EXPONENTIAL INTEGRATOR FOR DIAGONAL-DECAY DELTA ATTENTION: A NEGATIVE RESULT ON LENGTH EXTRAPOLATION

FARS
Analemma
fars@analemma.ai

ABSTRACT

Linear attention mechanisms like diagonal-decay delta attention (KDA) use Euler discretization with L2 normalization on keys and queries, which may discard useful key-norm information that could serve as a signal-strength channel for improved length extrapolation. We propose replacing the Euler coefficient with an exact exponential integrator derived from continuous-time dynamics, which provides bounded coefficients that enable stable training without L2 normalization. Experiments on three synthetic long-context tasks (Palindrome, MQAR, Stack) show that the exponential integrator achieves numerical stability without normalization (0 NaN/divergence across 27 runs). However, it does not improve accuracy at length extrapolation: the proposed method underperforms the baseline on Palindrome (-0.81 pp) and Stack (-2.98 pp) at $4\times$ extrapolation. Ablation analysis reveals that neither the integrator alone nor the removal of L2 normalization provides accuracy benefits. This negative result suggests that discretization error is not the primary bottleneck for length extrapolation in delta-rule attention.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Scaling Transformer context lengths is increasingly important for practical applications such as code assistants, agents, and long-document question answering. However, standard softmax attention has quadratic time and memory complexity in sequence length, motivating the development of subquadratic alternatives including linear attention mechanisms (Choromanski et al., 2020) and state-space models (Gu et al., 2021; Gu & Dao, 2023).

Delta-rule linear attention has emerged as a promising approach that updates a per-head state matrix using an error-correction rule derived from online regression (Yang et al., 2024b; Schlag et al., 2021). Recent variants such as Gated Delta Networks (Yang et al., 2024a) and Kimi Delta Attention (KDA) (Zhang et al., 2025) extend this framework with diagonal decay gates and achieve strong performance on long-context tasks. However, these methods implement the underlying continuous-time dynamics using first-order Euler discretization and require L2 normalization on keys and queries to ensure numerical stability.

We hypothesize that L2 normalization may discard useful key-norm information that could serve as a "signal-strength channel" for improved length extrapolation. Error-Free Linear Attention (EFLA) (Lei et al., 2025) shows that for rank-1 delta-rule dynamics, an exact exponential integrator can be computed in closed form, eliminating discretization error. We propose applying this exponential integrator to the delta-rule substep of KDA, which provides bounded coefficients that may enable stable training without L2 normalization.

Our contributions are as follows:

[1] https://gitlab.com/fars-a/exponential-integrator-delta-attention

- We derive an exponential-integrator coefficient for diagonal-decay delta attention that provides implicit stability control through bounded contraction when key norms are large.
- We demonstrate that this integrator enables numerically stable training without L2 normalization on keys and queries, with 0 NaN/divergence failures across 27 experimental runs.
- We report a negative result: despite achieving stability, the proposed method does not improve accuracy at length extrapolation, suggesting that discretization error is not the primary bottleneck for this task.

2 METHOD

2.1 BACKGROUND: DELTA-RULE LINEAR ATTENTION

Linear attention mechanisms replace the softmax normalization in standard attention with a linear recurrence, enabling $O(n)$ time complexity for sequence modeling (Choromanski et al., 2020; Schlag et al., 2021). The delta-rule variant maintains a state matrix $S_t \in \mathbb{R}^{d_k \times d_v}$ that is updated at each time step using an error-correction rule derived from online regression (Yang et al., 2024b). Given key $k_t \in \mathbb{R}^{d_k}$, query $q_t \in \mathbb{R}^{d_k}$, and value $v_t \in \mathbb{R}^{d_v}$, the delta-rule update minimizes the reconstruction loss $\mathcal{L}_t(S) = \frac{1}{2}\|S^\top k_t - v_t\|^2$ via a single gradient step with learning rate β_t, yielding:

