Supplementary material for Dirichlet Simplex Nest and Geometric Inference

Mikhail Yurochkin [1 2 *] Aritra Guha [3 *] Yuekai Sun [3] XuanLong Nguyen [3]

1. Proofs of Theorem 1 and 2

In this section we present the proofs of main theorems as described in the main text. We will first reintroduce some notations for the reader's convenience.

Notation Let $\lambda_{\max}(A)$ and $\lambda_{min}(A)$ denote the largest and smallest non-zero singular values of the matrix A. We use $f(\cdot)$ to denote the density of $\mathbb{Q}$ with respect to Lebesgue measure on the $K-1$ dimensional subspace containing the simplex $\mathscr{B}$. Let $g(\cdot)$ be the density of $\mathbb{P}$ with respect to the Lebesgue measure on the $K-1$ dimensional space containing the eigenvectors of Σ_{tot}^K, where Σ_{tot}^K is best $K-1$-rank approximation matrix of $\Sigma_{tot} := BSB^T + \epsilon_0 I_D$ and $\epsilon_0 I_D$ is a uniform upper bound on Cov$[x_i \mid \theta]$. Let Σ be the population covariance matrix with Σ^K as the best $K-1$ rank approximation. Note that

$$\begin{aligned}\Sigma = \text{Cov}(X_i) &= \mathbb{E}[\text{Cov}(X_i|\mu_i)] + \text{Cov}(\mathbb{E}[X_i|\mu_i]) \\ &\leq \epsilon_0 I_D + BSB^T. \end{aligned} \tag{1}$$

The following is a standard assumption to ensure the consistency of the k-means procedure embedded in our algorithm:

(a.1) Pollard's regularity criterion (PRC): The Hessian matrix of the function $c \mapsto \mathbb{Q}\phi_{BSB^T}(\cdot, c)$ evaluated at c^* for all optimizer c^* of $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ is positive definite, with minimum eigenvalue $\lambda_0 > 0$.

Theorem 1. Consider the noiseless setting, i.e., $F(\cdot \mid \mu) = \delta_\mu$. Suppose that $\mathscr{B} = \text{Conv}(\beta_1, \dots, \beta_K)$ is the true simplex, while $(\beta_{1n}, \dots, \beta_{Kn})$ are the vertex estimates obtained by VLAD algorithm. Moreover, we assume that the error in the Monte Carlo estimates of the extension parameters is negligible. Provided that condition (a.1) holds, then

$$\min_\pi \|(\beta_{\pi_{(1)}n}, \dots, \beta_{\pi_{(K)}n}) - (\beta_1, \dots, \beta_K)\| = O_{\mathbb{P}}(n^{-1/2})$$

where the minimization is taken over all permutations π of $\{1, \dots, K\}$.

Proof. First, we note that under the assumption of the noiseless setting, by following along the lines of the proof of Lemma 2 in main text, it can be seen that if $c^* = (c_1^*, \dots, c_K^*)$ optimize Eq. (1) in the main text and v_k's are such that $(v_1, \dots, v_K)$ form the empirical CVT centroids of Δ^{K-1}, then $c_i^* = BPv_i + c_0$, where c_0 is the population centroid.

Next, the convergence of the empirical CVT centroids to the corresponding population CVT centroids occurs at rate $O_{\mathbb{P}}(\frac{1}{\sqrt{n}})$ rate following Pollard (1982). The consistency of the extreme points of the Dirichlet Simplex Nest follows by the continuous mapping theorem since

$$\frac{\|Pe_k\|_2}{\|Pv_k\|_2} = \frac{\|e_k - \frac{1}{K}\mathbf{1}_K\|_2}{\|v_k - \frac{1}{K}\mathbf{1}_K\|_2} = \frac{\|B(e_k - \frac{1}{K}\mathbf{1}_K)\|_2}{\|B(v_k - \frac{1}{K}\mathbf{1}_K)\|_2}, \tag{2}$$

where $e_1, \dots, e_K$ are the canonical basis vectors on $\mathbb{R}^K$ denoting the vertices of Δ^{K-1}.

Finally, the knowledge of α enables us to compute $\frac{\|e_k - \frac{1}{K}\mathbf{1}_K\|_2}{\|v_k - \frac{1}{K}\mathbf{1}_K\|_2}$. This concludes the proof. □

It is considerably more challenging to establish the error bounds for our algorithm in the general setting where the observations are noisy. First, let us define the following:

$$\begin{aligned}\mathscr{C}_{\mathbb{P}_n} =& \{c^* : c^* = \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \mathbb{P}_n \phi_{(\Sigma_n)^K}(\cdot, c) \\ &= \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \frac{1}{n}\sum_{i=1}^n \phi_{(\Sigma_n)^K}(\tilde{X}_i, c)\}, \\ \mathscr{C}_{\mathbb{Q}} =& \{c^* : c^* = \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \mathbb{Q}\phi_{BSB^T}(\cdot, c)\}.\end{aligned}$$

Recall the following assumptions from the main text:

(a.2) The Hessian matrix of the function $c \mapsto \mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ evaluated at c^* for all optimizer c^* of $\mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ is uniformly positive definite with minimum eigenvalue bounded below from some $\lambda_0 > 0$, for all $(\Sigma)^K$ such that $(\Sigma - BSB^T) \leq \tilde{\epsilon} I_D$, for some $\tilde{\epsilon} > 0$.

[*] Equal contribution [1] IBM Research, Cambridge [2] MIT-IBM Watson AI Lab [3] Department of Statistics, University of Michigan. Correspondence to: Mikhail Yurochkin <mikhail.yurochkin@ibm.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

(b) There exists $\epsilon_0 > 0$ such that $\epsilon_0 I_D - \mathrm{Conv}(X|\theta)$ is positive semi-definite uniformly over $\theta \in \Delta^{K-1}$.

