Notes on Blum and Haghtalab (2016)

Karl Stratos

1 Setting

Draw a distribution over k topics $w \in \Delta^{k-1}$ from an unknown prior $\mathcal{P}$. Draw a document consisting of two paragraphs $(x^1, x^2) \in \mathbb{R}^n \times \mathbb{R}^n$ from an unknown distribution $\mathcal{D}^w$ over $\mathcal{X}^w \times \mathcal{X}^w$, where $\mathcal{X}^w \subseteq \mathbb{R}^n$ is the set of paragraphs that have the topic distribution w according to some membership function $f : \mathbb{R}^n \to \Delta^{k-1}$,

$$\mathcal{X}^w := \{x \in \mathbb{R}^n : \; f(x) = w\}$$

Note that $\mathcal{D}^w$ permits arbitrary correlation between paragraphs and feature dimensions (but there will be an additional assumption made on $\mathcal{D}^w$ later for a technical reason). The goal is to learn f given samples from $\mathcal{D}^w$.

2 Approach

Much of the paper assumes that f is linear. There is some $V = [v_1 \ldots v_k]^\top \in \mathbb{R}^{k \times n}$ such that the topic distribution of any paragraph x is given by

$$f(x) = Vx$$

where $f_i(x) = v_i^\top x$ is the probability of the i-th topic in x. The vectors $v_1 \ldots v_k$ are assumed to be linearly independent.

2.1 Subspace Identification

The plan is to identify $\mathrm{span}\,\{v_1 \ldots v_k\} \subseteq \mathbb{R}^n$ and discover that projecting paragraphs onto this subspace reveals their latent topics. Since the two paragraphs in a document (x^1, x^2) are drawn from the same w, they must satisfy

$$Vx^1 = Vx^2$$

It follows that $\mathrm{span}\,\{v_1 \ldots v_k\} \subseteq \mathrm{null}\left\{x^1 - x^2 : (x^1, x^2) \sim \mathcal{D}^w\right\}$, and thus the dimension of $\mathrm{span}\left\{x^1 - x^2 : (x^1, x^2) \sim \mathcal{D}^w\right\}$ is at most $n - k$. An extra assumption is imposed on $\mathcal{D}^w$ to ensure that the dimension is exactly $n - k$: for any subspace $Z \subseteq \mathrm{span}\left\{x^1 - x^2 : \; (x^1, x^2) \sim \mathcal{D}^w\right\}$ with $\dim(Z) < n - k$, there is some chance $\xi > 0$ that $x^1 - x^2$ will fall outside Z. This gives us

$$\mathrm{span}\,\{v_1 \ldots v_k\} = \mathrm{null}\left\{x_i^1 - x_i^2 : i \leq O\left(\frac{n-k}{\xi} \log \frac{n}{\delta}\right)\right\}$$

with probability at least $1 - \delta$ (Lemma A.1) and thus we can estimate the projection operator onto $\mathrm{span}\,\{v_1 \ldots v_k\}$.

2.2 Convex Hull

Let $\Delta \subset \mathbb{R}^n$ denote the convex hull of $a_1 \dots a_k \in \mathbb{R}^n$. We claim that if $A = [a_1 \dots a_k] := V^+$, then the projection of $x \in \mathcal{X}^w$ onto span $\{v_1 \dots v_k\}$ is in Δ.

Proof. Denote SVD of V as $\widetilde{U}\widetilde{\Sigma}\widetilde{V}^\top$. Then $A = \widetilde{V}\widetilde{\Sigma}^{-1}\widetilde{U}^\top$. Since $\widetilde{V}$ is an orthonormal basis of span $\{v_1 \dots v_k\}$, the projection operator onto span $\{v_1 \dots v_k\}$ is given by AA^+. Thus the projection of any $x \in \mathcal{X}^w$ onto span $\{v_1 \dots v_k\}$ has the form

$$x_{||} = AA^+x = AVx = \sum_{i=1}^{k} (v_i^\top x) a_i$$

The claim follows from the probabilistic structure of $v_i^\top x$. □

To ensure that each vertex a_i is eventually included in projected samples, a purity assumption is added: $P_{w \sim \mathcal{P}}(w = e_i) \geq \xi > 0$ for all $i = 1 \dots k$. Then $a_1 \dots a_k$ are included in $m \geq \frac{1}{\xi} \log \frac{k}{\delta}$ projected samples with probability at least $1 - \delta$.

2.3 Algorithm

Input: $S = \left\{(x_i^1, x_i^2) \sim \mathcal{D}^w\right\}_{i=1}^{m}$ where $w \sim \mathcal{P}$, number of topics k

1. Let $X^1 := [x_1^1 \dots x_m^1]$ and $X^2 := [x_1^2 \dots x_m^2]$.
2. Compute the k left singular vectors $U \in \mathbb{R}^{n \times k}$ of $X^1 - X^2$ corresponding to the smallest k singular values.
3. Obtain a convex hull $S_{||} := \left\{ UU^\top x_i^j \,:\, i \in [m],\ j \in [2] \right\}$.
4. Compute k vertices $A = [a_1 \dots a_k] \in \mathbb{R}^{n \times k}$ of $S_{||}$.

Output: $V = A^+$

It's clear that the algorithm succeeds in recovering V with probability at least $1 - \delta$ given $O\left(\frac{n-k}{\xi} \log \frac{n}{\delta} + \frac{1}{\xi} \log \frac{k}{\delta}\right)$ sample documents.

2.4 Problems with This Approach

We assumed that a paragraph x drawn from $\mathcal{D}^w$ satisfies $w = Vx$. This is an extremely unrealistic assumption. Note that we are enforcing *two* requirements:

1. There is a linear model V returning topic probabilities.
2. There is no noise in samples: *every* sample paragraph x *exactly* satisfies $w = Vx$.

