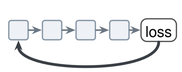
Cumulative-Goodness Free-Riding in Forward-Forward Networks: Real, Repairable, but Not Accuracy-Dominant

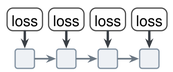
Amirhossein Yousefiramandi | Independent Researcher | NeurIPS 2026 | Project page: amirhossein-yousefi.github.io/ff-free-riding

NEURAL INFORMATION PROCESSING SYSTEMS

Background

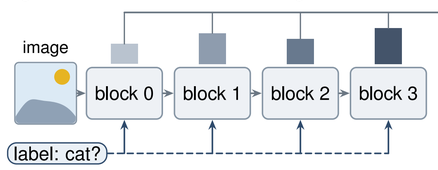


Backpropagation: errors flow back through all layers

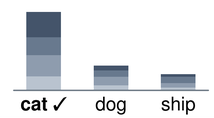


Forward-Forward (FF): each block learns from its own loss; no gradient between blocks

goodness: each block's score for (image, label), high for the true label, low for wrong ones



prediction: label with the largest summed goodness



inside our blocks: self-attention → feed-forward layer → goodness head

Q: If later blocks free-ride on earlier ones, is that why FF lags backpropagation?

In **cumulative-goodness** FF, a block's training score includes the earlier blocks' goodness, so later blocks can inherit a partly solved task and get little signal to improve their own margin: they **free-ride**.

Expected: repair → healthier deep blocks → higher accuracy

Key idea: cumulative goodness

history weight

$$G^{(d)} = g^{(d)} + \gamma \cdot \sum_{j<d} \mathrm{sg}\left(g^{(j)}\right) \quad \textit{training only}$$

$G^{(d)}$: block d's training score

$g^{(d)}$: its own goodness

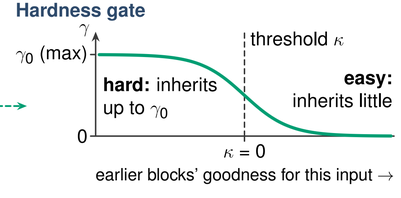
$\sum_{j<d} \mathrm{sg}\left(g^{(j)}\right)$: history: earlier blocks' goodness, detached: no gradient flows back (sg = stop-gradient)

History rules

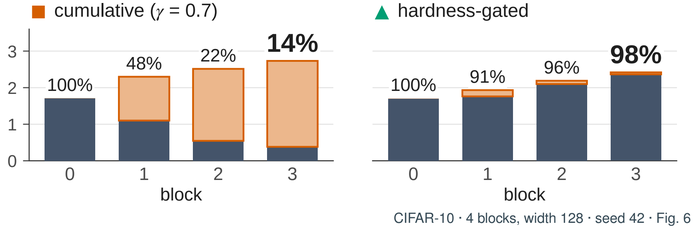
- **cumulative:** $\gamma = 0.7$ always
- **history-free:** $\gamma = 0$
- **hardness-gated:** this curve

Prediction never uses γ.

Hardness gate



training score (true label) = own goodness + inherited history (% = own share)

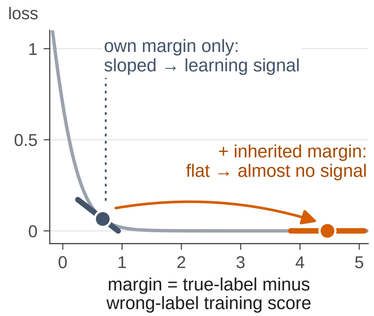


CIFAR-10 · 4 blocks, width 128 · seed 42 · Fig. 6

Free-riding: with $\gamma = 0.7$, block 3 makes only 14% of its training score itself.

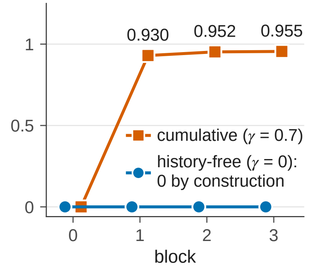
Free-riding is real

(a) Inherited margin flattens the loss



(b) Measured: free-riding index

share of the loss slope lost to history



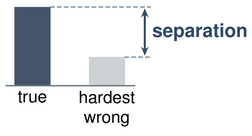
CIFAR-10 · 4 blocks, width 128 · seed 42 · (a) dots: block-3 batch means (Tab. 4)

1 Result 1: History provably flattens the cumulative loss

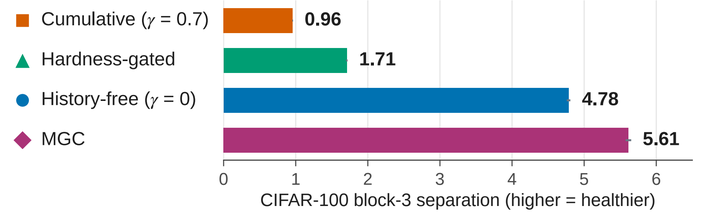
Theorem 3.1 (softplus loss): block d's learning signal from the cumulative loss shrinks exponentially with the earlier blocks' total margin when both it and block d's own margin are positive (∼96% of test images).

