Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss

Kaidi Cao
Stanford University
kaidicao@stanford.edu

Colin Wei
Stanford University
colinwei@stanford.edu

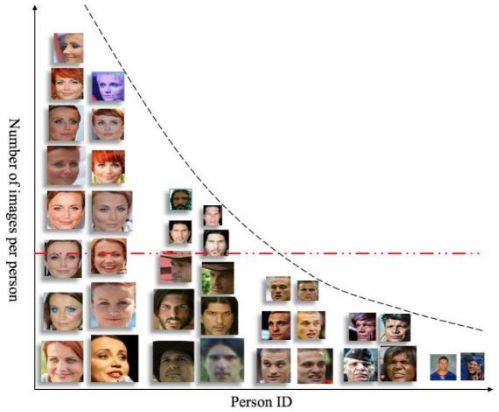
Adrien Gaidon
Toyota Research Institute
adrien.gaidon@tri.global

Nikos Arechiga
Toyota Research Institute
nikos.arechiga@tri.global

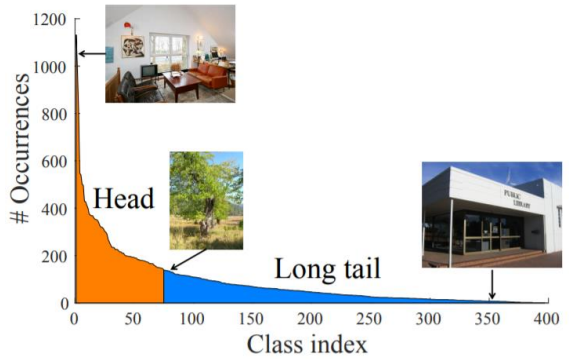
Tengyu Ma
Stanford University
tengyuma@stanford.edu

NIPS 2019

Class-Imbalance Problem

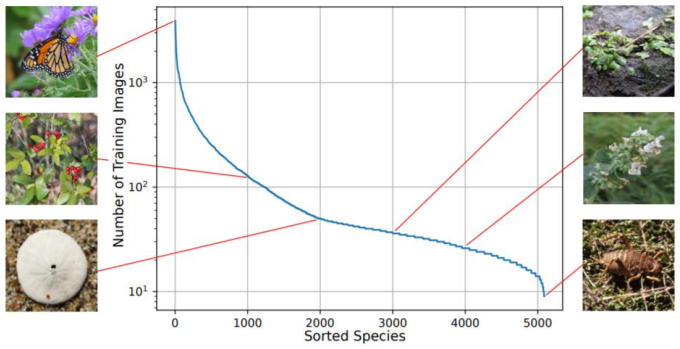


Faces [Zhang et al. 2017]

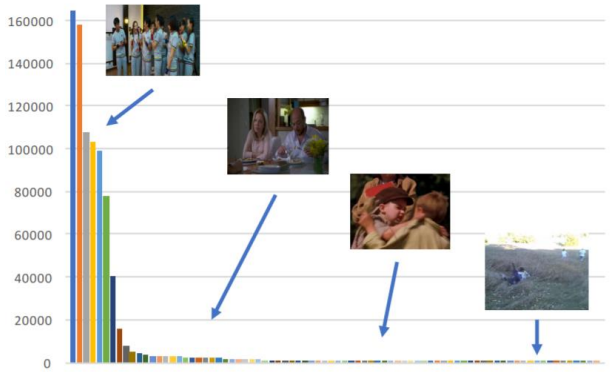


Places [Wang et al. 2017]

- **Long-tailed data is everywhere.**
- **Class-imbalance issue**: DNNs perform well on balanced data, but always fail on long-tailed tasks
- Intuitive solutions: re-sampling, re-weighting



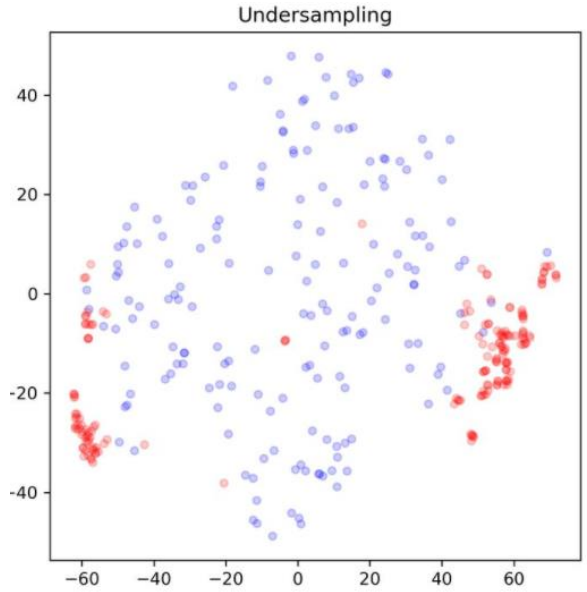
Species [Van Horn et al. 2019]



