We thank all reviewers for encouraging our work on the following strengths: 1) Balanced Softmax is simple yet effective; 2) our theoretical analysis shows inspiring insights; 3) our experiments are extensive and performance achieves SOTA. We will answer the major points below and address all remaining ones in the final version.

Reviewer #1:

Q1: Explanation about the mismatch (1/4 and 1) between the theory (Theorem 2 and Corollary 2.1) and practice.

A1: We used the slow rate $1/n^{\frac{1}{2}}$ in the derivation of Theorem 2 (see Sup. Mat.). [3] discussed that deep neural networks can improve the convergence rate. When the convergence rate used in Theorem 2 is $1/n^2$, the factor in Corollary 2.1 will be 1 and aligns with Balanced Softmax. We leave further discussions on the convergence rate to future works.

Reviewer #2:

Q1: Eqn.3 and Eqn.4 are very similar to [3, A, B], ... particularly similar to Eqn.11 in [B].

A1: We progress the line of works [3, A, B] by introducing novel probabilistic insights that also bring significant empirical improvements. Eqn.11 in [B] is generic (a superset of most loss engineerings like [3, 29, A]), it uses bi-level optimization to find the unknown logit adjustment $\xi_{p,j}$ of each class, leaves a large search space and a hard optimization landscape. We directly derive the optimal logit adjustment ($\xi_{p,j} = n_j$) with a solid probabilistic grounding (Theorem 1). Moreover, none of [3, A, B] touches the core observation of our work: the link between Softmax and the Bayesian inference under data-imbalanced scenarios. We will add a discussion on [3, A, B] in the final version.

Q2: Meta sampler has a similar idea to [12,24,27].

A2: [12,24,27] 's idea is to use meta-learning to find each training sample's importance towards model training, while we proposed Meta Sampler as a viable solution to the over-balance issue described in line 151-165. Moreover, none of the existing works extend from reweight to resample (Meta Sampler outperforms Meta Reweighter by a large margin on CIFAR10-LT); theirs are instance-based and ours is class-based (fewer parameters and simpler optimization landscape).

Q3: The analysis does not imply proposed softmax... adding the margin term into the loss won't affect the learning.

A3: We did not suggest to add a margin constant into the loss term, instead, we use Corollary 2.1 to show that the optimal margin can be achieved by a proper loss parameterization, i.e., the 1/4 variant of Balanced Softmax.

Q4: The authors argued that *re-sampling techniques* can be harmful to model training, but finally still apply it.

A4: The argument is for *Class Balanced Sampling*, but not for all *re-sampling techniques* (line 151-165). Please kindly refer to R3Q1 for why we need Meta Sampler as a learnable re-sampling technique to complement Balanced Softtmax.

Q5: When to start the meta sampler leads to a mother hyper-parameter.

A5: We apply the Meta Sampler from the very beginning of the training (epoch 0) like any other re-sampling strategy (e.g., Class Balanced Sampling), thus when to start Meta Sampler is not a mother hyper-parameter in our method.

Q6: Meta Sampler makes the contributions vague; include experimental results w/ and w/o the Meta Sampler.

A6: Meta Sampler is complementary to Balanced Softmax (line 38-39), which can be supported by the ablations on CIFAR-LT (Table 5). We provide more results on LVIS with only Balanced Softmax: AP_m:26.3, AP_f:28.8, AP_c:27.3, AP_r:16.2, AP_b:27.0. Compared to experiments in Table 4, the results show that BALMS works better as a whole.

Q7: The authors' baseline softmax results are much higher than those reported in other papers.

A7: Our baseline softmax results align with the most recent paper [29] (Table 7, CIFAR-100-LT), which is published on CVPR 2020. Please kindly refer to R3Q3 for why we retrain all compared methods on the baseline.

Reviewer #3:

Q1: Motivation for the additional (class) meta sampling is lacking.

A1: We need Meta Sampler to appropriately re-sample according to Balanced Softmax's effect on gradients. The 'over-balance' analysis shows a hypothesized case: when the training loss *infinitely approaches* 0 (line 160-162), Balanced Softmax will cast an inverse weight $1/n_j$ to gradients (its combination with Class Balanced Sampler makes the overall weight approach $1/n_j^2$, i.e., over-balanced). However, when the training loss does not *infinitely approach* 0 (in actual training), Balanced Softmax's effect on gradients can be viewed as variables between 1 and $1/n_j$. Therefore, we need to explicitly estimate the optimal sample rate to keep the gradient always being balanced weighted at $1/n_j$.

Q2: Why decoupled training is necessary?

A2: Decoupled training is not necessary. We used the technique in our work to: 1) align with recent research results ([15] ICLR 2020, [33] CVPR 2020) to benefit future study, and to 2) save the computational cost of Meta Sampler.

Q3: The quoted CIFAR results are difficult to compare with prior work.

A3: We retrained all compared methods since prior works chose different baselines and cannot be fairly compared with. We used the highest softmax baseline ([29], CVPR 2020), and it is more challenging and revealing to achieve performance gain on a higher baseline. Following the suggestions, we will specify more details on baseline variants.

Reviewer #4:

Q1: The 1/4 factor in the generalization bound is a bit unsatisfactory.

A1: The mismatch can be reasonably explained. Please kindly refer to our discussion on convergence rates in R1Q1.

Q2: Could the authors explain the source of this cost (Meta Sampler), and how the approach scales in practice?

A2: Meta Sampler involves a second-order optimization, it usually doubles the computational graph and triples the forward/backward times. Thus, end-to-end training with it is slower. In practice, with decoupled training, we only optimize for the linear classifier, which greatly reduced the #parameters in the loop and makes the cost acceptable.