Leveraging Union of Subspace Structure to Improve Constrained Clustering: Supplementary Material

Anonymous Authors[1]

In this document, we provide the proofs to Theorem 1 and Corollary 1, which appear in Section 3.1 of the main document. We also explain the optional UoS-Explore initialization phase of the SUPERPAC algorithm.

1. Proofs of Technical Results

Theorem 1. *Consider two d-dimensional subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$. Let $y = x + n$, where $x \in \mathcal{S}_1$ and $n \sim \mathcal{N}(0, \sigma^2 I_D)$. Define*

$$\mu(y) = \frac{\operatorname{dist}(y, \mathcal{S}_1)}{\operatorname{dist}(y, \mathcal{S}_2)}.$$

Then

$$\frac{(1-\varepsilon)\sqrt{\sigma^2(D-d)}}{(1+\varepsilon)\sqrt{\sigma^2(D-d) + \operatorname{dist}(x, \mathcal{S}_2)^2}} \leq \mu(y)$$

and

$$\mu(y) \leq \frac{(1+\varepsilon)\sqrt{\sigma^2(D-d)}}{(1-\varepsilon)\sqrt{\sigma^2(D-d) + \operatorname{dist}(x, \mathcal{S}_2)^2}},$$

with probability at least $1 - 4e^{-c\varepsilon^2(D-d)}$, where c is an absolute constant.

Proof. The proof relies on theorem 5.2.1 from (Vershynin, 2016), restated below.

Theorem 2. *(Concentration on Gauss space) Consider a random vector $X \sim \mathcal{N}(0, \sigma^2 I_D)$ and a Lipschitz function $f : \mathbb{R}^D \to \mathbb{R}$. Then for every $t \geq 0$,*

$$\mathbb{P}\left\{|f(X) - \mathbb{E}f(X)| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{\sigma^2 \|f\|_{\mathrm{Lip}}^2}\right),$$

where $\|f\|_{\mathrm{Lip}}$ is the Lipschitz constant of f.

First consider the numerator and note that $y - P_1 y = P_1^{\perp} y \sim \mathcal{N}(0, \sigma^2 P_1^{\perp})$ with

$$\mathbb{E}\left\|P_1^{\perp} y\right\|^2 = \sigma^2(D-d).$$

Let $f(z) = \|Pz\|_2$, where P is an arbitrary projection matrix. In this case, $\|f\|_{\mathrm{Lip}} = 1$, as f is a composition of 1-Lipschitz functions, which is also 1-Lipschitz. Further, by Exercise 5.2.5 of (Vershynin, 2016), we can replace $\mathbb{E}\|X\|_2$ by $\left(\mathbb{E}\|X\|_2^2\right)^{1/2}$ in the concentration inequality. Applying Thm. 2 to the above, we see that

$$\mathbb{P}\left\{\left|\left\|P_1^{\perp} y\right\| - \sqrt{\sigma^2(D-d)}\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{\sigma^2}\right). \tag{1}$$

Similarly, for the denominator, note that $y - P_2 y = P_2^{\perp} y \sim \mathcal{N}(P_2^{\perp} x, \sigma^2 P_2^{\perp})$ with

$$\mathbb{E}\left\|P_2^{\perp} y\right\|^2 = \sigma^2(D-d) + \gamma^2.$$

Since $P_2^{\perp} y$ is no longer centered, we let $g(z) = z + P_2^{\perp} x$, which also has $\|g\|_{\mathrm{Lip}} = 1$. Applying Thm. 2 to the centered random vector $\bar{y} \sim \mathcal{N}(0, \sigma^2 P_2^{\perp})$ with Lipschitz function $h = f \circ g$, we have that

$$\mathbb{P}\left\{\left|\left\|P_2^{\perp} y\right\| - \sqrt{\sigma^2(D-d) + \gamma^2}\right| \geq t\right\} \leq 2\exp\left(-\frac{ct^2}{\sigma^2}\right). \tag{2}$$

