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:

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

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" to minimize the scalar products are

$$\Delta W_Q = -T W_K$$

$$\Delta W_K = -T^\top W_Q$$

Finally, we have: (🔥 verify, circular equation?)

$$\mathrm{d}W_Q^* = -\alpha T W_K$$

$$\mathrm{d}W_K^* = -\alpha T^\top W_Q$$

$$\alpha = \frac{\gamma}{\left\|X\left(W_Q\,\mathrm{d}W_K^{*\top} + \mathrm{d}W_Q^* W_K^\top\right)X^\top\right\|}$$