Backward Pass - DeltaNet

Ryan Chu

forward:

① $G = KK^T$
$A = I + \mathrm{tril}(\mathrm{diag}(b)G, -1)$

② $B_K = \mathrm{diag}(b)K$, $B_V = \mathrm{diag}(b)V$

③ $W = A^{-1}B_K$, $U = A^{-1}B_V$

④ $X = U - WS^T$

⑤ $P = (QK^T) \odot M$

⑥ $O = QS^T + PX$

⑦ $S^+ = S + X^T K$

let $\bar{x}$ be $\frac{\partial L}{\partial x}$, upstream $\bar{S}^+$, $\bar{O}$

⑦ $\bar{S} \mathrel{+}= \bar{S}^+$, $\bar{X} \mathrel{+}= K(\bar{S}^+)^T$, $\bar{K} \mathrel{+}= X\bar{S}^+$

⑥ $\bar{Q} \mathrel{+}= \bar{O}S$, $\bar{S} \mathrel{+}= \bar{O}^T Q$
$\bar{P} \mathrel{+}= \bar{O}X^T$, $\bar{X} \mathrel{+}= P^T\bar{O}$

⑤ let $T = QK^T \rightarrow \bar{T} = \bar{P} \odot M$
$\bar{Q} \mathrel{+}= \bar{T}K$, $\bar{K} \mathrel{+}= \bar{T}^T Q$

④ $\bar{U} \mathrel{+}= \bar{X}$, $\bar{W} \mathrel{+}= -\bar{X}S$, $\bar{S} \mathrel{+}= -\bar{X}^T W$

③ $\bar{B}_K = A^{-T}\bar{W}$, $\bar{B}_V = A^{-T}\bar{U}$
$\bar{A} \mathrel{+}= -A^{-T}\bar{W}W^T - A^{-T}\bar{U}U^T$

$Y = A^{-1}B \rightarrow AY = B$

$A(dY) + (dA)Y = dB$

$dY = -A^{-1}(dA)Y + A^{-1}dB$

$G := \frac{dL}{dY}$ By Riesz, $dL = \langle G, dY \rangle = \mathrm{tr}(G^T dY)$

$dL = \mathrm{tr}\left(G^T\left[-A^{-1}(dA)Y + A^{-1}dB\right]\right)$

$= -\mathrm{tr}(G^T A^{-1}(dA)Y) + \mathrm{tr}(G^T A^{-1} dB)$

$\rightarrow -\mathrm{tr}((A^{-T}G)^T (dA) Y)$

$= -\mathrm{tr}(Y(A^{-T}G)^T(dA))$ [*]

$\therefore \frac{dL}{dB} = A^{-T}G$

$\frac{dL}{dA} = -A^{-T}GY^T$ $\quad = -\mathrm{tr}((A^{-T}GY^T)^T(dA))$

lower mask = $L_{rs} = \mathbb{1}[r>s]$

$A_{rs} = \delta_{rs} + L_{rs}(b_r G_{rs})$

for A on diagonal or upper is constant so those $\bar{b}$ & $\bar{G} = 0$

$r>s$: $A_{rs} = b_r G_{rs}$

$G = KK^T$ $\quad dG = dK K^T + K dK^T$

$\bar{G} := \frac{dL}{dG}$ $\quad dL = \langle \bar{G}, dG \rangle$

$= \langle \bar{G}, dK K^T \rangle + \langle \bar{G}, K dK^T \rangle$

$= \langle \bar{G}K, dK \rangle + \langle \bar{G}^T K, dK \rangle$

$= \langle (\bar{G} + \bar{G}^T)K, dK \rangle$

[*]: trace is cyclic

② $B_K \in \mathbb{R}^{c \times D}$ $\quad (B_K)_{id} = b_i K_{id}$

$\bar{K}_{id} \mathrel{+}= b_i(\bar{B}_K)_{id}$

$\bar{V}_{id} \mathrel{+}= b_i(\bar{B}_V)_{id}$

$\bar{b}_i \mathrel{+}= \sum_{d=1}^{D}\left((\bar{B}_K)_{id}K_{id} + (\bar{B}_V)_{id}V_{id}\right)$

① $\bar{b}_i \mathrel{+}= \sum_{s<i} \bar{A}_{is}G_{is}$

$\bar{G}_{is} = \bar{A}_{is}\frac{\partial A_{is}}{\partial G_{is}} = \begin{cases} \bar{A}_{is}b_i, & i>s \\ 0, & \text{o/w} \end{cases} = \mathrm{tril}(\mathrm{diag}(b)\bar{A}, -1)$

$\bar{K} \mathrel{+}= (\bar{G} + \bar{G}^T)K$

update $\bar{S}^+ \leftarrow \bar{S}$, then go previous chunk