A. Proofs of Results in Body

Proof of Lemma 1. Observe that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{N}\sim q^m}\left[\ell(y, f(x); \mathcal{N})\right] &= \mathop{\mathbb{E}}_{\mathcal{N}\sim q^m}\left[\phi(f_y(x)) + \sum_{i=1}^{m} w_{y,s_i}\cdot\varphi(-f_{s_i}(x))\right] \\
&= \phi(f_y(x)) + \sum_{i=1}^{m}\mathop{\mathbb{E}}_{s_i\sim q}\left[w_{y,s_i}\cdot\varphi(-f_{s_i}(x))\right] \\
&= \phi(f_y(x)) + m\cdot\mathop{\mathbb{E}}_{y'\sim q}\left[w_{yy'}\cdot\varphi(-f_{y'}(x))\right] \\
&= \phi(f_y(x)) + m\cdot\sum_{y'\neq y} q(y' \mid y)\cdot w_{yy'}\cdot\varphi(-f_{y'}(x)) \\
&= \phi(f_y(x)) + m\cdot\sum_{y'\neq y}\rho_{yy'}\cdot\varphi(-f_{y'}(x)),
\end{aligned}$$

where $\rho_{yy'} = m\cdot w_{yy'}\cdot q(y'|y)$. Similarly,

$$\begin{aligned}
\mathop{\mathbb{V}}_{\mathcal{N}\sim q^m}\left[\ell(y, f(x); \mathcal{N}, w)\right] &= \mathop{\mathbb{V}}_{\mathcal{N}\sim q^m}\left[\phi(f_y(x)) + \sum_{i=1}^{m} w_{y,s_i}\cdot\varphi(-f_{s_i}(x))\right] \\
&= \mathop{\mathbb{V}}_{\mathcal{N}\sim q^m}\left[\sum_{i=1}^{m} w_{y,s_i}\cdot\varphi(-f_{s_i}(x))\right] \\
&= m\cdot\mathop{\mathbb{V}}_{y'\sim q}\left[w_{y,y'}\cdot\varphi(-f_{y'}(x))\right] \\
&= m\cdot\mathop{\mathbb{E}}_{y'\sim q}\left[\left[w_{y,y'}\cdot\varphi(-f_{y'}(x))\right]^2\right] - m\cdot\left[\mathop{\mathbb{E}}_{y'\sim q}\left[w_{y,y'}\cdot\varphi(-f_{y'}(x))\right]\right]^2 \\
&= m\cdot\sum_{y'\neq y} q(y' \mid y)\cdot\left[w_{y,y'}\cdot\varphi(-f_{y'}(x))\right]^2 - \\
&\qquad m\cdot\left[\sum_{y'\neq y} q(y' \mid y)\cdot w_{y,y'}\cdot\varphi(-f_{y'}(x))\right]^2 \\
&= \sum_{y'\neq y} w_{yy'}\cdot\rho_{yy'}\cdot\varphi(-f_{y'}(x))^2 - \frac{1}{m}\cdot\left(\sum_{y'\neq y}\rho_{yy'}\cdot\varphi(-f_{y'}(x))\right)^2.
\end{aligned}$$

□

Proof of Lemma 2. Observe that

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{N}\sim q^m}\left[\sum_{y'\in\mathcal{N}} w_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}\right] &= m\cdot\mathop{\mathbb{E}}_{y'\sim q}\left[w_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}\right] \\
&= m\cdot\sum_{y'\neq y} q_{y'}\cdot w_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}.
\end{aligned}$$

By Jensen's inequality, the expected loss is thus bounded by

$$\log\left[1 + \sum_{y'\neq y} m\cdot q_{y'}\cdot w_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}\right].$$

□

Proof of Theorem 3. Let $G_{y,m}(x) \doteq e^{f_y(x)} + \sum_{y' \in \mathcal{N}} w_{yy'} e^{f_{y'}(x)}$ be an estimate of the partition function. From Lemma 6 (Appendix B),

$$\begin{aligned}
&\sqrt{m}\left(\log G_{y,m}(x) - \log \mu_y(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2(x)/\mu_y^2(x)\right) \\
\implies &\sqrt{m}\left(-f_y(x) + \log G_{y,m}(x) - [-f_y(x) + \log \mu_y(x)]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2(x)/\mu_y^2(x)\right) \\
\implies &\sqrt{m}\left(\ell(y, f(x); \mathcal{N}, w) - \ell_m^{q,w}(y, f(x))\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2(x)/\mu_y^2(x)\right),
\end{aligned}$$

where

$$\ell(y, f(x); \mathcal{N}, w) = -f_y(x) + \log\left[e^{f_y(x)} + \sum_{y' \in \mathcal{N}} w_{yy'} e^{f_y(x)}\right],$$

$$\ell_m^{q,w}(y, f(x)) = -f_y(x) + \log\left[e^{f_y(x)} + \sum_{y' \neq y} \rho_{yy'} e^{f_y(x)}\right],$$

and $\rho_{yy'} \doteq m \cdot q_{y'} \cdot w_{yy'}$ from Lemma 2. This implies that for sufficiently large m and for a given $x \in \mathcal{X}$, we have

$$\mathbb{E}_{\mathcal{N} \sim q^m}\left[\left(\ell(y, f(x); \mathcal{N}, w) - \ell_m^{q,w}(y, f(x))\right)^2\right] = \frac{\sigma_y^2(x)}{m \cdot \mu_y^2(x)} + o_p(1).$$

where $o_p(1)$ is a random variable that converges to 0 in probability.

