Supplementary material for "Long tail learning via logit adjustment"

A PROOFS OF RESULTS IN BODY

Proof of Theorem 1. Denote $\eta_y(x) = \mathbb{P}(y \mid x)$. Suppose we employ a margin $\Delta_{yy'} = \log \frac{\delta_{y'}}{\delta_y}$. Then, the loss is

$$\ell(y, f(x)) = -\log \frac{\delta_y \cdot e^{f_y(x)}}{\sum_{y' \in [L]} \delta_{y'} \cdot e^{f_{y'}(x)}} = -\log \frac{e^{f_y(x) + \log \delta_y}}{\sum_{y' \in [L]} e^{f_{y'}(x) + \log \delta_{y'}}}.$$

Consequently, under constant weights $\alpha_y = 1$, the Bayes-optimal score will satisfy $f^*_y(x) + \log \delta_y = \log \eta_y(x)$, or $f^*_y(x) = \log \frac{\eta_y(x)}{\delta_y}$.

Now suppose we use generic weights $\alpha \in \mathbb{R}^L_+$. The risk under this loss is

$$\begin{aligned}
\mathbb{E}_{x,y} \left[\ell_\alpha(y, f(x))\right] &= \sum_{y \in [L]} \pi_y \cdot \mathbb{E}_{x|y=y} \left[\ell_\alpha(y, f(x))\right] \\
&= \sum_{y \in [L]} \pi_y \cdot \mathbb{E}_{x|y=y} \left[\ell_\alpha(y, f(x))\right] \\
&= \sum_{y \in [L]} \pi_y \cdot \alpha_y \cdot \mathbb{E}_{x|y=y} \left[\ell(y, f(x))\right] \\
&\propto \sum_{y \in [L]} \bar{\pi}_y \cdot \mathbb{E}_{x|y=y} \left[\ell(y, f(x))\right],
\end{aligned}$$

where $\bar{\pi}_y \propto \pi_y \cdot \alpha_y$. Consequently, learning with the weighted loss is equivalent to learning with the original loss, on a distribution with modified base-rates $\bar{\pi}$. Under such a distribution, we have class-conditional distribution

$$\bar{\eta}_y(x) = \bar{\mathbb{P}}(y \mid x) = \frac{\mathbb{P}(x \mid y) \cdot \bar{\pi}_y}{\bar{\mathbb{P}}(x)} = \eta_y(x) \cdot \frac{\bar{\pi}_y}{\pi_y} \cdot \frac{\mathbb{P}(x)}{\bar{\mathbb{P}}(x)} \propto \eta_y(x) \cdot \alpha_y.$$

Consequently, suppose $\alpha_y = \frac{\delta_y}{\pi_y}$. Then, $f^*_y(x) = \log \frac{\bar{\eta}_y(x)}{\delta_y} = \log \frac{\eta_y(x)}{\pi_y} + C(x)$, where $C(x)$ does not depend on y. Consequently, $\operatorname{argmax}_{y \in [L]} f^*_y(x) = \operatorname{argmax}_{y \in [L]} \frac{\eta_y(x)}{\pi_y}$, which is the Bayes-optimal prediction for the balanced error.

In sum, a consistent family can be obtained by choosing any set of constants $\delta_y > 0$ and setting

$$\begin{aligned}
\alpha_y &= \frac{\delta_y}{\pi_y} \\
\Delta_{yy'} &= \log \frac{\delta_{y'}}{\delta_y}.
\end{aligned}$$

□

Proof of Theorem 2. We establish a more general result in Lemma 3 of the next section, which allows for a temperature parameter in the loss. This allows for interpolating between the standard softmax cross-entropy and margin based losses. □

B ON THE CONSISTENCY OF BINARY MARGIN-BASED LOSSES

It is instructive to study the pairwise margin loss (11) in the binary case. Endowing the loss with a temperature parameter $\gamma > 0$, we get[2]

$$\begin{aligned}\ell(+1, f) &= \frac{\omega_{+1}}{\gamma} \cdot \log(1 + e^{\gamma \cdot \delta_{+1}} \cdot e^{-\gamma \cdot f}) \\ \ell(-1, f) &= \frac{\omega_{-1}}{\gamma} \cdot \log(1 + e^{\gamma \cdot \delta_{-1}} \cdot e^{\gamma \cdot f})\end{aligned} \tag{13}$$

for constants $\omega_{\pm 1}, \gamma > 0$ and $\delta_{\pm 1} \in \mathbb{R}$. Here, we have used $\delta_{+1} = \Delta_{+1,-1}$ and $\delta_{-1} = \Delta_{-1,+1}$ for simplicity. The choice $\omega_{\pm 1} = 1, \delta_{\pm 1} = 0$ recovers the temperature scaled binary logistic loss. Evidently, as $\gamma \to +\infty$, these converge to weighted hinge losses with variable margins, i.e.,

$$\begin{aligned}\ell(+1, f) &= \omega_{+1} \cdot [\delta_{+1} - f]_+ \\ \ell(-1, f) &= \omega_{-1} \cdot [\delta_{-1} + f]_+.\end{aligned}$$

We study two properties of this family losses. First, under what conditions are the losses Fisher consistent for the balanced error? We shall show that in fact there is a simple condition characterising this. Second, do the losses preserve properness of the original binary logistic loss? We shall show that this is always the case, but that the losses involve fundamentally different approximations.

