We sincerely thank the reviewers for their time and constructive comments. We try to focus on one point raised by the reviewers in each paragraph as follows.

The proposed algorithm in this paper is to improve the solution efficiency of the sparse learning problems given by equation (1) in the main file. As discussed at the beginning of the supplemental file, Thunder outperforms existing solvers is mainly because of the passive feature recruiting strategies, sampling method for feature recruiting, and the safe stop condition regarding feature recruiting employed by the algorithm. These strategies can ensure Thunder has a smaller active set during algorithm updating, effective feature screening, efficient feature recruiting, and algorithm safety guarantee. According to our complexity study in Theorem 1 and Section 3.2, maintaining a small active set is crucial to active/working set type of algorithms. The efficiency of Thunder is based on strong theoretical support rather than engineering tricks.

The prediction and feature selection accuracy relies on the selection of λ. This λ selection problem has been studied by the statistics community, and it is beyond the scope of this paper. In this paper, we focus on optimization methodologies that can further scale up the solutions of sparse learning given one particular λ. As the problem is convex, duality gap is usually used to measure the precision of the solutions regarding a particular λ value.

The correlation between features may affect the efficiency of Thunder, but it does not impact the algorithm's safety. Here safety means the algorithm final step active set does not miss any features in the optimal active set $\bar{\mathcal{A}}$ of the problem. According to the derivation in Section 2.1, the stop condition regarding feature recruiting given in Lemma 1 ensures that the final active set is a super set that contains the optimal active set. If the condition in Lemma 1 is not met, Algorithm 2 will not stop feature recruiting. Each operation and updating of Algorithm 2 will decrease (or not change) the duality gap of the original problem, and the problem is convex. The duality gap will become smaller and smaller and then the algorithm can distinguish all active features according to Lemma 1. The safety of the algorithm is guaranteed by the safety of the operation at each step. As shown in the experiments, Thunder can outperform existing solvers on all three large real-world data sets, i.e., Finance, KDD2010, and URL. The results on these real-world data sets prove the advantages and effectiveness of Thunder under different data correlation scores.

According to the proof of Theorem 1, the algorithm complexity is given by $O\left(u\frac{\bar{L}^2}{\gamma^2}\left(\eta\bar{p}\log\frac{\bar{Q}}{\varepsilon_D}+c_1\eta\bar{p}H+|\bar{\mathcal{A}}|\log\frac{\varepsilon_D}{\varepsilon}\right)\right)$. Here $c_1=\frac{1}{H-1}\log\frac{\Pi_{i=1}^{H-1}Q_{h+1}(\beta_h)}{\Pi_{i=1}^{H-1}Q_h(\beta_h)}=\frac{1}{H-1}\log\frac{\Pi_{i=1}^{H-1}Q_{h+1}(\beta_h)}{\Pi_{i=1}^{H-1}\left(Q_h(\beta_{h-1})-K_1d_h\right)}$, and d_h is the average step size of the primal sub-problem. With $\eta=1+\frac{np\varsigma}{uK_1}+\frac{np(1-\varsigma)+p\log p}{uK_1K_2}$, after derivation we can get the optimal approximation of K_1 given by $a\sqrt{np/u}$, and a is a constant value. In the algorithm, we can set K_1 proportional to $\sqrt{np/u}$. Experimentally, the performance of Thunder is not sensitive to the value of K_2. We agree with the reviewers that we will include detailed theoretical analysis as well as the experimental study regarding the selection of K_1 and K_2 in the next version.

Similarly, in our experiments, the feature partition ratio ς does not affect Thunder's performance very significantly. As long as the size of $\mathcal{R}_t^1$ is more than around 1.5 times of $\mathcal{A}_t$, the performance of Thunder does not change a lot regarding ς. Thunder is not very sensitive to either ς or K_2 is because that the operations on $\mathcal{A}_t$ and $\mathcal{R}_t^1$, and the inner loop updating takes the main part of the algorithm. Another reason is that the sampling strategy utilized by Thunder can significantly reduce the feature recruiting and condition checking complexity resulted from the features outside of $\mathcal{A}_t$. The current algorithm complexity analysis in the supplemental file ignores the sampling steps. We will improve the complexity analysis along with the detailed sensitivity study regarding $\mathcal{R}_t^1$ and $\mathcal{R}_t^2$ ratio in the next version.

To recruit an active feature $x_i\in\mathcal{R}_t^1$, we need to evaluate its activity with $|x_i^\top\theta^*|$. However, here θ^* is unknown optimal dual variable, we have to use the current θ_t in hand to approximate the feature's activity. As mentioned above, we employ passive feature recruiting strategies, and it means that we only perform the recruiting operation when we are pretty sure about the features' activity. Give a feature $x_i\in\mathcal{R}_t^1$, if its activity ($|x_i^\top\theta_t|$) lower bound is larger than the upper bounds of most features in $\mathcal{R}_t^1$, we can say that we are confident about its activity, and then move it to the active set $\mathcal{A}_t$. The purpose of the proposed sampling strategy is to reduce the cost induced by the condition checking step in the feature recruiting operation. Instead of comparing the lower bound of $|x_i^\top\theta_t|$ with most feature's upper bound, we do the comparison with a small subset of it. The sampling strategy does not reduce or break the algorithm's accuracy and safety, and it is because that the algorithm's safety is guaranteed by the safe stop condition regarding feature recruiting. We will take the reviewers' suggestions and show more results on the effectiveness of sampling.

We thank the reviewers again for their insightful comments on writing. We will improve the figures, descriptions of the algorithm, term definition, notations, and writing based on their suggestions. We will include the papers listed by the reviewers in the reference.