LONG-TAIL LEARNING VIA LOGIT ADJUSTMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world classification problems typically exhibit an *imbalanced* or *long-tailed* label distribution, wherein many labels have only a few associated samples. This poses a challenge for generalisation on such labels, and also makes naïve learning biased towards dominant labels. In this paper, we present a statistical framework that unifies and generalises several recent proposals to cope with these challenges. Our framework revisits the classic idea of *logit adjustment* based on the label frequencies, which encourages a large *relative margin* between logits of rare positive versus dominant negative labels. This yields two techniques for long-tail learning, where such adjustment is either applied post-hoc to a trained model, or enforced in the loss during training. These techniques are statistically grounded, and practically effective on four real-world datasets with long-tailed label distributions.

1 INTRODUCTION

Real-world classification problems typically exhibit a *long-tailed* label distribution, wherein most labels are associated with only a few samples (Van Horn & Perona, 2017; Buda et al., 2017; Liu et al., 2019). Owing to this paucity of samples, generalisation on such labels is challenging; moreover, naïve learning on such data is susceptible to an undesirable bias towards dominant labels. This problem has been widely studied in the literature on learning under *class imbalance* (Kubat et al., 1997; Chawla et al., 2002; He & Garcia, 2009), and the related problem of *cost-sensitive learning* (Elkan, 2001).

Recently, long-tail learning has received renewed interest in the context of neural networks. Two active strands of work involve post-hoc normalisation of the classification weights (Zhang et al., 2019; Kim & Kim, 2019; Kang et al., 2020; Ye et al., 2020), and modification of the underlying loss to account for varying class penalties (Zhang et al., 2017; Cui et al., 2019; Cao et al., 2019; Tan et al., 2020). Each of these strands are intuitive, and have proven empirically successful. However, they are not without limitation: e.g., weight normalisation crucially relies on the weight norms being smaller for rare classes; however, this assumption is sensitive to the choice of optimiser (see §2). On the other hand, loss modification sacrifices the *consistency* that underpins the canonical softmax cross-entropy (see §5.1). Consequently, such techniques may prove suboptimal even in simple settings (see §6.1).

In this paper, we establish a statistical framework for long-tail learning that offers a unified view of post-hoc normalisation and loss modification techniques, while overcoming their limitations. Our framework revisits the classic idea of *logit adjustment* based on label frequencies (Provost, 2000; Zhou & Liu, 2006; Collell et al., 2016), which encourages a large *relative margin* between a pair of rare positive and dominant negative labels. Such adjustment can be achieved by shifting the learned logits post-hoc, or augmenting the softmax cross-entropy with a *pairwise label margin* (cf. (11)). While similar in nature to recent techniques, our logit adjustment approaches additionally have a firm statistical grounding: they are *Fisher consistent* for minimising the *balanced error* (cf. (2)), a common metric in long-tail settings which averages the per-class errors. This statistical grounding translates into strong empirical performance on four real-world datasets with long-tailed label distributions.

In summary, our contributions are:

(i) we establish a statistical framework for long-tail learning (§3) based on *logit adjustment* that provides a unified view of post-hoc correction and loss modification

(ii) we present two realisations of logit adjustment, applied either post-hoc (§4.1) or during training (§5.1); unlike recent proposals (Table 1), these are consistent for minimising the balanced error

(iii) we confirm the efficacy of the proposed logit adjustment techniques compared to several baselines on four real-world datasets with long-tailed label distributions (§6).

Method	Procedure	Consistent?	Reference
Weight normalisation	Post-hoc weight scaling	×	(Kang et al., 2020)
Adaptive margin	Softmax with rare +ve upweighting	×	(Cao et al., 2019)
Equalised margin	Softmax with rare -ve downweighting	×	(Tan et al., 2020)
Logit-adjusted threshold	Post-hoc logit translation	✓	This paper (cf. (9))
Logit-adjusted loss	Softmax with logit translation	✓	This paper (cf. (10))

Table 1: Comparison of approaches to long-tail learning. Weight normalisation re-scales the classification weights; by contrast, we *add* per-label offsets to the logits. Margin approaches uniformly increase the margin between a rare positive and all negatives (Cao et al., 2019), or decrease the margin between all positives and a rare negative (Tan et al., 2020) to prevent rare labels' gradient suppression. By contrast, we increase the margin between a *rare* positive and a *dominant* negative.

2 Problem setup and related work

Consider a multiclass classification problem with instances $\mathcal{X}$ and labels $\mathcal{Y} = [L] \doteq \{1, 2, \ldots, L\}$. Given a sample $S = \{(x_n, y_n)\}_{n=1}^N \sim \mathbb{P}^N$ for unknown distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$, our goal is to learn a scorer $f \colon \mathcal{X} \to \mathbb{R}^L$ that minimises the misclassification error $\mathbb{P}_{x,y}\big(y \notin \operatorname{argmax}_{y' \in \mathcal{Y}} f_{y'}(x)\big)$. Typically, one minimises a surrogate loss $\ell \colon \mathcal{Y} \times \mathbb{R}^L \to \mathbb{R}$ such as the softmax cross-entropy,

$$\ell(y, f(x)) = \log\Big[\sum\nolimits_{y' \in [L]} e^{f_{y'}(x)}\Big] - f_y(x) = \log\Big[1 + \sum\nolimits_{y' \neq y} e^{f_{y'}(x) - f_y(x)}\Big]. \tag{1}$$

We may view the resulting softmax probabilities $p_y(x) \propto e^{f_y(x)}$ as estimates of $\mathbb{P}(y \mid x)$.