B.1 CONSISTENCY OF THE BINARY PAIRWISE MARGIN LOSS

Given a loss ℓ, its *Bayes optimal* solution is $f^* \in \operatorname{argmin}_{f\colon \mathcal{X} \to \mathbb{R}} \mathbb{E}\left[\ell(y, f(x))\right]$. For consistency with respect to the balanced error in the binary case, we require this optimal solution f^* to satisfy $f^*(x) > 0 \iff \eta(x) > \pi$, where $\eta(x) \doteq \mathbb{P}(y = 1 \mid x)$ and $\pi \doteq \mathbb{P}(y = 1)$ (Menon et al., 2013). This is equivalent to a simple condition on the weights ω and margins δ of the pairwise margin loss.

Lemma 3. *The losses in (13) are consistent for the balanced error iff*

$$\frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{\sigma(\gamma \cdot \delta_{+1})}{\sigma(\gamma \cdot \delta_{-1})} = \frac{1 - \pi}{\pi},$$

where $\sigma(z) = (1 + \exp(z))^{-1}$.

Proof of Lemma 3. Denote $\eta(x) \doteq \mathbb{P}(y = +1 \mid x)$, and $\pi \doteq \mathbb{P}(y = +1)$. From Lemma 4 below, the pairwise margin loss is proper composite with invertible link function $\Psi\colon [0, 1] \to \mathbb{R} \cup \{\pm\infty\}$. Consequently, since by definition the Bayes-optimal score for a proper composite loss is $f^*(x) = \Psi(\eta(x))$ (Reid & Williamson, 2010), to have consistency for the balanced error, from (14), (15), we require

$$\begin{aligned}\Psi^{-1}(0) = \pi &\iff \frac{1}{1 - \frac{\ell'(+1,0)}{\ell'(-1,0)}} = \pi \\ &\iff 1 - \frac{\ell'(+1,0)}{\ell'(-1,0)} = \frac{1}{\pi} \\ &\iff -\frac{\ell'(+1,0)}{\ell'(-1,0)} = \frac{1 - \pi}{\pi} \\ &\iff \frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{\sigma(\gamma \cdot \delta_{+1})}{\sigma(\gamma \cdot \delta_{-1})} = \frac{1 - \pi}{\pi}.\end{aligned}$$

□

From the above, some admissible parameter choices include:

- $\omega_{+1} = \frac{1}{\pi}, \omega_{-1} = \frac{1}{1-\pi}, \delta_{\pm 1} = 1$; i.e., the standard weighted loss with a constant margin

[2] Compared to the multiclass case, we assume here a scalar score $f \in \mathbb{R}$. This is equivalent to constraining that $\sum_{y \in [L]} f_y = 0$ for the multiclass case.

- $\omega_{\pm 1} = 1$, $\delta_{+1} = \frac{1}{\gamma} \cdot \log \frac{1-\pi}{\pi}$, $\delta_{-1} = \frac{1}{\gamma} \cdot \log \frac{\pi}{1-\pi}$; i.e., the unweighted loss with a margin biased towards the rare class, as per our logit adjustment procedure

The second example above is unusual in that it requires scaling the margin with the temperature; consequently, the margin disappears as $\gamma \to +\infty$. Other combinations are of course possible, but note that one cannot arbitrarily choose parameters and hope for consistency in general. Indeed, some *inadmissible* choices are naïve applications of the margin modification or weighting, e.g.,

- $\omega_{+1} = \frac{1}{\pi}$, $\omega_{-1} = \frac{1}{1-\pi}$, $\delta_{+1} = \frac{1}{\gamma} \cdot \log \frac{1-\pi}{\pi}$, $\delta_{-1} = \frac{1}{\gamma} \cdot \log \frac{\pi}{1-\pi}$; i.e., combining *both* weighting and margin modification
- $\omega_{\pm 1} = 1$, $\delta_{+1} = \frac{1}{\gamma} \cdot (1-\pi)$, $\delta_{-1} = \frac{1}{\gamma} \cdot \pi$; i.e., specific margin modification

Note further that the choices of Cao et al. (2019); Tan et al. (2020) do not meet the requirements of Lemma 3.

