpha}_{rls} := \frac{1}{2} \min_{\alpha \in \Delta_K} \left\| \widehat{D}^{-1/2}(\widehat{r} - \widehat{A}\alpha) \right\|_2^2. \tag{22}$$

with $\widehat{D}$ defined in (14). Note that if we do not restrict the feasible set in Δ_K, then $\widehat{\alpha}_{rls}$ becomes our proposed $\widehat{\alpha}$. For completeness, we establish the full ℓ_1-norm error bound of $\widehat{\alpha}_{rls} - \alpha$ in Appendix A.6 and only state its simplified version below for known A.

Lemma 2. *Under Assumption 1 and Eq. (12), for any $t \geq 0$, with probability at least $1 - t$, the estimator $\widehat{\alpha}_{rls}$ in Eq. (22) with $\widehat{A} = A$ satisfies*

$$\|\widehat{\alpha}_{rls} - \alpha\|_1 \leq \min \left\{ 2, \frac{8\sqrt{2}}{\lambda_K(M)} \sqrt{(\|\alpha\|_0 \|A^\top D^{-1} r\|_\infty + 1) \frac{\|\alpha\|_0 \log(2K/t)}{N}} \right\}$$

where $\|\alpha\|_0$ denotes the number of non-zero entries of α and $\|A^\top D^{-1} r\|_\infty \leq 1$.

Comparing this result to Corollary 2, we observe that we no longer need condition (20): since $\widehat{\alpha}_{rls} \in \Delta_K$, we have the trivial bound $\|\widehat{\alpha}_{rls} - \alpha\|_1 \leq 2$. Consequently, a similar condition to (20) in Theorem 3 can be assumed without loss of generality. Also, the fact that $\widehat{\alpha}_{rls} \in \Delta_K$ implies that the bound of $\|\widehat{\alpha}_{rls} - \alpha\|_1$ depends on $\|\alpha\|_0$ rather than K. Provided that $\|\alpha\|_0 \|A^\top D^{-1} r\|_\infty \leq C$, for instance, when $\|\alpha\|_\infty \leq C/\|\alpha\|_0$, for some constant $C > 0$, we have

$$\|\widehat{\alpha}_{rls} - \alpha\|_1 \leq \min \left\{ 2, \frac{C'}{\lambda_K(M)} \sqrt{\frac{\|\alpha\|_0 \log(2K/t)}{N}} \right\}$$

with some $C' = C'(C) > 0$. This bound is potentially faster than that in part (b) of Corollary 2 when α is sparse. Although this rate analysis supports the validity of estimating W^S by using restricted least squares estimators of the mixture weights, our discussion in Section 3.3.1 reveals that our proposed *unrestricted* least squares estimator enjoys faster speed of convergence to a normal limit, making it amenable to derive asymptotically valid confidence intervals the resulting estimator of W^S.

Both $\widehat{\alpha}$ and $\widehat{\alpha}_{rls}$ are well defined for any data distributions, and analyzed under the multinomial sampling scheme assumption. One can also use this assumption to construct another natural estimator of α, the likelihood-based MLE estimator, $\widehat{\alpha}_{mle} \in \Delta_K$, recently studied in [6]. Under certain conditions on A and α, [6] shows that $\|\widehat{\alpha}_{mle} - \alpha\|_1 = \mathcal{O}_{\mathbb{P}}(\sqrt{s/N})$ for known A. The MLE estimator, however, also suffers slow speed of convergence to a normal limit similar to $\widehat{\alpha}_{rls}$ (see, Section 3.3.1, for a detailed discussion).

3.1.3 Finite sample error bounds for the Sketched Wasserstein Distance estimates

Combining Theorem 2 and Theorem 3 yields the rate of convergence of our proposed estimator $\widehat{W}^S(r, s)$, given by Eq. (15).

Theorem 4. *Grant Assumption 1 and Eq. (12). Assume that $\widehat{d}$ satisfies Eq. (11) and $\widehat{A}$ satisfies Eq. (16) and Eq. (17), and assume condition (19). Fix any $t > 0$, further assume (20) holds. Then, with probability at least $1 - t$, the estimator $\widehat{W}^S(r, s)$ defined in Eq. (15) satisfies*

$$\begin{aligned} |\widehat{W}^S(r, s) - W^S(r, s)| \ &\leq \ \frac{16 \operatorname{diam}(\mathcal{X})}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}} \\ &\quad + 8 \operatorname{diam}(\mathcal{X}) \|M^{-1}\|_{1,\infty} (\epsilon_{1,\infty} + 4 \epsilon_{\infty,1}) + 2 \epsilon_d. \end{aligned}$$

Proof. The proof follows by combining Theorem 2 and Theorem 3 and using the bound $\|\widehat{r} - r\|_1 + \|\widehat{s} - s\|_1 \leq 4$. $\qquad\square$

One salient feature of estimating W^S is that the bound only depends on properties of the metric space $\mathcal{X}$ rather than on any properties of the ambient space $\mathcal{Y}$. In particular, for known A and provided that $\lambda_K^{-1}(M) = \mathcal{O}(1)$, the ambient dimension $p = |\mathcal{Y}|$ does not appear in our bounds, as shown in Corollary 3 below.

Moreover, Theorem 4 shows that the error of estimating the mixture components $\mathcal{A}$ and their distance $d(A_k, A_{k'})$ enters the bound of estimating W^S additively. In some cases, it is possible to simplify the resulting bounds when d satisfies additional properties. For example, if d extends to a metric on $\mathcal{A} \cup \widehat{\mathcal{A}}$ which satisfies

$$d(A, A') \leq L\|A - A'\|_1, \qquad \text{for any } A, A' \in \mathcal{A} \cup \widehat{\mathcal{A}}, \tag{23}$$

then we have

$$\epsilon_d \leq 2 \max_k d(A_k, \widehat{A}_k) \leq 2L\epsilon_{1,\infty},$$

implying that the ϵ_d term can be absorbed into the $\epsilon_{1,\infty}$ term. The condition in Eq. (23) holds, for instance, if d is the Wasserstein distance (with $L = \text{diam}(\mathcal{X})$), or an ℓ_q norm on $\mathbb{R}^p$ (with $L = 1$).

For $\widehat{A}$ independent of $\widehat{r}$ and $\widehat{s}$ as well as known A, the bound in Theorem 4 can be immediately improved, according to Corollary 2. We include the result for completeness.

Corollary 3. *Grant Assumption 1 and Eq. (12). Fix any $t > 0$ and assume Eq. (20) holds.*

(a) *Assume $\widehat{A}$ is independent of $\widehat{r}$ and condition (19) holds, and that $\widehat{d}$ satisfies Eq. (11). Then, with probability at least $1 - t$,*

$$|\widehat{W}^S(r,s) - W^S(r,s)| \leq \frac{16 \text{ diam}(\mathcal{X})}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}}$$
$$+ 8 \text{ diam}(\mathcal{X}) \|M^{-1}\|_{1,\infty} \epsilon_{1,\infty} + 2 \epsilon_d.$$

(b) *Assume $\widehat{A} = A$ and $\widehat{d} = d$. Then, with probability at least $1 - t$,*

$$|\widehat{W}^S(r,s) - W^S(r,s)| \leq \frac{16 \text{ diam}(\mathcal{X})}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}}.$$

3.2 Minimax lower bound for known A

In this section, we prove a minimax lower bound, showing that our estimator of $W^S(r,s)$ is nearly rate optimal when A is known. Our lower bound applies to a stronger observation model, where we have direct access to i.i.d. samples from the measures α and β on $\mathcal{X}$. For known A, there exists a *transition* (in the sense of Le Cam [35]) from this model to the observation model where only samples from r and s are observed; therefore, a lower bound for the direct observation model implies a lower bound for the setting considered elsewhere in Section 3 as well. However, we choose to prove a lower bound for the stronger model because it in fact applies outside the context of mixtures models to the estimation of the Wasserstein distance on *any* finite metric space.

In the remainder of this section, we denote by $(\mathcal{X}, d)$ an arbitrary finite metric space of cardinality K, and let

$$\kappa_{\mathcal{X}} := \frac{\max_{x,x' \in \mathcal{X}} d(x, x')}{\min_{x \neq x' \in \mathcal{X}} d(x, x')}.$$

The quantity $\kappa_{\mathcal{X}}$ is equal to 1 if and only if d is the discrete metric d_0 on $\mathcal{X}$; more generally, $\kappa_{\mathcal{X}}$ denotes the minimum distortion of any metric embedding of $(\mathcal{X}, d)$ into $(\mathcal{X}, d_0)$.

Our main lower bound is the following.

Theorem 5. *Let $(\mathcal{X}, d)$ and $\kappa_{\mathcal{X}}$ be as above. There exist positive universal constants C, C' such that if $N \geq CK$, then*

$$\inf_{\widehat{W}} \sup_{\alpha, \beta \in \mathscr{P}(\mathcal{X})} \mathbb{E}_{\alpha, \beta} |\widehat{W} - W(\alpha, \beta; d)| \geq C' \text{diam}(\mathcal{X}) \left(\kappa_{\mathcal{X}}^{-2} \sqrt{\frac{K}{N(\log K)^2}} \vee \frac{1}{\sqrt{N}} \right),$$

where the infimum is taken over all estimators constructed from N i.i.d. samples from α and β, respectively.

Proof. The proof appears in Appendix A.7 □

As mentioned above, Theorem 5 shows that our estimators match the minimax optimal rate, up to logarithmic factors, when A is known, $\kappa_{\mathcal{X}} = \mathcal{O}(1)$ and $\lambda_K^{-1}(M) = \mathcal{O}(1)$. To our knowledge, Theorem 5 is the first lower bound for rates of estimation of the Wasserstein distance with nearly optimal dependence on the cardinality of the underlying metric space.

We do not know whether the logarithmic factors can be removed, nor whether the dependence on $\kappa_{\mathcal{X}}$ is optimal. However, it is possible to show that the dependence on $\kappa_{\mathcal{X}}$ cannot be removed in general, since there are finite metric spaces for which $\kappa_{\mathcal{X}} \to \infty$ for which the Wasserstein distance can be estimated at a rate independent of K [see, e.g., 57]. For example, if $\mathcal{X}$ is isomorphic to a subset of $[0, 1]$, then the minimax rate is $N^{-1/2}$, independent of K.

Our lower bound follows a strategy of [41] and is based on a reduction to estimation in the total variation distance, for which we prove a minimax lower bound based on the method of "fuzzy hypotheses" [53]. Though sharp lower bounds for estimating the total variation distance are known [29], the novelty of our bounds involves the fact that we must design priors which exhibit a large *multiplicative* gap in the functional of interest, whereas prior techniques [29, 60] are only able to control the magnitude of the additive gap between the values of the functionals.

3.3 The Asymptotic Distribution of Sketched Wasserstein Distance Estimators

Recall that

$$\widehat{W}^S(r, s) := \sup_{f \in \widehat{\mathcal{F}}} f^\top (\widehat{\alpha} - \widehat{\beta}), \tag{24}$$

where $\widehat{\mathcal{F}}$ is defined after Eq. (15).

In this section we first determine the limiting distribution of $\widehat{\alpha} - \widehat{\beta}$, then derive the limiting distribution of $\widehat{W}^S(r, s)$ by applying an appropriate version of the functional δ-method, since $W^S(r, s)$ is Hadamard directionally-differentiable (see Proposition 2 in Appendix A.10). The deviation of $\widehat{\mathcal{F}}$ from $\mathcal{F}$ is controlled as an intermediate step.

A strategy similar in spirit was used in [47, 51] to obtain the $\sqrt{N}$ limit in distribution of $W(\widehat{r}, \widehat{s}; \rho)$, the plug in estimator of the *standard* Wasserstein distance when $r, s \in \mathscr{P}(\mathcal{Y})$ for a finite or countable space $\mathcal{Y}$ equipped with a metric ρ. Here, $\widehat{r}$ and $\widehat{s}$ are empirical estimates of two discrete distributions r and s, whose dimension does *not* grow with N. This strategy cannot readily be extended to the case of practical interest in which p is growing with N, as in this scenario $\widehat{r}$ and $\widehat{s}$ do not have well defined distributional limits.

In contrast, even if $p = p(N)$, we show below that when $r, s \in \mathcal{S}$, then $\widehat{W}^S(r, s)$ does have a $\sqrt{N}$ distributional limit, as long as the number of mixtures K is independent of N. In

Section 3.4 we give a concrete example to topic models, in which case this assumption on K translates to saying that the number of topics K in the entire corpus only depends on how many documents we observe in the corpus, rather than the length of each individual document.

Throughout this section, K is assumed to be fixed and independent of N, while p as well as the parameters in r, s and A is allowed to grow with N.

3.3.1 Asymptotic normality of mixture weight estimators

Perhaps surprisingly, the limiting distribution of estimators of mixture weights in discrete mixture models is largely unexplored. In this section we show that the weighted least squares estimator $\widehat{\alpha}$ in (13) is asymptotically normal when the mixtures are either known, or are estimated from the data.

Let $r \in \mathcal{S}$ with $r = A\alpha$. The following theorem establishes the asymptotic normality of $\widehat{\alpha} - \alpha$. To state our result, let

$$V_r = A^\top D^{-1}\Sigma_r D^{-1}A, \qquad \text{with} \quad \Sigma_r = \operatorname{diag}(r) - rr^\top. \tag{25}$$

Theorem 6. *Under Assumption 1 and condition (12), let $\widehat{A}$ be any estimator such that Eq. (16) and Eq. (17) hold. Assume $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N})$, $\lambda_K^{-1}(V_r) = o(N)$ and*

$$(\epsilon_{1,\infty} + \epsilon_{\infty,1})\sqrt{N/\lambda_K(V_r)} = o(1), \quad as \ N \to \infty. \tag{26}$$

Then, we have the following convergence in distribution as $N \to \infty$,

$$\sqrt{N}(A^+\Sigma_r A^{+\top})^{-1/2}(\widehat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}_K(0, \mathbf{I}_K).$$

Proof. The proof can be found in Appendix A.9. $\qquad\qquad\square$

Remark 4. We begin by observing that if p is finite, independent of N, and A is known, the conclusion of the theorem follows trivially from the fact that $\sqrt{N}(\widehat{r} - r) \xrightarrow{d} \mathcal{N}_K(0, \Sigma_r)$ and $\widehat{\alpha} = A^+\widehat{r}$, under no conditions beyond the multinomial sampling scheme assumption. Our theorem generalizes this to the practical situation when A is estimated, from a data set that is potentially dependent on that used to construct $\widehat{\alpha}$, and when $p = p(N)$.

Condition $\lambda_K^{-1}(V_r) = o(N)$ is a mild regularity condition that is needed even if A is known, but both $p = p(N)$ and parameters in A and r are allowed to grow with N.

If V_r is rank deficient, so is the asymptotic covariance matrix $A^+\Sigma_r A^{+\top}$. This happens, for instance, when the mixture components A_k have disjoint supports and the weight vector α of $r = A\alpha$ is sparse. Nevertheless, a straightforward modification of our analysis leads to the same conclusion, except that the inverse of $A^+\Sigma_r A^{+\top}$ gets replaced by a generalized inverse and the limiting covariance matrix becomes $\mathbf{I}_s$ with $s = \operatorname{rank}(A^+\Sigma_r A^{+\top})$ and zeroes elsewhere.

Condition (26), on the other hand, ensures that the error of estimating the mixture components becomes negligible. We provide a sufficient condition for (26) in the topic model setting discussed in Section 3.4.

As mentioned in Section 3.1.2, an alternative point estimator of $W^S(r, s)$ can be based on estimators of the mixture weights that are restricted to the probability simplex Δ_K. Specializing to the restricted weighted least squares estimator $\widehat{\alpha}_{rls}$ given by Eq. (22) with $\widehat{A} = A$, the K.K.T. conditions of (22) imply

$$\sqrt{N}(\widehat{\alpha}_{rls} - \alpha) = \sqrt{N}A^+(\widehat{r} - r) + \sqrt{N}M^{-1}(\lambda - \mu\mathbb{1}_K). \tag{27}$$

Here $\lambda \in \mathbb{R}_+^K$ and $\mu \in \mathbb{R}$ are Lagrange variables corresponding to the restricted optimization in Eq. (22). We see that the first term on the right-hand side in Eq. (27) is asymptotically normal,

while the second one does not always converge to a normal limit. Thus, it is natural to consider a modified estimator

$$\widetilde{\alpha} := \widehat{\alpha}_{rls} - M^{-1}\left(\lambda - \mu \mathbb{1}_K\right)$$

by removing the second term in Eq. (27). Interestingly, the new estimator $\widetilde{\alpha}$ is nothing but our proposed estimator $\widehat{\alpha}$. This suggests that $\sqrt{N}(\widehat{\alpha} - \alpha)$ converges to a normal limit faster than that of $\sqrt{N}(\widehat{\alpha}_{rls} - \alpha)$. Figure 2 in Section 4.1 gives an instance of this fact.

Moreover, the MLE estimator $\widehat{\alpha}_{mle} \in \Delta_K$ of α as studied in [6] also suffers the slow speed of convergence to a normal limit, which can be readily seen from displays (E.2) – (E.3) in the Appendix of [6].

3.3.2 The Limiting distribution of the W^S estimator

As announced above, for any $r, s \in \mathcal{S}$ with $r = A\alpha$ and $s = A\beta$, we derive in the limiting distribution of our proposed estimator $\widehat{W}^S(r, s)$, based on a generic estimator $\widehat{A}$ of A, and the weighted least squares estimators $\widehat{\alpha}$ and $\widehat{\beta}$ of the mixture weights.

For any $r, s \in \mathcal{S}$, let

$$Z_{rs} \sim \mathcal{N}_K(0, Q_{rs}) \tag{28}$$

where we assume that

$$Q_{rs} := \lim_{N \to \infty} A^+ \left(\Sigma_r + \Sigma_s\right) A^{+\top} \in \mathbb{R}^{K \times K}$$

exists, and that we rely on the fact that K is independent of N to define the limit, but that the model parameters A, r, and s may depend on $p = p(N)$. Similar to Eq. (25), we write $V_s = A^\top D^{-1} \Sigma_s D^{-1} A$.

