AGE OF LEARNING: TEMPORAL PERSISTENCE OF PREDICTION ERRORS AS A LEARNING SIGNAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Current machine learning algorithms primarily rely on instantaneous signals such as loss, margin, and prediction confidence to characterize model behavior. These signals indicate how difficult a prediction is at the current optimization step, but they do not capture how long the model has remained incorrect. We study this temporal dimension of learning and introduce Age of Learning (AoL), a learning-state variable that measures the persistence of prediction errors over time. AoL increases while an error remains unresolved and resets when a correct prediction is achieved, thereby distinguishing persistent under-learning from transient mistakes. We develop AoL-based training strategies for both offline and streaming settings. In offline learning, sample-level AoL is accumulated over training and aggregated into class-level states that guide adaptive reweighting and resampling. In streaming learning, where full historical access is unavailable, we maintain lightweight class-level AoL states using current and buffered observations. Across long-tailed classification settings, AoL improves or matches standard training baselines, with larger benefits when learning difficulty persists over time. Multi-seed streaming experiments further show reproducible gains under temporally stable imbalance. Analysis of class frequency, loss, and margin shows that AoL is related to conventional difficulty measures but captures additional information about error duration. These results suggest that temporal persistence provides a useful complementary signal for characterizing and controlling learning dynamics in imbalanced and non-stationary environments.

1 INTRODUCTION

Modern machine learning and deep learning models (Bishop, 2006; Goodfellow et al., 2016) are commonly trained using instantaneous signals such as loss (Cui et al., 2019), margin (Cao et al., 2019), and prediction confidence (Lin et al., 2017), which quantify how well a model currently fits a sample, while dataset-level statistics such as class frequency characterize the training distribution. These signals support a wide range of learning strategies, including class reweighting (Cui et al., 2019), hard-example mining (Lin et al., 2017), margin-based objectives, and adaptive sampling (Hacohen & Tuytelaars, 2025). However, they leave out a different aspect of learning: how long a prediction error has remained unresolved.

This temporal information can matter when learning difficulty persists across optimization steps. Consider two samples that are currently misclassified with similar loss or margin. One may have only recently become incorrect, whereas the other may have remained misclassified for many consecutive epochs. Instantaneous difficulty measures treat these cases similarly despite their different learning histories. The same issue arises at the class level: classes with similar frequencies or current losses may exhibit very different histories of persistent under-learning. Such behavior is particularly relevant in long-tailed learning (Cui et al., 2019; Liu et al., 2019; Cao et al., 2019) and streaming data (Aljundi et al., 2019; Csaba et al., 2024; Wang et al., 2025). In long-tailed datasets (Zhang et al., 2023), rare classes may remain misclassified for extended periods due to insufficient representation, while in streaming scenarios certain classes may be neglected or become stale as the data distribution evolves. In both cases, it is not only the error magnitude that matters, but also its temporal persistence across successive optimization steps throughout training.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This work studies error persistence as an explicit learning signal. We introduce Age of Learning (AoL), a temporal learning-state variable that measures how long prediction error persists during training. For each sample or class, AoL increases when misclassification occurs and resets when correct classification is achieved. We illustrate the core idea in Figure 1.

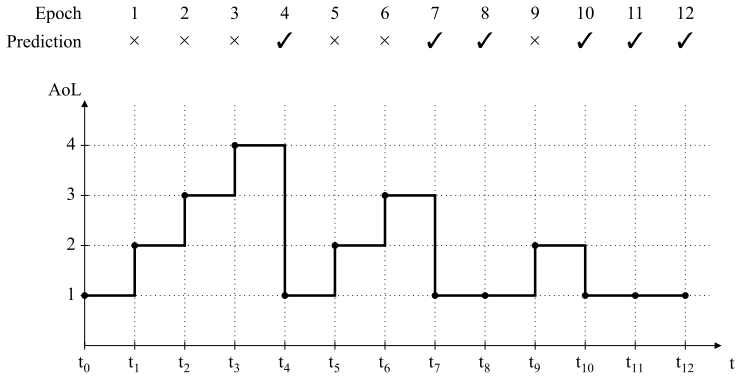


Figure 1: Illustration of Age of Learning (AoL). AoL increases when a sample or class remains misclassified and resets when correct classification is achieved. Rather than measuring instantaneous error severity, AoL captures error duration as a temporal learning-state variable that evolves with model updates to guide subsequent optimization.

AoL complements instantaneous difficulty signals by measuring how long prediction errors persist. A closely related temporal concept is Age of Incorrect Information (AoII) (Maatouk et al., 2020), which measures how long a receiver's estimate remains incorrect relative to an external source in communication systems. In contrast, AoL defines an endogenous temporal state of the learning process itself: it is derived from a trainable model's prediction trajectory, evolves with model updates, and feeds back into subsequent optimization. This enables persistent prediction error to serve directly as a learning-state variable for adaptive training.

We first formulate AoL in an offline setting, where sample-level AoL is computed and aggregated into class-level signals that guide training, particularly in long-tail scenarios. We then extend AoL to streaming learning, where full historical access is unavailable and sample-level tracking is infeasible, by maintaining lightweight per-class states updated from the current stream and optional buffered data. We evaluate AoL on long-tailed classification tasks in both offline and streaming settings. The results show that, in offline settings, AoL-based training improves or matches standard baselines; in streaming scenarios, it enables adaptive reweighting based on persistent misclassification and shows clear advantages when learning difficulty evolves with temporal consistency. More broadly, our results suggest that temporal behavior can serve as a useful optimization signal beyond static dataset statistics or instantaneous prediction errors. The main contributions of this work are:

- We introduce AoL, a temporal learning-state variable that explicitly represents the persistence of prediction errors.
- We develop AoL formulations for both offline and streaming learning, including sample-level and class-level states and lightweight online updates under restricted historical access.
- We show how AoL can be incorporated into adaptive training through loss reweighting and weighted resampling, providing concrete mechanisms for testing whether temporal persistence is useful for optimization.
- We empirically analyze the relationship between AoL and class frequency, loss, and margin, and show that AoL captures temporal information not determined by these conventional signals.
- We demonstrate reproducible improvements in long-tailed offline and streaming settings, with the strongest benefits arising when learning difficulty itself persists over time.

2 Related Work

We review prior work most relevant to AoL, including long-tail learning, sample reweighting, online learning, learning dynamics, and age-based metrics, and clarify the positioning of AoL relative to these areas across offline and streaming settings.

Long-tail learning. Long-tail learning addresses highly imbalanced class distributions (Zhang et al., 2021; Alshammari et al., 2022; Shi et al., 2023; Kunstner et al., 2024; Sanchez Aimar et al., 2025; Li et al., 2025; Cortes et al., 2025). Existing approaches include class-sensitive loss design and reweighting based on label frequency or effective sample size, margin-based objectives (Cui et al., 2019; Cao et al., 2019), decoupled representation and classifier learning (Kang et al., 2020), and resampling or hard-example emphasis (Shrivastava et al., 2016; Lin et al., 2017; Chou et al., 2020; Shi et al., 2023). These approaches primarily address imbalance through class statistics, margin design, or data-level interventions. AoL addresses a different dimension: classes with similar frequencies may exhibit very different learning trajectories, and AoL measures how long under-learning persists during optimization rather than inferring importance from sample count alone.

Sample reweighting. Example-weighting and selection methods construct importance from various signals. Meta-reweighting methods learn weights through meta-gradients and/or a clean validation or meta set (Ren et al., 2018; Shu et al., 2019); optimal-transport reweighting targets distributional rebalancing towards a balanced meta distribution (Guo et al., 2022); importance sampling primarily reduces stochastic-gradient variance while maintaining an unbiased estimator of the original objective (Katharopoulos & Fleuret, 2018). Other approaches construct weighting from loss-based objectives or related instantaneous difficulty signals (Yi et al., 2021; Sow et al., 2025), while data-selection methods use early-training gradient or error scores (Paul et al., 2021). AoL instead derives importance from the duration of unresolved prediction errors across updates rather than from static statistics or instantaneous difficulty. We incorporate this temporal signal into training through loss reweighting and weighted resampling.

Online learning. Online and streaming learning studies how models can be updated when data arrive sequentially, often under strict memory and access constraints (Fini et al., 2020; Flaspohler et al., 2021; Wu et al., 2024; Zhang et al., 2025; Sridharan & Yoo, 2025; Daniely et al., 2025; Liu et al., 2026). A large body of work addresses catastrophic forgetting (Lopez-Paz & Ranzato, 2017; Kirkpatrick et al., 2017) and non-stationarity (Monteleoni & Jaakkola, 2003; Zhang et al., 2025) through replay mechanisms (Aljundi et al., 2019), memory buffers (Hacohen & Tuytelaars, 2025), or continual-learning regularization (Csaba et al., 2024). Replay-based methods retain past samples to support learning. Our streaming formulation maintains a class-level AoL state to guide training emphasis and can use a replay buffer to estimate current class performance. In long-tailed streams, infrequently observed classes may remain under-learned for long periods.

Learning dynamics. Prior work studies temporal learning behavior through forgetting events and related sample-level dynamics (Toneva et al., 2019; Hacohen & Tuytelaars, 2025; Zheng et al., 2026; Jin et al., 2026). Example-forgetting methods record transitions between learned and forgotten states and show that prediction trajectories reveal information beyond final loss or accuracy. AoL shares this view but measures a different property: forgetting statistics characterize transitions between correctness states, whereas AoL measures the duration of consecutive incorrectness. Repeated short error episodes and one long uninterrupted error episode may therefore produce similar transition counts but substantially different AoL trajectories.