(c) There exists M_0 such that for all $M > M_0$,

$$\int_{\mathcal{B}(\sqrt{M},c_0)^c} \|x - c_0\|_2^2 g(x)\mathrm{d}x \le \frac{k_1}{M},$$

for some universal constant k_1, where $\mathcal{B}(\sqrt{M}, c_0)$ is a ball of radius $\sqrt{M}$ around the population centroid, c_0.

The assumptions (b) and (c) are very general assumptions and satisfied by a vast array of noise distributions, especially those with subexponential tails. In particular, the noise distributions considered in this work all satisfy these assumptions.

We restate the second main theorem of the paper.

Theorem 2. Suppose that $\mathscr{B} = \mathrm{Conv}(\beta_1, \dots, \beta_K)$ is the true simplex, while $(\beta_{1n}, \dots, \beta_{Kn})$ are the vertex estimates obtained by VLAD algorithm. Assume the error in the Monte Carlo estimates of the extension parameter is negligible. Provided that conditions (a.2), (b) and (c) hold, then

$$\min\nolimits_\pi \|(\beta_{\pi_{(1)}n}, \dots, \beta_{\pi_{(K)}n}) - (\beta_1, \dots, \beta_K)\|_2 = O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right) + O_{\mathbb{P}}(n^{-1/2}), \qquad (3)$$

where the minimization is over all permutations π of $\{1, \dots, K\}$.

Proof. The proof proceeds by the following steps:

First, in **Step 1**, we show that it is enough to restrict attention to the population estimates instead of empirical estimates. Next, in **Step 2**, we show that the k-means objectives for distributions of μ_i's and x_i's are close. **Step 3** shows that the objective values at the respective minimizers are also close to each other for the distributions considered in **Step 2**. Finaly, **Step 4** uses the strong convexity condition of (a.2) to bound the distance between respective k-means centers, and **Step 5** translates this bound to the estimation of the simplex vertices.

In that regard,

Step 1: Following Pollard (1982), the empirical estimates of CVT centroids optimizing $\mathbb{P}\phi_{\Sigma^K}(\cdot, c)$ converges to the corresponding population estimate at rate $O_{\mathbb{P}_n}(n^{-1/2})$. Thus it is enough to restrict attention to the population estimates.

Step 2: We will show that for all ϵ_0 sufficiently small,

$$|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)| = O(\epsilon_0^{1/3})$$

uniformly over $c \in \mathscr{B}^K$.

Since $\mathbb{Q}$ denotes the distribution corresponding to μ_i's, this distribution places its entire mass inside the simplex, therefore all minimizers of the function $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ lie inside $\mathscr{B}^K$. We can hence restrict our attention to $c \in \mathscr{B}^K$. By assumption (b), we have $BSB^T \le \Sigma^K$. Thus, it is enough to establish a bound for $|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)|$ $\forall\, c \in \mathscr{B}^K$.

$$\begin{aligned}|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)| &\le |\mathbb{P}\phi_{\Sigma^K}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)| \\ &\quad + |\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)|.\end{aligned} \qquad (4)$$

Step 2.1: Now, to bound the second term on the right hand side of Eq. (4) we use,

$$\begin{aligned}&|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)| \\ &\quad\le \int |\phi_{BSB^T}(x, c) - \phi_{\Sigma^K}(x, c)| f(x)\mathrm{d}x \\ &\quad\le \lambda_{\max}([BSB^T]^\dagger - [\Sigma^K]^\dagger) \\ &\quad\le \lambda_{\max}([BSB^T]^\dagger - [(BSB^T + \epsilon_0 I_D^K]^\dagger) \\ &\quad\le \frac{\epsilon_0}{\lambda_{\min}(BSB^T)\lambda_{\min}(BSB^T + \epsilon_0 I_D^K)},\end{aligned}$$

where $B^\dagger$ denotes the pseudo-inverse of B, and I_D^K is the matrix with top $K-1$ diagonal elements as 1, the rest zeros.

Step 2.2: Turning to the first term on right hand side of Eq. (4), we note that $\|\beta_i - \beta_j\|^2 \le \frac{K-1}{K}\lambda_{\max}(BSB^T)$. Therefore a compact ball of radius $a\lambda_{\max}(BSB^T)$ around the centroid c_0 of the simplex $\mathscr{B}$ for all sufficiently large constants $a > \frac{K-1}{K}$ contains the simplex completely. Consider a ball $\mathcal{B}(\sqrt{M}, c_0)$ of radius $\sqrt{M}$, with $M = a\lambda_{\max}(BSB^T)$ around the centroid c_0, the scalar a to be chosen later. For any $M > 0$,

$$\begin{aligned}|\mathbb{P}\phi_{\Sigma^K}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)| &\le \left| \int_{\mathcal{B}(\sqrt{M},c_0)^c} \phi_{\Sigma^K}(x, c) g(x)\mathrm{d}x \right| \\ &\quad + \left| \int_{\mathcal{B}(\sqrt{M},c_0)} \phi_{\Sigma^K}(x, c)[g(x) - f(x)]\mathrm{d}x \right|.\end{aligned} \qquad (5)$$

Step 2.2.1: For the first term on the right hand side of Eq. (5), we see that,

$$\begin{aligned}&\int_{\mathcal{B}(\sqrt{M},c_0)^c} \phi_{\Sigma^K}(x, (c_1, \dots, c_K)) g(x)\mathrm{d}x \\ &\le \min_{i \in \{1,\dots,K\}} \int_{\mathcal{B}(\sqrt{M},c_0)^c} \|x - c_i\|_{\Sigma^K}^2 g(x)\mathrm{d}x \\ &\le \max 2\|c_i - c_0\|_2^2 \mathbb{P}(X \in \mathcal{B}(\sqrt{M}, c_0)^c) \\ &\qquad + \frac{2}{\lambda_{\min}(BSB^T)} \int_{\mathcal{B}(\sqrt{M},c_0)^c} \|x - c_0\|_2^2 g(x)\mathrm{d}x.\end{aligned} \qquad (6)$$

The first inequality follows from Fatou's lemma, while the second follows from the fact that $\|a+b\|_2^2 \leq 2(\|a\|_2^2+\|b\|_2^2)$.