The second is relaxed in the next section (to a limited extent) by occassionally allowing some sampling noise. But the first, which is a much more fundamental issue, is left unaddressed. It undermines the whole premise of the paper that we are dodging the need to make simplifying assumptions in modeling the generative process: we *are* making enormous simplifying assumptions in modeling the discriminative process. A linear model is not even required to produce probabilities let alone capture the discriminative power required to disambiguate topics. A much more realistic choice of f would be a nonlinear function with a softmax on top.

3 Approach with Noise

3.1 Sampling Noise

We now assume that instead of receiving pristine samples $(x^1, x^2) \sim \mathcal{D}^w$, we receive $(\hat{x}^1, \hat{x}^2) \in \mathbb{R}^n \times \mathbb{R}^n$ where

$$(\hat{x}^1, \hat{x}^2) = \begin{cases} (x^1, x^2) & \text{with probability } p_0 \\ (x^1 + e^1, x^2 + e^2) & \text{with probability } 1 - p_0 \end{cases}$$

where $e^1, e^2 \in \mathbb{R}^n$ are independent draws from $\mathcal{N}(0, \sigma^2 I_n)$. Standard concentration results give us that we can still recover span $\{v_1 \dots v_k\}$ even if $p_0 = 0$. Let P be the projection onto span $\{v_1 \dots v_k\}$, Let $\widehat{P}$ be the estimated projection from m samples perturbed by $\mathcal{N}(0, \sigma^2 I_n)$. Given $\epsilon, \delta > 0$, if m is large enough (Lemma A.2),

$$P\left(\left|\left|P - \widehat{P}\right|\right|_2 \leq \epsilon\right) \geq 1 - \delta$$

Denoising. It's possible to identify noisy samples by the following argument. We will assume $\left|\left|P - \widehat{P}\right|\right|_2 \leq \epsilon'/(8M)$ for convenience. Let $z \in \mathbb{R}^n$ be any point at least ϵ' away from Δ. Consider the probability of a sample paragraph x falling in the $(\epsilon'/2)$-ball around z (thus outside Δ) upon projection.

$$P\left(\left|\left|\widehat{P}x - z\right|\right|_2 \leq \frac{\epsilon'}{2}\right) = \underbrace{P\left(x \text{ is not noisy} \wedge \left|\left|\widehat{P}x - z\right|\right|_2 \leq \frac{\epsilon'}{2}\right)}_{[1]} + \underbrace{P\left(x \text{ is noisy} \wedge \left|\left|\widehat{P}x - z\right|\right|_2 \leq \frac{\epsilon'}{2}\right)}_{[2]}$$

As for [1], since we then have Px inside Δ,

$$\left|\left|\widehat{P}x - z\right|\right|_2 \geq ||Px - z||_2 - \left|\left|P - \widehat{P}\right|\right|_2 ||x||_2 > \frac{\epsilon'}{2}$$

Thus [1] is zero and the $(\epsilon'/2)$-ball around z only contains noisy points. As for [2], since the noise $\mathcal{N}(0, \sigma^2 I_n)$ projected onto a k-dimensional subspace by P can be thought of as $\mathcal{N}(0, \sigma^2 I_k)$,

$$\begin{aligned} P\left(x \text{ is noisy} \wedge \left|\left|\widehat{P}x - z\right|\right|_2 \leq \frac{\epsilon'}{2}\right) &\leq P\left(||Px - z||_2 \leq \frac{\epsilon'}{2}\right) \\ &\leq P_{y \sim \mathcal{N}(0, \sigma^2 I_k)}\left(||y||_2 \leq \frac{3\epsilon'}{2}\right) \\ &\leq \delta \in \left[\exp\left(-\frac{k}{16}\right), 1\right] \end{aligned}$$

for $\epsilon' \leq (2\sigma/3)\sqrt{k(1 - \sqrt{(16/k)\log(1/\delta)})}$ (Corollary A.13). Let $\delta = p_0\gamma/4$ where γ is to be defined later. Then by the correspondence between probability mass and number of samples for balls (Lemma A.8), given $m = \Omega(\frac{k}{p_0\gamma} \log \frac{1}{\delta})$ we can claim that any point that has fewer than $mp_0\gamma/2$ neighbors within its $(\epsilon'/2)$-neighborhood is at least ϵ' far from Δ; also, any point $(\epsilon'/4)$ close to a_i is has more than $mp_0\gamma/2$ neighbors within its $(\epsilon'/2)$-neighborhood by Eq. (1).

Let $\widehat{S}_{||}$ denote the set of projected samples after denoising. By the above argument, given enough samples, with high probability each $x_{||} \in \widehat{S}_{||}$ satisfies dist $\left(x_{||}, \Delta\right) \leq \epsilon'$.

3.2 Vertex Noise

We no longer receive samples of pure topics $w = e_i$, but instead assume that

$$P_{w \sim \mathcal{P}}(||w - e_i||_1 \leq \epsilon) \geq g(\epsilon) \qquad \qquad \forall i \in [k]$$

where g is some polynomial function. This still allows us to guarantee that for each a_i, we receive samples close to it. If x is a non-noisy paragraph,

$$||Px - a_i||_2 = \left|\left|\sum_{j=1}^{k}(w_j - [e_i]_j)a_j\right|\right|_2 \leq \sum_{j=1}^{k}(w_j - [e_i]_j)\,||a_j||_2 \leq \frac{\epsilon'}{8}$$

with probability at least $\gamma := g(\frac{\epsilon'}{8\alpha k})$ where we assume bounded $||a_i|| \leq \alpha$ for convenience. For $\left|\left|P - \widehat{P}\right|\right|_2 \leq \epsilon'/(8M)$, we have $\left|\left|\widehat{P}x - a_i\right|\right|_2 \leq \frac{\epsilon'}{4}$ with probability at least $g(\frac{\epsilon'}{8\alpha k})$. Thus for each $i \in [k]$,

$$P\left(\left|\left|\widehat{P}x - a_i\right|\right|_2 \leq \frac{\epsilon'}{4}\right) \geq P\left(x \text{ is not noisy} \wedge \left|\left|\widehat{P}x - a_i\right|\right|_2 \leq \frac{\epsilon'}{4}\right) \geq p_0\gamma \qquad (1)$$

We have shown that given enough samples, with high probability each vertex a_i has some $\hat{a}_i \in \widehat{S}_{||}$ satisfying $||\hat{a}_i - a_i||_2 \leq \epsilon'$.

