Disentangling Labeling and Sampling Bias for Learning in Large-output Spaces

Ankit Singh Rawat [1] **Aditya Krishna Menon** [1] **Wittawat Jitkrittum** [1] **Sadeep Jayasumana** [1] **Felix X. Yu** [1]
Sashank J. Reddi [1] **Sanjiv Kumar** [1]

Abstract

Negative sampling schemes enable efficient training given a large number of classes, by offering a means to approximate a computationally expensive loss function that takes all labels into account. In this paper, we present a new connection between these schemes and loss modification techniques for countering *label imbalance*. We show that different negative sampling schemes *implicitly* trade-off performance on dominant versus rare labels. Further, we provide a unified means to explicitly tackle both *sampling bias*, arising from working with a subset of all labels, and *labeling bias*, which is inherent to the data due to label imbalance. We empirically verify our findings on long-tail classification and retrieval benchmarks.

1. Introduction

Classification problems with a large number of labels arise in language modelling (Mikolov et al., 2013; Levy & Goldberg, 2014), recommender systems (Covington et al., 2016; Xu et al., 2016), and information retrieval (Agrawal et al., 2013; Prabhu & Varma, 2014). Such *large-output* problems pose a core challenge: losses such as the softmax cross-entropy can be prohibitive to optimise, as they depend on the *entire* set of labels. Several works have thus devised *negative sampling* schemes for efficiently and effectively approximating such losses (Bengio & Senecal, 2003; 2008; Blanc & Rendle, 2018; Ruiz et al., 2018; Bamler & Mandt, 2020).

Broadly, negative sampling techniques *sample* a subset of "negative" labels, which are used to contrast against the observed "positive" labels. One further applies a suitable *weighting* on these "negatives", which ostensibly corrects the *sampling bias* introduced by the dependence on a random subset of labels. Intuitively, such bias assesses how closely a scheme approximates the unsampled loss on the full label set. This bias is well understood for *sampled softmax* schemes (see, e.g., Bengio & Senecal (2008)); surprisingly, however, far less is understood about other popular schemes, e.g., within-batch and uniform sampling (cf. §2.2).

In this paper, we systematically study the sampling bias of generic negative sampling schemes, with two main findings. First, we precisely characterise the *implicit losses* optimised by such schemes (§3). This reveals that, e.g., uniform and within-batch sampling do *not* faithfully approximate the unsampled loss; thus, measured by the yardstick of sampling bias, such schemes appear woefully sub-optimal.

Intriguingly, however, our analysis further reveals a novel *labeling bias* perspective on negative sampling. Specifically, we show that when the label distribution is skewed — as is common with a large number of labels (Jain et al., 2016) — different negative sampling schemes *implicitly* trade-off performance on *dominant versus rare* labels (§4). Consequently, while certain schemes may incur a sampling bias, they may prove highly effective in modeling rare labels.

Put together, our analysis reveals that negative sampling can account for *both* sampling and labeling bias, with the latter done *implicitly*. Note that the two concerns fundamentally differ: the former results from working with a random subset of labels, while the latter results from an inherent property of the data. By disentangling these concerns, we arrive at a unified means to *explicitly* tackle both biases (§4.2).

In summary, our contributions are:

(i) we precisely characterise the effect of generic negative sampling schemes by explicating the *implicit* losses they optimise (§3; Theorem 3). This shows that popular schemes such as uniform and within-batch negative sampling (Hidasi et al., 2016; Wang et al., 2017) do *not* faithfully approximate the unsampled loss.

(ii) we provide a novel connection between negative sampling and long-tail learning, and show that negative sampling can *implicitly* trade off performance on dominant versus rare labels (§4.1; Proposition 4). This implies that, e.g., within-batch sampling can boost performance on rare labels despite not approximating the unsampled loss. Further, this yields a unified means to tailor sam-

[1]Google Research, New York, USA. Correspondence to: Ankit Singh Rawat <ankitsrawat@google.com>, Aditya Krishna Menon <adityakmenon@google.com>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

pling schemes to focus on particular label slices (§4.2).

(iii) we empirically verify our analysis of sampling schemes' implicit dominant versus rare label tradeoffs (§5).

2. Background and Notation

Consider a multiclass retrieval[1] setting, where given an instance $x \in \mathcal{X}$, we want to recover its relevant label $y \in \mathcal{Y} = [L] \doteq \{1, 2, \ldots, L\}$. Let $S_N \doteq \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$ be a sample generated from the underlying data distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$. We seek to learn a scorer $f \colon \mathcal{X} \to \mathbb{R}^L$ via S_N such that it minimises the *retrieval risk* at some $r \in \mathbb{N}_+$:

$$L_{\text{ret}}(f) = \mathbb{P}_{x,y}\big(y \notin \text{top}_r(f)\big), \tag{1}$$

where $\text{top}_r(f)$ denotes the r indices with the highest scores under the scorer f. When $r = 1$, the retrieval risk in (1) reduces to minimising the standard *misclassification error*

$$L_{\text{err}}(f) = \mathbb{P}_{x,y}\big(y \notin \text{argmax}_{y' \in [L]}\, f_{y'}(x)\big). \tag{2}$$

To achieve this, a common strategy is to minimise a surrogate loss $\ell \colon \mathcal{Y} \times \mathbb{R}^L \to \mathbb{R}_+$, where $\ell(y, f(x))$ is the loss incurred for predicting $f(x)$ when the true label is y. A ubiquitous choice is the softmax-based cross-entropy loss a.k.a. softmax cross-entropy loss,

$$\ell(y, f(x)) = -f_y(x) + \log\left[\sum\nolimits_{y' \in [L]} e^{f_{y'}(x)}\right] \tag{3}$$

$$= \log\left[1 + \sum_{y' \neq y} e^{f_{y'}(x) - f_y(x)}\right]. \tag{4}$$

Equivalently, this is the log-loss under the *softmax distribution* $p_y(x) \propto \exp(f_y(x))$ with *logits* $f_y(x)$. Alternately, one may use a *decoupled* loss (Zhang, 2004)

$$\ell(y, f(x)) = \phi(f_y(x)) + \sum_{y' \neq y} \varphi(-f_{y'}(x)), \tag{5}$$

where $\phi, \varphi \colon \mathbb{R} \to \mathbb{R}_+$ are *margin losses* for binary classification, e.g., hinge loss. In (3) and (5), we may consider y to be a "positive" label, and each $y' \neq y$ a "negative" label.

In *large-output* settings where number of labels L is large, there are at least two additional considerations in designing a loss. First, most labels may have only a few associated samples (He & Garcia, 2009; Buda et al., 2017). Second, computing the losses (3) and (5) may be prohibitively expensive owing to their linear complexity in L (Bengio & Senecal, 2008; Morin & Bengio, 2005; Gutmann & Hyvärinen, 2012; Bose et al., 2018; Reddi et al., 2019). These problems have been separately addressed in the literatures on *long-tail learning* and *negative sampling*.

[1] The general multilabel retrieval setting, where each instance x potentially has multiple relevant labels $\mathbf{y} \subset \mathcal{Y}$, can be reformulated as a multiclass retrieval problem via standard reductions (Wydmuch et al., 2018; Reddi et al., 2019; Menon et al., 2019).

2.1. Long-tail Learning Methods

In *long-tail* learning settings, $\pi_y \doteq \mathbb{P}(y)$ is highly skewed, and so many labels have few associated samples. Here, achieving low misclassification error (2) may mask poor performance on rare classes, which is undesirable. To cope with this problem, existing approaches include post-processing model outputs (Fawcett & Provost, 1996; Zhou & Liu, 2006), re-balancing the training data (Chawla et al., 2002; Wallace et al., 2011), and modifying the loss function (Xie & Manski, 1989; Morik et al., 1999). One strategy is to encourage larger classification margins for rare labels. Such approaches augment the softmax cross-entropy with *pairwise label margins* $\rho_{yy'}$ (Menon et al., 2020):

$$\ell^{\rho}_{\text{mar}}(y, f(x)) = \log\left[1 + \sum\nolimits_{y' \neq y} \rho_{yy'} \cdot e^{f_{y'}(x) - f_y(x)}\right], \tag{6}$$

where $\log \rho_{yy'}$ represents the desired gap or "margin" between scores for y and y'. For suitable $\rho_{yy'}$, this loss allows for greater emphasis on rare classes' predictions. For example, Cao et al. (2019) proposed an "adaptive" loss with $\rho_{yy'} \propto \pi_y^{-1/4}$. This upweights rare "positive" labels y to encourage a larger gap $f_y(x) - f_{y'}(x)$ for such labels. Tan et al. (2020) proposed an "equalised" loss with $\rho_{yy'} = F(\pi_{y'})$, for increasing $F \colon [0, 1] \to \mathbb{R}_+$. This downweights rare "negatives" y', thus preventing $f_{y'}$ from being too small. Finally, Khan et al. (2018); Ren et al. (2020); Menon et al. (2020); Wang et al. (2021) proposed a "logit adjusted" loss with $\rho_{yy'} = \frac{\pi_{y'}}{\pi_y}$. This encourages larger margin between positives y that are *relatively* rare compared to negative y'.