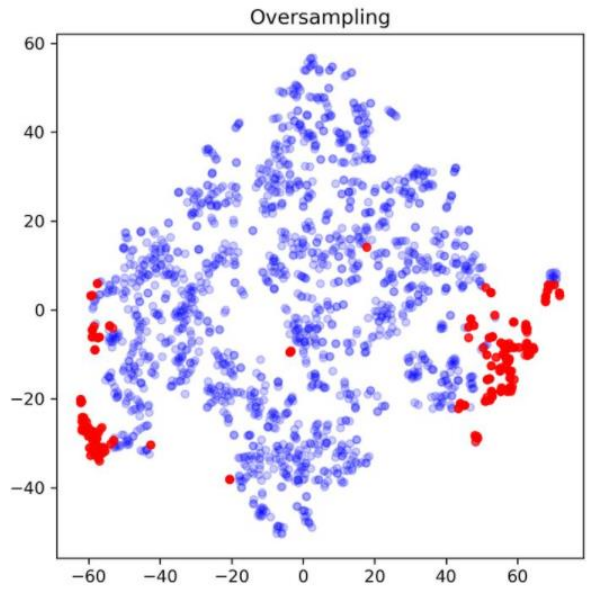
Actions [Zhang et al. 2019]

Limitations of Resampling Methods

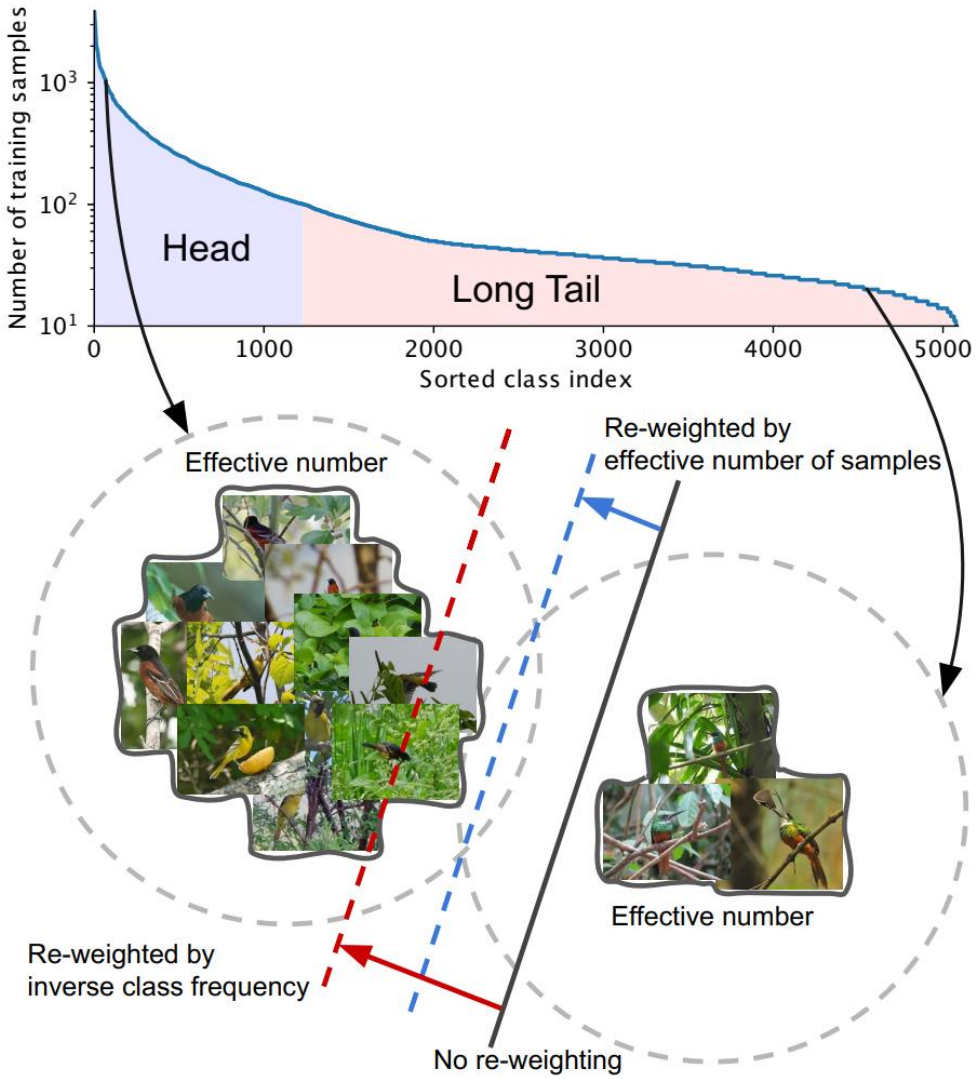
- Resampling: down-sampling for majority class
  - Expensive information is lost
  - Training multiple DNNs is costly