Letting $t = \varepsilon\sqrt{\sigma^2(D-d)}$ in (1) and $t = \varepsilon\sqrt{\sigma^2(D-d) + \gamma^2}$ in (2) yields

$$(1-\varepsilon)\sqrt{\sigma^2(D-d)} \leq \left\|P_1^{\perp} y\right\| \leq (1+\varepsilon)\sqrt{\sigma^2(D-d)}$$

and

$$\begin{aligned}(1-\varepsilon)\sqrt{\sigma^2(D-d) + \gamma^2} &\leq \left\|P_2^{\perp} y\right\| \\ &\leq (1+\varepsilon)\sqrt{\sigma^2(D-d) + \gamma^2},\end{aligned}$$

each with probability at least $1 - 2\exp\left(-c\varepsilon^2(D-d)\right)$ (since $\gamma > 0$). Applying the union bound gives the statement of the theorem. □

Corollary 1. *Suppose $x_1 \in \mathcal{S}_1$ is such that*

$$\operatorname{dist}(x_1, \mathcal{S}_2)^2 = \sin^2(\phi_1) + \delta\left(\frac{1}{d}\sum_{i=1}^{d} \sin^2(\phi_i)\right) \tag{3}$$

*Equal contribution [1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

for some small $\delta \geq 0$; that is, x_1 is close to the intersection of $\mathcal{S}_1$ and $\mathcal{S}_2$. Let x_2 be a random point in $\mathcal{S}_1$ generated as $x_2 = U_1 w$ where U_1 is a basis for $\mathcal{S}_1$ and $w \sim \mathcal{N}(0, \frac{1}{d} I_d)$. We observe $y_i = x_i + n_i$, where $n_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2$. If there exists $\tau > 1$ such that

$$\delta < \frac{5}{7} - \frac{1}{\tau}$$

and

$$\tau \left(\sin^2(\phi_1) + \frac{1}{6}\sigma^2 \left(D - d\right) \right) < \frac{1}{d} \sum_{i=1}^{d} \sin^2(\phi_i) \ , \quad (4)$$

that is, the average angle is sufficiently larger than the smallest angle, then

$$\mathbb{P}\left\{\mu(y_1) > \mu(y_2)\right\} \geq 1 - e^{-c\left(\frac{7}{100}\right)^2 ds} - 4e^{-c\left(\frac{1}{50}\right)^2 (D-d)}$$

where $\mu(y)$ is defined as in Thm. 1, c is an absolute constant, and $s = \frac{1}{d}\sum_{i=1}^{d} \sin^2(\phi_i)$.

Proof. We have from Thm. 1 that

$$\mu(y_2) \leq \frac{(1+\varepsilon)\sqrt{\sigma^2(D-d)}}{(1-\varepsilon)\sqrt{\sigma^2(D-d)+\gamma_2^2}}$$

and

$$\frac{(1-\varepsilon)\sqrt{\sigma^2(D-d)}}{(1+\varepsilon)\sqrt{\sigma^2(D-d)+\gamma_1^2}} \leq \mu(y_1)$$

with probability at least $1 - 4e^{-c\varepsilon^2(D-d)}$. Therefore if we get the upper bound of $\mu(y_2)$ to be smaller than the lower bound of $\mu(y_1)$, we are done. Rearranging this desired inequality we see that we need

$$\gamma_1^2 < \beta^4 \gamma_2^2 - (1-\beta^4)\sigma^2(D-d). \quad (5)$$

where $\beta = (1-\varepsilon)/(1+\varepsilon)$. Let ε be such that $\beta^4 = 5/6$, and let $\gamma_1^2 = \sin^2(\phi_1) + \delta s$ as in the theorem. Then we wish to select δ to satisfy

$$\delta < \frac{\frac{5}{6}\gamma_2^2 - \sin^2(\phi_1) - \frac{1}{6}\sigma^2(D-d)}{s}. \quad (6)$$

