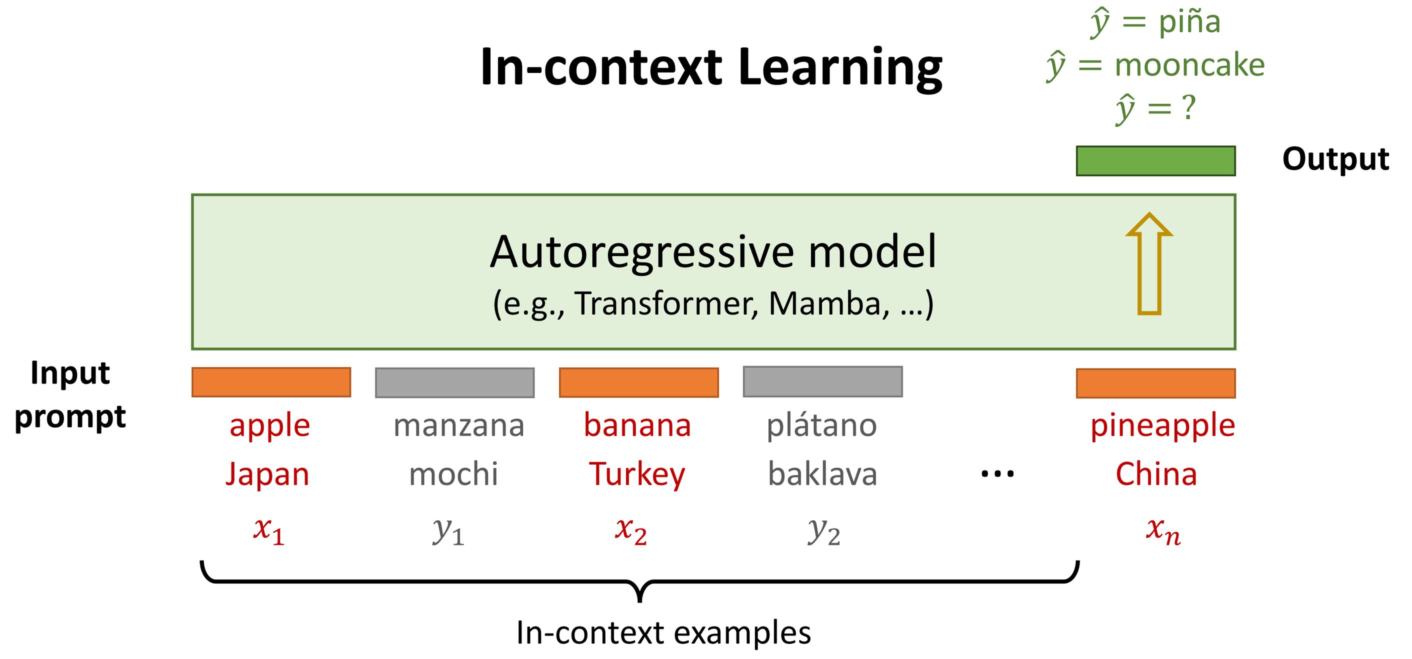
Gating is Weighting: Understanding Gated Linear Attention through In-context Learning

Yingcong Li[⋆,1,5] Davoud Ataee Tarzanagh[⋆,2] Ankit Singh Rawat[3] Maryam Fazel[4] Samet Oymak[1]

Equal contribution[⋆] University of Michigan[1] Samsung SDS Research America[2] Google Research NYC[3] University of Washington[4] New Jersey Institute of Technology[5]

Motivation



Q1: *What optimization algorithms are implemented by different model architectures in ICL?*

Key Findings on ICL Implementations:

Linear attention $\Longrightarrow$ Preconditioned Gradient Descent (PGD)[1,2]

State-space model/H3 $\Longrightarrow$ Sample-weighted PGD (WPGD)[1]

Our work: Gated linear attention $\Longrightarrow$ Data-dependent WPGD (DWPGD)[4]