- Resampling: up-sampling for minority class
  - Overfitting to minority class



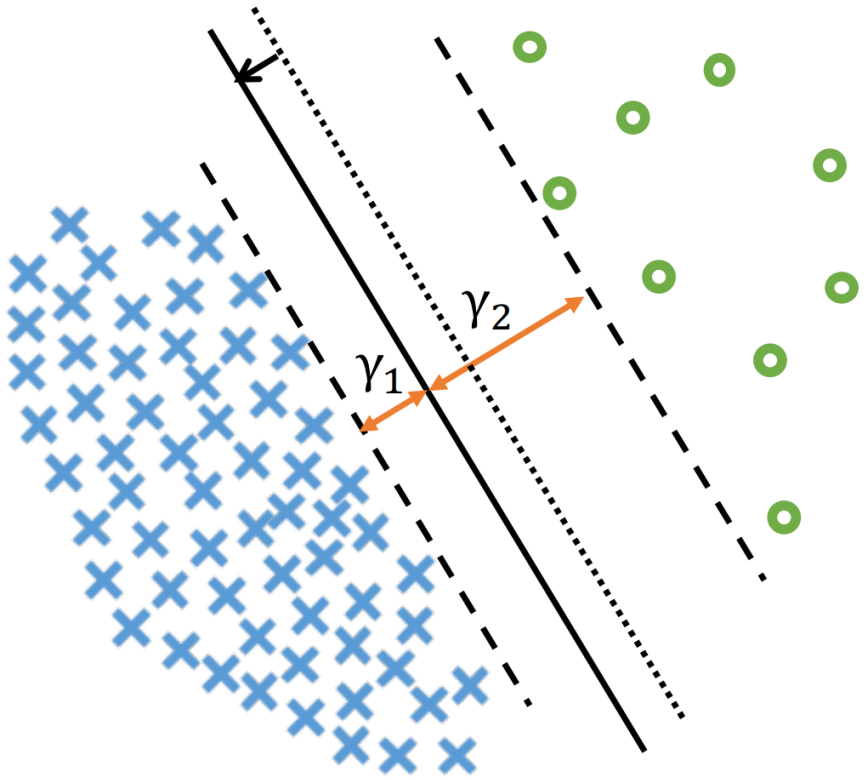
Limitations of Reweighting Methods



- Reweighting
  - Re-weighting loss by inverse class frequency cannot perform well.
  - There is information overlap among data.
  - Difficult to optimize when weights are large
- Solution: effective number
  - Inversely proportional to the effective number of samples.

A New Viewpoint: Margin

Core idea: Shifting the boundary by adjusting the margin.

