We thank all reviewers for their constructive feedback! We are encouraged they find the proposed view of **self-distillation (SD)** and **label-smoothing (LS)** as MAP insightful ([R2], [R3], [R4]), that relating accuracy to **confidence diversity (CD)** interesting ([R1], [R4]), that the theoretically inspired **Beta-smoothing (BS)** is useful ([R2], [R3]) and that experiments are carefully designed to verify our hypothesis ([R2], [R3]). We are also pleased [R4] endorsed the importance of improved calibration. We address reviewers comments below, and will incorporate all feedback.

[R1]: **Relationships between CD, MAP, SD, and LS**: We show in section 5 that SD and LS can be unified under the framework of amortized MAP estimation (Eqn 6). The difference between SD and LS amounts to different choices of priors. Specifically, SD corresponds to using instance-specific priors. We argue that the instance-specific prior is crucial for better generalization, and empirically verify the claim through comparisons of LS, BS and SD (See [R4] for further discussion). This explains why SD outperforms LS. Please refer to our response to [R3] for discussion on CD. **BS works worse than SD**: This is because the instance-specific priors obtained through BS are not as good as those from a pre-trained teacher model which can more accurately capture the relative uncertainties among samples. Note SD demands more computational resources. This does not contradict our unifying view of LS and SD as MAP estimation, and further demonstrates the importance of a good prior. **Some technical details are not described clearly**: Intuitively, samples that are easier to learn should be assigned a more confident label. As such, both confidence and beta samples are sorted in ascending orders so that samples with higher confidence will get more confident beta labels. Lastly, we want to apologize for any confusions, and will improve on clarity of the paper in general.

[R2]: We appreciate positive comments, and will update relevant works and improve the presentation accordingly.

[R3]: **Alternate formulations for CD**: CD is a measure of cross-sample variability of the prediction confidence in the ground truth label. One can alternatively compute the variance of prediction confidence. Average predictive uncertainty (PU) is the entropy of the softmax vectors on average (Eqn 4). PU will be high if all predictions are close to uniform. A model can have high PU, but small CD if the cross-sample variability of predictions is small. We hypothesize that, due to the expressivity of NNs, it is insufficient to only regularize PU (e.g., via LS), as all samples can have predictions close to the smoothed labels, thereby overfitting. Some samples are likely more representative than others, and should be assigned soft labels with higher confidence. As such, instance-specific priors should be used so that different samples have different soft labels. **Experimental section is somewhat weak**: We plan on adding more experiments (See our response to [R4] for some preliminary results). We will also include additional experiments with NLP tasks; We have experiments with standard distillation in Appendix 9.6; Only one generation is carried out for all SD experiments.

[R4]: **Ablation studies on CD**: BS serves as an ablation study to demonstrate the importance of CD. In BS, the soft labels on average are the same as that of LS, but have variability among the labels with samples from the Beta distribution, thereby promoting CD. Please refer to [R1] on the implementation of BS. We have previously explored using CD directly in training, but did not obtain good results. This could have been caused by difficulty in estimating CD accurately using a small mini-batch size of e.g. 32. Naively promoting CD during early stage of training could also have harmed learning. **On Bayesian view of distillation**: The MAP framework provides a unifying view to many recently proposed regularization methods, and serves as a guide for the more principled design of regularization techniques. For instance, it offers insights on improving distillation. Specifically, from this perspective, models with better accuracy may not be better teachers (as empirically observed in recent literature). Instead, what really matters is a model that captures the relative uncertainty among training samples. We hope to explore further on alternative ways of obtaining better instance-specific priors. **Distillation improves calibration seems incremental**: We agree. However, we stress that SD only improves calibration when temperature scaling is not applied to student model during distillation, as suggested by our MAP framework. The commonly used distillation loss will not lead to such improvements (See Sec. 3.1). **Lack of comparison to self-training**: We will add this experiment, and have some preliminary results with ResNet. Following the implementation of arxiv.org/abs/1703.01780, we obtain accuracy of $73.9\%/53.2\%$ with self-training as compared to $75.2\%/55.6\%$ with BS on CIFAR100/CUB200 respectively. **Sweeping over different values of smoothing**: Preliminary results on ResNet with CIFAR100 (LS: 74.0% BS: 74.3% SD: 75.3% when $\epsilon = 0.1$ and LS: 75.2% BS: 75.1% SD: 75.9% when $\epsilon = 0.3$) and CUB200 (LS: 50.4% BS: 53.5% SD: 55.5% when $\epsilon = 0.1$ and LS: 56.2% BS: 56.9% SD: 57.3% when $\epsilon = 0.3$) suggest that instance-specific priors are beneficial for different values of smoothing. We will incorporate further experiments. **Contradicting Fig. 1**: We acknowledge the large fluctuation in accuracy and will rephrase accordingly. Nevertheless, a comparison of Fig. 1 and 2 shows that the max accuracy achieved in 10 generations do agree with that of predictive uncertainty (PU) and CD, supporting our hypothesis that BAN's improvements mainly come from increased PU and CD. **Combining Fig. 1 and 2**: Good suggestion! Based on some preliminary results (not shown due to limited space), the correlation between accuracy and CD holds. In addition, the degree of PU also can influence results. From the MAP perspective, PU quantifies the overall confidence level of the instance-specific priors while CD measures the variability among them. There is an optimal CD and PU for best performance. **Related works section seem thin**: We will extensively revise the related works section. Particularly, our paper differs from arxiv.org/abs/1909.11723 in that their work does not highlight the importance of instance-specific regularization. We also provide a general MAP framework and a careful empirical comparison of LS and SD.