

# E79: Coupled Memory-Modulation Matrix System

A Mathematical Analysis of Mutual Gating Control

*Formal verification in Lean 4 with Mathlib*

## 1 Introduction

E79 represents the culmination of 79 architectural experiments in recurrent neural network design. Its key innovation is **mutual gating control**: two  $n \times n$  matrix states where each controls the other's forgetting dynamics.

This document provides:

1. Complete mathematical specification of E79
2. Analysis of how M modulates S (and vice versa)
3. Jacobian and gradient flow analysis
4. Key insights from the Lean formalization
5. Testable predictions and open questions

## 2 Mathematical Specification

### 2.1 State Definition

E79 maintains two matrix states:

$$\mathbf{S} \in \mathbb{R}^{n \times n} \quad \text{Content Memory (primary associative storage)} \quad (1)$$

$$\mathbf{M} \in \mathbb{R}^{n \times n} \quad \text{Modulation Memory (controls S's gating)} \quad (2)$$

Total state:  $2n^2$  real values. For  $n = 32$ , this is 2048 elements.

### 2.2 Input Vectors

At each timestep, E79 receives:

- $\mathbf{k} \in \mathbb{R}^n$ : Key vector for content addressing
- $\mathbf{v} \in \mathbb{R}^n$ : Value to store
- $\mathbf{q} \in \mathbb{R}^n$ : Query for output
- $\mathbf{m} \in \mathbb{R}^n$ : Modulation key for M addressing

## 2.3 The E79 Update Rule

**Input:** State  $(S, M)$ , vectors  $(k, v, q, m)$ , biases  $(b_S, b_M)$

**Step 1: Normalize keys**

$$\hat{k} = \frac{k}{\|k\|_2}, \quad \hat{m} = \frac{m}{\|m\|_2} \quad (3)$$

**Step 2: M controls S's decay gates (M  $\rightarrow$  S coupling)**

$$g_{\text{row}}^S = \sigma(M\hat{k} + b_S) \in (0, 1)^n \quad (4)$$

$$g_{\text{col}}^S = \sigma(M^\top \hat{k} + b_S) \in (0, 1)^n \quad (5)$$

**Step 3: S delta rule update with M-controlled gating**

$$\delta_S = v - S\hat{k} \quad (6)$$

$$S' = (g_{\text{row}}^S g_{\text{col}}^S) \odot S + \delta_S \hat{k}^\top \quad (7)$$

**Step 4: S controls M's decay gates (S  $\rightarrow$  M coupling)**

$$g_{\text{row}}^M = \sigma(S\hat{m} + b_M) \quad (8)$$

$$g_{\text{col}}^M = \sigma(S^\top \hat{m} + b_M) \quad (9)$$

**Step 5: M delta rule update (M predicts S's changes)**

$$\delta_M = \delta_S - M\hat{m} \quad (10)$$

$$M' = (g_{\text{row}}^M g_{\text{col}}^M) \odot M + \delta_M \hat{m}^\top \quad (11)$$

**Step 6: Output with self-gating**

$$o = (S'q) \odot \text{silu}(S'q) \quad (12)$$

**Return:** New state  $(S', M')$ , output  $o$

Algorithm 1: E79 Forward Pass - Mutual Gating Control

## 3 Key Insight 1: Factorized Gating is Rank-Deficient Control

### 3.1 The Factorized Gate Structure

The decay applied to S has the form:

$$\text{Gate}_{ij} = g_{\text{row},i}^S \times g_{\text{col},j}^S \quad (13)$$

This is a **rank-1 outer product**:

$$G^S = g_{\text{row}}^S (g_{\text{col}}^S)^\top \in \mathbb{R}^{n \times n} \quad (14)$$

**Theorem** (Rank Deficiency). The factorized gate  $G^S = g_{\text{row}}^S (g_{\text{col}}^S)^\top$  has rank at most 1.

This means  **$2n$  parameters control  $n^2$  decay rates**.

*Proof.* Any outer product  $uv^\top$  has rank  $\leq 1$  since all columns are scalar multiples of  $u$ . □

## 3.2 Consequences of Rank Deficiency

**Key Point:** You cannot independently control each element's decay. If row  $i$  decays quickly ( $g_{\text{row},i}^S$  small), then **all elements in row  $i$**  decay quickly, regardless of column.