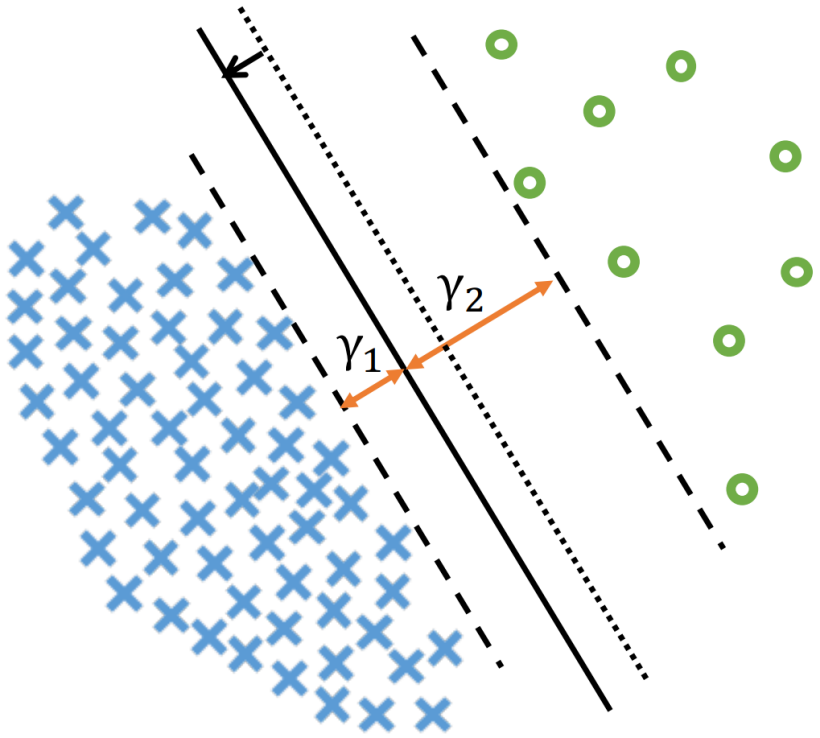


- γ_1 (γ_2) represents the margin toward class 1 (2).

How to find the optimal margin?

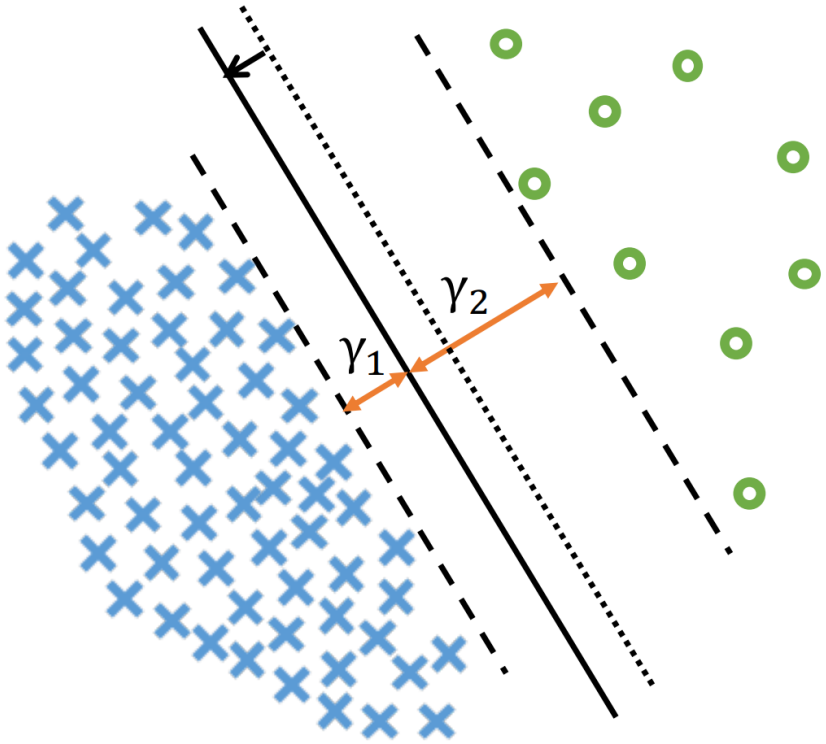
Label-Distribution-Aware Margin Loss

- Revisit the margin theory.
  - Define the margin of an example (x, y)
    - $\gamma(x, y) = f(x)_y - \max_{j \neq y} f(x)_j$
  - Define the margin for each class
    - $\gamma_j = \min_{i \in s_j} \gamma(x_i, y_i)$
  - Margin generalization bound
    - $\text{test error} \lesssim \frac{1}{\gamma_{min}} \sqrt{\frac{C(\mathcal{F})}{n}}$



Label-Distribution-Aware Margin Loss

- Find the optimal margin
  - Imbalanced setting:
    - $\frac{1}{2}\text{err}[\text{class1}] + \frac{1}{2}\text{err}[\text{class2}] \lesssim (\frac{1}{\gamma_1\sqrt{n_1}} + \frac{1}{\gamma_2\sqrt{n_2}})\sqrt{C(\mathcal{F})}$
  - Adjust γ_1 and γ_2 by shifting the boundary
    - $\min \frac{1}{\gamma_1\sqrt{n_1}} + \frac{1}{\gamma_2\sqrt{n_2}} \qquad \text{s.t.} \quad \gamma_1 + \gamma_2 = C$
  - Optimal solution: $\gamma_j = \frac{C}{n_j^{1/4}}$
  - Balanced case: $\gamma_1 = \gamma_2$



