Improved Bounds for Separating Nearby Clusters with Sum-of-Norms Clustering

Michael Thomas
michael.s.thomas@duke.edu

Summary of the Proposal

The goal of this research project is to improve the distance bound required to recover the clusters from two unit discs next to each other using the sum-of-norms clustering method. This problem has been approached in the continuous setting by Dunlap and Mourrat in [1]. They presented general results for the uniform measure on $B_1(-re_1) \cup B_1(re_1) \subseteq \mathbf{R}^d$ which is essentially the uniform distribution on two unit d-balls centered at $-re_1$ and re_1 respectively. The lower bound they obtained was $r > 2^{1-\frac{1}{d}}$, which is generalizable to higher dimensions d. In particular, I have been told by Professor Dunlap that this bound for unit discs in $\mathbf{R}^2$ could be improved. Perhaps by removing the generalization to d dimensions and focusing on d=2, I can improve upon this specific bound. The paper by Professor Dunlap provides many useful conditions for the proof. Additionally, I can empirically model the results in Python using an existing library by Daniel Duckworth described in this blogpost.

Background

The clustering method I am examining was independently introduced by Pelckmans et al. (2005); Hocking et al. (2011); Lindsten et al. (2011) and is called "convex clustering shrinkage," "clusterpath," or "sum-of-norms (SON) clustering" [2]. The algorithm works by minimizing the following loss function that balances data fidelity and cluster separation

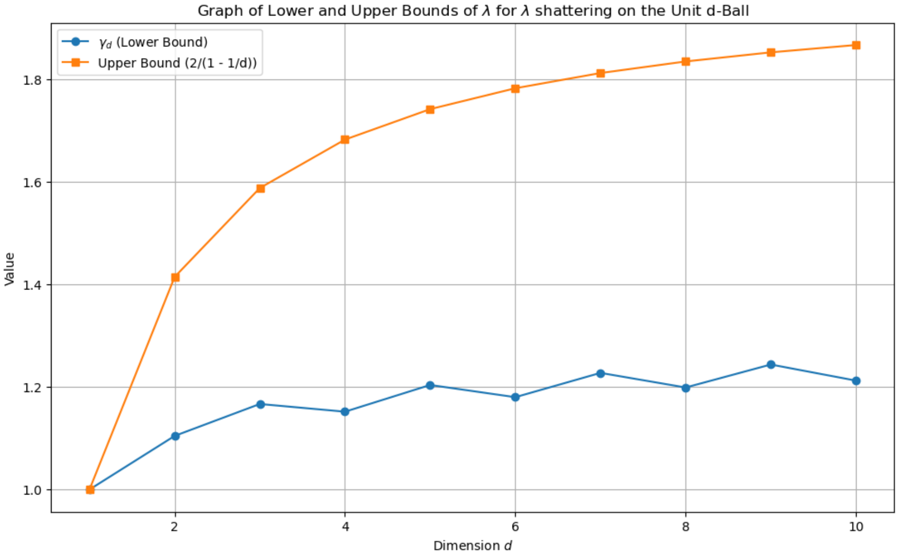
$$\frac{1}{N}\sum_{n=1}^{N}|y_n - x_n|^2 + \frac{\lambda}{N^2}\sum_{k,n=1}^{N}|y_k - y_n|$$

where $x_1, \ldots, x_N \in \mathbf{R}^d$ are datapoints, $y_1, \ldots, y_N \in \mathbf{R}^d$ are the corresponding predictions and the set $\{y_1, \ldots, x_N\}$ has cardinality K. $|\cdot|$ represents the Euclidian norm.

The SON clustering algorithm has several advantages over the more common k-means clustering method implemented with Lloyd's algorithm. Namely, k-means corresponds to non-convex optimization, so there are many local minima. The SON objective function is strongly convex due to the first summation, meaning that there exists a unique optimizer which doesn't depend on an initialization as Lloyd's algorithm does. An important property of SON clustering is that it induces a tree of clusters, meaning that as λ increases, clusters may fuse but they will never break. This property was proven by Chiquet, Gutierrez and Rigaill (CGR) (2017).

For the specific proposition regarding the distance bound for nearby balls, Proposition 5.5 in [1], two conditions are required for λ given $r > 2^{1-\frac{1}{d}}$. The first is that λ must be large enough that the cluster centers at $-re_1$ and re_1 can be separated by SON clustering. Additionally, the parameter λ must be large enough that each ball is not separated into multiple clusters. This leads to the bound that $\lambda \in (2 * 2^{2-\frac{1}{d}}, 2r)$ where $r > 2^{1-\frac{1}{d}}$ for the unit d-balls as described above. In particular, the lower bound on r comes from the upper bound for λ such that the points in each ball are clustered to a single point. In the literature, this term is known as the bound to be λ-cohesive. This bound itself is contained within a range which I have graphed below.

As you can see, at $d = 2$ there is a wide range for what this value might be. Due to this uncertainty, the upper bound for λ to be λ-cohesive is used in Proposition 5.5 for the distance bound on balls. Hopefully the bound can be improved for certain values of d, leading to a better distance bound on the discs. To be specific, I hope to improve on the bounds given in Proposition 4.6 of the paper.



References

[1] Alexander Dunlap and Jean-Christophe Mourrat. Sum-of-norms clustering does not separate nearby balls. *arXiv preprint arXiv:2104.13753*, 2022.

[2] Stephen Vavasis, Tao Jiang, Samuel Tan, and Sabrina Zhai. Recent progress in sum-of-norms clustering. Combinatorics & Optimization, University of Waterloo, 2023.