We make two final remarks. First, the above only considers consistency of the result of loss minimisation. For *any* choice of weights and margins, we may apply suitable post-hoc correction to the predictions to account for any bias in the optimal scores. Second, as $\gamma \to +\infty$, any *constant* margins $\delta_{\pm 1} > 0$ will have no effect on the consistency condition, since $\sigma(\gamma \cdot \delta_{\pm 1}) \to 1$. The condition will be wholly determined by the weights $\omega_{\pm 1}$. For example, we may choose $\omega_{+1} = \frac{1}{\pi}$, $\omega_{-1} = \frac{1}{1-\pi}$, $\delta_{+1} = 1$, and $\delta_{-1} = \frac{\pi}{1-\pi}$; the resulting loss will not be consistent for finite γ, but will become so in the limit $\gamma \to +\infty$.

B.2 Properness of the pairwise margin loss

In the above, we appealed to the pairwise margin loss being proper composite, in the sense of Reid & Williamson (2010). Intuitively, this specifies that the loss has Bayes-optimal score of the form $f^*(x) = \Psi(\eta(x))$, where Ψ is some invertible function, and $\eta(x) = \mathbb{P}(y = 1 \mid x)$. We have the following general result about properness of *any* member of the pairwise margin family.

Lemma 4. *The losses in (13) are proper composite, with link function*

$$\Psi(p) = \frac{1}{\gamma} \cdot \log \left[\left(\frac{a \cdot b}{q} - c \right) \pm \sqrt{\left(\frac{a \cdot b}{q} - c \right)^2 + 4 \cdot \frac{a}{q}} \right] - \log 2,$$

where $a = \frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{e^{\gamma \cdot \delta_{+1}}}{e^{\gamma \cdot \delta_{-1}}}$, $b = e^{\gamma \cdot \delta_{-1}}$, $c = e^{\gamma \cdot \delta_{+1}}$, *and* $q = \frac{1-p}{p}$.

Proof of Lemma 4. The above family of losses is proper composite iff the function

$$\Psi^{-1}(f) = \frac{1}{1 - \frac{\ell'(+1,f)}{\ell'(-1,f)}} \tag{14}$$

is invertible (Reid & Williamson, 2010, Corollary 12). We have

$$\begin{aligned} \ell'(+1, f) &= -\omega_{+1} \cdot \frac{e^{\gamma \cdot \delta_{+1}} \cdot e^{-\gamma \cdot f}}{1 + e^{\gamma \cdot \delta_{+1}} \cdot e^{-\gamma \cdot f}} \\ \ell'(-1, f) &= +\omega_{-1} \cdot \frac{e^{\gamma \cdot \delta_{-1}} \cdot e^{\gamma \cdot f}}{1 + e^{\gamma \cdot \delta_{-1}} \cdot e^{\gamma \cdot f}}. \end{aligned} \tag{15}$$

The invertibility of Ψ^{-1} is immediate. To compute the link function Ψ, note that

$$\begin{aligned} p = \frac{1}{1 - \frac{\ell'(+1,f)}{\ell'(-1,f)}} &\iff \frac{1}{p} = 1 - \frac{\ell'(+1,f)}{\ell'(-1,f)} \\ &\iff -\frac{\ell'(+1,f)}{\ell'(-1,f)} = \frac{1-p}{p} \\ &\iff \frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{e^{\gamma \cdot \delta_{+1}} \cdot e^{-\gamma \cdot f}}{1 + e^{\gamma \cdot \delta_{+1}} \cdot e^{-\gamma \cdot f}} \cdot \frac{1 + e^{\gamma \cdot \delta_{-1}} \cdot e^{\gamma \cdot f}}{e^{\gamma \cdot \delta_{-1}} \cdot e^{\gamma \cdot f}} = \frac{1-p}{p} \end{aligned}$$

$$\iff \frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{e^{\gamma\cdot\delta_{+1}}}{e^{\gamma\cdot\delta_{-1}}} \cdot \frac{1}{e^{\gamma\cdot f} + e^{\gamma\cdot\delta_{+1}}} \cdot \frac{1 + e^{\gamma\cdot\delta_{-1}} \cdot e^{\gamma\cdot f}}{e^{\gamma\cdot f}} = \frac{1-p}{p}$$
$$\iff a \cdot \frac{1 + b\cdot g}{g^2 + c\cdot g} = q,$$

where $a = \frac{\omega_{+1}}{\omega_{-1}} \cdot \frac{e^{\gamma\cdot\delta_{+1}}}{e^{\gamma\cdot\delta_{-1}}}$, $b = e^{\gamma\cdot\delta_{-1}}$, $c = e^{\gamma\cdot\delta_{+1}}$, $g = e^{\gamma\cdot f}$, and $q = \frac{1-p}{p}$. Thus,

$$\begin{aligned} a \cdot \frac{1 + b\cdot g}{g^2 + c\cdot g} = q &\iff \frac{g^2 + c\cdot g}{1 + b\cdot g} = \frac{a}{q} \\ &\iff g^2 + \left(c - \frac{a\cdot b}{q}\right)\cdot g - \frac{a}{q} = 0 \\ &\iff g = \frac{\left(\frac{a\cdot b}{q} - c\right) \pm \sqrt{\left(\frac{a\cdot b}{q} - c\right)^2 + 4\cdot\frac{a}{q}}}{2}. \end{aligned}$$

