Thank you for helpful comments and suggestions. We will address the concerns raised by the reviewers.

Response to Reviewer#1

(Q1) A scope of negative result is unclear.
(A1) The negative result of the separation-based approach comes from Corollary 5, which indicates a necessary condition for a surrogate loss of the 0-1-c loss to be rejection calibrated. In our paper, we demonstrate that if a surrogate loss is MPC or APC using the exponential loss, it will not satisfy this condition and therefore it is not calibrated. Note that our negative result is not limited to our demonstration. If one designs a new surrogate loss, one may use this corollary to verify whether that surrogate loss satisfies this necessary condition and may further check if it is sufficient for rejection calibration with Theorem 4.

We discuss the difficulty to satisfy Corollary 5 when the problem becomes multiclass in Lines 144–158. In Corollary 5, the condition under the supremum and infimum operators is identical, which is $\boldsymbol{\eta}\colon\ \max_y \eta_y = 1 - c$. In the binary case, if $\max_y \eta_y = 0.7$, then the class probability of the other class is uniquely determined as 0.3, since they must sum to one. This makes it easier to tune the hyperparameters to satisfy the necessary condition as illustrated in Eq. (9) for the binary case. On the other hand in the multiclass setting, even if we know $\max_y \eta_y = 0.7$, the class probabilities of the other classes can be *any positive values as long as they are sum to* 0.3. This makes it more difficult to tune the hyperparameters as we also showed in Eq. (9) that no values of α, β can satisfy both equations in (6). We will improve the clarity of our writing in Lines 144–158.

(Q2) Relationship to deep learning literature and the reviewer's suggested work.
(A2) To enable the use of deep learning in multiclass classification with rejection is definitely an important future research direction. We checked Eq. (5) in "Learning Confidence for Out-of-Distribution Detection in Neural Networks" and found that the proposed objective function looks similar to the separation-based approach but still different. The rejection criterion is not discussed but instead they have the confidence value. Moreover, they modified the predictions while training with their Eq.(2) and this mechanism may make the analysis difficult. Without their Eq. (2), with proper formalization of their Eq. (5) as a surrogate loss, and with proper formalization of the rejection criterion, e.g., $c < 0.5$, it should be possible to use our condition. Regarding the problem addressed by "On calibration of modern neural networks", a failure of class probability estimation generally comes from (1) model misspecification, (2) inappropriate surrogate loss, (3) insufficient optimization, and (4) overfitting. Since our excess risk bound holds for general models in learning with reject option, our result suggests that the reason (2) is excluded and other reasons become dominant.
As a future work, it would be interesting to explore the possibility of using our results in the separation-based approach to help design a better method for training two deep neural networks simultaneously for tackling the problem along this line of research more effectively.

Response to Reviewer#2

(Q3) State the impossibility theorem clearly.
(A3) If the impossibility theorem means the negative result of the separation-based approach, please see (A1).

(Q4) Significance of the theory for the confidence-based approach.
(A4) As pointed out, it is not surprising that the class probabilities can be estimated by using strictly proper composite losses. However, to the best of our knowledge, derivation of excess risk bounds has different difficulty, particularly when the multiclass case and/or the reject option are considered. For the OVA loss, even though some parts of the proof (in Appendix A.1) reduces to the same line as that of Yuan+ [29] for the binary case, other parts needs analysis characteristic to the multiclass case. Furthermore, for the CE loss, the derivation of an excess risk bound is much more difficult. This can be seen from not only the fact that the analysis of [9][29] does not apply to the multiclass case, but also from [19] where the difficulty of CE loss for the multiclass case is actually discussed. Although one may argue that the proof techniques we used are not complicated, our results are novel and relevant for multiclass classification with reject option since these surrogate losses are well-known and used in the literature. Apart from our bounds, we are only aware of the bounds by Ramaswamy+ [20], which focused on other surrogate losses specially designed depending on the rejection cost.

(Q5) Experimental results are not impressive.
(A5) The purpose of the experiments is not to show the superiority of new methods, but to verify our theoretical findings. We successfully verify the sub-optimality of the separation-based APC and MPC experimentally as Reviewer#2 mentioned. Furthermore, another important message from the experiments is that CE is a strong baseline in this problem, which the existing theoretically-grounded work of multiclass learning with rejection could not outperform (OVA+hinge by Ramaswamy+ [20]). As one of the references from Reviewer#1 suggested that using the confidence value from CE may not be highly reliable, CE is still the best way to tackle this problem in our experiments. Our experiments help illustrate that there are still many things to be done to advance the research in this direction, which is highly relevant for critical tasks where rejection is preferred over misclassification.