We thank all reviewers for the constructive comments. Due to the space limitation, we only address major review concerns and will incorporate other suggestions in a revised version of the current work.

R1-Q1: Motivation of using the SoRR loss. Traditional approaches to handle outliers focus on the design of robust *individual losses* that apply to individual training samples, notable examples include the Huber loss and the capped hinge loss as pointed out by the reviewer, or give pre-determined weights to individual training sample. We take a different approach to this work. Instead of changing the definition of individual losses, our method builds in robustness at the aggregate loss level, using the AoRR loss. The resulting learning algorithm is more flexible and allows the user to choose an individual loss form that is relevant to the learning problem. On the other hand, the weights on the individual training sample under AoRR loss is determined automatically.

R1-Q2: Related work of (Chang, Yu, & Yang, KDD 2017). This work addresses the outliers for multiclass SVM by using a hyperparameter to cap the values of the individual losses. This approach is *different* from ours since we directly address the original top-k multiclass SVM problem using our TKML loss *without* introducing this convex surrogate. Furthermore, it has been shown that the loss used in this work is not multi-class top-k consistent as shown in [34], while our TKML loss is consistent (see lines 265-270).

R1-Q3: Comparison with maximum loss. Although the sensitivity to outliers of the maximum loss can be alleviated with the tricks used in [27], our work on the AoRR aggregate loss provides a general mechanism to exclude the influence of potential outliers in the training data. We use the straightforward maximum loss, as it is a special case of the AoRR loss ($m = 0, k = 1$) and can better reflect the continuum of performance change while we adjust k and m. As such, we do not want to give any special treatment to this special case of AoRR loss.

R1-Q4: Performance comparison with the AT$_k$ loss. The advantage of the AoRR aggregate loss is its improved robustness to outliers. On datasets that do not have a significant fraction of outliers, its performance is expected to be similar to the AT$_k$ loss. On the other hand, when the dataset contains outliers, AoRR can significantly outperform existing aggregate losses, as demonstrated on datasets Monk and Phoneme (Table 1, with an average difference/standard deviation in the accuracy of 4.07%/0.05% and 2.22%/0.18%, respectively).

R1-Q5: Value of k when comparing with the AT$_k$ loss. Experiments on the synthetic data aim to show that the AoRR loss can exclude the influence of outliers. To exemplify this case, we chose to have one outlier in the synthetic data, and correspondingly, use $k = 2$. Furthermore, we perform an additional experiment to the request of R1 showing misclassification rate w.r.t different values of k in Figure 1. This result shows that for values other than 2, the AoRR loss still exhibits an advantage over the AT$_k$ loss.

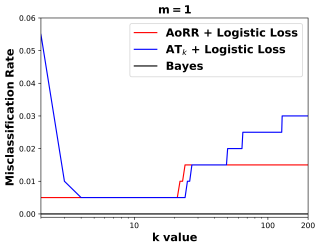

Figure 1: *Misclassification rates of the AT$_k$ loss (blue) and AoRR loss (red) for different k values on a balanced but multi-modal synthetic dataset.*

R1-Q6: Computation complexity of optimizing AoRR. For large datasets, the original DC algorithm may take a longer time to run. However, we can use a stochastic version of DCA to optimize our problem according to Thi, Hoai An Le, et al. "Stochastic DCA for minimizing a large sum of DC functions with application to Multi-class Logistic Regression." arXiv preprint arXiv:1911.03992 (2019). The authors have proved it is much more efficient and the convergence of it to a critical point is guaranteed with probability 1.

R1-Q7: Grid search for m and k. Our experiments are based on a grid search for selecting the value of k and m which may not be ideal in practice for large-scale datasets. For this reason, we provided a well-know adaptive setting method there, and its feasibility had been discussed in [13]. Since k and m are hyper-parameters which are chosen by cross-validation (CV), we did not include the time of searching for them into the overall time complexity, as commonly practiced in machine learning.

R1-Q8: Strategy to choose k. We suggested a strategy to choose k, but for the experiments we performed, we found that a grid search is simpler and often yields comparable performance.

R2-Q1: Novelty of this work. Although Theorem 1 is related with Lemma 1 in [10], there are some fundamental differences. The average top-k function in [10] is convex with regards to its inputs. In our work, we show that the SoRR can be expressed as the difference of two sum of top-k functions, which shows that it is not convex but can be solved with the DC algorithm. This connection has not been studied previously. In addition, based on the SoRR approach, we describe in details of new types of aggregate loss that are more robust to the presence of outliers in training data, and new types of individual losses for multi-class multi-label learning.

R4-Q1: Related work of (Nie, et.al., AAAI 2017). The capped hinge loss in this work is designed as an individual loss. Since the AoRR loss is at the aggregate level, we did not compare it as related work.

R4-Q2: Related work of (Z. Zhou, NSR 2018). The outliers in Figure 3 are different from the case of inaccurate supervision as described in (Z.-H Zhou, NSR, 2018), which assumes errors in the labels of training data. In our study of aggregate losses, we assume the training labels are *accurate* for all training samples, including the outliers.

R4-Q3: Value of k when comparing with the AT$_k$ loss. Please refer to our answers to R1-Q5.