
Brouillon 2

Léo Burgund

May 19, 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

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

We define

$$\begin{aligned} dS_{\text{linear}} &:= X(W_Q dW_K^\top + dW_Q W_K^\top)X^\top \\ dS_{\text{full}} &:= XW_Q dW_K^\top X^\top + X dW_Q W_K^\top X^\top + X dW_Q dW_K^\top X^\top \end{aligned}$$

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.

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:

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

1.1 Linear approach, $dS = dS_{\text{linear}}$

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

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

🔥 The following is false, to change..

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 for the problem (1.2), while respecting the norm constraint.

We then have the closed form expressions

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

1.2 Quadratic approach, $dS = dS_{\text{full}}$

We can define

$$\begin{aligned} dS(x) &= X(W_Q + x dW_Q)(W_K + x dW_K)^\top X^\top - XW_QW_K^\top X^\top \\ &= X(xW_Q dW_K^\top + x dW_Q W_K^\top + x^2 dW_Q dW_K^\top)X^\top \end{aligned}$$

Using first order approximation, should we study: (with $G = \nabla_S \mathcal{L}(S)$)

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

1.3 Problem A

We have $X \in \mathbb{R}^{d_s \times d_e}$, $G = \nabla_S \mathcal{L}(S) \in \mathbb{R}^{d_s \times d_s}$, W_Q and $W_K \in \mathbb{R}^{d_e \times d_k}$, $d_e > d_k$, $\gamma \in (0, \infty)$.

The problem is:

$$\arg \min_{\gamma, dS(\gamma)} \langle G, dS(\gamma) \rangle$$

We have

$$\begin{aligned} \langle G, dS(\gamma) \rangle &= \langle G, X(\gamma W_Q dW_K^\top + \gamma dW_Q W_K^\top + \gamma^2 dW_Q dW_K^\top)X^\top \rangle \\ &= \gamma \langle X^\top G X, W_Q dW_K^\top + dW_Q W_K^\top + \gamma dW_Q dW_K^\top \rangle \end{aligned}$$

Let $T = X^\top G X$, $R(\gamma) = W_Q dW_K^\top + dW_Q W_K^\top + \gamma dW_Q dW_K^\top$

We have $\text{rank}(R(\gamma)) = d_k < \text{rank}(T)$

The problem now is:

$$\arg \min_{\gamma, R(\gamma)} \gamma \langle T, R(\gamma) \rangle \text{ s.t. } \text{rank}(T) > \text{rank}(R(\gamma))$$

1.4 Problem B

We have a self-attention block. X is the input, d_s the sequence length, d_e the embedding size, d_k the key/query size.

We have $X \in \mathbb{R}^{d_s \times d_e}$, $G = \nabla_S \mathcal{L}(S) \in \mathbb{R}^{d_s \times d_s}$, W_Q and $W_K \in \mathbb{R}^{d_e \times d_k}$, $d_e > d_k$.

The idea is start with a low d_k hence low expressivity, and “grow new neurons”, by increasing d_k by p .

Let $Z' = (W_Q + \gamma dW_Q)(W_K + \gamma dW_K)^\top$, $\text{rank}(Z') = d_k$.

We want to find the augmented matrix Z , such that $\text{rank}(Z) = d_k + p$. We basically concatenate p new columns to the matrices $(W_Q + \gamma dW_Q)$ and $(W_K + \gamma dW_K)$, to augment their expressive possibility.

 Question: What would be the best expression for Z , to respect the previously introduced “step” γ ?

$$Z = [W_Q + \gamma dW_Q \mid \gamma W_Q^{\text{new}}][W_K + \gamma dW_K \mid \gamma W_K^{\text{new}}]^\top?$$

 Would augmenting Z' into Z cause problems with the first order approximation?

The problem is:

$$\arg \min_Z \langle X^\top G X, Z - W_Q W_K^\top \rangle$$

1.5 Study of dS

1.5.1 Brouillon: Searching for bounds

We can find an upper bound for the quadratic term, we have, according to [Lemma 1.2](#):

$$\|dS_{\text{quad}}\|_F := \|X dW_Q dW_K^\top X^\top\|_F \leq \|X\|_2 \|dW_Q dW_K^\top\|_F$$

we have

$$\|dW_Q dW_K^\top\|_F \leq \|dW_Q\|_F \|dW_K^\top\|_2 = \|dW_Q\|_F \|dW_K\|_2 \leq \|dW_Q\|_F \|dW_K\|_F$$

hence,

$$\|dS_{\text{quad}}\|_F \leq \|X\|_2 \|dW_Q\|_F \|dW_K\|_F$$

We also have an upper bound for the linear term, (useless?)