$$S_t = (I - \beta_t k_t k_t^\top) S_{t-1} + \beta_t k_t v_t^\top. \tag{1}$$

This recurrence corresponds to an explicit Euler discretization of the continuous-time dynamical system $\frac{dS}{d\tau} = -k_t k_t^\top S + k_t v_t^\top$ with step size β_t (Lei et al., 2025). Recent work extends this formulation with diagonal decay gates, where Kimi Delta Attention (KDA) applies a per-channel decay $D_t = \mathrm{Diag}(\alpha_t^{\mathrm{gate}})$ before the delta update (Zhang et al., 2025; Yang et al., 2024a):

$$S_t = (I - \beta_t k_t k_t^\top) D_t S_{t-1} + \beta_t k_t v_t^\top. \tag{2}$$

To ensure numerical stability, KDA applies L2 normalization to keys and queries, constraining $\|k_t\| = 1$. While effective, this normalization discards key-norm information that could potentially serve as a learnable signal-strength channel for improved length extrapolation.

2.2 EXPONENTIAL INTEGRATOR FOR THE DELTA-RULE SUBSTEP

We propose replacing the Euler discretization in the delta-rule substep with an exact exponential integrator derived from the continuous-time dynamics. For the rank-1 system $\frac{dS}{d\tau} = -k_t k_t^\top S + k_t v_t^\top$ with constant (k_t, v_t) over the integration interval, the matrix exponential $\exp(-\beta_t k_t k_t^\top)$ admits a closed-form solution due to the rank-1 structure of $k_t k_t^\top$ (Lei et al., 2025; Chen et al., 2018).

Since $k_t k_t^\top$ has a single non-zero eigenvalue $\lambda_t = \|k_t\|^2$, the matrix exponential simplifies to:

$$\exp(-\beta_t k_t k_t^\top) = I - \frac{1 - \exp(-\beta_t \lambda_t)}{\lambda_t} k_t k_t^\top. \tag{3}$$

This yields the exact discrete-time update:

$$S_t = \left(I - \tilde{\alpha}_t k_t k_t^\top\right) \tilde{S}_{t-1} + \tilde{\alpha}_t k_t v_t^\top, \tag{4}$$

where $\tilde{S}_{t-1} = D_t S_{t-1}$ is the decayed state and the exponential-integrator coefficient is:

$$\tilde{\alpha}_t = \frac{1 - \exp(-\beta_t \|k_t\|^2)}{\|k_t\|^2}. \tag{5}$$

The key property of this coefficient is its bounded behavior: as $\|k_t\| \to \infty$, $\tilde{\alpha}_t \to 1/\|k_t\|^2 \to 0$, providing implicit stability control. In contrast, the Euler coefficient $\alpha_t = 1 - \exp(-\|k_t\|^2)$ approaches 1 as $\|k_t\| \to \infty$, which can lead to numerical instability when key norms are large. This bounded contraction property suggests that the exponential integrator may enable stable training without L2 normalization on keys and queries. Figure 1 illustrates the comparison between Euler discretization and the proposed exponential integrator.

For numerical stability in implementation, we compute $\tilde{\alpha}_t$ using the `expm1` function: $\tilde{\alpha}_t = -\mathrm{expm1}(-\beta_t \lambda_t)/\max(\lambda_t, \epsilon)$, and use the limit $\tilde{\alpha}_t \leftarrow \beta_t$ when $\lambda_t < \epsilon$ since $\lim_{\lambda \to 0}(1 - e^{-\beta\lambda})/\lambda = \beta$.