The setting of *learning under class imbalance* or *long-tail learning* is where the distribution $\mathbb{P}(y)$ is highly skewed, so that many rare (or "tail") labels have a low probability of occurrence. Here, the misclassification error is not a suitable measure of performance: a trivial predictor which classifies every instance to the majority label will attain a low misclassification error. To cope with this, a natural alternative is the balanced error (Chan & Stolfo, 1998; Brodersen et al., 2010; Menon et al., 2013), which averages each of the per-class error rates: under a uniform label distribution

$$\operatorname{BER}(f) \doteq \frac{1}{L} \sum\nolimits_{y \in [L]} \mathbb{P}_{x|y}\big(y \notin \operatorname{argmax}_{y' \in \mathcal{Y}} f_{y'}(x)\big). \tag{2}$$

This can be seen as implicitly using a *balanced* class-probability function $\mathbb{P}^{\text{bal}}(y \mid x) \propto \frac{1}{L} \cdot \mathbb{P}(x \mid y)$, as opposed to the native $\mathbb{P}(y \mid x) \propto \mathbb{P}(y) \cdot \mathbb{P}(x \mid y)$ that is employed in the misclassification error.

Broadly, extant approaches to coping with class imbalance modify:

(i) the *inputs* to a model, for example by over- or under-sampling (Kubat & Matwin, 1997; Chawla et al., 2002; Wallace et al., 2011; Mikolov et al., 2013; Mahajan et al., 2018; Yin et al., 2018);

(ii) the *outputs* of a model, for example by post-hoc correction of the decision threshold (Fawcett & Provost, 1996; Collell et al., 2016) or weights (Kim & Kim, 2019; Kang et al., 2020); or

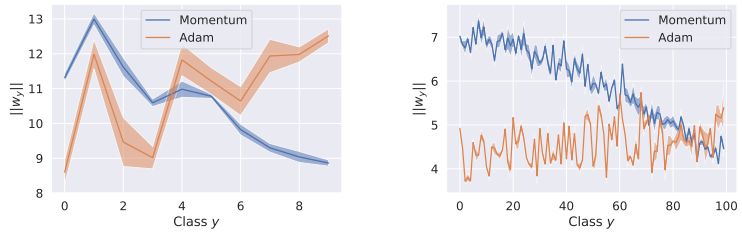
(iii) the *training procedure* of a model, for example by modifying the loss function (Zhang et al., 2017; Cui et al., 2019; Cao et al., 2019; Tan et al., 2020; Jamal et al., 2020).

One may easily combine approaches from the first stream with those from the latter two. Consequently, we focus on the latter two in this work, and describe some representative recent examples from each.

Post-hoc weight normalisation. Suppose $f_y(x) = w_y^\top \Phi(x)$ for classification weights $w_y \in \mathbb{R}^D$ and representations $\Phi \colon \mathcal{X} \to \mathbb{R}^D$, as learned by a neural network. (We may add per-label bias terms to f_y by adding a constant feature to Φ.) A fruitful avenue of exploration involves decoupling of representation and classifier learning (Zhang et al., 2019). Concretely, we first learn $\{w_y, \Phi\}$ via standard training on the long-tailed training sample S, and then for $x \in \mathcal{X}$ predict the label

$$\operatorname{argmax}_{y \in [L]} w_y^\top \Phi(x) / \nu_y^\tau = \operatorname{argmax}_{y \in [L]} f_y(x) / \nu_y^\tau, \tag{3}$$

for $\tau > 0$, where $\nu_y = \mathbb{P}(y)$ in Kim & Kim (2019); Ye et al. (2020) and $\nu_y = \|w_y\|_2$ in Kang et al. (2020). Intuitively, either choice of ν_y upweights the contribution of rare labels through *weight*



(a) CIFAR-10-LT. (b) CIFAR-100-LT.

Figure 1: Mean and standard deviation of per-class weight norms $\|w_y\|_2$ over 5 runs for a ResNet-32 under momentum and Adam optimisers. We use long-tailed ("LT") versions of CIFAR-10 and CIFAR-100, and sort classes in descending order of frequency; the first class is $100\times$ more likely to appear than the last class (see §6.2). Both optimisers yield comparable balanced error. However, the weight norms have incompatible trends: under momentum, the norms are strongly correlated with class frequency, while with Adam, the norms are *anti-correlated* or *independent* of the class frequency. Consequently, weight normalisation under Adam is ineffective for combatting class imbalance.

normalisation. The choice $\nu_y = \|w_y\|_2$ is motivated by the observations that $\|w_y\|_2$ tends to correlate with $\mathbb{P}(y)$. Further to the above, one may enforce $\|w_y\|_2 = 1$ during training (Kim & Kim, 2019).

