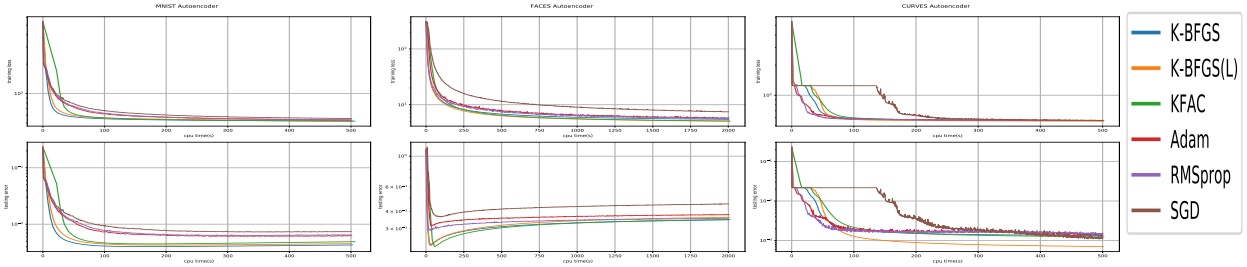
We thank all the reviewers for their time and for raising several interesting questions. We also appreciate that the reviewers carefully read the paper, catching typos, and making useful suggestions. Please see our responses below.

Reviewer #1: The three minor issues raised by the reviewer will be addressed in the revision.

Reviewer #2: In response to [8.Additional feedback], the most significant difference between the proposed methods and a standard BFGS approach is that our approach requires much less storage and work per iteration (see Sec 1 and Tables 1,2). Also, our BFGS updates depend on variables that are computed in the course of computing the stochastic gradients, rather than the gradients themselves. The most challenging part of our work involves the development of our BFGS updates so that they provide good approximations to the inverse of a Kronecker-factored block-diagonal approximation of the Hessian matrix. The relationship of our methods to the Fisher-Rao natural gradient (NG) method is that we approximate the Hessian instead of the Fisher matrix. As in the KFAC method, we use a block-diagonal Kronecker product approximation, where the two sub-blocks in the Kronecker product of each diagonal block is further approximated using BFGS-based updating. We will update the description of our methods in our revision to better explain the connections and differences between them and Fisher-Rao and other BFGS methods. We thank the reviewer for alerting us to the papers on Wasserstein-based NG methods and will include them in the introduction.

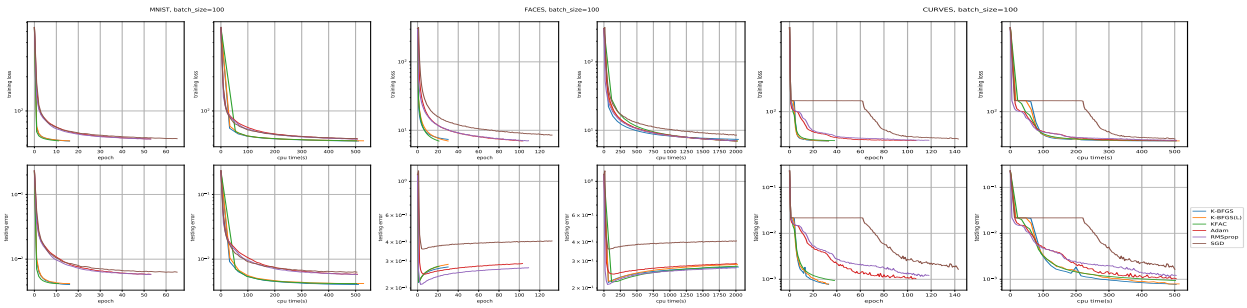
Reviewer #3: The generalization performance of our algorithms is demonstrated in the following figures, which we will add to the paper. The upper (lower) row plots the training loss (testing error - i.e., mean squared error on the test set), respectively. Hyper-parameters (HPs) are set as in Fig 1 of the paper.



Regarding additional HPs, our experiments show that the key HPs for our methods are learning rate and damping constant, as is the case for KFAC and the adaptive gradient methods. Our methods are relatively insensitive to the other HPs, which can be set to default values. We will also highlight how our algorithms are different from related prior work and add more discussion on the structural approximations made and their limitations and validity in a DL context. Please see the response to Reviewer #2.

Our K-BFGS can be extended to other architectures, such as CNNs. Due to both time and space limitations, we focused only on fully-connected NNs, but our preliminary studies indicate that our approach works equally well on CNNs and we are currently working on this extension. We follow standard QN notation, where B denotes the Hessian and H its inverse. A distributed version of the proposed method would definitely be of interest and we are looking into this.

Reviewer #4: We have repeated our experiments using mini-batches of size 100 for all algorithms (see the figures below, where HPs are optimally tuned for batch sizes of 100) and our proposed methods continue to demonstrate advantageous performance, both in training and testing. These results show that our approach works as well for relatively small mini-batch sizes of 100, as those of size 1000, which are less noisy, and hence is robust in the stochastic setting employed to train DNNs.



The reviewer's understanding of the need for two forward-backward passes is correct. We will add this explanation to the paper. Please see the response to Reviewer #2 for our numerical results on testing error.