□

As a sanity check, suppose $a = b = c = \gamma = 1$. This corresponds to the standard logistic loss. Then,

$$\Psi(p) = \log \frac{\left(\frac{1}{q} - 1\right) \pm \sqrt{\left(\frac{1}{q} - 1\right)^2 + 4\cdot\frac{1}{q}}}{2} = \log \frac{p}{1-p},$$

which is the standard logit function.

Figure 5 and 6 compares the link functions for a few different settings:

- the balanced loss, where $\omega_{+1} = \frac{1}{\pi}$, $\omega_{-1} = \frac{1}{1-\pi}$, and $\delta_{\pm 1} = 1$
- an unequal margin loss, where $\omega_{\pm 1} = 1$, $\delta_{+1} = \frac{1}{\gamma}\cdot\log\frac{1-\pi}{\pi}$, and $\delta_{-1} = \frac{1}{\gamma}\cdot\log\frac{\pi}{1-\pi}$
- a balanced + margin loss, where $\omega_{+1} = \frac{1}{\pi}$, $\omega_{-1} = \frac{1}{1-\pi}$, $\delta_{+1} = 1$, and $\delta_{-1} = \frac{\pi}{1-\pi}$.

The property $\Psi^{-1}(0) = \pi$ for $\pi = \mathbb{P}(y = 1)$ holds for the first two choices with any $\gamma > 0$, and the third choice as $\gamma \to +\infty$. This indicates the Fisher consistency of these losses for the balanced error. However, the precise way this is achieved is strikingly different in each case. In particular, each loss implicitly involves a fundamentally different link function.

To better understand the effect of parameter choices, Figure 7 illustrates the conditional Bayes risk curves, i.e.,

$$\underline{\mathrm{L}}(p) = p\cdot\ell(+1, \Psi(p)) + (1-p)\cdot\ell(+1, \Psi(p)).$$

We remark here that for the balanced error, this function takes the form $\underline{\mathrm{L}}(p) = p\cdot[\![p < \pi]\!] + (1-p)\cdot[\![p > \pi]\!]$, i.e., it is a "tent shaped" concave function with a maximum at $p = \pi$.

For ease of comparison, we normalise this curves to have a maximum of 1. Figure 7 shows that simply applying unequal margins does *not* affect the underlying conditional Bayes risk compared to the standard log-loss; thus, the change here is purely in terms of the link function. By contrast, either balancing the loss or applying a combination of weighting and margin modification results in a closer approximation to the conditional Bayes risk curve for the cost-sensitive loss with cost π.

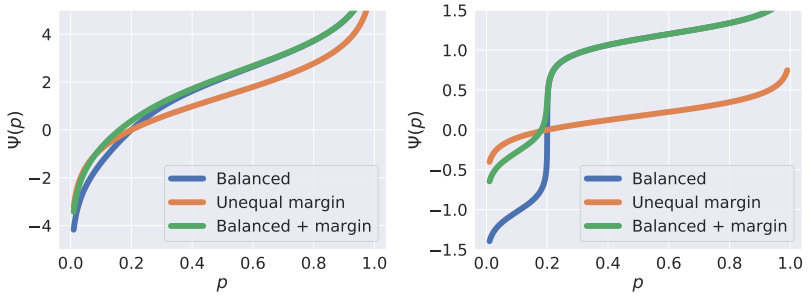


Figure 5: Comparison of link functions for various losses assuming $\pi = 0.2$, with $\gamma = 1$ (left) and $\gamma = 8$ (right). The balanced loss uses $\omega_y = \frac{1}{\pi_y}$. The unequal margin loss uses $\delta_y = \frac{1}{\gamma} \cdot \log \frac{1-\pi}{\pi}$. The balanced + margin loss uses $\delta_{-1} = \frac{\pi}{1-\pi}$, $\delta_{+1} = 1$, $\omega_{+1} = \frac{1}{\pi}$.