Another popular strategy is to re-weight the individual samples, so that rare labels receive a higher penalty; however, these are generally outperformed by margin approaches (Cao et al., 2019). Further, we shall demonstrate the latter have a strong connection to negative sampling.

2.2. Negative Sampling Methods

At a high level, negative sampling proceeds as follows: given an example (x, y), we draw an i.i.d. sample $\mathcal{N} = \{y'_1, \ldots, y'_m\}$ of m "negative" labels from some distribution $q \in \Delta_{[L]}$. We now compute a loss using $\mathcal{N}$ in place of $[L]$, applying suitable *weights* $w_{yy'} \geq 0$. For example, applied to the softmax cross-entropy in (3), we get

$$\ell(y, f(x); \mathcal{N}, w) = \log\left[1 + \sum_{y' \in \mathcal{N}} w_{yy'} \cdot e^{f_{y'}(x) - f_y(x)}\right]. \tag{7}$$

Similarly, the decoupled loss in (5) becomes

$$\ell(y, f(x); \mathcal{N}, w) = \phi(f_y(x)) + \sum_{y' \in \mathcal{N}} w_{yy'} \cdot \varphi(-f_{y'}(x)). \tag{8}$$

Remark 1. The distribution q may or may not include the positive label y amongst the negative sample $\mathcal{N}$. In the latter case, $q_y = 0$. See Appendix C for further discussion.

Sampling distribution q	Description
Current model probabilities	Model-based negatives
Uniform distribution	Uniform negatives
Training distribution	Within-batch negatives

Weight $w_{yy'}$	Description
$1/(m \cdot q_{y'})$	Importance weighting
$1/m$	Constant weighting
$q_y/q_{y'}$	Relative weighting

Table 1: Summary of sampling distributions $q \in \Delta_{[L]}$ for an example (x, y), and weights $w_{yy'}$ on the negatives y'. Here, m is the number of sampled negatives. Note that one may exclude the positive label y either while sampling, or via weighting.

Remark 2. We write q, w for brevity; in general, each of these may additionally depend on (x, y).

There are several canonical (q, w) choices in the literature. For the sampling distribution q, popular choices are:

- **Model-based**. Here, q is taken to be the current model probabilities; e.g., when training $f\colon \mathcal{X} \to \mathbb{R}^L$ to minimise the softmax cross-entropy (3), $q_{y'} \propto \exp(f_{y'}(x))$, or an efficient approximation thereof (Bengio & Senecal, 2008; Blanc & Rendle, 2018; Rawat et al., 2019; Sun et al., 2019).
- **Uniform**. Here, q is uniform over all labels (Hidasi et al., 2016; Wang et al., 2017), i.e., $q_{y'} = \frac{1}{L}$ for every $y' \in [L]$.
- **Within-batch**. Suppose we perform minibatch-based model training, so that we iteratively make updates based on a random subset $\{(x_b, y_b)\}_{b=1}^B$ of training samples. In within-batch negative sampling, we use all labels in the minibatch as our negatives, i.e., $\mathcal{N} = \bigcup_{b=1}^B \{y_b\}$ (Chen et al., 2017; Broscheit et al., 2020). In expectation, this is equivalent to $q = \pi$, the training label distribution.

For the weighting scheme $w_{yy'}$, popular choices are:

- **Importance weighting**. Here, $w_{yy'} = \frac{1}{m \cdot q_{y'}}$; this choice is motivated from the importance weighting identity, as shall subsequently be explicated.
- **Constant**. Here, $w_{yy'}$ is a constant for all sampled negatives; typically, this is just $w_{yy'} \equiv \frac{1}{m}$.
- **Relative weighting**. Here, w is a variant of importance weighting, given by the density ratio $w_{yy'} = q_y/q_{y'}$.

We summarize these common choices for (q, w) in Table 1. Particular combinations of (q, w) yield common negative sampling schemes. For example, the *sampled softmax* (Bengio & Senecal, 2008; Blanc & Rendle, 2018) employs a model-based sampling distribution q, coupled with the importance weighting w. Within-batch and uniformly sampled negatives are typically coupled with a constant weight (Hidasi et al., 2016; Wang et al., 2017; Broscheit et al., 2020). Yi et al. (2019b) employed within-batch sampling with relative weighting. Appendix C discusses further variants.

2.3. Efficiency versus Efficacy of Negative Sampling

There are two primary considerations in assessing a negative sampling scheme. The first is *efficiency*: to facilitate large-scale training, it ought to be cheap to sample from the distribution q. Uniform and within-batch negatives are particularly favourable from this angle, and have thus enjoyed wide practical use (Hidasi et al., 2016; Yi et al., 2019a).

The other side of the coin is *efficacy*: to ensure good model performance, the weights w ostensibly ought to yield a good approximation to the original (unsampled) loss. Sampled softmax schemes have been extensively studied through this lens; surprisingly, however, less attention has been spent on uniform and within-batch sampling, despite their popularity.

We now study the efficacy of generic sampling schemes, with the aim of understanding their *sampling bias* (i.e., deviation from the unsampled loss). In a nutshell, we explicate the *implicit loss* that a given sampling scheme optimises (§3). Beyond elucidating their sampling bias, this gives a novel *labeling bias* view of negative sampling, and connects them to the long-tail literature (§2.1). In particular, the implicit losses bear strong similarity with those designed to ensure balanced performance for long-tail settings. We thus propose a unified approach to design negative sampling schemes that overcome sampling *and* labeling bias (§4).

3. The Implicit Losses of Negative Sampling

Given a negative sampling scheme parameterised by (q, w), we now ask: what loss does this scheme *implicitly* optimise? Comparing this *implicit loss* with the unsampled loss (e.g., the softmax cross-entropy loss (3)) quantifies the *sampling bias*, which we shall see is non-trivial for popular schemes.

3.1. Implicit Losses for Generic Negative Samplers

Negative sampling schemes are characterised by the sampling distribution q and weighting scheme w. Each (q, w) pair defines the sampled loss minimised during the training based on a random subset of negatives $\mathcal{N}$ (cf. (7) and (8)). Such sampled losses are random variables, owing to their dependence on $\mathcal{N}$. We may thus study their corresponding *expected losses*, which we view as the *implicit loss* being optimised during training. The form of this loss helps us understand the effect of sampling on model performance.

In the following, we assume for simplicity that $q_y = 0$, so that the positive label y is excluded from the negative sample. Further, we assume that $\mathcal{N}$ is sampled *with* replacement. We first consider the setting of decoupled losses (cf. (8)), and

Sampling distribution	Weighting $w_{yy'}$	Implicit loss	Comment
Uniform	Constant ($\frac{1}{m}$)	$\log\left[1+\frac{1}{L}\sum_{y'\neq y} e^{f_{y'}(x)-f_y(x)}\right]$	Softmax with downweighted negatives
Uniform	Importance ($\frac{1}{m\cdot q_{y'}}$)	$\log\left[1+\sum_{y'\neq y} e^{f_{y'}(x)-f_y(x)}\right]$	Softmax cross-entropy
Uniform	Relative ($\frac{q_y}{q_{y'}}$)	$\log\left[1+\frac{m}{L}\sum_{y'\neq y} e^{f_{y'}(x)-f_y(x)}\right]$	Softmax with downweighted negatives
Uniform	"Tail" ($\frac{\pi_{y'}}{m\cdot q_{y'}\cdot\pi_y}$)	$\log\left[1+\sum_{y'\neq y} \frac{\pi_{y'}}{\pi_y}\cdot e^{f_{y'}(x)-f_y(x)}\right]$	Logit-adjusted loss of Menon et al. (2020)$^{\downarrow}$
Within-batch	Constant ($\frac{1}{m}$)	$\log\left[1+\sum_{y'\neq y} \pi_{y'}\cdot e^{f_{y'}(x)-f_y(x)}\right]$	Equalised loss of Tan et al. (2020)$^{\downarrow}$
Within-batch	Importance ($\frac{1}{m\cdot q_{y'}}$)	$\log\left[1+\sum_{y'\neq y} e^{f_{y'}(x)-f_y(x)}\right]$	Softmax cross-entropy
Within-batch	Relative ($\frac{q_y}{q_{y'}}$)	$\log\left[1+m\cdot\pi_y\cdot\sum_{y'\neq y} e^{f_{y'}(x)-f_y(x)}\right]$	Softmax with upweighted head labels$^{\uparrow}$
Within-batch	"Tail" ($\frac{\pi_{y'}}{m\cdot q_{y'}\cdot\pi_y}$)	$\log\left[1+\sum_{y'\neq y} \frac{\pi_{y'}}{\pi_y}\cdot e^{f_{y'}(x)-f_y(x)}\right]$	Logit-adjusted loss of Menon et al. (2020)$^{\downarrow}$

Table 2: Implicit losses for sampled softmax cross-entropy $\log\left[1+\sum_{y'\in\mathcal{N}} w_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}\right]$ on example (x,y). Here, negatives $\mathcal{N}=\{y'_1,\ldots,y'_m\}$ are sampled from q, and weighted with $w_{yy'}$. Different (q,w) from Table 1, and a "tail" weighting from (16), yield implicit losses (cf. Lemma 2) which equate to pairwise margin losses (6) from the long-tail literature, e.g., Tan et al. (2020); Menon et al. (2020). Losses denoted $^{\uparrow}$ ($^{\downarrow}$) are expected to benefit "head" ("tail") classes, i.e., yield better classification on data points with frequent (rare) labels. For the case of decoupled losses, see Table 3 (Appendix).

then the more challenging softmax cross-entropy.

