First of all, we would thank all the reviewers for their careful reading and thoughtful comments. We address their comments and questions below. We will prepare our final version based on their comments.

———- Response to Reviewer 1: ———-

- This method seems to only be applicable to disjoint groups of coefficients. Many applications require overlapping groups (e.g.,...). Does this method trivially extend? It does not seem to, at least to me.

Answer: Our method is relatively easy to be extended for overlapping groups by using overlap norm, which associates to the parameter vector a specific decomposition [a]. Since we can still use BCD for the overlap norm with a slight modification, our method is also available to the BCD with the overlapping group.

- This method focuses only on increasing the efficiency of the block-coordinate descent. Other methods, such as iterative thresholding for group-sparse inference, which does not use a full internal block iteration and instead only performs one step of thresholding per block, can be used instead (e.g.,...). Do any of the benefits transfer? Depending on the number of iterations it takes to compute, these methods could be comparable, or be much slower. It would be useful for the authors to compare to these other methods

Answer: Yes, our ideas of the upper bound and the candidate group set is applicable to other methods such as the iterative thresholding with a slight modification of Eq. (6) and Eq. (8). This is because the computations of the bound and the set are based on the difference between the reference parameter vector and the current parameter vector, which can also be handled in the iterative thresholding.

- I would like to see comparisons to other

We compared our method with an iterative thresholding-style algorithm. We shows the result for $\alpha = 0.2$ in Figure 1 as a example. Our method outperformed Iterative Thresholding. The similar results have been observed for other α. We will add all the results to the final version of our paper. We thank the reviewer for this constructive comment.

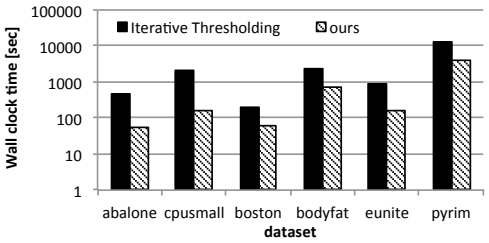

Figure 1: Comparison with iterative thresholding.($\alpha = 0.2$)

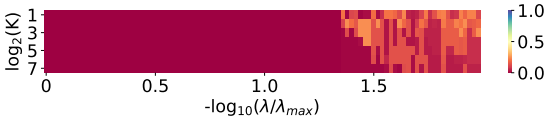

Figure 2: The ratio of active variables during the optimization of our method. K represents the number of the main loop of BCD using the upper bound. ($\alpha = 0.2$)

———- Response to Reviewer 2: ———-

- a figure giving the ratio of activated variable could be useful to compare the selection process (see fig. 4 in [1])

Figure 2 shows the result of the selection process on the boston dataset by following Figure 4 in [1]. K represents the number of the main loop of BCD using the upper bound. Our method effectively skips the variables even if λ is small. We will add the results of other datasets in the final version of our paper. We appreciate your insightful advice.

- clarify at the beginning that the method only consider non-overlapping groups

We will revise the paper to reflect this comment. We thank the reviewer for your important advice.

- what are some possible extension of the proposed algorithm ?

Our method can be flexibly extended to various types of group structured data. For examples, our method can be naturally extended to overlapping groups by using overlap norm [a]. In addition, our method is also applicable for tree-structured data since Tree-structured Group Lasso [b] uses a similar equation to Eq. (4) of SGL.

———- Response to Reviewer 3: ———-

- Provide a more clear proof of Theorem 2. (Related comment: This proof seems to ignore a major difference in the group updating order...)

Theorem 2 holds regardless of the group updating order because it guarantees *the converged value of the objective function* rather than *the converged parameter vector*. Since the problem of SGL is either convex or strongly convex depending on the condition of X, the *optimal value of the objective function* is unique (while *the optimal parameter vector* may be non-unique). In addition, because we assume that the original BCD converges to the solution in the theorem, the BCD using the upper bound also converges even if it starts with the initial parameter resulting from the BCD on the candidate group set. We will add this explanation to the paper in detail.

- Derive a convergence rate and compare that with the original BCD algorithm.

Our method performs the first BCD on the candidate group set and then performs the second BCD on all the groups using the upper bound. We would like to analyze the first BCD in the future work because our own criterion of Eq. (8) is nontrivial for the convergence analysis. However, since the updating order of the second BCD is the same as the original BCD, the second BCD achieves at least the same convergence rate as that of the original BCD with the initial parameter vector resulting from the first BCD.

- State Lemma 4 in a more clear way. (Related question: Does Lemma 4 mean that the candidate group set contains *ALL* groups whose parameters must be nonzeros?)

No, the candidate group set contains a subset of the nonzero groups *identified by using the lower bound of Lemma 3.2* as shown in the proof of Lemma 4. The lower bound confidently identifies groups with *nonzero* vectors while the upper bound of Lemma 1 identifies groups with *zero* vectors. We are sorry for the confusion. We will modify Lemma 4 to add the aforementioned explanation.

[a] L. Jacob, G. Obozinski, J. Vert. Group lasso with overlap and graph lasso. In ICML, 2009.

[b] R. Jenatton, J. Mairal, G. Obozinski, F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. In ICML, 2010.