Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Bc. Tomáš Domes	
Thesis title	Streaming Algorithms for Estimating Quantiles with Novel Error Guarantees	
Year submitted	2024	
Subject area	Computer Science - Discrete Models and Algorithms	
Review author	Pavel Veselý	Advisor
Department	Computer Science Institute of Charles University	

Thesis contributions. The thesis studies streaming algorithms for estimating data distributions, captured by ranks and quantiles (that generalize the median and percentiles). This a significant topic in theory with lots of practical applications. In particular, Tomáš has focused on randomized algorithms with the relative error guarantee, which for a query rank k, requires the algorithm to return any $(k \pm \varepsilon \cdot k)$-th smallest stream item, for an error parameter $\varepsilon > 0$ that determines the space used, i.e., the size of the data structure. The state-of-the-art data structure for this problem is ReqSketch (Cormode et al., JACM'23) that uses an intricate sampling strategy, using so-called compactors arranged into a logarithmic number of levels.

The main contribution is a new data structure called JaggedSketch that improves upon ReqSketch in several ways:

- The new sketch has improved error bound for high ranks for which the original ReqSketch analysis is in fact not tight.
- One can specify a (small) set of important ranks, together with their relative importance, and the algorithm will have a substantially better error guarantee for these ranks than that of ReqSketch, while not being worse for other ranks. In the compactor hierarchy, these important ranks form "jags", hence the name of the sketch.
- For a specific setting of parameters, the sketch uses space $O(1/\varepsilon)$ (for constant failure probability), while having a near-relative error. This improves significantly upon the KLL sketch (Karnin et al., FOCS 2016) that also uses space $O(1/\varepsilon)$ but only provides a weaker uniform $\pm n$ error guarantee.

The thesis text, including mathematical proofs, is written very clearly, with all necessary definitions and details. Furthermore, Chapter 1 explains the algorithms from the previous work and the new ideas in the thesis in a concise way, without giving too much detail. Chapter 2 then delves into the analysis of the sketch in the streaming setting, assuming the foreknowledge of the stream length and later showing how to remove the assumption by incurring an additional small space factor.

The improvements in theoretical bounds are supported by an extensive experimental evaluation, using a prototype Python implementation. The results demonstrate that JaggedSketch, when given the same space as ReqSketch, has substantially better error for selected important ranks while not being worse for other ranks. The effect of changing various parameters of the sketch is also evaluated in detail.

The main question left to future work is how exactly to merge JaggedSketches of two datasets and how to analyze the merge operation. However, due to introducing important ranks, such an analysis would be substantially more involved for JaggedSketch than the already intricate analysis of mergeability of ReqSketch.

Let me stress that, except for an initial idea from our consultant Jakub Tětek to use compactors of polynomially decreasing size, all the main theoretical ideas and practical improvements came from Tomáš. He also extended the mathematical analysis of ReqSketch in the streaming setting to JaggedSketch which required overcoming several new technical challenges.

Summary. Overall, this thesis makes a substantial progress in streaming quantile estimation, both theoretical and practical, and the results have a potential to be extended into a paper for a prestigious conference/workshop on algorithms or the theory of database systems (e.g., PODS), where the previous work was presented. Therefore, without any hesitation, I recommend to accept the thesis and suggest grade 1.

September 3, 2024 Pavel Veselý