3.3 Recovering Approximate Vertices

The final recovery step extracts $C \subset \widehat{S}_{||}$ such that

- For each $i \in [k]$, there is some $\hat{a} \in C$ with $||\hat{a} - a_i|| \leq \epsilon'$.
- For each $\hat{a} \in C$, there is some $i \in [k]$ with $||\hat{a} - a_i|| \leq \epsilon'$.

Then we can trivially recover $\hat{a}_1 \dots \hat{a}_k$ such that $||\hat{a}_i - a_i|| \leq \epsilon'$ by clustering C with threshold ϵ' (i.e., put two points in the same cluster iff their distance is at most ϵ') and returning any point from each of the resulting k clusters. For the correctness of this algorithm, we make a separability assumption: $||a_i - a_j|| \geq 3\epsilon'$.

The key for extracting such C is the result that[1]

$$d\left(x_{||}\right) := \text{dist}\left(x_{||}, \text{CH}\left(\widehat{S}_{||} \backslash B_{6r\epsilon'}\left(x_{||}\right)\right)\right)$$

is at least $2\epsilon'$ if $\left|\left|x_{||} - a_i\right|\right| \leq \epsilon'$ for some $i \in [k]$ and less than $2\epsilon'$ otherwise. Then we can easily calculate $C = \left\{x_{||} \in \widehat{S}_{||} : \;\; d\left(x_{||}\right) \geq 2\epsilon'\right\}$.

Proposition 3.1. *Let* $\hat{a}_i \in \widehat{S}_{||}$ *satisfy* $||\hat{a}_i - a_i||_2 \leq \epsilon'$ *for some* $i \in [k]$. *Then* $d(\hat{a}_i) \geq 2\epsilon'$.

Proof.

$$\text{dist}\left(\hat{a}_i, \text{CH}\left(\widehat{S}_{||} \backslash B_{6r\epsilon'}\left(\hat{a}_i\right)\right)\right) \geq \text{dist}\left(\hat{a}_i, \text{CH}\left(\Delta \backslash B_{5r\epsilon'}\left(\hat{a}_i\right)\right)\right) - \epsilon' \geq 2\epsilon'$$

[1] The constant r is defined to be the smallest value such that $\text{dist}\left(a_i, \text{CH}\left(\Delta \backslash B_{r\epsilon'}\left(a_i\right)\right)\right) \geq \epsilon'$. Note that $r \geq 1$. The vertices of Δ are "sharp" if r is small and thus more easily identified.

The first inequality follows from Lemma A.7. The second inequality follows from the fact that

$$\operatorname{dist}\left(\hat{a}_i, \operatorname{CH}\left(\Delta \backslash B_{5r\epsilon'}\left(\hat{a}_i\right)\right)\right) \geq \underbrace{\operatorname{dist}\left(a_i, \operatorname{CH}\left(\Delta \backslash B_{4r\epsilon'}\left(a_i\right)\right)\right)}_{\geq 4\epsilon'} - \underbrace{\left|\left|\hat{a}_i - a_i\right|\right|}_{\leq \epsilon'} \geq 3\epsilon'$$

which can be argued as follows. Let

$$q = \underset{\bar{q} \in \operatorname{CH}\left(\Delta \backslash B_{5r\epsilon'}\left(\hat{a}_i\right)\right)}{\arg\min} \left|\left|\hat{a}_i - \bar{q}\right|\right| \qquad t = \underset{\bar{t} \in \operatorname{CH}\left(\Delta \backslash B_{4r\epsilon'}\left(a_i\right)\right)}{\arg\min} \left|\left|a_i - \bar{t}\right|\right|$$

Note that $\operatorname{CH}\left(\Delta \backslash B_{5r\epsilon'}\left(\hat{a}_i\right)\right) \subseteq \operatorname{CH}\left(\Delta \backslash B_{4r\epsilon'}\left(a_i\right)\right)$ since $B_{4r\epsilon'}\left(a_i\right) \subseteq B_{5r\epsilon'}\left(\hat{a}_i\right)$. Then

$$\begin{aligned} \operatorname{dist}\left(\hat{a}_i, \operatorname{CH}\left(\Delta \backslash B_{5r\epsilon'}\left(\hat{a}_i\right)\right)\right) &= \left|\left|q - \hat{a}_i\right|\right| = \left|\left|q - a_i - \left(\hat{a}_i - a_i\right)\right|\right| \\ &\geq \left|\left|q - a_i\right|\right| - \left|\left|\hat{a}_i - a_i\right|\right| \\ &\geq \left|\left|t - a_i\right|\right| - \left|\left|\hat{a}_i - a_i\right|\right| \\ &= \operatorname{dist}\left(a_i, \operatorname{CH}\left(\Delta \backslash B_{4r\epsilon'}\left(a_i\right)\right)\right) - \left|\left|\hat{a}_i - a_i\right|\right| \end{aligned}$$

□

Proposition 3.2. *Let $\hat{x} \in \widehat{S}_{||}$ satisfy $\left|\left|\hat{x} - a_i\right|\right|_2 \geq 8r\epsilon'$ for all $i \in [k]$. Then $d(\hat{x}) \leq 2\epsilon'$.*