□

Proof of Proposition 4. From (11), for a given (q, w), the implicit loss for the sampled softmax cross-entropy is

$$\ell_m^{q,w}(y, f(x)) = \log\left[1 + \sum_{y' \neq y} \rho_{yy'} \cdot e^{f_{y'}(x) - f_y(x)}\right],$$

where $\rho_{yy'} = m \cdot q_{y'} \cdot w_{yy'}$. This exactly equals the pairwise margin loss (6). Thus, for a fixed $\rho_{yy'}$, picking $w_{yy'} = \frac{\rho_{yy'}}{m \cdot q_{y'}}$ guarantees the implicit and pairwise margin losses coincide. □

Proof of Proposition 5. This is a simple consequence of the fact that when applying importance weighting $\mathop{\mathbb{E}}_{q}\left[\frac{p(x)}{q(x)} \cdot f(x)\right]$ to approximate an expectation $\mathop{\mathbb{E}}_{p}[f(x)]$, the minimum variance choice of q is $q^*(x) \propto p(x) \cdot f(x)$ (e.g., see Alain et al. (2015)). □

B. A Helper Lemma

Lemma 6. *Define $G_{y,m}(x) \doteq e^{f_y(x)} + \sum_{y' \in \mathcal{N}} w_{yy'} e^{f_{y'}(x)}$ as a random variable that estimates the partition function, where $\mathcal{N} = \{y'_1, \ldots, y'_m\} \overset{i.i.d.}{\sim} q$. Assume $q \in \Delta_{[L]}$ has $q_y = 0$, and $w_{yy'} = \frac{1}{m}\eta_{yy'}$ where $\eta_{yy'}$ is independent of m. Let $\mu_y(x) := \mathbb{E}_{\mathcal{N} \sim q^m} G_{y,m}(x)$, and $\sigma_y^2(x) = \mathbb{V}_{y' \sim q}[\eta_{yy'} e^{f_{y'}(x)}]$, both assumed to be strictly positive and finite. Then, for any $x \in \mathcal{X}$, the following statements hold:*

1. $\sqrt{m}(G_{y,m}(x) - \mu_y(x)) \xrightarrow{d} \mathcal{N}(0, \sigma_y^2(x))$;
2. *If $\mu_y(x) > 0$, then* $\sqrt{m}\left(\log G_{y,m}(x) - \log \mu_y(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2(x)/\mu_y^2(x)\right)$.

Here, $\xrightarrow{d}$ denotes convergence in distribution as $m \to \infty$.

Proof. We first note that

$$\mathbb{E}_{y' \sim q}[G_{y,m}(x)] = e^{f_y(x)} + m\mathbb{E}_{y' \sim q} w_{yy'} e^{f_y(x)}$$

$$= e^{f_y(x)} + \sum_{y' \neq y} m \cdot q_{y'} \cdot w_{yy'} e^{f_y(x)}$$
$$= e^{f_y(x)} + \sum_{y' \neq y} q_{y'} \eta_{yy'} e^{f_y(x)} = \mu_y(x),$$

where we use the fact that samples in $\mathcal{N}$ are i.i.d. Equivalently $G_{y,m}(x) = e^{f_y(x)} + \frac{1}{m}\sum_{i=1}^{m} \eta_{y,y'_i} e^{f_{y'}(x)}$. Note that $e^{f_y(x)}$ is not random, and $\frac{1}{m}\sum_{i=1}^{m} \eta_{y,y'_i} e^{f_{y'}(x)}$ is an average of i.i.d. random variables. By the central limit theorem, it follows that $\sqrt{m}(G_{y,m}(x) - \mu_y(x)) \xrightarrow{d} \mathcal{N}(0, \sigma_y^2(x))$ where $\sigma_y^2(x) = \mathbb{V}_{y' \sim q}[\eta_{yy'} e^{f_{y'}(x)}]$.

Recall the Delta method which states that if a sequence of random variables $(X_m)_{m \in \mathbb{Z}_+}$ satisfies $\sqrt{m}(X_m - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ for some $\theta \in \mathbb{R}$ and $\sigma^2 > 0$, then $\sqrt{m}(g(X_m) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 g'(\theta)^2)$ for any function g whose derivative $g'(\theta)$ exists at θ, and is non-zero. Choosing $g(x) = \log(x)$ and applying the Delta method to $\sqrt{m}(G_{y,m}(x) - \mu_y(x)) \xrightarrow{d} \mathcal{N}(0, \sigma_y^2(x))$ gives

$$\sqrt{m}\left(\log G_{y,m}(x) - \log \mu_y(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_y^2(x)/\mu_y^2(x)\right)$$

□

C. Additional Discussion on Negative Sampling Schemes

Remark 4. For the softmax cross-entropy, one way of reasoning about the weighting on negative samples is as a *logit correction*. Observe that (7) may be rewritten

$$\ell(y, f(x); \mathcal{N}) = \log\left[1 + \sum\nolimits_{y' \in \mathcal{N}} e^{\bar{f}_{y'}(x) - f_y(x)}\right], \tag{18}$$

where $\bar{f}_{y'}(x) = f_{y'}(x) - \log w_{yy'}$ are *corrected* versions of the original logits or scores. Note further that if $\mathcal{N}$ can include the positive label y, this is tantamount to additionally correcting the positive logit as well.