Decoupled losses. For a generic (q,w), we now characterise the behavior of the sampled decoupled loss (8).

Lemma 1. *Pick any (q,w) where $q\in\Delta_{[L]}$ has $q_y=0$. For the sampled decoupled loss ℓ in (8) with $m\in\mathbb{Z}_+$ negatives,*

$$\mathop{\mathbb{E}}_{\mathcal{N}\sim q^m}\left[\ell(y,f(x);\mathcal{N},w)\right]=\underbrace{\phi(f_y(x))+\sum_{y'\neq y}\rho_{yy'}\cdot\varphi(-f_{y'}(x))}_{:=\ell_m^{q,w}(y,f(x))} \quad (9)$$

$$\mathop{\mathbb{V}}_{\mathcal{N}\sim q^m}\left[\ell(y,f(x);\mathcal{N},w)\right]=\sum_{y'\neq y} w_{yy'}\cdot\rho_{yy'}\cdot\varphi(-f_{y'}(x))^2-\frac{1}{m}\cdot\Big(\sum_{y'\neq y}\rho_{yy'}\cdot\varphi(-f_{y'}(x))\Big)^2, \quad (10)$$

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

The proof of Lemma 1 can be found in Appendix A. Note that one can view $\ell_m^{q,w}(y,f(x))$ as the implicit loss being optimized during the training. Compared to the unsampled loss ℓ in (5), we see that for general (q,w), $\ell_m^{q,w}\neq\ell$ owing to the former involving non-constant weights $\rho_{yy'}$ on each negative. Thus, negative sampling introduces a *sampling bias* in general. Eliminating this bias requires that $\rho_{yy'}=1$, which for a given sampling distribution q requires that $w_{yy'}=\frac{1}{m\cdot q_{y'}}$, i.e., the importance weighting w in Table 1.

Softmax cross-entropy loss. For the softmax cross-entropy, a closed-form for the implicit loss is elusive owing to the non-linear log in (7). Nonetheless, to draw qualitative conclusions about different negative sampling schemes, in this setting we treat the following natural upper bound on the expected loss $\mathbb{E}_{\mathcal{N}}\left[\ell(y,f(x);\mathcal{N},w)\right]$ as the implicit loss.

Lemma 2. *Pick any (q,w) where $q\in\Delta_{[L]}$ has $q_y=0$. For the sampled softmax cross-entropy ℓ in (7),*

$$\mathop{\mathbb{E}}_{\mathcal{N}\sim q^m}\left[\ell(y,f(x);\mathcal{N},w)\right]\leq\underbrace{\log\left[1+\sum_{y'\neq y}\rho_{yy'}\cdot e^{f_{y'}(x)-f_y(x)}\right]}_{:=\ell_m^{q,w}(y,f(x))}, \quad (11)$$

where $\rho_{yy'}\doteq m\cdot q_{y'}\cdot w_{yy'}$ for $m\in\mathbb{Z}_+$ negatives.

See Appendix A for the proof. Similar to the decoupled case, the implicit loss $\ell_m^{q,w}$ (i.e., the upper bound in Lemma 2) diverges from the softmax cross-entropy in general; thus, there is sampling bias. The two coincide for the importance weights $w_{yy'}=\frac{1}{m\cdot q_{y'}}$. In fact, when $q_{y'}\propto e^{f_{y'}}$, this choice of weight ensures *equality* in (11). For general q, it further guarantees asymptotic unbiased estimate of the *gradient* of the softmax cross-entropy as $m\to\infty$ (Bengio & Senecal, 2008). See also Rawat et al. (2019) for non-asymptotic bounds on the gradient bias.

Before proceeding, it is natural to ask if the upper bound $\ell_m^{q,w}$ in (11) is tight enough to base any further analysis upon. The following result establishes that for large m, $\ell_m^{q,w}$ indeed approximates the true behavior of the sampled softmax cross-entropy loss in (7) in a squared-error sense. (See Appendix A for the proof.)

Theorem 3. *Pick any (q,w) such that $q\in\Delta_{[L]}$ has $q_y=0$, and $w_{yy'}$ is expressible as $\frac{1}{m}\cdot\eta_{yy'}$ for $\eta_{yy'}$ independent of m. Let ℓ be the softmax cross-entropy (3), $\ell_m^{q,w}$ be as per Lemma 2, $\mu_y(x)\doteq e^{f_y(x)}+\mathbb{E}_{y'\sim q}[\eta_{yy'}\cdot$*

$e^{f_{y'}(x)}]$, *and* $\sigma_y^2(x) \doteq \mathbb{V}_{y' \sim q}\left[\eta_{yy'} \cdot e^{f_{y'}(x)}\right]$. *Assume that* $\mu_y(x), \sigma_y^2(x) \in (0, +\infty)$. *Then, for large enough* m,

$$\mathbb{E}_{\mathcal{N}}\left[\left(\ell(y, f(x); \mathcal{N}, w) - \ell_m^{q,w}(y, f(x))\right)^2\right] = \frac{\sigma_y^2(x)}{m \cdot \mu_y^2(x)} + E_m, \quad (12)$$

where $E_m = o_p(1)$ *converges to* 0 *in probability.*

In Theorem 3, $\mu_y(x)$ equals the standard partition function $\sum_{y'} e^{f_{y'}(x)}$ under the importance weighting $w_{yy'} = \frac{1}{m \cdot q_{y'}}$. The above guarantees convergence of the sampled loss to $\ell_m^{q,w}$, the upper bound in (11), *provided* $w_{yy'}$ takes the form $\eta_{yy'}/m$; equally, in the language of Lemma 2, this requires that $\rho_{yy'}/q_{y'}$ is a constant independent of m. The *rate* of convergence is governed by $\sigma_y^2(x)/\mu_y^2(x)$, which can be seen as an inverse signal-to-noise ratio. Intuitively, when the losses $e^{f_{y'}(x)}$ on negative labels $y' \neq y$ are highly concentrated (i.e., $\sigma_y^2(x)$ is small relative to $\mu_y(x)$), sampling even a few negatives can reliably estimate the true loss.

3.2. Implicit Losses of Existing Negative Samplers

Equipped with our understanding of the implicit losses for generic negative sampling methods, we now focus on specific methods that are prevalent in practice, namely, uniform and within-batch sampling distributions q with different weights w. Table 2 explicates the impact of specific choices of (q, w) through their implicit losses. For example, under uniform sampling, a constant weight of $\frac{1}{m}$ implicitly down-weights negatives in the standard loss by a factor of $\frac{1}{L}$. Indeed, except for the case of $w_{yy'} = \frac{1}{m \cdot q_{y'}}$, the implicit sampling losses subtly differ from the softmax cross-entropy: in such losses, the "negative" labels receive a variable penalty. Thus, in general, such schemes result in a sampling bias.

3.3. Discussion and Implications

We reiterate that while negative sampling schemes such as uniform and within-batch sampling (with constant weights) are popular, they have primarily been motivated as a cheap means of approximating the original (unsampled) loss. Quantifying *how* well they approximate this loss, however, has not (to our knowledge) been previously explicated.

Having specified such schemes' implicit losses above, we make two remarks. First, per §3.2 and Table 2, these schemes induce a sampling bias, and do *not* faithfully approximate the original loss. This appears to suggest that, despite their popularity, such schemes' efficacy is limited.

However, our second observation hints at a different conclusion: note that the implicit loss in (11) bears a striking similarity to the softmax with pairwise label margin (6) from the long-tail learning literature. This raises an intriguing, hitherto unexplored question: could such negative sampling schemes be effective at modelling *tail* labels?

4. Negative Sampling Meets Long-tail Learning

Leveraging the implicit loss derivation above, we now show that negative sampling schemes can mitigate *labeling* bias by implicitly trading-off performance on dominant versus rare classes. Subsequently, we present a unified view of negative sampling that addresses *both* labeling and *sampling* bias.

4.1. Implicit Sampling Losses: Head versus Tail Tradeoffs

In §3.2, we have seen how from a *sampling bias* perspective, to obtain an unbiased estimate of the original loss, the only admissible choice of (q, w) is $w_{yy'} = \frac{1}{m \cdot q_{y'}}$. Interestingly, from a *labeling bias* perspective, a different picture emerges: other choices of (q, w) may *implicitly* trade off performance on rare ("tail") versus dominant ("head") classes, which can be desirable. We illustrate this point via two examples.

Example: within-batch sampling and constant weights. Consider within-batch sampling, so that $q = \pi$ for training label distribution π. Coupled with a constant $w_{yy'} = \frac{1}{m}$, the sampled softmax cross-entropy (11) has implicit loss:

$$\ell_m^{q,w}(y, f(x)) = \log\left[1 + \sum_{y' \neq y} \pi_{y'} \cdot e^{f_{y'}(x) - f_y(x)}\right]. \quad (13)$$

Clearly, this disagrees with the standard softmax cross-entropy (3). However, the loss down-weights the contribution of rare labels when they are encountered as negatives. Since rare labels infrequently appear as positives during training, such down-weighting prevents them from being overwhelmed by negative gradients. The implicit loss (and thus the sampling scheme) is therefore expected to boost classification performance on samples with rare labels.

