
Brouillon 2

Léo Burgund

May 13, 2025

1 Problem

Goal:

$$\min_f \mathcal{L}(f).$$

We will study the variations of the loss made by the variations of S , with other parameters fixed. Hence we will study

$$\arg \min_S \mathcal{L}(S)$$

with

$$S = XW_Q W_K^\top X^\top$$

First order approximation:

$$\mathcal{L}(S + dS) = \mathcal{L}(S) + \langle G, dS \rangle + o(\|dS\|)$$

with $G = \nabla_S \mathcal{L}(S)$, and

$$\begin{aligned} dS &= X(W_Q + dW_Q)(W_K + dW_K)^\top X^\top - XW_Q W_K^\top X^\top \\ &= XW_Q dW_K^\top X^\top + X dW_Q W_K^\top X^\top + X dW_Q dW_K^\top X^\top \\ &= X(W_Q dW_K^\top + dW_Q W_K^\top)X^\top + o(\|dW_Q\| \cdot \|dW_K\|) \end{aligned} \quad (1.1)$$

🔥 Link between the two approximations $o(\|dS\|)$ and $o(\|dW_Q\| \cdot \|dW_K\|)$, is it okay to do the later as we did the former?

We will consider that $dS = X(W_Q dW_K^\top + dW_Q W_K^\top)X^\top$.

—
We will attempt to resolve the following problem:

$$\arg \min_{dS} \langle G, dS \rangle \quad \text{s.t. } \|dS\| \leq \gamma$$

with $\gamma \in \mathbb{R}_+$.

γ is similar to the learning rate, and constrains dS to respect the first order approximation.

🔥 Then γ must always be small? How small?

—
The solution dS has a norm $\|dS\| = \gamma$ when there exists a dS such that $\langle G, dS \rangle \leq 0$.

We make the hypothesis that we can always find such a dS .

We then have the following problem:

$$\begin{aligned} &\arg \min_{dS} \langle G, dS \rangle \quad \text{s.t. } \|dS\| = \gamma \\ &\left(\Leftrightarrow \gamma \cdot \arg \min_{dS} \langle G, dS \rangle \quad \text{s.t. } \|dS\| = 1 \right) \end{aligned} \quad (1.2)$$

We have

$$\begin{aligned}
\langle G, dS \rangle &= \langle G, X(W_Q dW_K^\top + dW_Q W_K^\top) X^\top \rangle \\
&= \langle X^\top G X, W_Q dW_K^\top + dW_Q W_K^\top \rangle \quad , \text{ let } T = X^\top G X \\
&= \langle T, W_Q dW_K^\top \rangle + \langle T, dW_Q W_K^\top \rangle \\
&= \langle dW_Q, T W_K \rangle + \langle dW_K, T^\top W_Q \rangle
\end{aligned}$$

Linear in dW_Q, dW_K .

The problem now is

$$\arg \min_{dW_Q, dW_K} \langle dW_Q, T W_K \rangle + \langle dW_K, T^\top W_Q \rangle \quad \text{s.t.} \quad \|X(W_Q dW_K^\top + dW_Q W_K^\top) X^\top\| = \gamma$$

Hence the “raw directions” of steepest descent to minimize the scalar products are

$$\Delta W_Q^{(0)} = -T W_K$$

$$\Delta W_K^{(0)} = -T^\top W_Q$$

We define the linear operator

$$\mathcal{A}(\Delta W_Q^{(0)}, \Delta W_K^{(0)}) := X(W_Q \Delta W_K^\top + \Delta W_Q W_K^\top) X^\top$$

and

$$dS^{(0)} := \mathcal{A}(\Delta W_Q^{(0)}, \Delta W_K^{(0)}), \quad \rho := \|dS^{(0)}\|_F$$

We make the hypothesis that $\rho \neq 0$, as we just have to skip the update if it is 0.

We define

$$\alpha := \frac{\gamma}{\rho}$$

and

$$\Delta W_Q := \alpha \Delta W_Q^{(0)}, \quad \Delta W_K := \alpha \Delta W_K^{(0)}$$

We then have

$$\|\mathcal{A}(\Delta W_Q, \Delta W_K)\|_F = \alpha \rho = \gamma$$

so the pair $\Delta W_Q, \Delta W_K$ have the best minimizing direction of problem (1.2), while respecting the norm constraint.