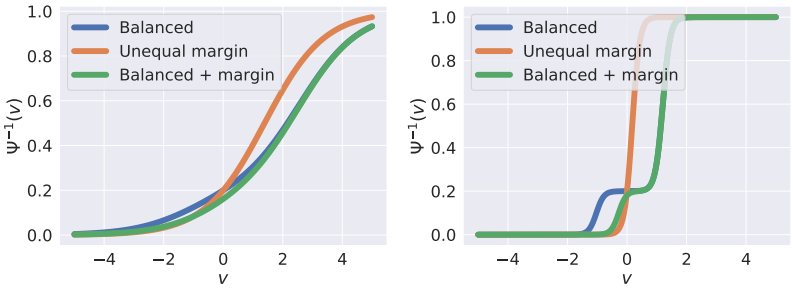


Figure 6: Comparison of link functions for various losses assuming $\pi = 0.2$, with $\gamma = 1$ (left) and $\gamma = 8$ (right). The balanced loss uses $\omega_y = \frac{1}{\pi_y}$. The unequal margin loss uses $\delta_y = \frac{1}{\gamma} \cdot \log \frac{1-\pi_y}{\pi_y}$. The balanced + margin loss uses $\delta_{-1} = \frac{\pi}{1-\pi}$, $\delta_{+1} = 1$, $\omega_{+1} = \frac{1}{\pi}$.

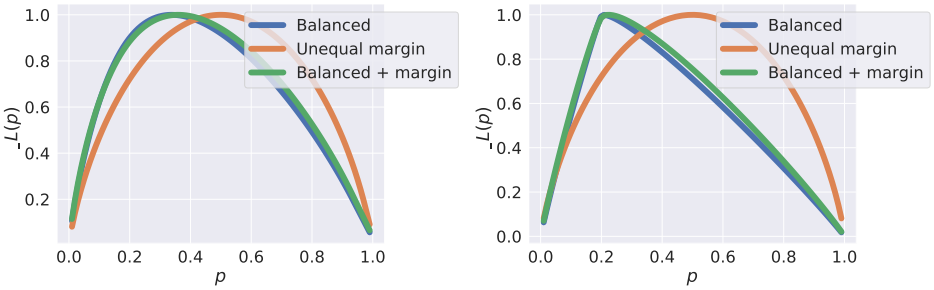


Figure 7: Comparison of conditional Bayes risk functions for various losses assuming $\pi = 0.2$, with $\gamma = 1$ (left) and $\gamma = 8$ (right). The balanced loss uses $\omega_y = \frac{1}{\pi_y}$. The unequal margin loss uses $\delta_y = \frac{1}{\gamma} \cdot \log \frac{1-\pi_y}{\pi_y}$. The first balanced + margin loss uses $\delta_{-1} = \pi$, $\delta_{+1} = 1$, $\omega_{+1} = \frac{1}{\pi}$. The second balanced + margin loss uses $\delta_{-1} = \frac{\pi}{1-\pi}$, $\delta_{+1} = 1$, $\omega_{+1} = \frac{1}{\pi}$.

C EXPERIMENTAL SETUP

Intending a fair comparison, we use the same setup for all the methods for each dataset. All networks are trained with SGD with a momentum value of 0.9. Unless otherwise specified, linear learning rate warm-up is used in the first 5 epochs to reach the base learning rate, and a weight decay of 10^{-4} is used. Other dataset specific details are given below.

CIFAR-10 and CIFAR-100: We use a CIFAR ResNet-32 model trained for 200 epochs. The base learning rate is set to 0.1, which is decayed by 0.1 at the 160th epoch and again at the 180th epoch. Mini-batches of 128 images are used.

We also use the standard CIFAR data augmentation procedure used in previous works such as Cao et al. (2019); He et al. (2016), where 4 pixels are padded on each size and a random 32×32 crop is taken. Images are horizontally flipped with a probability of 0.5.

ImageNet: We use a ResNet-50 model trained for 90 epochs. The base learning rate is 0.4, with cosine learning rate decay. We use a batch size of 512 and the standard data augmentation comprising of random cropping and flipping as described in Goyal et al. (2017). Following Kang et al. (2020), we use a weight decay of 5×10^{-4} on this dataset.

iNaturalist: We again use a ResNet-50 and train it for 90 epochs with a base learning rate of 0.4 and cosine learning rate decay. The data augmentation procedure is the same as the one used in ImageNet experiment above. We use a batch size of 512.

D ADDITIONAL EXPERIMENTS

We present here additional experiments:

(i) we present a detailed table of results with a more complex base architecture and number of training epochs for ImageNet-LT and iNaturalist;

(ii) we present results for CIFAR-10 and CIFAR-100 on the STEP profile (Cao et al., 2019) with $\rho = 100$.

(iii) we present results on synthetic data with varying imbalance ratios.

D.1 RESULTS WITH MORE COMPLEX BASE ARCHITECTURE

Table 3 presents results when using a ResNet-152, trained for either 90 or 200 epochs, on the larger ImageNet-LT and iNaturalist datasets. Consistent with the findings in Kang et al. (2020), training with a more complex architecture for longer generally yields significant gains. Logit adjustment, potentially when combined with the adaptive margin, is generally superior to baselines with the sole exception of results for ResNet-152 with 200 epochs on iNaturalist.