Other age-based metrics. Temporal notions of "age" have also been studied in communication and status-update systems (Sun et al., 2017; Kadota et al., 2018; 2019; Yates et al., 2021). Age of Information (AoI) (Kaul et al., 2012; Kam et al., 2016) measures the freshness of received information, while Age of Incorrect Information (AoII) (Maatouk et al., 2020) measures how long a receiver estimate remains incorrect relative to an external process. AoII is the closest conceptual antecedent to AoL because both assign temporal significance to persistent incorrectness. However, the state, dynamics, and role are different: AoII is defined by disagreement between an external source and a receiver estimate and is used to guide communication or update policies, whereas AoL is an endogenous learning-state variable derived from a trainable model's prediction trajectory relative to supervision and feeds back into subsequent optimization. AoL formulates persistent prediction error as a temporal state for characterizing and guiding learning.

3 Age of Learning

In this section, we formalize AoL as a temporal learning-state variable and describe how it can be incorporated into model training. We first introduce the quantities in the standard offline setting and then extend the formulation to streaming learning.

3.1 OFFLINE AOL

We first define AoL in an offline setting, where the full dataset is accessible during training. Let $i \in \{1, \ldots, N\}$ index training samples, and $e \in \{1, \ldots, E\}$ denote the training epoch index. The model prediction for sample i at epoch e is given by:

$$\hat{y}_i^{(e)} = \arg\max_k f_{\theta^{(e)}}(\boldsymbol{x}_i)_k, \quad k \in \{1, \ldots, C\} \tag{1}$$

where $f_{\theta^{(e)}} : \mathcal{X} \to \mathbb{R}^C$ denotes a neural network parameterized by $\theta^{(e)}$ at epoch e, which maps an input $\boldsymbol{x}_i$ to a C-dimensional logit vector, where C is the total number of classes, and $f_{\theta^{(e)}}(\boldsymbol{x}_i)_k$ represents the logit corresponding to class k. The predicted label $\hat{y}_i^{(e)}$ is therefore the class index with the maximum logit value. Next, we define a binary correctness indicator $c_i^{(e)} = \mathbb{I}[\hat{y}_i^{(e)} = y_i]$, where $y_i \in \{1, \ldots, C\}$ denotes the ground-truth label of sample i and $\mathbb{I}[\cdot]$ denotes the indicator function. Thus, $c_i^{(e)} = 1$ indicates that the model correctly classifies sample i at epoch e, while $c_i^{(e)} = 0$ indicates a misclassification.

At the beginning of training, the AoL of sample i is initialized as one. At the end of each epoch, the AoL is updated based on the prediction correctness such that $a_i^{(e)} = 1$ if the sample is correctly classified, otherwise $a_i^{(e)} = a_i^{(e-1)} + 1$, as follows:

$$a_i^{(e)} = \begin{cases} 1, & \text{if } c_i^{(e)} = 1, \\ a_i^{(e-1)} + 1, & \text{otherwise}, \end{cases} \tag{2}$$

which can be equivalently written in a compact form as:

$$a_i^{(e)} = (1 - c_i^{(e)})(a_i^{(e-1)} + 1) + c_i^{(e)}. \tag{3}$$

By definition, the instantaneous AoL state $a_i^{(e)}$ records the duration of the current consecutive misclassification episode. Here we use a one-based convention in which 1 is the minimum AoL state. This gives all samples equal positive initial states and avoids a degenerate zero denominator when AoL is subsequently normalized to construct training weights.

Instantaneous AoL may fluctuate as predictions switch between correct and incorrect states. To obtain a smoother learning-state variable that retains information from previous error episodes, we maintain an aggregated AoL, denoted by $\bar{a}_i^{(e)}$, using an exponential moving average (EMA):

$$\bar{a}_i^{(1)} = a_i^{(1)}, \tag{4}$$

$$\bar{a}_i^{(e)} = (1 - \rho)\bar{a}_i^{(e-1)} + \rho a_i^{(e)}, \qquad e \geq 2, \tag{5}$$

where $\rho \in [0, 1]$ controls the trade-off between historical memory and responsiveness to the current error episode. A larger ρ makes $\bar{a}_i^{(e)}$ respond more strongly to recent prediction behavior, whereas a smaller ρ produces a smoother state with longer temporal memory. The two limiting cases are illustrative: when $\rho = 1$, then $\bar{a}_i^{(e)} = a_i^{(e)}$, only the current error episode is retained; when $\rho = 0$, then $\bar{a}_i^{(e)} = \bar{a}_i^{(1)}$, and the aggregated state remains fixed at its initialization. Thus, nonzero intermediate values of ρ explicitly combine current error persistence with historical learning behavior.

Compared to the instantaneous AoL $a_i^{(e)}$, which measures *how long the current incorrect state has persisted*, the aggregated AoL $\bar{a}_i^{(e)}$ summarizes *how persistent incorrectness has been over the learning process*. In the following section, we use the aggregated AoL to construct sample-level and class-level training signals.

3.2 AOL-GUIDED REWEIGHTING AND RESAMPLING

We now describe how the resulting AoL signals are used to guide the optimization process. Since larger AoL values indicate persistent misclassifications, the training process should prioritize samples or classes with higher AoL to improve their contribution to the model. The aggregated AoL $\bar{a}_i^{(e)}$ serves as an indicator for the accumulated misclassifications, and it is subsequently used to construct

training weights. To ensure scale invariance, we normalize the aggregated AoL across all N training samples and define a sample-level weight for each training sample i as:

$$w_i^{(e)} = \frac{\bar{a}_i^{(e)}}{\frac{1}{N}\sum_{j=1}^{N} \bar{a}_j^{(e)}}. \tag{6}$$

This normalization preserves the overall magnitude of the loss while relatively emphasizing samples with larger AoL values, i.e., those whose misclassifications have remained for longer periods.

Furthermore, we consider a class-level formulation, which is the primary setting used in our method. Let $\mathcal{I}_c$ denote the set of samples belonging to class c, and $|\mathcal{I}_c|$ denote the number of elements in $\mathcal{I}_c$. The class-level AoL is defined as:

$$\bar{a}_c^{(e)} = \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \bar{a}_i^{(e)}, \tag{7}$$

which represents the average aggregated AoL over all samples in class c. Meanwhile, the corresponding class-level weight for each class c is defined as:

$$w_c^{(e)} = \frac{\bar{a}_c^{(e)}}{\frac{1}{C}\sum_{k=1}^{C} \bar{a}_k^{(e)}}. \tag{8}$$

Compared to the sample-level weights, which directly reflect individual sample dynamics but may suffer from high variance, the class-level weights aggregate information across samples within the same class and therefore provide a more stable signal in long-tailed settings. In practice, both variants are evaluated in the experiments, and the latter generally yields more consistent improvements.

Next we incorporate the AoL-based weights into the training objective. We consider two commonly used classification losses: Softmax cross-entropy and Sigmoid cross-entropy. For the Softmax cross-entropy, the per-sample loss is defined as:

$$\ell_i^{\text{Softmax}} = -\log \frac{\exp(z_{i,y_i})}{\sum_{k=1}^{C} \exp(z_{i,k})}, \tag{9}$$

where $z_{i,k} = f_\theta(\boldsymbol{x}_i)_k$ denotes the logit corresponding to class k. For the Sigmoid cross-entropy, we adopt a one-vs-all formulation:

$$\ell_i^{\text{Sigmoid}} = -\sum_{k=1}^{C} \Big[y_{i,k} \log \sigma(z_{i,k}) + (1 - y_{i,k}) \log(1 - \sigma(z_{i,k})) \Big], \tag{10}$$

where $y_{i,k}$ is the one-hot label and $\sigma(\cdot)$ is the Sigmoid function.

To incorporate the AoL information, we apply loss reweighting by multiplying each sample's loss with its corresponding weight. This can be written in a unified form as:

$$\mathcal{L}^{(e)} = \frac{\sum_{i \in \mathcal{B}} w_i^{(e)} \, \ell_i}{\sum_{i \in \mathcal{B}} w_i^{(e)}}, \tag{11}$$

where ℓ_i denotes either ℓ_i^{Softmax} or ℓ_i^{Sigmoid}, and $\mathcal{B}$ denotes the mini-batch. When using class-level weighting, the weight is given by $w_i^{(e)} = w_{y_i}^{(e)}$, where $i \in \{1, \ldots, N\}, y_i \in \{1, \ldots, C\}$. This assigns the same weight to all samples belonging to one class, i.e., $w_i^{(e)} = w_c^{(e)}$ for all $i \in \mathcal{I}_c$.

In addition to loss reweighting, we also consider incorporating AoL-based weighting into the data sampling process. Specifically, we employ a weighted random sampler with replacement, where the sampling probability of each sample i at epoch e is proportional to its weight $w_i^{(e)}$. In the sample-level setting, $w_i^{(e)}$ is directly used as the sampling weight. In the class-level setting, the sampling weight is obtained by assigning the corresponding class weight to each sample, i.e., $w_i^{(e)} = w_{y_i}^{(e)}$.

Overall, loss reweighting adjusts the contribution of each sample in the objective function, while weighted resampling modifies the data distribution seen by the model. Both mechanisms are guided by the AoL-based signals, encouraging the model to prioritize samples or classes whose incorrectness has become persistent. Computing offline AoL requires one additional forward-only evaluation per epoch, together with $O(N)$ sample-level states and $O(C)$ class-level states. The current and aggregated AoL can be maintained incrementally, without storing the full prediction history or requiring an additional backward pass.