Suppose that the noise distribution is subexponential for all latent locations $\theta \in \mathscr{B}$. Combining this with the Chebyshev inequality and condition (c), Eq. (6) can be re-written as:

$$\begin{aligned}
&\int_{\mathcal{B}(\sqrt{M},c_0)^c} \phi_{\Sigma^K}(x,(c_1,\dots,c_K))g(x)\mathrm{d}x \\
&\leq \tilde{C}\lambda_{\max}(BSB^T)\frac{Var(X)}{M} + \frac{2k_1}{\lambda_{\min}(BSB^T)M} \qquad (7)\\
&\leq \tilde{C}\frac{2(K-1)\lambda_{\max}^2(BSB^T)}{M} + \frac{2k_1}{\lambda_{\min}(BSB^T)M}
\end{aligned}$$

for some universal constant k_1.

Step 2.2.2: For the second term on the right hand side on Eq. (5), we use the following result.

Claim 1. For $M = a\lambda_{\max}(BSB^T)$, when centroids $c_i \in \mathscr{B}\ \forall\ i$, $\phi_{\Sigma^K}(x, c = (c_1,\dots,c_K))$ as a function of x is Lipschitz on $\mathcal{B}(\sqrt{M},c_0)$, with Lipschitz constant $\frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}$.

Now using the above result, we can easily extend $\phi_{\Sigma^K}(x, c = (c_1,\dots,c_K))$ to a Lipschitz function on the entire domain. For the particular choice of a,

$$\begin{aligned}
&\left|\int_{B(\sqrt{M},c_0)} \phi_{(\Sigma)^K}(x,c)(g(x)-f(x))\mathrm{d}x\right| \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} \sup_{\|l\|_{Lip}\leq 1}\left|\int l(x)(g(x)-f(x))\mathrm{d}x\right| \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} W_1(g,f) \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} \sqrt{(K-1)\epsilon_0}.
\end{aligned} \qquad (8)$$

In the above, $\|l\|_{Lip}$ denotes the Lipschitz constant of the function $l(\cdot)$. The second inequality in the above equation follows from Kantorovich-Rubinstein duality while for the last inequality, we use the definition of the Wasserstein distance and take (X,μ) as the coupling with densities $X \sim g$ and $\mu \sim f$ marginally (cf. (Villani, 2008)). Then, for any upper bound M_1 on the variance of $\|X-\mu\|_2$, $W_2(g,f) \leq M_1$, and we use the fact that $\sqrt{(K-1)\epsilon_0}$ forms such an upper bound.

Now, for the noise level $\epsilon_0 > 0$ sufficiently small, there exists $\epsilon > 0$, which is dependent on ϵ_0, such that the open interval $\left(C'\frac{(K-1)\lambda_{\max}^2(BSB^T)}{\epsilon}, \frac{\lambda_{\min}^2(BSB^T)}{\lambda_{\max}(BSB^T)(K-1)\epsilon_0}\epsilon^2/16\right)$ is non-empty for any fixed constant C'. Whenever a is chosen in this range, $|\mathbb{Q}\phi_{BSB^T}(\cdot,c) - \mathbb{P}\phi_{\Sigma^K}(\cdot,c)| \leq \epsilon$. Note that we can choose $\epsilon = O(\epsilon_0^{1/3})$ and $a = O(\epsilon_0^{-1/3})$ to satisfy the above condition.

Step 3: In this step, we show that objective function values for k-means corresponding to that of the population distributions of x_i's and μ_i's are close. Notice that the bounds obtained in Step 2 are uniform over $c \in \mathscr{B}$. For ease of writing, we denote $R_q(c) = \mathbb{Q}\phi_{BSB^T}(\cdot,c)$ and $R_p(c) = \mathbb{P}\phi_{\Sigma^K}(\cdot,c)$. Also, let $\operatorname{argmin} R_p(c) = c_p$ and $\operatorname{argmin} R_q(c) = c_q$. Then, for ϵ_0 sufficiently small, it follows from the discussion above that

$$\begin{aligned}
&|R_q(c_p) - R_q(c_q)| \\
&= |R_q(c_p) - R_p(c_p) + R_p(c_q) - R_q(c_q) + R_p(c_p) - R_p(c_q)| \\
&\leq |R_q(c_p) - R_p(c_p) + R_p(c_q) - R_q(c_q)| = O(\epsilon_0^{1/3}).
\end{aligned} \qquad (9)$$

