Lattice-Based Methods Surpass Sum-of-Squares in Clustering

Ilias Zadik IZADIK@MIT.EDU
Department of Mathematics, Massachusetts Institute of Technology
Min Jae Song MINJAE.SONG@NYU.EDU
Courant Institute of Mathematical Sciences, New York University
Alexander S. Wein AWEIN@CIMS.NYU.EDU
Algorithms and Randomness Center, Georgia Institute of Technology
Joan Bruna BRUNA@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, New York University
Center for Data Science, New York University
Center for Computational Mathematics, Flatiron Institute

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

In this work we show that for an important case of the canonical clustering task of a d-dimensional Gaussian mixture with unknown (and possibly degenerate) covariance, a lattice-based polynomial-time method can provably succeed with the statistically-optimal sample complexity of $d + 1$ samples. This is in contrast with the evidence of "computational hardness" for this task, as suggested by the previously established failure of low-degree methods and the Sum-of-Squares hierarchy to succeed with access to $\tilde{o}(d^{3/2})$ and $\tilde{o}(d^2)$ samples, respectively. [1]

Keywords: statistical-to-computational gaps, lattice basis reduction, sum-of-squares lower bounds, average-case complexity

Clustering is a fundamental primitive in unsupervised learning which gives rise to a rich class of computationally-challenging inference tasks. In this work, we focus on the canonical task of clustering d-dimensional Gaussian mixtures with unknown (and possibly degenerate) covariance. Recent works (Ghosh et al., 2020; Mao and Wein, 2021; Davis et al., 2021) have established lower bounds against the class of low-degree polynomial methods and the sum-of-squares (SoS) hierarchy for recovering the clusters with access to $\tilde{o}(d^2)$ and $\tilde{o}(d^{3/2})$ samples, respectively. Prior work on many similar inference tasks portends that such lower bounds strongly suggest the presence of an inherent statistical-to-computational gap for clustering, that is, a parameter regime where the clustering task is *statistically* possible but no *polynomial-time* algorithm succeeds.

One special case of the clustering task we consider is equivalent to the problem of finding a planted hypercube vector in an otherwise random subspace. We show that, perhaps surprisingly, this particular clustering model *does not exhibit* a statistical-to-computational gap, despite the aforementioned low-degree and SoS lower bounds. To achieve this, we give an algorithm based on Lenstra–Lenstra–Lovász (LLL) lattice basis reduction (Lenstra et al., 1982) which achieves the statistically-optimal sample complexity of $d + 1$ samples, building upon the use of LLL in the seminal papers Lagarias and Odlyzko (1985); Frieze (1986) and in the more recent inference settings (Zadik and Gamarnik, 2018; Gamarnik et al., 2021; Andoni et al., 2017; Song et al., 2021). This result extends the class of problems whose conjectured statistical-to-computational gaps can be "closed" by "brittle" polynomial-time algorithms, highlighting the crucial but subtle role of noise in the onset of statistical-to-computational gaps.

1. Extended abstract. Full version appears as [arXiv:2112.03898,v2].

References

Alexandr Andoni, Daniel Hsu, Kevin Shi, and Xiaorui Sun. Correspondence retrieval. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 105–126. PMLR, 2017.

Damek Davis, Mateo Diaz, and Kaizheng Wang. Clustering a mixture of gaussians with unknown covariance. *arXiv preprint arXiv:2110.01602*, 2021.

Alan M. Frieze. On the Lagarias-Odlyzko algorithm for the subset sum problem. *SIAM J. Comput.*, 15:536–539, 1986.

David Gamarnik, Eren C. Kızıldağ, and Ilias Zadik. Inference in high-dimensional linear regression via lattice basis reduction and integer relation detection. *IEEE Transactions on Information Theory*, pages 1–1, 2021. doi: 10.1109/TIT.2021.3113921.

Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for Sherrington-Kirkpatrick via planted affine planes. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–965. IEEE, 2020.

J. C. Lagarias and A. M. Odlyzko. Solving low-density subset sum problems. *J. ACM*, 32(1): 229–246, January 1985. ISSN 0004-5411. doi: 10.1145/2455.2461. URL https://doi.org/10.1145/2455.2461.

Arjen Klaas Lenstra, Hendrik Willem Lenstra, and László Lovász. Factoring polynomials with rational coefficients. *Mathematische Annalen*, 261(4):515–534, 1982.

Cheng Mao and Alexander S Wein. Optimal spectral recovery of a planted vector in a subspace. *arXiv preprint arXiv:2105.15081*, 2021.

Min Jae Song, Ilias Zadik, and Joan Bruna. On the cryptographic hardness of learning single periodic neurons. *arXiv preprint arXiv:2106.10744*, 2021.

Ilias Zadik and David Gamarnik. High dimensional linear regression using lattice basis reduction. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.