Remark 5. Further to Remark 4, the difference between the importance and relative weighting schemes may be understood as follows. Suppose we employ the softmax cross-entropy with explicit exclusion of the positive label from $\mathcal{N}$. Further, if we modify the positive logit to $\tilde{f}_y(x)$, and negative logit to $\bar{f}_y(x)$:

$$\ell(y, f(x); \mathcal{N}) = \log\left[1 + \sum_{y' \in \mathcal{N} - \{y\}} e^{\bar{f}_{y'}(x) - \tilde{f}_y(x)}\right].$$

By setting $\tilde{f}_y(x) = f_y(x)$, and $\bar{f}_{y'}(x) = f_{y'}(x) - \log(m \cdot q_{y'})$, we obtain the importance weighting scheme. On the other hand, if we additionally set $\tilde{f}_y(x) = f_y(x) - \log(m \cdot q_y)$, — i.e., apply *positive* logit correction as well — then we arrive at the relative weighting scheme.

Remark 6. As stated, the sampling distribution q may place non-zero mass on the "positive" label y; thus, one may include y amongst the "negative" labels. As this is intuitively undesirable, the domain of q may be additionally restricted so as to exclude this possibility. Further, one may explicitly discount this label from consideration by zeroing out its weight; e.g., we may apply $w_{yy'} = [\![y' \neq y]\!]$ in place of constant weighting. This is similar yet *distinct* to forcing q to exclude y from its sampling domain, as the former implicitly modifies the distribution of negatives. In practice, however, the two approaches have similar performance.

D. Expected Decoupled Losses under Negative Sampling

Table 3 summarises expected losses under negative sampling for the decoupled case.

E. Details of Long-tail Experiments

For all datasets, we use SGD with momentum 0.9. Dataset specific settings are given below.

Sampling distribution	Weighting	Expected loss on negatives	Comment
Uniform	Constant ($\frac{1}{m}$)	$\frac{1}{L-1}\sum_{y' \neq y} \varphi(-f_{y'}(x))$	Scaled decoupled loss
Uniform	Importance weighting ($\frac{L}{m}$)	$\sum_{y' \neq y} \varphi(-f_{y'}(x))$	Decoupled loss
Uniform	Relative weighting (1)	$\frac{m}{L}\sum_{y' \neq y} \varphi(-f_{y'}(x))$	Scaled decoupled loss
Uniform	$\frac{L}{m} \cdot \frac{\pi_{y'}}{\pi_y}$	$\sum_{y' \neq y} \frac{\pi_{y'}}{\pi_y} \cdot \varphi(-f_{y'}(x))$	Tail-heavy loss
Within-batch	Constant ($\frac{1}{m}$)	$\sum_{y' \neq y} \pi_{y'} \cdot \varphi(-f_{y'}(x))$	Tail-heavy loss
Within-batch	Importance weighting ($\frac{1}{m \cdot \pi_{y'}}$)	$\sum_{y' \neq y} \varphi(-f_{y'}(x))$	Decoupled loss
Within-batch	Relative weighting ($\frac{\pi_y}{\pi_{y'}}$)	$m \cdot \pi_y \cdot \sum_{y' \neq y} \varphi(-f_{y'}(x))$	Head-heavy loss
Within-batch	$\frac{1}{m \cdot \pi_y}$	$\sum_{y' \neq y} \frac{\pi_{y'}}{\pi_y} \cdot \varphi(-f_{y'}(x))$	Tail-heavy loss

Table 3: Expectation of loss of negatives $\sum_{y' \in \mathcal{N}} w_{yy'} \cdot \varphi(-f_{y'}(x))$ for an example (x, y). Here, negatives $\mathcal{N}$ are sampled from q with $q_y > 0$, and weighting scheme w satisfies $w_{yy} = 0$. Different choices of (q, w) yield upper bounds which resemble losses from the long-tail learning literature, such as the equalised loss of Tan et al. (2020) and the logit-adjusted loss of Menon et al. (2020).

CIFAR-100: We use a CIFAR ResNet-56 with weight decay of 10^{-4} trained for 256 epochs, using a minibatch size of 128. We use a stepwise annealed learning rate, with a base learning rate of 0.1 that is decayed by 0.1 at the 160th epoch, and by 0.01 at the 180th epoch. We apply standard CIFAR data augmentation per Cao et al. (2019); He et al. (2016).

ImageNet: We use a ResNet-50 with weight decay of 5×10^{-4} trained for 90 epochs, using a minibatch size of 512. We use a cosine learning rate with a base learning rate of 0.4. We apply standard ImageNet data augmentation per Goyal et al. (2017).

F. Additional Results: Long-tail Datasets

We present additional results on the long-tail learning benchmarks using a contrastive loss, and compare the overall (non-sliced) balanced errors of various methods.