3.3 ONLINE AOL

We now extend AoL to streaming learning, where data arrive sequentially and the full training set is no longer repeatedly accessible. In this setting, maintaining a persistent sample-level state for every previously observed example is generally impractical. To address this, we introduce an online formulation of AoL that maintains a lightweight class-level state using the current data stream together with an optional memory buffer.

We use $t \in \{1, \ldots, T\}$ to denote the streaming update step, in contrast to epoch index e used in the offline training. Let $\mathcal{B}^{(t)}$ denote the mini-batch observed at step t, and $\mathcal{M}^{(t)}$ a memory buffer containing a subset of previously observed samples. The buffer is updated according to a chosen memory policy and is subject to a fixed storage budget. We define the observation set used to estimate the current class-level learning status as:

$$\mathcal{S}^{(t)} = \begin{cases} \mathcal{M}^{(t)}, & \text{if a memory buffer is used,} \\ \mathcal{B}^{(t)}, & \text{otherwise.} \end{cases} \tag{12}$$

For each class c represented in $\mathcal{S}^{(t)}$, we compute the current class-level classification accuracy as:

$$r_c^{(t)} = \frac{\sum_{(x,y)\in\mathcal{S}^{(t)}} \mathbb{I}[y = c]\mathbb{I}[\hat{y}^{(t)}(x) = c]}{\sum_{(x,y)\in\mathcal{S}^{(t)}} \mathbb{I}[y = c]}, \tag{13}$$

where $\hat{y}^{(t)}(x)$ denotes the prediction of the current model at step t for input x.

Unlike the offline formulation, where sample-level error episodes can be tracked explicitly, the streaming setting provides only partial and time-varying observations of each class. We therefore update a class-level AoL state according to whether the estimated class accuracy falls below a threshold $\tau^{(t)}$:

$$a_c^{(t)} = \begin{cases} a_c^{(t-1)} + 1, & \text{if } r_c^{(t)} < \tau^{(t)}, \\ 1, & \text{if } r_c^{(t)} \geq \tau^{(t)}, \\ a_c^{(t-1)}, & \text{if class } c \text{ is not observed in } \mathcal{S}^{(t)}. \end{cases} \tag{14}$$

All class-level states are initialized as one. The threshold $\tau^{(t)}$ specifies the accuracy level at which an under-learning episode is considered to have ended. It may be chosen according to task requirements or adapted to the observed data. If an observed class has estimated accuracy below $\tau^{(t)}$, its class-level AoL increases, indicating persistent under-learning. Once its estimated accuracy reaches or exceeds $\tau^{(t)}$, the current under-learning episode is considered resolved, and the state resets to one. If the class is not observed, its state remains unchanged. Thus, persistence is measured over observed class evaluations, rather than elapsed streaming time.

To smooth fluctuations in the estimated class-level learning status, we obtain online class weights by applying the same EMA aggregation and mean normalization as in the offline setting to the class-level AoL states. Each observed sample is then weighted by its ground-truth class for online loss reweighting. This online AoL formulation provides a simple and practical mechanism for updating class-level learning states in streaming scenarios. It requires only $O(C)$ additional AoL state beyond the memory buffer and can operate directly on the current mini-batch when no buffer is available. Its computational overhead depends on the size of the observation set $\mathcal{S}^{(t)}$ and the frequency of AoL evaluations. More algorithmic details are provided in Algorithm 2, Appendix A.

4 EXPERIMENTS

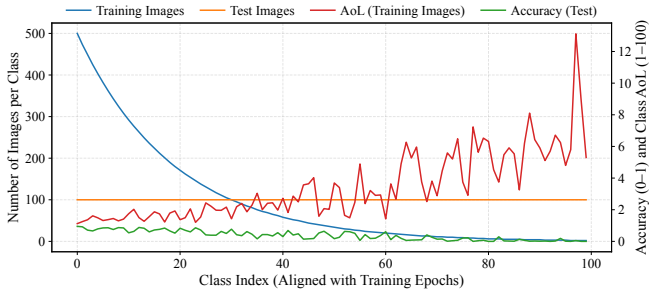
In this section, we first examine the information captured by AoL, then evaluate it in offline and streaming long-tailed settings, including both stable and dynamically evolving streams.

4.1 MOTIVATING OBSERVATION

We first examine whether AoL captures information beyond class frequency and conventional instantaneous difficulty signals. We use CIFAR-100 (Krizhevsky & Hinton, 2009) with imbalance



(a) The CIFAR-100 Dataset (Balanced)



(b) The Long-Tailed CIFAR-100 Dataset (IF = 200)

Figure 2: Relationship between class frequency, class-level AoL, and per-class test accuracy on CIFAR-100 under (a) balanced IF=1 and (b) long-tailed IF=200 settings. The blue and orange curves represent the numbers of training and test images per class (left y-axis). The red curve shows class-level AoL within $[1, 100]$, and the green curve shows the final per-class test accuracy within $[0, 1]$ (right y-axis). Under IF=1, all classes have identical training frequency but exhibit different AoL values. Under IF=200, it is highly apparent that accuracy is negatively related to AoL, and classes with similar tail frequencies also exhibit different AoL states, indicating that AoL captures variation in persistent under-learning beyond class frequency alone.

factor (IF) 1 and 200, representing balanced and highly imbalanced settings, respectively. The long-tailed training set is constructed as:

$$n_r = n_{\max}\mu^{-\frac{r}{(C-1)}}, \quad r = 0, \ldots, C-1, \tag{15}$$

where n_r is the number of training samples for the class at long-tail rank r, $n_{\max}$ is the number of samples in the head class, C is the number of classes, and μ is the target IF. We train ResNet-34 (He et al., 2016) for 100 epochs with batch size 128.

Figure 2 compares class frequency, class-level AoL, and final per-class test accuracy. Under IF=1, all classes have identical training frequencies but different AoL values, showing that class frequency alone does not determine the temporal learning state. This pattern becomes more pronounced under IF=200, where several tail classes with identical or nearly identical sample counts exhibit markedly different AoL values. Thus, prediction-error persistence can differ even among classes with comparable data scarcity. Across both settings, larger class-level AoL tends to correspond to lower final test accuracy, with Pearson correlations of -0.8231 for IF=1 and -0.8225 for IF=200. These strong negative correlations show that AoL captures persistent under-learning beyond class counts alone.

Correlation analysis. We further compare AoL with conventional instantaneous difficulty signals. For IF-1, the Pearson correlations of class-level AoL with class-level training loss/margin are 0.7841/-0.7757, and those with class-level test loss/margin are 0.6352/-0.6882. For IF-200, the corresponding correlations are 0.6697/-0.6633 on the training set and 0.4300/-0.3883 on the test set. Persistently difficult classes therefore tend to have larger losses and smaller margins, but the correlations are not perfect: loss and margin characterize the current prediction state, whereas AoL additionally captures how long incorrectness persists. Hence, AoL is related to, but distinct from, class frequency, loss, and margin. More detailed AoL dynamics are provided in Appendix D.

4.2 Offline Settings

We next evaluate AoL in offline long-tailed classification against standard Softmax and Sigmoid losses and their class-balanced (CB) counterparts. For fair and reproducible comparison, we conduct a PyTorch (Paszke et al., 2019) reimplementation based on the official code of CB (Cui et al., 2019). We train CIFAR ResNet-32 for 200 epochs using SGD with momentum 0.9, weight decay 2×10^{-4}, batch size 128, and an initial learning rate of 0.1, with 5-epoch warmup and step decay at epochs

Table 1: Offline long-tail classification results on CIFAR-100 and CIFAR-10 under different IFs. Top-1 test accuracy (%) is reported. The better result between the sample- and class-level AoL is presented in the table. Superscript s denotes that the result is achieved at the sample-level; the results without the superscript are obtained at the class-level.

CIFAR-100				
Method	**IF-400**	**IF-100**	**IF-10**	**IF-1**
Softmax	31.50	40.10	57.15	71.59
Softmax + CB	32.04	40.40	57.48	71.59
Softmax + AoL	**32.43**	**40.51**	**57.98**	**71.67**
Sigmoid	**33.00**	40.65	57.63	71.33
Sigmoid + CB	32.86	40.81	**58.44**	71.33
Sigmoid + AoL	**33.00**	**40.84**	58.07	**71.56**

CIFAR-10				
Method	**IF-400**	**IF-100**	**IF-10**	**IF-1**
Softmax	60.32	72.20	86.72	92.98
Softmax + CB	61.31	73.59	86.74	92.98
Softmax + AoL	**63.94**^{s}	**73.74**^{s}	**86.80**	**93.19**
Sigmoid	60.17	71.85	86.88	93.00
Sigmoid + CB	64.45	**74.96**	**87.28**	93.00
Sigmoid + AoL	**66.62**	72.76	87.27	**93.39**

160 and 180. All CIFAR-based experiments run on an NVIDIA TITAN X (Pascal) GPU. Training uses random crop and horizontal flip, while AoL is computed on the corresponding unaugmented training-evaluation set to avoid augmentation noise in its updates. Following the long-tailed CIFAR protocol, each imbalanced training set is constructed using the exponential rule in Section 4.1. To isolate temporal persistence from CB weighting, the main experiments evaluate AoL independently of CB and focus on loss reweighting, where AoL adjusts each training example's contribution to the objective. Table 1 reports the selected AoL results, while additional results for both sample-level and class-level AoL are provided in Appendix B.1.

Overall, AoL generally improves or matches the corresponding baselines across datasets and imbalance levels, with larger gains under severe long-tail settings. Under Softmax, AoL improves CIFAR-100 from 31.50% to 32.43% and CIFAR-10 from 60.32% to 63.94% at IF-400, while giving the strongest Softmax results at IF-100, IF-10, and IF-1 on both datasets. Under Sigmoid, AoL likewise improves or matches the baseline across all selected settings, including an increase from 60.17% to 66.62% on CIFAR-10 at IF-400. Its effectiveness at IF-1 further shows that the resulting training emphasis is not determined by class frequency alone.

