We thank all the reviewers for their comments and thoughts on the paper. We particularly thank the reviewers for the encouragement provided about the originality of the work, and for recognizing that this work has the potential of initiating the study of many problems in this field. We also hope that this work will encourage the study of many other statistical problems under limited memory constraints, and hopefully develop tools like Fano's inequality that work under memory constraints. We would like to start by addressing the key concern that most reviewers have, which is not showing a lower bound.

Comments on the lower bound: We do believe that the lower bound is not very easy to obtain for this problem, and we have spent time pursuing various approaches to prove a non-trivial lower bound, which were futile. Proving lower bounds even in the streaming model is a hard task, and we believe that the problem is equally hard, if not harder in the statistical setting where we aim to prove trade-off between memory and samples. One of the approaches we pursued was to consider the hard case that was used to prove the sample complexity lower bounds in WuYang, and JiaoHanVenkatWeissman, and reduce the problem to a hypothesis testing problem, which we were unable to prove. We also studied Raz's lower bounds, but currently, do not see a way to apply those techniques for the delicate trade-offs we aim for. In fact, only very recently (COLT 2019, after we submitted the paper), a lower bound on the sample-memory trade-offs were proved in certain regimes, for very specific problem of testing uniformity of distributions. In this light, we expect our work to interest many researchers to study statistical inference tasks under memory constraints, and we hope the reviewers appreciate the same.

We now provide responses to the individual reviewers.

Reviewer 1 Comment : "Overall, the trade-off is not clear ... must be at least Y"."

Please see the discussion above. We will also go through the paper carefully and fix typos.

Reviewer 2 We thank the reviewer for their comments and for pointing out typos. We will fix them.

Reviewer 3 We thank the reviewer for the detailed response and multiple suggestions for improvement.

1) For the lower bound part, please see the discussion above.
2) Extending the algorithm for larger space: We are not sure if there is a trade-off in this case. In fact, one clean open question is to characterize the smallest space needed by an algorithm that is sample optimal, namely uses $k/\varepsilon \log k$ samples. We are unaware of any algorithm that requires a strictly sub-linear space (say $k^{0.1}$ space). Our guess here is that in this case, a sample complexity linear in k is still unavoidable, which can be achieved with constant space using our algorithm. But this again relates to the lower bound part, which we think is hard and leave as future work.
3.1) The ideas in Algorithm 1 as well as the intervals algorithms can indeed be extended to any functional that can be expressed as an expectation of certain functions of the probability mass including power sums. However, we agree the way we select the intervals in this work is more specific to the problem of entropy estimation.
3.2) We agree with the comment, and we did observe that our simple algorithm does need the $\log^2(k)$ term. Consider a distribution that has one element with probability $1/2$, and the remaining elements have probability about $1/2k$. This will correspond to estimating the mean of a distribution that takes values $\log(2)$, and $\log(2k)$ with equal probability, thus requiring $\log^2(k)$ iterations.
3.3) This is a good question, and again relates to the lower bound remark discussed above. In addition, we think that if it is possible to solve general problems with much smaller space, and a tiny overhead in samples, they might be worth considering.
4) This is a great point. We thank the reviewer for pointing this out. We use the clipping step so that the range of our final estimate can be bounded better with fewer iterations. If we use the unclipped estimate as our final estimate, we would have to bound the concentration of the unclipped estimate with more interations. For example, in the two-intervals algorithm, because of the clipping step the range is reduced to $\log(N_i) - \log(1/4l_i) = \log(4N_il_i)$. The range of the unclipped estimate can only be bounded by $\log(N_i) - \log(N_i/(N_i + 1)) = \log(N_i + 1)$. This might be a trick that is only useful in the proof but currently we don't see an easy way to bound the concentration of the unclipped estimate without using super-linear number of samples. We will further investigate this line of thought.
5) The suggested algorithm would still work, but analyzing the performance would require more considerations because estimates from one interval are now dependent on the estimates from the other intervals. We agree that our algorithms are not sample efficient in terms of constants. However we would like to emphasize that our algorithms have the same sample complexity as that of the improvements suggested.

We thank the reviewers again. Hopefully we have answered their questions and concerns. We hope that the new line of questions proposed in this work takes precendence in the decision over the shortcomings.