NEURIPS 2026 · MAIN CONFERENCE

A Theoretical Bridge Between Long-Tailed Recognition and Continual Learning

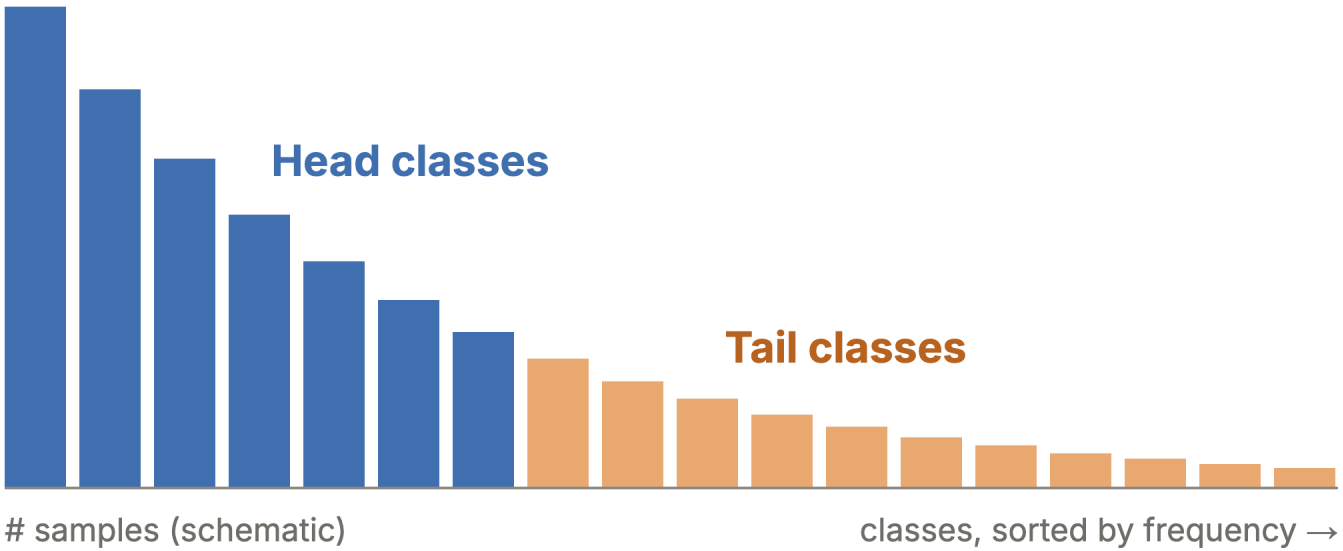
Long-tailed recognition admits a continual-learning formulation: learn the **Head**, then learn the **Tail** while preventing forgetting.

Mahdiyar Molahasani · Michael Greenspan · Ali Etemad

Electrical and Computer Engineering, Queen's University, Canada · m.molahasani@queensu.ca · mahdiyar.ai

1 THE PROBLEM

Head classes dominate the gradients.



A few **Head** classes hold most samples; many **Tail** classes have very few.

IMBALANCE FACTOR

IF = largest class ÷ smallest class

Training drifts toward the Head.

Multi-stage LTR training already looks like sequential learning.

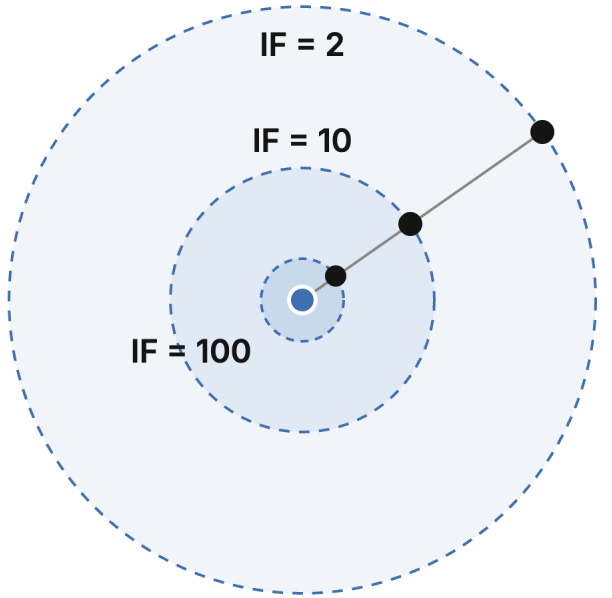
Can continual learning be harnessed to solve long-tailed recognition?