Table 3: Test set balanced error (averaged over 5 trials) on real-world datasets with more complex base architectures. Employing a ResNet-152 systematically improves all methods' performance, with logit adjustment remaining superior to existing approaches. The final row reports the results of combining logit adjustment with the adaptive margin loss of Cao et al. (2019), which yields further gains on iNaturalist.

	ImageNet-LT		**iNaturalist**		
Method	**ResNet-50**	**ResNet-152**	**ResNet-50**	**ResNet-152**	**ResNet-152**
			90 epochs	**90 epochs**	**200 epochs**
ERM	53.11	53.30	38.66	35.88	34.38
Weight normalisation ($\tau = 1$) (Kang et al., 2020)	52.00	51.49	48.05	45.17	45.33
Weight normalisation ($\tau = \tau^*$) (Kang et al., 2020)	49.37	48.97	34.10	31.85	30.34
Learnable weight scaling (LWS) (Kang et al., 2020)	52.30	49.50	34.10	30.90	27.90
Classifier retraining (cRT) (Kang et al., 2020)	52.70	49.90	34.80	31.50	28.80
Adaptive (Cao et al., 2019)	52.15	53.34	35.42	31.18	29.46
Equalised (Tan et al., 2020)	54.02	51.38	38.37	35.86	34.53
Logit adjustment post-hoc ($\tau = 1$)	49.66	49.25	33.98	31.46	30.15
Logit adjustment post-hoc ($\tau = \tau^*$)	49.56	49.15	33.80	31.08	29.74
Logit adjustment loss ($\tau = 1$)	48.89	47.86	33.64	31.15	30.12
Logit adjustment plus adaptive loss ($\tau = 1$)	51.25	50.46	31.56	29.22	28.02

D.2 PER-CLASS ERROR RATES

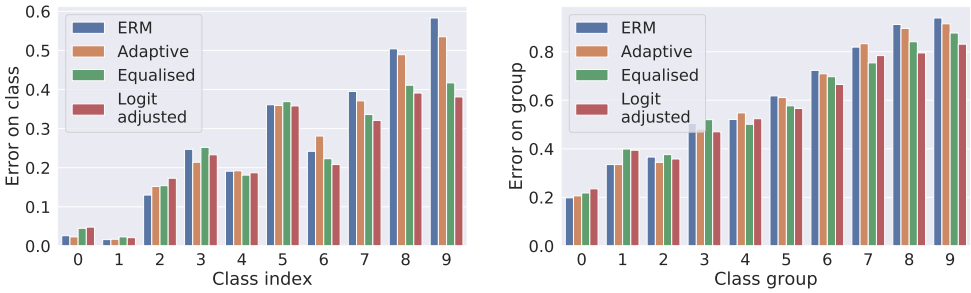
Figure 8 breaks down the per-class accuracies on CIFAR-10, CIFAR-100, and iNaturalist. On the latter two datasets, for ease of visualisation, we aggregate the classes into ten groups based on their frequency-sorted order (so that, e.g., group 0 comprises the top $\frac{L}{10}$ most frequent classes). As expected, dominant classes generally see a lower error rate with all methods. However, the logit adjusted loss is seen to systematically improve performance over ERM, particularly on rare classes.

D.3 RESULTS ON CIFAR-LT WITH STEP-100 PROFILE

Table 4 summarises results on the STEP-100 profile. Here, with $\tau = 1$, weight normalisation slightly outperforms logit adjustment. However, with $\tau > 1$, logit adjustment is again found to be superior (54.80); see Figure 9.

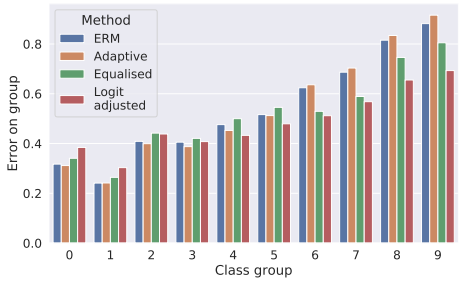
D.4 RESULTS ON SYNTHETIC DATA WITH VARYING IMBALANCE RATIO

Figure 10 shows results on the synthetic data of §6.1 for varying choice of $\mathbb{P}(y = +1)$. As expected, we see that as $\mathbb{P}(y = +1)$ increases, all methods become equitable in terms of performance. We

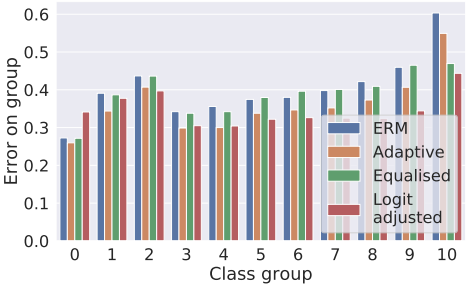


(a) CIFAR-10.

(b) CIFAR-100.



(c) CIFAR-100.



(d) iNaturalist.

