Coordinated Sampling for Inner Product Sketching: Recent Advances and Research Gaps

Aécio Santos *

CWI, Amsterdam, The Netherlands
aecio.santos@nyu.edu

The inner product is a fundamental operation for measuring the similarity between vectors. Its efficient estimation is a core problem in data analysis, with broad applications ranging from estimating correlations and join sizes in databases to powering machine learning algorithms and AI applications. For decades, the standard for estimating inner products from compressed "sketches" of data has been linear sketching algorithms like CountSketch. These methods provide strong guarantees, typically bounding the estimation error by a factor of the vectors' norms (e.g., $\epsilon\|a\|_2\|b\|_2$), which is effective for dense data.

This talk explores a recent paradigm shift that challenges this long-held standard. We will discuss new findings, initiated by Bessa et al. [1], demonstrating that coordinated weighted sampling can achieve stronger error guarantees, matching the guarantees of linear sketching for dense data and significantly improving it for sparse vectors that are common in many real-world applications. We will primarily cover the simple and computationally efficient algorithms from Daliri et al. [2], which extend threshold and priority sampling to the problem of inner product sketching. We will also review the state-of-the-art theoretical and empirical results of these sampling-based approaches and discuss their modern applications, such as estimating data correlations and discovering joinable tables [3], and well as sparse vector search for retrieval-augmented generation applications [5]. The talk will conclude with a brief discussion of research gaps and ongoing work in this area.

*This work was done while at New York University, in collaboration with Aline Bessa, Majid Daliri, Juliana Freire, Christopher Musco, Haoxiang Zhang, and others.

References

[1] Bessa, A., Daliri, M., Freire, J., Musco, C., Musco, C., Santos, A., & Zhang, H. (2023). Weighted minwise hashing beats linear sketching for inner product estimation. In Proceedings of the 42nd ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (pp. 169-181).

[2] Daliri, M., Freire, J., Musco, C., Santos, A., & Zhang, H. (2024). Sampling Methods for Inner Product Sketching. Proceedings of the VLDB Endowment, 17(9), 2185-2197.

[3] Santos, A., Bessa, A., Chirigati, F., Musco, C., & Freire, J. (2021, June). Correlation sketches for approximate join-correlation queries. In Proceedings of the 2021 International Conference on Management of Data (pp. 1531-1544).

[4] Santos, A., Bessa, A., Musco, C., & Freire, J. (2022, May). A sketch-based index for correlated dataset search. In 2022 IEEE 38th International Conference on Data Engineering (ICDE) (pp. 2928-2941). IEEE.

[5] Bruch, S., Nardini, F. M., Rulli, C., & Venturini, R. (2024). Efficient inverted indexes for approximate retrieval over learned sparse representations. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 152-162).