Q2: *What are **Gated Linear Attention (GLA)** architectures?*

Definition 1 (GLA): Given a sequence of (query, key, value) embeddings $(\boldsymbol{q}_i, \boldsymbol{k}_i, \boldsymbol{v}_i)_{i=1}^n$, the GLA recurrence is given by

$$\boldsymbol{S}_i = \boldsymbol{G}_i \odot \boldsymbol{S}_{i-1} + \boldsymbol{v}_i \boldsymbol{k}_i^\top, \quad \text{and} \quad \boldsymbol{o}_i = \boldsymbol{S}_i \boldsymbol{q}_i, \quad i \in \{1, \ldots, n\}. \tag{GLA}$$

Here, the *gating variable* $\boldsymbol{G}_i$ is applied to the state $\boldsymbol{S}_{i-1}$ through the Hadamard product $\odot$.

Note: When $\boldsymbol{G}_i \equiv \mathbf{1}$, (GLA) reduces to causal linear attention.

Popular GLA Architectures[3]:

Model[3]	Parameterization
Mamba (Gu & Dao, 2023)	$\boldsymbol{G}_t = \exp(-(\mathbf{1}^\top \boldsymbol{\alpha}_t) \odot \exp(A))$, $\boldsymbol{\alpha}_t = \text{softplus}(\boldsymbol{x}_t W_{\alpha_1} W_{\alpha_2})$
Mamba-2 (Dao & Gu, 2024)	$\boldsymbol{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}$, $\gamma_t = \exp(-\text{softplus}(\boldsymbol{x}_t W_\gamma) \exp(a))$
mLSTM (Beck et al., 2024; Peng et al., 2021)	$\boldsymbol{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}$, $\gamma_t = \sigma(\boldsymbol{x}_t W_\gamma)$
Gated Retention (Sun et al., 2024)	$\boldsymbol{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}$, $\gamma_t = \sigma(\boldsymbol{x}_t W_\gamma)^{\frac{1}{\tau}}$
DFW (Mao, 2022; Pramanik et al., 2023)	$\boldsymbol{G}_t = \boldsymbol{\alpha}_t \boldsymbol{\beta}_t^\top$, $\boldsymbol{\alpha}_t = \sigma(\boldsymbol{x}_t W_\alpha)$, $\boldsymbol{\beta}_t = \sigma(\boldsymbol{x}_t W_\beta)$
GateLoop (Katsch, 2023)	$\boldsymbol{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}$, $\boldsymbol{\alpha}_t = \sigma(\boldsymbol{x}_t W_{\alpha_1}) \exp(\boldsymbol{x}_t W_{\alpha_2} \boldsymbol{i})$
HGRN-2 (Qin et al., 2024b)	$\boldsymbol{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}$, $\boldsymbol{\alpha}_t = \gamma + (1-\gamma)\sigma(\boldsymbol{x}_t W_\alpha)$
RWKV-6 (Peng et al., 2024)	$\boldsymbol{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}$, $\boldsymbol{\alpha}_t = \exp(-\exp(\boldsymbol{x}_t W_\alpha))$
Gated Linear Attention (GLA)	$\boldsymbol{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}$, $\boldsymbol{\alpha}_t = \sigma(\boldsymbol{x}_t W_{\alpha_1} W_{\alpha_2})^{\frac{1}{\tau}}$

Our Goal: Develop a **mathematical understanding** of the GLA mechanism through the lens of in-context learning and optimization.

[1] Li Y, Rawat AS, Oymak S. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. NeurIPS 2024.
[2] Ahn K, Cheng X, Daneshmand H, Sra S. Transformers learn to implement preconditioned gradient descent for in-context learning. NeurIPS 2023.
[3] Yang S, Wang B, Shen Y, Panda R, Kim Y. Gated linear attention transformers with hardware-efficient training. ICML 2024.
[4] Li Y, Tarzanagh DA, Fazel M, Oymak S. Gating is weighting: Understanding gated linear attention through in-context learning. COLM 2025.