Step 4: In this step, we show that $\|\operatorname{argmin}_c \mathbb{P}\phi_{\Sigma^K}(\cdot,c) - \operatorname{argmin}_c \mathbb{Q}\phi_{BSB^T}(\cdot,c)\|_2 \to 0$ as $\epsilon_0 \to 0$. The intuition behind this is that since the functions $\mathbb{Q}\phi_{BSB^T}(\cdot,c)$ and $R_p(c) = \mathbb{P}\phi_{\Sigma^K}(\cdot,c)$ are point-wise close, and their minimized values are also close to one another, therefore, the points of minima must also be close. By a standard strong convexity argument, employing condition (a.2), for ϵ_0 sufficiently small, we get,

$$\begin{aligned}
&\|\operatorname*{argmin}_c \mathbb{P}\phi_{\Sigma^K}(\cdot,c) - \operatorname*{argmin}_c \mathbb{Q}\phi_{BSB^T}(\cdot,c)\|_2 \\
&\qquad = O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right).
\end{aligned} \qquad (10)$$

Step 5 : Finally, the error bound for the simplex vertices follows from a continuous mapping theorem's argument in a similar manner to that of the proof for Theorem 1. □

Claim 1. For $M = a\lambda_{\max}(BSB^T)$, when centroids $c_i \in \mathscr{B}\ \forall\ i$, $\phi_{\Sigma^K}(x, c = (c_1,\dots,c_K))$ as a function of x is Lipschitz on $\mathcal{B}(\sqrt{M},c_0)$, with Lipschitz constant $\frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}$.

Proof of Claim 1.

$$\begin{aligned}
&\frac{|\phi_{\Sigma^K}(x, c = c_1,\dots,c_K) - \phi_{\Sigma^K}(y, c = c_1,\dots,c_K)|}{\|x-y\|} \\
&\leq \max_{i\in\{1,\dots,K\}} \frac{|\|x-c_i\|_{\Sigma^K} - \|y-c_i\|_{\Sigma^K}|}{\|x-y\|_2} \\
&\leq \sup \frac{2\|x-y\|}{\lambda_{\min}(BSB^T)} \leq \frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}.
\end{aligned} \qquad (11)$$

□

2. Consistent estimation of concentration parameter

In this section we first provide several easy calculations required for the estimating equations for some commonly used noise distributions.

Lemma 1. Depending on the data generating distribution, the covariance matrix of the DSN model is given as follows.

(a) Gaussian data: $\Sigma = BS(\alpha)B^T + \sigma^2 I_d$, provided that $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$.

(b) Poisson data: $\Sigma = BS(\alpha)B^T + \text{Diag}(\sum_i B_i/K)$, provided that $x_{ij}|\mu_i \overset{ind}{\sim} Poi(\mu_{ij})$, where B_i denotes the i^{th} column of B and $\text{Diag}(a)$ is a diagonal matrix with the i^{th} diagonal element denoting the i^{th} element of the vector a. Here, $\mu_i = (\mu_{i1}, \dots, \mu_{iD})$.

(c) Multinomial data: $\Sigma = (1 - \frac{1}{N})BS(\alpha)B^T + \frac{1}{N}\text{Diag}(\sum_i B_i/K) - \frac{1}{N}(\sum_i B_i/K)(\sum_i B_i/K)^T$, provided that $x_i|\mu_i \sim \text{Multinomial}(N, \mu_{i1}, \mu_{iD})$. Here, $\mu_i = (\mu_{i1}, \dots, \mu_{iD})$ is a probability vector. (N resembles the number of words per document in the LDA model).

Proof. We compute $Cov(x_i)$ for each of the models. Note that $Cov(X_i) = \mathbb{E}(Cov(x_i|\mu_i)) + Cov(\mathbb{E}(x_i|\mu_i))$ from the tower property of conditional covariance, and $Cov(\mathbb{E}(x_i|\mu_i)) = BS(\alpha)B^T$ for all the models. Therefore we just need the computation for $\mathbb{E}(Cov(x_i|\mu_i))$ for each of the models.

For the Gaussian model, $\mathbb{E}(Cov(x_i|\mu_i)) = \sigma^2 I_D$.

For the Poisson model, $\mathbb{E}(Cov(x_i|\mu_i)) = \mathbb{E}(\mu_i) = B\mathbb{E}(\theta_i) = \text{Diag}(\sum_i B_i/K)$, where the second equality follows as $\mu_i = B\theta_i$ by the model, and the last equality follows because $\theta_i \sim \text{Dir}(\alpha)$.

For the multinomial model, $\mathbb{E}(Cov(x_i|\mu_i)) = \frac{1}{N}\mathbb{E}(\text{Diag}(\mu_i)) - \frac{1}{N}Cov(\mu_i\mu_i^T) = \frac{1}{N}(\text{Diag}(\sum_i B_i/K) - BS(\alpha)B^T)$ from which the result follows.

□

Equation (6) in the main text, for estimating α uses the data covariance matrix, $\hat{\Sigma}_n$. While this gives the correct estimating equation in the noiseless scenario, but for the noisy version we need to use $\tilde{\Sigma}_n$ instead where $\tilde{\Sigma}_n$ is a consistent estimator for $BS(\alpha)B^T$. The estimator estimator for different noise distributions can be obtained via the above lemma.

2.1. Proof of consistency

The proof of consistency of the proposed estimate for the Dirichlet concentration parameter is given as follows.

