Supplemental Material

A. Additional information about $p(\epsilon)$

We provide a more formal definition of the probability $p(\epsilon)$. Consider the empirical distribution $\widehat{p_c(z)}$ of n samples $z_1, \ldots, z_n$ from class c. This empirical distribution $\hat{p}$ can be viewed by itself as a random variable because it receives different values for different instantiations of the random sample. $\hat{p}$ is distributed over the simplex Δ^d where d is the dimension of the representation of z. As a result, the ϵ-ball around $\widehat{\mu_c}$ is also a random variable, since different random samples yield different centroids and balls.

For any given ϵ, some of these balls may cover the true distribution p and some may not, depending on the instances z_i drawn. $p(\epsilon)$ denotes the probability that such an ϵ-ball covers the true distribution.

B. Tightness of the bound

We derive a lower bound of our loss and evaluate empirically that the bounds are tight.

The lower bound is easily derived in a very similar way to the derivation of the upper bound in Theorem 1 using the triangle inequality. The bound has the following form

$$-\log P(z|\mu_c) \geq -\log \frac{e^{-d(\widehat{\mu_c},z)+2\varepsilon_c}}{\sum_{z' \in Z} e^{-d(\widehat{\mu_c},z')+2\varepsilon_c \delta(z',c)}} \qquad (20)$$

We further estimated empirically the relative magnitude of the bound gap $\frac{|Eq.17-Eq.20|}{Eq.18}$, averaged over all samples. For CIFAR100-LT it was 0.07, suggesting that the bounds are quite tight in our case.

C. DRO-LT Sample count $\epsilon/\sqrt{n}$

Given a set of samples z_i, their mean is known to have a standard deviation of $\sigma/\sqrt{n}$, where σ is the standard deviation of the sample distribution $p(z)$. In our case, σ is not known.

The DRO-LT variant that we call **Sample count** $\epsilon/\sqrt{n}$, can be viewed as assuming that all classes share the same standard deviation of their sample distribution σ, which we tune as a hyper-parameter, and the uncertainty about class centroid only varies by the number of samples.

D. Loss trade-off parameter λ

Figure 6 quantifies the effect of the trade-off parameter λ (Eq. 19) on the validation accuracy. The model was trained on CIFAR100-LT with an imbalance factor of 100 and with our DRO-LT loss (**Learnable** ε variant). It shows that training with DRO-LT alone ($\lambda = 0$) is not enough and leads to poor accuracy. Combining the robustness loss with a discriminative loss (cross-entropy) gives the best results and suggests high-quality feature representations and a discriminative classifier.

E. Robustness vs Performance

Figure 7 explores the effect of different uncertainty radii (ε) on the validation accuracy of a model trained on CIFAR100-LT with an imbalance factor 100. The radius is set to be equal for all classes in our DRO-LT loss (**Shared** ε variant). It shows that the accuracy is maintained over a large range of ε values. Setting ε to very small values nullifies the robustness and reduces accuracy. At the same time, very large values of ε cause the worst-case centroid in the ϵ-ball to be too far from μ_c making the bound too loose and again reduces the accuracy.

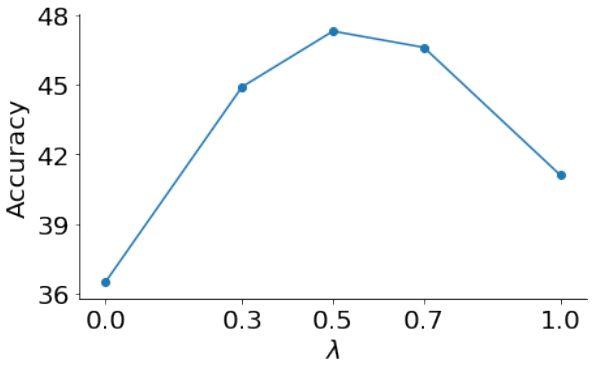

Figure 6: Validation accuracy of a model trained on CIFAR100-LT (imbalance factor 100) with different loss trade-off parameter (λ) values between standard cross-entropy loss and our DRO-LT loss (Learnable ε variant).