Applying concentration with γ_2^2, we have that $\gamma_2^2 \geq (1-\xi)^2 s$ with probability at least $1 - e^{-c\xi^2 ds}$ where c is an absolute constant. Therefore taking ξ to be such that $(1-\xi)^2 = 6/7$, we require

$$\delta < \frac{\frac{5}{7}s - \sin^2(\phi_1) - \frac{1}{6}\sigma^2(D-d)}{s} = \frac{5}{7} - \frac{1}{\tau}$$

where we used the definition of τ in the theorem. To quantify the probability we need the appropriate values for ε and ξ; we lower bound both with simple fractions: $1/50 < \varepsilon$ where $((1-\varepsilon)/(1+\varepsilon))^4 = \beta = 5/6$ and $7/100 < \xi$ where $(1-\xi)^2 = 6/7$. Applying the union bound with the chosen concentration values implies that $\mu(y_1) > \mu(y_2)$ holds with probability at least $1 - e^{-c\left(\frac{7}{100}\right)^2 ds} - 4e^{-c\left(\frac{1}{50}\right)^2 (D-d)}$. $\square$

Algorithm 1 UoS-Explore

Input: $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$: data, K: number of subspaces, d: dimension of subspaces, A: affinity matrix, maxQueries: maximum number of pairwise comparisons
Estimate Labels: $\hat{C} \leftarrow$ SpectralClustering(A,K)
Calculate Margin: Calculate margin and set
  $x_\vee \leftarrow \arg\max_{x \in \mathcal{X}} \hat{\mu}(x)$ (most confident point)
Initialize Certain Sets: $Z_1 \leftarrow x_\vee$, $\mathcal{Z} \leftarrow \{Z_1\}$,
  numQueries $\leftarrow 0$, $n_c \leftarrow 1$
while $n_c < K$ **and** numQueries $<$ maxQueries **do**
  Obtain Test Point: Choose x_T as point of maximum margin such that $\hat{C}(x_T) \neq \hat{C}(x \in Z_k)$ for any k. If no such x_T exists, choose x_T at random.
  Assign x_T to Certain Set:
    Sort $\{Z_1, \cdots, Z_{n_c}\}$ in order of most likely must-link (via subspace residual for x_T), query x_T against representatives from Z_k until must-link constraint is found or $k = n_c$. If no must-link constraint found, set $\mathcal{Z} \leftarrow \{Z_1, \cdots, Z_{n_c}, \{x_T\}\}$ and increment n_c.
end while

2. UoS-Explore Algorithm

In this section, we describe the process of initializing the certain sets. Note that this step is not necessary, as we could initialize all certain sets to be empty, but we found it led to improved performance experimentally. A main distinction between subspace clustering and the general clustering problem is that in the UoS model points can lie arbitrarily far from each other but still be on or near the same subspace. For this reason, the Explore algorithm from (Basu et al., 2004) is unlikely to quickly find points from different clusters in an efficient manner. Here we define an analogous algorithm for the UoS case, termed UoS-Explore, with pseudocode given in Algorithm 1. The goal of UoS-Explore is to find K certain sets, each containing as few points as possible (ideally a single point), allowing us to more rapidly assign test points to certain sets in the SUPERPAC algorithm. We begin by selecting our test point x_T as the most certain point, or the point of *maximum* margin and placing it in its own certain set. We then iteratively select x_T as the point of maximum margin that (1) is not in any certain set and (2) has a different cluster estimate from all points in the certain sets. If no such point exists, we choose uniformly at random from all points not in any certain set. This point is queried against a single representative from each certain set according to the UoS model as above until either a must-link is found or all set representatives have been queried, in which case x_T is added to a new certain set. This process is repeated until either K certain sets have been created or a terminal number of queries have been used. As points of maximum margin are more likely to be correctly clustered than other points in the set, we expect that by choosing points whose estimated labels indicate they do not belong to any current certain set, we will quickly find a point with no must-link constraints. In our simulations, we found that this algorithm finds at least one point from each cluster in nearly the lower limit of $K(K-1)/2$ queries on the Yale dataset.

References

Basu, Sugato, Banerjee, Arindam, and Mooney, Raymond J. Active semi-supervision for pairwise constrained clustering. In *Proc. SIAM Int. Conf. on Data Mining*, 2004.

Vershynin, Roman. *A Course in High Dimensional Probability*. 2016. URL www-personal.umich.edu/~romanv/teaching/2015-16/626/HDP-book.pdf.