Loss modification. A classic means of coping with class imbalance is to *balance* the loss, wherein $\ell(y, f(x))$ is weighted by $\mathbb{P}(y)^{-1}$ (Xie & Manski, 1989; Morik et al., 1999): e.g., applied to (1),

$$\ell(y, f(x)) = \frac{1}{\mathbb{P}(y)} \cdot \log \Big[1 + \sum\nolimits_{y' \neq y} e^{f_{y'}(x) - f_y(x)}\Big]. \tag{4}$$

While intuitive, balancing has minimal effect in separable settings: solutions that achieve zero training loss will necessarily remain optimal even under weighting (Byrd & Lipton, 2019). Intuitively, one would instead like to shift the separator closer to a dominant class. Li et al. (2002); Wu et al. (2008); Masnadi-Shirazi & Vasconcelos (2010) thus proposed to add *per-class margins* into the hinge loss. Cao et al. (2019) similarly proposed to add a per-class margin into the softmax cross-entropy:

$$\ell(y, f(x)) = \log \Big[1 + \sum\nolimits_{y' \neq y} e^{\delta_y} \cdot e^{f_{y'}(x) - f_y(x)}\Big], \tag{5}$$

where $\delta_y \propto \mathbb{P}(y)^{-1/4}$. This upweights rare "positive" y to encourage a larger gap $f_y(x) - f_{y'}(x)$, i.e., the margin between y and any "negative" $y' \neq y$. Separately, Tan et al. (2020) proposed

$$\ell(y, f(x)) = \log \Big[1 + \sum\nolimits_{y' \neq y} e^{\delta_{y'}} \cdot e^{f_{y'}(x) - f_y(x)}\Big], \tag{6}$$

where $\delta_{y'} \leq 0$ is an non-decreasing transform of $\mathbb{P}(y')$. Note that in the original softmax cross-entropy with $\delta_{y'} = 0$, a rare label often receives a strong *inhibitory* gradient signal as it disproportionately appear as a negative for dominant labels. This can be modulated by letting $\delta_{y'} \ll 0$.

Limitations of existing approaches. Each of the above methods are intuitive, and have shown strong empirical performance. However, a closer analysis identifies some subtle limitations.

Limitations of weight normalisation. Post-hoc weight normalisation with $\nu_y = \|w_y\|_2$ per Kang et al. (2020) is motivated by the observation that the weight norm $\|w_y\|_2$ tends to correlate with $\mathbb{P}(y)$. However, this assumption is highly dependent on the choice of optimiser, as Figure 1 illustrates: for ResNet-32 models trained on long-tailed versions of CIFAR-10 and CIFAR-100, when using the Adam optimiser, the norms are either *anti*-correlated or *independent* of $\mathbb{P}(y)$. Weight normalisation thus cannot achieve the desired effect of boosting rare labels' scores. One may hope to side-step this by simply using $\nu_y = \mathbb{P}(y)$; unfortunately, this choice has more subtle limitations (see §4.2).

Limitations of loss modification. Enforcing a per-label margin per (5) and (6) is intuitive, as it allows for shifting the decision boundary away from rare classes. However, when doing so, it is important to ensure *Fisher consistency* (Lin, 2004) (or *classification calibration* (Bartlett et al., 2006)) of the resulting loss for the balanced error. That is, the minimiser of the expected loss (equally, the empirical risk in the infinite sample limit) should result in a minimal balanced error. Unfortunately, both (5) and (6) are *not* consistent in this sense, even for binary problems; see §5.1, §6.1 for details.

3 LOGIT ADJUSTMENT FOR LONG-TAIL LEARNING: A STATISTICAL VIEW

The above suggests that there is scope for improving performance on long-tail problems, both in terms of post-hoc correction and loss modification. We now show how a statistical perspective on the problem suggests simple procedures of each type, which overcome the limitations discussed above.

Recall that our goal is to minimise the balanced error (2). A classical result is that the *best possible* or *Bayes-optimal* scorer for this problem, i.e., $f^* \in \operatorname{argmin}_{f\colon \mathcal{X} \to \mathbb{R}^L} \operatorname{BER}(f)$, satisfies the following (Menon et al., 2013), (Koyejo et al., 2014, Corollary 4), (Collell et al., 2016, Theorem 1):

$$\operatorname{argmax}_{y \in [L]} f^*_y(x) = \operatorname{argmax}_{y \in [L]} \mathbb{P}^{\text{bal}}(y \mid x) = \operatorname{argmax}_{y \in [L]} \mathbb{P}(x \mid y), \tag{7}$$

where $\mathbb{P}^{\text{bal}}$ is the balanced class-probability as per §2. In words, the Bayes-optimal prediction is the label under which the given instance $x \in \mathcal{X}$ is most likely. Consequently, for fixed class-conditionals $\mathbb{P}(x \mid y)$, varying the *class priors* $\mathbb{P}(y)$ arbitrarily will not affect the optimal scorers. This is intuitively desirable: the balanced error is agnostic to the level of imbalance in the label distribution.

To further probe (7), suppose the underlying class-probabilities $\mathbb{P}(y \mid x) \propto \exp(s^*_y(x))$, for scorer $s^* \colon \mathcal{X} \to \mathbb{R}^L$. Since by definition $\mathbb{P}^{\text{bal}}(y \mid x) \propto \mathbb{P}(y \mid x)/\mathbb{P}(y)$, (7) becomes

$$\operatorname{argmax}_{y \in [L]} \mathbb{P}^{\text{bal}}(y \mid x) = \operatorname{argmax}_{y \in [L]} \exp(s^*_y(x))/\mathbb{P}(y) = \operatorname{argmax}_{y \in [L]} s^*_y(x) - \ln \mathbb{P}(y), \tag{8}$$

i.e., we translate the optimal logits based on the class priors. While s^* is of course unknown in practice, (8) immediately suggests two means of optimising for the balanced error:

(i) train a model to estimate the standard $\mathbb{P}(y \mid x)$ (e.g., by minimising the standard softmax-cross entropy on the long-tailed data), and then explicitly modify its logits post-hoc as per (8)

(ii) train a model to estimate the balanced $\mathbb{P}^{\text{bal}}(y \mid x)$, whose logits are implicitly modified as per (8).