Theorem 3. Assume that function $\varphi(\tilde{\alpha}) = \frac{\gamma(\tilde{\alpha})^2}{K(K\tilde{\alpha}+1)}$ is monotonically increasing in $\tilde{\alpha}$, where $\gamma(\tilde{\alpha})$ is the extension parameter corresponding to $\tilde{\alpha}$. Let $\alpha_0 \in \mathscr{C}$ be the true concentration parameter for some compact set $\mathscr{C}$. Let $\hat{\alpha}_n = \text{argmin}_{\alpha \in \mathscr{C}} \|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - \tilde{\Sigma}_n\|$, where $\tilde{\Sigma}_n$ is a consistent estimator of $BS(\alpha)B^T$. Then,

$$\|\hat{\alpha}_n - \alpha_0\| \xrightarrow{\mathbb{P}} 0. \tag{12}$$

Proof. Notice that $\|\tilde{\Sigma}_n - BS(\alpha_0)B^T\| = o_P(1)$. Also, $\|\hat{B}(\gamma(\alpha)) - B(\gamma(\alpha))\| = O_P(n^{-1/2})$ for all $\alpha \in \mathscr{C}$. Therefore $\|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - B(\gamma(\alpha))S(\alpha)B(\gamma(\alpha))^T\| = O_P(n^{-1})$ for all $\alpha \in \mathscr{C}$. By monotonicity of the function φ, $BS(\alpha_0)B^T - B(\gamma(\alpha))S(\alpha)B(\gamma(\alpha))^T$ as a function of α is injective for all $\alpha \in \mathscr{C}$. Therefore, $\|\hat{B}(\gamma(\alpha_0))S(\alpha_0)\hat{B}(\gamma(\alpha_0))^T - \tilde{\Sigma}_n\| = o_P(1)$, by triangle inequality. The statement of the theorem then follows by employing a subsequence argument. □

2.2. Identifiability of the concentration parameter

In the statement of Theorem 3, we require a condition which amounts to a identifiability condition of the parameter α. In this section, we provide empirical evidence that the DSN model with unknown concentration parameter α is identifiable from second moments.

As we shall see, the identifiability of α boils to the invertibility of a scalar function. Recall the covariance matrix of a $\mathsf{Dir}(\alpha)$ distribution is

$$S(\alpha) = \frac{I_K - P_K}{K(K\alpha + 1)},$$

where $P_K = \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$ is the projector onto $\mathsf{span}\{\mathbf{1}_K\}$. Let $B(\gamma) = \gamma(C - \mu) + \mu$ be the γ-extension of the (scaled) K-means centroids C from the center of the DSN $\mu = \frac{1}{K}B\mathbf{1}_K$. The question of the identifiability of the concentration parameter boils down to whether there are distinct α_1 and α_2 such that

$$\begin{aligned} &B(\gamma(\alpha_1))S(\alpha_1)B(\gamma(\alpha_1))^T \\ &\quad = B(\gamma(\alpha_2))S(\alpha_2)B(\gamma(\alpha_2))^T, \end{aligned} \tag{13}$$

where $\gamma(\alpha)$ is the extension parameter that corresponds to concentration parameter α. As long as C has full column rank, we may pre and post-multiply (13) by $C^\dagger$ and $(C^\dagger)^T$ respectively to see that (13) is equivalent to

$$\begin{aligned} &(\gamma(\alpha_1)(I_K - P_K) + P_K)S(\alpha_1)(\gamma(\alpha_1)(I_K - P_K) + P_K) \\ &\quad = (\gamma(\alpha_2)(I_K - P_K) + P_K)S(\alpha_2)(\gamma(\alpha_2)(I_K - P_K) + P_K). \end{aligned}$$

Recalling $S(\alpha)$ is a scalar multiple of $I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$, we see that (13) is equivalent to whether there are distinct α_1

and α_2 such that

$$\frac{\gamma(\alpha_1)^2}{K(K\alpha_1+1)} = \frac{\gamma(\alpha_2)^2}{K(K\alpha_2+1)}.$$

This is equivalent to the invertibility of the function

$$\varphi(\alpha) = \frac{\gamma(\alpha)^2}{K(K\alpha+1)}. \tag{14}$$

Figure 1 shows this function for $K = 10$ over a range of reasonable values of α. We see that the function is in fact invertible.

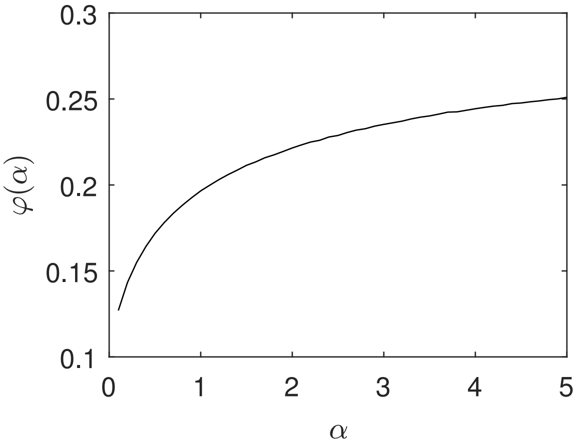


Figure 1: Empirical study of α identifiability.

Although Figure 1 suggests (14) is invertible, we do not have a rigorous proof. The main challenge is obtaining precise control on the growth of (13). Inspecting Figure 1 shows that $\varphi(\alpha)$ is almost flat as soon as α exceeds $\frac{5}{2}$. Intuitively, this is a consequence of the hardness of distinguishing between DSNs with large α's (and correspondingly large extension parameters). Mathematically, it is hard to obtain precise control on the growth of $\varphi(\alpha)$ because it is not possible to evaluate $\gamma(\alpha)$ explicitly. Although it is possible to show that

$$\gamma(\alpha) = \frac{1-\frac{1}{K}}{\int_{V_k} e_k^T \theta p_\alpha(x)dx - \frac{1}{K}}, \tag{15}$$

where $V_k = \{\theta \in \Delta^{K-1} : \operatorname{argmax}\{\theta_l : l \in [K]\} = k\}$ is the k-th Voronoi cell in a centroidal Voronoi tessellation of Δ^{K-1}, e_k is the k^{th} canonical basis vector and p_α is the $\mathsf{Dir}(\alpha)$ density, it is hard to evaluate the integral. We defer an investigation of the identifiability of the concentration parameter to future work.

