Learn2Mix: Training Neural Networks Using Adaptive Data Integration

Shyam Venkatasubramanian[1], Vahid Tarokh[1]

[1]Duke University

Outline

Motivation: Faster, More Robust Training

- Standard training (k classes) implicitly uses *fixed* class proportions per batch $\tilde{\alpha}$.
- Classes differ in difficulty — harder classes deserve more emphasis during training.
- **Question:** Can we adapt batch class proportions using real-time error rates to **accelerate convergence** and **handle imbalance**?

Key Idea (Learn2Mix)

Adapt class proportions each epoch via instantaneous class-wise error rates:

$$\alpha^t \leftarrow \alpha^{t-1} + \gamma\left(\frac{\mathcal{L}(\theta^{t-1})}{1^T\mathcal{L}(\theta^{t-1})} - \alpha^{t-1}\right), \quad \text{and:} \quad \mathcal{L}(\theta^t, \alpha^t) = \sum_{i=1}^{k} \alpha_i^t \mathcal{L}_i(\theta^t).$$

Learn2Mix in Practice

Per-epoch loop:

1. Build batches by sampling $\alpha_i^t M$ examples from each class set J_i and combining (batch size M).
2. Compute per-class empirical losses $\mathcal{L}_i(\theta^t)$ from the batches.
3. Aggregate empirical loss and update parameters:

$$\mathcal{L}(\theta^t, \alpha^t) = \sum_{i=1}^{k} \alpha_i^t \, \mathcal{L}_i(\theta^t), \qquad \theta^{t+1} = \theta^t - \eta \, \nabla_\theta \mathcal{L}(\theta^t, \alpha^t).$$

4. **Update mixing parameters** using normalized class losses:

$$\mathbb{L}_i^t = \frac{\mathcal{L}_i(\theta^t)}{\sum_{j=1}^{k} \mathcal{L}_j(\theta^t)}, \quad \text{and:} \quad \alpha^{t+1} = \alpha^t + \gamma\big(\mathbb{L}^t - \alpha^t\big).$$

- Mixing rate $\gamma \in (0,1)$ controls how quickly proportions adapt.
- Works for classification, regression, reconstruction (group data into k partitions).

Learn2Mix in Practice (continued)

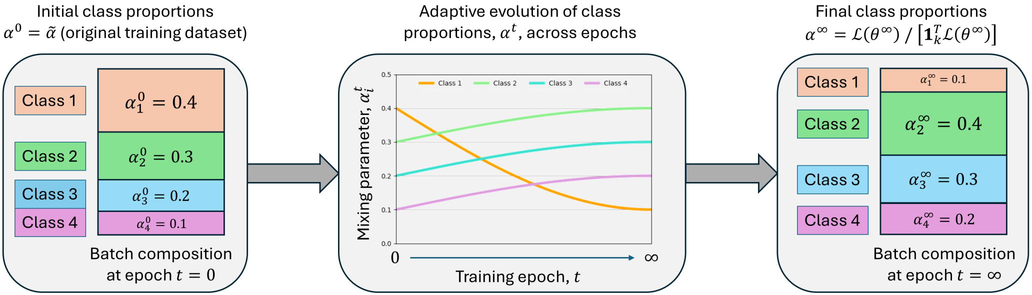


Illustration of the learn2mix training mechanism. The class-wise composition of batches is adaptively modified during training using instantaneous class-wise error rates.

Theoretical Results

Assume each per-class loss $\mathcal{L}_i(\theta)$ is μ_i-strongly convex with L_i-Lipschitz gradients. Suppose $\mu^* = \min_i \mu_i$, $L^* = \max_i L_i$, and let θ^* denote the optimal model parameters.

Proposition (Mixing Convergence): If $\eta \in (0, 2/L^*)$ and $\gamma \in (0,1)$, then:

$$\lim_{t\to\infty} \theta^t = \theta^*, \quad \text{and:} \quad \lim_{t\to\infty} \alpha^t = \alpha^* = \frac{\mathcal{L}(\theta^*)}{1^T\mathcal{L}(\theta^*)}.$$

Proposition (Faster Convergence): If the following condition holds:

$$(\star) \quad \left[\left(\frac{\mu^*}{2} - L^*\right)\|\theta^t - \theta^*\|^2 + \tilde{\alpha}^\top(\mathcal{L}(\theta^t) - \mathcal{L}(\theta^*))\right]\left[\|\theta^t - \theta^*\| - (\mathcal{L}(\theta^t) - \mathcal{L}(\theta^*))\right] > 0,$$

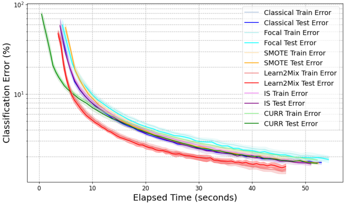
then for any learning rate $\eta > 0$ and mixing rate $\gamma \in (0, \beta]$:

$$\left\|\theta^t - \eta\nabla_\theta\mathcal{L}(\theta^t, \alpha^t) - \theta^*\right\| \le \left\|\theta^t - \eta\nabla_\theta\mathcal{L}(\theta^t, \tilde{\alpha}) - \theta^*\right\|.$$

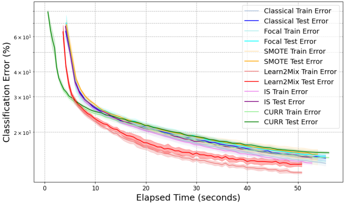
Interpretation: When $(\star)$ holds, reweighting toward higher-error classes (Learn2Mix) yields a one-step update that is *closer* to θ^* than the update under classical training with the same η.