F.1. Results on Contrastive Loss

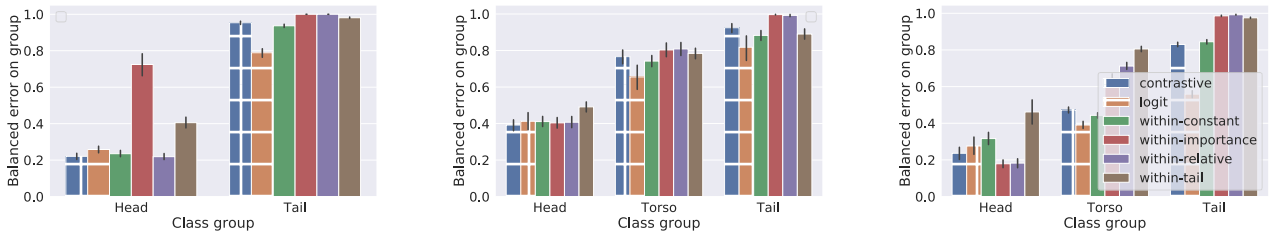
Figure 3 shows results using the contrastive loss on the long-tail benchmarks. Here, the performance of different sampling schemes is more variable compared to the softmax cross-entropy. In particular, on Tail classes, the performance of sampling is generally poor compared to the baseline. Note that the latter is the de-facto choice of loss function for long-tail settings. Consequently, the default hyperparameters (e.g., learning rate and batch size) are generally attuned to this loss. Further tuning of these may improve the results for the contrastive loss.

F.2. Balanced Error Plots

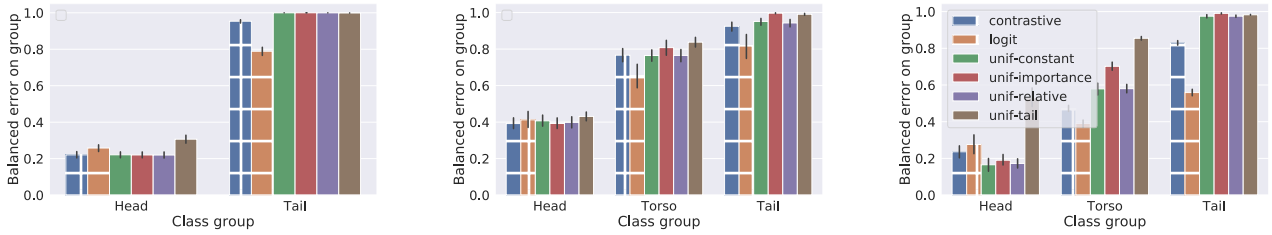
Figures 4 and 5 present the balanced errors of the various choices of sampling and weighting schemes, for the softmax cross-entropy and contrastive loss, respectively. We see that the gains of within-batch sampling with constant weighting are such that it can improve over the standard loss using *all* the labels. In general, performance is superior using the softmax cross-entropy versus contrastive loss; this is in keeping with the former's extensive use as a foundation in long-tail problems.

F.3. Results with Varying Number of Sampled Negatives

We present results where the number of sampled negatives varies from $\{32, 64, 128, 256\}$ on ImageNet-LT, using the softmax cross-entropy. Figure 6 shows that with fewer sampled negatives, performance tends to slightly degrade, as expected. However, even with a modest number of negatives, the general trends seen in the body are reflected.

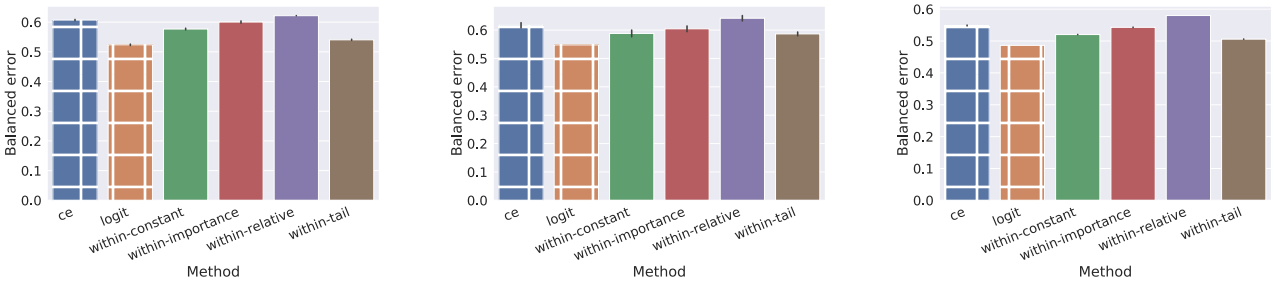


(a) CIFAR-100-LT Step (contrastive loss). (b) CIFAR-100-LT Exp (contrastive loss). (c) ImageNet-LT (contrastive loss).