Such *logit adjustment* techniques — which have been a classic approach to class-imbalance (Provost, 2000) — neatly align with the post-hoc and loss modification streams discussed in §2. However, unlike most previous techniques from these streams, logit adjustment is endowed with a clear statistical grounding: by construction, the optimal solution under such adjustment coincides with the Bayes-optimal solution (7) for the balanced error, i.e., it is *Fisher consistent* for minimising the balanced error. We now study each of the techniques (i) and (ii) in turn.

4 POST-HOC LOGIT ADJUSTMENT

We now propose a post-hoc logit adjustment scheme for a classifier trained on long-tailed data. We further show this has subtle advantages over recent weight normalisation schemes.

4.1 THE POST-HOC LOGIT ADJUSTMENT PROCEDURE

When employing the softmax cross-entropy to train a neural network, we aim to approximate the underlying $\mathbb{P}(y \mid x)$ with $p_y(x) \propto \exp(f_y(x))$ for logits $f_y(x) = w_y^\top \Phi(x)$. Given learned $\{w, \Phi\}$, one typically predicts the label $\operatorname{argmax}_{y \in [L]} f_y(x)$, i.e., the most likely label under the model's $\mathbb{P}(y \mid x)$. In *post-hoc logit adjustment*, we propose to instead predict, for suitable $\tau > 0$:

$$\operatorname{argmax}_{y \in [L]} \exp(w_y^\top \Phi(x))/\pi_y^\tau = \operatorname{argmax}_{y \in [L]} f_y(x) - \tau \cdot \log \pi_y, \tag{9}$$

where $\pi \in \Delta_L$ (for simplex Δ) are estimates of the class priors $\mathbb{P}(y)$, e.g., the empirical frequencies on the training sample S. Effectively, (9) adds a label-dependent offset to each of the logits.

When $\tau = 1$, this can be seen as applying (8) with a plugin estimate of $\mathbb{P}(y \mid x)$, i.e., $p_y(x) \propto \exp(w_y^\top \Phi(x))$. When $\tau \neq 1$, this can be seen as applying (8) to *temperature scaled* estimates $\bar{p}_y(x) \propto \exp(\tau^{-1} \cdot w_y^\top \Phi(x))$. To unpack this, recall that (8) justifies post-hoc logit thresholding given access to the true probabilities $\mathbb{P}(y \mid x)$. In practice, high-capacity neural networks often produce uncalibrated estimates of these probabilities (Guo et al., 2017). Temperature scaling is a means to calibrate the estimates, and is routinely employed for distillation (Hinton et al., 2015).

One may treat τ as a tuning parameter to be chosen based on holdout calibration, e.g., the expected calibration error (Murphy & Winkler, 1987; Guo et al., 2017), probabilistic sharpness (Gneiting et al.,

2007; Kuleshov et al., 2018), or a proper scoring rule such as the log-loss or squared error (Gneiting & Raftery, 2007). One may alternately fix $\tau = 1$ and aim to learn inherently calibrated probabilities, e.g., via label smoothing (Szegedy et al., 2016; Müller et al., 2019).

Post-hoc logit adjustment with $\tau = 1$ is not a new idea in the classical label imbalance literature (Fawcett & Provost, 1996; Provost, 2000; Maloof, 2003; Zhou & Liu, 2006; Collell et al., 2016); however, it has had limited exploration in the recent long-tail learning literature. Further, the case $\tau \neq 1$ is important in practical usage of neural networks, owing to their typical lack of probabilistic calibration (Guo et al., 2017). Interestingly, post-hoc logit adjustment also has an important advantage over recently proposed weight normalisation techniques, as we now discuss.

4.2 Comparison to post-hoc weight normalisaiton

Recall that weight normalisation involves learning logits $f_y(x) = w_y^\top \Phi(x)$, and then post-hoc normalising the weights via w_y / ν_y^τ for $\tau > 0$. We demonstrated in §2 that using $\nu_y = \|w_y\|_2$ may be ineffective when using adaptive optimisers. However, even with $\nu_y = \pi_y$, there is a subtle contrast to post-hoc logit adjustment: while the former performs a *multiplicative* update to the logits, the latter performs an *additive* update. The two techniques may thus yield different orderings over labels, since

$$w_1^\top \Phi(x)/\pi_1 < w_2^\top \Phi(x)/\pi_2 \not\Rightarrow \not\Leftarrow \exp(w_1^\top \Phi(x))/\pi_1 < \exp(w_2^\top \Phi(x))/\pi_2.$$

Observe that if a rare label y has *negative* score $w_y^\top \Phi(x) < 0$, and there is another label with positive score, then it is *impossible* for the weight normalisation to give y the highest score. By contrast, under logit adjustment, $w_y^\top \Phi(x) - \ln \pi_y$ will be lower for dominant classes, regardless of the original sign. Weight normalisation is thus *not* consistent for the balanced error, unlike logit adjustment.