Theorem 7. *Under Assumption 1 and the multinomial sampling assumption Eq. (12), let $\widehat{d}$ and $\widehat{A}$ be any estimators such that Eq. (11), Eq. (16) and Eq. (17) hold. Assume $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N})$, $\epsilon_d \sqrt{N} = o(1)$, $\max\{\lambda_K^{-1}(V_r), \lambda_K^{-1}(V_s)\} = o(N)$ and*

$$(\epsilon_{1,\infty} + \epsilon_{\infty,1})\sqrt{N / \min\{\lambda_K(V_r), \lambda_K(V_s)\}} = o(1), \quad as\ N \to \infty. \tag{29}$$

Then, as $N \to \infty$, we have the following convergence in distribution,

$$\sqrt{N}\left(\widehat{W}^S(r, s) - W^S(r, s)\right) \xrightarrow{d} \sup_{f \in \mathcal{F}'(r,s)} f^\top Z_{rs} \tag{30}$$

where

$$\mathcal{F}'(r, s) := \mathcal{F} \cap \left\{f \in \mathbb{R}^K : f^\top(\alpha - \beta) = W^S(r, s)\right\}. \tag{31}$$

The proof of Theorem 7 can be found in Appendix A.10. It builds on the asymptotic normality results of our estimators of the mixture weights in Theorem 6 and then applies a variant of the δ-method that is suitable for Hadamard-directionally differentiable functions. The second step requires to verify the W^S distance in (7) is Hadamard-directionally differentiable, which is shown in Proposition 2 of Appendix A.10.

Comparing to Theorem 6, Theorem 7 additionally requires $\epsilon_d\sqrt{N} = o(1)$; as in Section 3.1.3, this requirement can be subsumed into the requirement on $\epsilon_{1,\infty}$ under Eq. (23).

Our results can be readily generalized to the cases where $\widehat{r}$ and $\widehat{s}$ have different sample sizes. We refer to Appendix B for its precise statement.

For completeness, we include the following corollary pertaining to the case where A is estimated from the data, but K, p, r, and s do not grow with N. In this situation, the conditions of Theorem 7 can be greatly simplified.

Corollary 4. *Under Assumption 1 and and the multinomial sampling assumption Eq. (12),* *assume that p and the entries in A, r and s do not depend on N. Let $\widehat{d}$ and $\widehat{A}$ be any estimators* *such that Eq. (11), Eq. (16) and Eq. (17) hold. Assume $(\epsilon_d + \epsilon_{1,\infty} + \epsilon_{\infty,1})\sqrt{N} = o(1)$. Then, as* *$N \to \infty$, we have the following convergence in distribution,*

$$\sqrt{N}\left(\widehat{W}^S(r,s) - W^S(r,s)\right) \xrightarrow{d} \sup_{f \in \mathcal{F}'(r,s)} f^\top Z_{rs}$$

with $\mathcal{F}'(r,s)$ defined in Eq. (31).

3.4 Application to W^S estimation in topic models

Topic modelling is a popular statistical tool for discovering hidden thematic structure in a variety of applications such as text mining, natural language processing, social science and so on [9]. In topic models, the bag-of-word representation is widely used, which assumes the access of observed frequencies of p words (in a pre-specified vocabulary) over a collection of n documents. For each document $i \in [n]$, let $\widehat{r}^{(i)} \in \Delta_p$ denote the observed frequencies of all words and further let N_i be the length of this document. Topic models assume that $N_i\widehat{r}^{(i)} \sim$ Multinomial$_p(N_i; r^{(i)})$, and each $r^{(i)} = A\alpha^{(i)} \in \mathcal{S}$ is a mixture discrete distribution with mixture components $\mathcal{A} = \{A_1, \ldots, A_K\}$ corresponding to K topic distributions. Each topic distribution $A_k \in \Delta_p$ is itself a discrete distribution of p words for the kth topic. The mixture weight $\alpha^{(i)} \in \Delta_K$ further represents proportions of K topics in document i.

One main usage of topic models is to quantify (dis)similarity between documents based on thematic structures. Such similarity, or distance, between documents find applications in various downstream analyses such as document clustering, document retrieval and document classification. Although there have been many useful metrics proposed for quantifying similarity between topic distributions, distances between documents that are able to capture thematic dissimilarity are relatively scarce. Our proposed Sketched Wassertein Distance, W^S, allows us to lift any metric defined at the level of topic distributions to a corresponding new distance at the document level. Estimation and inference of such new distance can be done by our developed approach and theory. We refer to Section 5 for concrete examples in a real data analysis.

In the rest of this section, we provide explicit rates of $\epsilon_{\infty,1}$ and $\epsilon_{1,\infty}$ for estimating A under topic models, hence offer a complete analysis of both estimation and inference of W^S. For ease of presentation, we consider $N \asymp N_i$ for all $i \in [n]$ and fixed K throughout this section.

Since the seminal paper [10], estimation of the topic distributions (A) has been studied extensively from the Bayesian perspective (see, [9], for a comprehensive review of this class of techniques). A recent list of papers [2, 3, 4, 7, 8, 31, 33] have proposed provably fast algorithms of estimating A from the frequentist point of view under the following anchor word assumption,

Assumption 3. For each $k \in [K]$, there exists at least one $j \in [p]$ such that $A_{jk} > 0$ and $A_{jk'} = 0$ for all $k' \neq k$.

Assumption 3 is also served for identifiability of A coupled with other regularity conditions on the topic proportions. Empirically, [17] has shown that Assumption 3 holds in a wide range of data sets in topic models. We notice that Assumption 3 implies Assumption 1.

Under topic models and Assumption 3, [7, 31] have established minimax lower bounds of $\epsilon_{1,\infty} = \|\widehat{A} - A\|_{1,\infty}$ over the parameter space

$$\Theta_A := \{A : A_k \in \Delta_p, \forall k \in [K], \ A \text{ satisfies Assumption 3}, \ \|A_{j\cdot}\|_1 \geq c/p, \forall j \in [p]\}.$$

with $c \in (0,1)$ being some absolute constant. Using the fact that $\sum_j \|A_{j\cdot}\|_1 = K$, the lower bounds of $\epsilon_{1,\infty}$ further imply a lower bound of $\epsilon_{\infty,1} = \max_j(\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_1 / \|A_{j\cdot}\|_1)$ for any $A \in \Theta_A$. Concretely, for some absolute constants $c_0 > 0$ and $c_1 \in (0,1)$, we have

$$\inf_{\widehat{A}} \sup_{A \in \Theta_A} \mathbb{P}\left\{\epsilon_{1,\infty} \geq c_0 \sqrt{\frac{p}{nN}}\right\} \geq c_1, \quad \inf_{\widehat{A}} \sup_{A \in \Theta_A} \mathbb{P}\left\{\epsilon_{\infty,1} \geq c_0 \sqrt{\frac{p}{nN}}\right\} \geq c_1.$$

On the other hand, [7] proposed a computationally efficient estimator of A that achieves the lower bounds of both $\epsilon_{1,\infty}$ [7, Corollary 8] and $\epsilon_{\infty,1}$ [6, Theorem J.1] up to a logarithmic factor of $L := n \vee p \vee N$ under suitable conditions (see, for instance, [6, Appendix J.1]).

Consequently, the following corollary provides one concrete instance for which conclusions in Theorem 7 hold for the proposed W^S estimator based on $\widehat{A}$ in [7].

Corollary 5. *Grant topic model assumptions and Assumption 3 as well as conditions listed in [6, Appendix J.1]. For any $r^{(i)}, r^{(j)} \in \mathcal{S}$ with $i, j \in [n]$, suppose $p = o(n)$ and*

$$\max\left\{\lambda_K^{-1}(M), \lambda_K^{-1}(V_{r^{(i)}}), \lambda_K^{-1}(V_{r^{(j)}})\right\} = \mathcal{O}(1). \tag{32}$$

For any d satisfying Eq. (23) in W^S, as $N \to \infty$,

$$\sqrt{N}\left(\widehat{W}^S(r^{(i)}, r^{(j)}) - W^S(r^{(i)}, r^{(j)})\right) \xrightarrow{d} \sup_{f \in \mathcal{F}'(r^{(i)}, r^{(j)})} f^\top Z_{r^{(i)} r^{(j)}}. \tag{33}$$

Here $\mathcal{F}'(r^{(i)}, r^{(j)})$ and $Z_{r^{(i)} r^{(j)}}$ are defined in Eq. (31) and Eq. (28), respectively, with $r^{(i)}$ and $r^{(j)}$ in lieu of r and s.

Under regularity conditions in (32), condition (26) in Theorem 6 as well as condition (29) in Theorem 7 translate to $p = o(n)$, a condition that typically holds in many applications of topic models.

4 Simulations

We conduct simulation studies in this section to support our theoretical findings. In Section 4.1, we verify the asymptotic normality of our proposed estimator of the mixture weight. Speed of convergence in distribution of our W^S estimator is evaluated in Section 4.2 while estimation of the limiting distribution of the W^S estimator is discussed in Section 4.3.

We use the following data generating mechanism. The mixture weights are generated uniformly from Δ_K (equivalent to the symmetric Dirichlet distribution with parameter equal to 1). For the mixture components in A, we first generate its entries as i.i.d. samples from $\text{Unif}(0,1)$ and then normalize each column to the unit sum. Samples of $r = A\alpha$ are generated according to the multinomial sampling scheme in Eq. (12). We take the distance d in W^S to be the total variation distance, that is, $d(A_k, A_\ell) = \frac{1}{2}\|A_k - A_\ell\|_1$ for each $k, \ell \in [K]$. To simplify presentation, we consider both A and d known throughout the simulations.

4.1 Asymptotic normality of the proposed estimator of the mixture weight

We verify the asymptotic normality of our proposed weighted least squares estimator (WLS) in (13) and compare its speed of convergence to a normal limit with two other estimators mentioned in Section 3.1.2: the weighted restrictive least squares estimator (WRLS) in Eq. (22) and the maximum likelihoood estimator (MLE).

Since A is known, it suffices to consider one mixture distribution $r = A\alpha \in \mathcal{S}$. We fix $K = 5$ and $p = 1000$ and vary $N \in \{25, 50, 100, 300, 500\}$. For each setting, we obtain estimates of α

for each estimator based on 500 repeatedly generated multinomial samples. Figure 2 depicts the QQ-plots (quantiles of the estimates after centering and standardizing versus quantiles of the standard normal distribution) of all three estimators for the 4th coordinate of α. It is clear that WLS converges to a normal limit much faster than WRLS and MLE, corroborating our discussion in Section 3.3.1.
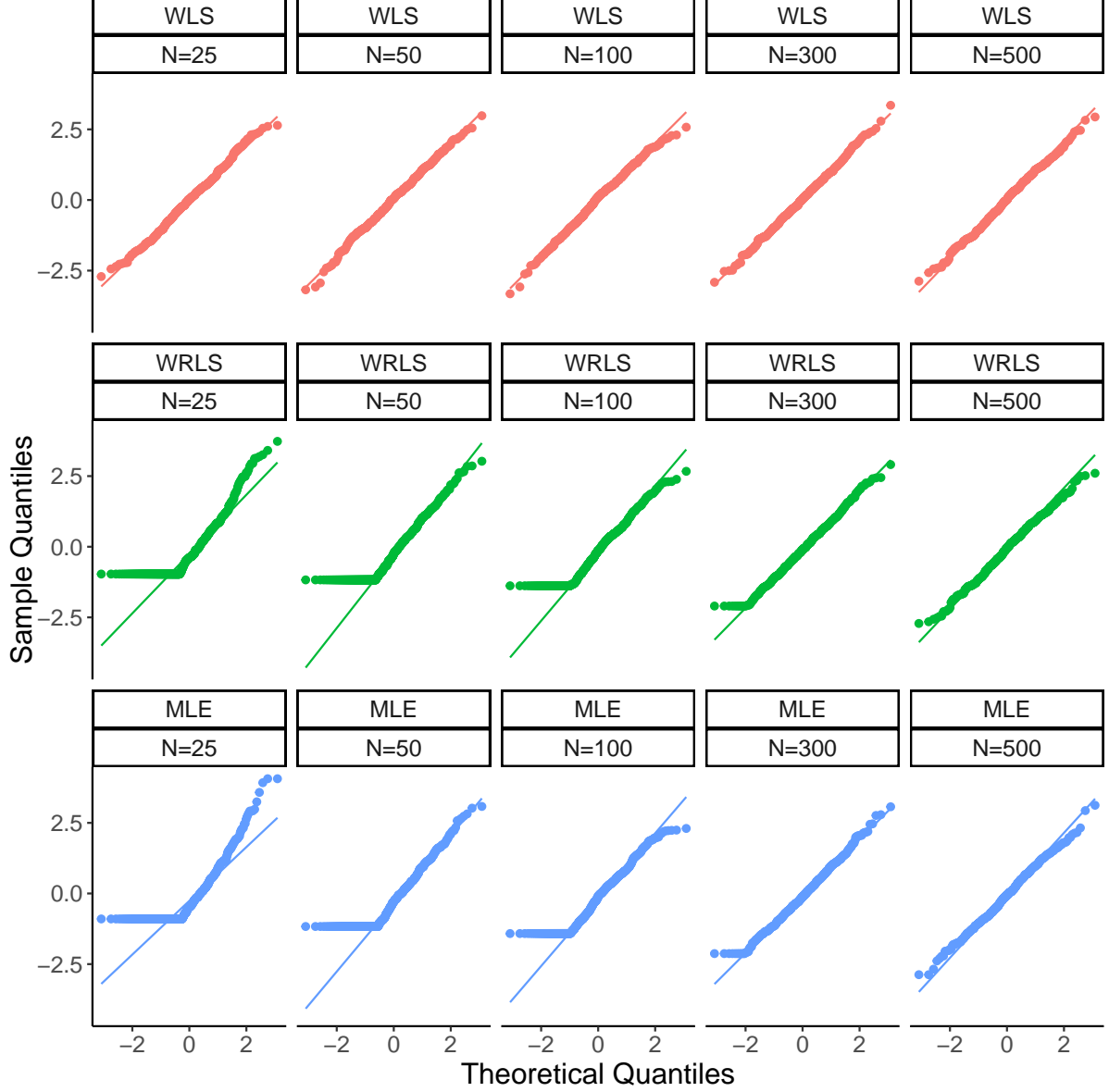

Figure 2: QQ-plots of the estimates of $\alpha_4 \approx 0.18$ for WLS, WRLS and MLE

4.2 Speed of convergence in distribution of the W^S estimator

We focus on the null case, $r = s$, to show how the speed of convergence of our W^S estimator in Theorem 7 depends on N, p and K. Specifically, we consider (i) $K = 10$, $p = 300$ and $N \in \{10, 100, 1000\}$, (ii) $N = 100$, $p = 300$ and $K \in \{5, 10, 20\}$ and (iii) $N = 100$, $K = 10$ and $p \in \{50, 100, 300, 500\}$. For each combination of K, p and N, we first generate 10 mixture distributions $r^{(1)}, \ldots, r^{(10)} \in \mathcal{S}$. Then for each of $r^{(i)}$ with $i \in \{1, 2, \ldots, 10\}$, to evaluate the

speed of convergence of our estimator, we compute the Kolmogorov-Smirnov (KS) distance, as well as the p-value of the two-sample KS test, between our estimates $\sqrt{N}\,\widehat{W}^S(r^{(i)}, r^{(i)})$ and its limiting distribution. Because the latter is not given in closed form, we mimic the theoretical c.d.f. by a c.d.f. based on $20{,}000$ draws from it. Finally, both the averaged KS distances and p-values over $i \in \{1, 2, \dots, 10\}$ are shown in Table 1. We can see that the speed of convergence of our W^S estimator gets faster as N increases or K decreases, but seems unaffected by the ambient dimension p (we note here that the sampling variations across different settings of varying p are larger than those of varying N and K). In addition, our W^S estimator already seems to converge in distribution to the established limit as soon as $N \geq K$.

Table 1: The averaged KS distances and p-values of the two sample KS test between our estimates and samples of the limiting distribution

| | $K = 10, p = 300$ | | | $N = 100, p = 300$ | | | $N = 100, K = 10$ | | | |
	$N = 10$	$N = 100$	$N = 1000$	$K = 5$	$K = 10$	$K = 20$	$p = 50$	$p = 100$	$p = 300$	$p = 500$
KS distances ($\times 10^{-2}$)	1.02	0.94	0.87	0.77	0.92	0.97	0.95	0.87	1.01	0.92
P-values of KS test	0.34	0.39	0.50	0.58	0.46	0.36	0.37	0.47	0.29	0.48

4.3 Estimation of the limiting distribution of the W^S estimator

To fully utilize Theorem 7 for inference of W^S, one needs to estimate the limiting distribution of Eq. (33). Bootstrap [19] is a powerful tool for this purpose. However, since the Hadamard-directional derivative of W^S is non-linear, [18] shows that the classical bootstrap fails to consistently estimate the limiting distribution in Eq. (33). Nevertheless, the same paper shows that a version of m-out-of-N bootstrap, for $m/N \to 0$ and $m \to \infty$, can be used instead. In our simulation, we follow [47] by setting $m = N^\gamma$ for some $\gamma \in (0, 1)$. To evaluate its performance, consider the null case, $r = s$, and choose $K = 10$, $p = 500$, $N \in \{100, 500, 1000, 3000\}$ and $\gamma \in \{0.1, 0.3, 0.5, 0.7\}$. The number of repetitions of Bootstrap is set to 500 while for each setting we repeat 200 times. The KS distance is used to evaluate the closeness between Bootstrap samples and the limiting distribution. Again, we approximate the theoretical c.d.f. of the limiting distribution based on 20,000 draws from it. Figure 3 shows the KS distances for various choices of N and γ. As we we can see, the result of m-out-of-N bootstrap gets better as N increases. Furthermore, it suggests $\gamma = 0.3$ which gives the best performance over all choices of N.