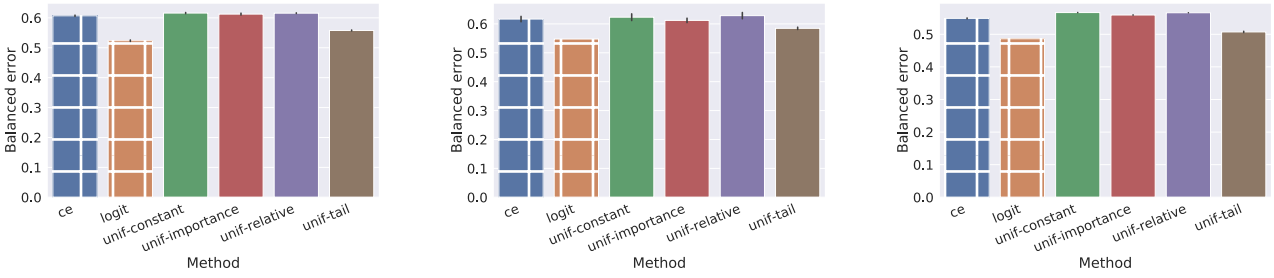


(d) CIFAR-100-LT Step (contrastive loss). (e) CIFAR-100-LT Exp (contrastive loss). (f) ImageNet-LT (contrastive loss).

Figure 3: Results on head, torso and tail labels on long-tail learning benchmarks, using the contrastive loss. Fig. 3a - 3c show the performance of within-batch negative sampling along with baseline loss functions. Fig. 3d - 3f illustrate the performance of uniform negative sampling.



(a) CIFAR-100-LT (Step). (b) CIFAR-100-LT (Exp). (c) ImageNet-LT.

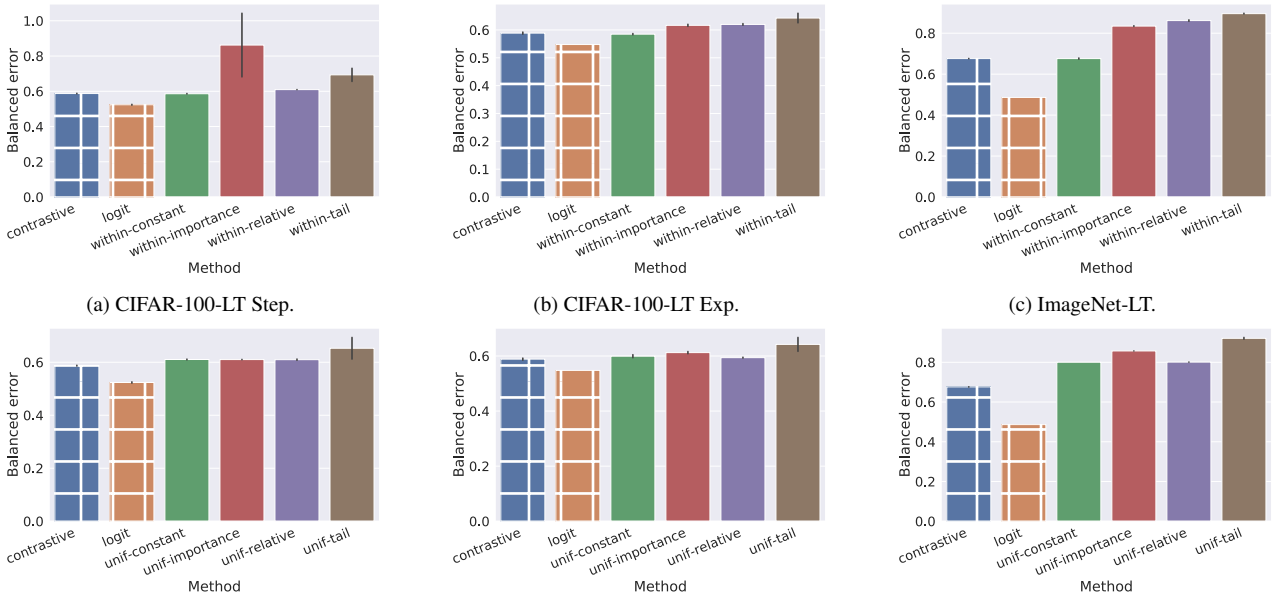


(d) CIFAR-100-LT (Step). (e) CIFAR-100-LT (Exp). (f) ImageNet-LT.

Figure 4: Balanced error on long-tail learning benchmarks using the softmax cross-entropy. We present results for within-batch (`within`) negative sampling (Fig. 4a - 4c) and uniform (`unif`) negative sampling (Fig. 4d - 4f) , using the constant weight (`const`), importance weighting (`importance`), and relative weighting (`relative`) schemes from Table 1.

