Brouillon 2

Léo Burgund

May 25, 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

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

We define

$$\begin{aligned} \mathrm{d}S_{\text{linear}} &:= X\big(W_Q \,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\big)X^\top \\ \mathrm{d}S_{\text{full}} &:= XW_Q \,\mathrm{d}W_K^\top X^\top + X \,\mathrm{d}W_Q W_K^\top X^\top + X \,\mathrm{d}W_Q \,\mathrm{d}W_K^\top X^\top \end{aligned}$$

—

We will attempt to resolve the following problem:

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

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

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

—

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

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

We then have the following problem:

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

1.1 Linear approach, $\mathrm{d}S = \mathrm{d}S_{\text{linear}}$

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

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

🔥 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

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

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

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

$$\Delta W_Q^\star = -\frac{\gamma}{\rho} T W_K$$

$$\Delta W_K^\star = -\frac{\gamma}{\rho} T^\top W_Q$$

1.2 Quadratic approach, $\mathrm{d}S = \mathrm{d}S_{\text{full}}$

We can define

$$\begin{aligned}\mathrm{d}S(x) &= X\big(W_Q + x\,\mathrm{d}W_Q\big)\big(W_K + x\,\mathrm{d}W_K\big)^\top X^\top - XW_QW_K^\top X^\top \\ &= X\big(xW_Q\,\mathrm{d}W_K^\top + x\,\mathrm{d}W_QW_K^\top + x^2\,\mathrm{d}W_Q\,\mathrm{d}W_K^\top\big)X^\top\end{aligned}$$

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

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

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:

$$\arg\min_{\gamma,\,\mathrm{d}S(\gamma)} \langle G, \mathrm{d}S(\gamma)\rangle$$

We have

$$\begin{aligned}\langle G, \mathrm{d}S(\gamma)\rangle &= \big\langle G, X\big(\gamma W_Q\,\mathrm{d}W_K^\top + \gamma\,\mathrm{d}W_QW_K^\top + \gamma^2\,\mathrm{d}W_Q\,\mathrm{d}W_K^\top\big)X^\top\big\rangle \\ &= \gamma\big\langle X^\top GX, W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_QW_K^\top + \gamma\,\mathrm{d}W_K\,\mathrm{d}W_K^\top\big\rangle\end{aligned}$$

Let $T = X^\top GX$, $R(\gamma) = W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_QW_K^\top + \gamma\,\mathrm{d}W_K\,\mathrm{d}W_K^\top$

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

The problem now is:

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

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 $Z' = \big(W_Q + \gamma\,\mathrm{d}W_Q\big)\big(W_K + \gamma\,\mathrm{d}W_K\big)^\top$, $\operatorname{rank}(Z') = d_k$.

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

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

$$Z = \Big[W_Q + \gamma\,\mathrm{d}W_Q \mid \gamma W_Q^{\text{new}}\Big]\big[W_K + \gamma\,\mathrm{d}W_K \mid \gamma W_K^{\text{new}}\big]^\top ?$$

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

The problem is:

$$\arg\min_Z \big\langle X^\top GX, Z - W_QW_K^\top\big\rangle$$

1.5 Study of $\mathrm{d}S$

1.5.1 Brouillon: Searching for bounds

We can find an upper bound for the quadratic term, we have, according to Lemma 1.2:

$$\left\|\mathrm{d}S_{\text{quad}}\right\|_F := \left\|X\,\mathrm{d}W_Q\,\mathrm{d}W_K^\top X^\top\right\|_F \le \|X\|_2\left\|\mathrm{d}W_Q\,\mathrm{d}W_K^\top\right\|_F$$

we have

$$\left\|\mathrm{d}W_Q\,\mathrm{d}W_K^\top\right\|_F \le \left\|\mathrm{d}W_Q\right\|_F\left\|\mathrm{d}W_K^\top\right\|_2 = \left\|\mathrm{d}W_Q\right\|_F\|\mathrm{d}W_K\|_2 \le \left\|\mathrm{d}W_Q\right\|_F\|\mathrm{d}W_K\|_F$$

hence,

$$\left\|\mathrm{d}S_{\text{quad}}\right\|_F \le \|X\|_2\left\|\mathrm{d}W_Q\right\|_F\|\mathrm{d}W_K\|_F$$

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

$$\begin{aligned}\|\mathrm{d}S_{\text{linear}}\|_F := \left\|X\left(W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right)X^\top\right\|_F &\le \left\|XW_Q\,\mathrm{d}W_K^\top X^\top\right\|_F + \left\|X\,\mathrm{d}W_Q W_K^\top X^\top\right\|_F \\ &\le \|X\|_2\left\|W_Q\right\|_F\|\mathrm{d}W_K\|_F + \|X\|_2\left\|\mathrm{d}W_Q\right\|_F\|W_K\|_F \\ &\le \|X\|_2\left(\left\|W_Q\right\|_F\|\mathrm{d}W_K\|_F + \left\|\mathrm{d}W_Q\right\|_F\|W_K\|_F\right)\end{aligned}$$

And a lower bound,

$$\|\mathrm{d}S_{\text{linear}}\| \ge \left|\left\|XW_Q\,\mathrm{d}W_K^\top X^\top\right\|_F - \left\|X\,\mathrm{d}W_Q W_K^\top X^\top\right\|_F\right|$$

Hence

$$\frac{\left\|\mathrm{d}S_{\text{quad}}\right\|}{\|\mathrm{d}S_{\text{linear}}\|} \le \frac{\|X\|_2\left\|\mathrm{d}W_Q\right\|_F\|\mathrm{d}W_K\|_F}{\left|\left(\left\|XW_Q\,\mathrm{d}W_K^\top X^\top\right\|_F - \left\|X\,\mathrm{d}W_Q W_K^\top X^\top\right\|_F\right)\right|}$$

1.5.2 Brouillon: Other bound attempt

Applying Lemma 1.2, we have

$$\frac{\sigma_{\min}(X)^2}{\sigma_{\max}(X)^2}\frac{\left\|\mathrm{d}W_Q\,\mathrm{d}W_K^\top\right\|_F}{\left\|W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right\|_F} \le \frac{\left\|\mathrm{d}S_{\text{quad}}\right\|}{\|\mathrm{d}S_{\text{linear}}\|} \le \frac{\sigma_{\max}(X)^2}{\sigma_{\min}(X)^2}\frac{\left\|\mathrm{d}W_Q\,\mathrm{d}W_K^\top\right\|_F}{\left\|W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right\|_F}$$

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

$$\frac{\left\|\mathrm{d}S_{\text{quad}}\right\|}{\|\mathrm{d}S_{\text{linear}}\|} \approx \frac{\left\|\mathrm{d}W_Q\,\mathrm{d}W_K^\top\right\|_F}{\left\|W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right\|_F}$$

1.5.3 Direct form

We also have the direct form

$$\frac{\left\|\mathrm{d}S_{\text{quad}}\right\|}{\|\mathrm{d}S_{\text{linear}}\|} = \frac{\left\|X\,\mathrm{d}W_Q\,\mathrm{d}W_K^\top X^\top\right\|_F}{\left\|X\left(W_Q\,\mathrm{d}W_K^\top + \mathrm{d}W_Q W_K^\top\right)X^\top\right\|_F}$$

We can consider two different approaches, either picking $\mathrm{d}S_{\text{full}}$ or $\mathrm{d}S_{\text{linear}}$.

🔥 How and when to choose $\mathrm{d}S_{\text{full}}$ or $\mathrm{d}S_{\text{linear}}$?

Appendix A

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

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

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,*

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

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

Then

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

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

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