Long-tail learning via logit adjustment

Aditya Krishna Menon

Sadeep Jayasumana

Ankit Singh Rawat

Himanshu Jain

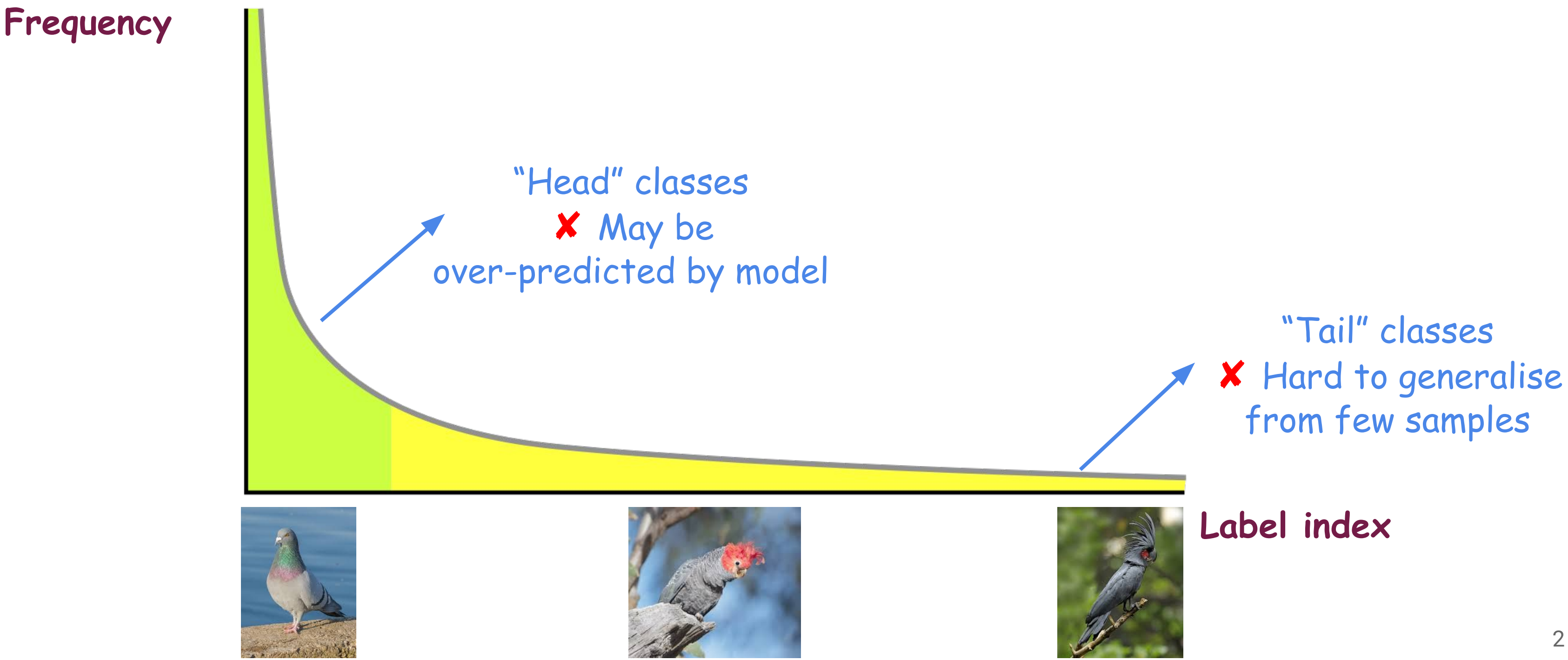
Andreas Veit

Sanjiv Kumar

Google Research

Long-tail learning

Classification where the label distribution is skewed

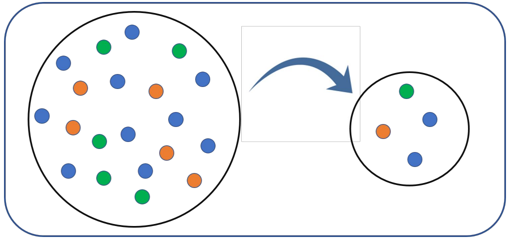


Summary of our work

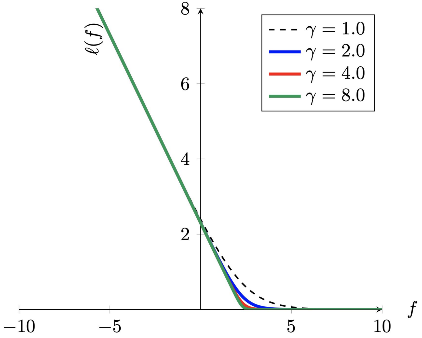
- ✓ A statistical perspective of long-tail learning
- ✓ Unifies and generalises existing approaches
- ✓ Yields new post-hoc and loss modification approaches

Existing approaches

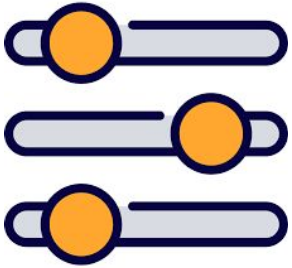
General strategies



Data sampling



Loss modification



Post-hoc adjustment

Weight normalisation

For instance x, a neural model computes logits

$$f_y(x) = w_y^\top \Phi(x)$$

✘ Norms may be smaller for rare classes!

Weight normalisation [Kang et al., ‘20]:

(1) Learn w, Φ via standard ERM
(2) Post-hoc normalise the weight norms

Decouples representation and classifier learning

Loss modification

Enforce varying margin depending on label frequency:

Softmax:

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

Adaptive margin:

[Cao et al., 2019]

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

e.g., 1/P(y)

Encourage higher margin between rare +ve and all -ves

Equalised loss:

[Tan et al., 2020]

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

e.g., P(y')

Prevent rare -ves from having gradient overwhelmed