F.4. Results on iNaturalist 2018

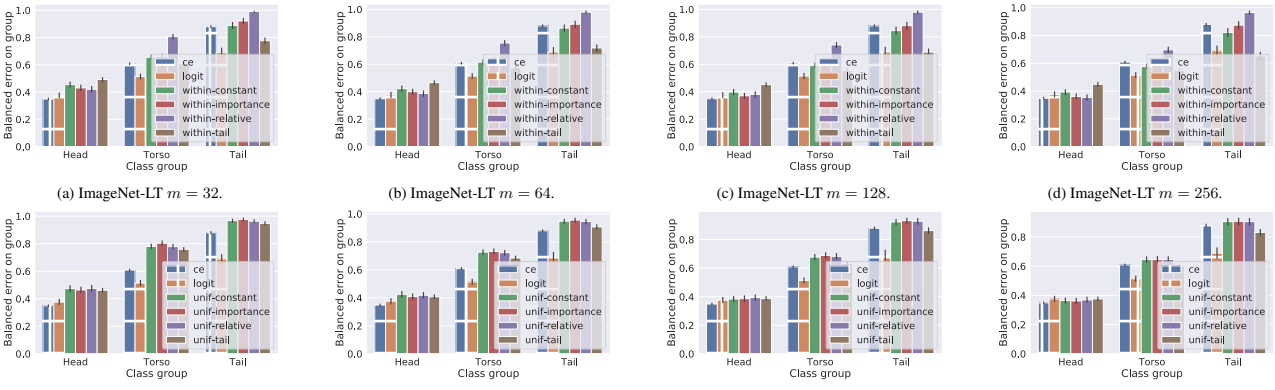
See Figure 7.



(a) CIFAR-100-LT Step. (b) CIFAR-100-LT Exp. (c) ImageNet-LT.

(d) CIFAR-100-LT Step. (e) CIFAR-100-LT Exp. (f) ImageNet-LT.

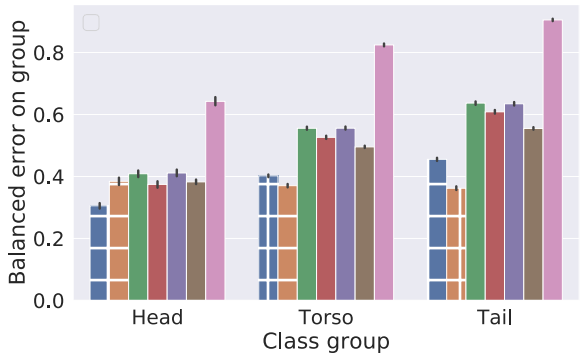
Figure 5: Balanced error on long-tail learning benchmarks using the contrastive loss. We present results for within-batch (`within`) negative sampling (Fig. 5a - 5c) and uniform (`unif`) negative sampling (Fig. 5d - 5f), using the constant weight (`const`), importance weighting (`importance`), and relative weighting (`relative`) schemes from Table 1.



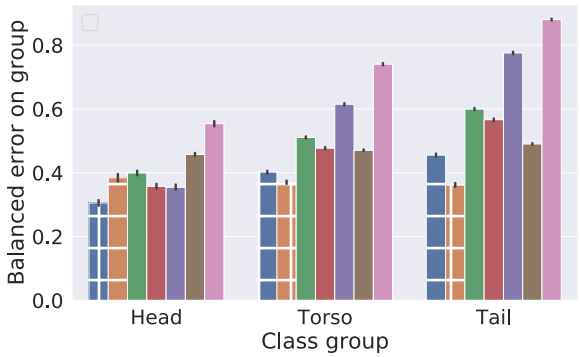
(a) ImageNet-LT $m = 32$. (b) ImageNet-LT $m = 64$. (c) ImageNet-LT $m = 128$. (d) ImageNet-LT $m = 256$.

(e) ImageNet-LT $m = 32$. (f) ImageNet-LT $m = 64$. (g) ImageNet-LT $m = 128$. (h) ImageNet-LT $m = 256$.

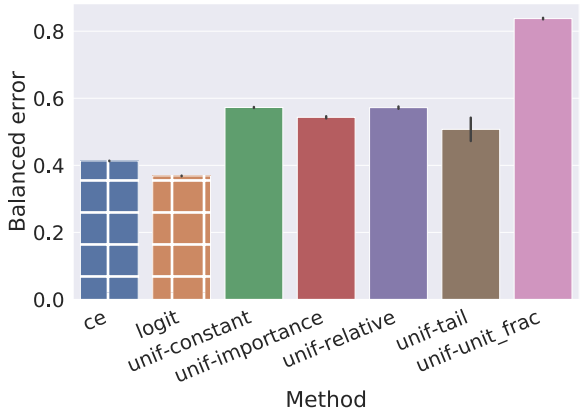
Figure 6: Results on head, torso and tail labels on ImageNet-LT, using varying number of sampled negatives. Fig. 6a - 6d show the performance of within-batch negative sampling along with baseline loss functions. Fig. 6e - 6h illustrate the performance of uniform negative sampling.



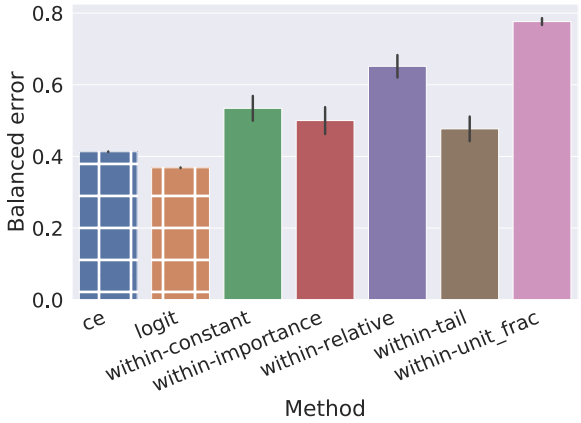
(a) CE, Uniform



(b) CE, within batch



(c) CE. Uniform. Balanced error.



(d) CE, within batch. Balanced error.

