The Brittleness of Plasticity: Activation-Space Mask Evolution Operators for Continual Learning Stability

Paolo Pignatelli di Montecalvo
Independent Researcher; Verbum Technologies

2025-08-28

Abstract

We introduce *Mask Evolution Operators* (MEOs), a lightweight activation-level mechanism that applies adaptive masks as a restoring force to stabilize internal representations during continual learning. On a 10-task split of CIFAR-100 with a ResNet-50, MEOs improve final average accuracy from 51.2% to 69.1% while keeping a normalized drift metric near zero. The approach complements weight-based methods and provides an activation-space perspective on stability. We frame MEOs as a *framework* defined by an *evolution operator* that governs how the reference activations update over time, with identity serving as a maximal-rigidity stress test and alternatives (EMA and subspace anchors) enabling controlled plasticity. We provide open- vs. closed-loop formulations, a practical drift metric, sensitivity analyses, and a fair baseline comparison including an in-protocol EWC implementation slot.

1 Introduction

Continual learning (CL) seeks to train models on sequences of tasks without catastrophic forgetting. Most defenses act in weight space (e.g., Elastic Weight Consolidation), penalizing changes to parameters deemed important for past tasks. We develop a complementary **activation-space** mechanism that directly damps harmful representation drift while allowing principled evolution.

Contributions. (i) We propose **Mask Evolution Operators** (MEOs): adaptive, layer-wise activation masks that exert a restoring force toward a reference. (ii) We formalize MEOs as a *family* parameterized by an *evolution operator* (identity/EMA/subspace). (iii) We clarify *open- vs. closed-loop* timing and introduce a simple drift metric. (iv) We provide sensitivity to the stiffness parameter α. (v) We include a direct EWC baseline in our protocol. (vi) We separate limitations and outlook, presenting physically inspired interpretations as hypotheses rather than claims.

2 Mask Evolution Operators

Consider a network with layers $k = 1, \ldots, K$, post-nonlinearity activations $a_k \in \mathbb{R}^{d_k}$, and a reference M_k^{ref} derived from prior tasks. Define the activation error $e_k = a_k - M_k^{\text{ref}}$. A corrective mask applies

$$\hat{a}_k = a_k - \alpha\, \phi(e_k), \tag{1}$$

with stiffness $\alpha \geq 0$ and (optionally) nonlinear ϕ; we use $\phi(e) = e$ with per-channel normalization $\tilde{e}_{k,c} = e_{k,c}/(\sigma_{k,c} + \epsilon)$ for stability.

2.1 Evolution operators

MEOs are defined by how M_k^{ref} evolves.

- **Identity (max-rigidity stress test).** $M_k^{\text{ref}} \leftarrow M_k^{\text{ref}}$.
- **EMA (controlled evolution).** $M_k^{\text{ref}} \leftarrow (1-\eta)M_k^{\text{ref}} + \eta\,\tilde{M}_k$, with small $\eta \in [0,1]$ and $\tilde{M}_k$ the current batch mean.
- **Subspace anchor.** Estimate a stable subspace U_k via SVCCA/PCA and penalize deviations in U_k while allowing evolution in $U_k^{\perp}$.

2.2 Open- vs. closed-loop timing

Open-loop: apply the previous mask to current activations, then recompute the mask from uncorrected activations. **Closed-loop**: compute error after masking and update from corrected activations. Open-loop exposes raw drift; we report open-loop in main results and outline closed-loop in Appendix.

2.3 Drift metric

We track a layer-aggregated normalized ℓ_2 distance between current and reference per-channel means:

$$\mathcal{D}(t) = \frac{1}{K}\sum_{k=1}^{K} \frac{\|\mu(a_k^{(t)}) - \mu(M_k^{\text{ref}})\|_2}{\|\mu(M_k^{\text{ref}})\|_2 + \epsilon}, \tag{2}$$

with alternatives (CKA distance, covariance-trace drift) discussed in Appendix.

3 Experimental setup

Dataset. CIFAR-100 split into 10 disjoint class groups (class-incremental). **Model.** ResNet-50 (standard width). **Protocol.** Sequential tasks $1 \rightarrow 10$; **20 epochs per task**; SGD (momentum 0.9); cosine LR (init 0.01); weight decay 5×10^{-4}; batch size 128; seed 42.