$$\begin{aligned} \|dS_{\text{linear}}\|_F &:= \|X(W_Q dW_K^\top + dW_Q W_K^\top)X^\top\|_F \leq \|XW_Q dW_K^\top X^\top\|_F + \|X dW_Q W_K^\top X^\top\|_F \\ &\leq \|X\|_2 \|W_Q\|_F \|dW_K\|_F + \|X\|_2 \|dW_Q\|_F \|W_K\|_F \\ &\leq \|X\|_2 \left(\|W_Q\|_F \|dW_K\|_F + \|dW_Q\|_F \|W_K\|_F \right) \end{aligned}$$

And a lower bound,

$$\|dS_{\text{linear}}\| \geq \left| \|XW_Q dW_K^\top X^\top\|_F - \|X dW_Q W_K^\top X^\top\|_F \right|$$

Hence

$$\frac{\|dS_{\text{quad}}\|}{\|dS_{\text{linear}}\|} \leq \frac{\|X\|_2 \|dW_Q\|_F \|dW_K\|_F}{\left| \left(\|XW_Q dW_K^\top X^\top\|_F - \|X dW_Q W_K^\top X^\top\|_F \right) \right|}$$

1.5.2 Brouillon: Other bound attempt

Applying [Lemma 1.2](#), we have

$$\frac{\sigma_{\min}(X)^2}{\sigma_{\max}(X)^2} \frac{\|dW_Q dW_K^\top\|_F}{\|W_Q dW_K^\top + dW_Q W_K^\top\|_F} \leq \frac{\|dS_{\text{quad}}\|}{\|dS_{\text{linear}}\|} \leq \frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^2} \frac{\|dW_Q dW_K^\top\|_F}{\|W_Q dW_K^\top + dW_Q W_K^\top\|_F}$$

Hence if $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ are close,

$$\frac{\|dS_{\text{quad}}\|}{\|dS_{\text{linear}}\|} \approx \frac{\|dW_Q dW_K^\top\|_F}{\|W_Q dW_K^\top + dW_Q W_K^\top\|_F}$$

1.5.3 Direct form

We also have the direct form

$$\frac{\|dS_{\text{quad}}\|}{\|dS_{\text{linear}}\|} = \frac{\|X dW_Q dW_K^\top X^\top\|_F}{\|X(W_Q dW_K^\top + dW_Q W_K^\top)X^\top\|_F}$$

We can consider two different approaches, either picking dS_{full} or dS_{linear} .

🔥 How and when to choose dS_{full} or dS_{linear} ?

Appendix A

Lemma 1.1. *Let $M \in \mathbb{R}^{m \times n}$, $\sigma_1 \geq \dots \geq \sigma_{\min(m,n)}$ its singular in decreasing order, and $M = U\Sigma V^\top$ its SVD decomposition.*

$$\begin{aligned}\|M\|_F^2 &= \text{tr}(MM^\top) = \text{tr}(U\Sigma V^\top V\Sigma^\top U^\top) = \text{tr}(U\Sigma\Sigma^\top U^\top) = \text{tr}(U^\top U\Sigma\Sigma^\top) \\ &= \text{tr}(\Sigma\Sigma^\top) = \|\Sigma\|_F^2 = \sum_i^{\min(m,n)} \sigma_i^2 \\ &\geq \sigma_1 = \|M\|_2^2\end{aligned}$$

Hence $\|M\|_F \geq \|M\|_2$.

Lemma 1.2. *We know (bound on the Rayleigh quotient) that for any symmetric positive semidefinite matrix M and any vector x ,*

$$x^\top M x \leq \lambda_{\max}(M) \|x\|^2.$$

Let $A \in \mathbb{R}^{m \times p}$, $B \in \mathbb{R}^{p \times n}$, (a_1, \dots, a_m) the row vectors of A .

Then

$$\begin{aligned}\|AB\|_F^2 &= \text{tr}(A(BB^\top)A^\top) \\ &= \sum_{i=1}^m a_i(BB^\top)a_i^\top \\ &\leq \sum_{i=1}^m \lambda_{\max}(BB^\top) \|a_i\|^2 = \lambda_{\max}(BB^\top) \text{tr}(AA^\top) = \|B\|_2^2 \|A\|_F^2\end{aligned}$$

Hence $\|AB\|_F \leq \|B\|_2 \|A\|_F$.

We can prove the same way that $\|AB\|_F \geq \sigma_{\min}(B) \|A\|_F$.

The same reasoning can be applied to prove $\|AB\|_F \leq \|A\|_2 \|B\|_F$.

Moreover, let $C \in \mathbb{R}^{n \times o}$, we have

$$\|ABC\|_F = \|(AB)C\|_F \leq \|AB\|_F \|C\|_2 \leq \|A\|_2 \|B\|_F \|C\|_2.$$