5 Real data analysis

In this section we illustrate our developed theory on a real data set [44] that uses the bag-of-word representation of 5811 articles published in NeurIPS from 1987 to 2015. After removing stop words, rare words that occur in less than 200 articles and common words that appear in more than 80% articles, the vocabulary size is $p = 3734$. After further removing short documents that in total have less than 150 words in the remaining vocabulary, we obtain $n = 5799$ articles with the averaged document length equal to 1493.

We apply the TOP algorithm in [7] that we have discussed in Section 3.4 and find $\widehat{K} = 26$ mixture components as well as the estimated topic distributions $\widehat{A}$. Since each topic distribution A_k represents a conditional probability of p words given topic k, we use its most frequent words for interpretation. Table 2, for instance, lists the top 10 most frequent words of some selected topics.

To proceed, we take some popular distances between topic distributions to demonstrate that our proposed W^S can lift any such distance at the topic level to a meaningful new distance at
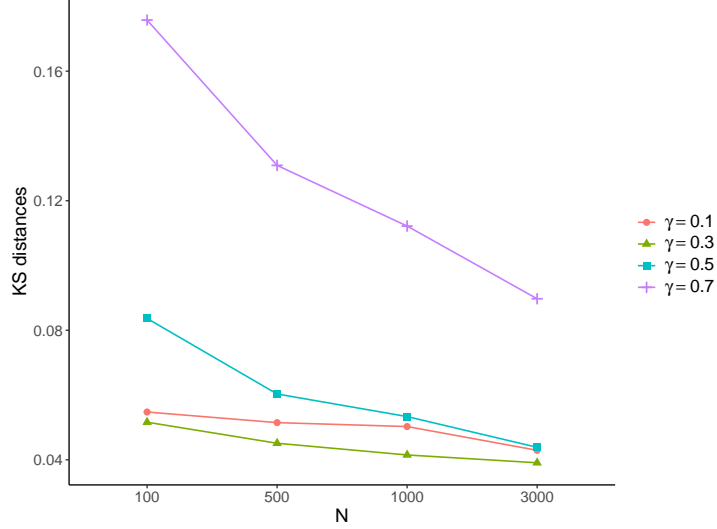
Figure 3: The averaged KS distances of m-out-of-N bootstrap with $m = N^\gamma$

the document level. Our experiments indicate that the resulting distances successfully capture thematic similarity between documents. We consider the following choices of d at the topic level.

(a) A natural choice of d is the total variation distance (TV) defined as $d(A_k, A_\ell) = \frac{1}{2}\|A_k - A_\ell\|_1$ for each $k, \ell \in [K]$. The W^S based on this choice is also considered in [6] and is found to successfully yield thematically meaningful pairwise document distances.

(b) Similarly, one can also consider the ℓ_2 norm distance (L2) between columns of A, defined as, $d(A_k, A_\ell) = \frac{1}{2}\|A_k - A_\ell\|_2$.

(c) Another popular choice of d is the 1-Wasserstein distance $d(A_k, A_\ell) = W(A_k, A_\ell; D^w)$ where $D^w \in \mathbb{R}_+^{p \times p}$ is a distance matrix based on a pre-trained word embedding (see, for instance, [64]).

(d) All three distances mentioned above use all words to compute dissimilarity between two topics. In many applications, it is more appealing to measure topic similarity by using a few frequent words, for both computation and interpretability considerations. The Jaccard distance based on the overlap between the most frequent words in two topics, for instance, is commonly used in practice to measure similarity among topics [27, 28]. Let $S_{T,k}$ denote the set of the top T most frequent words in topic k, that is, the set of coordinates of A_k with the T largest values. The Jaccard distance between topic k and topic ℓ is defined as

$$d(A_k, A_\ell) = J(A_k, A_\ell; T) = 1 - \frac{|S_{T,k} \cap S_{T,\ell}|}{|S_{T,k} \cup S_{T,\ell}|}.$$

(e) The Average Jaccard distance (AJ) is an improved version of the Jaccard distance which takes into account the order of the selected words [23, 49]. It is defined as

$$d(A_k, A_\ell) = AJ(A_k, A_\ell; T) = \frac{1}{T}\sum_{t=1}^{T} J(A_k, A_\ell; t).$$

22

Table 2: The top 10 most frequent words (from top to bottom) and interpretation of 8 selected topics

Topic 2	Topic 3	Topic 7	Topic 19	Topic 22	Topic 23	Topic 24	Topic 25
boosting	causal	estimator	regret	tensor	topic	tree	supervised
error	causes	estimators	rounds	rank	document	trees	labeled
loss	models	estimation	round	matrix	documents	node	training
training	logical	distribution	algorithms	norm	topics	nodes	semi
bound	variables	mse	bound	decomposition	lda	models	unlabeled
algorithms	graph	sample	online	latent	models	root	classification
margin	distribution	error	setting	order	words	structure	class
weak	structure	density	loss	low	word	graph	performance
examples	experiments	estimate	let	decompositions	distribution	probability	kernel
theorem	cause	probability	price	mode	latent	search	network
let	noise	log	arm	matrices	dirichlet	pruning	error
classifiers	variable	theorem	best	analysis	inference	distribution	graph
distribution	linear	matrix	strategy	vectors	variational	algorithms	input
bounds	inference	variance	game	method	blei	variables	methods
class	nodes	let	theorem	error	sampling	training	features
log	effect	gaussian	lemma	convex	probability	parent	method
regret	observed	entropy	optimal	vector	parameters	leaf	label
classifier	probability	random	value	theorem	log	decision	examples
generalization	may	kernel	section	completion	prior	inference	matrix
probability	noisy	optimal	fixed	models	distributions	let	feature

We take all 402 papers published in year 2015 and compute their pairwise W^S based on each choice of d from (a), (b) and (e)[1] via the weighted restricted least squares estimators of the mixture. We employ the restricted version of the least squares estimators in light of Lemma 2, which shows that they can have better performance when the mixture weights are sparse, as is common in topic models. We also consider using different numbers of frequent words in the Average Jaccard distance. Specifically, we consider AJ_T with $T \in \{5, 10, 25, 50\}$.

We find that these choices of d all successfully capture thematic similarity between documents. For illustration purpose, we focus on 11 articles, as listed in Table 3, which are related with at most two topics. This is done via thresholding the estimated mixture weights at level 0.01. Appendix C contains the estimated W^S between all pairs of the selected documents by choosing d from TV, AJ_5 and AJ_25. We see qualitatively similar results for all choices of d. For further interpretation, we collect the top 3 most similar and dissimilar pairwise $\widehat{W}^S$ for each topic distance in Table 4. All three choices of d yield the same three pairs of documents that are similar. Looking at document 1 and document 5, they are both related with *density / distribution estimation* due to their association with Topic 7 in Table 2. Regarding document 2 and document 10, although their titles seem irrelevant, they both cover topic 19, which is related with *online learning and regret optimization*, suggesting their distance to be close. Moving to the pair of document 7 and document 9, the similarity between them is already seen from their titles. Indeed we find that both of them are related with topic 22 which can be interpreted as *matrix / tensor decomposition*. On the other hand, it is clear to see from the titles of documents in Table 3 that the top 3 largest pairwise distances listed in Table 4 for all choices of d are as expected. However, we notice that using the Average Jaccard distance gives rise to an estimator

---

[1]Since it is shown in [6] that the W^S of using (c) gives qualitatively similar result as that based on (a), we omit its result here. Similarly, we only consider the Average Jaccard distance due to its better performance than the Jaccard distance [23, 49].

Table 3: The selected papers and their associated topics

Paper ID	Paper names	Topics
Doc 0	Online Gradient Boosting	2, 19
Doc 1	Competitive Distribution Estimation: Why is Good-Turing Good	7, 19
Doc 2	Online Learning with Adversarial Delays	2, 19
Doc 3	Semi-supervised Convolutional Neural Networks for Text Categorization via Region Embedding	23, 25
Doc 4	Efficient Non-greedy Optimization of Decision Trees	2, 24
Doc 5	Nonparametric von Mises Estimators for Entropies, Divergences and Mutual Informations	7
Doc 6	Logarithmic Time Online Multiclass prediction	2, 24
Doc 7	Fast and Guaranteed Tensor Decomposition via Sketching	22, 23
Doc 8	Local Causal Discovery of Direct Causes and Effects	3, 24
Doc 9	Sparse and Low-Rank Tensor Decomposition	22
Doc 10	Revenue Optimization against Strategic Buyers	19

$\widehat{W}^S$ that better discriminates between these dissimilar documents. Nevertheless, thematic similarity between documents seem to be preserved by W^S for all aforementioned topic distances, providing a justification of their popularity.

Table 4: $\widehat{W}^S$ based on TV, AJ_5 and AJ_25 of the top 3 most similar and dissimilar pairs of documents

	(Doc 1, Doc 5)	(Doc 2, Doc 10)	(Doc 7, Doc 9)	(Doc 7, Doc 10)	(Doc 2, Doc 8)	(Doc 8, Doc 10)
TV	.002	.008	.035	.509	.513	.513

	(Doc 1, Doc 5)	(Doc 2, Doc 10)	(Doc 7, Doc 9)	(Doc 8, Doc 9)	(Doc 5, Doc 8)	(Doc 8, Doc 10)
AJ_5	.004	.021	.069	1	1	1

	(Doc 1, Doc 5)	(Doc 2, Doc 10)	(Doc 7, Doc 9)	(Doc 8, Doc 10)	(Doc 2, Doc 3)	(Doc 3, Doc 10)
AJ_25	.003	.018	.067	.993	.997	0.999

A Proofs

A.1 Proof of Theorem 1

Proof. We will prove

$$
\begin{aligned}
W^S(r, s) &= \inf_{\alpha, \beta \in \Delta_K : r = A\alpha, s = A\beta} \inf_{\gamma \in \Gamma(\alpha, \beta)} \sum_{\ell, k \in [K]} \gamma_{k\ell} d(A_k, A_\ell) \\
&= \inf_{\alpha, \beta \in \Delta_K : r = A\alpha, s = A\beta} W(\alpha, \beta; d)
\end{aligned}
\tag{34}
$$

under Assumption 2. In particular, Theorem 1 follows immediately under Assumption 1.

To prove (34), let us denote the right side of (34) by $D(r, s)$ and write $W(\alpha, \beta) := W(\alpha, \beta; d)$ for simplicity. We first show that D satisfies **R1** and **R2**. To show convexity, fix $r, r', s, s' \in \mathcal{S}$, and let $\alpha, \alpha', \beta, \beta' \in \Delta_K$ satisfy

$$
\begin{aligned}
r &= A\alpha \\
r' &= A\alpha' \\
s &= A\beta \\
s' &= A\beta'.
\end{aligned}
\tag{35}
$$

Then for any $\lambda \in [0, 1]$, it is clear that $\lambda r + (1 - \lambda)r' = A(\lambda \alpha + (1 - \lambda)\alpha')$, and similarly for s. Therefore

$$
\begin{aligned}
D(\lambda r + (1 - \lambda)r', \lambda s + (1 - \lambda)s') &\leq W(\lambda \alpha + (1 - \lambda)\alpha', \lambda \beta + (1 - \lambda)\beta') \\
&\leq \lambda W(\alpha, \beta) + (1 - \lambda)W(\alpha', \beta'),
\end{aligned}
$$

where the second inequality uses the convexity of the Wasserstein distance [55, Theorem 4.8].

Taking infima on both sides over $\alpha, \alpha', \beta, \beta' \in \Delta_K$ satisfying Eq. (35) shows that $D(\lambda r + (1 - \lambda)r', \lambda s + (1 - \lambda)s') \leq \lambda D(r, s) + (1 - \lambda)D(r', s')$, which establishes that D' satisfies **R1**. To show that it satisfies **R2**, we note that Assumption 2 implies that if $A_k = A\alpha$, then we must have $\alpha = \delta_{A_k}$. Therefore

$$
D(A_k, A_\ell) = \inf_{\alpha, \beta \in \Delta_K : A_k = A\alpha, A_\ell = A\beta} W(\alpha, \beta) = W(\delta_{A_k}, \delta_{A_\ell}),
\tag{36}
$$

Recalling that the Wasserstein distance is defined as an infimum over couplings between the two marginal measures and using the fact that the only coupling between δ_{A_k} and δ_{A_ℓ} is $\delta_{A_k} \times \delta_{A_\ell}$, we obtain

$$
W(\delta_{A_k}, \delta_{A_\ell}) = \int d(A, A') \, \mathrm{d}(\delta_{A_k} \times \delta_{A_\ell})(A, A') = d(A_k, A_\ell),
$$

showing that D also satisfies **R2**. As a consequence, by Definition (1), we obtain in particular that $W^S(r, s) \geq D(r, s)$ for all $r, s \in \mathcal{S}$.

To show equality, it therefore suffices to show that $W^S(r, s) \leq D(r, s)$ for all $r, s \in \mathcal{S}$. To do so, fix $r, s \in \mathcal{S}$ and let α, β satisfy $r = A\alpha$ and $s = A\beta$. Let $\gamma \in \Gamma(\alpha, \beta)$ be an arbitrary coupling between α and β, which we identify with an element of $\Delta_{K \times K}$. The definition of $\Gamma(\alpha, \beta)$ implies that, $\sum_{\ell=1}^{K} \gamma_{k\ell} = \alpha_k$ for all $k \in [K]$, and since $r = A\alpha$, we obtain $r = \sum_{\ell, k \in [K]} \gamma_{k\ell} A_k$. Similarly, $s = \sum_{\ell, k \in [K]} \gamma_{k\ell} A_\ell$. The convexity of W^S (**R1**) therefore implies

$$
W^S(r, s) = W^S\left(\sum_{\ell, k \in [K]} \gamma_{k\ell} A_k, \sum_{\ell, k \in [K]} \gamma_{k\ell} A_\ell\right) \leq \sum_{\ell, k \in [K]} \gamma_{k\ell} W^S(A_k, A_\ell).
$$

Since W^S agrees with d on the original mixture components (**R2**), we further obtain

$$W^S(r, s) \leq \sum_{\ell, k \in [K]} \gamma_{k\ell}\, d(A_k, A_\ell)\,.$$

Finally, we may take infima over all α, β satisfying $r = A\alpha$ and $s = A\beta$ and $\gamma \in \Gamma(\alpha, \beta)$ to obtain

$$
\begin{aligned}
W^S(r, s) &\leq \inf_{\alpha, \beta \in \Delta_K : r = A\alpha, s = A\beta} \inf_{\gamma \in \Gamma(\alpha, \beta)} \sum_{\ell, k \in [K]} \gamma_{k\ell} d(A_k, A_\ell) \\
&= \inf_{\alpha, \beta \in \Delta_K : r = A\alpha, s = A\beta} W(\alpha, \beta) \\
&= D(r, s)\,.
\end{aligned}
$$

Therefore $W^S(r, s) \leq D(r, s)$, establishing that $W^S(r, s) = D(r, s)$, as claimed. $\square$

A.2 Proof of Corollary 1

Proof. Under Assumption 1, the map $\tau : \alpha \mapsto r = A\alpha$ is a bijection between Δ_K and $\mathcal{S}$. Since $(\mathcal{X}, d)$ is a (finite) metric space, (Δ_K, W) is a complete metric space [55, Theorem 6.18], and Theorem 1 establishes that the map from (Δ_K, W) to $(\mathcal{S}, W^S)$ induced by τ is an isomorphism of metric spaces. In particular, $(\mathcal{S}, W^S)$ is a complete metric space, as desired. $\square$

A.2.1 A counterexample to Corollary 1 under Assumption 2

Let $A_1, \ldots, A_4$ be probability measures on the set $\{1, 2, 3, 4\}$, represented as elements of Δ_4 as

$$
\begin{aligned}
A_1 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0, 0) \\
A_2 &= (0, 0, \tfrac{1}{2}, \tfrac{1}{2}) \\
A_3 &= (\tfrac{1}{2}, 0, \tfrac{1}{2}, 0) \\
A_4 &= (0, \tfrac{1}{2}, 0, \tfrac{1}{2})\,.
\end{aligned}
$$

We equip $\mathcal{X} = \{A_1, A_2, A_3, A_4\}$ with any metric d satisfying the following relations:

$$
\begin{aligned}
d(A_1, A_2), d(A_3, A_4) &< 1 \\
d(A_1, A_3) &= 1\,.
\end{aligned}
$$

These distributions satisfy Assumption 2 but not Assumption 1, since $A_1 \in \text{span}(A_2, A_3, A_4)$. Let $r = A_1$, $t = A_3$, and $s = \frac{1}{2}(A_1 + A_2) = \frac{1}{2}(A_3 + A_4)$. Then $W^S(r, t) = d(A_1, A_3) = 1$, but $W^S(r, s) + W^S(s, t) = \frac{1}{2}d(A_1, A_2) + \frac{1}{2}d(A_3, A_4) < 1$. Therefore, Assumption 2 alone is not strong enough to guarantee that W^S as defined in Eq. (6) is a metric.

A.3 Proof of Theorem 2

Proof. Recall Eq. (7) and Eq. (8). By adding and subtracting terms, we have

$$|\bar{W}^S(r, s) - W^S(r, s)| \leq \mathrm{I} + \mathrm{II}$$

where

$$\mathrm{I} := \left| \sup_{f \in \bar{\mathcal{F}}} f^\top (\bar{\alpha} - \bar{\beta}) - \sup_{f \in \mathcal{F}} f^\top (\bar{\alpha} - \bar{\beta}) \right|,$$

$$\mathrm{II} := \left| \sup_{f \in \mathcal{F}} f^\top (\bar{\alpha} - \bar{\beta}) - \sup_{f \in \mathcal{F}} f^\top (\alpha - \beta) \right|.$$

To bound the term I, Lemma 3, stated and proved in Appendix A.3.1, shows that

$$\left| \sup_{f \in \bar{\mathcal{F}}} f^\top (\bar{\alpha} - \bar{\beta}) - \sup_{f \in \mathcal{F}} f^\top (\bar{\alpha} - \bar{\beta}) \right| \le d_H(\mathcal{F}, \bar{\mathcal{F}}) \|\bar{\alpha} - \bar{\beta}\|_1$$

where, for two subsets A and B of $\mathbb{R}^K$, $d_H(A, B)$ is the Hausdorff distance defined as

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \|a - b\|_\infty, \ \sup_{b \in B} \inf_{a \in A} \|a - b\|_\infty \right\}. \tag{37}$$

Together with Lemma 4 which provides an upper bound of $d_H(\mathcal{F}, \bar{\mathcal{F}})$, we obtain

$$\mathrm{I} \le \max_{k, k' \in [K]} \left| d(A_k, A_{k'}) - \bar{d}(k, k') \right| \|\bar{\alpha} - \bar{\beta}\|_1 \le \epsilon_d \left(2 + \epsilon_\alpha + \epsilon_\beta \right), \tag{38}$$

where the second inequality is due to Eq. (10), Eq. (11), triangle inequality and $\|\alpha - \beta\|_1 \le 2$. We also assume the identity label permutation for Eq. (11).