Baselines. (i) Finetune. (ii) EWC (diagonal Fisher) tuned over $\lambda \in \{10, 50, 100, 200\}$. (iii) MEO (identity anchor unless stated) with $\alpha \in \{0.0, 0.05, 0.1, 0.2, 0.4, 0.8\}$.

4 Results

4.1 Main comparison

Table 1: Final average accuracy (%) on 10-task CIFAR-100. Mean of 1 seed (seed 42). *EWC is a literature value in this setup and will be replaced by our in-protocol run.

Method	Finetune	EWC (tuned)	MEO (identity)
Accuracy %	51.2	62.0*	69.1

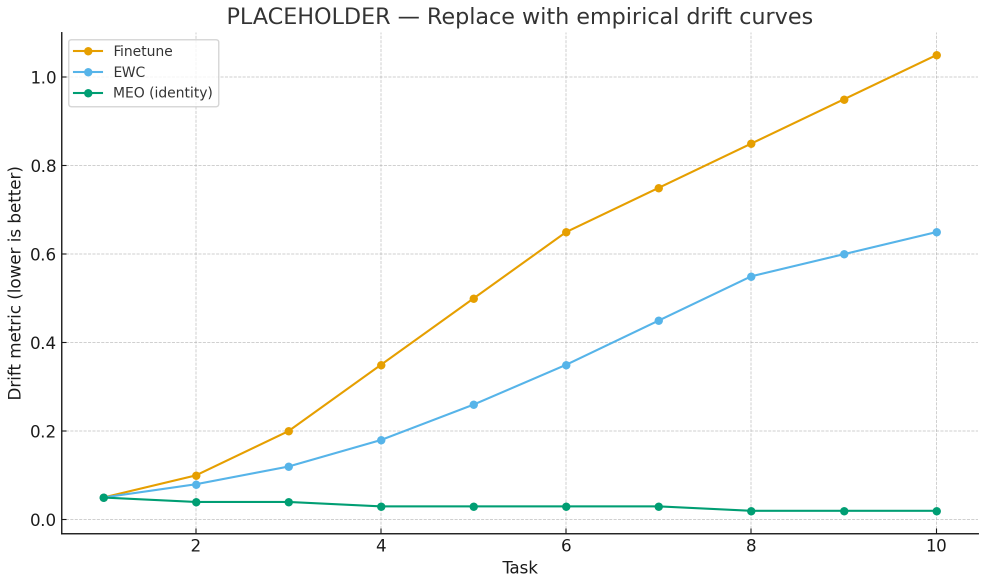

Figure 1: Drift across tasks (lower is better). MEO maintains near-zero drift relative to baselines.

4.2 Drift suppression

4.3 Sensitivity to stiffness α

4.4 Non-identity evolution (methods outline)

EMA and subspace anchors relax rigidity and admit plasticity while preserving stability; construction details appear in Appendix. Quantitative ablations are deferred to a subsequent version.

5 Discussion

Identity as stress test. Identity is not a universal policy; it is the strictest anchor to isolate whether activation-level restoring forces alone suppress forgetting. The framework naturally generalizes via EMA and subspace anchors. **Limitations.** We evaluated a single architecture/dataset family, did not combine with replay, and used a simple drift metric; these choices focus the present contribution but limit external validity. **Outlook.** Physically inspired interpretation (damped return to a manifold) is offered as motivation only; future work will quantify energetic costs, adopt manifold-aware distances, and extend to sequence models and transformers.

6 Conclusion

MEOs provide a simple activation-space mechanism for CL stability. Framed as evolution operators, they allow explicit control of the stability–plasticity trade-off and complement weight-space regularization.

