We thank the reviewers for their helpful feedback. We are encouraged that you note AUM's simplicity—"works with any classifier out-of-the-box" (R2), "it can save a lot of computational demands" (R4)—as we believe this differentiates our method from existing ones. We would also emphasize our results on supposedly-clean real-world datasets; for example, a 1.2% reduction in error on CIFAR100 (without synthetic noise) simply by removing data.

It seems that R3, as they admit themself, is "confused" by our submission and contribution. *Clearly our use of the label margin concept is not intended to be a novelty in and of itself, as it has been prominent in the ML literature for decades with uses in hundreds of publications.* We could cite [Wang et al., CVPR 2018] (as suggested by R3) but we find the earlier work more appropriate, like [Vapnik, 1995], [Bartlett, NeurIPS 1997], or [Weinberger and Saul, JMLR 2009]. The novelty of our method is using the label margin as part of an intuitive and reliable metric to identify mislabeled data, which we are the first to do. Additionally, we clearly discuss/compare to Co-Teaching in Sec. 2/Table 1, and note that the other reviewers find our paper "well written" (R4), "enjoyable to read" (R1), and "very clear" (R2). However, we do agree with R3's point concerning the subsampled Clothing1M dataset (see response to R4).

R1. Thank you for your supportive comments and interesting remarks. Per your suggestions, we will discuss the connection between double descent and detecting dataset poisoning in Sec. 5. (*"Effect of data augmentation."*) AUM performs comparably with/without augmentation. On CIFAR10 (40% noise) *with* augmentation, AUM reduces error from 43% to 12%; *without* augmentation, it reduces error from 51% to 20%. (*"Line 266 Why is* 2% *top 1 error not significant?"*) This is a typo; thank you for catching. On ImageNet the standard network achieves a top-1 error of 24.2% (as in Table 3, not 22.2% as in line 226). Thus the difference between AUM and standard training is 0.2%.

R2. Thank you for positive feedback and detailed questions. We hope to address them here and in the camera ready.

Fair comparisons with other methods: Our results in Fig. 3 indicate that AUM identifies mislabeled samples with higher precision and recall than other methods. Therefore, if we use our training procedure (remove identified data, send remaining data to a base learner) AUM should outperform existing identification methods.

"Do the removed samples introduce new problem?": This is an interesting point. Empirically, some classes in e.g. WebVision are less likely to be mislabeled (e.g. GOLDFISH) than others (e.g. WATER OUZEL). The number of samples removed per class varied from 109 to 828. We note that this dataset was already imbalanced (class size ranging from 701 to 5688); therefore it is unlikely AUM introduced a new imbalance problem. We will discuss this more in Sec. 5.

"How to choose a good set of [threshold] samples?": We choose $N/(C+1)$ threshold samples (for C classes) uniformly at random from the training set (see line 135). This way, the threshold class size equals the average class size, though this strategy might need adjustment for extremely imbalanced datasets. We are unclear what you mean by "the assigned logit value of threshold samples will be biased." Threshold samples approximate mislabeled samples, as a large threshold logit cannot be learned through generalization (i.e. no "true" positives exist) and therefore must be memorized.

R3. *"Analyses about the difference AUM and original margin"*: AUM is more robust and consistent than the margin at any given epoch. Averaging across epochs increases the "signal to noise ratio." See Fig. S1, S3 in the appendix.

R4. Thank you for pointing us to Yi and Wu's PENCIL paper! Although quite different from AUM, it is clearly relevant in this context and we will of course include it in the camera ready version.

"The authors claimed that non-uniform label noise is not too common in practice": Sorry, this is not what we meant in line 313 ("this particular high-noise setting is not too common"). We were referring to 40% pair-wise asymmetric noise, which is an extreme and synthetic setting. We completely agree that non-uniform noise does exist, and our method is able to successfully reduce error on real-world (non-uniform) noisy datasets (Table 3).

Supporting the claim "AUM is less prone to confusing difficult samples for mislabeled data": Table 3 provides evidence for this claim, though we agree the text in Sec. 4 should emphasize this more. On CIFAR10/100/ImageNet (without any synthetic noise, Table 3), INCV and DY-Bootstrap achieve worse performance than standard training, suggesting that some of the large-loss samples removed by these methods were actually "good" data. Our method actually improves accuracy over standard training, suggesting that AUM is not removing these difficult (but beneficial) data.

Clothing1M results: On the full Clothing1M dataset, AUM achieves 29.0% error (standard training achieves 31.1%). Originally we trained on a 100K subset due to a limited computational budget; we will update Table 3 with the full dataset results. While AUM does not achieve SOTA on this task (compared to e.g. PENCIL), we emphasize that it consistently improves error both on noisy datasets (WebVision, Clothing1M) and clean datasets (CIFAR, TinyImageNet).

"Data cleansing has been widely exploited in the literature of label noise." We are not sure which "data cleansing" methods you are referring to, but please let us know which additional baselines we should include in the final version. Currently, we compare AUM's identification performance against INCV and DY-Boostrap (Fig. 3 and 6).