Label-Distribution-Aware Margin Loss

- Loss functions with efficient margins
  - Enforce $\gamma_i = C \cdot n_i^{-1/4}$
  - Extension of multi-class hinge loss

$$\mathcal{L}_{\text{LDAM-HG}}((x, y); f) = \max(\max_{j \neq y}\{z_j\} - z_y + \Delta_y, 0)$$

$$\text{where } \Delta_j = \frac{C}{n_j^{1/4}} \text{ for } j \in \{1, \ldots, k\}$$

  - Extension of cross entropy

$$\mathcal{L}_{\text{LDAM}}((x, y); f) = -\log \frac{e^{z_y - \Delta_y}}{e^{z_y - \Delta_y} + \sum_{j \neq y} e^{z_j}}$$

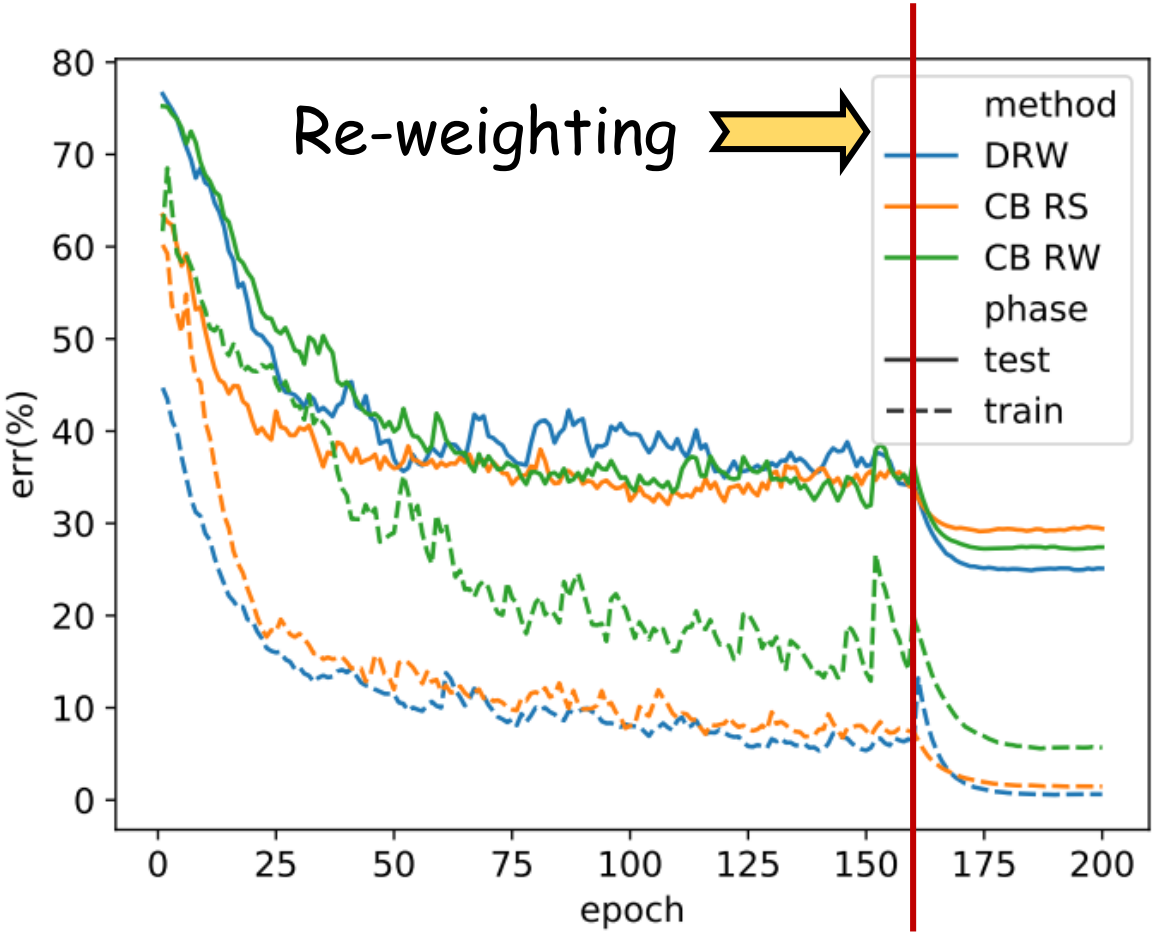
$$\text{where } \Delta_j = \frac{C}{n_j^{1/4}} \text{ for } j \in \{1, \ldots, k\}$$