To bound from above II, by triangle inequality and adding and subtracting terms, we have

$$\mathrm{II} \ \le \sup_{f \in \mathcal{F}} f^\top (\bar{\alpha} - \alpha) + \sup_{f \in \mathcal{F}} f^\top (\bar{\beta} - \beta).$$

Observing that $f \in \mathcal{F}$ implies

$$\|f\|_\infty \le \max_{k \in [K]} d(A_1, A_k) \le \mathrm{diam}(\mathcal{X}),$$

we have

$$\begin{aligned} \sup_{f \in \mathcal{F}} f^\top (\bar{\alpha} - \alpha) &\le \sup_{f : \|f\|_\infty \le \mathrm{diam}(\mathcal{X})} f^\top (\bar{\alpha} - \alpha) \\ &\le \mathrm{diam}(\mathcal{X}) \sup_{f : \|f\|_\infty \le 1} f^\top (\bar{\alpha} - \alpha) \\ &= \mathrm{diam}(\mathcal{X}) \ \|\bar{\alpha} - \alpha\|_1 \\ &\le \mathrm{diam}(\mathcal{X}) \ \epsilon_\alpha, \end{aligned}$$

implying that

$$\mathrm{II} \le \mathrm{diam}(\mathcal{X}) \ (\epsilon_\alpha + \epsilon_\beta).$$

Combining with Eq. (38) concludes

$$|\bar{W}^S(r, s) - W^S(r, s)| \le 2\epsilon_d + (\epsilon_d + \mathrm{diam}(\mathcal{X}))(\epsilon_\alpha + \epsilon_\beta),$$

completing the proof. $\qquad \square$

A.3.1 Lemmata used in the proof of Theorem 2

We first state the following lemmata that are used in the proof of Theorem 2.

Lemma 3. *For any $\alpha, \beta \in \mathbb{R}^K$ and any subsets $A, B \subseteq \mathbb{R}^K$,*

$$\left| \sup_{f \in A} f^\top(\alpha - \beta) - \sup_{f \in B} f^\top(\alpha - \beta) \right| \le d_H(A, B)\|\alpha - \beta\|_1.$$

Proof. Since the result trivially holds if either A or B is empty, we consider the case that A and B are non-empty sets. Pick any $\alpha, \beta \in \mathbb{R}^K$. Fix $\varepsilon > 0$, and let $f^* \in A$ be such that $f^{*\top}(\alpha - \beta) \ge \sup_{f \in A} f^\top(\alpha - \beta) - \varepsilon$. By definition of $d_H(A, B)$, there exists $\widehat{f} \in B$ such that $\|\widehat{f} - f^*\|_\infty \le d_H(A, B) + \varepsilon$. Then

$$
\begin{aligned}
\sup_{f \in A} f^\top(\alpha - \beta) - \sup_{f \in B} f^\top(\alpha - \beta) &\le f^{*\top}(\alpha - \beta) - \sup_{f \in B} f^\top(\alpha - \beta) + \varepsilon \\
&\le (f^* - \widehat{f})^\top(\alpha - \beta) + \varepsilon \\
&\le d_H(A, B)\|\alpha - \beta\|_1 + \varepsilon(1 + \|\alpha - \beta\|_1).
\end{aligned}
$$

Since $\varepsilon > 0$ was arbitrary, we obtain

$$\sup_{f \in A} f^\top(\alpha - \beta) - \sup_{f \in B} f^\top(\alpha - \beta) \le d_H(A, B)\|\alpha - \beta\|_1.$$

Analogous arguments also yield

$$\sup_{f \in B} f^\top(\alpha - \beta) - \sup_{f \in A} f^\top(\alpha - \beta) \le d_H(A, B)\|\alpha - \beta\|_1$$

completing the proof. $\qquad\square$

Lemma 4. *Under conditions in Theorem 2,*

$$d_H(\mathcal{F}, \bar{\mathcal{F}}) \le \max_{k,k' \in [K]} \left| d(A_k, A_{k'}) - \bar{d}(k, k') \right|$$

holds up to a label permutation.

Proof. For two subsets A and B of $\mathbb{R}^K$, we recall the Hausdorff distance

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} \|a - b\|_\infty, \ \sup_{b \in B} \inf_{a \in A} \|a - b\|_\infty \right\}.$$

Without loss of generality, we assume the identity label permutation for Eq. (11). For notational simplicity, we write

$$E_{kk'} = \left| d(A_k, A_{k'}) - \bar{d}(k, k') \right|, \qquad \forall \, k, k' \in [K]. \tag{39}$$

We first prove

$$\sup_{f \in \mathcal{F}} \inf_{\widehat{f} \in \bar{\mathcal{F}}} \|\widehat{f} - f\|_\infty \le \max_{k,k' \in [K]} \left| d(A_k, A_{k'}) - \bar{d}(k, k') \right| = \max_{k,k' \in [K]} E_{kk'}. \tag{40}$$

For any $f \in \mathcal{F}$, we know that $f_1 = 0$ and $f_k - f_\ell \le d(A_k, A_\ell)$ for all $k, \ell \in [K]$. Define

$$\widehat{f}_k = \min_{\ell \in [K]} \left\{ \bar{d}(k, \ell) + f_\ell + E_{\ell 1} \right\}, \qquad \forall \, k \in [K]. \tag{41}$$

It suffices to show $\widehat{f} \in \bar{\mathcal{F}}$ and

$$\|\widehat{f} - f\|_\infty \le \max_{k,k' \in [K]} E_{kk'}. \tag{42}$$

Clearly, for any $k, k' \in [K]$, the definition in Eq. (41) yields

$$
\begin{aligned}
\widehat{f}_k - \widehat{f}_{k'} &= \min_{\ell \in [K]} \left\{ \bar{d}(k,\ell) + f_\ell + E_{\ell 1} \right\} - \min_{\ell' \in [K]} \left\{ \bar{d}(k',\ell') + f_{\ell'} + E_{\ell' 1} \right\} \\
&= \min_{\ell \in [K]} \left\{ \bar{d}(k,\ell) + f_\ell + E_{\ell 1} \right\} - \left\{ \bar{d}(k',\ell^*) + f_{\ell^*} + E_{\ell^* 1} \right\} \qquad \text{for some } \ell^* \in [K] \\
&\le \bar{d}(k,\ell^*) - \bar{d}(k,\ell^*) \\
&\le \bar{d}(k,k').
\end{aligned}
$$

Also notice that

$$
\begin{aligned}
\widehat{f}_1 &= \min \left\{ \min_{\ell \in [K]\setminus\{1\}} \left\{ \bar{d}(1,\ell) + f_\ell + E_{\ell 1} \right\}, \ \bar{d}(1,1) + f_1 + E_{11} \right\} \\
&= \min \left\{ \min_{\ell \in [K]\setminus\{1\}} \left\{ \bar{d}(1,\ell) + f_\ell + E_{\ell 1} \right\}, \ 0 \right\},
\end{aligned}
$$

by using $f_1 = 0$ and $E_{11} = 0$. Since, for any $\ell \in [K] \setminus \{1\}$, by the definition of $\mathcal{F}$,

$$\bar{d}(1,\ell) + f_\ell + E_{\ell 1} \ge \bar{d}(1,\ell) - d(A_1, A_\ell) + E_{\ell 1} \ge 0,$$

we conclude $\widehat{f}_1 = 0$, hence $\widehat{f} \in \bar{\mathcal{F}}$.

Finally, to show Eq. (42), we have $f_1 = \widehat{f}_1 = 0$ and for any $k \in [K] \setminus \{1\}$,

$$\widehat{f}_k - f_k = \min_{\ell \in [K]} \left\{ \bar{d}(k,\ell) + f_\ell + E_{\ell 1} - f_k \right\} \le E_{k1}$$

and, conversely,

$$
\begin{aligned}
\widehat{f}_k &- f_k \\
&= \min_{\ell \in [K]} \{ \bar{d}(k,\ell) + f_\ell - f_k + E_{\ell 1} \} \\
&\ge \min_{\ell \in [K]} \{ \bar{d}(k,\ell) - d(A_k, A_\ell) + E_{\ell 1} \} \\
&\ge \min_{\ell \in [K]} \{ -E_{k\ell} + E_{\ell 1} \}.
\end{aligned}
$$

The last two displays imply

$$
\begin{aligned}
\|\widehat{f} - f\|_\infty &\le \max_{k \in [K]} \max \left\{ E_{k1}, \max_{\ell \in [K]} \{ E_{k\ell} - E_{\ell 1} \} \right\} \\
l &= \max_{k,\ell \in [K]} \{ E_{k\ell} - E_{\ell 1} \} \\
&\le \max_{k,\ell} E_{k\ell},
\end{aligned}
$$

completing the proof of Eq. (40).

By analogous arguments, we also have

$$\max_{\widehat{f} \in \bar{\mathcal{F}}} \min_{f \in \mathcal{F}} \|\widehat{f} - f\|_\infty \le \max_{k,k' \in [K]} E_{kk'},$$

completing the proof. $\qquad \square$

A.4 Proof of Lemma 1

Recall that $M = A^\top D^{-1} A$ with $D = \text{diag}(\|A_{1\cdot}\|_1, \ldots, \|A_{p\cdot}\|_1)$. Claim (1) follows immediately from Assumption 1 and $\min_{j \in [p]} \|A_{j\cdot}\|_1 > 0$. To show (2), note that $M \mathbb{1}_K = \mathbb{1}_K$ follows from the definition. Also, $\lambda_1(M) \leq \|M\|_{\infty,1} = 1$ and

$$\lambda_1(M) \geq \frac{1}{K} \mathbb{1}_K^\top M \mathbb{1}_K = 1$$

together imply $\lambda_1(M) = 1$. By definition,

$$\lambda_K(M) = \inf_{\|v\|_2=1} \sum_{j=1}^p \frac{(A_{j\cdot}^\top v)^2}{\|A_{j\cdot}\|_1} \geq \frac{\lambda_K(A^\top A)}{\|A\|_{\infty,1}}$$

proving claim (3). Finally, the last claim follows from the fact $\lambda_1(M^{-1}) \leq \|M^{-1}\|_{\infty,1} \leq \sqrt{K} \lambda_1(M^{-1})$ and $\lambda_1(M^{-1}) = \lambda_K^{-1}(M)$.

A.5 Proof of Theorem 3 & Corollary 2

Proof of Theorem 3. Without loss of generality, we assume the identity label permutation for Eq. (17). Let

$$A^+ = M^{-1} A^\top D^{-1}, \qquad \widehat{A}^+ = \widehat{M}^{-1} \widehat{A}^\top \widehat{D}^{-1}, \tag{43}$$

with $M = A^\top D^{-1} A$ and $\widehat{M} = \widehat{A}^\top \widehat{D}^{-1} \widehat{A}$. By definition,

$$\|\widehat{\alpha} - \alpha\|_1 = \|\widehat{A}^+ \widehat{r} - A^+ r\|_1 \leq \|\widehat{A}^+ (\widehat{r} - r)\|_1 + \|(\widehat{A}^+ - A^+) r\|_1. \tag{44}$$

For the second term on the right hand side, using $r = A\alpha$ gives

$$
\begin{aligned}
\|(\widehat{A}^+ - A^+) r\|_1 &= \|(\widehat{A}^+ A - \mathbf{I}_K)\alpha\|_1 \\
&\leq \|\widehat{A}^+ (A - \widehat{A})\|_{1,\infty} \|\alpha\|_1 && \text{by } \widehat{A}^+ \widehat{A} = \mathbf{I}_K \\
&\leq \|\widehat{A}^+\|_{1,\infty} \|\widehat{A} - A\|_{1,\infty} && \text{by } \|\alpha\|_1 = 1 \\
&\leq \|\widehat{A}^+\|_{1,\infty} \, \epsilon_{1,\infty}.
\end{aligned}
$$

Since

$$\|\widehat{A}^+\|_{1,\infty} = \max_{j \in [p]} \frac{\|(\widehat{M})^{-1} \widehat{A}_{j\cdot}\|_1}{\|\widehat{A}_{j\cdot}\|_1} \leq \|(\widehat{M})^{-1}\|_{1,\infty} \leq 2\|M^{-1}\|_{\infty,1}$$

where Lemma 6 is invoked to obtain the last step, we have

$$\|(\widehat{A}^+ - A^+) r\|_1 \leq 2\|M^{-1}\|_{\infty,1} \epsilon_{1,\infty}. \tag{45}$$

Regarding $\|\widehat{A}^+ (\widehat{r} - r)\|_1$, adding and subtracting terms gives

$$\|\widehat{A}^+ (\widehat{r} - r)\|_1 \leq \|A^+ (\widehat{r} - r)\|_1 + \|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1.$$

Invoke Lemma 5 to obtain that, with probability at least $1 - t$,

$$\|A^+ (\widehat{r} - r)\|_1 \leq \max_{j \in [p]} \|(A^+)_j\|_2 \left(\sqrt{\frac{2K \log(2K/t)}{N}} + \frac{4K \log(2K/t)}{N} \right). \tag{46}$$

Since

$$\max_{j \in [p]} \|(A^+)_j\|_2 = \max_{j \in [p]} \frac{\|(M)^{-1} A_{j\cdot}\|_2}{\|A_{j\cdot}\|_1} \leq \|M^{-1}\|_{\text{op}} = \frac{1}{\lambda_K(M)},$$

together with Eq. (20), the following holds with probability at least $1 - t$

$$\|A^+(\widehat{r} - r)\|_1 \leq \frac{2}{\lambda_K(M)} \sqrt{\frac{2K \log(2K/t)}{N}}. \tag{47}$$

It remains to bound from above $\|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1$. To this end, notice that

$$
\begin{aligned}
\widehat{A}^+ - A^+ &= \widehat{M}^{-1}\widehat{A}^\top(\widehat{D}^{-1} - D^{-1}) + \widehat{M}^{-1}(\widehat{A} - A)^\top D^{-1} \\
&\quad + (\widehat{M}^{-1} - M^{-1})A^\top D^{-1} \\
&= \widehat{M}^{-1}\widehat{A}^\top \widehat{D}^{-1}(D - \widehat{D})D^{-1} + \widehat{M}^{-1}(\widehat{A} - A)^\top D^{-1} \\
&\quad + \widehat{M}^{-1}(M - \widehat{M})M^{-1}A^\top D^{-1}.
\end{aligned} \tag{48}
$$

We obtain

$$
\begin{aligned}
&\|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 \\
&\leq \|\widehat{M}^{-1}\|_{1,\infty}\|\widehat{A}^\top \widehat{D}^{-1}\|_{1,\infty}\|(\widehat{D} - D)D^{-1}\|_{1,\infty}\|\widehat{r} - r\|_1 \\
&\quad + \|\widehat{M}^{-1}\|_{1,\infty}\|(\widehat{A} - A)^\top D^{-1}\|_{1,\infty}\|\widehat{r} - r\|_1 \\
&\quad + \|\widehat{M}^{-1}(M - \widehat{M})\|_{1,\infty}\|A^+(\widehat{r} - r)\|_1 \\
&\leq 2\|\widehat{M}^{-1}\|_{1,\infty}\epsilon_{\infty,1}\|\widehat{r} - r\|_1 + \|\widehat{M}^{-1}(M - \widehat{M})\|_{1,\infty}\|A^+(\widehat{r} - r)\|_1 \\
&\leq 4\|M^{-1}\|_{1,\infty}\epsilon_{\infty,1}\|\widehat{r} - r\|_1 + 2\|M^{-1}\|_{\infty,1}(2\epsilon_{\infty,1} + \epsilon_{1,\infty})\|A^+(\widehat{r} - r)\|_1
\end{aligned} \tag{49}
$$

where in the last line we invoke Lemma 6. Plugging in Eq. (47) further gives

$$
\begin{aligned}
\|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 &\leq 4\|M^{-1}\|_{1,\infty}\epsilon_{\infty,1}\|\widehat{r} - r\|_1 \\
&\quad + \frac{4\|M^{-1}\|_{\infty,1}}{\lambda_K(M)}(2\epsilon_{\infty,1} + \epsilon_{1,\infty})\sqrt{\frac{2K \log(2K/t)}{N}}.
\end{aligned} \tag{50}
$$

Combining Eq. (50), Eq. (45) and Eq. (47) yields

$$\|\widehat{\alpha} - \alpha\|_1 \leq \frac{4}{\lambda_K(M)}\sqrt{\frac{2K \log(2K/t)}{N}} + 2\|M^{-1}\|_{\infty,1}\left(\epsilon_{1,\infty} + 2\epsilon_{\infty,1}\|\widehat{r} - r\|_1\right)$$

with probability $1 - t$. The proof is complete. $\qquad\square$

Proof of Corollary 2. It suffices to prove part (a) as part (b) follows immediately by using $\epsilon_{1,\infty} = \epsilon_{\infty,1} = 0$.

