

Operator Derivation for Gated OJA Rule

Jiaxi Hu, Moonshot AI
hujiaxi@moonshot.cn

<https://github.com/fla-org/flash-linear-attention>

1 QKVO Rotary Position Embedding

The mechanism of multiplicative positional encodings like RoPE can be analyzed through a generalized attention formulation:

$$\mathbf{o}_t = \sum_{j=1}^t \mathbf{v}_j \exp \left((\mathbf{R}_j \mathbf{k}_j)^\top (\mathbf{R}_t \mathbf{q}_t) \right) = \sum_{j=1}^t \mathbf{v}_j \exp \left(\mathbf{k}_j^\top \left(\prod_{s=i+1}^t \mathbf{R}_s \right) \mathbf{q}_t \right) \quad (1)$$

where the position relationship between the t -th query \mathbf{q}_t and the i -th key \mathbf{k}_i is reflected by the cumulative matrix products. RoPE defines the transformation matrix \mathbf{R}_j as a block diagonal matrix composed of $d_k/2$ 2D rotation matrices $\mathbf{R}_j^k = \begin{pmatrix} \cos(j\theta_k) & -\sin(j\theta_k) \\ \sin(j\theta_k) & \cos(j\theta_k) \end{pmatrix}$ with **per-2-dimensional** angular frequency θ_k . Due to the properties of rotation matrices, i.e., $\mathbf{R}_{t-i} = \mathbf{R}_t^\top \mathbf{R}_i$, absolute positional information \mathbf{R}_t and \mathbf{R}_i can be applied separately to \mathbf{q}_t and \mathbf{k}_i , which are then transformed into relative positional information $t - i$ encoded as $\prod_{j=i+1}^t \mathbf{R}_j = \begin{pmatrix} \cos((t-i)\theta_k) & -\sin((t-i)\theta_k) \\ \sin((t-i)\theta_k) & \cos((t-i)\theta_k) \end{pmatrix}$.

What would happen if we applied a rotation matrix to \mathbf{VO} ?

$$\mathbf{o}_t = \mathbf{R}_t^\top \sum_{j=1}^t (\mathbf{R}_j \mathbf{v}_j) \exp(\mathbf{k}_j^\top \mathbf{q}_t) = \sum_{j=1}^t \left(\prod_{s=i+1}^t \mathbf{R}_s \right) \mathbf{v}_j \exp(\mathbf{k}_j^\top \mathbf{q}_t) \quad (2)$$

1.1 Gated Delta Rule as QK Position Encodings

$$\mathbf{o}_t = \sum_{j=1}^t \mathbf{v}_j \left(\mathbf{k}_j^\top \left(\prod_{s=j+1}^t \text{diag}(\boldsymbol{\alpha}_s) (\mathbf{I} - \beta_s \mathbf{k}_s \mathbf{k}_s^\top) \right) \mathbf{q}_t \right) \quad (3)$$

1.2 Gated Oja Rule as VO Position Encodings

$$\mathbf{o}_t = \sum_{j=1}^t \left(\prod_{s=j+1}^t (\mathbf{I} - \beta_s \mathbf{v}_s \mathbf{v}_s^\top) \text{diag}(\boldsymbol{\alpha}_s) \right) \mathbf{v}_j (\mathbf{k}_j^\top \mathbf{q}_t) \quad (4)$$

1.3 Gated Oja Rule with Value Gate

For a standard Gated OjaNet with a diagonal value gate,

$$\mathbf{S}_t = \text{Diag}(\boldsymbol{\alpha}_t) \mathbf{S}_{t-1} + \beta_t \mathbf{v}_t (\mathbf{k}_t^\top - \mathbf{v}_t^\top \text{Diag}(\boldsymbol{\alpha}_t) \mathbf{S}_{t-1}^\top) = (\mathbf{I} - \beta_t \mathbf{v}_t \mathbf{v}_t^\top) \text{Diag}(\boldsymbol{\alpha}_t) \mathbf{S}_{t-1} + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$$

By partially expanding the recurrence, we have

$$\mathbf{S}_{[t]}^r = \underbrace{\left(\prod_{i=1}^r \left(\mathbf{I} - \beta_{[t]}^i \mathbf{v}_{[t]}^i \mathbf{v}_{[t]}^{i\top} \right) \text{Diag}(\boldsymbol{\alpha}_{[t]}^i) \right)}_{:= \mathbf{D}_{[t]}^r \text{ ("pseudo" memory decay)}} \mathbf{S}_{[t]}^0 + \underbrace{\sum_{i=1}^r \left(\left(\prod_{j=i+1}^r \left(\mathbf{I} - \beta_{[t]}^j \mathbf{v}_{[t]}^j \mathbf{v}_{[t]}^{j\top} \right) \text{Diag}(\boldsymbol{\alpha}_{[t]}^j) \right) \beta_{[t]}^i \mathbf{v}_{[t]}^i \mathbf{k}_{[t]}^{i\top} \right)}_{:= \mathbf{H}_{[t]}^r \text{ ("pseudo" Incremental memory)}}$$

