Response to Reviewer 2

February 11, 2017

We appreciate the helpful feedback from the reviewer. We have addressed your questions and comments. Below we give a point-by-point response to each of the questions:

1. Last paragraph of page 4: you assume that (3) for the training set is strictly convex in θ. I am wondering if the strict convexity assumption would exclude some interesting high-dimensional cases. Can this assumption be removed? Otherwise, it should be listed as a separate condition along with Conditions 1-3 on page 7, as it is a strong condition.

   We thank the reviewer for finding this mistake. The strict convexity assumption is unnecessary, so we have removed it from the paper. Our method only requires the conditions specified in Theorem 1. The conditions are applicable to many popular penalized regression settings, including high-dimensional problems.

2. Page 7, I found the statement of Definition 3 somewhat disconnected with the rest of the paper. For example, you assume the $n \times p$ matrix B has orthonormal columns. Does this require p is smaller than (or equal to) n? Can you modify this definition directly using the subspace defined earlier?

   We apologize for the misleading notation in Definition 3. n is supposed to be an arbitrary number, not the training size. For clarification, we have replaced n with a different variable q.

3. Examples in Section 2.4. It was said that details are included in the Appendix. I found those details in the Appendix are still very brief. Could you provide more details on how to check Conditions 1-3 for each example in the Appendix? And how about the strong convexity assumption? Is it satisfied, too?

   We have included more details in Section A.3 of the Appendix on how to check Conditions 1-3 for each example. We have also removed the strong convexity assumption. We only require invertibility of the Hessian matrix of the training criterion at its minimizer.

4. Examples in Section 3. Could you add false positive and false negative in each of the tables? And how about the computational speed? The p in each of the examples is still small. Could you provide some general comments on how well the new algorithm can handle large p?

   We updated Table 6 in Section 4 with false positive and false negative rates. We also streamlined the simulation results in Section 3. We originally included the percent of correctly identified nonzero features, but this metric is misleading. The goal of minimizing the validation error is to minimize the generalization error, not to model recovery or identification of the nonzero features. We have therefore removed these columns.

   We compare efficiency of the methods by the number of times the methods needed to solve the inner training criterion. We didn't include computation time since it heavily depends upon the solver used to minimize the training criterion.

We have found that our algorithm scales well with p. The reason is that in non-smooth optimization problems, the gradient of the validation loss often only depends on the nonzero features in the model. Therefore as long as the problem is reasonably sparse, the gradient can be computed quickly.