Corrigendum to *"One-Pass Additive-Error Subset Selection for ℓ_p Subspace Approximation"*

Amit Deshpande ✉
Microsoft Research, Bengaluru, India

Rameshwar Pratap ✉
Indian Institute of Technology, Mandi, H.P., India

This note clarifies certain similarities and differences between Bachem *et al.* [2] and our work [4] that are important but were missing in the related work section of our paper. Bachem *et al.* [2] consider the k-means clustering problem, whereas we consider the ℓ_p subspace approximation problem. k-means++ [1] gives a $O(\log k)$ approximation, in expectation, to the k-means problem using k passes of D^2 sampling. k-means|| [3] gives a $O(1)$ approximation, in expectation, to the k-means problem by using $O(\log n)$ passes of D^2 sampling and picking $O(k)$ points in each pass. They empirically show that only a small constant number of passes suffice in practice. Bachem *et al.* [2] give an MCMC algorithm to reduce the running time of k-means++ from $O(nkd)$ to $O\left(nd + \frac{1}{\epsilon}k^2 d \log \frac{k}{\epsilon}\right)$, while retaining its $O(\log k)$ approximation guarantee within a small additive error. Reducing the number of passes is not the focus of Bachem *et al.* [2], however, a naïve implementation of their algorithm takes two passes, a single-pass pre-processing step followed by a single-pass MCMC subroutine.

We consider adaptive sampling w.r.t. subspaces [5, 6] instead of D^2 sampling w.r.t. subsets [1]. Previous work has shown that $\tilde{O}\left((k/\epsilon)^p\right)$ passes of adaptive ℓ_p norm sampling to pick k points in each pass, gives an additive error guarantee, in expectation, for ℓ_p subspace approximation problem [6]. We give a single-pass algorithm to implement the above while retaining its additive error guarantee. Our algorithm and the proof of its approximation guarantee are inspired by Bachem *et al.* [2]. The steps in the proofs of Lemma 1 and 2 are identical to those in the corresponding lemmas of Bachem *et al.* [2], except the distributions used are different – we consider adaptive sampling w.r.t. subspaces instead of D^2 sampling w.r.t. subsets used in Bachem *et al.* [2]. Our algorithm combines and implements the corresponding pre-processing and MCMC subroutines for adaptive ℓ_p norm sampling in a single pass over the input. Our result implicitly indicates that one can achieve a $O(1)$ approximation with an additive error for the k-means problem by applying a single-pass implementation of Bachem *et al.* [2] to k-means|| instead of k-means++.

References

1 David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM, 2007. URL: `http://dl.acm.org/citation.cfm?id=1283383.1283494`.

2 Olivier Bachem, Mario Lucic, Seyed Hamed Hassani, and Andreas Krause. Fast and provably good seedings for k-means. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 55–63, 2016. URL: `https://proceedings.neurips.cc/paper/2016/hash/d67d8ab4f4c10bf22aa353e27879133c-Abstract.html`.

3 Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *Proc. VLDB Endow.*, 5(7):622–633, 2012. `doi:10.14778/2180912.2180915`.

49th International Colloquium on Automata, Languages, and Programming (ICALP 2022).
Editors: Mikołaj Bojańczyk, Emanuela Merelli, and David P. Woodruff;
Article No. 51; pp. 51:1–51:2

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

4 Amit Deshpande and Rameshwar Pratap. One-pass additive-error subset selection for p subspace approximation. In Mikolaj Bojanczyk, Emanuela Merelli, and David P. Woodruff, editors, *49th International Colloquium on Automata, Languages, and Programming, ICALP 2022, July 4-8, 2022, Paris, France*, volume 229 of *LIPIcs*, pages 51:1–51:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. doi:10.4230/LIPIcs.ICALP.2022.51.

5 Amit Deshpande, Luis Rademacher, Santosh S. Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. *Theory Comput.*, 2(12):225–247, 2006. doi:10.4086/toc.2006.v002a012.

6 Amit Deshpande and Kasturi Varadarajan. Sampling-based dimension reduction for subspace approximation. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, STOC '07, pages 641–650, New York, NY, USA, 2007. Association for Computing Machinery. doi:10.1145/1250790.1250884.