Weight normalisation: limitation

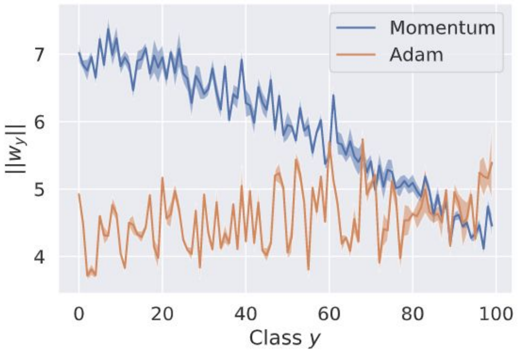
$$f_y(x) = w_y^T \Phi(x)$$

✘ Norms may be smaller for rare classes!

CIFAR-10

Classes ordered by frequency

CIFAR-100



Classes ordered by frequency

Weight norms don't correlate with *P*(*y*) when using Adam

A statistical framework

Balanced error

Typically, measure misclassification error:

$$\mathrm{ERR}(h) = \mathbb{P}(y \neq h(x))$$
$$= \sum_{i \in [L]} \mathbb{P}(y = i) \cdot \mathbb{P}(y \neq h(x) \mid y = i)$$

✘ Can do very well by predicting majority class!

Under class imbalance, can measure balanced error:

$$\mathrm{BER}(h) = \sum_{i \in [L]} \frac{1}{L} \cdot \mathbb{P}(y \neq h(x) \mid y = i)$$

✓ Treat all classes equally

Statistical view of long-tail learning

Bayes-optimal prediction:

$$\text{argmax}_{y \in [L]} \mathbb{P}^{\text{bal}}(y \mid x)$$

Balanced label distribution

Equivalently, if **P**(y | x) $\propto$ exp($s^*_y(x)$),

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

"Optimal" logits

Increase score for rare classes

Strategies for long-tail learning

Bayes-optimal solution suggests two strategies:

(1) Estimate $\mathbf{P}(y \mid x)$, and adjust logits post-hoc

(2) Directly estimate $\mathbf{P}_{\text{bal}}(y \mid x)$ by inherently adjusting logits

Post-hoc logit adjustment

Standard prediction:

$$\text{argmax}_{y \in [L]} \exp(w_y^{\mathrm{T}} \Phi(x)) = \text{argmax}_{y \in [L]} f_y(x)$$