Figure 8: Per-class error rates of loss modification techniques. For (b) and (c), we aggregate the classes into 10 groups. ERM displays a strong bias towards dominant classes (lower indices). Our proposed logit adjusted softmax loss achieves significant gains on rare classes (higher indices).

Table 4: Test set balanced error (averaged over 5 trials) on CIFAR-10-LT and CIFAR-100-LT under the STEP-100 profile; lower is better. On CIFAR-100-LT, weight normalisation edges out logit adjustment. See Figure 9 for a demonstrated that tuned versions of the same outperfom weight normalisation.

Method	**CIFAR-10-LT**	**CIFAR-100-LT**
ERM	36.54	60.23
Weight normalisation ($\tau = 1$)	30.86	55.19
Adaptive	34.61	58.86
Equalised	31.42	57.82
Logit adjustment post-hoc ($\tau = 1$)	28.66	55.82
Logit adjustment (loss)	27.57	55.52

generally find a consistent trend in the relative performance of the various methods, which matches the results in the body.

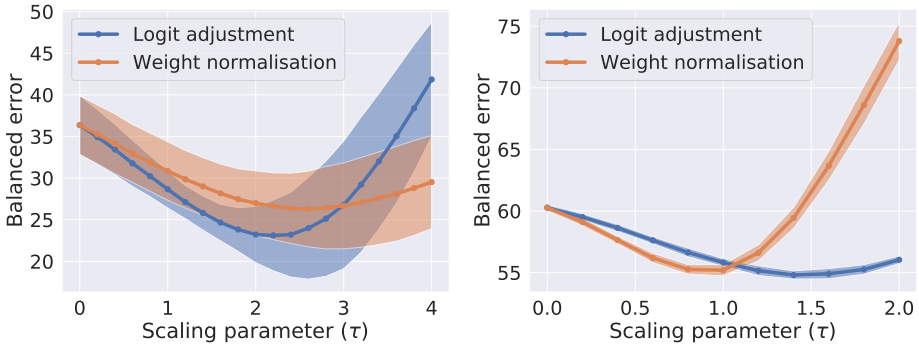
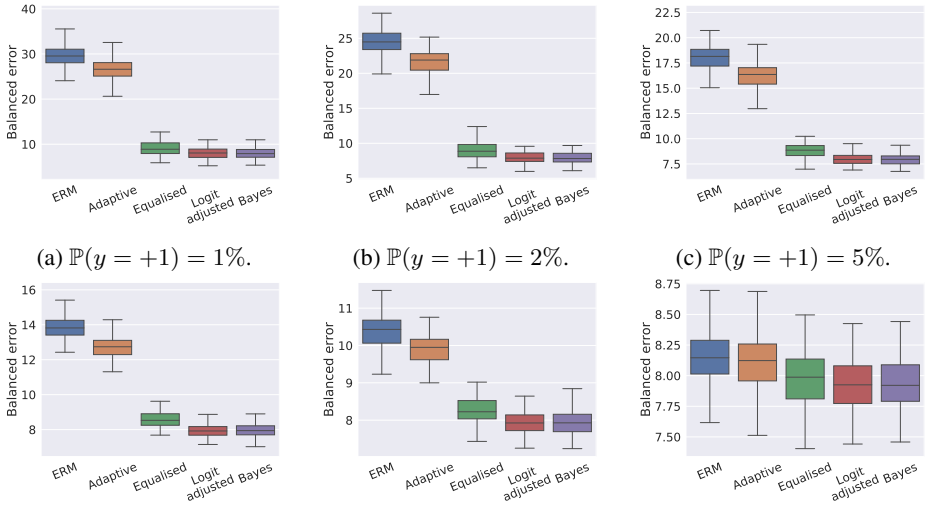


Figure 9: Post-hoc adjustment on STEP-100 profile, CIFAR-10 and CIFAR-100. Logit adjustment outperforms weight normalisation with suitable tuning.



(a) $\mathbb{P}(y = +1) = 1\%$. (b) $\mathbb{P}(y = +1) = 2\%$. (c) $\mathbb{P}(y = +1) = 5\%$.

(d) $\mathbb{P}(y = +1) = 10\%$. (e) $\mathbb{P}(y = +1) = 20\%$. (f) $\mathbb{P}(y = +1) = 40\%$.

Figure 10: Results on synthetic data with varying imbalance ratio.

E DOES WEIGHT NORMALISATION INCREASE MARGINS?

Suppose that one uses SGD with a momentum, and finds solutions where $\|w_y\|_2$ tracks the class priors. One intuition behind normalisation of weights is that, drawing inspiration from the binary case, this ought to increase the classification margins for tail classes. Unfortunately, as discussed below, this intuition is *not* necessarily borne out.