The proof of part (a) follows from that of Theorem 3 except that we can use Lemma 5 to bound $\|\widehat{A}^+(\widehat{r} - r)\|_1$ in Eq. (44) directly. Specifically, since $\widehat{A}$ is independent of $\widehat{r}$ and so is $\widehat{A}^+$, invoke Lemma 5 to obtain that, for any $t \in (0, 1)$, with probability at least $1 - t$,

$$\|\widehat{A}^+(\widehat{r} - r)\|_1 \leq \max_{j \in [p]} \|(\widehat{A}^+)_j\|_2 \left(\sqrt{\frac{2K \log(2K/t)}{N}} + \frac{4K \log(2K/t)}{N}\right).$$

The result follows by noting that

$$\max_{j \in [p]} \|(\widehat{A}^+)_j\|_2 = \max_{j \in [p]} \frac{\|(\widehat{M})^{-1}\widehat{A}_{j\cdot}\|_2}{\|\widehat{A}_{j\cdot}\|_1} \leq \frac{1}{\lambda_K(\widehat{M})} \leq \frac{2}{\lambda_K(M)}$$

from Lemma 6. $\qquad\square$

A.5.1 Lemmata used in the proof of Theorem 3

Lemma 5. *For any fixed matrix $B \in \mathbb{R}^{K \times p}$, with probability at least $1 - t$ for any $t \in (0, 1)$, we have*

$$\|B(\widehat{r} - r)\|_1 \leq \max_{j \in [p]} \|B_j\|_2 \left(\sqrt{\frac{2K \log(2K/t)}{N}} + \frac{4K \log(2K/t)}{N} \right).$$

Proof. Start with $\|B(\widehat{r} - r)\|_1 = \sum_{k \in [K]} |B_{k\cdot}^\top (\widehat{r} - r)|$ and pick any $k \in [K]$. Note that

$$N\widehat{r} \sim \mathrm{Multinomial}_p(N; r).$$

By writing

$$\widehat{r} - r = \frac{1}{N} \sum_{t=1}^{N} (Z_t - r)$$

with $Z_t \sim \mathrm{Multinomial}_p(1; r)$, we have

$$B_{k\cdot}^\top (\widehat{r} - r) = \frac{1}{N} \sum_{t=1}^{N} B_{k\cdot}^\top (Z_t - r).$$

To apply the Bernstein inequality, recalling that $r \in \Delta_p$, we find that $B_{k\cdot}^\top (Z_t - r)$ has zero mean, $B_{k\cdot}^\top (Z_t - r) \leq 2\|B_{k\cdot}\|_\infty$ for all $t \in [N]$ and

$$\frac{1}{N} \sum_{t=1}^{N} \mathrm{Var}\left(B_{k\cdot}^\top (Z_t - r) \right) \leq B_{k\cdot}^\top \mathrm{diag}(r) B_{k\cdot}.$$

An application of the Bernstein inequality gives, for any $t \geq 0$,

$$\mathbb{P}\left\{ B_{k\cdot}^\top (\widehat{r} - r) \leq \sqrt{\frac{t\, B_{k\cdot}^\top \mathrm{diag}(r) B_{k\cdot}}{N}} + \frac{2t\|B_{k\cdot}\|_\infty}{N} \right\} \geq 1 - 2e^{-t/2}.$$

Choosing $t = 2 \log(2K/\epsilon)$ for any $\epsilon \in (0, 1)$ and taking the union bounds over $k \in [K]$ yields

$$\|B(\widehat{r} - r)\|_1 \leq \sqrt{\frac{2 \log(2K/\epsilon)}{N}} \sum_{k \in [K]} \sqrt{B_{k\cdot}^\top \mathrm{diag}(r) B_{k\cdot}} + \frac{4 \log(2K/\epsilon)}{N} \sum_{k \in [K]} \|B_{k\cdot}\|_\infty$$

with probability exceeding $1 - \epsilon$. The result thus follows by noting that

$$\sum_{k \in [K]} \sqrt{B_{k\cdot}^\top \mathrm{diag}(r) B_{k\cdot}} \leq \sqrt{K} \sqrt{\sum_{k \in [K]} \sum_{j \in [p]} r_j B_{kj}^2} \leq \sqrt{K} \max_{j \in [p]} \|B_j\|_2$$

and $\sum_{k \in [K]} \|B_{k\cdot}\|_\infty \leq K \max_{j \in [p]} \|B_j\|_2$. $\qquad\square$

Lemma 6. *Under Assumption 1 and condition (19), on the event that Eq. (16) and Eq. (17) hold, one has*

$$\lambda_K(\widehat{M}) \geq \lambda_K(M)/2, \qquad \|\widehat{M}^{-1}\|_{\infty,1} \leq 2\|M^{-1}\|_{\infty,1}$$

and

$$\max\left\{ \|(\widehat{M} - M)\widehat{M}^{-1}\|_{1,\infty},\ \|(\widehat{M} - M)\widehat{M}^{-1}\|_{\infty,1} \right\} \leq 2\|M^{-1}\|_{\infty,1}(2\epsilon_{\infty,1} + \epsilon_{1,\infty}).$$

Moreover, $\lambda_K(\widehat{M}) \geq \lambda_K(M)/2$ is guaranteed by Assumption 1 and $4\epsilon_{\infty,1} + 2\epsilon_{1,\infty} \leq \lambda_K(M)$.

Proof. We assume the identity label permutation for Eq. (16) and Eq. (17). First, by using $\|M\|_{\mathrm{op}} \leq \|M\|_{\infty,1}$ for any symmetric matrix M, we find

$$\|M^{-1}\|_{\infty,1} \geq \|M^{-1}\|_{\mathrm{op}} = \lambda_K^{-1}(M) \geq 1. \tag{51}$$

Then condition (19) guarantees, for any $j \in [p]$,

$$\|\widehat{A}_{j\cdot}\|_1 \geq \|A_{j\cdot}\|_1 \left(1 - \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_1}{\|A_{j\cdot}\|_1}\right) \geq \frac{1}{2}\|A_{j\cdot}\|_1. \tag{52}$$

To prove the first result, by Weyl's inequality, we have

$$\lambda_K(\widehat{M}) \geq \lambda_K(M) - \|\widehat{M} - M\|_{\mathrm{op}}.$$

Using $\|M\|_{\mathrm{op}} \leq \|M\|_{\infty,1}$ again yields

$$
\begin{aligned}
\|\widehat{M} - M\|_{\mathrm{op}} &\leq \|\widehat{M} - M\|_{\infty,1} \\
&\leq \|\widehat{A}^{\top}\widehat{D}^{-1}(\widehat{A} - A)\|_{\infty,1} + \|\widehat{A}^{\top}(\widehat{D}^{-1} - D^{-1})A\|_{\infty,1} \\
&\quad + \|(\widehat{A} - A)^{\top}D^{-1}A\|_{\infty,1} \\
&\overset{(i)}{\leq} \|\widehat{D}^{-1}(\widehat{A} - A)\|_{\infty,1} + \|\widehat{D}^{-1}(\widehat{D} - D)D^{-1}A\|_{\infty,1} \\
&\quad + \|(\widehat{A} - A)^{\top}\|_{\infty,1}\|D^{-1}A\|_{\infty,1} \\
&\overset{(ii)}{\leq} 2\max_{j \in [p]} \frac{\|\widehat{A}_{j\cdot} - A_{j\cdot}\|_1}{\|\widehat{A}_{j\cdot}\|_1} + \max_{k \in [K]} \|\widehat{A}_k - A_k\|_1,
\end{aligned} \tag{53}
$$

where the step (i) uses $\|\widehat{A}^{\top}\|_{\infty,1} = 1$ and the step (ii) is due to $\|D^{-1}A\|_{\infty,1} = \max_j \|A_{j\cdot}\|_1/\|A_{j\cdot}\|_1 = 1$. Invoke Eq. (17) to obtain

$$\|\widehat{M} - M\|_{\mathrm{op}} \leq 2\epsilon_{\infty,1} + \epsilon_{1,\infty} \overset{(19)}{\leq} \frac{1}{2\|M^{-1}\|_{\infty,1}}.$$

Together with (51), we conclude $\lambda_K(\widehat{M}) \geq \lambda_K(M)/2$. Moreover, we readily see that $\lambda_K(\widehat{M}) \geq \lambda_K(M)/2$ only requires $\lambda_K(M)/2 \geq 2\epsilon_{\infty,1} + \epsilon_{1,\infty}$.

To prove the second result, start with the decomposition

$$
\begin{aligned}
\widehat{M}^{-1} - M^{-1} &= M^{-1}(M - \widehat{M})\widehat{M}^{-1} \\
&= M^{-1}A^{\top}D^{-1}(A - \widehat{A}) + M^{-1}A^{\top}(D^{-1} - \widehat{D}^{-1})\widehat{A}\widehat{M}^{-1} \\
&\quad + M^{-1}(A - \widehat{A})^{\top}\widehat{D}^{-1}\widehat{A}\widehat{M}^{-1}.
\end{aligned} \tag{54}
$$

It then follows from dual-norm inequalities, $\|A^{\top}\|_{\infty,1} = 1$ and $\|\widehat{D}^{-1}\widehat{A}\|_{\infty,1} = 1$ that

$$
\begin{aligned}
\|\widehat{M}^{-1}\|_{\infty,1} &\leq \|M^{-1}\|_{\infty,1} + \|\widehat{M}^{-1} - M^{-1}\|_{\infty,1} \\
&\leq \|M^{-1}\|_{\infty,1} + \|M^{-1}\|_{\infty,1}\|D^{-1}(A - \widehat{A})\|_{\infty,1} \\
&\quad + \|M^{-1}\|_{\infty,1}\|(D^{-1} - \widehat{D}^{-1})\widehat{A}\|_{\infty,1}\|\widehat{M}^{-1}\|_{\infty,1} \\
&\quad + \|M^{-1}\|_{\infty,1}\|(A - \widehat{A})^{\top}\|_{\infty,1}\|\widehat{M}^{-1}\|_{\infty,1} \\
&\leq \|M^{-1}\|_{\infty,1} + \|M^{-1}\|_{\infty,1}\|D^{-1}(A - \widehat{A})\|_{\infty,1} \\
&\quad + \|M^{-1}\|_{\infty,1}\|D^{-1}(\widehat{D} - D)\|_{\infty,1}\|\widehat{M}^{-1}\|_{\infty,1} \\
&\quad + \|M^{-1}\|_{\infty,1}\|(A - \widehat{A})^{\top}\|_{\infty,1}\|\widehat{M}^{-1}\|_{\infty,1} \\
&\leq \|M^{-1}\|_{\infty,1} + \|M^{-1}\|_{\infty,1}\epsilon_{\infty,1} \\
&\quad + \|M^{-1}\|_{\infty,1}(\epsilon_{\infty,1} + \epsilon_{1,\infty})\|\widehat{M}^{-1}\|_{\infty,1}.
\end{aligned}
$$

Since, similar to Eq. (51), we also have $\|\widehat{M}^{-1}\|_{\infty,1} \geq 1$. Invoke condition Eq. (19) to conclude

$$\|\widehat{M}^{-1}\|_{\infty,1} \leq 2\|M^{-1}\|_{\infty,1},$$

completing the proof of the second result.

Finally, Eq. (53) and the second result immediately give

$$\|(\widehat{M} - M)\widehat{M}^{-1}\|_{\infty,1} = \|\widehat{M} - M\|_{\infty,1}\|\widehat{M}^{-1}\|_{\infty,1} \leq 2\|M^{-1}\|_{\infty,1}(2\epsilon_{\infty,1} + \epsilon_{1,\infty}).$$

The same bound also holds for $\|(\widehat{M} - M)\widehat{M}^{-1}\|_{1,\infty}$ by the symmetry of $\widehat{M} - M$ and $\widehat{M}^{-1}$. $\qquad\square$

A.6 Proof of Lemma 2

We will prove that with probability at least $1 - t$, the following holds

$$\|\widehat{\alpha}_{rls} - \alpha\|_1 \leq \min\left\{ 2, \frac{8\sqrt{2}}{\lambda_K(M)}\sqrt{(\|\alpha\|_0\|A^\top D^{-1}r\|_\infty + 1)\frac{\|\alpha\|_0 \log(2K/t)}{N}} \right. \tag{55}$$
$$\left. + \frac{8\sqrt{\|\alpha\|_0}(\epsilon_{1,\infty} + 2\epsilon_{\infty,1})}{\lambda_K(M)} \right\}.$$

In particular, when A is known, the above bound yields Lemma 2.

Proof of Eq. (55). Fix any $t > 0$ and we may assume $\|\alpha\|_0 \log(2K/t) \leq N$ and

$$2\epsilon_{1,\infty} + 4\epsilon_{\infty,1} \leq \lambda_K(M).$$

The latter together with Lemma 6 implies $\lambda_K(\widehat{M}) \geq \lambda_K(M)/2$ and $\widehat{D}$ is non-singular. It suffices to show that, with probability at least $1 - t$ for any $t > 0$,

$$\|\widehat{\alpha}_{rls} - \alpha\|_2 \leq \min\left\{ 2, \frac{4\sqrt{2}}{\lambda_K(M)}\sqrt{\frac{(\|\alpha\|_0\|A^\top D^{-1}r\|_\infty + 1)\log(2K/t)}{N}} + \frac{4(\epsilon_{1,\infty} + 2\epsilon_{\infty,1})}{\lambda_K(M)} \right\}. \tag{56}$$

Indeed, by the fact that $\alpha, \widehat{\alpha}_{rls} \in \Delta_K$, the basic arguments (see, for instance, [6]) yield

$$\|\widehat{\alpha}_{rls} - \alpha\|_1 \leq 2\sqrt{\|\alpha\|_0}\|\widehat{\alpha}_{rls} - \alpha\|_2$$

which in conjunction with Eq. (56) proves Eq. (55).

By definition in Eq. (22), the standard arguments give

$$\begin{aligned}
\|\widehat{D}^{-1/2}\widehat{A}(\widehat{\alpha}_{rls} - \alpha)\|_2^2 &\leq 2\left|(\widehat{\alpha}_{rls} - \alpha)^\top \widehat{A}^\top \widehat{D}^{-1}(\widehat{r} - \widehat{A}\alpha)\right| \\
&\leq 2\left|(\widehat{\alpha}_{rls} - \alpha)^\top A^\top D^{-1}(\widehat{r} - r)\right| + 2\left|(\widehat{\alpha}_{rls} - \alpha)^\top \widehat{A}^\top \widehat{D}^{-1}(\widehat{A} - A)\alpha\right| \\
&\quad + 2\left|(\widehat{\alpha}_{rls} - \alpha)^\top (\widehat{A}^\top \widehat{D}^{-1} - AD^{-1})(\widehat{r} - r)\right| \\
&:= \mathrm{I} + \mathrm{II} + \mathrm{III}.
\end{aligned}$$

We first bound each term on the right-hand side separately.

Start with

$$\mathrm{I} \leq 2\|\widehat{\alpha}_{rls} - \alpha\|_1\|A^\top D^{-1}(\widehat{r} - r)\|_\infty \leq 2\sqrt{\|\alpha\|_0}\|\widehat{\alpha}_{rls} - \alpha\|_2 \max_{k \in [K}|A_k^\top D^{-1}(\widehat{r} - r)|.$$

Pick any $k \in [K]$ and observe that

$$A_k^\top D^{-1}(\widehat{r} - r) = \frac{1}{N} \sum_{i=1}^N A_k^\top D^{-1}(Z_i - r) := \frac{1}{N} \sum_{i=1}^N Y_i$$

with $Z_i \sim \text{Multinomial}_p(1; r)$. Since, for all $i \in [n]$, $\mathbb{E}[Y_i] = 0$, $|Y_i| \le \|D^{-1}A\|_\infty \|Z_i - r\|_1 \le 2$ and

$$\mathbb{E}[Y_i^2] \le A_k^\top D^{-1} \text{diag}(r) D^{-1} A_k = \sum_{j=1}^p \frac{A_{jk}^2 r_j}{\|A_{j\cdot}\|_1^2} \le \|A^\top D^{-1} r\|_\infty,$$

an application of Bernstein inequality implies that, for any $t > 0$,

$$\mathbb{P}\left\{ |A_k^\top D^{-1}(\widehat{r} - r)| \le \sqrt{\frac{\|A^\top D^{-1} r\|_\infty t}{N}} + \frac{2t}{3N} \right\} \le 2e^{-t/2}.$$

Taking the union bounds over $k \in [K]$ gives that for any $t > 0$, with probability at least $1 - t$,

$$\text{I} \le 2\sqrt{\|\alpha\|_0} \|\widehat{\alpha}_{rls} - \alpha\|_2 \left(\sqrt{\frac{2\|A^\top D^{-1} r\|_\infty \log(2K/t)}{N}} + \frac{2\log(2K/t)}{N} \right)$$

$$\le 2\sqrt{2\|\alpha\|_0} \|\widehat{\alpha}_{rls} - \alpha\|_2 \sqrt{\left(\|A^\top D^{-1} r\|_\infty + \frac{1}{\|\alpha\|_0} \right) \frac{\log(2K/t)}{N}}. \tag{57}$$

To bound II, we find

$$\begin{aligned}
\text{II} &\le 2\|\widehat{\alpha}_{rls} - \alpha\|_\infty \|\widehat{A}^\top \widehat{D}^{-1}(\widehat{A} - A)\alpha\|_1 \\
&\le 2\|\widehat{\alpha}_{rls} - \alpha\|_2 \|\widehat{A}^\top \widehat{D}^{-1}\|_{1,\infty} \|(\widehat{A} - A)\alpha\|_1 \\
&\le 2\|\widehat{\alpha}_{rls} - \alpha\|_2 \, \epsilon_{1,\infty}.
\end{aligned} \tag{58}$$

Similarly,

$$\text{III} \le 2\|\widehat{\alpha}_{rls} - \alpha\|_2 \|(\widehat{A}^\top \widehat{D}^{-1} - A^\top D^{-1})(\widehat{r} - r)\|_1.$$

Since

$$\begin{aligned}
&\|(\widehat{A}^\top \widehat{D}^{-1} - A^\top D^{-1})(\widehat{r} - r)\|_1 \\
&\le \|\widehat{A}^\top \widehat{D}^{-1}(D - \widehat{D}) D^{-1}(\widehat{r} - r)\|_1 + \|(\widehat{A} - A)^\top D^{-1}(\widehat{r} - r)\|_1 \\
&\le \|(D - \widehat{D}) D^{-1}(\widehat{r} - r)\|_1 + \|(\widehat{A} - A)^\top D^{-1}(\widehat{r} - r)\|_1 \\
&\le 2\epsilon_{\infty,1} \|\widehat{r} - r\|_1
\end{aligned}$$

by using Eq. (17) in the last step, we obtain

$$\text{III} \le 4\|(\widehat{\alpha}_{rls} - \alpha)\|_2 \, \epsilon_{\infty,1}. \tag{59}$$

Finally, since

$$\|\widehat{D}^{-1/2} \widehat{A}(\widehat{\alpha}_{rls} - \alpha)\|_2^2 \ge \lambda_K(\widehat{M}) \|\widehat{\alpha}_{rls} - \alpha\|_2^2 \ge \frac{\lambda_K(M)}{2} \|\widehat{\alpha}_{rls} - \alpha\|_2^2,$$

collecting Eq. (57), Eq. (58) and Eq. (59) concludes the bound in (56), hence completes the proof of Lemma 2. $\qquad\square$

A.7 Proof of Theorem 5

Proof. In this proof, the symbols C and c denote universal positive constants whose value may change from line to line. Without loss of generality, we may assume by rescaling the metric that $\operatorname{diam}(\mathcal{X}) = 1$. The lower bound $\operatorname{diam}(\mathcal{X}) N^{-1/2}$ follows directly from Le Cam's method [see 41, proof of Theorem 4]. We therefore focus on proving the lower bound $\kappa_{\mathcal{X}}^{-2} \sqrt{K/(N(\log K)^2)}$, and may assume that $K/(\log K)^2 \geq C \kappa_{\mathcal{X}}^4$. In proving the lower bound, we may also assume that β is fixed to be the uniform distribution over $\mathcal{X}$, which we denote by ρ, and that the statistician obtains N i.i.d. observations from the unknown distribution α alone. We write

$$R_N := \inf_{\widehat{W}} \sup_{\alpha \in \Delta_K} \mathbb{E}_\alpha |\widehat{W} - W(\alpha, \rho; d)|$$

for the corresponding minimax risk.