4.3 Online Settings

We then consider online settings, where data arrive sequentially and the complete dataset is unavailable. We evaluate AoL under two scenarios: streaming learning with relatively stable long-tail distributions and dynamic streaming with evolving long-tail structures. In the first scenario, we construct a long-tailed CIFAR-10 training set and make only 5% of the samples accessible at each step. Following Section 3.3, previously observed samples are maintained in a fixed-capacity buffer using reservoir sampling. Let n be the number of samples observed so far and K the buffer capacity. Once the buffer is full, each new sample is inserted with probability $P_{\text{insert}} = K/n$ and, if inserted, replaces a uniformly sampled buffer element. In the second scenario, a long-tailed subset is dynamically sampled from a balanced source pool according to Equation (15). A random permutation π_t maps long-tail ranks to class identities, so the head-tail assignment evolves over time. Under Period-1, this assignment changes every step, whereas under Period-10 it remains fixed for 10 consecutive steps before being resampled. We use a learning rate of 0.1 and Softmax loss reweighting in all online CIFAR-related experiments. At each update, the threshold $\tau^{(t)}$ is set to the median accuracy of the classes represented in $\mathcal{S}^{(t)}$.

The single-seed results for the stable and Period-1 settings are reported in Table 4, Appendix B.1. Under restricted accessibility, AoL improves CIFAR-10 by 3.32 and 8.70 points for IF-200 and IF-100, respectively, while matching the baseline for IF-1. Under Period-1, AoL improves the results for IF-50 and IF-10 but provides no gain at IF-200, indicating that class-level AoL estimates can become unstable when the stream changes too rapidly and observations are sparse. Table 2 reports the three-seed Period-10 evaluation. On CIFAR-10, AoL improves the baseline in all nine seed and IF pairs, with mean gains of $+3.10 \pm 0.51$, $+1.42 \pm 1.15$, and $+0.92 \pm 0.24$ points for IF-200, IF-50, and IF-10, respectively. On CIFAR-100, AoL improves or matches the baseline across all tested seeds, with corresponding mean gains of $+2.31 \pm 0.89$, $+1.42 \pm 0.91$, and $+0.02 \pm 0.03$. These results show that when class-level structure persists over multiple steps, AoL gains are reproducible across random seeds rather than specific to a single run.

Table 2: Period-10 online AoL results on CIFAR-10 and CIFAR-100 over three seeds. Top-1 test accuracy (%) is reported; the summary rows show mean ± sample SD.

CIFAR-10

IF	Seed	Softmax	Softmax + AoL	Gain
200	42	77.85	**81.44**	+3.59
	1024	76.52	**79.10**	+2.58
	2026	75.69	**78.83**	+3.14
	Mean	76.69±1.09	**79.79±1.44**	+3.10±0.51
50	42	82.67	**85.36**	+2.69
	1024	84.11	**84.55**	+0.44
	2026	83.29	**84.42**	+1.13
	Mean	83.36±0.72	**84.78±0.51**	+1.42±1.15
10	42	86.93	**87.58**	+0.65
	1024	86.22	**87.25**	+1.03
	2026	86.42	**87.51**	+1.09
	Mean	86.52±0.37	**87.45±0.17**	+0.92±0.24

CIFAR-100

IF	Seed	Softmax	Softmax + AoL	Gain
200	42	42.78	**44.18**	+1.40
	1024	42.61	**45.79**	+3.18
	2026	42.97	**45.31**	+2.34
	Mean	42.79±0.18	**45.09±0.83**	+2.31±0.89
50	42	49.61	**50.65**	+1.04
	1024	48.73	**51.19**	+2.46
	2026	49.95	**50.72**	+0.77
	Mean	49.43±0.63	**50.85±0.29**	+1.42±0.91
10	42	**57.31**	**57.31**	0.00
	1024	**56.75**	**56.75**	0.00
	2026	56.63	**56.69**	+0.06
	Mean	56.90±0.36	**56.92±0.34**	+0.02±0.03

Large-scale evaluation. To further evaluate the performance of AoL on larger-scale datasets, we evaluate Places365 (Zhou et al., 2018) and ImageNet-1k (Russakovsky et al., 2015) under the same Period-10 protocol. For both datasets, we use the original full training set at 224×224 resolution as the source pool, ResNet-34 with seed 42 on four NVIDIA GH200 Grace Hopper Superchips, a learning rate of 0.01, and a buffer capacity of $500{,}000$, while all other experimental settings remain unchanged. At IF-200, IF-50, and IF-10, the Softmax baseline/AoL accuracies are $37.23/39.79\%$, $42.31/43.79\%$, and $46.95/47.24\%$ on Places365, yielding gains of $+2.56$, $+1.48$, and $+0.29$ points, respectively. On ImageNet-1k, the corresponding accuracies are $46.74/48.31\%$, $54.46/54.81\%$, and $61.30/61.30\%$, with gains of $+1.57$, $+0.35$, and 0.00 points. Across both datasets, the gains are larger under more severe imbalance; this is consistent with the trend shown in Table 2.

5 Limitations and Future Work

The current AoL formulation intentionally uses binary correctness to isolate error duration and therefore does not capture error severity, confidence, or semantic type. Consequently, persistently mislabeled, ambiguous, or out-of-distribution samples may accumulate large AoL even when emphasizing them does not improve generalization. Class-level aggregation may reduce sensitivity to isolated noise but does not eliminate systematic noise. The online formulation also relies on sufficient temporal consistency: when class-level structure changes too rapidly or observations are sparse, as in the severe Period-1 setting, the accumulated AoL state can become less informative. Our evaluation is intentionally limited to the controlled Softmax/Sigmoid baselines to isolate temporal persistence; otherwise, additional mechanisms and hyperparameters would confound the attribution of AoL. See Appendix C for a fuller discussion of the evaluation scope. Natural extensions include confidence- or uncertainty-aware AoL, robust variants for noisy labels, adaptive temporal aggregation, and task-specific learning-state definitions for structured or generative models.

6 Conclusion

We introduced Age of Learning (AoL), a temporal learning-state variable that captures how long prediction errors remain unresolved during training. Unlike instantaneous signals such as loss, margin, and confidence, AoL represents error duration and provides a complementary view of learning difficulty. We developed AoL formulations for both offline and streaming settings, including sample-level and class-level states and lightweight online updates under restricted historical access. Across long-tailed classification experiments, AoL generally improves or matches the corresponding baselines, with reproducible gains in streaming settings where class-level learning difficulty persists over multiple updates. Analysis of class frequency, loss, and margin further shows that AoL is related to conventional difficulty signals but is not determined by them. These results support temporal error persistence as a useful additional signal for characterizing and guiding learning dynamics. Rather than replacing existing measures of difficulty, AoL provides a simple yet effective way to incorporate the history of unresolved prediction errors into adaptive training.

AI USE STATEMENT

Generative AI tools were used for language editing, rewriting, presentation refinement, and limited coding and debugging assistance, including LaTeX. They were not used for research methodology or experiment design, data generation or analysis, mathematical claims or proofs, or result interpretation. All AI-assisted work was strictly reviewed and verified by the authors. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmark datasets and does not involve human subjects, private or personally identifiable data, or the release of new datasets. We are not aware of additional ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We provide the algorithmic definitions, implementation details, hyperparameters, data construction procedures, random seeds, and additional experimental results in the main paper and appendix. An anonymized implementation is provided as supplementary material to facilitate reproduction of the reported results. All code will be released.

REFERENCES

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e562cd9c0768d5464b64cf61da7fc6bb-Paper.pdf.

Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6897–6907, June 2022.

Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/621461af90cadfdaf0e8d4cc25129f91-Paper.pdf.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In Adrien Bartoli and Andrea Fusiello (eds.), *Computer Vision – ECCV 2020 Workshops*, pp. 95–110, Cham, 2020. Springer International Publishing. ISBN 978-3-030-65414-6.

Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SkZUo6Xg61.

Botos Csaba, Wenxuan Zhang, Matthias Müller, Ser-Nam Lim, Mohamed Elhoseiny, Philip Torr, and Adel Bibi. Label delay in online continual learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119976–120012. Curran Associates, Inc., 2024. doi: 10.52202/079017-3813. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d8f5f134febb4bd74d8f79e338de382c-Paper-Conference.pdf.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Amit Daniely, Idan Mehalel, and Elchanan Mossel. Online learning of neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oedSIxYWql.

Enrico Fini, Stéphane Lathuilière, Enver Sangineto, Moin Nabi, and Elisa Ricci. Online continual learning under extreme memory constraints. In *Computer Vision – ECCV 2020*, pp. 720–735, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58604-1.

Genevieve E Flaspohler, Francesco Orabona, Judah Cohen, Soukayna Mouatadid, Miruna Oprescu, Paulo Orenstein, and Lester Mackey. Online learning with optimism and delay. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3363–3373. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/flaspohler21a.html.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Dandan Guo, Zhuo Li, Meixi Zheng, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Learning to re-weight examples with optimal transport for imbalanced classification. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25517–25530. Curran Associates, Inc., 2022. doi: 10.52202/068431-1850. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a39a9aceda771cded859ae7560530e09-Paper-Conference.pdf.