Example: within-batch sampling and relative weights. Suppose again that $q = \pi$, but that we instead use the relative weighting scheme $w_{yy'} = \frac{\pi_y}{\pi_{y'}}$. Here, the implicit loss for the sampled softmax cross-entropy is

$$\ell_m^{q,w}(y, f(x)) = \log\left[1 + \sum_{y' \neq y} \pi_y \cdot e^{f_{y'}(x) - f_y(x)}\right]. \quad (14)$$

Compared to (13), this has a subtly different negative weighting of π_y, rather than $\pi_{y'}$. This places greater emphasis on *positive* labels that occur more frequently in the training set, and thus has an *opposite* effect to (13): we encourage large model scores for *dominant* compared to rare labels.

4.2. Relating Long-tail and Implicit Sampling Losses

To generalise the above, recall that in the *long-tail learning* problem (§2.1), several popular losses to cope with label imbalance are special cases of the softmax cross-entropy with pairwise margin (6), i.e., for a given $\rho_{yy'} \geq 0$,

$$\ell^{\rho}_{\text{mar}}(y, f(x)) = \log \left[1 + \sum_{y' \neq y} \rho_{yy'} \cdot e^{f_{y'}(x) - f_y(x)} \right].$$

Intriguingly, both the implicit sampling losses from §4.1 are special cases of this loss; in particular, (13) is a special case of the equalised loss of Tan et al. (2020). Thus, such negative sampling schemes *implicitly* cope with label imbalance. This connection is no accident: the following establishes a simple equivalence between the two loss families.

Proposition 4. *Pick any $\rho_{yy'} \geq 0$, and let ℓ^{ρ}_{mar} be the softmax cross-entropy with pairwise margin ρ (6). Given any $q \in \Delta_{[L]}$ with $q_y = 0$ and $m \in \mathbb{Z}_+$, suppose we set*

$$w_{yy'} = \frac{\rho_{yy'}}{m \cdot q_{y'}}. \tag{15}$$

Then, the implicit $\ell^{q,w}_m$ from (11) satisfies:

$$(\forall x \in \mathcal{X}, z \in \mathbb{R}^L)\, \ell^{q,w}_m(y, z) = \ell^{\rho}_{\text{mar}}(y, z).$$

See Appendix A for the proof. Proposition 4 illustrates an intimate connection between negative sampling schemes, and losses for long-tail learning. Remarkably, while the former is ostensibly a means of coping with sampling bias, it can *implicitly* address the labeling bias problem arising from a skewed distribution of classes.

4.3. Discussion and Implications

For generic (q, w), we may understand the performance impact of negative sampling by identifying the implicit losses (Lemma 1 and 2) optimised during training. Both §4.1 and §4.2 demonstrate the dual utility of these implicit losses. We comment on these and further implications of our results.

Overcoming sampling and labeling bias. From §4.1, common negative sampling schemes *implicitly* trade-off performance on dominant versus rare classes. Boosting performance on samples with rare labels is an important goal (Van Horn & Perona, 2017; Buda et al., 2017), and relates to ensuring *fairness* (Sagawa et al., 2020).

Proposition 4 also gives a way to *explicitly* control both sampling *and* labelling bias: i.e., operate on a subset of labels (for computational efficiency), but also ensure good performance on rare labels (to ensure "fairness" across classes). Given a target set of label margins $\rho_{yy'}$ (as in (6)) — which can suitably balance dominant versus rare class performance — one may use (15) to pick a suitable combination of (q, w) to achieve this balance. For example, suppose we wish to approximate the logit-adjustment loss of Menon et al. (2020) (see also Khan et al. (2018); Ren et al. (2020); Wang et al. (2021)), where $\rho_{yy'} = \frac{\pi_{y'}}{\pi_y}$. Then, for a given q, we may set

$$w_{yy'} = \frac{\pi_{y'}}{m \cdot q_{y'} \cdot \pi_y}. \tag{16}$$

Within-batch sampling can boost the tail. Within-batch sampling has the virtue of simplicity: it creates negatives from each sampled minibatch, which involves minimal additional overhead. Our above analysis further shows that with a constant weighting, within-batch sampling can implicitly boost performance on rare classes. Thus, while this method may not be competitive in terms of standard *class-weighted* error, we expect it to have good *balanced* error (where dominant and rare labels are treated equally).

Bias-variance trade-off. The above apparently settles the issue of what a good choice of (q, w) is: for a target $\rho_{yy'}$, (q, w) should satisfy (15). However, how can one choose amongst the infinitude of such pairs? One desideratum is to minimise the *variance* of the loss estimate. Towards this, the following result deals with the more amenable setting involving a decoupled loss.

Proposition 5. *Pick any $\rho_{yy'} \geq 0$, and let $w_{yy'}$ satisfy (15). The sampling distribution q^* with minimum loss variance is*

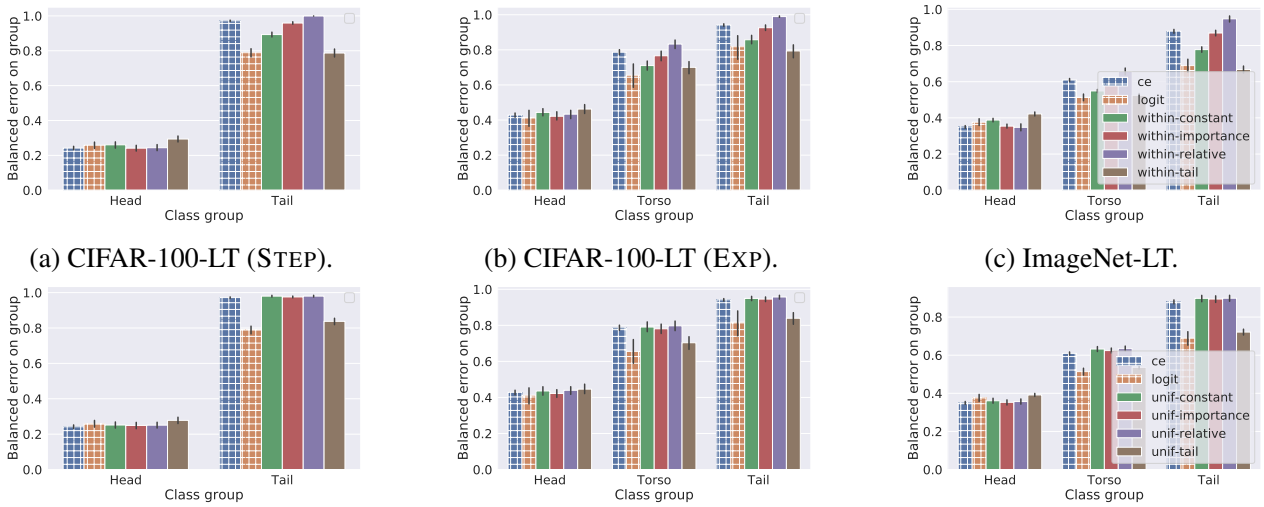
$$(\forall x \in \mathcal{X})\, q^*(y' \mid x, y) \propto \rho_{yy'} \cdot \varphi(-f_{y'}(x)). \tag{17}$$

See Appendix A for the proof. According to (17), to minimise the variance of the loss estimate, one would like to sample the negative that contribute the most to the original decoupled loss. However, this would require evaluating the loss on all classes, which defeats the purpose of sampling the negatives in the first place. Investigating efficient approximations (e.g., per Bamler & Mandt (2020)) is an interesting direction for future work.

Remark 3. In this paper, we focus on the bias and variance of the loss estimate resulting from the underlying sampling method. However, the quality of the corresponding gradient estimates also plays a critical role in the training performance. The bias and variance of the gradient estimate is an active topic of research, both in the context of negative sampling and mini-batch selection (see, e.g., Bengio & Senecal (2008); Rawat et al. (2019); Katharopoulos & Fleuret (2018)).

5. Experiments

We now present experiments on benchmarks for both long-tail learning and retrieval, illustrating our main finding: existing negative sampling schemes, such as within-batch sampling with constant weighting, implicitly trade-off performance on dominant versus rare labels. Further, we show the



(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 1: Balanced error on Head, Torso and Tail labels on long-tail learning benchmarks; see §5.1 for definitions of the STEP and EXP profiles. We employ within-batch (`within`) negative sampling in Fig. 1a - 1c and uniform (`unif`) negative sampling in Fig. 1d - 1f, using the constant, importance, relative, and "tail" weighting schemes from Table 2. We also include the results of standard softmax cross-entropy (`ce`) and logit adjustment (`logit`), which use *all* labels, and are thus shaded. Amongst the standard schemes from Table 1, using within-batch negatives with a constant weighting performs the best on Tail classes. This is despite such a scheme providing a biased estimate of the cross-entropy; indeed, on CIFAR-100-LT, sampling does *better* than the cross-entropy on tail classes, despite the latter employing all labels during training. Our newly proposed "tail" weighting scheme (16) consistently outperforms all schemes on the Tail, confirming the viability of handling sampling *and* labeling bias.

viability of designing custom sampling schemes to specifically target performance on rare labels.