Proof. We know that there is some $x = \sum_{i=1}^k \alpha_i a_i \in \Delta$ such that $||x - \hat{x}|| \leq \epsilon'$, and that there are $\hat{a}_i \in \widehat{S}_{||}$ satisfying $\left|\left|\hat{a}_i - a_i\right|\right|_2 \leq \epsilon'$. We argue that $x' := \sum_{i=1}^k \alpha_i \hat{a}_i$

1. Lies inside $\operatorname{CH}\left(\widehat{S}_{||} \backslash B_{6r\epsilon'}\left(\hat{x}\right)\right)$, and

2. Satisfies $\left|\left|\hat{x} - x'\right|\right|_2 \leq 2\epsilon'$,

which proves the claim. The first property follows because $\hat{a}_i$ are also sufficiently far away from $\hat{x}$:

$$\left|\left|\hat{x} - \hat{a}_i\right|\right| \geq \underbrace{\left|\left|\hat{x} - a_i\right|\right|}_{\geq 8r\epsilon'} - \underbrace{\left|\left|a_i - \hat{a}_i\right|\right|}_{\leq r\epsilon'} \geq 6r\epsilon'$$

so that $\hat{a}_1 \ldots \hat{a}_k \in \widehat{S}_{||} \backslash B_{6r\epsilon'}(\hat{x})$ and thus

$$x' \in \operatorname{CH}\left(\left\{\hat{a}_1 \ldots \hat{a}_k\right\}\right) \subseteq \operatorname{CH}\left(\widehat{S}_{||} \backslash B_{6r\epsilon'}\left(\hat{x}\right)\right)$$

The second property follows easily because

$$\left|\left|\hat{x} - x'\right|\right|_2 \leq \left|\left|\hat{x} - x\right|\right|_2 + \left|\left|x - x'\right|\right|_2 \leq \epsilon' + \sum_{i=1}^k \alpha_i \left|\left|a_i - \hat{a}_i\right|\right|_2 \leq 2\epsilon'$$

□

3.4 Algorithm

Input: access to noisy samples $(\hat{x}_i^1, \hat{x}_i^2) \sim \mathcal{D}^w$ where $w \sim \mathcal{P}$, number of topics k, model parameters r, M, error/confidence parameters ϵ, δ

1. Set $\epsilon' = \frac{\epsilon}{8r}$.

2. Take enough samples $\widehat{S}_1$ to ensure that the projection operator $\widehat{P}$ estimated from $\widehat{S}$ satisfies $\left|\left|P - \widehat{P}\right|\right| \leq \frac{\epsilon'}{8M}$ with probability at least $1 - \frac{\delta}{2}$.

3. Take enough fresh samples $\widehat{S}_2$ to ensure that their denoised projections

   $$\widehat{S}_{||} = \left\{ \widehat{P}\hat{x} :\ \hat{x} \in \widehat{S}_2,\ \left|\left\{x \in B_{\epsilon/16r}(\hat{x})\right\}\right| \geq \frac{p_0\gamma\left|\widehat{S}_2\right|}{2} \right\}$$

   satisfies

   - For all $\hat{x}_{||} \in \widehat{S}_{||}$, we have $\text{dist}(\hat{x}_{||}, \Delta) \leq \epsilon'$,
   - For all $i \in [k]$, we have some $\hat{a}_i \in \widehat{S}_{||}$ such that $||a_i - \hat{a}_i|| \leq \epsilon'$,

   with probability at least $1 - \frac{\delta}{2}$.

4. Construct

   $$C = \left\{x_{||} \in \widehat{S}_{||} :\ \text{dist}\left(x_{||}, \text{CH}\left(\widehat{S}_{||} \backslash B_{6r\epsilon'}\left(x_{||}\right)\right)\right) \geq 2\epsilon'\right\}$$

   and cluster into k groups with threshold ϵ. Assign any point in the i-th group as $\hat{a}_i$ to compute $\widehat{A} = [\hat{a}_1 \ldots \hat{a}_k]$.

Output: $\widehat{V} = \widehat{A}^+$

We use $\epsilon' = \frac{\epsilon}{8r}$ so that any cluster in C only contains points at most $8r\epsilon' = \epsilon$ away from each other (Proposition 3.2). Thus the algorithm returns $\hat{a}_1 \ldots \hat{a}_k$ such that

$$P(||a_i - \hat{a}_i|| \leq \epsilon \quad \forall i \in [k]) \geq 1 - \delta$$

A Lemmas

Lemma A.1. *If $m = O\left(\frac{n-k}{\xi}\log\frac{n}{\delta}\right)$ then*

$$P_{(x_1^1,x_1^2)\dots(x_m^1,x_m^2)\sim\mathcal{D}^w}\left(\dim\left(\operatorname{span}\left\{x_i^1 - x_i^2\right\}_{i=1}^m\right) \geq n-k\right) \geq 1-\delta$$

Proof. Define $Z_j := \operatorname{span}\left\{x_i^1 - x_i^2 : i \leq \frac{j}{\xi}\log\frac{n}{\delta}\right\}$ and claim $P(\dim(Z_j) < j) \leq j\frac{\delta}{n}$ for all $j = 0\dots n-k$.

$$\begin{aligned}
&P(\dim(Z_{j+1}) < j+1) \\
&\leq P\left(\dim(Z_j) < j \wedge \dim(Z_{j+1}) < j+1\right) + P\left(\dim(Z_j) \geq j \wedge \dim(Z_{j+1}) < j+1\right) \\
&\leq P\left(\dim(Z_j) < j\right) + P\left(\dim(Z_j) \geq j \wedge \text{all } \tfrac{1}{\xi}\log\tfrac{n}{\delta} \text{ samples fell in } Z_j\right) \\
&\leq j\frac{\delta}{n} + (1-\xi)^{\frac{1}{\xi}\log\frac{n}{\delta}} \leq (j+1)\frac{\delta}{n}
\end{aligned}$$