This constraint explains why E79 needs **two** coupled matrices:

- Single matrix with factorized gating has limited expressiveness
- The coupling between S and M compensates for each other's rank deficiency
- M can modulate S's gating to achieve richer decay patterns than either could alone

**Proposition** (Effective Degrees of Freedom). The factorized gate has  $2n - 1$  effective degrees of freedom (not  $2n$ , due to the constraint that scaling  $\mathbf{g}_{\text{row}}$  by  $c$  and  $\mathbf{g}_{\text{col}}$  by  $\frac{1}{c}$  gives the same result).

Compare to full gating:  $n^2$  degrees of freedom.

The ratio:  $\frac{2n-1}{n^2} \approx \frac{2}{n}$  for large  $n$ .

## 4 Key Insight 2: Bidirectional Jacobian Coupling

### 4.1 The Jacobian is NOT Lower-Triangular

**Key Point:** Unlike the simplified description, the actual E79 Jacobian is **fully coupled** in both directions.

The full E79 state is  $\mathbf{z} = \text{vec}([\mathbf{S}; \mathbf{M}]) \in \mathbb{R}^{2n^2}$ .

**Theorem** (Bidirectional Coupling). The Jacobian of the E79 update has the block structure:

$$\frac{\partial \mathbf{z}'}{\partial \mathbf{z}} = \begin{pmatrix} \mathbf{J}_{SS} & \mathbf{J}_{SM} \\ \mathbf{J}_{MS} & \mathbf{J}_{MM} \end{pmatrix} \quad (15)$$

where **both off-diagonal blocks are non-zero**:

- $\mathbf{J}_{SM} = \frac{\partial \mathbf{S}'}{\partial \mathbf{M}} \neq \mathbf{0}$ : M affects S' through gating
- $\mathbf{J}_{MS} = \frac{\partial \mathbf{M}'}{\partial \mathbf{S}} \neq \mathbf{0}$ : S affects M' through gating AND  $\delta_S$

*Proof. M  $\rightarrow$  S coupling:* From Equation 7, the gates  $\mathbf{g}_{\text{row}}^S, \mathbf{g}_{\text{col}}^S$  depend on M:

$$\mathbf{g}_{\text{row}}^S = \sigma(\mathbf{M}\hat{\mathbf{k}} + \mathbf{b}_S) \quad (16)$$

Therefore:

$$\frac{\partial \mathbf{S}'}{\partial \mathbf{M}} = \frac{\partial \mathbf{S}'}{\partial \mathbf{g}^S} \cdot \frac{\partial \mathbf{g}^S}{\partial \mathbf{M}} \neq \mathbf{0} \quad (17)$$

**S  $\rightarrow$  M coupling:** From Equation 11, the gates  $\mathbf{g}_{\text{row}}^M, \mathbf{g}_{\text{col}}^M$  depend on S, and  $\delta_M$  depends on  $\delta_S$  which depends on S:

$$\frac{\partial \mathbf{M}'}{\partial \mathbf{S}} \neq \mathbf{0} \quad (18)$$

□

## 4.2 Dynamical Systems Interpretation

**Insight:** E79 is a **fully coupled nonlinear dynamical system**, not a hierarchical cascade. The two matrices co-evolve and mutually regulate each other's dynamics.

This is qualitatively similar to:

- **Lotka-Volterra equations** (predator-prey dynamics)
- **Coupled oscillators** in physics
- **Mutual inhibition circuits** in neuroscience

## 5 Key Insight 3: Gradient Flow Analysis

### 5.1 How M Gets Gradients

**Theorem** (M Gradient Path). M influences the output through the gating path:

$$\text{Loss} \rightarrow \mathbf{o} \rightarrow \mathbf{S}' \rightarrow \mathbf{g}_{\text{row}}^S, \mathbf{g}_{\text{col}}^S \rightarrow \mathbf{M} \quad (19)$$

The gradient:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = \frac{\partial \mathcal{L}}{\partial \mathbf{S}'} \cdot \frac{\partial \mathbf{S}'}{\partial \mathbf{g}^S} \cdot \frac{\partial \mathbf{g}^S}{\partial \mathbf{M}} \quad (20)$$