Consider a scorer $f_y(x) = w_y^\top \Phi(x)$, where $w_y \in \mathbb{R}^d$ and $\Phi \colon \mathcal{X} \to \mathbb{R}^d$. The *functional* margin for an example (x, y) is (Koltchinskii et al., 2001)

$$\gamma_f(x, y) \doteq w_y^\top \Phi(x) - \max_{y' \neq y} w_{y'}^\top \Phi(x). \tag{16}$$

This generalises the classical binary margin, wherein by convention $\mathcal{Y} = \{\pm 1\}$, $w_{-1} = -w_1$, and

$$\gamma_{\mathrm{f}}(x, y) \doteq y \cdot w_1^\top \Phi(x) = \frac{1}{2} \cdot \left(w_y^\top \Phi(x) - w_{-y}^\top \Phi(x) \right), \tag{17}$$

which agrees with (16) upto scaling. One may also define the *geometric* margin in the binary case to be the distance of (x, y) from its classifier:

$$\gamma_{g,\mathrm{b}}(x) \doteq \frac{|w_1 \cdot \Phi(x)|}{\|w_1\|_2}. \tag{18}$$

Clearly, $\gamma_{g,\mathrm{b}}(x) = \frac{|\gamma_{\mathrm{f}}(x,y)|}{\|w_1\|_2}$, and so for fixed functional margin, one may increase the geometric margin by minimising $\|w_1\|_2$. However, the same is *not* necessarily true in the multiclass setting, since here the functional and geometric margins do not generally align (Tatsumi et al., 2011; Tatsumi & Tanino, 2014). In particular, controlling each $\|w_y\|_2$ does *not* necessarily control the geometric margin.

F BAYES-OPTIMAL CLASSIFIER UNDER GAUSSIAN CLASS-CONDITIONALS

Proof of (12). Suppose

$$\mathbb{P}(x \mid y) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{\|x-\mu_y\|^2}{2\sigma^2}\right)$$

for suitable μ_y and σ. Then,

$$\begin{aligned}
\mathbb{P}(x \mid y=+1) > \mathbb{P}(x \mid y=-1) &\iff \exp\left(-\frac{\|x-\mu_{+1}\|^2}{2\sigma^2}\right) > \exp\left(-\frac{\|x-\mu_{-1}\|^2}{2\sigma^2}\right) \\
&\iff \frac{\|x-\mu_{+1}\|^2}{2\sigma^2} < \frac{\|x-\mu_{-1}\|^2}{2\sigma^2} \\
&\iff \|x-\mu_{+1}\|^2 < \|x-\mu_{-1}\|^2 \\
&\iff 2 \cdot (\mu_{+1}-\mu_{-1})^\top x > \|\mu_{+1}\|^2 - \|\mu_{-1}\|^2.
\end{aligned}$$

Now use the fact that in our setting, $\|\mu_{+1}\|^2 = \|\mu_{-1}\|^2$. □

We now explicate that the class-probability function for the synthetic dataset in § 6.1 is exactly in the family assumed by the logistic regression. This implies that logistic regression is well-specified for this problem, and thus can perfectly model $\mathbb{P}(y=+1 \mid x)$ in the infinite sample limit. Note that

$$\begin{aligned}
\mathbb{P}(y=+1 \mid x) &= \frac{\mathbb{P}(x \mid y=+1) \cdot \mathbb{P}(y=+1)}{\mathbb{P}(x)} \\
&= \frac{\mathbb{P}(x \mid y=+1) \cdot \mathbb{P}(y=+1)}{\sum_{y'} \mathbb{P}(x \mid y') \cdot \mathbb{P}(y')} \\
&= \frac{1}{1 + \frac{\mathbb{P}(x|y=-1)\cdot\mathbb{P}(y=-1)}{\mathbb{P}(x|y=+1)\cdot\mathbb{P}(y=+1)}}.
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{\mathbb{P}(x \mid y=-1)}{\mathbb{P}(x \mid y=+1)} &= \exp\left(\frac{\|x-\mu_{+1}\|^2 - \|x-\mu_{-1}\|^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{\|\mu_{+1}\|^2 - \|\mu_{-1}\|^2 - 2\cdot(\mu_{+1}-\mu_{-1})^\top x}{2\sigma^2}\right) \\
&= \exp\left(\frac{-(\mu_{+1}-\mu_{-1})^\top x}{\sigma^2}\right).
\end{aligned}$$

Thus,

$$\mathbb{P}(y=+1 \mid x) = \frac{1}{1+\exp(-w_*^\top x + b_*)},$$

where $w_* = \frac{1}{\sigma^2} \cdot (\mu_{+1} - \mu_{-1})$, and $b_* = \log \frac{\mathbb{P}(y=-1)}{\mathbb{P}(y=+1)}$. This implies that a sigmoid model for $\mathbb{P}(y=+1 \mid x)$, as employed by logistic regression, is well-specified for the problem. Further, the bias term b_* is seen to take the form of the log-odds of the class-priors per (8), as expected.