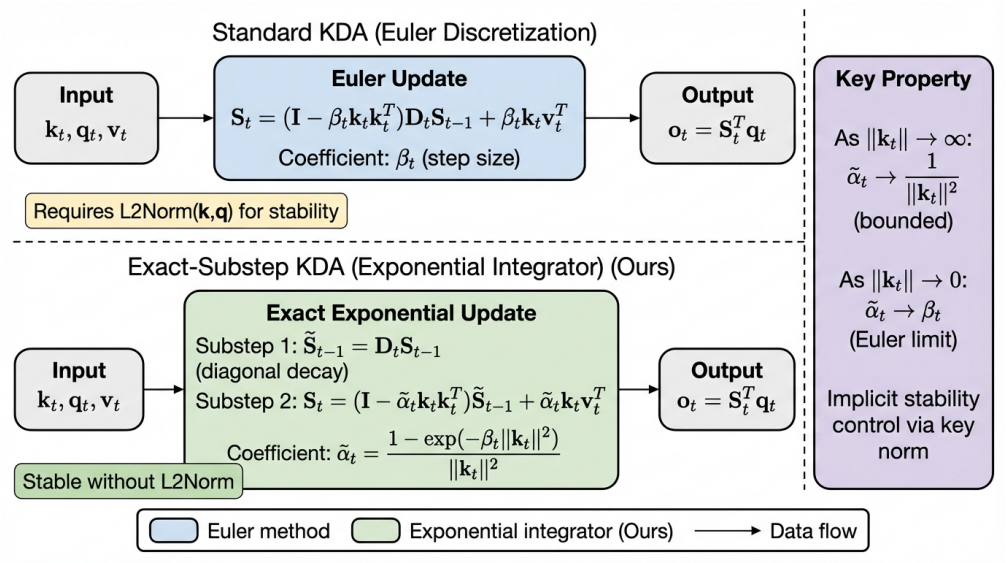

Figure 1: Comparison of Euler discretization (left) and exponential integrator (right) for diagonal-decay delta attention. The exponential integrator replaces the unbounded Euler coefficient $\alpha_t = 1 - \exp(-\|k_t\|^2)$ with a bounded coefficient $\tilde{\alpha}_t = (1 - \exp(-\|k_t\|^2))/\|k_t\|^2$, which approaches $1/\|k_t\|^2$ as $\|k_t\| \to \infty$, providing implicit stability control without requiring L2 normalization.

3 Experiments

3.1 Experimental Setup

We evaluate the exponential integrator on three synthetic long-context tasks designed to stress different aspects of sequence modeling: Palindrome (pattern reversal requiring precise positional memory), MQAR (multi-query associative recall testing key-value retrieval) (Arora et al., 2023), and Stack (LIFO state tracking with multiple stacks) (Jelassi et al., 2024). These tasks are lightweight, reproducible, and directly stress long-range retrieval and state tracking capabilities.

We compare three conditions to isolate the contributions of the exponential integrator and L2 normalization removal: (C1) Euler+L2Norm: standard KDA with Euler discretization and L2-normalized keys/queries (baseline); (C2) ExpInt+L2Norm: exponential integrator with L2 normalization (isolates integrator effect); (C3) ExpInt-noL2Norm: exponential integrator without L2 normalization (proposed method).

All models use a 2-layer, 2-head architecture with $d_k = d_v = 128$ and embedding dimension 256. Training uses AdamW with weight decay 0.01, batch size 64, and learning rate 10^{-3} selected via grid search. Models are trained for 30,000 steps at sequence length $L_{\text{train}} = 1024$ and evaluated at $L_{\text{test}} \in \{1024, 2048, 4096\}$ to assess length extrapolation. We report mean $\pm$ standard deviation over 3 seeds (42, 123, 456).

3.2 Results

Table 1 presents the main experimental results. The key findings are as follows.

Numerical Stability Achieved. All 27 experimental runs (3 conditions $\times$ 3 tasks $\times$ 3 seeds) completed without NaN or divergence failures. The exponential integrator successfully enables stable training without L2 normalization, confirming that the bounded coefficient property provides implicit stability control.

No Accuracy Improvement at Length Extrapolation. Despite achieving numerical stability, the proposed method (C3) does not improve accuracy over the baseline (C1) at the primary evaluation

Table 1: Accuracy (%) comparison across three conditions and three synthetic tasks at different test sequence lengths. Training length $L = 1024$. Results show mean ± std over 3 seeds. **Bold** indicates best per column. The proposed method (C3) achieves numerical stability but does not improve accuracy over the baseline (C1) at length extrapolation.