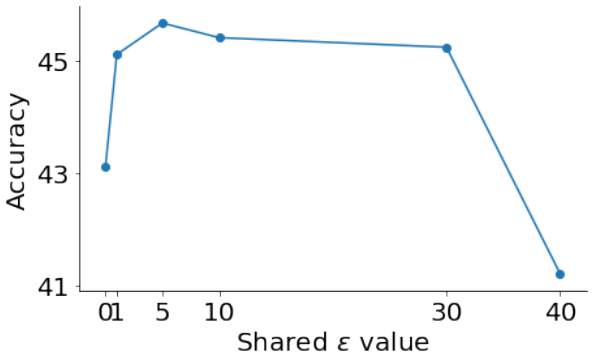

Figure 7: Validation accuracy of a model trained on CIFAR100-LT (imbalance factor 100) with different uncertainty radius ε. The radius is set to be equal for all classes in our **shared** ε variant.

F. More analysis on iNaturalist:

Table 5 compares DRO-LT with common long-tail methods on iNaturalist with accuracy broken by class frequency: many-shot ("Many"), medium-shot ("Med") and few-show ("Few"). It shows that improvement is larger at the head (Many) and tail (Few), but relatively small for most of the classes in this dataset (Med). We also provide the variance for 10 runs.

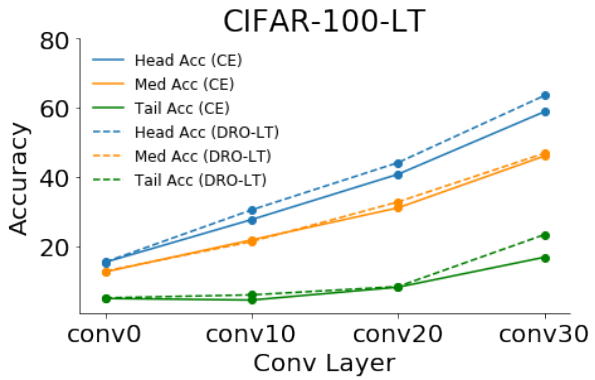

Figure 8: Accuracy of a nearest-centroid neighbor classifier when applied to convolutional layers 0, 10, 20, and 30 of a ResNet-32. Our model narrows the accuracy gap between head and tail classes. The model was trained on CIFAR-100-LT with an imbalance factor of 100. We compare a model trained with standard cross-entropy loss (solid line) and a model trained with DRO-LT (dashed line) where the loss is applied to the last convolutional layer (conv30). We report balanced validation accuracy for head classes (blue), medium classes (orange), and tail classes (green).

iNaturalist	Many	Med	Few	Acc
CE*	72.2	63.0	57.2	61.7
CB LWS	71.0	69.8	68.8	69.5
DRO-LT (ours)				
Shared ε	78.2	70.6	64.7	69.0 ± 0.2
$\varepsilon/\sqrt{n}$	71.0	68.9	**69.3**	69.1 ± 0.2
Learned ε	**73.9**	**70.6**	68.9	$\mathbf{69.7 \pm 0.1}$

Table 5: Top-1 accuracy on long-tailed iNaturalist with accuracy broken to Many-shot, Medium-shot and Few-shot classes. Our approaches improves the performance on head and tail classes.

G. Results for CIFAR-10-LT:

Table 6 compares our approach with common long-tail methods on CIFAR-10-LT [6]. Our method outperforms all baselines.

CIFAR-10-LT	Acc
CE	72.1
CB LWS [21]	73.5
LDAM DRW [6]	77.03
smDragon [35]	79.6
Learned ε (ours)	**82.6**

Table 6: Top-1 accuracy on long-tailed CIFAR-10 [6] with imbalance factor 100.

H. Balancing latent representations

Here, we provide more analysis on the imbalance of latent representation and compare our approach with a standard baseline.

Figure 8 shows the accuracy obtained with a nearest-centroid classifier when applied to layers 0, 10, 20, and 30 of a ResNet-32. We compare a model trained with standard cross-entropy loss (solid line) and a model trained with DRO-LT (dashed line) where the loss is applied to the last convolutional layer (conv30). We show that our model narrows the accuracy gap between head classes (blue) and tail classes (green), mostly in the last layer. This shows the effectiveness of our approach on the latent representations of the deep models and suggests that applying our loss to the rest of the layers might result in a more balanced model overall.