Main Results

- **Input prompt:** Construct input prompt as follows:

$$\boldsymbol{Z} = [\boldsymbol{z}_1 \ \ldots \ \boldsymbol{z}_n \ \boldsymbol{z}_{n+1}]^\top = \begin{bmatrix} \boldsymbol{x}_1 \ldots \boldsymbol{x}_n & \boldsymbol{x}_{n+1} \\ y_1 \ldots y_n & 0 \end{bmatrix}^\top.$$

Theorem 1: GLA $\Leftrightarrow$ DWPGD

Consider model construction $\boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{P}_k & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{W}_q = \begin{bmatrix} \boldsymbol{P}_q & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{W}_v = \begin{bmatrix} \mathbf{0}_{d\times d} & 0 \\ 0 & 1 \end{bmatrix}$, and take the last coordinate of the last token output denoted by $(\boldsymbol{o}_{n+1})_{d+1}$ as a prediction. Then, we have

$$f_{\texttt{GLA}}(\boldsymbol{Z}) := (\boldsymbol{o}_{n+1})_{d+1} = \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}, \quad \text{where} \quad \hat{\boldsymbol{\beta}} = \boldsymbol{P}_q(\boldsymbol{X}\boldsymbol{P}_k \odot \boldsymbol{\Omega})^\top \boldsymbol{y}.$$

Here, $\boldsymbol{\Omega} = [\boldsymbol{g}_{1:n+1} \ \cdots \ \boldsymbol{g}_{n:n+1}]^\top \in \mathbb{R}^{n\times d}$, where $\boldsymbol{g}_{i:n+1}, i \in [n]$ is given by

$$\boldsymbol{g}_{i:n+1} := \boldsymbol{g}_{i+1} \odot \boldsymbol{g}_{i+2} \cdots \boldsymbol{g}_{n+1} \in \mathbb{R}^d, \quad \text{and} \quad \boldsymbol{G}_i = \begin{bmatrix} * & * \\ \boldsymbol{g}_i^\top & * \end{bmatrix}.$$

Note: When $\boldsymbol{G}_i \equiv \mathbf{1}$, $\hat{\boldsymbol{\beta}} = \boldsymbol{P}_q \boldsymbol{P}_k^\top \boldsymbol{X}^\top \boldsymbol{y}$ reduces to *preconditioned gradient descent (PGD)*.

Optimization Landscape of WPGD

- **Learning problem:** Given $(\boldsymbol{x}, y, \boldsymbol{X}, \boldsymbol{y}) \sim \mathcal{D}$, learn optimal WPGD:

$$\mathcal{L}^\star_{\texttt{WPGD}} := \min_{\boldsymbol{P}\in\mathbb{R}^{d\times d}, \boldsymbol{\omega}\in\mathbb{R}^n} \mathbb{E}\left[\left(y - \boldsymbol{x}^\top \boldsymbol{P} \boldsymbol{X}^\top(\boldsymbol{\omega} \odot \boldsymbol{y})\right)^2\right].$$

- **Data model:** Correlated tasks $\boldsymbol{\beta}_i \sim \mathcal{N}(0, \boldsymbol{I})$ jointly Gaussian, $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, $y_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}_i, \sigma^2)$.

Theorem 2: Stationary Points

Define $h_1 : \mathbb{R}_+ \to \mathbb{R}_+$ and $h_2 : [1, \infty) \to \mathbb{R}_+$ as

$$h_1(\bar{\gamma}) := \left(\sum_{i=1}^n \frac{\lambda_i a_i^2}{(1+\lambda_i\bar{\gamma})^2}\right)\left(\sum_{i=1}^n \frac{a_i^2}{(1+\lambda_i\bar{\gamma})^2}\right)^{-1},$$

$$h_2(\gamma) := \left(1 + M\left(\sum_{i=1}^d \frac{s_i^2}{(M+s_i\gamma)^2}\right)\left(\sum_{i=1}^d \frac{s_i^3}{(M+s_i\gamma)^2}\right)^{-1}\right)^{-1},$$

where $\{s_i\}$, $\{\lambda_i\}$ are eigenvalues of $\boldsymbol{\Sigma}$, $\boldsymbol{R}$; $\{a_i\}$ from $\boldsymbol{r} = \boldsymbol{E}\boldsymbol{a}$; $M = \sigma^2 + \sum_i s_i$.