Rethinking Training Strategies

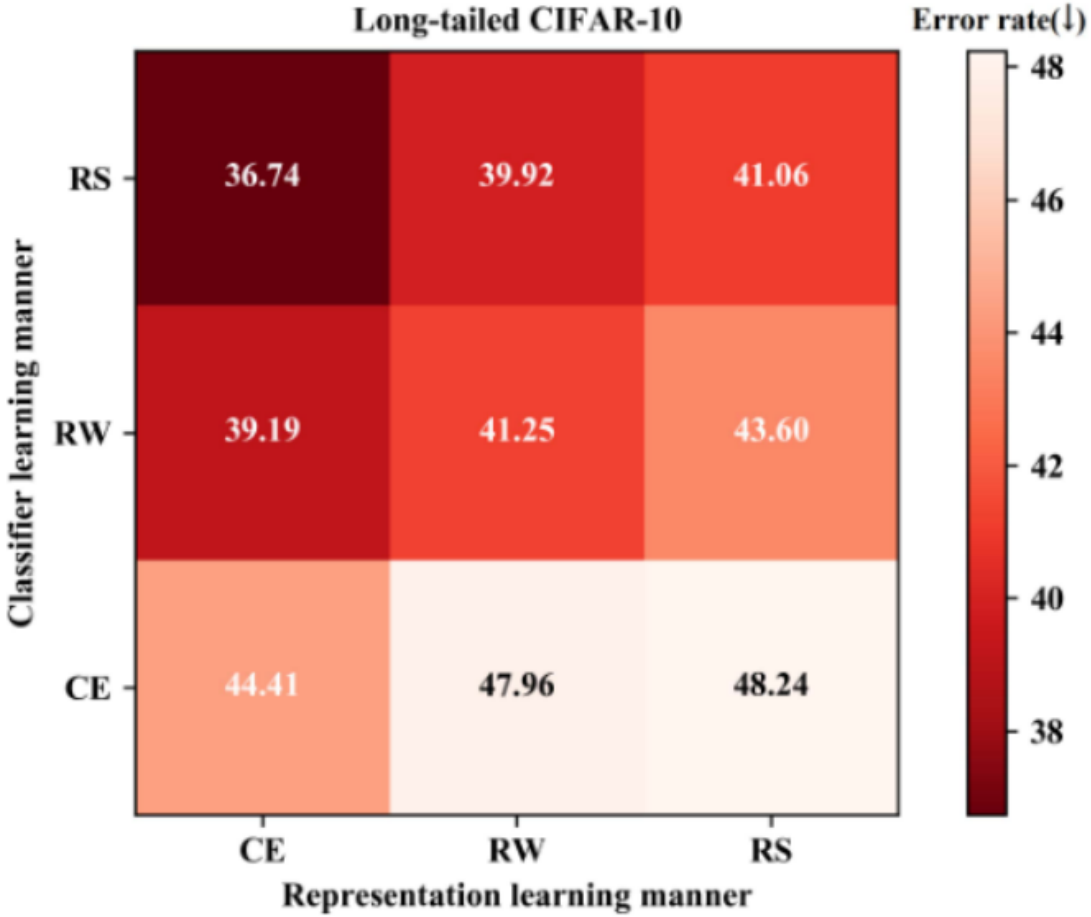
- ERM
  - Minority classes have worse training error.
- Re-sampling
  - Overfitting to minority classes.
- Re-weighting
  - Difficulties and instabilities in training.

Deferred Re-Weighting (DRW)

- Re-weighting (Re-sampling) only after annealing the learning rate.



Training: imbalanced
Testing: balanced



CVPR2020 [Zhou et al.]