Then, we employ the WY representation:

$$\begin{aligned} \mathbf{D}_{[t]}^r &= \text{Diag}(\boldsymbol{\alpha}_{[t]}^{1:r}) - \sum_{i=1}^r \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^i \mathbf{w}_{[t]}^{i\top} & \mathbf{w}_{[t]}^r &= \beta_{[t]}^r \left(\text{Diag}(\boldsymbol{\alpha}_{[t]}^{1:r}) \mathbf{v}_{[t]}^r - \sum_{i=1}^{r-1} \mathbf{w}_{[t]}^i (\mathbf{v}_{[t]}^{i\top} \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^r) \right) \\ \mathbf{H}_{[t]}^r &= \sum_{i=1}^r \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^i \mathbf{u}_{[t]}^{i\top} & \mathbf{u}_{[t]}^r &= \beta_{[t]}^r \left(\mathbf{k}_{[t]}^r - \sum_{i=1}^{r-1} \mathbf{u}_{[t]}^i (\mathbf{v}_{[t]}^{i\top} \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^r) \right) \end{aligned}$$

To maximize hardware efficiency, we apply the UT transform to reduce non-matmul FLOPs, which is crucial to enable better hardware utilization during training.

$$\begin{aligned} \mathbf{W}_{[t]} &= \mathbf{M}_{[t]} \text{Diag}(\beta_{[t]}^{1 \rightarrow C}) \left(\mathbf{A}_{[t]}^{1 \rightarrow C} \odot \mathbf{V}_{[t]} \right), & \mathbf{U}_{[t]} &= \mathbf{M}_{[t]} \text{Diag}(\beta_{[t]}^{1 \rightarrow C}) \mathbf{K}_{[t]} \\ \mathbf{M}_{[t]} &= \left(\mathbf{I} + \text{lower} \left(\text{Diag}(\beta_{[t]}^{1 \rightarrow C}) \left(\mathbf{V}_{[t]} \odot \mathbf{A}_{[t]}^{1 \rightarrow C} \right) \left(\frac{\mathbf{V}_{[t]}^\top}{\mathbf{A}_{[t]}^{1 \rightarrow C}} \right) \right) \right)^{-1} \end{aligned}$$

Then we have the following vector form:

$$\begin{aligned} \mathbf{S}_{[t]}^r &= \mathbf{D}_{[t]}^r \mathbf{S}_{[t]}^0 + \mathbf{H}_{[t]}^r = \text{Diag}(\boldsymbol{\alpha}_{[t]}^{1:r}) \mathbf{S}_{[t]}^0 + \sum_{i=1}^r \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^i \left(\mathbf{u}_{[t]}^i - (\mathbf{w}_{[t]}^{i\top} \mathbf{S}_{[t]}^0) \right) \\ \mathbf{o}_{[t]}^r &= \mathbf{S}_{[t]}^r \mathbf{q}_{[t]}^r = \text{Diag}(\boldsymbol{\alpha}_{[t]}^{1:r}) \mathbf{S}_{[t]}^0 \mathbf{q}_{[t]}^r + \sum_{i=1}^r \text{Diag}(\boldsymbol{\alpha}_{[t]}^{i:r}) \mathbf{v}_{[t]}^i \left(\mathbf{u}_{[t]}^i - (\mathbf{w}_{[t]}^{i\top} \mathbf{S}_{[t]}^0) \right) \mathbf{q}_{[t]}^r \end{aligned}$$

Equivalently, in matrix form:

$$\begin{aligned} \mathbf{S}_{[t+1]} &= \text{Diag}(\boldsymbol{\alpha}_{[t]}^{1:C}) \mathbf{S}_{[t]} + \left(\mathbf{A}_{[t]}^{i \rightarrow C} \odot \mathbf{V}_{[t]} \right)^\top (\mathbf{U}_{[t]} - \mathbf{W}_{[t]} \mathbf{S}_{[t]}) \\ \textbf{VO-PE: } \mathbf{O}_{[t]} &= \mathbf{A}_{[t]}^{1 \rightarrow C} \odot \left(\mathbf{Q}_{[t]} \mathbf{S}_{[t]}^\top + \text{Tril} \left(\mathbf{Q}_{[t]} \underbrace{(\mathbf{U}_{[t]} - \mathbf{W}_{[t]} \mathbf{S}_{[t]})^\top}_{\text{"pseudo"-key term}} \right) \frac{\mathbf{V}_{[t]}}{(\mathbf{A}_{[t]}^{1 \rightarrow C})} \right) \end{aligned}$$