Exact Flow Linear Attention: Exact Solution from Continuous-Time Dynamics

Jingdi Lei[1], Di Zhang[2], Soujanya Poria[1]

[1]Nanyang Technological University, [2]Fudan University,

Code Available

Motivation

- Softmax attention is **quadratic** in length; linear attention keeps a fixed-size state with $O(Ld^2)$ cost.
- **Delta-rule** linear attention (DeltaNet) is strong, but is usually viewed only as an online learning rule.
- Viewed as a numerical method it is only **first-order accurate**: error grows when dynamics are **stiff** (large key norms or step sizes).
- Existing fixes (decay, gating, adaptive forgetting) are **heuristic**.

Can we derive the exact update directly from the underlying continuous-time dynamics?

Key Insight: Delta Rule = Euler

One **explicit Euler step** of size β_t gives exactly the DeltaNet update:

$$\begin{aligned}\mathbf{S}_t &= \mathbf{S}_{t-1} + \beta_t(\mathbf{b}_t - \mathbf{A}_t\mathbf{S}_{t-1}) \\ &= (\mathbf{I} - \beta_t\mathbf{k}_t\mathbf{k}_t^\top)\mathbf{S}_{t-1} + \beta_t\mathbf{k}_t\mathbf{v}_t^\top\end{aligned}$$

The Euler step is only first-order accurate: it degrades for **large key norms or step sizes**, exactly where robustness matters.

$$\frac{d\mathbf{S}(t)}{dt} = -\mathbf{A}_t\,\mathbf{S}(t) + \mathbf{b}_t, \qquad \mathbf{A}_t = \mathbf{k}_t\mathbf{k}_t^\top, \quad \mathbf{b}_t = \mathbf{k}_t\mathbf{v}_t^\top$$

Under **zero-order hold** ($\mathbf{k}_t, \mathbf{v}_t$ constant within each token interval) the memory follows a linear ODE:

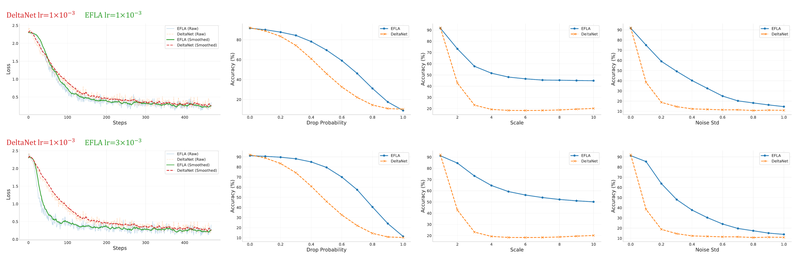


Figure 1: On sMNIST, EFLA converges faster than DeltaNet and stays more robust to dropout, intensity scaling and Gaussian noise.

Our Contributions

- Reinterpret delta-rule linear attention as the **explicit Euler discretization** of a ZOH continuous-time system.
- **EFLA** replaces the Euler step with the **exact closed-form flow**, made tractable by the **rank-1** dynamics.
- Same parameters, linear-time cost and chunkwise parallelism as DeltaNet, with **better robustness**, **lower perplexity** and **stronger downstream accuracy**.

EFLA: Exact Flow Linear Attention

1 Solve the ODE exactly over one step of length β_t:

$$\mathbf{S}_t = e^{-\beta_t\mathbf{A}_t}\,\mathbf{S}_{t-1} + \int_0^{\beta_t} e^{-(\beta_t-\tau)\mathbf{A}_t}\,\mathbf{b}_t\,d\tau$$

2 Rank-1 structure: $\mathbf{A}_t^2 = \lambda_t\,\mathbf{A}_t$ with $\lambda_t = \mathbf{k}_t{}^\top\mathbf{k}_t$, so the matrix exponential has a closed form (no $O(d^3)$ expm):

$$e^{-\beta_t\mathbf{A}_t} = \mathbf{I} - \frac{1-e^{-\beta_t\lambda_t}}{\lambda_t}\,\mathbf{A}_t$$

and, since $\mathbf{A}_t\mathbf{b}_t = \lambda_t\mathbf{b}_t$, the input integral collapses:

$$\int_0^{\beta_t} e^{-(\beta_t-\tau)\mathbf{A}_t}\,\mathbf{b}_t\,d\tau = \frac{1-e^{-\beta_t\lambda_t}}{\lambda_t}\,\mathbf{b}_t$$

3 EFLA update: the exact solution, in the same rank-1 form as DeltaNet:

$$\alpha_t = \frac{1-e^{-\beta_t\lambda_t}}{\lambda_t}$$

$$\mathbf{S}_t = (\mathbf{I} - \alpha_t\,\mathbf{k}_t\mathbf{k}_t^\top)\,\mathbf{S}_{t-1} + \alpha_t\,\mathbf{k}_t\mathbf{v}_t^\top$$

- **Small-step limit** recovers DeltaNet ($\alpha_t \to \beta_t$); EFLA removes the Euler error:

$$\alpha_t = \beta_t - \frac{1}{2}\beta_t^2\lambda_t + \mathcal{O}(\beta_t^3\lambda_t^2)$$