Formalized in Lean 4

Long-tailed training lands within O(1/√IF) of the Head-only solution.

$$\|\theta^* - \theta^*_H\| = \mathcal{O}\left(\frac{1}{\sqrt{\mathrm{IF}}}\right)$$

The model trained on all data ends up close to the model trained on the **Head** alone. The more imbalanced the data, the closer it lands.



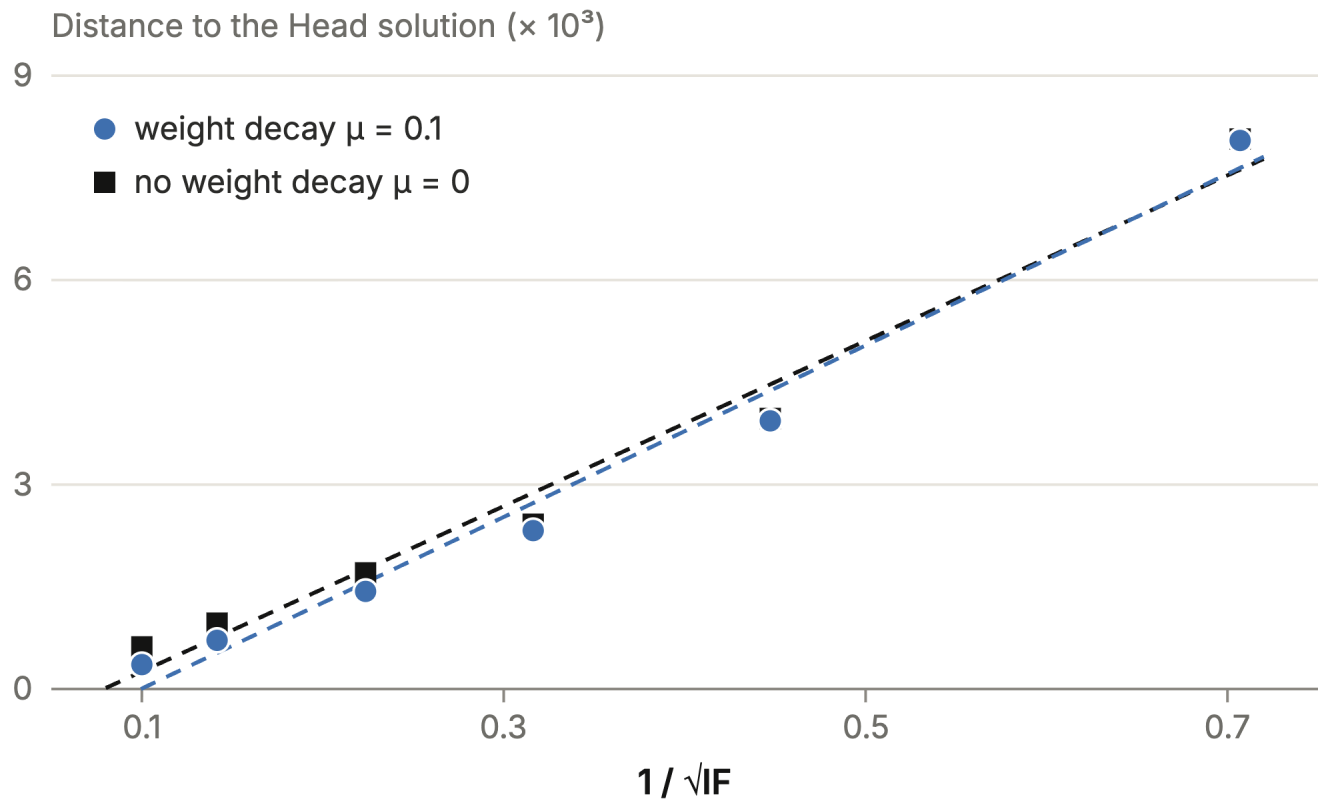
Convex + L²
Theorem 3.3

Deep nets via KŁ
Theorem 3.5

n partitions
Theorem 3.7

2 THEORY · EMPIRICAL CHECK

Verified: the distance is linear in 1/√IF.



