We thank the reviewers for their insightful comments and constructive feedback. As the reviewers mentioned, our work shows the following strengths. (1) The theoretical analysis is “profound and novel” [R3,R4,R5]. (2) Experiments are designed “thoroughly” and “carefully” which “verifies the feasibility” [R3,R4]. (3) The paper is “well-written and organized” [R3,R4]. We will answer the major points below and address all remaining ones in the final version.

[R3]: “For eq. (3) (4) (5), the first item on the right-hand side, $\sqrt{\frac{C(F)}{n_j}}$ or $\frac{C(F)}{\sqrt{n_j}}$?”
• This depends on how $C(F)$ is defined. If it is defined to be the Rademacher complexity, then the former is correct.

[R3]: I suggest the authors to polish up the Figure 1.
• Thanks for the suggestion! We’ll update with a better one for the final version.

[R3]: "“Hinge loss (HG) does not work well with 100 classes”, what you mean by not work well?"
• When trained on CIFAR-100, Hinge loss seems to suffer from optimization issues — the training accuracy is at most about 80%. Thus we didn’t report the test accuracy because the failure here is of a different nature.

[R4,R6]: “It is unclear to me why the loss function (10) enforces the desired margin in (9).”; “Provide a strong justification for the equation (10)”
• The Hinge loss in (10) achieves its minimum value zero only if the margin is at least Δ_y. Recall that the margin is defined to be $\gamma = z_y - \max_{j\neq y} z_j$. Therefore, Hinge loss = $\max\{\Delta_y - \gamma, 0\} = 0$ if and only if $\gamma \geq \Delta_j$. Hinge loss is a standard loss that encourages margins in the context of SVM.[1] We extend it to allow label-dependent margins.

[R4]: “I wonder what exactly is showing in Figure 2.”
• We visualize the distributions of the last-but-one layer of the neural network, which are referred to as the features. Please refer to the details in L230-L235. We will clarify more in the final version.

[R4,R6]: “If the second stage does not move the weight by much, shouldn’t the ERM with LDAM loss work well enough?”; “Provide a better why DRW is important?”
• We believe that the second stage with smaller learning rate serves as a fine-tuning-like process to capture sophisticated details in each class. Thus in the second stage, emphasizing rare examples are important, because without it, the training accuracies for all the classes can not be approximately 100%. (Relatively smaller movements in the second stage could also change the performance by more than a few percents.) With the initial large learning rate in the first stage, by contrast, the network learns the shared patterns/features shared across all tasks, and therefore it would be better to train with all the examples with uniform weights. Such phenomenon/intuitions were also observed[2] and justified in recent works[3]. We realized this from the ablation study in Fig. 6 in Appendix, which shows that the features learned in the first stage with ERM are better than those with re-weighting.

[R5]: "How to decide the hyperparameter C? How is the LDAM-HG-DRS in Table.1 implemented?"
• We tune C as a hyper-parameter for each dataset. In particular, we use $C = 0.5$ for all CIFAR-10 and CIFAR-100 experiments, and $C = 0.3$ for all iNaturalist experiments. Regarding the LDAM-HG-DRS implementation, we follow Eq. (10) to implement Hinge loss. Here DRS means the delayed re-sampling strategy.

[R5]: "CB+Softmax and LDAM seem to be quite similar"; “it seems that the main boost of performance is stemmed from the DRW (deferred re-weighting)”; additional baseline CB+DRW.
• We’d like first to clarify that CB only re-weights the losses, and therefore is a re-weighting scheme more similar to vanilla re-weighting than to LDAM (which is a new loss). DRW, a deferred re-weighting scheme that we proposed, is an improved version of CB or vanilla re-weighting, and is orthogonal to LDAM. In Tab. 2, we see that either using LDAM alone (4th row), or DRW alone (3rd row), on top of the ERM baseline, can outperform prior work. LDAM alone (3.5% improvement) is slightly more useful than DRW alone (2.6%), and together, they give 6.8% improvement. Thus we don’t agree that the main boost stems from DRW. We found CB+DRW does not outperform DRW alone, which also suggests that DRW is a better re-weighting scheme.

[R6]: Test the proposed method for more general machine learning tasks.
• Thank you for your suggestion. We selected these datasets (1) to compare with related works, (2) because they are challenging, (3) because they are representative of ubiquitous real-world dataset imbalance issues. Nonetheless, we add one additional sentiment analysis experiment on the Large Movie Review (IMDB) Dataset, a popular and standard task in NLP. We manually created an imbalanced training set by removing 90% of negative reviews. We train a 2-layer bidirectional LSTM with Adam optimizer. Test accuracy of different methods are listed as follows: ERM: 63.18, Re-weight: 76.34, Re-sample: 73.50, LDAM: 82.16. Thus our conclusions hold on other tasks. We will add this result to the final version of the paper.

[1] Wikipedia contributors. "Hinge loss." Wikipedia, The Free Encyclopedia.

[2] Nakkiran, Preetum, et al. "SGD on Neural Networks Learns Functions of Increasing Complexity."

[3] Li, Yuanzhi, et al. "Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks."