NeurIPS5595 author response: WOR and p's: Sketches for ℓ_p-Sampling Without Replacement

We appreciate your careful feedback on our submission. We will use your concerns and suggestions to improve the presentation to make it more accessible. Below we address the specific points raised by the reviewers.

R1: Though the problem setting may seem specific at first glance, we believe that the contribution is of broad interest:

- Weighted random sampling has been studied intensively for decades and applied across many disciplines. Without-replacement sampling has also been extensively studied and used across disciplines to improve performance on skewed data sets, so is not "fringe."
- Sampling is important in machine learning applications and optimization, for example, in learning algorithms that use importance sampling to select training examples or features. Performing aggregations efficiently on data that does not fit in memory (streaming, distributed, federated) is important. Data commonly lives on multiple devices, servers, cloud components, etc.
- The powers in the range $[0, 2]$ are important, well-studied, and applied to frequencies (say of training data) in practice. They are related to p-norms. When $p < 1$, high frequencies are mitigated (which is commonly done in weighting training examples by frequency, for example, in word models). The choice $p = 1$ corresponds to sampling by frequency. The choice $p = 2$ emphasizes larger entries. For example, we are likely to sample the "$\sqrt{n}$" in the vector $(1, 1, 1, \ldots, 1, \sqrt{n}, 1, 1, \ldots, 1)$, whereas for $p = 1$ we are not. This can lead to lower variance estimators for functions that are sensitive to large weights.

R1: Our work is novel in broader regimes than stated in the review. Support for negative updates is important, but our work is novel also when we only have positive updates. That is, we provide the first known WOR method for $p > 1$. Prior methods for related problems use different techniques, but we acknowledge that the paper should discuss this in more detail. In particular, prior WOR approaches for $p \leq 1$ are efficient to implement but do not allow samples to be coordinated because the randomization is not reproducible, and they do not support negative updates. Methods based on random projections are asymptotically efficient and support negative updates but have large hidden constants, do not support coordination of samples, and do not support WOR. Our more general problem required a different technique.

R2: The question of scope for NeurIPS/ICML is a fluid one as the conference greatly expanded in recent years. ICML/NeurIPS programs routinely include multiple papers that present or improve fundamental tools that are building blocks in ML implementations: streaming/distributed aggregation, sketching and sampling, matrix computations, and generally optimizing the computation/data transfer of the training process. The program also includes methods that build on these tools. For example, from NeurIPS 2019: (i) Communication-efficient Distributed SGD with Sketching, (ii) Extreme Classification in Log Memory using Count-Min Sketch: A Case Study of Amazon Search with 50M Products, and (iii) Sampling sketches for concave sublinear functions of frequencies. As for our specific contribution, weighted sampling is a fundamental tool that applies to many aspects of ML (this is discussed briefly in the introduction).

R2: Concerning using other heavy hitters methods: we did a proof of concept implementation and so far only used count-sketch, but our use of the HH sketch is indeed a black box. We agree with the reviewer that it would be nice to implement our method with the MG sketch. This will be advantageous in the regime $p \leq 1$ and for positive updates (this is what the MG sketch supports). In this regime it will be much more efficient than count-sketch (which we used in our experiments). Moreover, the error guarantee is for worst-case streams, but MG should be better in practice.

R2: About sampling versus finding heavy hitters: an rHH sketch alone only gives us heavy hitter keys but no information on the tail. Some datasets may not have heavy hitter keys. A weighted sample exposes the respective HH (if there are any) but is much more powerful. In particular, we can obtain unbiased estimates of sum queries and domain queries. For example, to estimate the total "weight" (sum of powers of frequencies) of keys that are not heavy, obtain an unbiased gradient update estimate, or estimate the value of a loss function (or other function of the form of a sum over keys). An HH sketch cannot provide that in general. Our careful scaling approach obtains a weighted random sample by (essentially) turning "random" keys into HH with probability proportional to a power of their frequency. A sample can also be more interpretable than a sketch, providing us representative examples of a dataset.

R3: About the exposition: we will work to make the introduction more accessible. The general area of the paper is streaming/distributed sketching/sampling. We know this community uses a somewhat different language than, say, a probability theorist. However, this is a broad area that is fundamental to efficient computation and with ML applications. In each ML conference there are several papers that use or develop sketching techniques. Our use of p in the context of ℓ_p norms/sampling/HH is standard in that community. The title is meant to allude to the book "War and Peace". Our apologies for the confusion, this can be addressed.