CIFAR-100 · RESNET-18

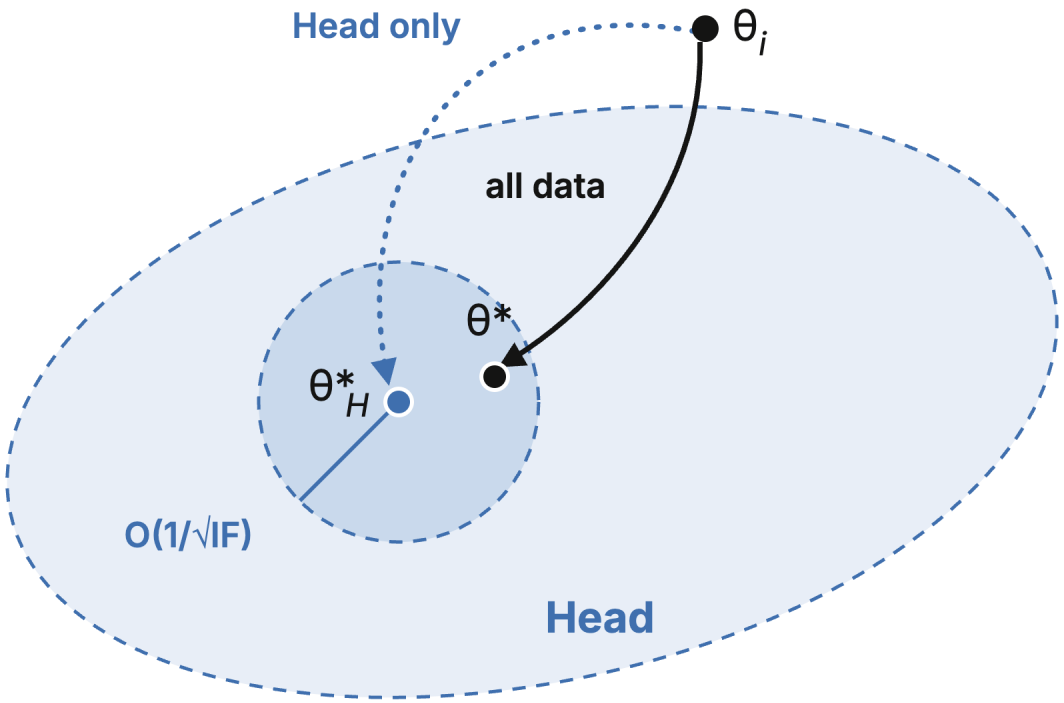
$R^2 = 0.98$

Linear fit (dashed) on a **non-convex** deep network, with and without weight decay.

Also linear on convex MNIST-LT: **$R^2 \geq 0.97$**.