We employ the method of "fuzzy hypotheses" [53], and, following a standard reduction [see, e.g 29, 60], we will derive a lower bound on the minimax risk by considering a modified observation model with Poisson observations. Concretely, in the original observation model, the empirical frequency count vector $Z := N\widehat{\alpha}$ is a sufficient statistic, with distribution

$$Z \sim \operatorname{Multinomial}_K(N; \alpha). \tag{60}$$

We define an alternate observation model under which Z has independent entries, with $Z_k \sim \operatorname{Poisson}(N\alpha_k)$, which we abbreviate as $Z \sim \operatorname{Poisson}(N\alpha)$. We write $\mathbb{E}'_\alpha$ for expectations with respect to this probability measure. Note that, in contrast to the multinomial model, the Poisson model is well defined for any $\alpha \in \mathbb{R}_+^K$.

Write $\Delta'_K = \{\alpha \in \mathbb{R}_+^K : \|\alpha\|_1 \geq 3/4\}$, and define

$$\tilde{R}_N := \inf_{\widehat{W}} \sup_{\alpha \in \Delta'_K} \mathbb{E}'_\alpha |\widehat{W} - W(\alpha/\|\alpha\|_1, \rho; d)|$$

The following lemma shows that R_N may be controlled by $\tilde{R}_{2N}$.

Lemma 7. *For all $N \geq 1$,*

$$\tilde{R}_{2N} \leq R_N + \exp(-N/12).$$

Proof. Fix $\delta > 0$, and for each $n \geq 1$ denote by $\widehat{W}_n$ a near-minimax estimator for the multinomial sampling model:

$$\sup_{\alpha \in \Delta_K} \mathbb{E}_\alpha |\widehat{W}_n - W(\alpha, \rho; d)| \leq R_n + \delta.$$

Since Z is a sufficient statistic, we may assume without loss of generality that $\widehat{W}_n$ is a function of the empirical counts Z.

We define an estimator for the Poisson sampling model by setting $\widehat{W}'(Z) = \widehat{W}_{N'}(Z)$, where $N' := \sum_{k=1}^K Z_k$. If $Z \sim \operatorname{Poisson}(2N\alpha)$ for $\alpha \in \mathbb{R}_+^K$, then, conditioned on $N' = n'$, the random variable Z has distribution $\operatorname{Multinomial}_K(n'; \alpha/\|\alpha\|_1)$. For any $\alpha \in \Delta'_K$, we then have

$$\mathbb{E}'_\alpha |\widehat{W}' - W(\alpha/\|\alpha\|_1, \rho; d)| = \sum_{n' \geq 0} \mathbb{E}'_\alpha \left[|\widehat{W}' - W(\alpha/\|\alpha\|_1, \rho; d)| \Big| N' = n' \right] \mathbb{P}_\alpha[N' = n']$$

$$= \sum_{n' \geq 0} \mathbb{E}'_\alpha \left[|\widehat{W}_{n'} - W(\alpha/\|\alpha\|_1, \rho; d)| \Big| N' = n' \right] \mathbb{P}_\alpha[N' = n']$$

$$= \sum_{n' \geq 0} \mathbb{E}_{\alpha/\|\alpha\|_1} \left[|\widehat{W}_{n'} - W(\alpha/\|\alpha\|_1, \rho; d)| \right] \mathbb{P}_\alpha[N' = n']$$

$$\leq \left(\sum_{n' \geq 0} R_{n'} \mathbb{P}_\alpha[N' = n'] \right) + \delta.$$

36

Since $N' \sim \mathrm{Poisson}(2N\|\alpha\|_1)$, and $R_{n'}$ is a non-increasing function of n' and satisfies $R_{n'} \leq 1$, standard tail bounds for the Poisson distribution show that if $\|\alpha\|_1 \geq 3/4$, then for some universal positive constant C,

$$\mathbb{E}'_\alpha |\widehat{W}' - W(\alpha/\|\alpha\|_1, \rho; d)| \leq P_\alpha[N' < N] + \sum_{n' \geq N} R_{n'} \mathbb{P}_\alpha[N' = n'] + \delta \leq e^{-N/12} + R_N + \delta \,.$$

Since $\alpha \in \Delta'_K$ and $\delta > 0$ were arbitrary, taking the supremum over α and infimum over all estimators $\widehat{W}_N$ yields the claim. $\qquad \square$

It therefore suffices to prove a lower bound on $\tilde{R}_N$. Fix an $\epsilon \in (0, 1/4)$, $\varkappa \geq 1$ and a positive integer L to be specified. We employ the following proposition.

Proposition 1. *There exists a universal positive constant c_0 such that for any $\varkappa \geq 1$ and positive integer L, there exists a pair of mean-zero random variables X and Y on $[-1, 1]$ satisfying the following properties:*

- $\mathbb{E}X^\ell = \mathbb{E}Y^\ell$ *for* $\ell = 1, \ldots, 2L - 2$
- $\mathbb{E}|X| \geq \varkappa \mathbb{E}|Y| \geq c_0 L^{-1} \varkappa^{-1}$

Proof. A proof appears in Appendix A.8 $\qquad \square$

We define two priors μ_0 and μ_1 on $\mathbb{R}_+^K$ by letting

$$\mu_1 = \mathrm{Law}\left(\frac{1}{K} + \frac{\epsilon}{K} X_1, \ldots, \frac{1}{K} + \frac{\epsilon}{K} X_K \right)$$
$$\mu_0 = \mathrm{Law}\left(\frac{1}{K} + \frac{\epsilon}{K} Y_1, \ldots, \frac{1}{K} + \frac{\epsilon}{K} Y_K \right) \,,$$

where $X_1, \ldots, X_K$ are i.i.d. copies of X and $Y_1, \ldots, Y_K$ are i.i.d. copies of Y. Since $\epsilon < 1/4$ and $X_k, Y_k \geq -1$ almost surely, μ_1 and μ_0 are supported on Δ'_K.

Following [53], we then define "posterior" measures

$$\mathbb{P}_j = \int \mathbb{P}_\alpha \mu_j(d\alpha), \qquad j = 0, 1 \,,$$

where $\mathbb{P}_\alpha = \mathrm{Poisson}(N\alpha)$ is the distribution of the Poisson observations with parameter α.

The following lemma shows that μ_0 and μ_1 are well separated with respect to the values of the functional $\alpha \mapsto W(\alpha/\|\alpha\|_1, \rho; d)$.

Lemma 8. *Assume that $\varkappa \geq 7\kappa_\mathcal{X}$. Then there exists $r \in \mathbb{R}_+$. such that*

$$\mu_0(\alpha : W(\alpha/\|\alpha\|_1, \rho; d) \leq r) \geq 1 - \delta$$
$$\mu_1(\alpha : W(\alpha/\|\alpha\|_1, \rho; d) \geq r + 2s) \geq 1 - \delta$$

where $s = \frac{1}{2} \epsilon c_0 L^{-1} \varkappa^{-2}$ and $\delta = 2e^{-Kc_0^2/L^2\varkappa^4}$

Proof. We begin by recalling the following comparison inequality between the Wasserstein distance and the total variation distance: for any two probability measures $\nu, \nu' \in \Delta_K$,

$$\kappa_\mathcal{X}^{-1} \|\nu - \nu'\|_1 \leq 2W(\nu, \nu'; d) \leq \|\nu - \nu'\|_1 \,. \tag{61}$$

Note that

$$\int \|\alpha - \rho\|_1 \mu_0(d\alpha) = \mathbb{E} \sum_{k=1}^{K} \left| \frac{1}{K} + \frac{\epsilon}{K} Y_k - \frac{1}{K} \right| = \epsilon \mathbb{E}|Y|, \tag{62}$$

and by Hoeffding's inequality,

$$\mu_0(\alpha : \|\alpha - \rho\|_1 \geq \epsilon \mathbb{E}|Y| + t) \leq e^{-Kt^2/2\epsilon^2}. \tag{63}$$

Analogously,

$$\mu_1(\alpha : \|\alpha - \rho\|_1 \leq \epsilon \mathbb{E}|X| - t) \leq e^{-Kt^2/2\epsilon^2}. \tag{64}$$

Under either distribution, Hoeffding's inequality also yields that $\big|\|\alpha\|_1 - 1\big| \geq t$ with probability at most $e^{-Kt^2/2\epsilon^2}$. Take $t = \epsilon c_0 L^{-1} \varkappa^{-2}$. Then letting $\delta = 2e^{-Kc_0^2/L^2\varkappa^4}$, we have with μ_0 probability at least $1 - \delta$ that

$$\begin{aligned}
2W(\alpha/\|\alpha\|_1, \rho; d) &\leq \left\| \frac{\alpha}{\|\alpha\|_1} - \rho \right\|_1 \\
&\leq \left\| \alpha - \frac{\alpha}{\|\alpha\|_1} \right\|_1 + \|\alpha - \rho\|_1 \\
&= \big|\|\alpha\|_1 - 1\big| + \|\alpha - \rho\|_1 \\
&\leq \epsilon \mathbb{E}|Y| + 2\epsilon c_0 L^{-1} \varkappa^{-1}.
\end{aligned} \tag{65}$$

And, analogously, with μ_1 probability at least $1 - \delta$,

$$\begin{aligned}
2W(\alpha/\|\alpha\|_1, \rho; d) &\geq \kappa_{\mathcal{X}}^{-1} \left\| \frac{\alpha}{\|\alpha\|_1} - \rho \right\|_1 \\
&\geq \kappa_{\mathcal{X}}^{-1} \|\alpha - \rho\|_1 - \left\| \alpha - \frac{\alpha}{\|\alpha\|_1} \right\|_1 \\
&\geq \epsilon \kappa_{\mathcal{X}}^{-1} \mathbb{E}|X| - 2\epsilon c_0 L^{-1} \varkappa^{-1} \\
&\geq \epsilon \varkappa \kappa_{\mathcal{X}}^{-1} \mathbb{E}|Y| - 2\epsilon c_0 L^{-1} \varkappa^{-2},
\end{aligned} \tag{66}$$

where we have used that $\kappa_{\mathcal{X}} \geq 1$ and $\mathbb{E}|X| \geq \varkappa \mathbb{E}|Y|$ by construction. Therefore, as long as $\varkappa \geq 7\kappa_{\mathcal{X}}$, we may take $r = \frac{1}{2} \epsilon \mathbb{E}|Y| + \epsilon c_0 L^{-1} \varkappa^{-2}$, in which case

$$\begin{aligned}
r + 2s &= \frac{1}{2} \epsilon \mathbb{E}|Y| + 2\epsilon c_0 L^{-1} \varkappa^{-2} \\
&\leq \frac{1}{2} \epsilon \varkappa \kappa_{\mathcal{X}}^{-1} \mathbb{E}|Y| - 3\epsilon \mathbb{E}|Y| + 2\epsilon c_0 L^{-1} \varkappa^{-2} \\
&\leq \frac{1}{2} \epsilon \varkappa \kappa_{\mathcal{X}}^{-1} \mathbb{E}|Y| - \epsilon c_0 L^{-1} \varkappa^{-2},
\end{aligned}$$

where the last inequality uses that $\mathbb{E}|Y| \geq c_0 L^{-1} \varkappa^{-2}$. With this choice of r and s, we have by Eq. (65) and Eq. (66) that

$$\begin{aligned}
\mu_0(\alpha : W(\alpha/\|\alpha\|_1, \rho; d) \leq r) &\geq 1 - \delta \\
\mu_1(\alpha : W(\alpha/\|\alpha\|_1, \rho; d) \geq r + 2s) &\geq 1 - \delta,
\end{aligned}$$

as claimed. $\square$

We next bound the statistical distance between $\mathbb{P}_0$ and $\mathbb{P}_1$.

Lemma 9. *If $\epsilon^2 \leq (L+1)K/(4e^2N)$, then*

$$d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1) \leq K2^{-L}.$$

Proof. By construction of the priors μ_0 and μ_1, we may write

$$\mathbb{P}_j = (\mathbb{Q}_j)^{\otimes K}, \quad j = 0, 1,$$

where Q_0 and Q_1 are the laws of random variables U_0 and U_1 defined by

$$V_0 \mid \lambda \sim \mathrm{Poisson}(\lambda_0), \qquad \lambda_0 \overset{d}{=} \frac{N}{K} + \frac{N\epsilon}{K}Y$$

$$V_1 \mid \lambda \sim \mathrm{Poisson}(\lambda_1), \qquad \lambda_1 \overset{d}{=} NK + \frac{N\epsilon}{K}X.$$

By [29, Lemma 32], if $L + 1 \geq (2eN\epsilon/K)^2/(N/K) = 4e^2\epsilon^2N/K$, then

$$d_{\mathrm{TV}}(\mathbb{Q}_0, \mathbb{Q}_1) \leq 2 \left(\frac{eN\epsilon}{K\sqrt{N(L+1)/K}} \right)^{L+1} \leq 2^{-L}.$$

Therefore, under this same condition,

$$d_{\mathrm{TV}}(\mathbb{P}_0, \mathbb{P}_1) = d_{\mathrm{TV}}(\mathbb{Q}_0^{\otimes K}, \mathbb{Q}_1^{\otimes K}) \leq K d_{\mathrm{TV}}(\mathbb{Q}_0, \mathbb{Q}_1) \leq K2^{-L}.$$

$\square$

Combining the above two lemmas with [53, Theorem 2.15(i)], we obtain that as long as $\varkappa \geq 7\kappa_\mathcal{X}$ and $\epsilon^2 \leq (L+1)K/(4e^2N)$, we have

$$\inf_{\widehat{W}} \sup_{\alpha \in \Delta'_K} \mathbb{E}'_\alpha \mathbb{1} \left\{ |\widehat{W} - W(\alpha/\|\alpha\|_1, \rho; d)| \geq \frac{1}{2}\epsilon c_0 L^{-1}\varkappa^{-2} \right\} \geq \frac{1}{2}(1 - K2^{-L}) - 2e^{-Kc_0^2/L^2\varkappa^4}.$$

Let $\varkappa = 7\kappa_\mathcal{X}$, $L = \lceil \log_2 K \rceil + 1$, and set $\epsilon = c\sqrt{K/N}$ for a sufficiently small positive constant c. By assumption, $K/(\log K)^2 \geq C\varkappa^4$, and by choosing C to be arbitrarily small we may make $2e^{-Kc_0^2/L^2\varkappa^4}$ arbitrarily small, so that the right side of the above inequality is strictly positive. We therefore obtain

$$\tilde{R}_N \geq C\varkappa^{-2}\sqrt{\frac{K}{N(\log K)^2}},$$

and applying Lemma 7 yields

$$R_N \geq C\varkappa^{-2}\sqrt{\frac{K}{N(\log K)^2}} - \exp(-N/12).$$

Since we have assumed that the first term is at least $N^{-1/2}$, the second term is negligible as long as N is larger than a universal constant. This proves the claim.

$\square$

A.8 Proof of Proposition 1

Proof. First, we note that it suffices to construct a pair of random variables X' and Y' on $[0, (c_1 L \varkappa)^2]$ satisfying $\mathbb{E}(X')^k = \mathbb{E}(Y')^k$ for $k = 1, \ldots, L-1$ and $\mathbb{E}\sqrt{X'} \geq \varkappa \mathbb{E}\sqrt{Y'} \geq c_2$ for some positive constants c_1 and c_2. Indeed, letting $X = (c_1 L \varkappa)^{-1} \varepsilon \sqrt{X'}$ and $Y = (c_1 L \varkappa)^{-1} \varepsilon \sqrt{Y'}$ where ε is a Rademacher random variable independent of X' and Y' yields a pair (X, Y) with the desired properties. We therefore focus on constructing such a X' and Y'.