5.1. Results on Long-tail Benchmarks

We present results on long-tailed ("LT") versions of the CIFAR-100 and ImageNet datasets. All datasets feature a rapidly decaying label distribution. Note that negative sampling is hardly necessary on these datasets, as the number of labels L (100 and 1000 respectively) is not prohibitively large. Nonetheless, our aim is to verify that under standard long-tail learning setups, different negative sampling schemes trade-off performance on dominant versus rare classes; we do *not* seek to improve upon existing baselines.

For CIFAR-100, we downsample labels via the EXP and STEP profiles of Cui et al. (2019); Cao et al. (2019), with imbalance ratio $\max_y \mathbb{P}(y)/\min_y \mathbb{P}(y) = 100$. For ImageNet, we use the long-tailed version from Liu et al. (2019). We train a ResNet-56 for CIFAR and a ResNet-50 for ImageNet, using SGD with momentum; see Appendix E for details.

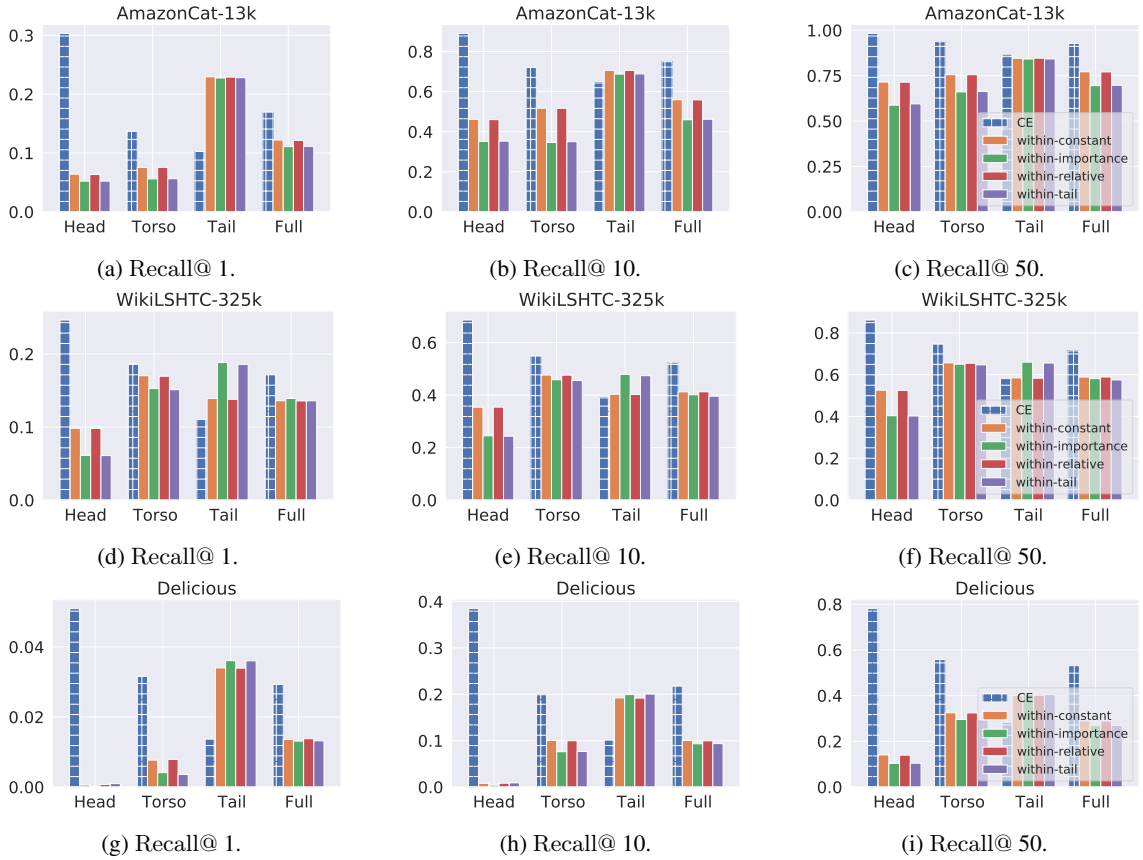
For each dataset, we consider the following techniques:

– standard training with the softmax cross-entropy (3), "cosine contrastive" loss (Hadsell et al., 2006) with margin $\rho = 0$ (i.e., a decoupled loss (cf. (5)) with $\phi(z) = (1-z)^2$ and $\varphi(z) = [z]_+^2$), and the logit adjustment loss of Menon et al. (2020), a representative long-tail learning baseline. These losses use *all* L labels.

– negative sampling using uniform & within-batch negatives; and constant, importance, & relative weighting. Additionally, we employ the "tail" weights for each sampler that mimic the logit-adjusted loss, as per Table 2. We use $m = 32$ negatives on CIFAR-100, and $m = 512$ negatives on ImageNet.

We report the test set *balanced* error, which averages the per-class error rates; this emphasises tail classes more. We slice this error by breaking down classes into three groups — "Head", "Torso", and "Tail" — depending on whether the number of associated samples is ≥ 100, between 20 and 100, or < 20 respectively, per Kang et al. (2020). By construction, for STEP profile, label frequencies take only two values. Accordingly, we refer to classes as "Head" and "Tail" classes. We confirm that different sampling techniques trade-off performance on dominant (Head) versus rare (Tail) classes.

Summary of findings. Figure 1 summarises our results using within-batch and uniform sampling, and the softmax cross-entropy. (See Appendix F for contrastive loss results.) We observe the following.



(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 2: Performance of softmax cross-entropy loss with within-batch negative sampling (cf. (7)) on AMAZONCAT-13K (Figure 2a - 2c), WIKILSHTC-325K (Fig. 2d - 2f), and DELICIOUS (Fig. 2g - 2i). These experiments utilize $m = 256$ negative for AMAZONCAT-13K and WIKILSHTC-325K, and $m = 64$ negatives for DELICIOUS. We report Recall@k for $k \in \{1, 10, 50\}$ on three subpopulations (Head, Torso, and Tail) and the entire test set (Full). We combine within-batch sampling with constant, importance, and relative weighting schemes. For reference, we include the results of standard softmax cross-entropy loss (`ce`). For all weight choices, within-batch sampling focuses on Tail subpopulation.

Within-batch sampling and *constant weighting helps on the tail*. Per our analysis, using within-batch negatives with a constant weighting — which we argued approximates the equalised loss of Tan et al. (2020) — performs well on tail classes. Conversely, with relative weighting, performance on head classes is superior. This is despite both schemes providing biased estimates of the softmax cross-entropy.

The sampling distribution matters. Note that different weighting schemes show markedly different trends on within-batch versus uniform sampling. This highlights that the choice of (q, w) crucially dictates the nature of the trade-off on dominant versus rare classes.

Handling sampling and labeling bias. Note that the weights given by (16) significantly improve performance on tail classes. This confirms that, given a target loss (in this case, logit adjustment), one can design sampling schemes to mimic the loss. Thus, as per our analysis, one can jointly handle sampling and labeling bias.

The limits of sampling. The logit adjustment baseline significantly outperforms the cross-entropy on tail classes — in keeping with Menon et al. (2020) — but also slightly improves over the "tail" sampling proposal which seeks to approximate it. While the latter does approximately minimise a similar loss, the variance introduced by working with a subset of labels comes at a price. In practice, one can reduce the variance by choosing as large a number of sampled labels as is computationally feasible.

5.2. Results on Retrieval Benchmarks

We next focus on the retrieval benchmarks from the extreme classification literature (Agrawal et al., 2013; Bengio et al., 2019), where due to a large number of labels it is common to employ negative sampling. In particular, we experiment with AMAZONCAT-13K and WIKILSHTC-325K datasets

from the extreme classification repository (Varma, 2018), with 13K and 325K labels, respectively. In addition, we also explored a small scale dataset DELICIOUS from the repository to make our conclusions more general. See Appendix G for various statistics of these datasets. Given these inherently multilabel datasets, we first construct their multiclass versions via a standard reduction procedure (Reddi et al., 2019; Wydmuch et al., 2018): for a multilabel example $(x, \{y_1, \ldots, y_t\})$ with t positive labels, we create t multiclass examples $\{(x, y_1), \ldots, (x, y_t)\}$.

We employ a fully connected neural network with a single hidden layer of width 512 using a linear activation as model architecture. We experiment with both the softmax cross-entropy and cosine contrastive loss, and employ the negative sampling techniques described in the previous section.

To assess the impact of different sampling and weighting schemes on different subpopulations of these datasets, we partition the multiclass test set ("Full") into three subpopultations based on the label frequency. These three subpopulations corresponds "Head", "Torso", and "Tail" labels, with decreasing frequency in training set. In particular, the Head subpopulation consists of those labels whose label frequencies are greater than the 66% quantile of the label frequencies in the training set. The label frequency for a label in the Torso subpopulation falls between the 66% and 33% quantile of the label frequencies in the training set. The remaining labels then constitute the Tail subpopulation.

Summary of findings. Figure 2 shows the results of the $\mathrm{Recall@}k$ metric for all three retrieval datasets when softmax cross-entropy loss and within-batch sampling are employed. We defer the results for uniform sampling to Appendix H. Furthermore, we also present the results for the $\mathrm{Precision@}k$ metric and its *propensity-scored* variant (Jain et al., 2016) on the original (multilabel) datasets in Appendix H. We make the following observations.