Guy Hacohen and Tinne Tuytelaars. Predicting the susceptibility of examples to catastrophic forgetting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21546–21569. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hacohen25a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Chenhan Jin, Shengze Xu, Qingsong Wang, Fan JIA, Dingshuo Chen, and Tieyong Zeng. OrderDP: A theoretically guaranteed lossless dynamic data pruning framework. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=e77QyyRQPz.

Igor Kadota, Abhishek Sinha, Elif Uysal-Biyikoglu, Rahul Singh, and Eytan Modiano. Scheduling policies for minimizing age of information in broadcast wireless networks. *IEEE/ACM Transactions on Networking*, 26(6):2637–2650, 2018. doi: 10.1109/TNET.2018.2873606.

Igor Kadota, Abhishek Sinha, and Eytan Modiano. Scheduling algorithms for optimizing age of information in wireless networks with throughput constraints. *IEEE/ACM Transactions on Networking*, 27(4):1359–1372, 2019. doi: 10.1109/TNET.2019.2918736.

Clement Kam, Sastry Kompella, Gam D. Nguyen, and Anthony Ephremides. Effect of message transmission path diversity on status age. *IEEE Transactions on Information Theory*, 62(3):1360–1374, 2016. doi: 10.1109/TIT.2015.2511791.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Angelos Katharopoulos and Francois Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2525–2534. PMLR, 2018. URL https://proceedings.mlr.press/v80/katharopoulos18a.html.

Sanjit Kaul, Roy Yates, and Marco Gruteser. Real-time status: How often should one update? In *2012 Proceedings IEEE INFOCOM*, pp. 2731–2735, 2012. doi: 10.1109/INFCOM.2012.6195689.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Frederik Kunstner, Alan Milligan, Robin Yadav, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30106–30148. Curran Associates, Inc., 2024. doi: 10.52202/079017-0948. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/350e718ff74062b4bac2c6ffd9e1ac66-Paper-Conference.pdf.

Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-SAM: Focal sharpness-aware minimization for long-tailed classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36624–36651. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25dk.html.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Mingyang Liu, Yongshan Chen, Zhiyuan Fan, Gabriele Farina, Asuman E. Ozdaglar, and Kaiqing Zhang. Online learning and equilibrium computation with ranking feedback. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lg6H2oJPky.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf.

Ali Maatouk, Saad Kriouile, Mohamad Assaad, and Anthony Ephremides. The age of incorrect information: A new performance metric for status updates. *IEEE/ACM Transactions on Networking*, 28(5):2215–2228, 2020. doi: 10.1109/TNET.2020.3005549.

Claire Monteleoni and Tommi Jaakkola. Online learning of non-stationary sequences. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/feecee9f1643651799ede2740927317a-Paper.pdf.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ac56f8fe9eea3e4a365f29f0f1957c55-Paper.pdf.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4334–4343. PMLR, 2018. URL https://proceedings.mlr.press/v80/ren18a.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Emanuel Sanchez Aimar, Nathaniel Helgesen, Yonghao Xu, Marco Kuhlmann, and Michael Felsberg. Flexible distribution alignment: Towards long-tailed semi-supervised learning with proper calibration. In *Computer Vision – ECCV 2024*, pp. 307–327, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72949-2.

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *Advances in Neural Information Processing Systems*, volume 36, pp. 75669–75687. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eeffa70bcbbd43f6bd067edebc6595e8-Paper-Conference.pdf.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/e58cc5ca94270acaceed13bc82dfedf7-Paper.pdf.

Daouda Sow, Herbert Woisetschläger, Saikiran Bulusu, Shiqiang Wang, Hans Arno Jacobsen, and Yingbin Liang. Dynamic loss-based sample reweighting for improved large language model pretraining. In *International Conference on Learning Representations*, volume 2025, pp. 89490–89520, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ded26b348d55953a4863d41540b7d5c4-Paper-Conference.pdf.

Karthik Sridharan and Seung Won Wilson Yoo. Online learning with unknown constraints. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=nqQcAhXGSy.

Yin Sun, Elif Uysal-Biyikoglu, Roy D. Yates, C. Emre Koksal, and Ness B. Shroff. Update or wait: How to keep your data fresh. *IEEE Transactions on Information Theory*, 63(11):7492–7508, 2017. doi: 10.1109/TIT.2017.2735804.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Honglian Wang, Sijing Tu, Lutz Oettershagen, and Aristides Gionis. Streaming stochastic submodular maximization with on-demand user requests. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=pjfMwXm61w.

Skyler Wu, Fred Lu, Edward Raff, and James Holt. Stabilizing linear passive-aggressive online learning with weighted reservoir sampling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FNOBf6JM7r.

Roy D. Yates, Yin Sun, D. Richard Brown, Sanjit K. Kaul, Eytan Modiano, and Sennur Ulukus. Age of information: An introduction and survey. *IEEE Journal on Selected Areas in Communications*, 39(5):1183–1210, 2021. doi: 10.1109/JSAC.2021.3065072.

Mingyang Yi, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Zhi-Ming Ma. Reweighting augmented samples by minimizing the maximal expected loss. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=9G5MIc-goqB.

Xing Zhang, Zuxuan Wu, Zejia Weng, Huazhu Fu, Jingjing Chen, Yu-Gang Jiang, and Larry S. Davis. Videolt: Large-scale long-tailed video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7960–7969, October 2021.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023. doi: 10.1109/TPAMI.2023.3268118.

Yu-Jie Zhang, Peng Zhao, and Masashi Sugiyama. Non-stationary online learning for curved losses: Improved dynamic regret via mixability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=TeHF8YjJaw.

Guodong Zheng, Peng Wang, Shengchao Hu, Quan Zheng, and Li Shen. Understanding the dynamics of forgetting and generalization in continual learning via the neural tangent kernel. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NE2yIxdo1w.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009.

The appendix provides algorithms, additional results, evaluation details, and visualizations of AoL.

A ALGORITHMS

In this section, we provide the algorithmic details of AoL in both offline and online settings. Algorithm 1 summarizes the offline procedure described in Section 3. After each training epoch, the model is evaluated on the unaugmented training set, and the sample-level AoL states are updated according to prediction correctness. These states are then aggregated using EMA and converted into normalized sample-level or class-level weights. The resulting weights are used in the following epoch for either loss reweighting or weighted resampling. Algorithm 2 summarizes the corresponding streaming-setting procedure, where class-level learning states are estimated from the current observation set. For each observed class, its accuracy determines whether the current AoL state increases or resets, while the states of unobserved classes remain unchanged. The resulting class-level AoL states are aggregated using the same EMA principle as in the offline setting, normalized across classes, and used for online loss reweighting. The specific memory policy is independent of the AoL update and determines only the observation set available at each streaming step.

Algorithm 1 Offline AoL

Require: Training set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, model f_θ, number of epochs E, EMA factor ρ, weighting level $q \in \{\text{SAMPLE}, \text{CLASS}\}$

Ensure: AoL-based sample weights $\{w_i^{(e)}\}_{i=1}^{N}$

Initialize $a_i^{(0)} \leftarrow 1$ and $w_i^{(0)} \leftarrow 1$ for all $i \in \{1, \dots, N\}$
for $e = 1$ to E **do**
  Train one epoch using $\{w_i^{(e-1)}\}_{i=1}^{N}$
  Evaluate $f_{\theta^{(e)}}$ on the unaugmented training set
  for each sample $i = 1, \dots, N$ **do**
    $\hat{y}_i^{(e)} \leftarrow \arg\max_{k \in \{1,\dots,C\}} f_{\theta^{(e)}}(\boldsymbol{x}_i)_k$
    $c_i^{(e)} \leftarrow \mathbb{I}[\hat{y}_i^{(e)} = y_i]$
    if $c_i^{(e)} = 1$ **then**
      $a_i^{(e)} \leftarrow 1$
    else
      $a_i^{(e)} \leftarrow a_i^{(e-1)} + 1$
    end if
    if $e = 1$ **then**
      $\bar{a}_i^{(e)} \leftarrow a_i^{(e)}$
    else
      $\bar{a}_i^{(e)} \leftarrow (1-\rho)\bar{a}_i^{(e-1)} + \rho a_i^{(e)}$
    end if
  end for
  if $q = \text{SAMPLE}$ **then**
    For each sample $i = 1, \dots, N$, set $w_i^{(e)} \leftarrow \dfrac{\bar{a}_i^{(e)}}{\frac{1}{N}\sum_{j=1}^{N} \bar{a}_j^{(e)}}$
  else
    for each class $c = 1, \dots, C$ **do**
      $\bar{a}_c^{(e)} \leftarrow \dfrac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} \bar{a}_i^{(e)}$
    end for
    For each class $c = 1, \dots, C$, set $w_c^{(e)} \leftarrow \dfrac{\bar{a}_c^{(e)}}{\frac{1}{C}\sum_{k=1}^{C} \bar{a}_k^{(e)}}$
    $w_i^{(e)} \leftarrow w_{y_i}^{(e)}$, for all i
  end if
  Use $\{w_i^{(e)}\}$ in epoch $e+1$ for loss reweighting, or set $p_i^{(e+1)} \propto w_i^{(e)}$ for weighted random sampling
end for

Algorithm 2 Online AoL

Require: Stream $\{\mathcal{B}^{(t)}\}_{t=1}^{T}$, model f_θ, EMA factor ρ, optional memory buffer $\mathcal{M}^{(t)}$
Ensure: Class weights $\{w_c^{(t)}\}_{c=1}^{C}$