By [60, Lemma 7] combined with [52, Section 2.11.1], there exists a universal positive constant c_1 such that we may construct two probability distributions μ_+ and μ_- on $[8\varkappa^2, (c_1 \varkappa L)^2]$ satisfying

$$\int x^k \, \mathrm{d}\mu_+(x) = \int x^k \, \mathrm{d}\mu_-(x) \quad k = 1, \ldots, L \tag{67}$$

$$\int x^{-1} \, \mathrm{d}\mu_+(x) = \int x^{-1} \, \mathrm{d}\mu_-(x) + \frac{1}{16\varkappa^2} \, . \tag{68}$$

We define a new pair of measures ν_+ and ν_- by

$$\nu_+(\mathrm{d}x) = \frac{1}{Z} \left(\frac{1}{x(x-1)} \mu_+(\mathrm{d}x) + \alpha_0 \delta_0(\mathrm{d}x) \right) \tag{69}$$

$$\nu_-(\mathrm{d}x) = \frac{1}{Z} \left(\frac{1}{x(x-1)} \mu_-(\mathrm{d}x) + \alpha_1 \delta_1(\mathrm{d}x) \right) , \tag{70}$$

where we let

$$Z = \int \frac{1}{x-1} \, \mathrm{d}\mu_+(x) - \int \frac{1}{x} \, \mathrm{d}\mu_-(x) \tag{71}$$

$$\alpha_0 = \int \frac{1}{x} \, \mathrm{d}\mu_+(x) - \int \frac{1}{x} \, \mathrm{d}\mu_-(x) = \frac{1}{32\varkappa^2} \tag{72}$$

$$\alpha_1 = \int \frac{1}{x-1} \, \mathrm{d}\mu_+(x) - \int \frac{1}{x-1} \, \mathrm{d}\mu_-(x) \, . \tag{73}$$

Since the support of μ_+ and μ_- lies in $[8\varkappa^2, (c_1 \varkappa L)^2]$, these quantities are well defined, and Lemma 10 shows that they are all positive. Finally, the definition of Z, α_0, and α_1 guarantees that ν_+ and ν_- have total mass one. Therefore, ν_+ and ν_- are probability measures on $[0, (c_1 \varkappa L)^2]$.

We now claim that

$$\int x^k \nu_+(\mathrm{d}x) = \int x^k \nu_-(\mathrm{d}x) \quad k = 1, \ldots, L \, . \tag{74}$$

By definition of ν_+ and ν_-, this claim is equivalent to

$$\int \frac{x^\ell}{x-1} \mu_+(\mathrm{d}x) = \int \frac{x^\ell}{x-1} \mu_-(\mathrm{d}x) + \alpha_1 \quad k = 0, \ldots, L-1 \, , \tag{75}$$

or, by using the definition of α_1,

$$\int \frac{x^\ell - 1}{x-1} \mu_+(\mathrm{d}x) = \int \frac{x^\ell - 1}{x-1} \mu_-(\mathrm{d}x) \quad k = 0, \ldots, L-1 \, . \tag{76}$$

But this equality holds due to the fact that $\frac{x^\ell - 1}{x-1}$ is a degree-$(\ell - 1)$ polynomial in x, and the first L moments of μ_+ and μ_- agree.

Finally, since $x \geq 1$ on the support of ν_-, we have

$$\int \sqrt{x}\, d\nu_-(x) \geq 1\,. \tag{77}$$

We also have

$$\int \sqrt{x}\, d\nu_+(x) = \frac{1}{Z} \int \frac{1}{\sqrt{x}(x-1)}\, d\mu_+(x)\,, \tag{78}$$

and by Lemma 11, this quantity is between $1/(8\varkappa)$ and $1/\varkappa$.

Letting $Y' \sim \nu_+$ and $X' \sim \nu_-$, we have verified that $X', Y' \in [0, (c_1 L \varkappa)^2]$ almost surely and $\mathbb{E}(X')^k = \mathbb{E}(Y')^k$ for $k = 1, \ldots, L$. Moreover, $\mathbb{E}\sqrt{X'} \geq 1 \geq \varkappa \mathbb{E}\sqrt{Y'} \geq 1/8$, and this establishes the claim. $\qquad \square$

A.8.1   Technical lemmata

Lemma 10. *If $\varkappa \geq 1$, then the quantities Z, α_0, α_1 are all positive, and $Z \in [\varkappa^{-2}/16, \varkappa^{-2}/8]$.*

Proof. First, $\alpha_0 = \frac{1}{16\kappa^2}$ by definition, and $Z \geq \alpha_0$ since $(x-1)^{-1} \geq x^{-1}$ on the support of μ_+. Moreover, for all $x \geq 8\varkappa^2 \geq 8$,

$$\left| \frac{1}{x-1} - \frac{1}{x} \right| = \frac{1}{x(x-1)} \leq \varkappa^{-2}/50\,. \tag{79}$$

Therefore $\alpha_1 \geq \alpha_0 - \varkappa^{-2}/50 > 0$ and $Z \leq \alpha_0 + \varkappa^{-2}/50 \leq \varkappa^{-2}/8$, as claimed. $\qquad \square$

Lemma 11. *If $\varkappa \geq 1$, then*

$$\frac{1}{Z} \int \frac{1}{\sqrt{x}(x-1)}\, d\mu_+(x) \in \left[\frac{1}{8\varkappa}, \frac{1}{\varkappa} \right] \tag{80}$$

Proof. For $x \geq 8\varkappa^2 \geq 8$,

$$\frac{1}{Z} \left| \frac{1}{\sqrt{x}(x-1)} - \frac{1}{x^{3/2}} \right| = \frac{1}{Z x^{3/2}(x-1)} \leq \frac{16\varkappa^2}{7(8\varkappa^2)^{3/2}} = \frac{1}{7\sqrt{2}\varkappa}\,. \tag{81}$$

Since

$$\frac{1}{Z} \int \frac{1}{x^{3/2}}\, d\mu_+(x) \leq \frac{1}{Z}(8\varkappa^2)^{-3/2} \leq \frac{1}{\sqrt{2}\varkappa}\,, \tag{82}$$

we obtain that

$$\frac{1}{Z} \int \frac{1}{\sqrt{x}(x-1)}\, d\mu_+(x) \leq \frac{1}{7\sqrt{2}\varkappa} + \frac{1}{\sqrt{2}\varkappa} = \varkappa^{-1}\,. \tag{83}$$

The lower bound bound follows from an application of Jensen's inequality:

$$\frac{1}{Z} \int \frac{1}{x^{3/2}}\, d\mu_+(x) \geq \frac{1}{Z} \left(\int x^{-1}\, d\mu_+(x) \right)^{3/2} \geq \frac{1}{Z}(16\varkappa^2)^{-3/2} \geq \frac{1}{8\varkappa}\,, \tag{84}$$

where we have used the fact that

$$\int x^{-1}\, d\mu_+(x) = \int x^{-1}\, d\mu_+(x) + \frac{1}{16\varkappa^2} \geq (16\varkappa^2)^{-1}\,. \tag{85}$$

$\qquad \square$

A.9 Proof of Theorem 6

Proof. By definition, we have

$$\widehat{\alpha} - \alpha = A^+(\widehat{r} - r) + (\widehat{A}^+ - A^+)(\widehat{r} - r) + (\widehat{A}^+ - A^+)r.$$

For notational convenience, let us write

$$Q_r = A^+ \Sigma_r A^{+\top} = M^{-1} V_r M^{-1}. \tag{86}$$

It suffices to show that, for any $u \in \mathbb{R}^K \setminus \{0\}$,

$$\sqrt{N} \frac{u^\top A^+ (\widehat{r} - r)}{\sqrt{u^\top Q_r u}} \xrightarrow{d} \mathcal{N}(0,1), \tag{87}$$

$$\sqrt{N} \frac{u^\top (\widehat{A}^+ - A^+)(\widehat{r} - r)}{\sqrt{u^\top Q_r u}} = o_{\mathbb{P}}(1), \tag{88}$$

$$\sqrt{N} \frac{u^\top (\widehat{A}^+ - A^+) r}{\sqrt{u^\top Q_r u}} = o_{\mathbb{P}}(1). \tag{89}$$

Define the $\ell_2 \to \ell_\infty$ condition number of V_r as

$$\kappa(V_r) = \inf_{u \neq 0} \frac{u^\top V_r u}{\|u\|_\infty^2}. \tag{90}$$

Notice that

$$\frac{\kappa(V_r)}{K} \leq \lambda_K(V_r) \leq \kappa(V_r). \tag{91}$$

A.9.1 Proof of Eq. (87)

By recognizing that the left-hand-side of Eq. (87) can be written as

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} \frac{u^\top A^+ (Z_t - r)}{\sqrt{u^\top Q_r u}} := \frac{1}{\sqrt{N}} \sum_{t=1}^{N} Y_t.$$

with Z_t, for $t \in [N]$, being i.i.d. Multinomial$_p(1; r)$. It is easy to see that $\mathbb{E}[Y_t] = 0$ and $\mathbb{E}[Y_t^2] = 1$. To apply the Lyapunov central limit theorem, it suffices to verify

$$\lim_{N \to \infty} \frac{\mathbb{E}[|Y_t|^3]}{\sqrt{N}} = 0.$$

This follows by noting that

$$\mathbb{E}[|Y_t|^3] \leq 2 \frac{\|u^\top A^+\|_\infty}{\sqrt{u^\top Q_r u}} \qquad \text{by } \mathbb{E}[Y_t^2] = 1, \|Z_t - r\|_1 \leq 2$$

$$\leq \frac{2\|u^\top A^\top D^{-1}\|_\infty}{\sqrt{u^\top V_r u}} \qquad \text{by Eq. (86)}$$

$$\leq \frac{2}{\sqrt{\kappa(V_r)}} \qquad \text{by Eq. (90)}$$

and by using $\lambda_K^{-1}(V_r) = o(N)$ and Eq. (91).

A.9.2 Proof of Eq. (88)

Fix any $u \in \mathbb{R}^K \setminus \{0\}$. Note that

$$
\frac{u^\top (\widehat{A}^+ - A^+)(\widehat{r} - r)}{\sqrt{u^\top Q_r u}} = \frac{u^\top M(\widehat{A}^+ - A^+)(\widehat{r} - r)}{\sqrt{u^\top V_r u}}
$$

$$
\leq \frac{\|u\|_\infty \|M(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1}{\|u\|_\infty \sqrt{\kappa(V_r)}} \qquad \text{by Eq. (90).}
$$

Further notice that $\|M(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1$ is smaller than

$$
\|M - \widehat{M}\|_{1,\infty} \|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 + \|\widehat{M}(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1.
$$

Using Eq. (49) in conjunction with Eq. (53) gives

$$
\|M - \widehat{M}\|_{1,\infty} \|(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1
$$
$$
\leq 2\|M^{-1}\|_{1,\infty}(2\epsilon_{\infty,1} + \epsilon_{1,\infty})\Big[2\epsilon_{\infty,1}\|\widehat{r} - r\|_1 + (2\epsilon_{\infty,1} + \epsilon_{1,\infty})\|A^+(\widehat{r} - r)\|_1\Big]. \qquad (92)
$$

On the other hand, from Eq. (48) and its subsequent arguments, it is straightforward to show

$$
\|\widehat{M}(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 \leq \|\widehat{A}^\top \widehat{D}^{-1}\|_{1,\infty}\|(\widehat{D} - D)D^{-1}\|_{1,\infty}\|\widehat{r} - r\|_1
$$
$$
+ \|(\widehat{A} - A)^\top D^{-1}\|_{1,\infty}\|\widehat{r} - r\|_1
$$
$$
+ \|M - \widehat{M}\|_{1,\infty}\|A^+(\widehat{r} - r)\|_1
$$
$$
\leq 2\epsilon_{\infty,1}\|\widehat{r} - r\|_1 + (2\epsilon_{\infty,1} + \epsilon_{1,\infty})\|A^+(\widehat{r} - r)\|_1. \qquad (93)
$$

By combining Eq. (92) and Eq. (93) and by using Eq. (26) together with the fact that

$$
\|M^{-1}\|_{\infty,1} \leq \sqrt{K}\|M^{-1}\|_{\mathrm{op}} = \frac{\sqrt{K}}{\lambda_K(M)} \qquad (94)
$$

to collect terms, we obtain

$$
\|M(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 \lesssim \epsilon_{\infty,1}\|\widehat{r} - r\|_1 + (2\epsilon_{\infty,1} + \epsilon_{1,\infty})\|A^+(\widehat{r} - r)\|_1.
$$

Finally, using $\|\widehat{r} - r\|_1 \leq 2$ and Eq. (46) concludes

$$
\|M(\widehat{A}^+ - A^+)(\widehat{r} - r)\|_1 = \mathcal{O}_\mathbb{P}\left(\epsilon_{\infty,1} + (\epsilon_{\infty,1} + \epsilon_{1,\infty})\lambda_K^{-1}(M)\sqrt{\frac{K\log(K)}{N}}\right),
$$

implying Eq. (88) under $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N})$, Eq. (26) and Eq. (91).

A.9.3 Proof of Eq. (89)

By similar arguments, we have

$$
\sqrt{N}\frac{u^\top(\widehat{A}^+ - A^+)r}{\sqrt{u^\top V_r u}} \leq \sqrt{N}\frac{\|M(\widehat{A}^+ - A^+)r\|_1}{\sqrt{\kappa(V_r)}}.
$$

Since, for $r = A\alpha$,

$$
\begin{aligned}
\|M(\widehat{A}^+ - A^+)r\|_1 &= \|M\widehat{A}^+(A - \widehat{A})\alpha\|_1 \\
&\leq \|M\widehat{A}^+\|_{1,\infty}\|(\widehat{A} - A)\alpha\|_1 \\
&\leq \left(\|\widehat{A}^\top\widehat{D}^{-1}\|_{1,\infty} + \|(\widehat{M} - M)\widehat{M}^{-1}\widehat{A}^\top\widehat{D}^{-1}\|_{1,\infty}\right)\|\widehat{A} - A\|_{1,\infty} \\
&\leq (1 + \|\widehat{M} - M\|_{1,\infty}\|\widehat{M}^{-1}\|_{1,\infty})\epsilon_{1,\infty} \\
&\leq \left[1 + 2\|M^{-1}\|_{1,\infty}(2\epsilon_{\infty,1} + \epsilon_{1,\infty})\right]\epsilon_{1,\infty}
\end{aligned}
$$

by using Lemma 6 and Eq. (53) in the last step, invoking Eq. (26), (94) and $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N})$ concludes

$$
\|M(\widehat{A}^+ - A^+)r\|_1 = \mathcal{O}(\epsilon_{1,\infty})
$$

implying Eq. (89). The proof is complete. $\qquad\square$

A.10   Proof of Theorem 7

Proof. Recall Eq. (7) and Eq. (8). We have

$$
\begin{aligned}
\widehat{W}^S(r,s) - W^S(r,s) &= \sup_{f\in\widehat{\mathcal{F}}} f^\top(\widehat{\alpha} - \widehat{\beta}) - \sup_{f\in\mathcal{F}} f^\top(\widehat{\alpha} - \widehat{\beta}) \\
&\quad + \sup_{f\in\mathcal{F}} f^\top(\widehat{\alpha} - \widehat{\beta}) - \sup_{f\in\mathcal{F}} f^\top(\alpha - \beta).
\end{aligned}
$$

From Eq. (38), we know

$$
\left|\sup_{f\in\widehat{\mathcal{F}}} f^\top(\widehat{\alpha} - \widehat{\beta}) - \sup_{f\in\mathcal{F}} f^\top(\widehat{\alpha} - \widehat{\beta})\right| \leq \epsilon_d\left(2 + \|\widehat{\alpha} - \alpha\|_1 + \|\widehat{\beta} - \beta\|_1\right).
$$

By invoking Theorem 3, recalling Eq. (94) and using $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N})$, $\epsilon_d\sqrt{N} = o(1)$, Eq. (29) as well as $\max\{\lambda_K(V_r), \lambda_K(V_s)\} \leq 1$, it follows

$$
\sqrt{N}\left|\sup_{f\in\widehat{\mathcal{F}}} f^\top(\widehat{\alpha} - \widehat{\beta}) - \sup_{f\in\mathcal{F}} f^\top(\widehat{\alpha} - \widehat{\beta})\right| = o_{\mathbb{P}}(1).
$$

It remains to show that, as $N \to \infty$,

$$
\sqrt{N}\left[\sup_{f\in\mathcal{F}} f^\top(\widehat{\alpha} - \widehat{\beta}) - \sup_{f\in\mathcal{F}} f^\top(\alpha - \beta)\right] \xrightarrow{d} \sup_{f\in\mathcal{F}'(r,s)} f^\top Z_{rs}.
$$

From Theorem 6, we know that

$$
\sqrt{N}\left[(\widehat{\alpha} - \widehat{\beta}) - (\alpha - \beta)\right] \xrightarrow{d} Z_{rs}, \quad \text{as } N \to \infty.
$$

Then for the function $h : \mathbb{R}^K \to \mathbb{R}$ defined as $h(x) = \sup_{f\in\mathcal{F}} f^\top x$, since Proposition 2 shows that h is Hadamard-directionally differentiable at $u = \alpha - \beta$ with derivative $h'_u : \mathbb{R}^K \to \mathbb{R}$ defined as

$$
h'_u(x) = \sup_{f\in\mathcal{F}:f^\top u=h(u)} f^\top x = \sup_{f\in\mathcal{F}'(r,s)} f^\top x,
$$

the result follows by an application of Theorem 8. $\qquad\square$

A.10.1 Hadamard directional derivative

To obtain distributional limits, we employ the notion of *directional Hadamard differentiability*. The differentiability of our functions of interest follows from well known general results [see, e.g., 11, Section 4.3]; we give a self-contained proof for completeness..

Proposition 2. *Define a function $f : \mathbb{R}^d \to \mathbb{R}$ by*

$$f(x) = \sup_{c \in \mathcal{C}} c^\top x, \tag{95}$$

for a convex, compact set $\mathcal{C} \subseteq \mathbb{R}^d$. Then for any $u \in \mathbb{R}^d$ and sequences $t_n \searrow 0$ and $h_n \to h \in \mathbb{R}^d$,

$$\lim_{n \to \infty} \frac{f(u + t_n h_n) - f(u)}{t_n} = \sup_{c \in \mathcal{C}_u} c^\top h, \tag{96}$$

where $\mathcal{C}_u := \{c \in \mathcal{C} : c^\top u = f(u)\}$. In particular, f is Hadamard-directionally differentiable.