Then $P(\dim(Z_{n-k}) < n-k) \leq (n-k)\frac{\delta}{n} < \delta$. □

Lemma A.2. $\left|\left|P - \widehat{P}\right|\right|_2$ *is close to zero with high probability given large enough m.*

Proof. We use Davis-Kahan: for any symmetric $B, \widehat{B} \in \mathbb{R}^{n\times n}$ with eigendecomposition $U\Lambda U^\top$ and $\widehat{U}\widehat{\Lambda}\widehat{U}^\top$ (in descending eigenvalues) such that $P = U_{-k}U_{-k}^\top$ and $\widehat{P} = \widehat{U}_{-k}\widehat{U}_{-k}^\top$,

$$\left|\left|P - \widehat{P}\right|\right|_2 \leq \frac{\left|\left|\widehat{B} - B\right|\right|_2}{\hat{\lambda}_{n-k}}$$

Let $D := X^1 - X^2$ and $\widehat{D} := \widehat{X}^1 - \widehat{X}^2$. We use $B = \mathbf{E}\left[(x^1 - x^2)(x^1 - x^2)\right]$ and

$$\widehat{B} = \frac{1}{m}\widehat{D}\widehat{D}^\top - 2\sigma^2 I_n$$

Note that we work with the bias-corrected estimate $\frac{1}{m}\widehat{D}\widehat{D}^\top - 2\sigma^2 I_n$ instead of $\frac{1}{m}\widehat{D}\widehat{D}^\top$ which is what we actually use to obtain $\widehat{P}$ in the algorithm.[2] But this is fine as they have the same eigenvectors.[3] We can then bound the nominator as

$$\left|\left|\widehat{B} - B\right|\right|_2 \leq \left|\left|\widehat{B} - \frac{1}{m}DD^\top\right|\right|_2 + \left|\left|\frac{1}{m}DD^\top - B\right|\right|_2$$

Each term can be bounded, assuming for convenience that paragraphs have a bounded norm $||x|| \leq M$ (Lemma A.3 and A.4). We can lower bound $\lambda_{n-k}(\frac{1}{m}\widehat{D}\widehat{D}^\top)$ by

[2] The bias comes from the noise covariance. Write $\widehat{D} = D + E$ so that each column of E is distributed as $\mathcal{N}(0, 2\sigma^2 I_n)$ and $(1/m)\mathbf{E}\left[EE^\top\right] = 2\sigma^2 I_n$. Then

$$\mathbf{E}\left[\frac{1}{m}\widehat{D}\widehat{D}^\top\right] = \mathbf{E}\left[\frac{1}{m}DD^\top\right] + 2\sigma^2 I_n$$

[3] This relies on the fact that the noise is spherical: $2\sigma^2 I_n = 2\sigma^2\widehat{U}\widehat{U}^\top$.

observing that

$$\begin{aligned}
&\lambda_{n-k}(B) - \lambda_{n-k}\left(\frac{1}{m}\widehat{D}\widehat{D}^\top\right) \\
&\leq \lambda_{n-k}(B) - \lambda_{n-k}\left(\frac{1}{m}DD^\top\right) + \lambda_{n-k}\left(\frac{1}{m}DD^\top\right) - \lambda_{n-k}\left(\frac{1}{m}\widehat{D}\widehat{D}^\top\right) \\
&\leq \underbrace{\left|\left|B - \frac{1}{m}DD^\top\right|\right|_2}_{\leq \frac{\delta_0}{4}} + \underbrace{\left|\left|\frac{1}{m}DD^\top - \frac{1}{m}\widehat{D}\widehat{D}^\top\right|\right|_2}_{\leq 2\sigma^2 + \frac{\delta_0}{4}}
\end{aligned}$$

where we use the fact that $\left|\lambda_i - \hat{\lambda}_i\right| \leq ||E||_2$. The last terms can be bounded using results in Lemma A.3 and A.4. This gives us

$$\lambda_{n-k}\left(\frac{1}{m}\widehat{D}\widehat{D}^\top\right) \geq \lambda_{n-k}(B) - \left(2\sigma^2 + \frac{\delta_0}{2}\right) \geq 4\sigma^2 + \frac{\delta_0}{2}$$

if we make another assumption that the covariance matrix of $x^1 - x^2$ is well conditioned: $\lambda_{n-k}(B) \geq 6\sigma^2 + \delta_0$. We can now claim that given large enough m, with high probability

$$\left|\left|P - \widehat{P}\right|\right|_2 \leq \frac{\left|\left|\widehat{B} - B\right|\right|_2}{\lambda_{n-k}\left(\widehat{B}\right)} \leq \frac{2\epsilon}{2\sigma^2 + \frac{\delta_0}{2}} \leq \frac{\epsilon}{\delta_0}$$

□

Lemma A.3. *$P(\left|\left|\frac{1}{m}DD^\top - B\right|\right|_2 \geq \epsilon) \leq \delta$ given $m = \Omega\left(\frac{M^4 + M^2\epsilon}{\epsilon^2}\log\frac{n}{\delta}\right)$ samples.*

Proof. Let $d_i = x_i^1 - x_i^2$ denote the i-th column of D and define $S_i := (d_i d_i^\top - B)/m$ so that $\sum_{i=1}^m S_i = \frac{1}{m}DD^\top - B$. Note that $\mathbf{E}[S_i] = 0$. Since $||d_i|| \leq 2M$ and thus $\left|\left|d_i d_i^\top\right|\right|_2 = ||d_i||^2 = 4M^2$,

$$||S_i||_2 = \frac{\left|\left|d_i d_i^\top - B\right|\right|_2}{m} \leq \frac{\left|\left|d_i d_i^\top\right|\right|_2 + ||B||_2}{m} \leq \frac{8M^2}{m}$$

$$\sum_{i=1}^m \left|\left|\mathbf{E}\left[S_i S_i^\top\right]\right|\right|_2 \leq \sum_{i=1}^m \mathbf{E}\left[\left|\left|S_i S_i^\top\right|\right|_2\right] \leq \sum_{i=1}^m \mathbf{E}\left[||S_i||_2 \, ||S_i||_2\right] \leq \frac{64M^4}{m}$$