Figure 7: **(a)**, **(b)**: Results on head, torso, and tail labels on iNaturalist 2018 using uniform (`unif`) (in **(a)**) and within-batch (`within`) (in **(b)**) negative sampling. **(c)**, **(d)**: Balanced error on iNaturalist 2018. We present results for uniform (`unif`) and within-batch (`within`) and negative sampling, using the constant weight (`const`), importance weighting (`importance`), and relative weighting (`relative`) schemes from Table 1.

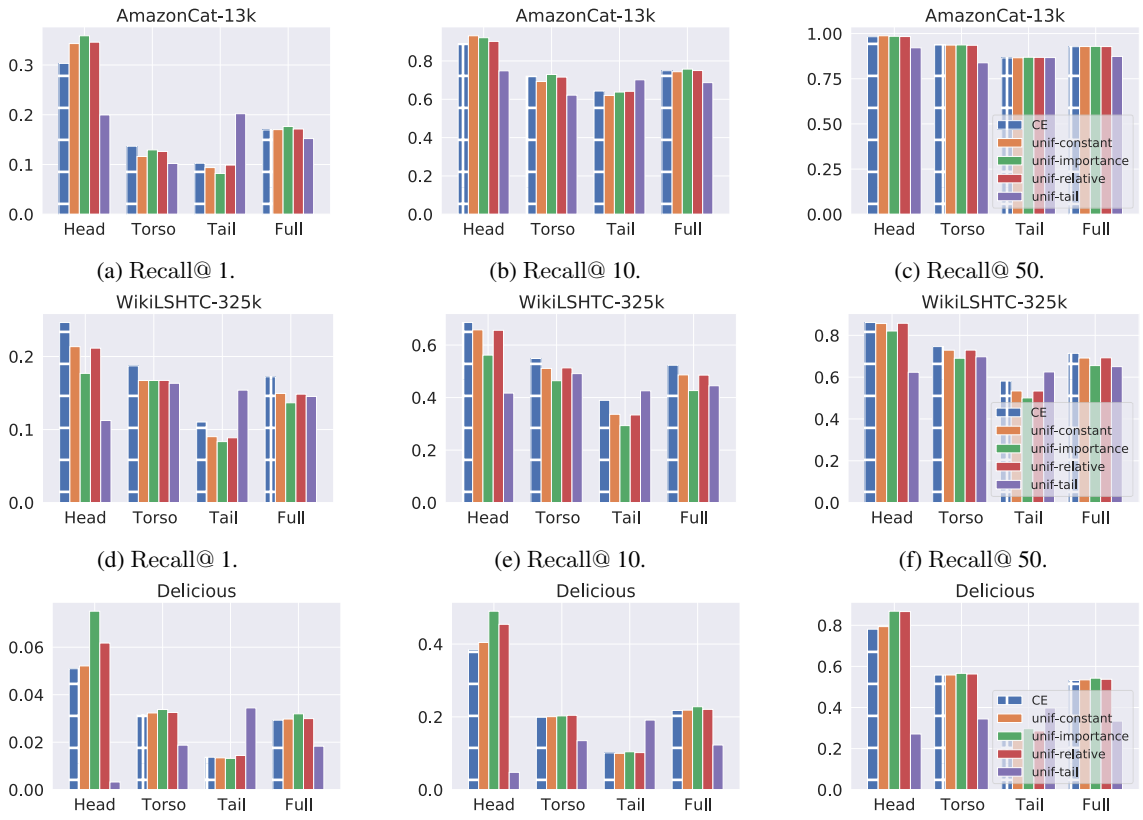
G. Retrieval Datasets

Table 4 presents the details of the retrieval benchmarks used in § 5.2.

Dataset	#Features	#Labels	#Train Points	#Test Points	Average #I/L	Average #L/I
DELICIOUS	500	983	12920	3185	311.61	19.03
AMAZONCAT-13K	203,882	13,330	1,186,239	306,782	448.57	5.04
WIKILSHTC-325K	1,617,899	325,056	1,778,351	587,084	17.46	3.19

Table 4: Summary of the extreme classification datasets used in this paper (Varma, 2018). #I/L is the number of instances per label, and #L/I is the number of labels per instance.

H. Additional Results: Retrieval Datasets

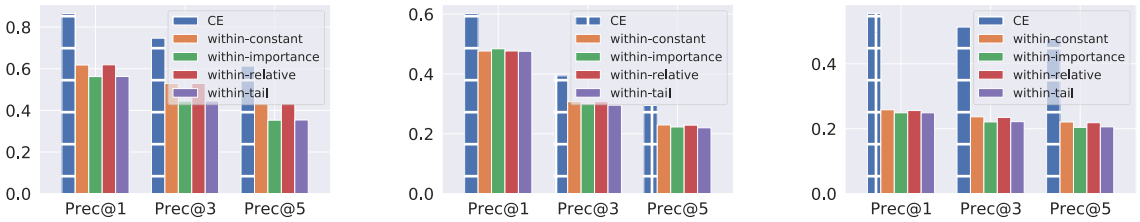


(a) Recall@ 1. (b) Recall@ 10. (c) Recall@ 50.