5 The logit adjusted softmax cross-entropy

We now show how to directly encode logit adjustment into the softmax cross-entropy. The resulting approach has an intuitive relation to existing loss modification techniques.

5.1 The logit adjusted loss

From §3, the second approach to optimising for the balanced error is to directly model $\mathbb{P}^{\text{bal}}(y \mid x) \propto \mathbb{P}(y \mid x)/\mathbb{P}(y)$. To do so, consider the following *logit adjusted softmax cross-entropy loss* for $\tau > 0$:

$$\ell(y, f(x)) = -\log \frac{e^{f_y(x) + \tau \cdot \log \pi_y}}{\sum_{y' \in [L]} e^{f_{y'}(x) + \tau \cdot \log \pi_{y'}}} = \log \left[1 + \sum\nolimits_{y' \neq y} \left(\frac{\pi_{y'}}{\pi_y} \right)^\tau \cdot e^{f_{y'}(x) - f_y(x)} \right]. \quad (10)$$

Given a scorer that minimises the above, we now predict $\operatorname{argmax}_{y \in [L]} f_y(x)$ as usual.

Compared to the standard softmax cross-entropy (1), the above applies a *label-dependent offset* to each logit. Compared to (9), we *directly* enforce the class prior offset while learning the logits, rather than doing this post-hoc. The two approaches have a deeper connection: observe that (10) is equivalent to using a scorer of the form $g_y(x) = f_y(x) + \tau \cdot \log \pi_y$, with $\operatorname{argmax}_{y \in [L]} f_y(x) = \operatorname{argmax}_{y \in [L]} g_y(x) - \tau \cdot \log \pi_y$. Consequently, one can equivalently view learning with this loss as learning a standard scorer $g(x)$, and post-hoc adjusting its logits. For non-convex objectives, as encountered in neural networks, the bias endowed by adding $\tau \cdot \log \pi_y$ to the logits is likely to result in a different local minima, typically with improved performance.

For more insight into the logit adjusted loss, consider the following *pairwise margin loss*

$$\ell(y, f(x)) = \alpha_y \cdot \log \left[1 + \sum\nolimits_{y' \neq y} e^{\Delta_{yy'}} \cdot e^{(f_{y'}(x) - f_y(x))} \right], \quad (11)$$

for label weights $\alpha_y > 0$, and *pairwise label margins* $\Delta_{yy'}$ representing the desired gap between scores for y and y'. For $\tau = 1$, our logit adjusted loss (10) corresponds to (11) with $\alpha_y = 1$ and $\Delta_{yy'} = \log \left(\frac{\pi_{y'}}{\pi_y} \right)$. This demands a larger margin between *rare* positive ($\pi_y \sim 0$) and *dominant* negative ($\pi_{y'} \sim 1$) labels, so that scores for dominant classes do not overwhelm those for rare ones.

Existing loss modification techniques can be viewed as special cases of (11). For example, $\alpha_y = 1/\pi_y$ and $\Delta_{yy'} = 0$ yields the balanced loss (4). When $\alpha_y = 1$, the choice $\Delta_{yy'} = \pi_y^{-1/4}$ yields (5). Finally, $\Delta_{yy'} = \log F(\pi_{y'})$ yields (6), where $F \colon [0,1] \to (0,1]$ is some non-decreasing function. These losses thus either consider the frequency of the positive y or negative y', but not *both*. Remarkably, the specific choice that leads to our loss in (10) has a firm statistical grounding: it ensures Fisher consistency (in the sense of, e.g., Bartlett et al. (2006)) for the balanced error.

Theorem 1. *For any $\delta \in \mathbb{R}_+^L$, the pairwise loss in (11) is Fisher consistent with weights and margins*

$$\alpha_y = \delta_y / \mathbb{P}(y) \qquad \Delta_{yy'} = \log\left(\delta_{y'}/\delta_y\right).$$

Letting $\delta_y = \pi_y$, we immediately deduce that the logit-adjusted loss of (10) is consistent, *provided* our π_y is a consistent estimate of $\mathbb{P}(y)$. Similarly, $\delta_y = 1$ recovers the classic result that the balanced loss is consistent. While Theorem 1 only provides a sufficient condition in multi-class setting, one can provide a necessary and sufficient condition that rules out other choices of Δ in the binary case.

Theorem 2. *Suppose $\mathcal{Y} = \{\pm 1\}$. Let $\delta_y \doteq \Delta_{y,-y}$, and $\sigma(z) = (1 + \exp(z))^{-1}$. Then, the pairwise margin loss in (11) is Fisher consistent for the balanced error iff*

$$\frac{\alpha_{+1}}{\alpha_{-1}} \cdot \frac{\sigma(\delta_{+1})}{\sigma(\delta_{-1})} = \frac{1 - \mathbb{P}(y = +1)}{\mathbb{P}(y = +1)}.$$

5.2 Discussion and Extensions

Our pairwise margin loss in (11) subsumes several existing loss-correction approaches in the literature. Further, it suggests the exploration of new choices of Δ. For example, we shall see the efficacy of combining the Δ implicit in the adaptive loss of Cao et al. (2019) with our proposed Δ. One may also generalise the formulation in Theorem 1, and employ $\Delta_{yy'} = \tau_1 \cdot \log \pi_y - \tau_2 \cdot \log \pi_{y'}$, where $\tau_1, \tau_2 > 0$. This interpolates between our loss ($\tau_1 = \tau_2$) and a version of the equalised loss ($\tau_1 = 0$).

