Efficient Sub-Gaussian Mean Estimation for Heavy-Tailed Distributions

Kayleigh Migdol, Professor Pravesh Kothari

Carnegie Mellon University

Problem Description

Sub-Gaussian distributions [1] as the Gaussian distribution are commonly taught, studied, and assumed. Because of the light tails, it is trivial to accurately estimate the mean with bounds by calculating the sample mean. Because of the light tails, it is possible to estimate the mean quite simply with bounds.
While sub-Gaussian distributions are commonly studied, heavier tailed distributions are more common in the real world. For example, the Cauchy and Pareto distributions. In this case, we cannot make the same assumptions (such as finite variance) and therefore cannot use theorems such as Hoeffding's Inequality as they require finite variance. In the event that the variance is undefined or infinity, it is impossible to provide a bound on the emperical mean. For example, in the case of the Cauchy distribution, the sample mean follows a cauchy distribution as well. Because the Cauchy distribution is unbounded, there are no bounds on the sample mean and thus we cannot give any bounds on our error.

Median of Means Method

One surprisingly simple method for this is the median of means [2]. In this, the data is randomly split up into bins and the emperical mean of each bin is calculated. The mean estimate is the median of these means. A bound for this can be found in Theorem 4.1 in the paper cited above. Specifically, they claim that the estimator is sub-Gaussian, allowing it to have the properties and therefore described above. It is able to run in $O(n\log(n))$ time.

MT Method

The MT estimator is proposed by Cherapanamjeri [3]. The algorithm simply bins the data and then takes the mean of each of the bins. The algorithm sets the initial expectation estimate to be the zero vector and then uses gradient descent to calculate the estimate. In order to do estimate the distance from the true mean and the gradient, it uses the optimization problem it names MT. MT specifically attempts to find the direction that maximizes the number of points a certain distance from the mean estimate. This direction is then the gradient used when updating the mean estimate. The process continues for a fixed amount of time. The runtime complexity of the overall algorithm is simply $\tilde{O}(n^4 + n^2 d)$ where n is the number of data points and d is the number of dimensions. A bound for the estimate can be found in Theorem 1 in the paper cited above.

Conclusion

While the mean of sub-Gaussian distributions can be estimated using the sample mean, it is not as simple for heavier tailed distributions. In order to have bounded estimates, more complicated methods must be used such as the median of means and MT methods. While these methods are computationally more expensive (the MT method specifically), it allows us to have bounds on the error in our estimate.

References

[1] Philippe Rigollet.
18.s997: High dimensional statistics lecture notes, 2015.

[2] Luc Devroye, Matthieu Lerasle, Gabor Lugosi, Roberto I Oliveira, et al.
Sub-gaussian mean estimators.
The Annals of Statistics, 44(6):2695–2725, 2016.

[3] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett.
Fast mean estimation with sub-gaussian rates.
arXiv preprint arXiv:1902.01998, 2019.

[4] Roman Vershynin.
High-dimensional probability: An introduction with applications in data science, volume 47.
Cambridge university press, 2018.

Contact Information

- Email: kmigdol@andrew.cmu.edu

Carnegie Mellon University