. . . and repairable

Separation: a block's own margin against the hardest of several sampled wrong labels, averaged.



- **History-free:** remove the history ($\gamma = 0$)
- **Hardness-gated:** gate it per example
- **MGC** (missing-gradient-compensated): add back, approximately, the signal history removes

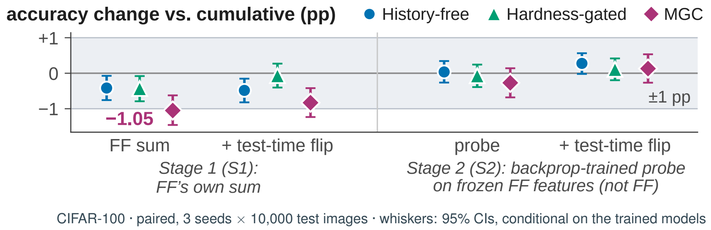


CIFAR-100 · 4 blocks, width 256 · 3 seeds · mean ± SD

2 Result 2: Removing or gating history restores separation

History-free and hardness-gated raise block-3 separation up to 5× on CIFAR-100. **Caveat:** MGC replaces the depth-scaled auxiliary that the history rules keep (a loss on each block's own margin, weighted up with depth); in a CIFAR-10 control, 92% of MGC's separation gain came from dropping it, not the compensation; its CIFAR-100 gain is not decomposed.

. . . but not accuracy-dominant



CIFAR-100 · paired, 3 seeds × 10,000 test images · whiskers: 95% CIs, conditional on the trained models

3 Result 3: Healthier deep blocks, yet no accuracy gain

History-free and hardness-gated: 95% CIs inside ±1 percentage point (pp), FF sum slightly below 0. MGC: an FF-sum cost the probe recovers.

Same network, goodness re-split
goodness moved between blocks

blocks 0–2 — **block 3: 14% of sum**

block 3 — **block 3: 98% of sum**

same sum for every label

0 / 10,000 predictions change

Retrained with a repair (probe + flip)



∼13% of predictions change between any two history rules; about as many fixed as broken

MGC vs. cumulative: 22.7%

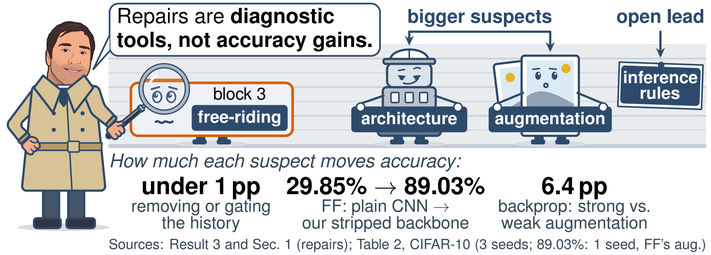
CIFAR-100 · re-split: cumulative model, seed 42, two arbitrary splits · retrained: 3 seeds

4 Result 4: The sum hides where the evidence is

Re-splitting the sum changes no prediction (Prop. 3.2). Repairs are not exact re-splits: they change what the network learns, and that accuracy stays put on these readouts (MGC's Stage-1 cost aside) is empirical.

Is free-riding why FF lags backpropagation?

Not the main bottleneck in our settings.



How much each suspect moves accuracy:

under 1 pp removing or gating the history

29.85% → 89.03% FF: plain CNN → our stripped backbone

6.4 pp backprop: strong vs. weak augmentation

Sources: Result 3 and Sec. 1 (repairs); Table 2, CIFAR-10 (3 seeds; 89.03%: 1 seed, FF's aug.)

Calibration: our FF backbone vs. the strongest prior FF result we list

CIFAR-10	90.50% vs. ASGE's 90.62%	**within 0.12 pp**
CIFAR-100	66.37% vs. ASGE's 65.42%	**+0.95 pp**
Tiny ImageNet	48.54% vs. SCFF's 35.67%†	**+12.9 pp**†

single crop; ours: $\gamma = 0$, 3 seeds · †SCFF is unsupervised: not like-for-like · Sec. 4.2, App. Tab. 29 calibration, not a leaderboard: architectures, readouts, and augmentation differ

Scope: softplus and squared-hinge (CIFAR-10, 4 blocks × 128 only) losses, one goodness statistic, co-designed backbones, CIFAR-10/100, Tiny ImageNet; controls reduce, not remove, a recipe confound.

Open: *ImageNet scale? Repairs at greater depth?*

Code

Paper

github.com/amirhossein-yousefi/ff-free-riding

arxiv.org/abs/2605.06240 · Hinton (2022), arXiv:2212.13345
ASGE, arXiv:2509.12394 · SCFF, Nature Communications 16:5978