Empirical Results: Classification

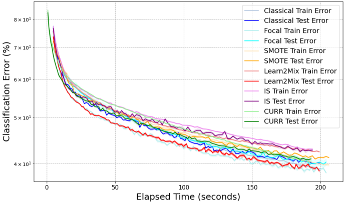
Learn2Mix yields faster convergence across various classification benchmarks.



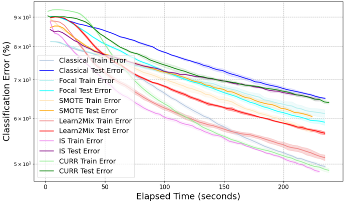
MNIST



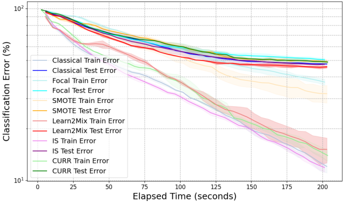
Fashion MNIST



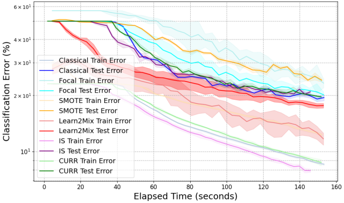
CIFAR-10



Imagenette (Imbalanced)



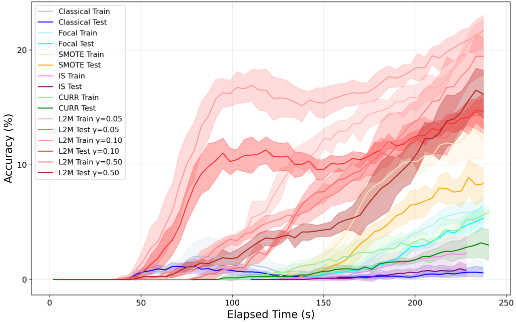
CIFAR-100 (Imbalanced)



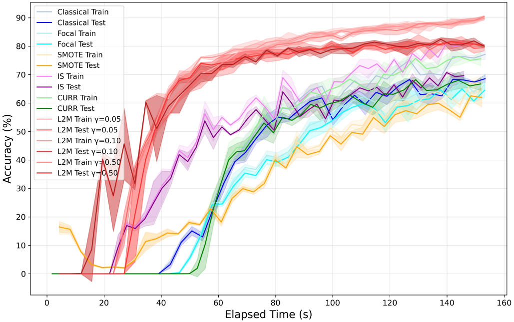
IMDB (Imbalanced)

Empirical Results: Tracking Worst-Class Performance

Learn2mix improves worst-class classification accuracies during training.



Imagenette (Imbalanced)

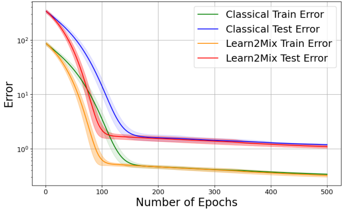


IMDB (Imbalanced)

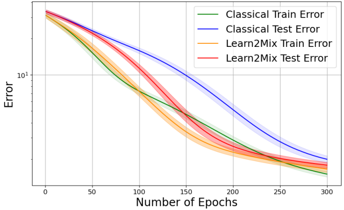
- Matches intuition: learn2Mix increases the proportion of harder classes during training
- Translates to stronger results for the most challenging classes.

Empirical Results: Regression & Reconstruction

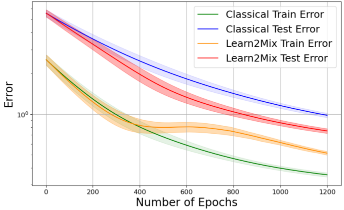
Learn2Mix yields faster convergence in regression and reconstruction settings.



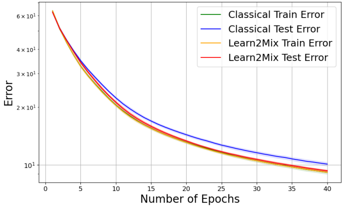
Mean Estimation



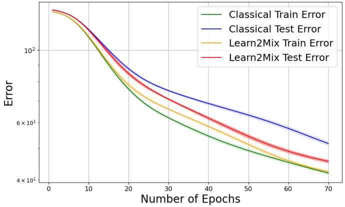
Wine Quality



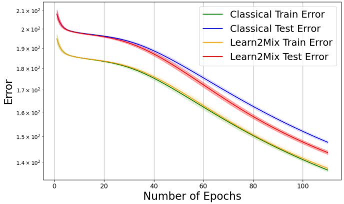
California Housing



MNIST (Reconstruction)



Fashion MNIST (Reconstruction)



CIFAR-10 (Reconstruction)

Conclusion

Takeaways

- **Learn2Mix:** reweights batches toward harder classes via current errors.
- **Theory:** converges to θ^*; mixing stabilizes at difficulty-aware α^*.
- **Practice:** faster convergence and stronger worst-class results across tasks.

Paper & Updates

Acknowledgements: Shyam Venkatasubramanian and Vahid Tarokh were supported in part by the U.S. Air Force Office of Scientific Research under award FA9550-21-1-0235.