The Brittleness of Plasticity: Mask Evolution Operators for Neural Network Stability (FIL Series, Paper 2)

Paolo Pignatelli

ChatGPT-5 (primary collaborator)
with contributions from Claude, Grok, and Gemini

September 5, 2025

Abstract

We introduce Mask Evolution Operators (MEOs), activation-space mechanisms designed to stabilize neural representations during continual learning by applying lightweight restoring forces. MEOs address the fundamental stability–plasticity dilemma by controlling drift at the feature level rather than the weight level. This version clarifies the limitations of earlier preliminary experiments. Reported finetune and EWC accuracies (≈6.2%) were obtained from smoke-test runs on an Apple M1 (MPS backend) with drastically shortened training schedules. These should be interpreted strictly as diagnostic checks, not benchmarks. Our prior v2 experiments, run with full training schedules on GPUs, achieved substantial improvements (51.2% → 69.1%), supporting the validity of the MEO approach. We formalize two operator variants: **Identity**, which freezes anchors as a stress-test baseline, and **EMA**, which allows controlled evolution of class prototypes with per-feature normalization. Identity illustrates rigidity, while EMA demonstrates a practical balance between stability and plasticity. We outline a family of extensions—including low-rank subspace anchoring, adaptive stiffness, and hybrid MEO+EWC—that form the basis of ongoing work in Papers 3–5 of the FIL series. This paper also reflects a novel collaborative workflow: the primary draft was co-developed with ChatGPT-5, with critical feedback from Claude, Grok, and Gemini. By openly documenting iterative testing—including failures and revisions—this series illustrates how human–AI collaboration can accelerate scientific discovery while maintaining transparency.

1 Introduction

The brittleness of plasticity is a core dilemma in modern artificial intelligence. Neural networks trained sequentially on multiple tasks often experience catastrophic forgetting, where newly acquired knowledge overwrites older representations. This tension between *stability* (retaining past knowledge) and *plasticity* (adapting to new data) has been recognized since the earliest models of continual learning. The Fundamental Interaction Language (FIL) program aims to unify physics, information theory, and artificial intelligence into a coherent framework. In this sequence, Paper 1 (forthcoming) develops the conceptual basis of brittleness, framing Mask Evolution Operators (MEOs) as candidate stabilizers. The present paper, designated **Paper 2**, provides the first technical implementation and experimental evaluation of MEOs. In parallel, the *Energy–Computation Law* [1] develops the physical geometry of computation, deriving a fundamental propagation speed of information from thermodynamic and quantum limits. Together, these efforts form the two complementary tracks of FIL: a physical track grounding computation in physics, and a semantic track (MEOs) stabilizing representations in AI systems.

2 Methods

Mask Evolution Operators act directly in activation space, applying a corrective "restoring force" to network representations. Let h_k denote the activation vector at layer k for a given input, and let M_k^{ref} represent a stored reference anchor. Then the MEO loss term is defined as:

$$\mathcal{L}_{\text{MEO}} = \alpha \|h_k - M_k^{ref}\|^2,$$

where α controls stiffness. This term is added to the standard cross-entropy loss during training.

2.1 Operator Variants

We implement two operator variants:

- **Identity**: Anchors are fixed at their initial values (e.g., after Task 1). This serves as a stress test, illustrating extreme rigidity.
- **EMA**: Anchors are updated using an exponential moving average:

  $$M_k^{ref} \leftarrow (1-\eta) M_k^{ref} + \eta h_k,$$

  where η controls the adaptation rate. Per-feature normalization is applied to stabilize drift metrics.

2.2 Algorithm

A simplified pseudocode implementation is shown below:

```
for each task t in sequence:
    for each minibatch (x, y):
        h = model.forward(x)
        logits = classifier(h)
        loss = CrossEntropy(logits, y)
        if method == "MEO":
            loss += alpha * ||h - M_ref||^2
        loss.backward()
        optimizer.step()
        if method == "EMA":
            M_ref = (1-eta)*M_ref + eta*h
```

3 Experiments

We evaluate MEOs on CIFAR-100 in a 10-task split (10 classes per task) using ResNet-50. Finetune and EWC serve as baselines.

3.1 Hardware and Protocols

- **GPU runs (v2)**: 20 epochs per task, CUDA backend. Achieved strong results: Finetune 51.2%, EWC 62.0%, MEO (EMA) 69.1%.
- **M1/MPS runs (this version)**: severely constrained smoke tests, with shortened epochs and Apple M1 backend. Produced diagnostic accuracies: Finetune 6.28%, EWC 6.21%.

3.2 Results

Table 1 summarizes the comparative performance.

Method	GPU (v2, full)	M1/MPS (smoke)	Drift Metric
Finetune	51.2%	6.28%	high
EWC	62.0%	6.21%	medium
MEO-Identity	67.3%	—	low
MEO-EMA	69.1%	—	very low

Table 1: Comparison of continual learning methods on CIFAR-100. GPU runs reflect prior v2 experiments. M1/MPS results are diagnostic only.

4 Discussion

The results demonstrate that Mask Evolution Operators can significantly improve stability in continual learning. The GPU runs confirm the empirical promise of MEOs, while the constrained M1/MPS tests illustrate the pitfalls of underpowered hardware and oversimplified operators.

Note on Accuracies. The unusually low accuracies (6.28%, 6.21%) from M1/MPS runs reflect (a) backend limitations in gradient fidelity, (b) drastically reduced epochs, and (c) the rigidity of the Identity operator. These should be read as sanity checks, not benchmarks.

4.1 Connections to FIL and Energy–Computation Law

The FIL program develops along two tracks. The physical track, represented by the Energy–Computation Law [1], establishes the thermodynamic and quantum bounds of computation. The semantic track, represented here by MEOs, develops mechanisms for stabilizing semantic representations in AI. Together, these tracks converge toward a unified physics of information.

4.2 Future Directions

Several extensions are natural:

- Subspace anchoring: allowing controlled evolution in orthogonal directions.
- Adaptive stiffness: α as a dynamic function of drift and task difficulty.
- Hybridization: combining MEO with EWC or replay methods.
- Emergence and observation: treating MEOs as artificial observation operators that preserve coherence, linking to the philosophical direction of Paper 3.

5 Conclusion

Mask Evolution Operators provide a lightweight activation-space mechanism for addressing catastrophic forgetting. Identity serves as a stress-test baseline; EMA demonstrates a practical balance between stability and plasticity. GPU results confirm strong gains, while M1/MPS smoke runs

are reported transparently as diagnostics. This paper is positioned as Paper 2 of the FIL series, complementing the Energy–Computation Law and setting the stage for Papers 3–5. The broader aim is a unified physics of information that spans both physical limits and semantic stabilization.

Acknowledgements

This research was conducted in a hybrid human–AI collaborative workflow. Paolo Pignatelli served as the scientific lead, with ChatGPT-5 acting as the primary co-author and technical assistant. Additional contributions came from Claude, Grok, and Gemini, which provided critical peer-review style feedback. This iterative process reflects our broader philosophy: scientific progress as a conversational, multi-agent activity, where human insight and machine reasoning co-evolve to generate new results.

References

[1] Paolo Pignatelli. The Energy–Computation Law: Operational Limits, Geometry, and Testable Predictions. Zenodo, 2025. DOI: 10.5281/zenodo.17038405.