For $\tau = -1$, a similar loss to (10) has been considered in the context of *negative sampling* for scalability (Yi et al., 2019): here, one samples a subset of negatives based on π, and corrects the logits to obtain an unbiased estimate of the loss based on all negatives (Bengio & Senecal, 2008). Losses of the general form (11) have also been explored for structured prediction (Pletscher et al., 2010).

Cao et al. (2019, Theorem 2) provides a rigorous generalisation bound for the adaptive margin loss under the assumption of separable data with binary labels. The inconsistency of the loss with respect to the balanced error concerns the more general scenario of non-separable multiclass data, which may occur, e.g., owing to label noise or limited model capacity. We shall subsequently demonstrate that encouraging consistency is not merely of theoretical interest, and can lead to gains in practice.

6 Experimental Results

We now present experiments confirming our main claims: (i) on simple binary problems, existing weight normalisation and loss modification techniques may not converge to the optimal solution (§6.1); (ii) on real-world datasets, our post-hoc logit adjustment generally outperforms weight normalisation, and one can obtain further gains via our logit adjusted softmax cross-entropy (§6.2).

6.1 Results on Synthetic Dataset

We begin with a binary classification task, wherein samples from class $y \in \{\pm 1\}$ are drawn from a 2D Gaussian with isotropic covariance and means $\mu_y = y \cdot (+1, +1)$. We introduce class imbalance by setting $\mathbb{P}(y = +1) = 5\%$. The Bayes-optimal classifier for the balanced error is (see Appendix F)

$$f^*(x) = +1 \iff \mathbb{P}(x \mid y = +1) > \mathbb{P}(x \mid y = -1) \iff (\mu_1 - \mu_{-1})^\top x > 0, \tag{12}$$

i.e., it is a linear separator passing through the origin. We compare this separator against those found by several margin losses based on (11): standard ERM ($\Delta_{yy'} = 0$), the adaptive loss (Cao et al., 2019) ($\Delta_{yy'} = \pi_y^{-1/4}$), an instantiation of the equalised loss (Tan et al., 2020) ($\Delta_{yy'} = \log \pi_{y'}$), and our logit adjusted loss ($\Delta_{yy'} = \log \frac{\pi_{y'}}{\pi_y}$). For each loss, we train an affine classifier on a sample of $10,000$ instances, and evaluate the balanced error on a test set of $10,000$ samples over 100 independent trials.

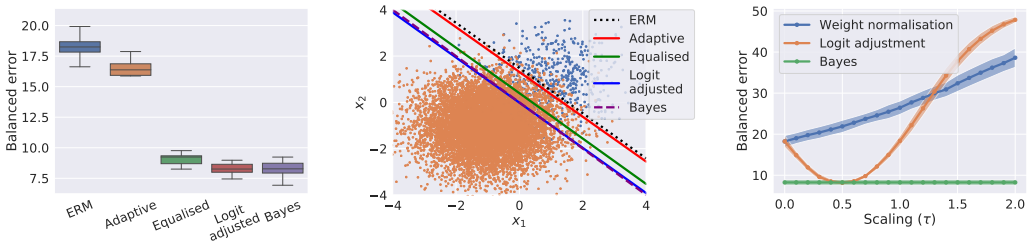


Figure 2: Results on synthetic binary classification problem. Our logit adjusted loss tracks the Bayes-optimal solution and separator (left & middle panel). Post-hoc logit adjustment matches the Bayes performance with suitable scaling (right panel); however, *any* weight normalisation fails.

Method	**CIFAR-10-LT**	**CIFAR-100-LT**	**ImageNet-LT**	**iNaturalist**
ERM	27.16	61.64	53.11	38.66
Weight normalisation ($\tau = \tau^*$) (Kang et al., 2020)	21.50	58.76	49.37	34.40*
Class-balanced (Cui et al., 2019)	25.43‡	60.40‡	53.21	35.84‡
Adaptive (Cao et al., 2019)	26.65†	60.40†	52.15	33.31
Adaptive + DRW (Cao et al., 2019)	22.97†	57.96†	49.85	32.00†
Equalised (Tan et al., 2020)	26.02	57.26	54.02	38.37
Logit adjustment post-hoc ($\tau = 1$)	22.60	58.24	49.66	33.98
Logit adjustment loss ($\tau = 1$)	22.33	56.11	48.89	33.64
Logit adjustment plus adaptive loss ($\tau = 1$)	22.42	55.92	51.25	31.56

Table 2: Test set balanced error (averaged over 5 trials) on real-world datasets. Here, †, *, ‡ are numbers for "LDAM + SGD" and "LDAM + DRW" from Cao et al. (2019, Table 2, 3); "τ-normalised" from Kang et al. (2020, Table 3, 7); and "Class-Balanced" from Cui et al. (2019, Table 2, 3). Here, $\tau = \tau^*$ refers to using the best possible tuning parameter τ; see Figure 3 for results on various τ.