*Proof.* From Equation 7:  $S'_{ij} = g_{\text{row},i}^S \cdot g_{\text{col},j}^S \cdot S_{ij} + (\delta_S)_i \hat{k}_j$

The gradient with respect to  $g_{\text{row},i}^S$ :

$$\frac{\partial S'_{ij}}{\partial g_{\text{row},i}^S} = g_{\text{col},j}^S \cdot S_{ij} \quad (21)$$

And  $g_{\text{row},i}^S = \sigma\left(\sum_l M_{il} \hat{k}_l + (b_S)_i\right)$ , so:

$$\frac{\partial g_{\text{row},i}^S}{\partial M_{il}} = \sigma'(\dots) \cdot \hat{k}_l \quad (22)$$

Composing via chain rule yields a non-zero path from Loss to M. □

### 5.2 Meta-Learning Interpretation

**Insight:** M receives gradients that encode: “If you had gated S differently, the output would have been better.”

This is **implicit meta-learning** — M learns to control S's forgetting based on task loss, without explicit meta-supervision.

## 6 Key Insight 4: Tied Keys Collapse the System

### 6.1 The Tied Keys Theorem

**Theorem** (Tied Keys Reduction). If  $m = k$  everywhere during training, then E79's expressive power collapses toward a single matrix.

Specifically, M cannot organize independently from S when using the same addressing.

*Proof.* With  $\hat{m} = \hat{k}$ , both matrices are updated and queried with the same key.

The residual:

$$\delta_M = v - S\hat{k} - M\hat{k} = v - (S + M)\hat{k} \quad (23)$$

The combined retrieval  $(S + M)\hat{k}$  acts like a single matrix.  $\square$

**Key Point:** The separate modulation key  $m$  is **essential** for E79 to be more than a single larger matrix.

**Testable prediction:** If trained weights satisfy  $W_m \approx W_k$ , then E79 is not utilizing its full capacity.

## 7 Key Insight 5: State Efficiency vs Attention

### 7.1 State Size Comparison

Model	State Size	Per-Step Cost	Scaling
E79	$2n^2$	$O(n^2)$	Fixed
Attention	$T \times d$	$O(T^2 d)$	Grows with $T$
E1 (vector)	$n$	$O(nd)$	Fixed

**Theorem** (Crossover Point). E79 uses less memory than attention when sequence length  $T$  exceeds:

$$T > \frac{2n^2}{d} \quad (24)$$

For  $n = 32, d = 512$ : crossover at  $T > 4$ .

E79 compresses arbitrarily long sequences into fixed  $2n^2$  state.

### 7.2 The Compression Tradeoff

**Insight:** E79 trades **sequence-length scaling** for **fixed-size compression**.

- Attention: Full context access,  $O(T^2)$  cost
- E79: Compressed context,  $O(1)$  state but lossy

E79's mutual gating helps determine **what to keep** in the limited state budget.

## 8 Key Insight 6: K-Level Generalization

### 8.1 The K-Level Hierarchy

**Definition** (K-Level Coupled Memory). For  $K \geq 1$ , define matrices  $M_0, M_1, \dots, M_{K-1} \in \mathbb{R}^{n \times n}$ :

$$r_0 = v - M_0 \hat{k}_0 \quad (\text{Level 0 residual}) \quad (25)$$

$$r_i = r_{i-1} - M_i \hat{k}_i \quad \text{for } i = 1, \dots, K-1 \quad (26)$$

Each level learns the residual of the previous level.

$K$	Description
1	Standard delta rule (E74). Single matrix.
2	E79. S + M with mutual gating.
3	Triple hierarchy. S + M + N.
$K$	Chain of $K$ mutually-gated residual predictors.

## 8.2 Diminishing Returns

**Theorem** (Residual Decay). If level  $i$  converges (learns to predict  $\mathbf{r}_{i-1}$  well), then:

$$\|\mathbf{r}_i\| \ll \|\mathbf{r}_{i-1}\| \quad (27)$$

Each additional level has diminishing marginal benefit.