Within-batch sampling helps on the tail. Our results show that within-batch sampling consistently boosts the performance on the Tail subpopulation. This is in contrast with the uniform sampling (cf. Appendix H).

Tradeoff of Head versus Tail labels. It is evident from our results that there is a trade-off in between the model's performance on Head versus Tail subpopulation. In particular, within batch sampling and uniform sampling (which is aligned with the standard softmax cross-entropy loss on the retrieval datasets) boost the Tail and Head respectively.

Overall, the results on the long-tail and retrieval benchmarks confirm the implicit trade-offs associated with negative sampling, and illustrate that through suitable choice of (q, w), one may effectively tackle both sampling and labeling bias.

6. Future Work

We have shown that negative sampling methods — while devised for the problem of learning in large-output spaces — implicitly trade-off performance on dominant versus rare classes, and in particular, minimise losses that cope with long-tail label distributions. Potential future work includes devising instance-dependent samplers that can mimic (17), e.g., per Bamler & Mandt (2020); Bose et al. (2018); Cai & Wang (2018); Sun et al. (2019), and considering the labeling bias effects of generic instance-dependent weights.

Acknowledgements

The authors thank Ed Chi for valuable discussions and suggestions.

References

Agrawal, R., Gupta, A., Prabhu, Y., and Varma, M. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, pp. 13–24, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320351.

Alain, G., Lamb, A., Sankar, C., Courville, A. C., and Bengio, Y. Variance reduction in SGD by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

Bamler, R. and Mandt, S. Extreme classification via adversarial softmax approximation. In *International Conference on Learning Representations*, 2020.

Bengio, S., Dembczynski, K., Joachims, T., Kloft, M., and Varma, M. Extreme Classification (Dagstuhl Seminar 18291). *Dagstuhl Reports*, 8(7):62–80, 2019. ISSN 2192-5283.

Bengio, Y. and Senecal, J.-S. Quick training of probabilistic neural nets by importance sampling. In Bishop, C. M. and Frey, B. J. (eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, volume R4 of *Proceedings of Machine Learning Research*, pp. 17–24. PMLR, 03–06 Jan 2003. URL http://proceedings.mlr.press/r4/bengio03a.html. Reissued by PMLR on 01 April 2021.

Bengio, Y. and Senecal, J. S. Adaptive importance sampling to accelerate training of a neural probabilistic language model. *Trans. Neur. Netw.*, 19(4):713–722, April 2008. ISSN 1045-9227.

Blanc, G. and Rendle, S. Adaptive sampled softmax with kernel based sampling. In *Proceedings of the 35th Inter-*

national Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, pp. 589–598, 2018.

Bose, A. J., Ling, H., and Cao, Y. Adversarial contrastive estimation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1021–1032, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1094. URL https://www.aclweb.org/anthology/P18-1094.

Broscheit, S., Gashteovski, K., Wang, Y., and Gemulla, R. Can we predict new facts with open knowledge graph embeddings? a benchmark for open link prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2296–2308, Online, July 2020. Association for Computational Linguistics.

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint arXiv:1710.05381*, 2017.

Cai, L. and Wang, W. Y. KBGAN: Adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1470–1480, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1133. URL https://www.aclweb.org/anthology/N18-1133.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research (JAIR)*, 16:321–357, 2002.

Chen, T., Sun, Y., Shi, Y., and Hong, L. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 767–776, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874.

Covington, P., Adams, J., and Sargin, E. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pp. 191–198, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340359.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Fawcett, T. and Provost, F. Combining data mining and machine learning for effective user profiling. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 8–13. AAAI Press, 1996.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Gutmann, M. U. and Hyvärinen, A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(11):307–361, 2012.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742, 2006.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hidasi, B., Karatzoglou, A., Baltrunas, L., and Tikk, D. Session-based recommendations with recurrent neural networks. In Bengio, Y. and LeCun, Y. (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Jain, H., Prabhu, Y., and Varma, M. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 935–944, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2525–2534. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/katharopoulos18a.html.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., and Togneri, R. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. doi: 10.1109/TNNLS.2017.2732482.

Levy, O. and Goldberg, Y. Neural word embedding as implicit matrix factorization. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2177–2185. Curran Associates, Inc., 2014.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2537–2546. Computer Vision Foundation / IEEE, 2019.

Menon, A. K., Rawat, A. S., Reddi, S., and Kumar, S. Multilabel reductions: what is my loss optimising? In *Advances in Neural Information Processing Systems 32*, pp. 10600–10611. Curran Associates, Inc., 2019.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pp. 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

Morik, K., Brockhausen, P., and Joachims, T. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 268–277, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-612-2.

Morin, F. and Bengio, Y. Hierarchical probabilistic neural network language model. In Cowell, R. G. and Ghahramani, Z. (eds.), *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pp. 246–252. PMLR, 06–08 Jan 2005. URL http://proceedings.mlr.press/r5/morin05a.html. Reissued by PMLR on 30 March 2021.

Prabhu, Y. and Varma, M. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pp. 263–272, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569.

Rawat, A. S., Chen, J., Yu, F. X. X., Suresh, A. T., and Kumar, S. Sampled softmax with random fourier features. In *Advances in Neural Information Processing Systems*, pp. 13857–13867, 2019.

Reddi, S. J., Kale, S., Yu, F., Holtmann-Rice, D., Chen, J., and Kumar, S. Stochastic negative mining for learning with large output spaces. In Chaudhuri, K. and Sugiyama, M. (eds.), *International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1940–1949. PMLR, 16–18 Apr 2019.

Ren, J., Yu, C., sheng, s., Ma, X., Zhao, H., Yi, S., and Li, h. Balanced meta-softmax for long-tailed visual recognition. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4175–4186. Curran Associates, Inc., 2020.

Ruiz, F., Titsias, M., Dieng, A. B., and Blei, D. Augment and reduce: Stochastic inference for large categorical distributions. In Dy, J. and Krause, A. (eds.), *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4403–4412, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgEQnRqYQ.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., and Yan, J. Equalization loss for long-tailed object recognition. *arXiv preprint arXiv:2003.05176*, 2020.

Van Horn, G. and Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

Varma, M. Extreme classification repository. Website, 8 2018. http://manikvarma.org/downloads/XC/XMLRepository.html.

Wallace, B., K.Small, Brodley, C., and Trikalinos, T. Class imbalance, redux. In *Proc. ICDM*, 2011.

Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C. C., and Lin, D. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Wang, Q., Mao, Z., Wang, B., and Guo, L. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, 2017.

Wydmuch, M., Jasinska, K., Kuznetsov, M., Busa-Fekete, R., and Dembczynski, K. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *Advances in Neural Information Processing Systems 31*, pp. 6355–6366. Curran Associates, Inc., 2018.

Xie, Y. and Manski, C. F. The logit model and response-based samples. *Sociological Methods & Research*, 17(3): 283–302, 1989.

Xu, Z., Chen, C., Lukasiewicz, T., Miao, Y., and Meng, X. Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, pp. 1921–1924, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340731.

Yi, X., Yang, J., Hong, L., Cheng, D. Z., Heldt, L., Kumthekar, A., Zhao, Z., Wei, L., and Chi, E. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 269–277, New York, NY, USA, 2019a. Association for Computing Machinery. ISBN 9781450362436.

Yi, X., Yang, J., Hong, L., Cheng, D. Z., Heldt, L., Kumthekar, A., Zhao, Z., Wei, L., and Chi, E. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, pp. 269–277, New York, NY, USA, 2019b. Association for Computing Machinery. ISBN 9781450362436.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

Zhou, Z.-H. and Liu, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(1), 2006.

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

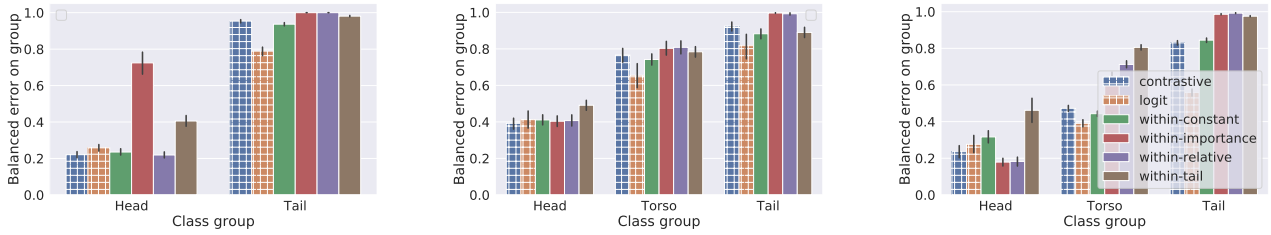
Figure 3 shows results using the contrastive loss on the long-tail benchmarks. Here, the performance of different sampling schemes is more variable compared to the softmax cross-entropy. In particular, on Tail classes, the performance of sampling is generally poor compared to the baseline. Note that 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

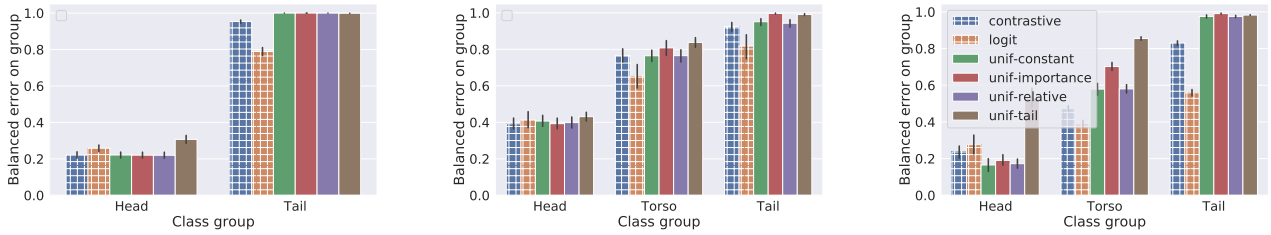
Figures 4 and 5 present the balanced errors of the various choices of sampling and weighting schemes, for the softmax cross-entropy and contrastive loss, respectively. We see that the gains of within-batch sampling with constant weighting are such that it can improve over the standard loss using *all* 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.

