Recovering imbalanced clusters via gradient-based projection pursuit

Martin Eppert[1], **Satyaki Mukherjee**[2], **Debarghya Ghoshdastidar**[1]

[1] *Technical University of Munich, Germany*
[2] *National University of Singapore, Singapore*

Abstract

Projection Pursuit is a classic exploratory technique for finding "interesting" projections of a dataset. We propose a method for identifying projections containing two imbalanced clusters using a gradient-based technique to optimize the projection index. As sample complexity is a major limiting factor in Projection Pursuit, we analyze our algorithm's sample complexity within a Planted Vector setting where we can observe that imbalanced clusters can be recovered more easily than balanced ones. This differs from previous literature that primarily focuses on the recovery of projections containing symmetric distributions such as balanced clusters [1] and Bernoulli-Rademacher [4]. While some work has addressed recovering skewed directions [3] the sample complexity required for such recovery has not been previously analyzed. Additionally, we give a generalized result that allows for the analysis of the sample complexity for a variety of data distributions and projection indices. We compare these results to computational lower bounds in the Low-Degree-Polynomial framework [2]. While there still is a small gap between the upper and lower bounds, we give empirical evidence that by numerically whitening the data before gradient ascent, the lower bound of the sample complexity can be achieved for large imbalances. Finally, we experimentally evaluate our method's applicability to multi-class classification problems on FashionMNIST.

Keywords

Gradient-Based Methods, Projection Pursuit, Optimization.

Acknowledgements

This work has been supported by the German Research Foundation (DFG) through DFG-ANR PRCI "ASCAI" (GH 257/3-1).

References

[1] Davis, D., Diaz, M. and Wang, K. (2021). Clustering a mixture of gaussians with unknown covariance. *ArXiv Preprint ArXiv:2110.01602.*

[2] Kunisky, D., Wein, A. and Bandeira, A. (2019). Notes on Computational Hardness of Hypothesis Testing: Predictions using the Low-Degree Likelihood Ratio.

[3] Loperfido, N. (2018). Skewness-based projection pursuit: A computational approach. *Computational Statistics And Data Analysis. 120*, 42–57.

[4] Mao, C. and Wein, A. (2022). Optimal Spectral Recovery of a Planted Vector in a Subspace.