Matrix Bernstein then gives us

$$P\left(\left|\left|\frac{1}{m}DD^\top - B\right|\right|_2 \geq \epsilon\right) \leq 2n \exp\left(\frac{-\epsilon^2/2}{\frac{64M^4}{m} + \frac{8M^2\epsilon}{3m}}\right) \leq \delta$$

Ignoring constants and solving for m, we have the result. □

Lemma A.4. *$P\left(\left|\left|\widehat{B} - \frac{1}{m}DD^\top\right|\right|_2 \geq \epsilon\right) \leq \delta$ given enough samples.*

Proof sketch. By writing $\widehat{D} = D + E$, we obtain

$$\left|\left|\widehat{B} - \frac{1}{m}DD^\top\right|\right|_2 \leq \left|\left|\frac{1}{m}EE^\top - 2\sigma^2 I_n\right|\right|_2 + \left|\left|\frac{1}{m}DE^\top\right|\right|_2 + \left|\left|\frac{1}{m}ED^\top\right|\right|_2$$

The first term is the estimation error of the covariance matrix of $\mathcal{N}(0, 2\sigma^2 I_n)$ given m samples, which can be bounded using the convergence properties of sample covariance of the Gaussian distribution (Corollary A.6). Each of the remaining two terms can be bounded using Matrix Bernstein. For instance, consider $Z = \frac{1}{m}DE^\top$. Let $S_i = \frac{1}{m}d_i e_i^\top$ where $d_i = x_i^1 - x_i^2$ and $e_i \sim \mathcal{N}(0, 2\sigma^2 I_n)$. Note that $Z = \sum_{i=1}^m S_i$ and $\mathbf{E}[S_i] = \frac{1}{m}\mathbf{E}[d_i]\,\mathbf{E}[e_i]^\top = 0$. Also,

$$||S_i||_2 = \frac{||d_i||\,||e_i||}{m} \leq \frac{2M\sigma\sqrt{n}\log(m/\delta)}{m} =: L$$

for all $i \in [m]$ with probability at least $1-\delta$. We used the fact that $||d_i|| \leq 2M$ and $||e_i|| \geq \sigma\sqrt{n}\log(1/\delta)$ with probability at most δ (Lemma A.10). The matrix variance statistic of Z is bounded as

$$v(Z) := \sum_{i=1}^m ||\mathbf{E}\left[S_i S_i^\top\right]|| \leq \sum_{i=1}^m \mathbf{E}\left[||S_i||^2\right] \leq \frac{4M^2\sigma^2 n(\log(m/\delta))^2}{m}$$

Putting together, we obtain $P(||Z||_2 \geq \epsilon) \leq \delta$ by solving for m that satisfies

$$2n\exp\left(\frac{-\epsilon^2/2}{v(Z) + L\epsilon/3}\right) \approx n\exp\left(\frac{-\epsilon^2 m}{c^2 + c\epsilon}\right) \leq \delta$$

where $c := M\sigma\sqrt{n}\log(m/\delta)$. Assuming that m is sufficiently large and ϵ small, we only consider the dominating terms to have

$$n\exp\left(\frac{-\epsilon^2 m}{c^2}\right) \leq \delta \qquad \Longleftrightarrow \qquad m \geq \frac{1}{\epsilon^2}M^2\sigma n\log\frac{n}{\delta}\left(\log\frac{m}{\delta}\right)^2$$

We can use $m = n^k$ (thus $\log m = k\log n$) for some k since $k\log n = o(n^k)$. Using this, we can simplify this expression to have

$$m = \Omega\left(\frac{M^2\sigma n\ \text{polylog}\frac{n}{\delta}}{\epsilon^2}\right)$$

□

Lemma A.5. *Let $E \in \mathbb{R}^{n\times m}$ where each column is drawn from $\mathcal{N}(0, I_n)$ independently. For any $\epsilon \in [\frac{n}{m} + 2\sqrt{\frac{n}{m}}, 1)$ and $\delta \in (0,1)$, given*

$$m = \Omega\left(\frac{n + \log\frac{1}{\delta} + \sqrt{n\log\frac{1}{\delta}}}{\epsilon^2}\right)$$

samples, we have $P\left(||\frac{1}{m}EE^\top - I_n|| \geq \epsilon\right) \leq \delta$.

Proof. The claim is equivalent to

$$P\left(\frac{\sigma_1^2}{m} \geq \epsilon + 1\right) \leq \delta$$

where σ_1 is the maximum singular value of E. We use Theorem B.3 with $t = \sqrt{(1+\epsilon)m} - \sqrt{m} - \sqrt{n} \geq 0$ to obtain

$$P\left(\frac{\sigma_1^2}{m} \geq \epsilon + 1\right) \leq 2\exp\left(-\frac{t^2}{2}\right) \leq \delta$$

The value of m must satisfy

$$m \geq \left(\frac{\sqrt{n} + \sqrt{2 \log \frac{2}{\delta}}}{\sqrt{1+\epsilon} - 1} \right)^2$$

Use the fact that $\sqrt{1+\epsilon} \geq 1 + \frac{\epsilon}{2} - \frac{\epsilon^2}{4}$ for all $\epsilon \in [0, \infty)$ to instead require m to satisfy

$$m \geq \left(\frac{\sqrt{n} + \sqrt{2 \log \frac{2}{\delta}}}{\frac{\epsilon}{2} - \frac{\epsilon^2}{4}} \right)^2 = \frac{16n + 32 \log \frac{2}{\delta} + 16\sqrt{8n \log \frac{2}{\delta}}}{\epsilon^2 - 4\epsilon^3 + \epsilon^4}$$

Since $\epsilon < 1$, the claim follows.