3 THE BRIDGE

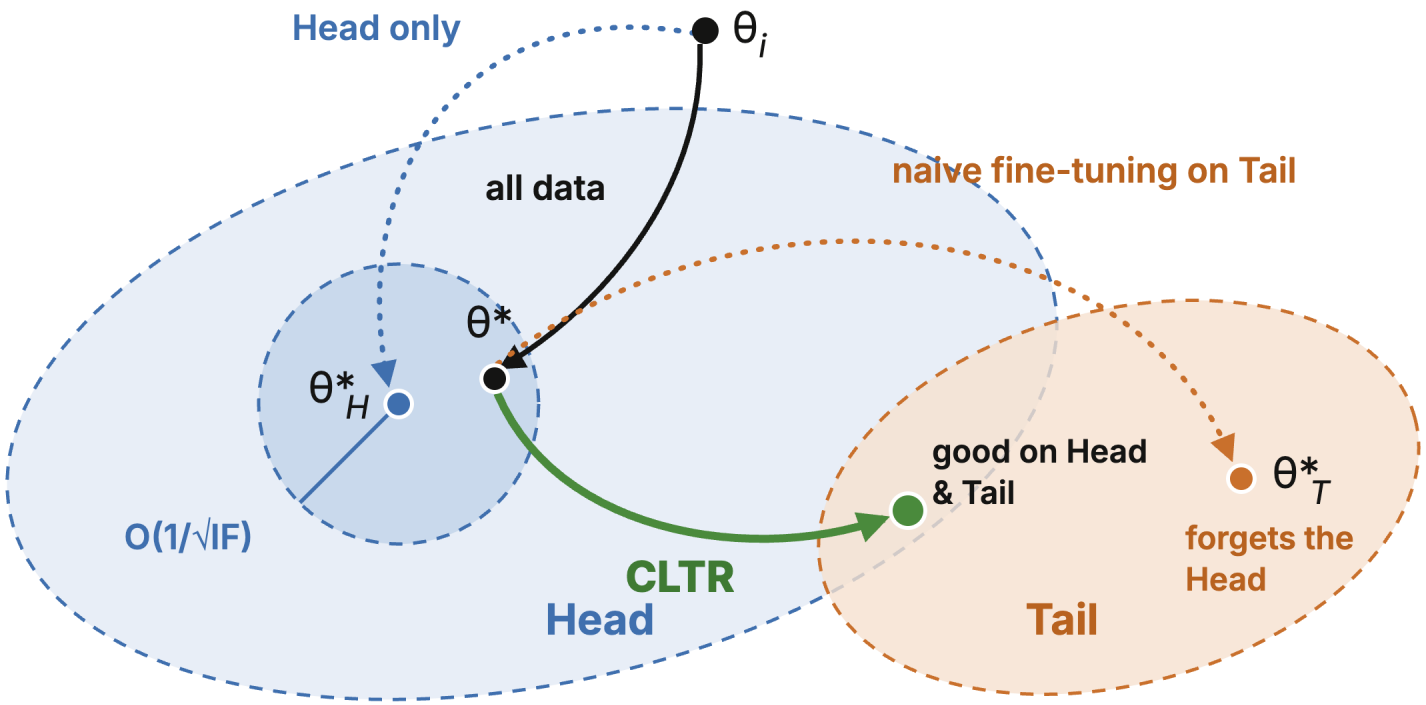
Training on all data already lands near the Head solution.



Imbalance pulls the full-data solution into the Head's basin. The Head is, in effect, already learned.

- θ^* trained on all data
- Head-only solution

Learn the Head first. Then learn the Tail without forgetting.



The rest is a **continual-learning** problem: Tail classes arrive as new tasks, and the Head must not be forgotten.

Proposition 3.9

Sort classes into partitions by size, each much larger than the next. The LTR objective can then be written as the objective of a continual-learning stream: partition 1 → 2 → … → N.

3 THE CLTR ALGORITHM

Three steps, with any off-the-shelf CL method.

1

Partition

Sort classes by count. Split into N groups with equal local imbalance, IF^(1/N) (Prop. 3.10).

2

Learn the Head

Train on the largest partition to reach the Head solution.

3

Add the Tail

Learn the remaining partitions as new tasks with a **standard continual-learning method**.

EVALUATED WITH

LwF · EWC · Modified EWC · GPM · SGP · FOSTER · TPL · DualPrompt

N = 2 for forgetting-sensitive methods (LwF, EWC, Modified EWC, GPM, SGP); N = 4 for FOSTER and TPL.

4 GUARANTEE

The CL objective upper-bounds the balanced loss.

$$\mathcal{L}_{\text{bal}}(\theta) \;\le\; \mathcal{L}_{\text{CL}}(\theta) + \frac{1}{2}\,\mathcal{L}_H(\theta_H^*)$$

$$\mathcal{L}_{\text{bal}} = \tfrac{1}{2}\mathcal{L}_H + \tfrac{1}{2}\mathcal{L}_T$$

$$\mathcal{L}_{\text{CL}} = \mathcal{L}_T + \alpha\, D_\Phi\big(h_\theta, z\big) + \beta\, D_\psi\big(\theta, \theta_H^*\big)$$

Holds near the Head solution if **β or α clears a threshold** (Assump. 3.11, Thm. 3.12).

Measured thresholds: **β ≈ 0.1**, **α ≈ 10^{-4}**.