The stationary point $(\boldsymbol{P}^\star, \boldsymbol{\omega}^\star)$ (up to rescaling) is:

$$\boldsymbol{P}^\star = \boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{\gamma^\star}{M}\cdot\boldsymbol{\Sigma} + \boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}^{-\frac{1}{2}}, \quad \boldsymbol{\omega}^\star = (h_2(\gamma^\star)\cdot\boldsymbol{R} + \boldsymbol{I})^{-1}\boldsymbol{r},$$

where $\gamma^\star$ is a fixed point of $h_1(h_2(\gamma)) + 1$.

Theorem 3: Global Uniqueness

Under mild spectral gap condition $\Delta_{\boldsymbol{\Sigma}} \cdot \Delta_{\boldsymbol{R}} < M + s_{\min}$:

T1. Mapping $h_1(h_2(\gamma)) + 1$ is a contraction with unique fixed point $\gamma^\star$.

T2. Loss has unique global minimum $(\boldsymbol{P}^\star, \boldsymbol{\omega}^\star)$ up to rescaling.

Key insight: Optimal $\boldsymbol{\omega}^\star = (h_2(\gamma^\star)\boldsymbol{R} + \boldsymbol{I})^{-1}\boldsymbol{r}$ depends on task correlations $\boldsymbol{R}$, $\boldsymbol{r}$, enabling context-aware weighting.

Optimization Landscape of GLA

Definition 2(Multi-task Prompt and GLA Objective): Consider prompt with K correlated tasks $(\boldsymbol{\beta}_k)_{k=1}^K$ and one query task $\boldsymbol{\beta}$. For each task $k \in [K]$, a prompt of length n_k is drawn, consisting of IID input-label pairs $\{(\boldsymbol{x}_i^k, y_i^k)\}_{i=1}^{n_k}$. Let $n := \sum_{k=1}^K n_k$.

$$\boldsymbol{Z} = \begin{bmatrix} \underbrace{\begin{matrix} \boldsymbol{x}_1^1 \cdots \boldsymbol{x}_{n_1}^1 & 0 \\ y_1^1 \cdots y_{n_1}^1 & 0 \\ 0 \ \cdots \ 0 & \boldsymbol{c}^1 \end{matrix}}_{\text{task } 1} & \cdots & \underbrace{\begin{matrix} \boldsymbol{x}_1^K \cdots \boldsymbol{x}_{n_K}^K & 0 \\ y_1^K \cdots y_{n_K}^K & 0 \\ 0 \ \cdots \ 0 & \boldsymbol{c}^K \end{matrix}}_{\text{task } K} & \begin{matrix} \boldsymbol{x} \\ 0 \\ 0 \end{matrix} \end{bmatrix}^\top. \tag{1}$$

Here, $\{\boldsymbol{c}^1, \cdots, \boldsymbol{c}^K\}$ are K **linearly independent** contextual features. The GLA optimization problem is described as:

$$\mathcal{L}^\star_{\texttt{GLA}} := \min_{\boldsymbol{P}_k, \boldsymbol{P}_q \in \mathbb{R}^{d\times d}, G \in \mathcal{G}} \mathcal{L}_{\texttt{GLA}}(\boldsymbol{P}_k, \boldsymbol{P}_q, G) \quad \text{where} \quad \mathcal{L}_{\texttt{GLA}}(\boldsymbol{P}_k, \boldsymbol{P}_q, G) = \mathbb{E}\left[\left(y - f_{\texttt{GLA}}(\boldsymbol{Z})\right)^2\right].$$

Here, $G(\cdot)$ represents the gating function and $\mathcal{G}$ denotes the function search space.