3. Computational cost of VLAD

In this section, we tally up the computational cost of VLAD. The dominant cost it that of computing the top K singular factors of the centered data matrix $\bar{X}$. This costs $O(DKn)$ floating point operations (FLOP's). The cost of the subsequent clustering step is asymptotically negligible compared to the cost of the SVD. Assuming each step of the K-means algorithm costs $O(Kn)$ FLOP's and the algorithm converges linearly, we see that the cost of obtaining an $O(\frac{1}{n})$-suboptimal solution is $O(Kn\log n)$. We discount the cost of Monte Carlo estimates of the extension parameter because it can be tabulated. Thus the computational cost of the algorithm is dominated by the cost of computing the SVD.

4. Experimental details and additional results

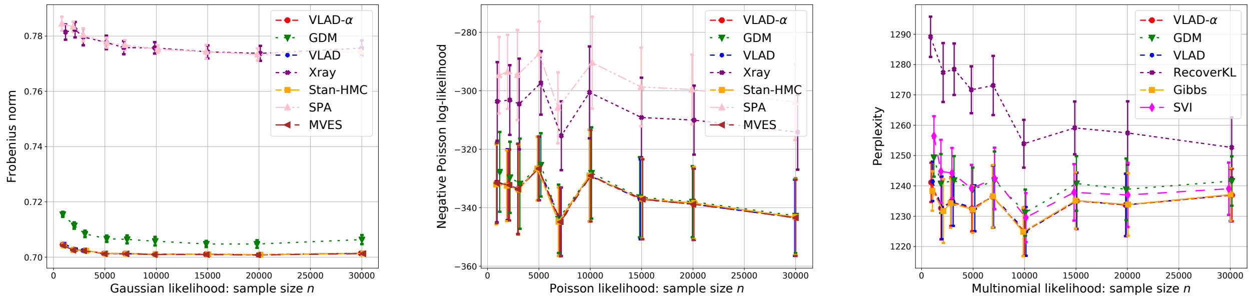
Additional results for convergence behavior We complement the results presented in Fig. 2 of the main text with the corresponding plots of the likelihood evaluated on a set of held out data. These results are summarized in Fig. 2. For all plots, the smaller value is better. We see that VLAD shows performance as good as HMC and Gibbs sampler at a much lower computational time. This supports our findings in the main text.

Additional results for geometry of the DSN Again, we further support our results of Fig. 3 of the main text with the corresponding held out data likelihood scores. Fig. 3 summarizes the results - VLAD shows competitive performance.

Additional results for varying Dirichlet prior In Figure 4 we demonstrate held out data likelihood corresponding to experiments of Fig. 4 of the main text with. We see that VLAD performs well in the whole range of analyzed values and likelihood kernels.

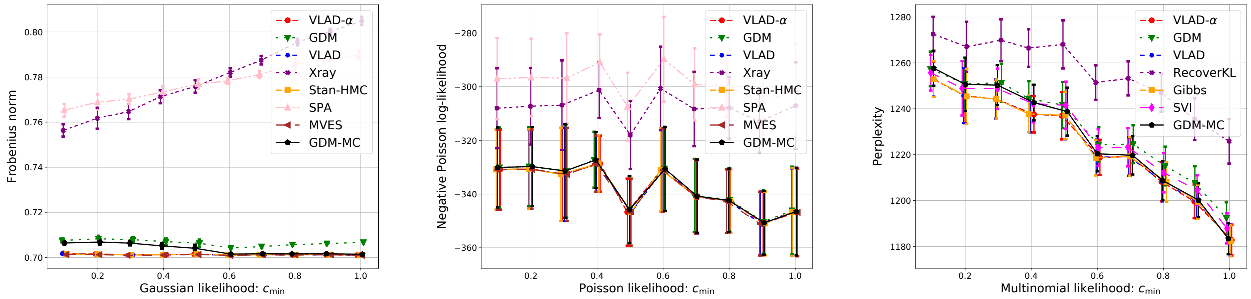
Data generation for simulations studies For all experiments, unless otherwise specified, we set $D = 500, K = 10, \alpha = 2, n = 10000$ (for LDA vocabulary size $D = 2000$). To generate DSN extreme points, for Gaussian data we sample $\beta_1, \ldots, \beta_K \sim \mathcal{N}(0, K)$; for Poisson data $\beta_1, \ldots, \beta_K \sim \text{Gamma}(\mathbf{1}, K\mathbf{1})$; for the LDA $\beta_1, \ldots, \beta_K \sim \text{Dir}_D(\eta)$ with $\eta = 0.1$. To ensure skewed geometry we further rescale extreme points towards their mean by uniform random factors between 0.5 and 1. To do so first compute the mean of extreme points $C = \frac{1}{K}\sum_k \beta_k$ and then rescale each one with $\beta_k = C + c_k(\beta_k - C)$, where $c_k \sim \text{Unif}(c_{\text{min}}, 1)$. Except for the DSN geometry experiment, we set $c_{\text{min}} = 0.5$.

Then we sample weights $\theta_i \sim \text{Dir}_K(\alpha)$ and data mean $\mu_i = \sum_k \theta_{ik}\beta_k$. For Gaussian data $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$, $\sigma = 1$; for Poisson data $x_i|\mu_i \sim \text{Pois}(\mu_i)$; for LDA we follow standard generating process (Blei et al., 2003) with 3000 words per document. All experiments were run for 20



(a) Frobenius norm for Gaussian data (b) Negative log-likelihood for Poison data (c) Perplexity for LDA data

Figure 2: Held out data performance for increasing sample size n



(a) Frobenius norm for Gaussian data (b) Negative log-likelihood for Poison data (c) Perplexity for LDA data

Figure 3: Held out data performance for varying DSN geometry

repetitions and mean was used in the plots along with half standard deviation error bars.