Initialize $a_c^{(0)} \leftarrow 1$ and $w_c^{(0)} \leftarrow 1$ for all $c \in \{1, \dots, C\}$
for $t = 1$ to T **do**
  Update the optional buffer $\mathcal{M}^{(t)}$ using the chosen memory policy
  If a buffer is used, set $\mathcal{S}^{(t)} \leftarrow \mathcal{M}^{(t)}$; otherwise, set $\mathcal{S}^{(t)} \leftarrow \mathcal{B}^{(t)}$
  Evaluate the current model on $\mathcal{S}^{(t)}$ and compute $r_c^{(t)}$ according to Eq. (13)
  Set $\tau^{(t)} \leftarrow \mathrm{median}\{r_c^{(t)} : c \text{ is represented in } \mathcal{S}^{(t)}\}$
  for each class c represented in $\mathcal{S}^{(t)}$ **do**
    if $r_c^{(t)} < \tau^{(t)}$ **then**
      $a_c^{(t)} \leftarrow a_c^{(t-1)} + 1$
    else
      $a_c^{(t)} \leftarrow 1$
    end if
  end for
  Keep $a_c^{(t)} \leftarrow a_c^{(t-1)}$ for classes not represented in $\mathcal{S}^{(t)}$
  Apply the EMA aggregation in Eqs. (4–5) to the class-level states $\{a_c^{(t)}\}$, yielding $\{\bar{a}_c^{(t)}\}$
  For each class $c = 1, \dots, C$, set $w_c^{(t)} \leftarrow \dfrac{\bar{a}_c^{(t)}}{\frac{1}{C}\sum_{k=1}^{C} \bar{a}_k^{(t)}}$
  Assign each observed sample the weight of its ground-truth class for online loss reweighting
end for

B Additional Experimental Details

B.1 Offline and Online Settings

In this subsection, we provide additional offline and online results. Table 3 presents additional offline results corresponding to Table 1. Table 3 reports additional offline results on CIFAR-100 and CIFAR-10 under different imbalance factors (IFs), including both loss reweighting and weighted resampling. For each setting, we evaluate the corresponding CB baselines together with the sample-level and class-level AoL variants and report the better AoL result, using the class-level result when their accuracies are equal. Table 4 reports additional single-seed results under both the access-constrained and evolving long-tail streaming settings. These CIFAR-10 experiments use ResNet-32, class-level AoL, a fixed learning rate of 0.1, and the same buffer and optimization protocols described in Section 4.3. In the access-constrained setting, the stream exposes only 5% of the long-tailed training samples at each epoch, while in the dynamic setting the long-tail structure is resampled over time from a balanced IF-1 source pool. These results further show that AoL benefits from temporally persistent class-level difficulty and sufficiently reliable buffer statistics.

B.2 Arrival Imbalance

In this subsection, we study a complementary offline setting that isolates arrival imbalance from class-frequency imbalance. Standard epoch-wise training with shuffle already induces a mild form of natural arrival sparsity, since samples from all classes are not guaranteed to appear uniformly in every batch. To amplify this effect, we introduce an arrival loader that, at the beginning of each epoch, uniformly retains a fraction $p \in \{1, 2, 5, 10, 20, 50, 100\}\%$ of the training samples. This construction induces a more temporally sparse condition: although the dataset remains class-balanced, some samples may not appear for several consecutive epochs. The retained subset of each epoch is the only data used for optimization and AoL updates in that epoch. We evaluate this setting on balanced CIFAR-100 and balanced CIFAR-10 using loss reweighting only, with the same ResNet-32 architecture, SGD optimizer, batch size, learning-rate schedule, momentum, and weight decay as in the previous offline experiments. For this arrival-specific extension of AoL, the age update is computed only from the samples that actually enter the current epoch training subset;

Table 3: Offline results on CIFAR-100 and CIFAR-10 with varying imbalance factors (IFs). Top-1 test accuracy (%) is reported. Panel (a) evaluates loss reweighting, and panel (b) evaluates weighted resampling. The better result between the sample- and class-level AoL is presented in the table; the class-level result is reported when they are the same. Superscript s denotes that the result is achieved at the sample-level; the results without the superscript are obtained at the class-level. In addition, β is the CB hyperparameter and ρ is the AoL factor. CB† denotes the CB variant with $\beta = 0.0$, and AoL† denotes the AoL variant with $\rho = 0.0$.

(a) Loss reweighting

CIFAR-100

Method	IF-400	IF-300	IF-200	IF-100	IF-50	IF-20	IF-10	IF-1
Softmax	31.50	**33.46**	35.77	40.10	44.85	51.26	57.15	71.59
Softmax + CB	32.04	32.54	35.83	40.40	**45.38**	**52.38**	57.48	71.59
β	0.9	0.9	0.9	0.9	0.9	0.99	0.99	–
Softmax + AoL	**32.43**	**33.46**	**36.08**	**40.51**	44.90	51.87	**57.98**	**71.67**
ρ	0.3	0.0	0.1	0.8	0.5	0.3	0.9	0.2
Sigmoid	**33.00**	34.34	36.21	40.65	45.33	52.21	57.63	71.33
Sigmoid + CB	32.86	34.13	36.67	40.81	45.03	**52.98**	**58.44**	71.33
β	0.9	0.9	0.9	0.9	0.9	0.99	0.99	–
Sigmoid + AoL	**33.00**	**34.50**^{s}	**36.70**^{s}	**40.84**	**45.41**	52.42	58.07	**71.56**
ρ	0.0	0.3^{s}	0.1^{s}	0.1	1.0	1.0	0.7	0.2

CIFAR-10

Method	IF-400	IF-300	IF-200	IF-100	IF-50	IF-20	IF-10	IF-1
Softmax	60.32	65.70	67.94	72.20	77.03	82.78	86.72	92.98
Softmax + CB	61.31	65.21	**69.56**	73.59	**78.14**	**84.18**	86.74	92.98
β	0.99	0.9	0.99	0.9999	0.999	0.999	0.99	–
Softmax + AoL	**63.94**^{s}	**66.80**	67.94	**73.74**^{s}	77.89	83.40	**86.80**	**93.19**
ρ	0.6^{s}	0.2	0.0	0.8^{s}	0.5	0.4	1.0	0.6
Sigmoid	60.17	62.84	65.76	71.85	77.40	**83.92**	86.88	93.00
Sigmoid + CB	64.45	**67.67**	**70.42**	**74.96**	**79.69**	**83.92**	**87.28**	93.00
β	0.999	0.99	0.9	0.9999	0.9999	0.9	0.9999	–
Sigmoid + AoL	**66.62**	67.46	68.17	72.76	78.50	**83.92**	87.27	**93.39**
ρ	0.2	1.0	0.2	0.2	0.9	0.0	0.8	0.8

(b) Weighted resampling

CIFAR-100

Method	IF-400	IF-300	IF-200	IF-100	IF-50	IF-20	IF-10	IF-1
Softmax	31.50	**33.46**	35.77	40.10	44.85	51.26	57.15	**71.59**
Softmax + CB†	31.78	33.18	**36.21**	**40.69**	**45.22**	51.80	57.28	70.70
Softmax + CB	31.50	33.13	35.33	40.27	44.17	51.89	57.17	70.70
β	0.9	0.9	0.9	0.9	0.9	0.9	0.9	–
Softmax + AoL†	31.78	33.18	**36.21**	**40.69**	**45.22**	51.80	57.28	70.70
Softmax + AoL	**32.26**	33.18	**36.21**	**40.69**	**45.22**	**52.13**	**57.55**	70.84
ρ	0.2	0.0	0.0	0.0	0.0	0.4	0.5	0.6
Sigmoid	**33.00**	**34.34**	36.21	40.65	**45.33**	52.21	57.63	**71.33**
Sigmoid + CB†	32.65	33.54	36.56	40.55	44.91	52.20	57.91	71.26
Sigmoid + CB	31.90	34.04	**36.84**	40.33	44.48	52.03	**58.08**	71.26
β	0.9	0.9	0.9	0.9	0.9	0.9	0.99	–
Sigmoid + AoL†	32.65	33.54	36.56	40.55	44.91	52.20	57.91	71.26
Sigmoid + AoL	32.65	33.96	36.56	**40.98**	45.32	**52.53**	58.02	71.26
ρ	0.0	0.2	0.0	0.6	0.8	0.9	0.4	0.0

CIFAR-10

Method	IF-400	IF-300	IF-200	IF-100	IF-50	IF-20	IF-10	IF-1
Softmax	60.32	65.70	67.94	72.20	77.03	82.78	86.72	92.98
Softmax + CB†	59.33	63.23	**69.20**	70.52	77.53	82.55	86.75	92.91
Softmax + CB	**63.17**	**68.25**	67.17	**73.13**	77.91	83.28	86.75	92.91
β	0.99	0.99	0.999	0.99	0.99	0.9999	0.9	–
Softmax + AoL†	59.33	63.23	**69.20**	70.52	77.53	82.55	86.75	92.91
Softmax + AoL	62.96	65.55^{s}	**69.20**	72.87^{s}	**78.80**	**83.75**	**86.77**	**93.16**
ρ	0.7	0.4^{s}	0.0	0.3^{s}	0.9	0.9	0.5	0.5
Sigmoid	60.17	62.84	65.76	71.85	77.40	**83.92**	86.88	93.00
Sigmoid + CB†	59.25	61.46	67.16	71.89	**78.47**	83.22	86.95	92.82
Sigmoid + CB	62.97	63.89	68.98	**73.30**	77.97	83.90	86.95	92.82
β	0.999	0.99	0.999	0.99	0.9	0.99	0.9, 0.9999	–
Sigmoid + AoL†	59.25	61.46	67.16	71.89	**78.47**	83.22	86.95	92.82
Sigmoid + AoL	**63.87**	**65.94**	**69.38**^{s}	72.58	**78.47**	83.90	**87.14**	**93.26**^{s}
ρ	0.5	0.5	0.7^{s}	0.7	0.0	0.3	0.9	0.5^{s}

Table 4: Online AoL results on CIFAR-10. Top-1 test accuracy (%) is reported. Panel (a) evaluates relatively stable long-tail streams where 5% of the training samples are randomly selected at each epoch; the large-buffer setting retains all observed samples, while the limited-buffer setting uses reservoir sampling with $K = 10000$. Panel (b) evaluates evolving long-tail streams generated from a balanced IF$= 1$ source pool with a limited buffer of $K = 50000$. AoL generally improves or matches Softmax, with larger gains under stronger imbalance and limited-buffer access. Period-10 also shows more consistent gains than Period-1.

