We thank all reviewers for their helpful and constructive comments. We'll further improve in the final version. Below we address their detailed comments.

To R1: Thanks for acknowledging our contributions.

To R2: We disagree on the judgement with our highest respect, due to the nontrivial technical differences and results. In particular, our contributions are: (1) We introduce generalization bounds of learning algorithms on various losses, i.e. HL, SA and HL. Besides, the inequalities between these (actual and surrogate) losses are introduced, which can help the analysis extend to other forms of hypothesis classes; (2) based on the theoretical analysis, we explain the phenomenon when in small label-space case, optimizing HL with its surrogate loss can have better performance on the SA measure than directly optimizing SA with its surrogate loss; and (3) the experimental results support our theoretical analysis.

Technique differences for bounds w.r.t. many measures: The analysis techniques for multi-class classification cannot be trivially extended to multi-label classification because we first need to analyze the relationships between these measures. Besides, it's nontrivial to analyze the relationship between HL and RL, especially for the second inequality (See Lemma 3 in Appendix C.1). Furthermore, as agreed by R4, our analysis can also be extended to other forms of hypothesis classes (e.g. neural networks [15,19]), because the inequalities among these (actual and surrogate) losses are independent of the hypothesis classes. More specifically, for multi-class classification, the performance of algorithms is often evaluated in terms of only one measure (e.g. stand zero-one loss [*1]). Hence, the generalization bound analysis of an algorithm is just provided for the measure [*1, *2, 17] that it aims to optimize. In comparison, for multi-label classification, the performance of algorithms is evaluated in terms of many measures simultaneously, such as HL, SA, and RL. Thus, this requires us to analyze the generalization bounds of an algorithm in terms of other measures in addition to the measure that it aims to optimize. We'll add the discussions in the final version.

Results: With the above analysis techniques, we obtain new theoretical results that are substantially different from the existing ones [11]. More specifically, [11] shows that SA and HL are conflicting measures — algorithms aiming to optimize HL would perform poorly if evaluated on SA, and vice versa. In comparison, we show that when in small label space case, optimizing HL with its surrogate loss can have better performance on the SA measure than directly optimizing SA with its surrogate loss.

To R3: Thanks for acknowledging our novelty and sorry for the unclear parts. We'll make the comparison and statements more precise in the final version. For the clarity of subsequent discussions, we distinguish two somewhat orthogonal approach paradigms [*3] w.r.t. a loss $L^{0/1}$ for MLC: 1) one paradigm first estimates the conditional probability $P(\mathbf{y}|\mathbf{x})$ and gets the classifier by the optimal strategy w.r.t. $L^{0/1}$; 2) the other one directly optimizes $L^{0/1}$ with its surrogate loss to find a classifier in a constrained parametric hypothesis space. In fact, the analysis in [11] is under the first paradigm, while we are under the second one. Below, we discuss the pros and cons of each one in detail. We'll add the discussions in the final version.

Pros and cons of the analysis in [11]: [11] can provide much insight for the first approach paradigm although there is still a gap between the actual $P(\mathbf{y}|\mathbf{x})$ and its estimated one through many parametric methods (e.g., probabilistic classifier chains, etc). In contrast, it may offer less insight for the second paradigm (e.g., binary relevance which directly optimizes HL with its surrogate loss). More specifically, [11] assumes that the hypothesis space is unconstrained to allow $P(\mathbf{y}|\mathbf{x})$ known, and gets the Bayes-optimal classifiers w.r.t. HL (i.e. $\mathbf{h}_H^*$) and SA (i.e. $\mathbf{h}_s^*$) by their corresponding optimal strategy. Then, it analyzes the regret (a.k.a excess risk) upper bounds of $\mathbf{h}_H^*$ and $\mathbf{h}_s^*$ in terms of SA (i.e., Proposition 4) and HL (i.e., Proposition 5) respectively, and finds the bounds are large, which concludes that HL and SA conflict with each other.

Pros and cons of our analysis: Our analysis can provide much insight for the second paradigm. In contrast, it may offer less insight for the first paradigm. More specifically, we have no assumption of the conditional independence of labels, and directly analyze the generalization bounds for the learning algorithms w.r.t. many measures. Although here we consider the kernel-based hypothesis class, which includes the linear and non-linear model by specifying different kernel functions, our analysis can be extended to other forms of hypothesis classes.

To R4: Thanks for acknowledging our contributions. Indeed, the learning algorithms optimize the loss with a surrogate loss rather than the actual one. We'll make this clearer and qualify the related conclusions more appropriately in the final version.

[*1] Lei et al. Multi-class svms: From tighter data-dependent generalization bounds to novel algorithms. NeurIPS 2015.

[*2] Lei et al. Data-dependent generalization bounds for multi-class classification. IEEE Trans. on Information Theory, 65(5):2995-3021, 2019.

[*3] Waegeman et al. On the bayes-optimality of f-measure maximizers. JMLR (15)3333-3388, 2014.