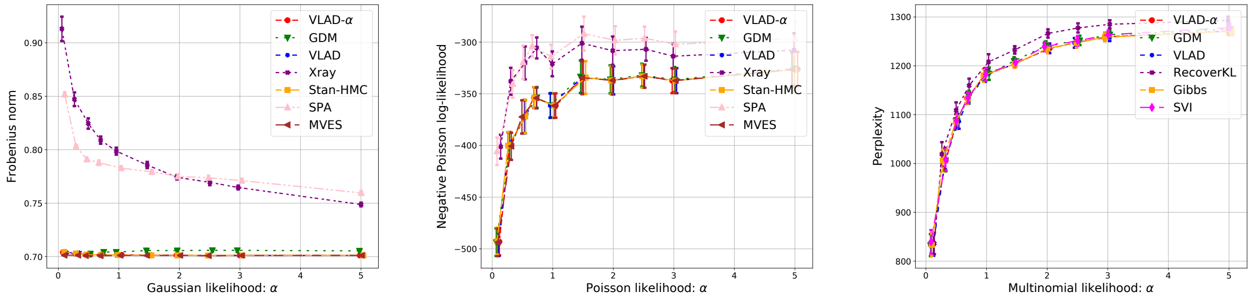
Baseline methods and algorithms setups We considered four separability based NMF algorithms: Xray (Kumar et al., 2013) with code from https://github.com/arbenson/mrnmf; MVES (Chan et al., 2009) with code from http://www.ee.nthu.edu.tw/cychi/source_code_download-e.php; Sequential Projection Algorithm (Gillis & Vavasis, 2014) that we implemented in Python; RecoverKL (Arora et al., 2013) for the LDA case with code from https://github.com/MyHumbleSelf/anchor-baggage.

Bayesian NMF approaches often assume positive weights without the simplex constraint imposed by the Dirichlet prior on weights. Incorporating the simplex constraint complicates the inference (Paisley et al., 2014) as Dirichlet distribution is not conjugate to popular choices of data likelihood such as Gaussian or Poisson. Therefore we are not aware of any implementation for DSN type of models outside of the LDA scenario. We instead chose to compare to automated Bayesian inference methods. We implemented DSN inference with Poison and Gaussian likelihoods in Stan (Carpenter et al., 2017) and considered all three supported estimation procedures: HMC with No U-Turn Sampler (Hoffman & Gelman, 2014), MAP optimization and (Kucukelbir et al., 2017) Automatic Differentiation Variational Inference. MAP optimization and ADVI performed poorly and we did not report their performance. HMC was always trained with true value of α and with knowledge of $\sigma = 1$ for the Gaussian scenario. Number of iterations was set to 80 for $n < 3000$, 60 for $n = 3000$ and 40 for $n > 3000$. We had to restrict number of iterations due to prohibitively long running time (40 iterations for $n = 30000$ took 3.5 hours for Gaussian likelihood and 14 hours for Poisson likelihood; VLAD took 7 seconds in both cases). For the LDA, we used Gibbs sampler (Griffiths & Steyvers, 2004) from https://github.com/lda-project/lda trained for 1000 iterations (1000 iterations for $n = 30000$ took 3.6 hours; VLAD took 3min). Gibbs sampler was trained with true values of α and η. We used Stochastic Variational Inference (Hoffman et al., 2013) implementation from scikit-learn (Pedregosa et al., 2011) and trained it with true values of α and η.

For the Geometric Dirichlet Means (Yurochkin & Nguyen, 2016) we used implementation from https://github.com/moonfolk/Geometric-Topic-Modeling with 8 K-means restarts and ++ initialization.

VLAD was implemented in Python using numpy SVD package and scikit-learn (Pedregosa et al., 2011) K-means clustering with 8 restarts and ++ initialization. The code is available at https://github.com/moonfolk/VLAD.

For the NYT data https://archive.ics.uci.