□

Corollary A.6. *Let $E \in \mathbb{R}^{n \times m}$ where each column is drawn from $\mathcal{N}(0, 2\sigma^2 I_n)$ independently. For any $\epsilon \in [\frac{2\sigma^2 n}{m} + 4\sqrt{\frac{\sigma n}{m}}, 2\sigma^2)$ and $\delta \in (0,1)$, given*

$$m = \Omega \left(\frac{\sigma^4 n + \sigma^4 \log \frac{1}{\delta} + \sqrt{\sigma^2 n \log \frac{1}{\delta}}}{\epsilon^2} \right)$$

samples, we have $P\left(||\frac{1}{m} EE^\top - I_n|| \geq \epsilon\right) \leq \delta$.

Lemma A.7. *If $\hat{x} \in \mathrm{CH}\left(\widehat{S}_{||} \backslash B_{\delta + r\epsilon'}(\hat{a}_i)\right)$, then there is some $x \in \mathrm{CH}\left(\Delta \backslash B_\delta(\hat{a}_i)\right)$ such that $||\hat{x} - x|| \leq \epsilon'$.*

Proof. Express $\hat{x} = \sum_{i=1}^{l} \alpha_i \hat{z}_i$ where $\hat{z}_i$ are vertices in $\mathrm{CH}\left(\widehat{S}_{||} \backslash B_{\delta + r\epsilon'}(\hat{a}_i)\right)$. We know that there exists $z_i \in \Delta$ such that $||z_i - \hat{z}_i|| \leq \epsilon' \leq r\epsilon'$. Thus they cannot be in $B_\delta(\hat{a}_i)$, and we can let $x = \sum_{i=1}^{l} \alpha_i z_i$. Moreover,

$$||\hat{x} - x|| \leq \sum_{i=1}^{l} \alpha_i ||z_i - \hat{z}_i|| \leq \epsilon'$$

□

Lemma A.8 (Claim 4.10 in B&H)**.** *Let D be any distribution over $\mathbb{R}^k$. Let $x_1 \dots x_m \sim D$ independently. Then $m = O(\frac{k}{\gamma} \log \frac{1}{\delta})$ is sufficient so that for any ball $B \subseteq \mathbb{R}^k$,*

- *If $P(x \in B) \geq 2\gamma$, the number of samples inside B is greater than γm,*
- *If $P(x \in B) \leq \gamma/2$, the number of samples inside B is smaller than γm,*

with probability at least $1 - \delta$.

Lemma A.9 (Large deviation)**.** *For $x \sim \mathcal{N}(0, I_n)$,*

$$P\left(||x||_2 \geq \sqrt{n + 3\log\frac{1}{\delta}}\right) \leq \delta \qquad \forall \delta \in (0,1]$$

$$P\left(||x||_2 \leq \sqrt{n - 3\log\frac{1}{\delta}}\right) \leq \delta \qquad \forall \delta \in (e^{-n/3}, 1]$$

Proof. For all $B \geq \sqrt{n}$,

$$\begin{aligned}
P\left(||x||_2^2 \geq B^2\right) &\leq P\left(e^{\lambda ||x||_2^2} \geq e^{\lambda B^2}\right) && \forall \lambda \in \mathbb{R} \\
&\leq \prod_{i=1}^{n} \mathbf{E}_{x_i \sim \mathcal{N}(0,1)}\left[e^{\lambda x_i^2}\right] e^{-\lambda B^2} && \text{Markov and independence} \\
&= (1-2\lambda)^{-n/2} e^{-\lambda B^2} && \forall \lambda \in [0, 1/2) \\
&\leq e^{\frac{n-B^2}{3}} && \text{by choosing } \lambda = \tfrac{1}{3}
\end{aligned}$$

which is at most δ if $B^2 \geq n + 3\log\frac{1}{\delta}$. Thus

$$P\left(||x||_2^2 \geq n + 3\log\frac{1}{\delta}\right) = P\left(||x||_2 \geq \sqrt{n + 3\log\frac{1}{\delta}}\right) \leq \delta$$

The other side can be similarly shown to be

$$P\left(||x||_2^2 \leq n - 3\log\frac{1}{\delta}\right) = P\left(||x||_2 \leq \sqrt{n - 3\log\frac{1}{\delta}}\right) \leq \delta$$

provided that $3\log\frac{1}{\delta} \leq n$. □

Corollary A.10 (Large deviation)**.** *For $x \sim \mathcal{N}(0, \sigma^2 I_n)$,*

$$P\left(||x||_2 \geq \sigma\sqrt{n}\log\frac{1}{\delta}\right) \leq \delta \qquad \forall \delta \in (0, e^{-1}]$$

Proof.

$$\begin{aligned}
P\left(||x||_2 \geq \sigma\sqrt{n}\log\frac{1}{\delta}\right) &\leq P\left(||x||_2 \geq \sigma\sqrt{n\log\frac{1}{\delta}}\right) \\
&= P\left(||z||_2 \geq \sqrt{n\log\frac{1}{\delta}}\right) && z \sim \mathcal{N}(0, I_n) \\
&\leq P\left(||z||_2 \geq \sqrt{n + 2\log\frac{1}{\delta}}\right) && \forall \delta \leq e^{-1}
\end{aligned}$$

The claim follows from Lemma A.9. □

Lemma A.11 (Small deviation)**.** *For $x \sim \mathcal{N}(0, I_n)$ and $\epsilon \in [0, 1]$,*

$$\begin{aligned}
P\left(||x||_2 \geq \sqrt{(1+\epsilon)n}\right) &\leq \exp\left(-\left(\frac{\epsilon^2}{2} + \frac{\epsilon^3}{2}\right)\frac{n}{2}\right) \\
P\left(||x||_2 \leq \sqrt{(1-\epsilon)n}\right) &\leq \exp\left(-\left(\frac{\epsilon^2}{2} + \frac{\epsilon^3}{2}\right)\frac{n}{2}\right)
\end{aligned}$$