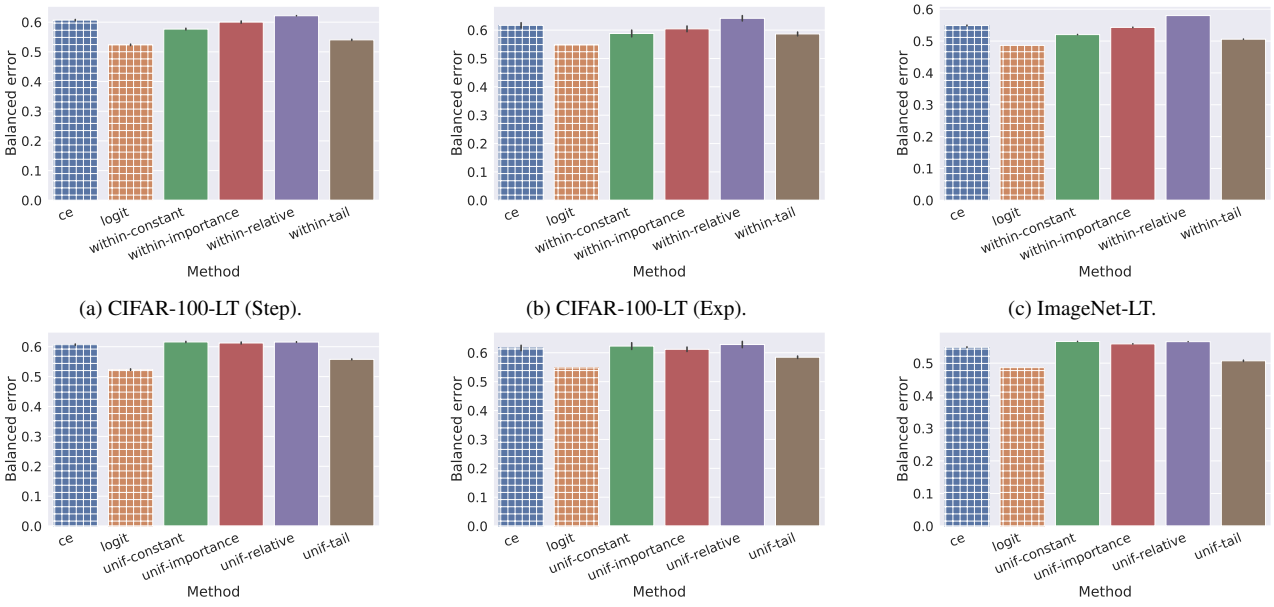


(a) CIFAR-100-LT Step (contrastive loss). (b) CIFAR-100-LT Exp (contrastive loss). (c) ImageNet-LT (contrastive loss).



(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.



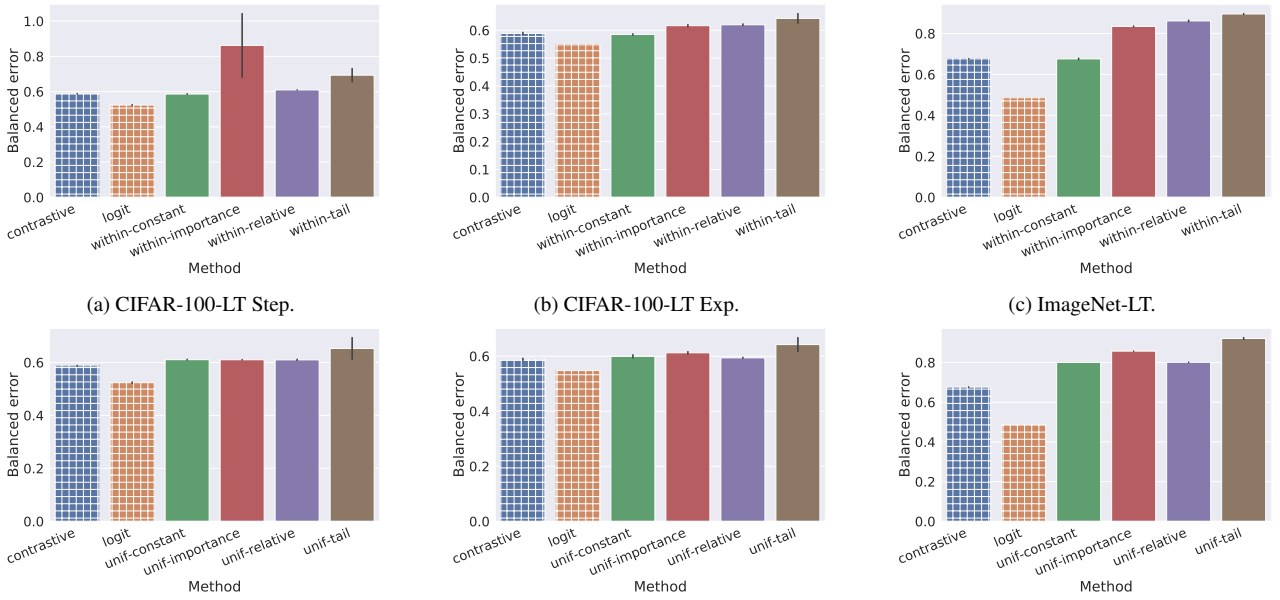
(a) CIFAR-100-LT (Step). (b) CIFAR-100-LT (Exp). (c) ImageNet-LT.

(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

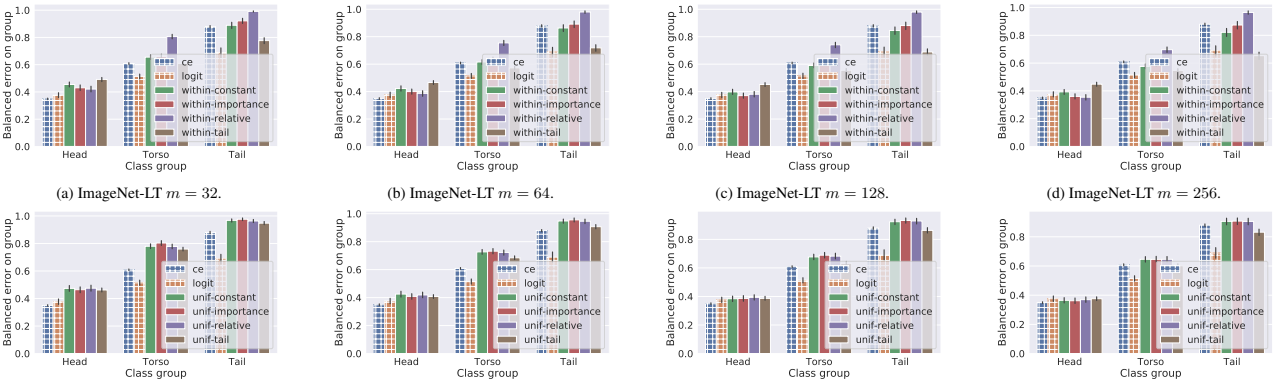
See Figure 7.



(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.

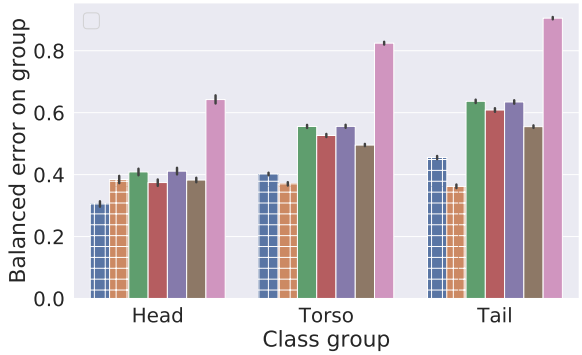
Figure 5: Balanced error on long-tail learning benchmarks using the contrastive loss. We present results for within-batch (`within`) negative sampling (Fig. 5a - 5c) and uniform (`unif`) negative sampling (Fig. 5d - 5f), using the constant weight (`const`), importance weighting (`importance`), and relative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



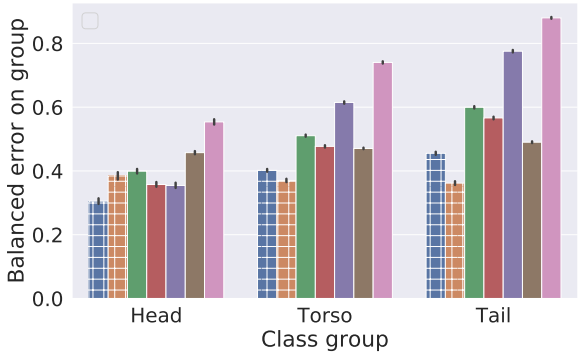
(a) ImageNet-LT $m = 32$. (b) ImageNet-LT $m = 64$. (c) ImageNet-LT $m = 128$. (d) ImageNet-LT $m = 256$.