(a) Online streams with 5% random selection at each epoch

	Large buffer (all observed)				Limited buffer $K = 10000$			
Method	**IF-200**	**IF-100**	**IF-50**	**IF-1**	**IF-200**	**IF-100**	**IF-50**	**IF-1**
Softmax	30.31	35.00	40.80	**80.76**	30.31	35.00	40.80	**80.76**
Softmax + AoL	**34.15**	**39.53**	**44.66**	**80.76**	**33.63**	**43.70**	**45.19**	**80.76**
ρ	0.2	0.1	0.1, 0.2	0.0	0.2	0.1	0.1	0.0

(b) Online dynamic long-tail streams with limited buffer size $K = 50000$

	Dynamic Long-Tail with Period-1					Dynamic Long-Tail with Period-10				
Method	**IF-200**	**IF-100**	**IF-50**	**IF-20**	**IF-10**	**IF-200**	**IF-100**	**IF-50**	**IF-20**	**IF-10**
Softmax	**54.03**	75.66	81.58	86.38	87.49	77.85	79.77	82.67	86.41	86.93
Softmax + AoL	**54.03**	**76.45**	**84.46**	**86.40**	**87.60**	**81.44**	**82.49**	**85.36**	**86.42**	**87.58**
ρ	0.0	0.1	0.2	0.4	0.6	0.1	0.5	0.4	0.9	0.8

Table 5: Offline arrival-imbalance ablation on balanced CIFAR-100 and CIFAR-10. Top-1 test accuracy (%) is reported. At the beginning of each epoch, an arrival loader uniformly retains a fraction $p \in \{1, 2, 5, 10, 20, 50, 100\}\%$ of the training set, while all other optimization settings remain identical to the standard offline setup. AoL generally improves or matches both Softmax and Sigmoid baselines across arrival fractions, with the largest gains under sparse arrivals.

(a) CIFAR-100 (Balanced)

Method	**1%**	**2%**	**5%**	**10%**	**20%**	**50%**	**100%**
Softmax	18.55	32.04	53.24	**58.67**	**66.74**	69.88	71.59
Softmax + AoL	**19.43**	**34.27**	**53.66**	**58.67**	**66.74**	**69.92**	**71.67**
ρ	0.4	0.4	0.7	0.0	0.0	0.6	0.2
Sigmoid	19.05	**35.55**	54.35	59.31	66.19	69.76	71.33
Sigmoid + AoL	**20.01**	**35.55**	**54.49**	**59.73**	**66.66**	**70.04**	**71.56**
ρ	0.4	0.0	1.0	0.9	0.7	0.2	0.2

(b) CIFAR-10 (Balanced)

Method	**1%**	**2%**	**5%**	**10%**	**20%**	**50%**	**100%**
Softmax	57.04	**75.11**	84.97	86.69	90.46	91.88	92.98
Softmax + AoL	**61.54**	**75.11**	**85.66**	**87.49**	**90.60**	**92.62**	**93.19**
ρ	0.9	0.0	0.2	0.1	0.3	0.7	0.6
Sigmoid	62.55	76.58	84.79	86.82	90.02	**92.42**	93.00
Sigmoid + AoL	**63.80**	**78.12**	**85.64**	**87.56**	**90.59**	**92.42**	**93.39**
ρ	0.4	1.0	0.5	1.0	0.6	0.0	0.8

samples that were previously observed but do not appear in the current epoch are incremented by one, while samples never observed remain NaN and are excluded from the aggregation.

Table 5(a) and Table 5(b) report balanced CIFAR-100 and CIFAR-10 results, respectively. The main purpose of this experiment is to examine whether AoL remains informative when class frequencies are unchanged but sample arrival becomes increasingly sparse over time. The results suggest that AoL is generally more helpful in the low-arrival regime, especially under the 1%–5% settings where supervision is temporally sparse. This may indicate that AoL captures not only class frequency but also the duration for which samples remain absent from optimization. As the arrival rate increases, the performance gap between AoL and the baseline generally becomes smaller, with several settings showing comparable performance. This suggests that the benefit of an explicit AoL signal may be most visible when temporal persistence of errors is a key limiting factor. Overall, the experiment suggests that AoL captures a complementary aspect of difficulty beyond static class balance, namely the persistence of under-exposure during training.

C Evaluation Scope and Experimental Design

Controlled baseline design. The primary goal of the experiments is to determine whether temporal error persistence provides a useful learning signal, rather than to rank AoL as a complete long-tailed recognition system. AoL is deliberately implemented as a lightweight training signal that can be attached to standard objectives. We therefore use standard Softmax and Sigmoid losses as unweighted references and CB Loss (Cui et al., 2019) as the principal frequency-based weighting reference. This comparison is particularly informative because CB and AoL modify the same part of the training pipeline, the emphasis assigned to training examples, while deriving that emphasis from fundamentally different information. Here, CB uses static class statistics through the effective number of samples, whereas AoL uses the temporal persistence of prediction errors. The backbone, base loss family, data construction, and optimization pipeline can therefore remain coherent, allowing the contribution of the learning signal itself to be examined with limited changes.

This design should not be interpreted as an attempt to establish a leaderboard ranking against state-of-the-art long-tailed recognition systems. Contemporary approaches may jointly modify representation learning, classifier training, logits or margins, sampling strategies, auxiliary objectives, or multi-stage optimization. *Comparing such complete systems would conflate the proposed temporal signal with several additional design choices and would answer a different question from the one studied here.* These methods are included in the related-work discussion to position AoL within the broader literature, whereas the empirical comparisons are intentionally structured to isolate the new learning signal. Similarly, loss reweighting and weighted resampling are used as controlled mechanisms for applying AoL rather than being presented as the methodological novelty themselves.

Hyperparameter selection and reporting. We follow the experimental convention of the original CB Loss study (Cui et al., 2019), in which hyperparameters are selected by cross-validation. For CB, we use the original candidate set $\beta \in \{0.9, 0.99, 0.999, 0.9999\}$. For AoL, we use the predefined grid $\rho \in \{0.0, 0.1, \ldots, 1.0\}$ under the same validation protocol. Sample-level and class-level AoL are treated as two predefined formulations rather than post-hoc modifications; the formulation selected by validation is reported, with the class-level formulation used in case of a tie. Test data are reserved for final accuracy reporting. To make the selection transparent, Table 3 reports the corresponding β and ρ values, and the superscript s explicitly identifies cases in which the sample-level AoL formulation performs better.

Experimental coverage. Although the principal comparison is intentionally controlled, the evaluation varies the factors that are directly relevant to the proposed temporal signal. The offline study covers CIFAR-10 and CIFAR-100, eight imbalance factors, Softmax and Sigmoid objectives, loss reweighting and weighted resampling, and both sample-level and class-level AoL formulations. The streaming study considers access-constrained streams, large and limited memory buffers, rapidly evolving Period-1 streams, more temporally persistent Period-10 streams, CIFAR-10 and CIFAR-100, and three random seeds in the main Period-10 evaluation. We further evaluate the same Period-10 formulation on the full Places365 and ImageNet-1k source pools and separately study balanced-data arrival sparsity and diagnostic AoL trajectories. These experiments are intended to test the proposed signal across imbalance severity, temporal persistence, observation constraints, dataset scale, and intervention mechanism rather than across unrelated system components.

Streaming scope and comparison protocol. The streaming experiments consider task-free classification over a fixed label space whose sample accessibility and class-frequency structure evolve over time. They are therefore distinct from task-incremental or class-incremental continual-learning protocols that introduce explicit task boundaries, new classes, delayed supervision, or additional task-specific assumptions. In the controlled streaming comparisons, AoL and the Softmax baseline use the same incoming data, memory policy, buffer capacity, architecture, optimizer, and training schedule; the difference is whether the class-level temporal state is used for loss reweighting. This matched design isolates the effect of AoL under the streaming conditions studied here. Specialized continual-learning systems remain compatible directions for incorporating AoL, but comparing complete systems with different replay, regularization, and task assumptions is outside the controlled question addressed in the present experiments.

Online computation and memory. An online AoL update is performed once after each training epoch in all reported experiments, rather than after every SGD mini-batch. At an AoL update, the current model is evaluated on the observation set $\mathcal{S}^{(t)}$ to estimate class-level accuracy and update the AoL states. This requires a forward-only evaluation and no additional backward pass, while the persistent AoL state itself requires only $O(C)$ memory beyond the data buffer. The computational cost therefore depends on both the observation set size and the chosen AoL evaluation frequency. The 500,000-sample buffer used in the large-scale Places365 and ImageNet-1k experiments is intended to evaluate scalability under a common large-scale Period-10 protocol, rather than to minimize memory demand. Memory-constrained behavior is examined separately through the limited-buffer CIFAR experiments, including $K = 10{,}000$ and $K = 50{,}000$.