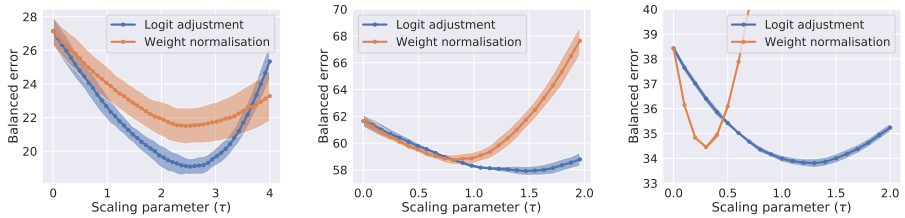
Figure 2 confirms that the logit adjusted margin loss attains a balanced error close to that of the Bayes-optimal, which is visually reflected by its learned separator closely matching that in (12). This is in line with our claim of the logit adjusted margin loss being consistent for the balanced error, unlike other approaches. Figure 2 also compares post-hoc weight normalisation and logit adjustment for varying scaling parameter τ (c.f. (3), (9)). Logit adjustment is seen to approach the performance of the Bayes predictor; *any* weight normalisation is however seen to hamper performance. This verifies the consistency of logit adjustment, and inconsistency of weight normalisation (§4.2).

6.2 Results on Real-World Datasets

We present results on the CIFAR-10, CIFAR-100, ImageNet and iNaturalist 2018 datasets. Following prior work, we create "long-tailed versions" of the CIFAR datasets by suitably downsampling examples per label following the Exp profile of Cui et al. (2019); Cao et al. (2019) with imbalance ratio $\rho = \max_y \mathbb{P}(y)/\min_y \mathbb{P}(y) = 100$. Similarly, we use the long-tailed version of ImageNet produced by Liu et al. (2019). We employ a ResNet-32 for CIFAR, and a ResNet-50 for ImageNet and iNaturalist. All models are trained using SGD with momentum; see Appendix C for more details. See also Appendix D.4 for results on CIFAR under the Step profile also considered in the literature.

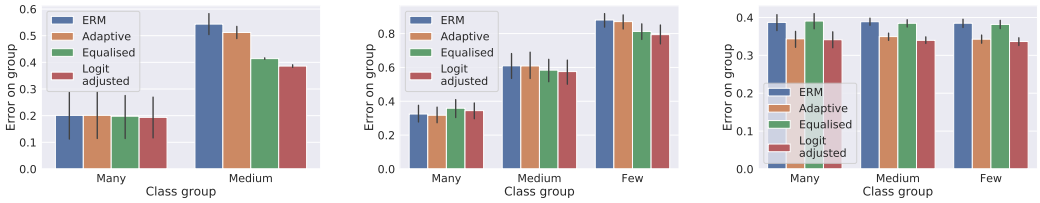
Baselines. We consider several representative baselines: (i) empirical risk minimisation (ERM) on the long-tailed data, (ii) post-hoc weight normalisation (Kang et al., 2020) per (3) (using $\nu_y = \|w_y\|_2$) applied to ERM, (iii) the class-balanced loss of Cui et al. (2019), (iv) the adaptive margin loss (Cao et al., 2019) per (5), including with the "deferred reweighting" (DRW) training scheme, and (v) the equalised loss (Tan et al., 2020) per (6), with $\delta_{y'} = F(\pi_{y'})$ for the threshold-based F of Tan et al. (2020). Where possible, we report numbers for the baselines from the respective papers.

We compare the above methods against our proposed post-hoc logit adjustment (9), and logit adjusted loss (10). For post-hoc logit adjustment, we fix the scalar $\tau = 1$ for our basic results; we analyse the effect of tuning this in Figure 3. We additionally evaluate a combination of our logit



(a) CIFAR-10. (b) CIFAR-100. (c) iNaturalist.

Figure 3: Comparison of balanced error for post-hoc correction techniques when varying scaling parameter τ (c.f. (3), (9)). Post-hoc logit adjustment consistently outperforms weight normalisation.



(a) CIFAR-10-LT (b) CIFAR-100-LT. (c) iNaturalist.

Figure 4: Comparison of per-group errors. We construct three groups of classes: "Many", comprising those with at least 100 training examples; "Medium", comprising those with at least 20 and at most 100 training examples; and "Few", comprising those with at most 20 training examples.

adjusted softmax cross-entropy with the adaptive margin of Cao et al. (2019); this uses (11) with $\Delta_{yy'} = \log(\pi_{y'}/\pi_y) + \pi_y^{-0.25}$. We do not perform *any* further tuning of our techniques.

Results and analysis. Table 2 demonstrates that our proposed logit adjustment techniques consistently outperform existing methods. Indeed, weight normalisation is generally improved significantly by post-hoc logit adjustment (e.g., 8% relative reduction on CIFAR-10). Similarly, loss correction techniques are generally outperformed by our logit adjusted softmax cross-entropy (e.g., 6% relative reduction on iNaturalist). We remark that Cao et al. (2019) observed that their loss benefits from a deferred reweighting scheme (DRW), wherein class-weighting is applied after a fixed number of epochs. Table 2 indicates this is consistently outperformed by suitable variants of logit adjustment.

While both logit adjustment techniques perform similarly, there is a slight advantage to the loss function version. Nonetheless, the strong performance of post-hoc logit adjustment corroborates the ability to decouple representation and classifier learning in long-tail settings (Zhang et al., 2019). Note that Table 2 only reports the results for scalar $\tau = 1$. In practice, tuning this parameter can significantly improve performance further. Figure 3 studies the effect of tuning τ for post-hoc weight normalisation (using $\nu_y = \|w_y\|_2$) and post-hoc logit adjustment. Even without *any* scaling, post-hoc logit adjustment generally offers superior performance to the best result from weight normalisation (cf. Table 2); with scaling, this is further improved. See Appendix D.4 for a plot on ImageNet-LT.