Logit adjusted prediction:

$$\text{argmax}_{y \in [L]} \exp(w_y^{\mathrm{T}} \Phi(x)) / \pi_y^{\tau} = \text{argmax}_{y \in [L]} f_y(x) - \tau \cdot \log \pi_y,$$

Scaling parameter (τ)

Estimate of $P(y)$ (π_y)

When $\tau > 1$, equivalent to temperature-scaling the probabilities

Comparison to weight normalisation

Logit adjustment performs additive correction:

$$\text{argmax}_{y \in [L]} \exp(w_y^{\mathrm{T}} \Phi(x))/\pi_y^{\tau} = \text{argmax}_{y \in [L]} f_y(x) - \tau \cdot \log \pi_y,$$

Weight normalisation performs multiplicative correction:

$$\text{argmax}_{y \in [L]} (w_y^{\mathrm{T}} \Phi(x))/\pi_y^{\tau} = \text{argmax}_{y \in [L]} f_y(x)/\pi_y^{\tau}$$

Logit adjusted loss

Logit adjusted softmax cross-entropy:

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

Add fixed offset to logits

> 1 when $P(y') > P(y)$, i.e., -ve is more common than +ve

Now predict argmax $f_y(x)$ as normal

A margin view

Consider the pairwise margin loss

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

Existing losses → Δ_{yy} depends on *y* or *y'*, but not both

Logit adjustment → $\Delta_{yy'}$ = log **P**(*y'*)/**P**(*y*) = log **P**(*y'*) - log **P**(*y*)

Enforces a relative margin between labels

Google

Experiments

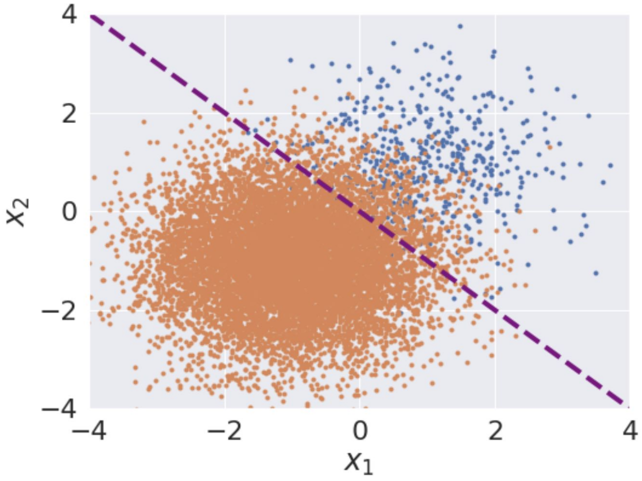
Google

Experiments: synthetic data

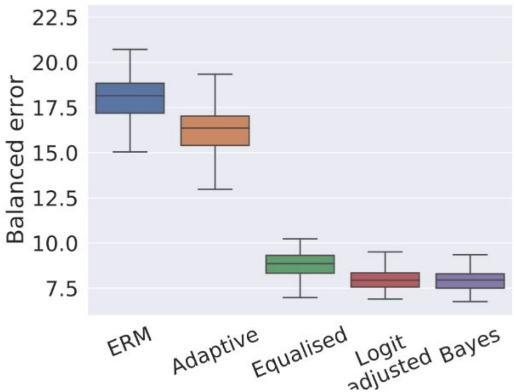
Consider data drawn from a mixture of isotropic Gaussians, with **P**(y = 1) = 5%

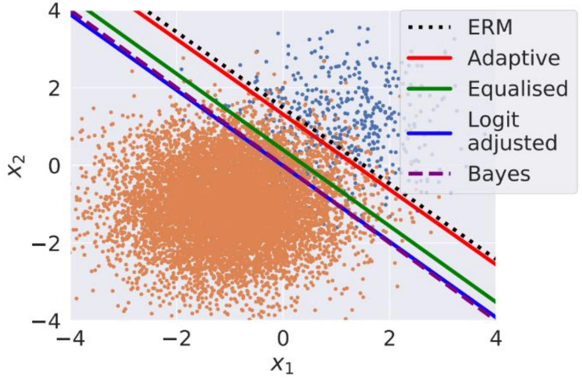
Bayes-optimal for balanced error: separator passing through origin