Given context examples $\{(\boldsymbol{X}_k, \boldsymbol{y}_k) := (\boldsymbol{x}_i^k, y_i^k)_{i=1}^{n_k}\}_{k=1}^K$, define the concatenated data $(\boldsymbol{X}, \boldsymbol{y})$:

$$\boldsymbol{X} = \left[\boldsymbol{X}_1^\top \cdots \boldsymbol{X}_K^\top\right]^\top \in \mathbb{R}^{n\times d}, \quad \text{and} \quad \boldsymbol{y} = \left[\boldsymbol{y}_1^\top \cdots \boldsymbol{y}_K^\top\right]^\top \in \mathbb{R}^n.$$

Theorem 4 (Optimization Equivalence): GLA $\Leftrightarrow$ DWPGD

Consider GLA with input prompt $\boldsymbol{Z}$ defined in (1). There exists a gating function $G(\cdot)$ such that the optimal risk $\mathcal{L}^\star_{\texttt{GLA}}$ obeys

$$\mathcal{L}^\star_{\texttt{GLA}} = \mathcal{L}^\star_{\texttt{WPGD}}.$$

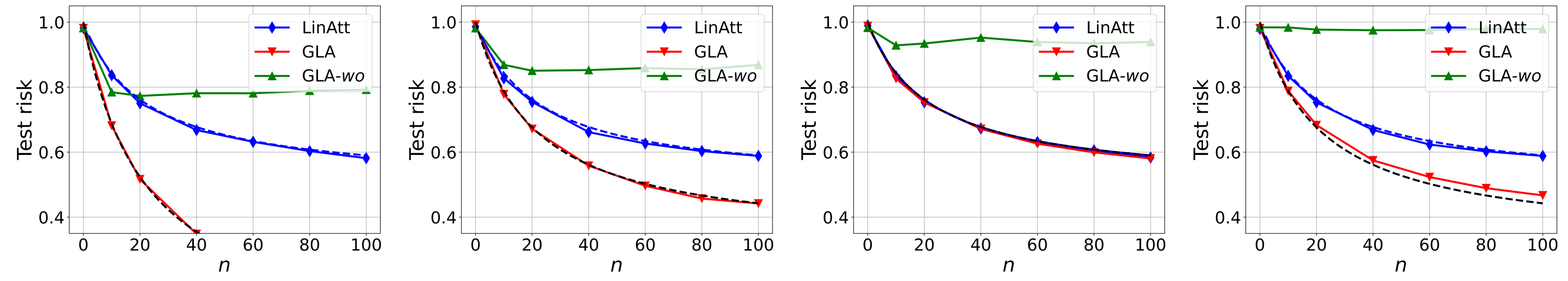
Corollary (Loss landscape of one-layer linear attention): Consider a single-layer linear attention following the same model constructions. Let $\boldsymbol{R}$ and $\boldsymbol{r}$ be the corresponding correlation matrix and vector. Suppose $\boldsymbol{\Sigma} = \boldsymbol{I}$. Then, the optimal risk obeys

$$\mathcal{L}^\star_{\texttt{ATT}} := \min_{\boldsymbol{P}\in\mathbb{R}^{d\times d}} \mathcal{L}_{\texttt{WPGD}}(\boldsymbol{P}, \boldsymbol{\omega} = \mathbf{1}) = d + \sigma^2 - \frac{d(\mathbf{1}^\top \boldsymbol{r})^2}{n(d+\sigma^2+1) + \mathbf{1}^\top \boldsymbol{R}\mathbf{1}}.$$

Experiments

Data setting:

- $K = 2$, $n_1 = n_2 = n/2$, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{I})$
- $(r_1, r_2) = (\text{corr\_coef}(\boldsymbol{\beta}_1, \boldsymbol{\beta}), \text{corr\_coef}(\boldsymbol{\beta}_2, \boldsymbol{\beta}))$ and $\text{corr\_coef}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = 0$.
- $G(\boldsymbol{z}) = \sigma(\boldsymbol{W}_g \boldsymbol{z})\mathbf{1}^\top$ where $\sigma(x) = (1 + e^{-z})^{-1}$ is the activation function.



(a) $(r_1, r_2) = (0, 1)$ (b) $(r_1, r_2) = (0.2, 0.8)$ (c) $(r_1, r_2) = (0.5, 0.5)$ (d) $(r_1, r_2) = (0.8, 0.2)$

- LinAtt: Linear attention; GLA: Gated linear attention; GLA-wo: GLA without contextual features