To better understand the gains, Figure 4 reports errors on a per-group basis, where following Kang et al. (2020) we construct three groups of classes — "Many", "Medium", and "Few" — comprising those with ≥ 100, between $(20, 100)$, and ≤ 20 training examples respectively. Logit adjustment shows consistent gains over all three groups. See Appendix D.3 for a finer-grained breakdown.

Discussion and extensions Table 2 shows the advantage of logit adjustment over recent proposals, under standard setups from the literature. Further improvements are possible by fusing complementary ideas, and we remark on a few such options. First, one may use a more complex base architecture; e.g., Kang et al. (2020) found gains by employing a ResNet-152, and training for 200 epochs. Table 3 (Appendix) confirms that logit adjustment similarly benefits from this choice, achieving a balanced error of $\mathbf{30.12}\%$ on iNaturalist, and $\mathbf{28.02}\%$ when combined with the adaptive margin. Second, the DRW training scheme (which applies to any loss) may result in further gains for our techniques. Third, incorporating developments in meta-learning (Wang et al., 2017; Jamal et al., 2020) is also a promising avenue. While further exploring such variants are of empirical interest, we hope to have illustrated the conceptual and empirical value of logit adjustment, and leave this for future work.

REFERENCES

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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

Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 3121–3124, Aug 2010.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *arXiv:1710.05381 [cs, stat]*, October 2017.

Jonathon Byrd and Zachary Chase Lipton. What is the effect of importance weighting in deep learning? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 872–881, 2019.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Philip K. Chan and Salvatore J. Stolfo. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. In *KDD-98 Workshop on Distributed Data Mining*, 1998.

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

Guillem Collell, Drazen Prelec, and Kaustubh R. Patil. Reviving threshold-moving: a simple plug-in bagging ensemble for binary and multiclass imbalanced data. *CoRR*, abs/1606.08698, 2016.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Tom Fawcett and Foster Provost. 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.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 243–268, 2007.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017.

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

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

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, 2020.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning, 2019.

Vladimir Koltchinskii, Dmitriy Panchenko, and Fernando Lozano. Some new bounds on the generalization error of combined classifiers. In T. K. Leen, T. G. Dietterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems 13*, pp. 245–251. MIT Press, 2001.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent binary classification with generalized performance metrics. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2744–2752, Cambridge, MA, USA, 2014. MIT Press.

Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1997.

Miroslav Kubat, Robert Holte, and Stan Matwin. Learning when negative examples abound. In Maarten van Someren and Gerhard Widmer (eds.), *Proceedings of the European Conference on Machine Learning (ECML)*, volume 1224 of *Lecture Notes in Computer Science*, pp. 146–153. Springer Berlin Heidelberg, 1997. ISBN 978-3-540-62858-3.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Yaoyong Li, Hugo Zaragoza, Ralf Herbrich, John Shawe-Taylor, and Jaz S. Kandola. The perceptron algorithm with uneven margins. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pp. 379–386, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. ISBN 1558608737.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73 – 82, 2004. ISSN 0167-7152.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. 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.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 185–201, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01216-8.

Marcus A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML 2003 Workshop on Learning from Imbalanced Datasets*, 2003.

Hamed Masnadi-Shirazi and Nuno Vasconcelos. Risk minimization, probability elicitation, and cost-sensitive SVMs. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 759–766, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Aditya Krishna Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 603–611, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 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.

Katharina Morik, Peter Brockhausen, and Thorsten Joachims. 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.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 4696–4705, 2019.

Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, 1987.

Patrick Pletscher, Cheng Soon Ong, and Joachim M. Buhmann. Entropy and margin maximization for structured output learning. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 83–98, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

Foster Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI-2000 Workshop on Imbalanced Data Sets*, 2000.

Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826, 2016.

J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, and J. Yan. Equalization loss for long-tailed object recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11659–11668, 2020.

Keiji Tatsumi and Tetsuzo Tanino. Support vector machines maximizing geometric margins for multi-class classification. *TOP*, 22(3):815–840, 2014.

Keiji Tatsumi, Masashi Akao, Ryo Kawachi, and Tetsuzo Tanino. Performance evaluation of multiobjective multiclass support vector machines maximizing geometric margins. *Numerical Algebra, Control & Optimization*, 1:151, 2011. ISSN 2155-3289.

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

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

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 7029–7039. Curran Associates, Inc., 2017.

Shan-Hung Wu, Keng-Pei Lin, Chung-Min Chen, and Ming-Syan Chen. Asymmetric support vector machines: Low false-positive learning under the user tolerance. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 749–757, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934.

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

Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning, 2020.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 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, 2019. Association for Computing Machinery. ISBN 9781450362436.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for deep face recognition with long-tail data. *CoRR*, abs/1803.09014, 2018.

Junjie Zhang, Lingqiao Liu, Peng Wang, and Chunhua Shen. To balance or not to balance: A simple-yet-effective approach for learning with long-tailed distributions, 2019.

X. Zhang, Z. Fang, Y. Wen, Z. Li, and Y. Qiao. Range loss for deep face recognition with long-tailed training data. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5419–5428, 2017.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18(1), 2006.