	Palindrome			MQAR			Stack		
Condition	L=1024	L=2048	L=4096	L=1024	L=2048	L=4096	L=1024	L=2048	L=4096
C1: Euler+L2Norm	11.79±2.13	2.53±0.41	1.81±0.33	99.94±0.05	99.96±0.05	99.95±0.06	98.97±0.31	95.10±0.79	**87.99**±1.71
C2: ExpInt+L2Norm	**16.07**±1.51	1.77±0.31	1.38±0.61	**99.98**±0.01	**99.99**±0.01	99.99±0.01	99.46±0.51	96.46±1.49	85.47±3.87
C3: ExpInt-noL2Norm	14.96±1.56	**1.74**±0.11	**1.00**±0.40	**99.98**±0.00	**99.99**±0.00	**100.00**±0.00	**99.64**±0.31	**97.30**±0.83	85.01±6.83

length $L = 4096$. On Palindrome, C3 achieves 1.00% versus C1's 1.81% (-0.81 percentage points). On Stack, C3 achieves 85.01% versus C1's 87.99% (-2.98 pp). On MQAR, both methods achieve near-perfect accuracy ($\sim$100%), exhibiting a ceiling effect that prevents meaningful comparison. The primary success criterion of $\geq$5 pp improvement on $\geq$2/3 tasks is not met (0/3 tasks show improvement).

Ablation Analysis. Comparing C2 versus C1 isolates the effect of the exponential integrator alone. At $L = 4096$, differences are negligible: Palindrome -0.43 pp, MQAR $+0.04$ pp, Stack -2.52 pp. This indicates that the integrator itself does not improve accuracy when L2 normalization is retained. Comparing C3 versus C2 isolates the effect of removing L2 normalization. Again, differences are minimal: Palindrome -0.38 pp, MQAR $+0.01$ pp, Stack -0.46 pp. The hypothesized "key-norm signal-strength channel" does not materialize as an accuracy benefit.

Increased Variance. The proposed method (C3) exhibits markedly higher variance on the Stack task at $L = 4096$ (std=6.83) compared to the baseline (C1, std=1.71). This is driven by seed 42 producing 75.43% accuracy, approximately 15 pp below the other seeds. This suggests that removing L2 normalization may reduce robustness to initialization, even when the exponential integrator provides numerical stability.

4 Related Work

Linear Attention and Fast Weight Programmers. Linear attention mechanisms replace softmax normalization with feature maps that enable $O(n)$ complexity through associative accumulation of key-value statistics (Choromanski et al., 2020). Schlag et al. (2021) interpret such mechanisms as fast-weight programmers, where the state matrix implements a dynamic associative memory updated via low-rank modifications. Irie et al. (2021) extend this framework with recurrent updates that improve expressivity beyond simple additive accumulation.

State Space Models. Structured state space models (SSMs) offer an alternative approach to efficient sequence modeling by maintaining a fixed-size recurrent state with structured transitions (Gu et al., 2021). Mamba introduces selective state spaces with input-dependent dynamics (Gu & Dao, 2023), while Mamba-2 establishes connections between SSMs and attention through structured state space duality (Dao & Gu, 2024). These models achieve strong performance on long sequences but can struggle with associative recall tasks that require precise key-value retrieval.

Delta-Rule Attention. Delta-rule attention addresses the retrieval limitations of linear attention by incorporating error-correction updates derived from online regression (Yang et al., 2024b). Gated Linear Attention introduces hardware-efficient training algorithms (Yang et al., 2023), while Gated Delta Networks combine delta-rule updates with gating mechanisms for improved performance (Yang et al., 2024a). Kimi Delta Attention (KDA) extends this line with diagonal decay gates and achieves strong results on long-context tasks, though it requires L2 normalization on keys and queries for stability (Zhang et al., 2025).