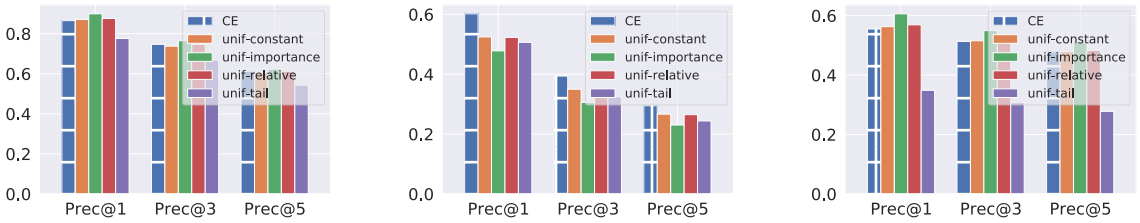
(d) Recall@ 1. (e) Recall@ 10. (f) Recall@ 50.

(g) Recall@ 1. (h) Recall@ 10. (i) Recall@ 50.

Figure 8: Performance of uniform negative sampling based cross-entropy loss (cf. (7)) on AMAZONCAT-13K (Figure 8a - 8c), WIKILSHTC-325K (Fig. 8d - 8f), and DELICIOUS (Fig. 8g - 8i). These experiments utilize $m = 256$ negative for AMAZONCAT-13K and WIKILSHTC-325K, and $m = 64$ negatives for DELICIOUS. We report the performance on three subpopulations (Head, Torso, and Tail) and the entire test set (Full), as measured by Recall@k for k = 1, 10, and 50. We combine uniform sampling with constant, importance, and relative weighting schemes. For reference, we include the results of standard softmax cross-entropy loss (`ce`). Note that for the retrieval datasets, the uniform sampling aligns with `ce` as it consistently focuses on Head, Torso, and Tail in that order. This is in contrast with the within-batch sampling (cf. Figure 2).

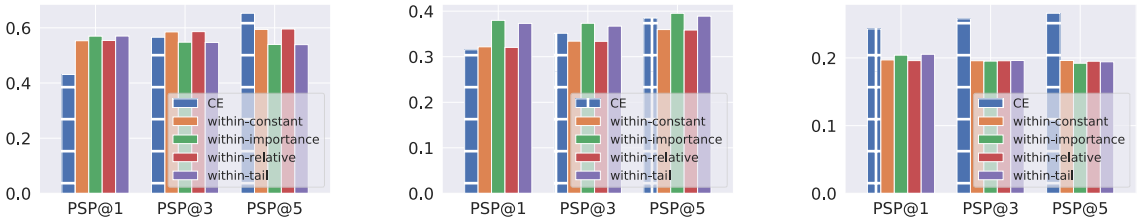


(a) AMAZONCAT-13K. (b) WIKILSHTC-325K. (c) DELICIOUS.

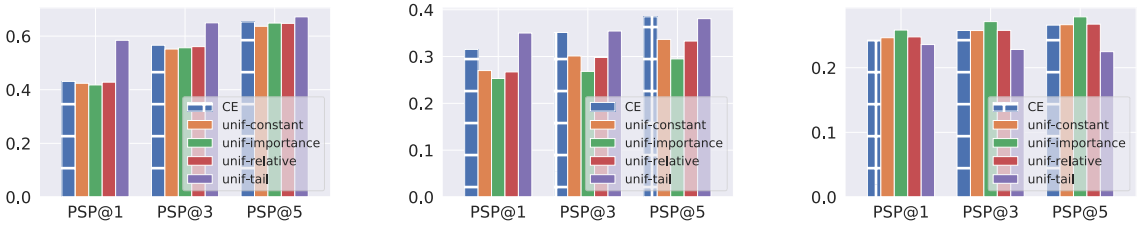


(d) AMAZONCAT-13K. (e) WIKILSHTC-325K. (f) DELICIOUS.

Figure 9: Performance of within-batch and uniform sampling on the entire original (multilabel) test sets of AMAZONCAT-13K, WIKILSHTC-325K, and DELICIOUS, as measured by Precision@k for k = 1, 3, and 5. These experiments utilize $m = 256$ negative for AMAZONCAT-13K and WIKILSHTC-325K, and $m = 64$ negatives for DELICIOUS. For reference, we include the results of standard softmax cross-entropy loss (ce). Note that these results are included here for completeness, as they are often reported in the literature as the performance measure. Since these results do not breakdown the performance on different subpopulations, they do not highlight the impact of sampling distribution and weighting schemes on various subpopulations.



(a) AMAZONCAT-13K. (b) WIKILSHTC-325K. (c) DELICIOUS.



(d) AMAZONCAT-13K. (e) WIKILSHTC-325K. (f) DELICIOUS.

Figure 10: Performance of within-batch and uniform sampling on the entire original (multilabel) test sets of AMAZONCAT-13K, WIKILSHTC-325K, and DELICIOUS, as measured by *propensity-scored* variant of Precision@k (Jain et al., 2016) or PSP@k for k = 1, 3, and 5. These experiments utilize $m = 256$ negative for AMAZONCAT-13K and WIKILSHTC-325K, and $m = 64$ negatives for DELICIOUS. For reference, we include the results of standard softmax cross-entropy loss (ce).