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:

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

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

$$\begin{aligned}
\mathrm{d}S &= X\big(W_Q + \mathrm{d}W_Q\big)(W_K + \mathrm{d}W_K)^\top X^\top - XW_Q W_K^\top X^\top \\
&= XW_Q \,\mathrm{d}W_K^\top X^\top + X\,\mathrm{d}W_Q W_K^\top X^\top + X\,\mathrm{d}W_Q \,\mathrm{d}W_K^\top X^\top \\
&= X\big(W_Q \,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\big)X^\top + o\big(\|\mathrm{d}W_Q\| \cdot \|\mathrm{d}W_K\|\big)
\end{aligned} \tag{1.1}$$

🔥 Link between the two approximations $o(\|\mathrm{d}S\|)$ and $o\big(\|\mathrm{d}W_Q\| \cdot \|\mathrm{d}W_K\|\big)$, is it okay to do the later as we did the former?

We will consider that $\mathrm{d}S = X\big(W_Q \,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\big)X^\top$.

—

We will attempt to resolve the following problem:

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

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

γ is similar to the learning rate, and constrains $\mathrm{d}S$ to respect the first order approximation.

🔥 Then γ must always be small? How small?

—

The solution $\mathrm{d}S$ has a norm $\|\mathrm{d}S\| = \gamma$ when there exists a $\mathrm{d}S$ such that $\langle G, \mathrm{d}S \rangle \leq 0$.

We make the hypothesis that we can always find such a $\mathrm{d}S$.

We then have the following problem:

$$\begin{gathered}
\arg\min_{\mathrm{d}S} \langle G, \mathrm{d}S \rangle \quad \text{s.t. } \|\mathrm{d}S\| = \gamma \\
\Big(\Longleftrightarrow \gamma \cdot \arg\min_{\mathrm{d}S} \langle G, \mathrm{d}S \rangle \ \text{ s.t. } \|\mathrm{d}S\| = 1\Big)
\end{gathered} \tag{1.2}$$

We have

$$
\begin{aligned}
\langle G, \mathrm{d}S \rangle &= \left\langle G, X\left(W_Q \,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right) X^\top \right\rangle \\
&= \left\langle X^\top G X, W_Q \,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top \right\rangle \quad \text{, let } T = X^\top G X \\
&= \left\langle T, W_Q \,\mathrm{d}W_K^\top \right\rangle + \left\langle T, \mathrm{d}W_Q W_K^\top \right\rangle \\
&= \left\langle \mathrm{d}W_Q, T W_K \right\rangle + \left\langle \mathrm{d}W_K, T^\top W_Q \right\rangle
\end{aligned}
$$

Linear in $\mathrm{d}W_Q, \mathrm{d}W_K$.

The problem now is

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

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

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

and

$$
\mathrm{d}S^{(0)} := \mathcal{A}\left(\Delta W_Q^{(0)}, \Delta W_K^{(0)}\right), \;\rho := \left\| \mathrm{d}S^{(0)} \right\|_F
$$

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

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

We then have

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

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

We the have the closed form expressions

$$
\begin{aligned}
\rho &= X\left(W_Q \left(-T^\top W_Q\right)^\top - T W_K W_K^\top\right) X^\top \\
&= -X\left(W_Q W_Q^\top T + T W_K W_K^\top\right) X^\top \\
\Delta W_Q^\star &= -\frac{\gamma}{\rho} T W_K \\
\Delta W_K^\star &= -\frac{\gamma}{\rho} T^\top W_Q
\end{aligned}
$$