**Conjecture.** There exists an optimal  $K^*$  that depends on:

1. **Task complexity:** Structure in residuals
2. **Training time:** Deeper hierarchies converge slower
3. **Compute budget:** Each level costs  $n^2$  parameters and  $O(n^2)$  compute

The benchmark showing  $n = 32$  optimal for 10-minute training suggests  $K = 2$  is near-optimal for that regime.

## 9 Testable Predictions

The formalization yields several experimentally testable predictions:

### 9.1 Prediction 1: Key Divergence

**Measure:**  $\frac{\|\mathbf{W}_m - \mathbf{W}_k\|_F}{\|\mathbf{W}_k\|_F}$

**Expected:** This should be significantly positive ( $> 0.1$ ) if E79 is utilizing both matrices effectively.

**If violated:** E79 has collapsed to approximately a single larger matrix.

### 9.2 Prediction 2: Gate Utilization

**Measure:** Variance of  $\mathbf{g}_{\text{row}}^S$  and  $\mathbf{g}_{\text{col}}^S$  across inputs.

**Expected:** High variance indicates M is actively controlling S's forgetting.

**If violated:** Gates are near-constant, reducing to fixed decay.

### 9.3 Prediction 3: Residual Decay Over Training

**Measure:**  $\frac{\|\delta_M\|}{\|\delta_S\|}$  over training.

**Expected:** Should decrease if M learns to predict S's errors.

**If violated:** M is not learning useful residual structure.

### 9.4 Prediction 4: Jacobian Spectral Radius

**Measure:** Largest eigenvalue magnitude of the coupled Jacobian.

**Expected:** Should be  $< 1$  for stability.

If **violated**: Risk of gradient explosion or state divergence.

## 10 Comparison to Related Architectures

Architecture	Coupling	Gating	State
LSTM	Hierarchical (cell/hidden)	Input-dependent	Vector
Transformer	None (parallel)	Attention weights	KV cache
Mamba/SSM	None	Input-dependent	Diagonal matrix
E79	<b>Mutual (bidirectional)</b>	<b>Cross-matrix</b>	<b>Full matrices</b>

**Insight:** E79 is unique in having **bidirectional mutual control** between memory systems. This is more like biological neural circuits (e.g., cortical-thalamic loops, hippocampal-prefrontal interactions) where populations mutually regulate each other.

## 11 Empirical Results Summary

From the benchmark (100M params, 10-minute training):

Model	Loss	tok/s	State
Mamba2	1.27	78.7K	SSM (parallel)
<b>E79 n=32</b>	<b>1.51</b>	<b>31.5K</b>	<b><math>2 \times 32^2 = 2048</math></b>
E1 (gated)	1.53	45.5K	vector
E42 (linear)	1.59	137K	vector
FLA-GDN	1.99	18.7K	matrix

Key observations:

- E79 beats E1 (1.51 vs 1.53): mutual gating helps
- n=32 optimal for 10-min training (larger n under-converged)
- 40% of Mamba2 throughput despite sequential scan

## 12 Summary of Formalization Insights

#	Insight
1	<b>Factorized gating is rank-deficient:</b> $2n$ params control $n^2$ decays. The coupling compensates.
2	<b>Jacobian is bidirectionally coupled:</b> Not hierarchical—true mutual control.
3	<b>Gradient flow enables meta-learning:</b> M learns “how to gate S” from task loss.
4	<b>Tied keys collapse the system:</b> Separate $m \neq k$ is essential.
5	<b>Fixed state beats attention for long sequences:</b> Crossover at $T > 2 \frac{n^2}{d}$ .
6	<b>K-level hierarchies have diminishing returns:</b> K=2 may be near-optimal.
7	<b>Mutual control resembles biological circuits:</b> Lotka-Volterra / coupled oscillator dynamics.

## 13 Open Questions

1. **Optimal K for K-level hierarchies:** Is K=2 optimal, or would K=3 help for harder tasks?
2. **Adaptive coupling:** Can we learn the coupling structure rather than hard-coding it?

3. **Parallel scan for coupled matrices:** Can we achieve Mamba2-like parallelism?
4. **Scaling laws:** How does E79 scale beyond 100M parameters?
5. **Biological analogs:** Are there neural circuits with similar mutual gating dynamics?
6. **Formal stability analysis:** Under what conditions is the coupled system guaranteed stable?