BOUND CHECKED AT 33 CHECKPOINTS · CIFAR100-LT · APP. F.5

33/33

CLTR (LwF)

holds everywhere

33/33

CLTR (EWC)

holds everywhere

0/33

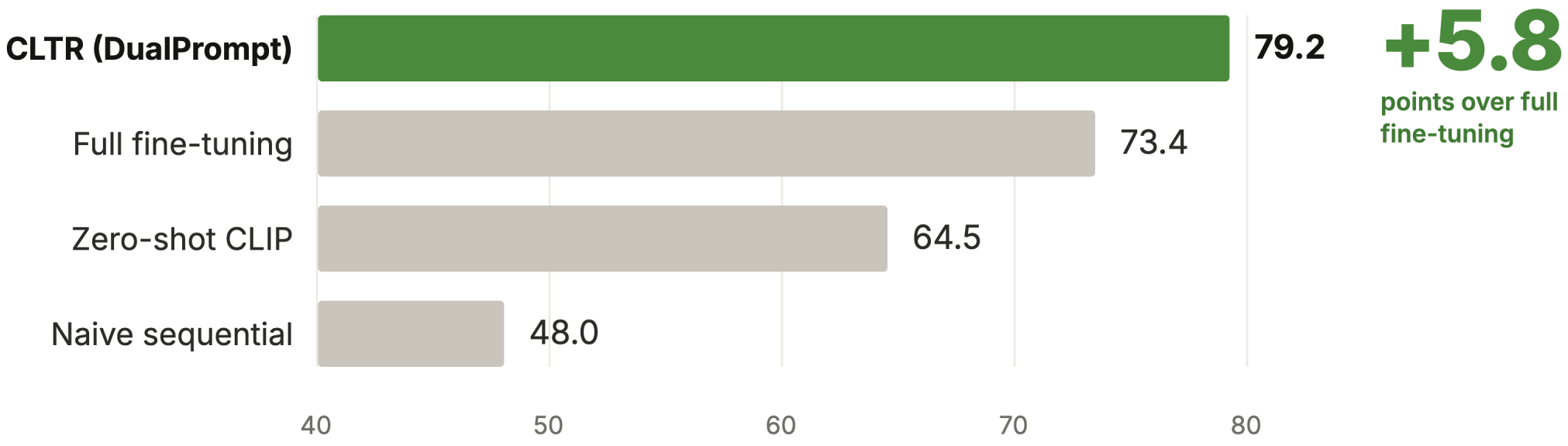
Naive sequential

violated everywhere

5 RESULTS · FOUNDATION MODEL

CLTR adapts CLIP better than full fine-tuning.

CLIP VIT-B/32 · CIFAR100-LT · IF 100 · TOP-1 ACCURACY (%)



Also +7.1 at IF 50 and +3.0 at IF 10 (Table 2). CLTR (DualPrompt) also uses balanced softmax, replay, distillation and logit adjustment.

Standard CL methods rival specialized LTR methods.

4 of 6

benchmark settings where a standard CL method (SGP or TPL) is **best or tied-best** against specialized LTR methods.

Setting	Best specialized LTR	Best CLTR
CIFAR100-LT · IF 100	53.4 · WD+Max	50.7 · SGP
CIFAR100-LT · IF 50	57.7 · WD+Max	58.0 · SGP · best
CIFAR100-LT · IF 10	68.7 · WD+Max	67.2 · SGP
CIFAR10-LT · IF 100	84.3 · BCL	84.7 · TPL · best
CIFAR10-LT · IF 50	87.2 · BCL	87.6 · TPL · best
ImageNet-LT	53.9 · WD+Max	53.9 · TPL · tied

Top-1 accuracy (%), Table 1. ResNet-32 on CIFAR, ResNeXt-50 on ImageNet-LT.

TAKEAWAYS

Long-tailed recognition is a continual-learning problem.

1. Long-tailed training lands within **O(1/√IF)** of the Head-only solution, in convex models and deep networks.

2. So LTR can be written as continual learning: learn the **Head**, then the **Tail** without forgetting.

3. With a provable upper bound, standard CL methods rival specialized LTR methods.

Code + Lean 4 proofs

github.com/MahdiyarMM/CLTR

Interactive note

mahdiyar.ai/posts/cltr