LDAM-DRW

Algorithm 1 Deferred Re-balancing Optimization with LDAM Loss

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$. A parameterized model f_θ

Initialize the model parameters θ randomly
for $t = 1$ to T_0 **do**
    $\mathcal{B} \leftarrow \text{SampleMiniBatch}(\mathcal{D}, m)$ ▷ a mini-batch of m examples
    $\mathcal{L}(f_\theta) \leftarrow \frac{1}{m} \sum_{(x,y)\in\mathcal{B}} \mathcal{L}_{\text{LDAM}}((x, y); f_\theta)$
    $f_\theta \leftarrow f_\theta - \alpha \nabla_\theta \mathcal{L}(f_\theta)$ ▷ one SGD step
    Optional: $\alpha \leftarrow \alpha/\tau$ ▷ anneal learning rate by a factor τ if necessary

for $t = T_0$ to T **do**
    $\mathcal{B} \leftarrow \text{SampleMiniBatch}(\mathcal{D}, m)$ ▷ A mini-batch of m examples
    $\mathcal{L}(f_\theta) \leftarrow \frac{1}{m} \sum_{(x,y)\in\mathcal{B}} n_y^{-1} \cdot \mathcal{L}_{\text{LDAM}}((x, y); f_\theta)$ ▷ standard re-weighting by frequency
    $f_\theta \leftarrow f_\theta - \alpha \frac{1}{\sum_{(x,y)\in\mathcal{B}} n_y^{-1}} \nabla_\theta \mathcal{L}(f_\theta)$ ▷ one SGD step with re-normalized learning rate

Experiments: IMDB

Table 1: Top-1 validation errors on imbalanced IMDB review dataset. Our proposed approach LDAM-DRW outperforms the baselines.

Approach	Error on positive reviews	Error on negative reviews	Mean Error
ERM	2.86	70.78	36.82
RS	7.12	45.88	26.50
RW	5.20	42.12	23.66
LDAM-DRW	4.91	30.77	17.84

Class-imbalance: remove 90% of negative reviews.

- ERM: all the examples have the same weights.
- RS: each example is sampled with probability proportional to **the inverse sample size of its class**.
- RW: we re-weight each sample by **the inverse of the sample size of its class**, and then re-normalize to make the weights 1 on average in the mini-batch

Experiments: CIFAR

Table 2: Top-1 validation errors of ResNet-32 on imbalanced CIFAR-10 and CIFAR-100. The combination of our two techniques, LDAM-DRW, achieves the best performance, and each of them individually are beneficial when combined with other losses or schedules.

Dataset	Imbalanced CIFAR-10				Imbalanced CIFAR-100			
Imbalance Type	long-tailed		step		long-tailed		step	
Imbalance Ratio	100	10	100	10	100	10	100	10
ERM	29.64	13.61	36.70	17.50	61.68	44.30	61.45	45.37
Focal [Lin et al., 2017]	29.62	13.34	36.09	16.36	61.59	44.22	61.43	46.54
LDAM	26.65	13.04	33.42	15.00	60.40	43.09	60.42	43.73
CB RS	29.45	13.21	38.14	15.41	66.56	44.94	66.23	46.92
CB RW [Cui et al., 2019]	27.63	13.46	38.06	16.20	66.01	42.88	78.69	47.52
CB Focal [Cui et al., 2019]	25.43	12.90	39.73	16.54	63.98	42.01	80.24	49.98
HG-DRS	27.16	14.03	29.93	14.85	-	-	-	-
LDAM-HG-DRS	24.42	12.72	24.53	12.82	-	-	-	-
M-DRW	24.94	13.57	27.67	13.17	59.49	43.78	58.91	44.72
LDAM-DRW	**22.97**	**11.84**	**23.08**	**12.19**	**57.96**	**41.29**	**54.64**	**40.54**

- HG: Hinge loss
- DRS: Deferred Re-Weighting
- M-DRW: cross entropy loss with uniform margin

Imbalance ratio: $\rho = \max_i\{n_i\}/\min_i\{n_i\}$.

Experiments: iNaturalist

Table 3: Validation errors on iNaturalist 2018 of various approaches. Our proposed method LDAM-DRW demonstrates significant improvements over the previous state-of-the-arts. We include ERM-DRW and LDAM-SGD for the ablation study.

Loss	Schedule	Top-1	Top-5
ERM	SGD	42.86	21.31
CB Focal [Cui et al., 2019]	SGD	38.88	18.97
ERM	DRW	36.27	16.55
LDAM	SGD	35.42	16.48
LDAM	**DRW**	**32.00**	**14.82**

Thanks