Proof. Use the Chernoff trick in Lemma A.9 to obtain

$$P\left(||x||_2^2 \geq (1+\epsilon)n\right) \leq (1-2\lambda)^{-n/2}\exp(-\lambda n(1+\epsilon)) \qquad \forall \lambda < \frac{1}{2}$$

We can easily get the bound $\exp(-\epsilon n/3)$ here by using $\lambda = 1/3$. But to obtain a tighter bound,

$$\begin{aligned}
\left(\frac{\exp(-2\lambda(1+\epsilon))}{1-2\lambda}\right)^{n/2} &\leq \left((1+\epsilon)e^{-\epsilon}\right)^{n/2} && \text{using } \lambda = \frac{\epsilon}{2(1+\epsilon)} \\
&\leq \left((1+\epsilon)\left(1-\epsilon+\frac{\epsilon^2}{2}\right)\right)^{n/2} && \text{using } e^{-x} \leq 1 - x + \frac{x^2}{2} \\
&\leq \exp\left(-\left(\frac{\epsilon^2}{2}+\frac{\epsilon^3}{2}\right)\frac{n}{2}\right)
\end{aligned}$$

The other side is similar. □

Corollary A.12 (Small deviation). *For $x \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\epsilon \in [0, 1]$,*

$$P\left(||x||_2 \leq \sigma\sqrt{(1-\epsilon)n}\right) \leq \exp\left(-\frac{n\epsilon^2}{16}\right)$$

Proof. This follows from Lemma A.11 since

$$\begin{aligned}
P\left(||x||_2 \leq \sigma\sqrt{(1-\epsilon)n}\right) &= P_{z\sim\mathcal{N}(0,I_n)}\left(||z||_2 \leq \sqrt{(1-\epsilon)n}\right) \\
&\leq \exp\left(-\left(\frac{\epsilon^2}{2}+\frac{\epsilon^3}{2}\right)\frac{n}{2}\right) \leq \exp\left(-\frac{n\epsilon^2}{16}\right)
\end{aligned}$$

□

Corollary A.13 (Small deviation). *For $x \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\delta \in [\exp(-k/16), 1]$,*

$$P\left(||x||_2 \leq \sigma\sqrt{\left(1-\sqrt{\frac{16}{k}\log\frac{1}{\delta}}\right)n}\right) \leq \delta$$

B Tools for Spectral Analysis

Theorem B.1 (Davis-Kahan). *Let $B, \widehat{B} \in \mathbb{R}^{n\times n}$ be symmetric with eigendecomposition $B = U\Lambda U^\top$ and $\widehat{B} = \widehat{U}\widehat{\Lambda}\widehat{U}^\top$ where $\Lambda = \operatorname{diag}(\lambda_1 \geq \ldots \geq \lambda_n)$ and $\widehat{\Lambda} = \operatorname{diag}(\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_n)$. For any $1 \leq i \leq j \leq n$,*

$$\left|\left|U_{i:j}U_{i:j}^\top - \widehat{U}_{i:j}\widehat{U}_{i:j}^\top\right|\right|_2 \leq \frac{\left|\left|\widehat{B}-B\right|\right|_2}{\inf\left\{\left|\hat{\lambda}-\lambda\right| : \ \lambda \in [\lambda_j, \lambda_i], \ \ \hat{\lambda} \in (-\infty, \hat{\lambda}_{j+1}] \cup [\hat{\lambda}_{i-1}, \infty)\right\}}$$

In particular, suppose B has rank $n-k$, thus $\lambda_{n-k+1} = \cdots = \lambda_n = 0$. Assume that $\widehat{B}$ has rank at least $n-k$, thus $\hat{\lambda}_{n-k} > 0$. Denote the projection onto $\operatorname{range}(U_{n-k+1:n})$ *by P and* $\operatorname{range}(\widehat{U}_{n-k+1:n})$ *by $\widehat{P}$. Then*

$$\left|\left|P-\widehat{P}\right|\right|_2 \leq \frac{\left|\left|\widehat{B}-B\right|\right|_2}{\hat{\lambda}_{n-k}}$$

Theorem B.2 (Matrix Bernstein)**.** *Let* $Z = \sum_{i=1}^{m} S_i$ *where all* $S_i \in \mathbb{R}^{d_1 \times d_2}$ *are independent,* $\boldsymbol{E}[S_i] = 0$*, and* $||S_i||_2 \leq L$*. Then for all* $\epsilon \geq 0$*,*

$$P(||Z||_2 \geq \epsilon) \leq (d_1 + d_2) \exp\left(\frac{-\epsilon^2/2}{v(Z) + L\epsilon/3}\right)$$

where $v(Z)$ *is the matrix variance statistic of* Z *(Tropp, 2015):*

$$\begin{aligned} v(Z) &:= \max\left\{||\boldsymbol{E}\left[ZZ^\top\right]||, ||\boldsymbol{E}\left[Z^\top Z\right]||\right\} \\ &= \max\left\{\sum_{i=1}^{m} ||\boldsymbol{E}\left[S_i S_i^\top\right]||, \sum_{i=1}^{m} ||\boldsymbol{E}\left[S_i^\top S_i\right]||\right\} \quad \textit{by independence \& centeredness} \end{aligned}$$

Theorem B.3 (Corollary 5.35, Vershynin (2010))**.** *Let* $A \in \mathbb{R}^{N \times n}$ *where* $A_{i,j} \sim \mathcal{N}(0,1)$ *independently. Denote the largest and smallest singular values of* A *by* $s_{\max}(A)$ *and* $s_{\min}(A)$*. Then for every* $t \geq 0$*, with probability at least* $1 - 2\exp(-t^2/2)$*,*

$$\sqrt{N} - \sqrt{n} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + t$$