- **Key norm controls decay**: the memory component along $\mathbf{k}_t$ contracts by $e^{-\beta t\lambda t}$, while the orthogonal subspace is untouched:

$$\mathbf{k}_t^\top(\mathbf{I} - \alpha_t\,\mathbf{k}_t\mathbf{k}_t^\top) = e^{-\beta_t\lambda_t}\,\mathbf{k}_t^\top$$

4 Chunkwise-parallel training (WY / UT transform, chunk $[i]$) works unchanged:

$$\mathbf{T}_{[i]} = (\mathbf{I} + \mathrm{StrictTril}(\mathrm{diag}(\boldsymbol{\alpha})\,\mathbf{K}_{[i]}\mathbf{K}_{[i]}^\top))^{-1}\mathrm{diag}(\boldsymbol{\alpha})$$

$$\mathbf{W}_{[i]} = \mathbf{T}_{[i]}\mathbf{K}_{[i]}, \qquad \mathbf{U}_{[i]} = \mathbf{T}_{[i]}\mathbf{V}_{[i]}$$

$$\mathbf{S}_{[i+1]} = \mathbf{S}_{[i]} + \mathbf{K}_{[i]}^\top(\mathbf{U}_{[i]} - \mathbf{W}_{[i]}\mathbf{S}_{[i]})$$

$$\mathbf{O}_{[i]} = \mathbf{Q}_{[i]}\mathbf{S}_{[i]} + (\mathbf{Q}_{[i]}\mathbf{K}_{[i]}^\top \odot \mathbf{M})(\mathbf{U}_{[i]} - \mathbf{W}_{[i]}\mathbf{S}_{[i]})$$

- **No extra parameters**, same $O(Ld^2)$ complexity: a drop-in replacement for the delta-rule state update.
- **Gated EFLA**: the gated delta rule also has an exact closed-form flow.

Algorithm: EFLA, recurrent form (per token)

1 $\lambda_t = \mathbf{k}_t^\top\mathbf{k}_t$ (clip at 10^{-12})
2 $\alpha_t = (1 - e^{-\beta t\lambda t}) / \lambda_t$ (computed with expm1)
3 $\mathbf{S}_t = (\mathbf{I} - \alpha_t\mathbf{k}_t\mathbf{k}_t^\top)\mathbf{S}_{t-1} + \alpha_t\mathbf{k}_t\mathbf{v}_t^\top$
4 $\mathbf{o}_t = \mathbf{S}_t^\top\mathbf{q}_t$ Tip: keys scaled by $\sqrt{d}$ work best.

Main Results

Language modeling on SlimPajama (340M: 8B tokens, 1.3B: 50B tokens). Avg = mean of 9 zero-shot tasks.

Model	Wiki ppl ↓	LMB ppl ↓	LMB acc ↑	Avg acc ↑
340M				
Mamba-2	36.90	105.69	20.7	42.0
DeltaNet	38.09	96.26	22.5	42.2
EFLA (ours)	**35.26**	**79.97**	**23.9**	**44.1**
Gated DeltaNet	34.93	72.46	25.5	43.9
Gated EFLA (ours)	**34.28**	**69.37**	**26.4**	**44.8**
1.3B				
DeltaNet	18.38	17.29	41.8	50.5
EFLA (ours)	**18.30**	**16.54**	**43.2**	**51.8**

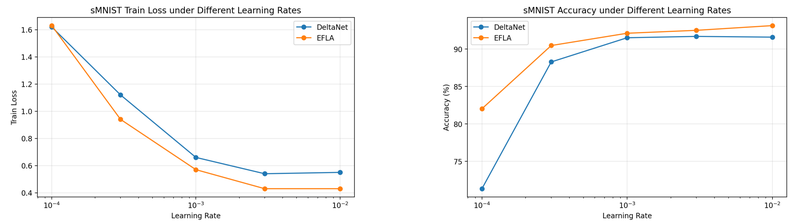
−17% LAMBADA ppl vs. DeltaNet (340M)

+1.9 avg acc (340M)

+1.3 avg acc (1.3B)

0 extra parameters

Robustness & Efficiency



(a) Training loss. (b) Accuracy.

Figure 2: sMNIST under different learning rates: EFLA has lower loss and higher accuracy at every LR.

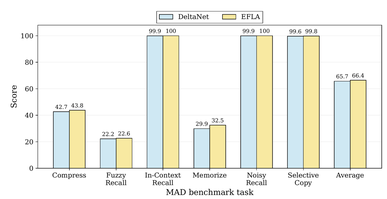


Figure 3: MAD benchmark (left).

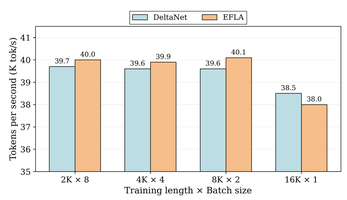


Figure 4: training throughput (right).

- **MAD**: avg 65.7 → **66.4**, largest gains on Memorize and Compress.
- **Throughput** is nearly identical to DeltaNet: no extra cost.

Don't approximate the ODE, **solve it**: exact flow at the cost of an Euler step!