# Instance-dependent Early Stopping

Suqin Yuan, Runqi Lin — The University of Sydney
Lei Feng — Southeast University
Bo Han — Hong Kong Baptist University
Tongliang Liu — The University of Sydney

## Can we early stop training at instance-level?

**Method:** Once the model has mastered an instance, the training on it should stop.

## 1. How to defined ‘mastered’ ?

- Intuitively, $\left|\Delta^2 Loss_i(w^{(t)})\right| < \delta$ means mastered.

  However, it fails to account for:

  1. Variation in sample difficulty
  2. Class (sub-class) distribution imbalance
  3. Conflicting optimization objectives between samples

- We using <u>loss plateaus</u> as a proxy measure:

  Performance on a particular sample do not increase during training, we consider that sample to have reached its learning potential under current conditions.

We use the second-order difference to identify the mastered instances, which quantifies the rate of change in the loss for sample $i$ across three consecutive epochs, $t^{th}$, $(t-1)^{th}$, and $(t-2)^{th}$ training epochs. The second-order difference is defined as:

$$\Delta^2 L_i(w^{(t)}) = [L_i(w^{(t)}) - L_i(w^{(t-1)})] - [L_i(w^{(t-1)}) - L_i(w^{(t-2)})]$$
$$= L_i(w^{(t)}) - 2L_i(w^{(t-1)}) + L_i(w^{(t-2)}).$$

$\left|\Delta^2 Loss_i(w^{(t)})\right| < \delta$ means mastered.

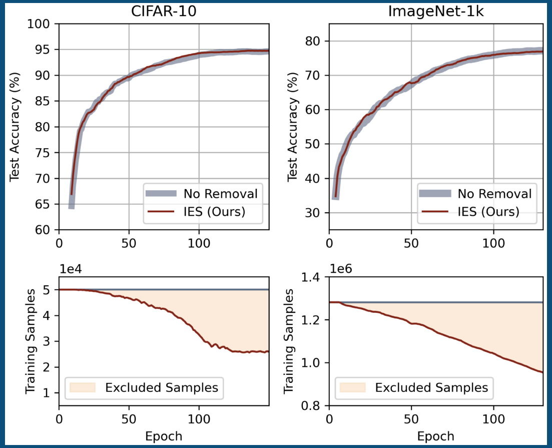
The approach provides a uniform standard to evaluate learning progress across samples, regardless of their loss values or inherent complexity.

## 2. Benefits from IES

- Computational Efficiency:
  - IES progressively exclude mastered samples from the training process, directly lowering computational requirements.
  - Sample assessment is a "free lunch" as it leverages information already collected during forward passes.
  - Every $k$ epochs, IES reintroduces all samples (including excluded samples) in forward pass to re-include those exhibiting loss fluctuations.
- Avoiding Over-Memorization [1,2]:
  - By limiting excessive repetition of easily learned examples, IES reduces harmful overfitting.
  - IES decreases the sharpness of the model's loss landscape. (Figure 4 in paper)
  - IES potentially improves both generalization performance and transferability of the resulting models. (Table 3 in paper)
- Robust to label noise and fairness.

[1] Do we need zero training loss after achieving zero training error? In ICML 2020. Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama.
[2] On the over-memorization during natural, robust and catastrophic overfitting. In ICLR 2024. Runqi Lin, Chaojian Yu, Bo Han, and Tongliang Liu.

## 3. Results:




Average Wall-time Speedup:
(lossless speedup)

CIFAR-10: ~1.4×
CIFAR-100: ~1.2×
ImageNet-1k: ~1.4×