Proof. Denote the right side of Eq. (96) by $g_u(h)$. For any $c \in \mathcal{C}_u$, the definition of f implies

$$\frac{f(u + t_n h_n) - f(u)}{t_n} \geq \frac{c^\top(u + t_n h_n) - c^\top u}{t_n} = c^\top h_n. \tag{97}$$

Taking limits of both sides, we obtain

$$\liminf_{n \to \infty} \frac{f(u + t_n h_n) - f(u)}{t_n} \geq c^\top h, \tag{98}$$

and since $c \in \mathcal{C}_u$ was arbitrary, we conclude that

$$\liminf_{n \to \infty} \frac{f(u + t_n h_n) - f(u)}{t_n} \geq g_u(h). \tag{99}$$

In the other direction, it suffices to show that any cluster point of $\frac{f(u + t_n h_n) - f(u)}{t_n}$ is at most $g_u(h)$. For each n, pick $c_n \in \mathcal{C}_{u + t_n h_n}$, which exists by compactness of $\mathcal{C}$. By passing to a subsequence, we may assume that $c_n \to \bar{c} \in \mathcal{C}$, and since

$$\bar{c}^\top u = \lim_{n \to \infty} c_n^\top(u + t_n h_n) = \lim_{n \to \infty} f(u + t_n h_n) = f(u), \tag{100}$$

we obtain that $\bar{c} \in \mathcal{C}_u$. On this subsequence, we therefore have

$$\limsup_{n \to \infty} \frac{f(u + t_n h_n) - f(u)}{t_n} \leq \lim_{n \to \infty} \frac{c_n^\top(u + t_n h_n) - c_n^\top u}{t_n} = \bar{c}^\top h \leq g_u(h), \tag{101}$$

proving the claim. $\qquad\square$

The following theorem is a variant of the δ-method that is suitable for Hadamard directionally differentiable functions.

Theorem 8 (Theorem 1, [45])**.** *Let f be a function defined on a subset F of $\mathbb{R}^d$ with values in $\mathbb{R}$, such that: (1) f is Hadamard directionally differentiable at $u \in F$ with derivative $f'_u : F \to \mathbb{R}$, and (2) there is a sequence of $\mathbb{R}^d$-valued random variables X_n and a sequence of non-negative numbers $\rho_n \to \infty$ such that $\rho_n(X_n - u) \xrightarrow{d} X$ for some random variable X taking values in F. Then, $\rho_n(f(X_n) - f(u)) \xrightarrow{d} f'_u(X)$.*

B Limiting distribution of our W^S estimator for two samples with different sample sizes

We generalize the results in Theorem 7 to cases where $\widehat{r}$ and $\widehat{s}$ are the empirical estimates based on N_r and N_s, respectively, i.i.d. samples of r and s. Write $N_{\min} = \min\{N_r, N_s\}$ and $N_{\max} = \max\{N_r, N_s\}$. Let $Z'_{rs} \sim \mathcal{N}_K(0, Q'_{rs})$ where

$$Q'_{rs} = \lim_{N_{\min} \to \infty} A^+ \left(\frac{N_s}{N_r + N_s} \Sigma_r + \frac{N_r}{N_r + N_s} \Sigma_s \right) A^{+\top}.$$

Theorem 9. *Under Assumption 1, let $\widehat{A}$ be any estimator such that Eq. (11), Eq. (16) and Eq. (17) hold. Assume K does not grow with N. Further assume $\lambda_K^{-1}(M) = \mathcal{O}(\sqrt{N_{\min}})$, $\max\{\lambda_K^{-1}(V_r), \lambda_K^{-1}(V_s)\} = o(N_{\min})$, $\epsilon_d \sqrt{N_{\max}} = o(1)$ and*

$$(\epsilon_{1,\infty} + \epsilon_{\infty,1}) \sqrt{N_{\max} / \min\{\lambda_K(V_r), \lambda_K(V_s)\}} = o(1), \quad as \ N \to \infty.$$

Then, as $N_{\min} \to \infty$, we have the following convergence in distribution,

$$\sqrt{\frac{N_r N_s}{N_r + N_s}} \left(\widehat{W}^S(r, s) - W^S(r, s) \right) \xrightarrow{d} \sup_{f \in \mathcal{F}'(r,s)} f^\top Z'_{rs}$$

with $\mathcal{F}'(r, s)$ defined in Eq. (31).

Proof. From Theorem 6, by recognizing that

$$\sqrt{\frac{N_r N_s}{N_r + N_s}} \left[(\widehat{\alpha} - \widehat{\beta}) - (\alpha - \beta) \right] \xrightarrow{d} Z'_{rs}, \quad as \ N_{\min} \to \infty,$$

the proof follows from the same arguments in the proof of Theorem 7. $\square$

C Additional results in Section 5

The following table gives all pairwise Sketched Wasserstein Distances between the selected papers in Table 3 for d chosen from TV, AJ_5 and AJ_25.

Table 5: Pairwise $\widehat{W}^S$ based on TV distance

	Doc 0	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Doc 0	0.00	0.36	0.36	0.38	0.39	0.36	0.33	0.48	0.48	0.47	0.36
Doc 1	0.36	0.00	0.43	0.36	0.37	0.00	0.37	0.40	0.41	0.40	0.43
Doc 2	0.36	0.43	0.00	0.47	0.44	0.43	0.43	0.51	0.51	0.51	0.01
Doc 3	0.38	0.36	0.47	0.00	0.33	0.36	0.34	0.39	0.41	0.42	0.47
Doc 4	0.39	0.37	0.44	0.33	0.00	0.37	0.06	0.44	0.39	0.44	0.45
Doc 5	0.36	0.00	0.43	0.36	0.37	0.00	0.37	0.40	0.41	0.39	0.44
Doc 6	0.33	0.37	0.43	0.34	0.06	0.37	0.00	0.44	0.40	0.45	0.44
Doc 7	0.48	0.40	0.51	0.39	0.44	0.40	0.44	0.00	0.48	0.04	0.51
Doc 8	0.48	0.41	0.51	0.41	0.39	0.41	0.40	0.48	0.00	0.48	0.51
Doc 9	0.47	0.40	0.51	0.42	0.44	0.39	0.45	0.04	0.48	0.00	0.51
Doc 10	0.36	0.43	0.01	0.47	0.45	0.44	0.44	0.51	0.51	0.51	0.00

Table 6: Pairwise $\widehat{W}^S$ based on Average Jaccard distance with $T = 5$

	Doc 0	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Doc 0	0.00	1.00	0.90	0.95	0.94	1.00	0.79	1.00	1.00	1.00	0.92
Doc 1	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Doc 2	0.90	1.00	0.00	1.00	0.98	1.00	0.97	1.00	1.00	1.00	0.02
Doc 3	0.95	1.00	1.00	0.00	1.00	1.00	0.99	0.94	1.00	1.00	1.00
Doc 4	0.94	1.00	0.98	1.00	0.00	1.00	0.15	1.00	0.97	1.00	1.00
Doc 5	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
Doc 6	0.79	1.00	0.97	0.99	0.15	1.00	0.00	1.00	0.97	1.00	1.00
Doc 7	1.00	1.00	1.00	0.94	1.00	1.00	1.00	0.00	1.00	0.07	1.00
Doc 8	1.00	1.00	1.00	1.00	0.97	1.00	0.97	1.00	0.00	1.00	1.00
Doc 9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.07	1.00	0.00	1.00
Doc 10	0.92	1.00	0.02	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Table 7: Pairwise $\widehat{W}^S$ based on Average Jaccard distance with $T = 25$

	Doc 0	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Doc 9	Doc 10
Doc 0	0.00	0.90	0.77	0.92	0.89	0.91	0.75	0.97	0.98	0.97	0.78
Doc 1	0.90	0.00	0.95	0.96	0.95	0.00	0.94	0.95	0.95	0.95	0.95
Doc 2	0.77	0.95	0.00	1.00	0.95	0.96	0.93	0.99	0.99	0.99	0.02
Doc 3	0.92	0.96	1.00	0.00	0.96	0.96	0.95	0.91	0.98	0.97	1.00
Doc 4	0.89	0.95	0.95	0.96	0.00	0.95	0.14	0.98	0.83	0.99	0.96
Doc 5	0.91	0.00	0.96	0.96	0.95	0.00	0.94	0.95	0.95	0.95	0.96
Doc 6	0.75	0.94	0.93	0.95	0.14	0.94	0.00	0.98	0.85	0.98	0.94
Doc 7	0.97	0.95	0.99	0.91	0.98	0.95	0.98	0.00	0.99	0.07	0.99
Doc 8	0.98	0.95	0.99	0.98	0.83	0.95	0.85	0.99	0.00	0.99	0.99
Doc 9	0.97	0.95	0.99	0.97	0.99	0.95	0.98	0.07	0.99	0.00	0.99
Doc 10	0.78	0.95	0.02	1.00	0.96	0.96	0.94	0.99	0.99	0.99	0.00

References

[1] ALTSCHULER, J. M., WEED, J. and RIGOLLET, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, eds.).
URL https://proceedings.neurips.cc/paper/2017/hash/491442df5f88c6aa018e86dac21d3606-Abstract.html

[2] ARORA, S., GE, R., HALPERN, Y., MIMNO, D. M., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*.

[3] ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012, IEEE 53rd Annual Symposium*. IEEE.

[4] BANSAL, T., BHATTACHARYYA, C. and KANNAN, R. (2014). A provable svd-based algorithm for learning topics in dominant admixture corpus. *Advances in neural information processing systems* **27**.

[5] BELLET, A., HABRARD, A. and SEBBAN, M. (2015). Metric learning. *Synthesis lectures on artificial intelligence and machine learning* **9** 1–151.

[6] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2021). Likelihood estimation of sparse topic distributions in topic models and its applications to wasserstein document distance calculations. *arXiv preprint arXiv:2107.05766* .

[7] BING, X., BUNEA, F. and WEGKAMP, M. (2020). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796.

[8] BING, X., BUNEA, F. and WEGKAMP, M. (2020). Optimal estimation of sparse topic models. *Journal of Machine Learning Research* **21** 1–45.

[9] BLEI, D. M. (2012). Introduction to probabilistic topic models. *Communications of the ACM* **55** 77–84.

[10] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* 993–1022.

[11] BONNANS, J. F. and SHAPIRO, A. (2000). *Perturbation analysis of optimization problems.* Springer Series in Operations Research, Springer-Verlag, New York.

[12] CHA, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical models and Methods in Applied Sciences* **1** 300–307.

[13] CHEN, Y., GEORGIOU, T. T. and TANNENBAUM, A. (2018). Optimal transport for gaussian mixture models. *IEEE Access* **7** 6269–6278.

[14] CUTURI, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States* (C. J. C. Burges, L. Bottou, Z. Ghahramani and K. Q. Weinberger, eds.).
URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[15] DAVIS, J. V., KULIS, B., JAIN, P., SRA, S. and DHILLON, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning.*

[16] DELON, J. and DESOLNEUX, A. (2020). A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM J. Imaging Sci.* **13** 936–970.
URL https://doi.org/10.1137/19M1301047

[17] DING, W., ISHWAR, P. and SALIGRAMA, V. (2015). Most large topic models are approximately separable. In *2015 Information Theory and Applications Workshop (ITA).*

[18] DÜMBGEN, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields* **95** 125–140.

[19] EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7** 1 – 26.
URL https://doi.org/10.1214/aos/1176344552

[20] GENEVAY, A., CHIZAT, L., BACH, F. R., CUTURI, M. and PEYRÉ, G. (2019). Sample complexity of sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan* (K. Chaudhuri and M. Sugiyama, eds.), vol. 89 of *Proceedings of Machine Learning Research.* PMLR.
URL http://proceedings.mlr.press/v89/genevay19a.html

[21] GOLDFELD, Z., GREENEWALD, K., NILES-WEED, J. and POLYANSKIY, Y. (2020). Convergence of smoothed empirical measures with applications to entropy estimation. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **66** 4368–4391.

[22] GOLDFELD, Z. and GREENEWALD, K. H. (2020). Gaussian-smoothed optimal transport: Metric structure and statistical efficiency. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]* (S. Chiappa and R. Calandra, eds.), vol. 108 of *Proceedings of Machine Learning Research.* PMLR.
URL http://proceedings.mlr.press/v108/goldfeld20a.html

[23] GREENE, D., O'CALLAGHAN, D. and CUNNINGHAM, P. (2014). How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases* (T. Calders, F. Esposito, E. Hüllermeier and R. Meo, eds.). Springer Berlin Heidelberg, Berlin, Heidelberg.

[24] HAN, Y., JIAO, J. and WEISSMAN, T. (2020). Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory* **1** 814–823.

[25] HEINRICH, P. and KAHN, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** 2844–2870.
URL https://doi.org/10.1214/17-AOS1641

[26] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex analysis and minimization algorithms. I*, vol. 305 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences].* Springer-Verlag, Berlin. Fundamentals.

[27] HUANG, A. ET AL. (2008). Similarity measures for text document clustering. In *Proceedings of*

the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol. 4.

[28] JACCARD, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist* **11** 37–50.

[29] JIAO, J., HAN, Y. and WEISSMAN, T. (2018). Minimax estimation of the L_1 distance. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **64** 6672–6706.

[30] KAYA, M. and BILGE, H. Ş. (2019). Deep metric learning: A survey. *Symmetry* **11** 1066.

[31] KE, T. Z. and WANG, M. (2017). A new svd approach to optimal topic estimation. *arXiv:1704.07016* .

[32] KIRK, W. and SHAHZAD, N. (2014). *Fixed point theory in distance spaces.* Springer, Cham.

[33] KLOPP, O., PANOV, M., SIGALLA, S. and TSYBAKOV, A. (2021). Assigning topics to documents by successive projections. *arXiv preprint arXiv:2107.03684* .

[34] KOLOURI, S., PARK, S. R., THORPE, M., SLEPCEV, D. and ROHDE, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine* **34** 43–59.

[35] LE CAM, L. (1986). *Asymptotic methods in statistical decision theory.* Springer Series in Statistics, Springer-Verlag, New York.
URL https://doi.org/10.1007/978-1-4612-4946-7

[36] MCCUTCHEON, A. L. (1987). *Latent class analysis.* 64, Sage.

[37] MCLACHLAN, G. and PEEL, D. (2000). *Finite mixture models.* Wiley Series in Probability and Statistics: Applied Probability and Statistics, Wiley-Interscience, New York.
URL https://doi.org/10.1002/0471721182

[38] MENA, G. and NILES-WEED, J. (2019). Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox and R. Garnett, eds.).
URL https://proceedings.neurips.cc/paper/2019/hash/5acdc9ca5d99ae66afdfe1eea0e3b26b-Abstract.html

[39] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics* **41** 370–400.

[40] NGUYEN, X. (2016). Borrowing strengh in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli* **22** 1535–1571.
URL https://doi.org/10.3150/15-BEJ703

[41] NILES-WEED, J. and RIGOLLET, P. (2022+). Estimation of wasserstein distances in the spiked transport model. *Bernoulli* To appear.

[42] PARR, W. C. and SCHUCANY, W. R. (1980). Minimum distance and robust estimation. *J. Amer. Statist. Assoc.* **75** 616–624.
URL http://links.jstor.org/sici?sici=0162-1459(198009)75:371<616:MDARE>2.0.CO;2-S&origin=MSN

[43] PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* **185** 71–110.

[44] PERRONE, V., JENKINS, P. A., SPANO, D. and TEH, Y. W. (2017). Poisson random fields for dynamic feature models. *Journal of Machine Learning Research* **18**.

[45] RÖMISCH, W. (2004). Delta method, infinite dimensional. *In Encyclopedia of Statistical Sciences* .

[46] SAHLER, W. (1968). A survey on distribution-free statistics based on distances between distribution functions. *Metrika* **13** 149–169.
URL https://doi.org/10.1007/BF02613382

[47] SOMMERFELD, M. and MUNK, A. (2018). Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 219–238.

[48] STEPHENS, M. A. (1974). Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association* **69** 730–737.

[49] SU, J., GREENE, D. and BOYDELL, O. (2016). Topic stability over noisy sources. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*. The COLING 2016 Organizing Committee, Osaka, Japan.

URL https://aclanthology.org/W16-3913

[50] Takahashi, W. (1970). A convexity in metric space and nonexpansive mappings. I. *Kodai Mathematical Seminar Reports* **22** 142–149.

[51] Tameling, C., Sommerfeld, M. and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability* **29** 2744–2781.

[52] Timan, A. F. (1994). *Theory of approximation of functions of a real variable*. Dover Publications, Inc., New York. Translated from the Russian by J. Berry, Translation edited and with a preface by J. Cossar, Reprint of the 1963 English translation.

[53] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics, Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[54] Vermunt, J. K. and Magidson, J. (2004). Latent class analysis. *The sage encyclopedia of social sciences research methods* **2** 549–553.

[55] Villani, C. (2009). *Optimal transport*, vol. 338 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin. Old and new.

[56] Weaver, N. (2018). *Lipschitz algebras*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. Second edition of [MR1832645].

[57] Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* **25** 2620–2648.

[58] Wolfowitz, J. (1953). Estimation by the minimum distance method. *Ann. Inst. Statist. Math., Tokyo* **5** 9–23.
URL https://doi.org/10.1007/bf02949797

[59] Wolfowitz, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88.
URL https://doi.org/10.1214/aoms/1177707038

[60] Wu, Y. and Yang, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics* **47** 857–883.

[61] Wu, Y. and Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** 1981–2007.
URL https://doi.org/10.1214/19-AOS1873

[62] Yang, L. and Jin, R. (2006). Distance metric learning: A comprehensive survey. *Michigan State Universiy* **2** 4.

[63] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F. and Solomon, J. M. (2019). Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox and R. Garnett, eds.).
URL https://proceedings.neurips.cc/paper/2019/hash/8b5040a8a5baf3e0e67386c2e3a9b903-Abstract.html

[64] Yurochkin, M., Claici, S., Chien, E., Mirzazadeh, F. and Solomon, J. M. (2019). Hierarchical optimal transport for document representation. *Advances in Neural Information Processing Systems* **32**.