Interpretation of "complementary." We use the term *complementary* to describe the information represented by AoL: class frequency describes the amount of available data, and loss or margin describes the current prediction state, whereas AoL describes how long an incorrect state has persisted. The correlation analysis in Section 4.1 directly examines this distinction. Complementarity in this sense does not imply that AoL must produce an additive gain when mechanically combined with every existing rebalancing method. Combining AoL with CB, for example, would require an additional design choice for composing temporal and frequency-based weights, together with corresponding normalization and hyperparameter selection. The main experiments therefore evaluate the signals separately so that the contribution of temporal persistence can be identified without introducing an additional combination rule.

Statistical scope of the results. The extensive offline tables are intended primarily as controlled mechanism studies across datasets, imbalance factors, base losses, weighting levels, and intervention mechanisms. Accordingly, small differences in single-run offline results are not interpreted as evidence of statistical significance. We use terms such as "matches" or "comparable" when the observed difference is negligible. Reproducibility across random initialization is examined explicitly in the three-seed Period-10 experiments in Table 2, where all individual seed results, means, and sample standard deviations are reported. This separates broad controlled coverage from the dedicated multi-seed robustness evaluation.

Large-scale evaluation and dataset scope. The Places365 and ImageNet-1k experiments use the original full training sets at 224×224 resolution as source pools and apply the same controlled Period-10 construction used in the streaming study. Their purpose is to test whether the AoL mechanism remains effective when the source dataset and memory buffer are substantially larger, without dataset-specific changes to the method. These experiments therefore provide evidence of computational and dataset-scale transfer under the studied controlled imbalance protocol; they are not presented as a state-of-the-art comparison on naturally occurring long-tailed benchmarks. Natural long-tail distributions constitute an additional evaluation axis that is orthogonal to the present controlled study.

Diagnostic visualizations. The heatmaps in Section D are intentionally generated from standard Softmax training in which AoL is logged but not fed back into optimization. This isolates the learning-state signal itself: an AoL-guided model would change the trajectory being diagnosed and would therefore mix the existence of the temporal pattern with the effect of the intervention. The CIFAR-10 visualization uses a power-law color normalization with $\gamma = 0.35$, $v_{\min} = 1$, and $v_{\max} = 100$, while the CIFAR-100 visualization uses $\gamma = 0.5$ with the same bounds. Raw AoL values are not numerically clipped before visualization; values above 100 simply saturate at the upper end of the displayed color scale. The nonlinear normalization is used only to preserve visual contrast among low- and medium-AoL regions and has no effect on any quantitative result.

Reproducibility. The supplementary material contains the code used for the reported offline and streaming AoL updates, long-tail and streaming data construction, hyperparameter configurations, and memory-buffer handling. Together with the reported architectures, optimization schedules, random seeds, selected β and ρ values, buffer capacities, and data-construction rules, this is intended to make the controlled comparisons independently reproducible. All code will be released.

D DIAGNOSTIC VISUALIZATIONS

To better understand the training dynamics captured by AoL, we visualize the aggregated class-level AoL trajectories on CIFAR-10 and CIFAR-100 under the long-tailed setting with imbalance factor 200. These visualizations are purely diagnostic: AoL is computed and logged from the training process, but it is not fed back into the optimization objective, nor is it used as a weighting term or sampling probability during training. We use a nonlinear color normalization in the heatmaps so that both low- and medium-AoL regions remain visually distinguishable.

Figure 3 shows the CIFAR-10 result. The head classes remain relatively dark and stable throughout training, whereas the tail classes exhibit noticeably larger AoL values in the early stage and more visible fluctuations over time. Although AoL is not used to guide optimization in this diagnostic experiment, the overall AoL level decreases as training progresses, suggesting that persistent error episodes become shorter over training. In the later stage of training, especially after approximately 160 epochs, the heatmap becomes visibly darker, and the separation between earlier and later training phases becomes more pronounced. This transition is consistent with the shift to a lower learning rate, although the exact boundary should be interpreted as a qualitative observation rather than a strict phase transition.

Figure 4 presents the corresponding CIFAR-100 visualization. Compared with CIFAR-10, CIFAR-100 exhibits a more pronounced head and tail difference: head classes generally stabilize earlier, while many tail classes retain higher AoL for longer. Several tail classes still show intermittent bright bands in the middle and late stages, indicating that their predictions remain less stable than those of head classes. Nevertheless, the global trend is still downward, and the latter part of training is substantially darker than the early part. This suggests that AoL captures a meaningful notion of learning persistence even when it is not explicitly used for optimization. That is, AoL can reveal class-dependent learning dynamics that are not fully reflected by instantaneous accuracy alone.

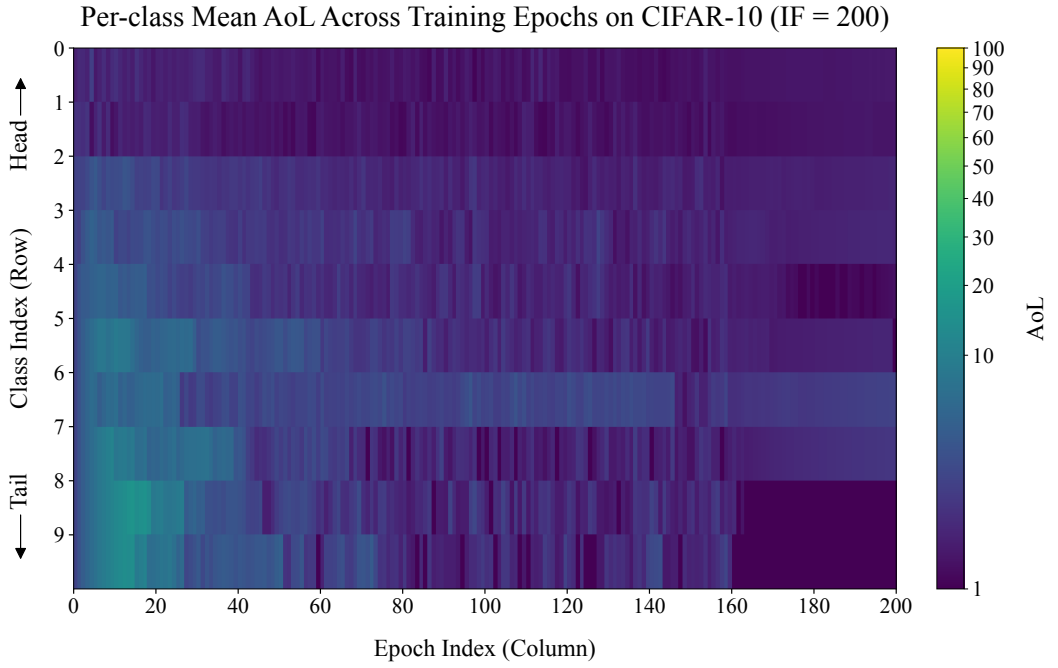


Figure 3: Per-class mean AoL across training epochs on CIFAR-10 (IF = 200) in the Softmax setting. The head classes are shown at the top and the tail classes at the bottom. AoL gradually decreases over training, and the later stage becomes noticeably darker, indicating reduced prediction-error persistence as optimization proceeds.

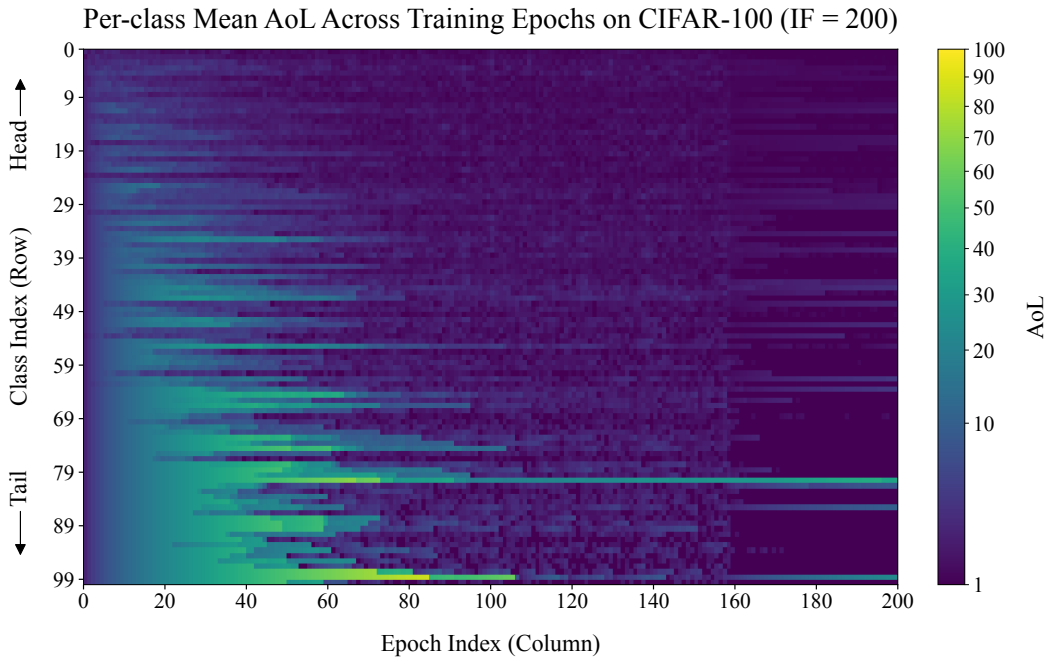


Figure 4: Per-class mean AoL across training epochs on CIFAR-100 (IF = 200) in the Softmax setting. Compared with CIFAR-10, the head-tail gap is more pronounced, and tail classes exhibit more persistent high-AoL regions, especially in the early and middle stages of training. The late-stage darkening is consistent with the transition to a lower learning rate.