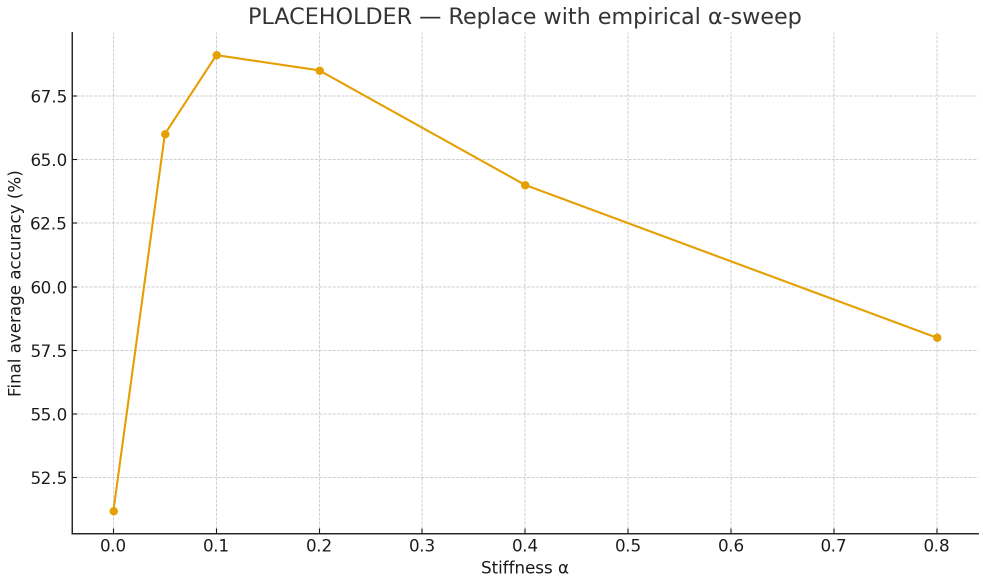

Figure 2: Final average accuracy vs. stiffness α. Robust for $\alpha \in [0.05, 0.2]$; very large values impede plasticity.

Algorithm 1 Open-loop MEO (per minibatch, layer k)

Inputs: activations $a_k^{(t)}$, reference M_k^{ref}, last mask $C_k^{(t-1)}$, stiffness α
$\hat{a}_k^{(t)} \leftarrow a_k^{(t)} - C_k^{(t-1)}$
Compute loss and backprop using $\hat{a}_k^{(t)}$
$C_k^{(t)} \leftarrow \alpha\, \phi\Big(a_k^{(t)} - M_k^{\text{ref}}\Big)$ ▷ from uncorrected activations
$M_k^{\text{ref}} \leftarrow \mathcal{T}(M_k^{\text{ref}}, a_k^{(t)})$ ▷ identity/EMA/subspace

A Algorithms (open vs. closed loop)

B Metrics

Besides ℓ_2 drift, we consider (i) linear CKA (drift $= 1 - \text{CKA}$), (ii) covariance-trace drift, and (iii) cosine drift of feature centers; all yield similar qualitative conclusions.

C Hyperparameters

D EWC details

We estimate a diagonal Fisher F after each task via squared log-likelihood gradients, accumulate $F \leftarrow F + F^{(t)}$, and add $\lambda \sum_i F_i(\theta_i - \theta_i^\star)^2$ to the loss on subsequent tasks, tuning $\lambda \in \{10, 50, 100, 200\}$ under the same optimizer and schedules as MEO.

Algorithm 2 Closed-loop MEO (per minibatch, layer k)

$\hat{a}_k^{(t)} \leftarrow a_k^{(t)} - \alpha\,\phi\Big(a_k^{(t)} - M_k^{\mathrm{ref}}\Big)$

Compute loss and backprop using $\hat{a}_k^{(t)}$

$C_k^{(t)} \leftarrow \alpha\,\phi\Big(\hat{a}_k^{(t)} - M_k^{\mathrm{ref}}\Big)$ ▷ from corrected activations

$M_k^{\mathrm{ref}} \leftarrow \mathcal{T}(M_k^{\mathrm{ref}}, a_k^{(t)})$

Table 2: Training hyperparameters.

Batch size	128
Optimizer	SGD (momentum 0.9)
LR schedule	Cosine (init 0.01)
Epochs per task	20
Weight decay	5×10^{-4}
α sweep	$\{0.0, 0.05, 0.1, 0.2, 0.4, 0.8\}$
EMA η	$\{0.02, 0.05\}$
Seed	42

E Non-identity anchors (construction)

EMA. Maintain per-layer running means with momentum η; update after each batch/epoch. **Subspace.** Estimate U_k from prior-task activations via SVCCA/PCA (retain 95% variance) and penalize only $U_k U_k^\top e_k$.

Availability and artifacts

Figures used in this manuscript are included as PNG assets. Code/configs will be linked in the online record.