To Reviewer 1

Thanks for the feedback! We would like to stress that our main contribution is a *novel theoretical result* (Theorems 1 and 2) that is of considerable interest to the machine learning community. For this reason, we deliberately structured the paper to highlight the technical analysis in the main body. We believe that proposing a new method which - for the first time - *completely eliminates inversion bias from distributed estimation* is significant enough to stand on its own. However, we do provide numerical experiments plotting the estimation error on four benchmark datasets, all of which clearly support our analysis. One of the plots is shown in Section 1.1, while the remaining plots and a detailed discussion of the experiments is shown in Appendix C. We completely agree that further empirical evaluation is needed, for example incorporating determinantal averaging into several different distributed second-order optimization methods (such as GIANT [WRKXM18] and DANE [SSZ14]) and comparing the effects, however this is beyond the scope of this work. For the final version of this paper, we will be happy to provide a conclusion section at the end, where we will emphasize that our main result holds very generally and is not just applicable to the Newton's method, the main motivation here, but also to other linear functions of inverse Hessians as well as inverse covariance matrices which are of interest in uncertainty quantification.

To Reviewer 3

Thanks for all the comments and suggestions! We will be sure to incorporate them into the final version. Also, we will definitely mention the extra references.

- Regarding line 158 and with/wihout replacement sampling, what we meant is that the sample we use for a single estimator is drawn without replacement. However, as you observed, when there is multiple estimators, the same point may be used in several of them.
- Regarding the proof of Lemma 7, the proof structure is the following: the inductive hypothesis for size n consists of both statements (a) and (b). We then prove statement (a) for size $n+1$ using the inductive hypothesis (employing both (a) and (b) for size n). Finally, we prove statement (b) for size $n+1$ by using (already proven) statement (a) for size $n+1$.
- Thanks for the feedback regarding the proof of Lemma 9. We will clarify it in the final version.

To Reviewer 4

Thanks for the comments and questions! We will address the comments in the final version. We answer your questions below:

- There are methods which sample estimators with probability proportional to certain determinantal weights (those weights are similar to ours, but not identical), such as volume sampling [DW17] (see lines 163-167 in the paper). However, unavoidably those sampling techniques are so computationally expensive that this cost alone makes them impractical in most distributed settings.
- We do not know whether the identities from Section 2 extend to elementary symmetric polynomials, however this is quite possible and it is a very interesting question indeed!

References

[DW17] Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 3087–3096, Long Beach, CA, USA, December 2017.

[SSZ14] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate Newton-type method. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1000–1008, Bejing, China, 22–24 Jun 2014. PMLR.

[WRKXM18] Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W Mahoney. Giant: Globally improved approximate newton method for distributed optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 2332–2342. Curran Associates, Inc., 2018.