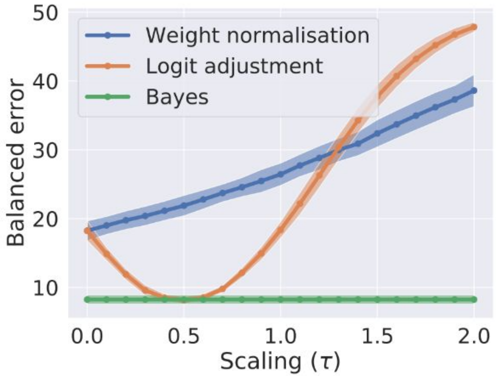
ERM will favour fewer mistakes on dominant class



Experiments: synthetic data







Converge to Bayes solution (consistency)

Weight normalisation fails: correct label has -ve score!

Experiments: real-world data

Method	CIFAR-10-LT	CIFAR-100-LT	ImageNet-LT	iNaturalist
ERM	27.16	61.64	53.11	38.66
Weight normalisation ($\tau = 1$) (Kang et al., 2020)	24.02	58.89	52.00	48.05
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

Google

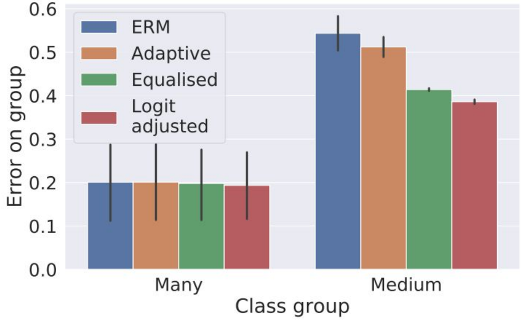
Logit adjustment techniques consistently perform well

Break-down of error rates

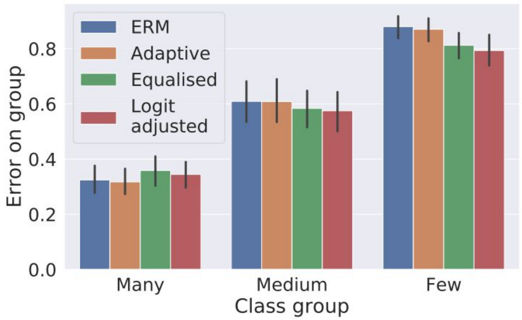
"Many": >= 100 examples

"Medium": [20, 100] examples

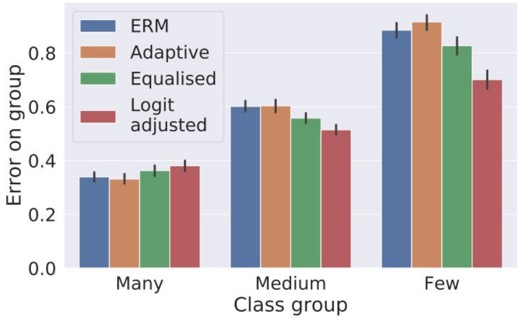
"Few": < 20 examples



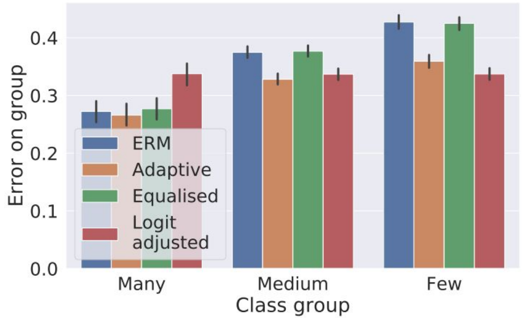
CIFAR10-LT



CIFAR100-LT



ImageNet-LT



iNaturalist

Sacrifice a little on "head" classes for gains on "tail" classes

Summary

Summary of our work

- ✓ A statistical perspective of long-tail learning
- ✓ Unify and generalise existing works
- ✓ New post-hoc and loss modification approaches