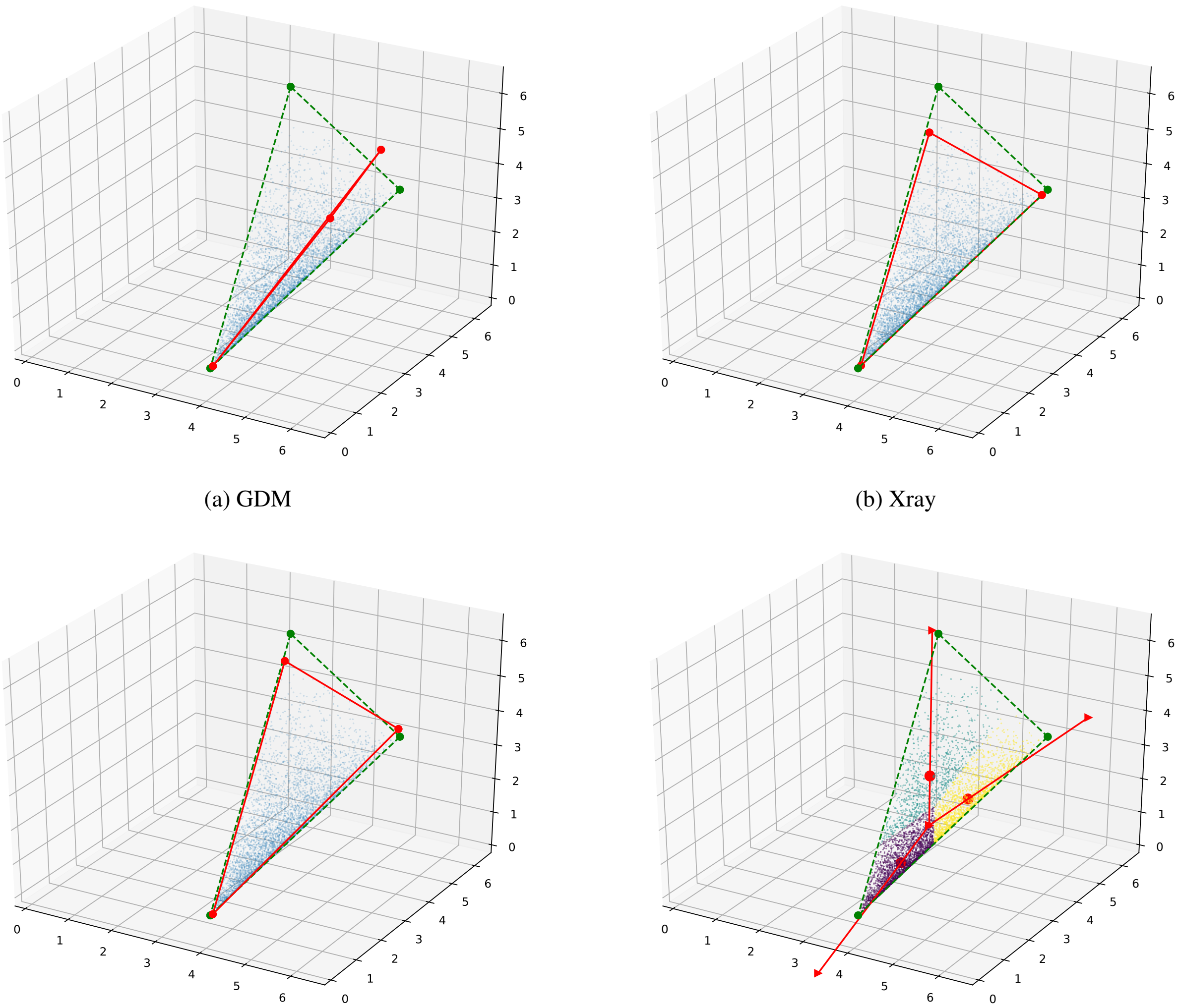
(a) Frobenius norm for Gaussian data (b) Negative log-likelihood for Poison data (c) Perplexity for LDA data

Figure 4: Held out data performance for increasing α

`edu/ml/datasets/bag+of+words` we trained Gibbs sampler with $\alpha = 0.1$ and $\eta = 0.1$ for 1000 iterations and SVI with default settings. For the stock data we trained HMC for 100 iterations with $\alpha = 0.05$.

5. On asymmetric Dirichlet prior

In our work we assumed that $\theta_i \sim \mathrm{Dir}_K(\alpha)$, where $\alpha \in \mathbb{R}_+$. When α is a scalar, the corresponding Dirichlet distribution is referred to as symmetric. More generally, $\alpha \in \mathbb{R}_+^K$ is a vector of parameters. Our algorithmic guarantees, such as alignment of CVT centroids of $\mathscr{B}$, extreme points and centroid of $\mathscr{B}$ and equivalence of extension parameters for all extreme points directions, fail for the general asymmetric case. Wallach et al. (2009) showed that more careful treatment of the parameter α can improve the quality of the LDA topics. Geometric treatment of the asymmetric Dirichlet distribution remains to be the question of future studies. To facilitate the discussion, here we visualize the problem using toy $D = 3, K = 3$ example (similar to Fig. 1 of the main text) with $\alpha = (0.5, 1.5, 2.5)$. Results of the four different algorithms are shown in Fig. 5. Note that for VLAD (Fig. 5d) we only show the directions of the line segments of the obtained sample CVT centroids and the data center, since we do not have a procedure for extension parameter estimation in the asymmetric Dirichlet case. We see that all of the algorithms fail with various degrees of error and notice that the directions obtained by VLAD no longer appear consistent, however do not deviate drastically from the truth. We propose to call such toy triangle experiment a *triangle test* and hope to "pass" the asymmetric Dirichlet *triangle test* in the future work.



(a) GDM

(b) Xray

(c) HMC

(d) VLAD

Figure 5: Asymmetric Dirichlet toy simplex learning: $n = 5000, D = 3, K = 3, \alpha = (0.5, 1.5, 2.5)$

References

Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288, 2013.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022, March 2003.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Chan, T.-H., Chi, C.-Y., Huang, Y.-M., and Ma, W.-K. A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing. *IEEE Transactions on Signal Processing*, 57(11):4418–4432, 2009.

Gillis, N. and Vavasis, S. A. Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, April 2014. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2013.226.

Griffiths, T. L. and Steyvers, M. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

Hoffman, M. D. and Gelman, A. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1): 430–474, 2017.

Kumar, A., Sindhwani, V., and Kambadur, P. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *International Conference on Machine Learning*, pp. 231–239, 2013.

Paisley, J. W., Blei, D. M., and Jordan, M. I. Bayesian nonnegative matrix factorization with stochastic variational inference., 2014.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Pollard, D. A central limit theorem for k-means clustering. *The Annals of Probability*, 10(4):919–926, 1982.

Villani, C. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wallach, H. M., Mimno, D. M., and McCallum, A. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pp. 1973–1981, 2009.

Yurochkin, M. and Nguyen, X. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.