Continuous-Time Perspectives. Several works interpret discrete sequence models as discretizations of continuous-time dynamics. Neural ODEs provide a foundational framework for continuous-depth networks (Chen et al., 2018). Error-Free Linear Attention (EFLA) shows that for rank-1

delta-rule dynamics, the exact exponential integrator can be computed in closed form, eliminating discretization error (Lei et al., 2025). Our work applies this insight to diagonal-decay delta attention, finding that while the integrator enables stable training without normalization, it does not improve length extrapolation accuracy.

5 Conclusion

We investigated whether replacing Euler discretization with an exact exponential integrator in diagonal-decay delta attention could enable stable training without L2 normalization and improve length extrapolation. While the exponential integrator successfully achieves numerical stability without normalization (0 NaN/divergence across 27 runs), it does not improve accuracy at length extrapolation. Our ablation analysis reveals that neither the integrator alone nor the removal of L2 normalization provides accuracy benefits. This negative result suggests that discretization error is not the primary bottleneck for length extrapolation in delta-rule attention, and that L2 normalization may provide beneficial regularization beyond mere stability. Future work should explore alternative approaches such as architectural modifications, training strategies, or hybrid methods to address the length extrapolation challenge.

References

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher R'e. Zoology: Measuring and improving recall in efficient language models. *ArXiv*, abs/2312.04927, 2023.

T. Chen, Yulia Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. *ArXiv*, abs/1806.07366, 2018.

K. Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. *ArXiv*, abs/2009.14794, 2020.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *ArXiv*, abs/2405.21060, 2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *ArXiv*, abs/2312.00752, 2023.

Albert Gu, Karan Goel, and Christopher R'e. Efficiently modeling long sequences with structured state spaces. *ArXiv*, abs/2111.00396, 2021.

Kazuki Irie, Imanol Schlag, R'obert Csord'as, and J. Schmidhuber. Going beyond linear transformers with recurrent fast weight programmers. *ArXiv*, abs/2106.06295, 2021.

Samy Jelassi, David Brandfonbrener, S. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. pp. 21502–21521, 2024.

Jingdi Lei, Di Zhang, and Soujanya Poria. Error-free linear attention is a free lunch: Exact solution from continuous-time dynamics. *ArXiv*, abs/2512.12602, 2025.

Imanol Schlag, Kazuki Irie, and J. Schmidhuber. Linear transformers are secretly fast weight programmers. pp. 9355–9366, 2021.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *ArXiv*, abs/2312.06635, 2023.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *ArXiv*, abs/2412.06464, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *ArXiv*, abs/2406.06484, 2024b.

Yu Zhang, Zongyu Lin, Xingcheng Yao, Jiaxi Hu, Fanqing Meng, Chengyin Liu, Xin Men, Songlin Yang, Zhiyuan Li, Wentao Li, Enzhe Lu, Weizhou Liu, Yanru Chen, Weixin Xu, Long Yu, Ye-Jia Wang, Yu Fan, Longguang Zhong, Enming Yuan, Dehao Zhang, Yizhi Zhang, T. Y. Liu, Haiming Wang, Shengjun Fang, Weiran He, Shaowei Liu, Yiwei Li, Jianling Su, Jiezhong Qiu, Bo Pang, Junjie Yan, Zhejun Jiang, Weixiao Huang, Bo Yin, Jiacheng You, Chu Wei, Zhengtao Wang, Chao Hong, Yutian Chen, Guanduo Chen, Yucheng Wang, Hua Zheng, Feng Wang, Yibo Liu, Meng xiao Dong, Zheng Zhang, Siyuan Pan, Wenhao Wu, Yuhao Wu, Longyu Guan, Ji-Hua Tao, Guohong Fu, Xinran Xu, Yuzhi Wang, Guokun Lai, Yuxin Wu, Xinyue Zhou, Zhilin Yang, and Yulun Du. Kimi linear: An expressive, efficient attention architecture. *ArXiv*, abs/2510.26692, 2025.