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

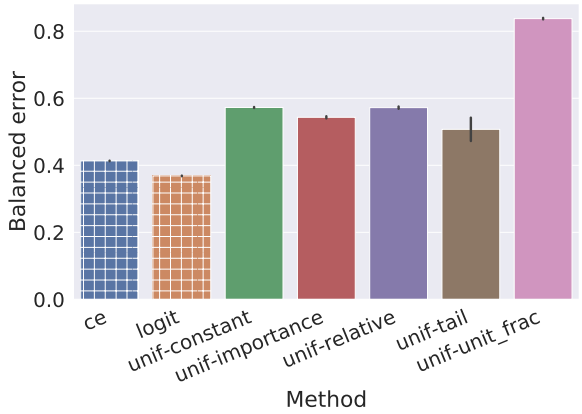
Figure 6: Results on head, torso and tail labels on ImageNet-LT, using varying number of sampled negatives. Fig. 6a - 6d show the performance of within-batch negative sampling along with baseline loss functions. Fig. 6e - 6h illustrate the performance of uniform negative sampling.



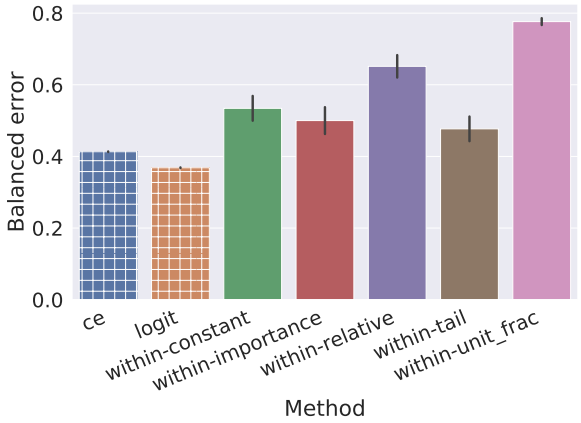
(a) CE, Uniform



(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.

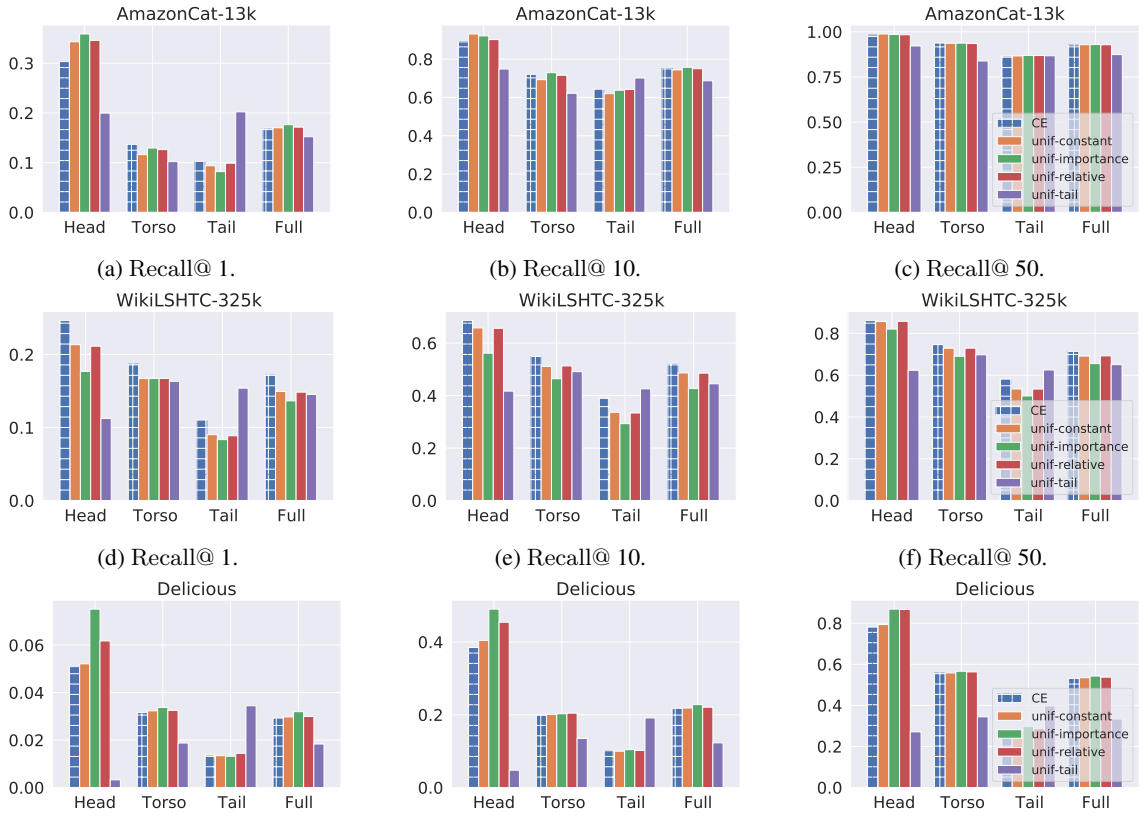
G. Retrieval Datasets

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

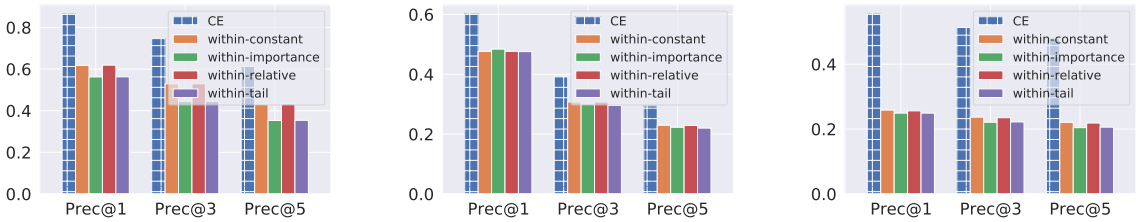


(a) Recall@ 1. (b) Recall@ 10. (c) Recall@ 50.

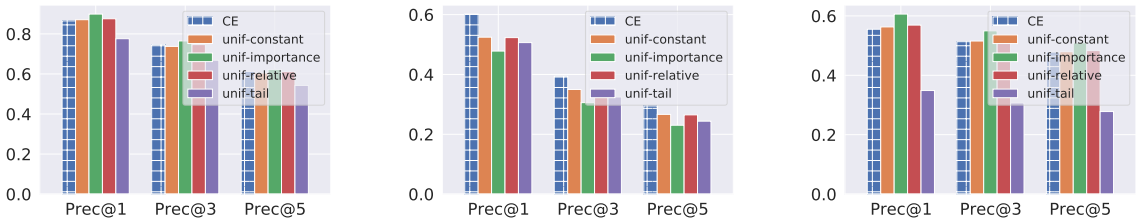
(d) Recall@ 1. (e) Recall@ 10. (f) Recall@ 50.

(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).

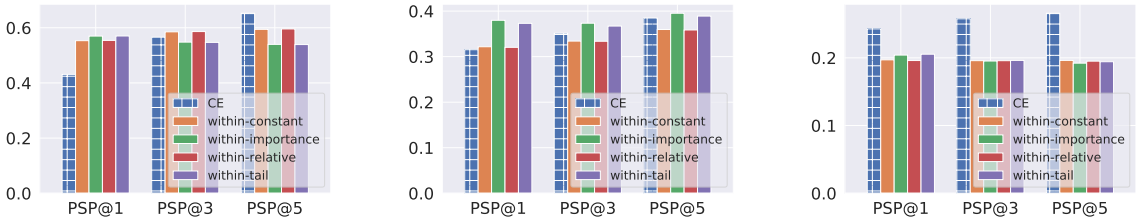


(a) AMAZONCAT-13K. (b) WIKILSHTC-325K. (c) DELICIOUS.

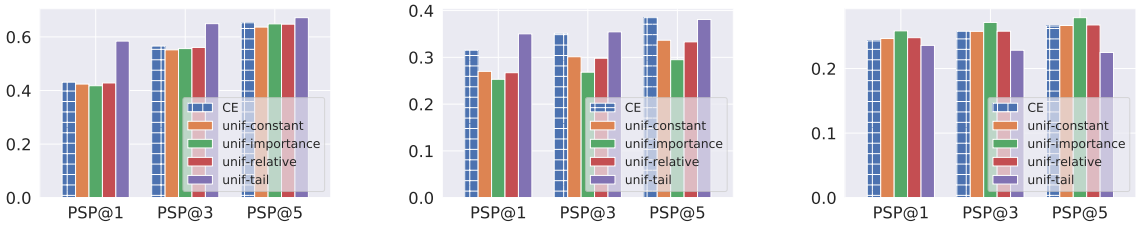


(d) AMAZONCAT-13K. (e) WIKILSHTC-325K. (f) DELICIOUS.

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).