Brouillon 4

Léo Burgund

May 27, 2025

1 Nomenclature

1.1 Dimensions

- b Mini-batch size
- d_e Embedding dimension
- d_s Sequence length
- d_k Query/Keys dimension
- d_v Value dimension
- h Number of heads

We make the hypothesis that $d_k < d_e < d_s$.

1.2 Matrix operations in a self-attention block

In the case of multi head attention, for each head $i = 1, ..., h$, we have:

- Input $X \in \mathbb{R}^{d_s \times d_e}$
- $W_{Q_i} \in \mathbb{R}^{d_e \times \frac{d_k}{h}}, Q_i := XW_{Q_i} \in \mathbb{R}^{d_s \times \frac{d_k}{h}}$
- $W_{K_i} \in \mathbb{R}^{d_e \times \frac{d_k}{h}}, K_i := XW_{K_i} \in \mathbb{R}^{d_s \times \frac{d_k}{h}}$
- $S_i := \frac{Q_i K_i^\top}{\sqrt{\frac{d_k}{h}}} \in \mathbb{R}^{d_s \times d_s}$
- $A_i := \text{softmax}_{\text{row}}(S)$
- $W_{V_i} \in \mathbb{R}^{d_e \times \frac{d_v}{h}}, V_i := XW_{V_i} \in \mathbb{R}^{d_s \times \frac{d_v}{h}}$
- $H_i := A_i V_i \in \mathbb{R}^{d_s \times \frac{d_v}{h}}$, $H = [H_1, ..., H_h] \in \mathbb{R}^{d_s \times d_v}$
- $W_O \in \mathbb{R}^{d_v \times d_e}$
- Output $Y := HW_O + X \in \mathbb{R}^{d_s \times d_e}$

For now, we study the case $h = 1$.

🔥 We omit the $\frac{1}{\sqrt{d_k}}$ scaling for the S matrix, it can cause problem with growing, so for growing we will make it a learnable parameter (and initialize it at $\frac{1}{\sqrt{d_{k_{\text{initial}}}}}$?). Or maybe scale with $\left(\frac{\sqrt{d_k}}{\sqrt{d_k+p}}\right)$? But need to maintain the same output for the model?

2 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

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

Let $G := \nabla_S \mathcal{L}(S)$.

We have

$$\operatorname{rk}(G) \leq d_s, \quad \operatorname{rk}(S) \leq d_k$$

We have the first order approximation, with the introduction of $\gamma \in \mathbb{R}_+^*$, similar to a step size:

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

We consider the problem

$$\underset{\mathrm{d}S}{\arg\min} \, \langle G, \mathrm{d}S \rangle_F \text{ s.t. } \|\mathrm{d}S\|_F \leq \gamma$$

Let $Z = W_Q W_K^\top$, with $\operatorname{rk}(Z) \leq d_k$, we then have

$$\begin{aligned} S &= X W_Q W_K^\top X^\top \\ &= X Z X^\top \end{aligned}$$

and

$$\begin{aligned} \mathrm{d}S &= X\left(W_{Q_{+1}} W_{K_{+1}}\right) X^\top - X W_Q W_K^\top X^\top \\ &= X(Z + \mathrm{d}Z) X^\top - X Z X^\top \\ &= X \, \mathrm{d}Z X^\top \end{aligned}$$

Moreover,

$$\operatorname{rk}(\mathrm{d}S) = \operatorname{rk}\left(X \, \mathrm{d}Z X^\top\right) = \operatorname{rk}(\mathrm{d}Z) \leq d_k < d_e < d_s$$

Hence, the problem becomes

$$\begin{aligned} \mathrm{d}Z^\star &= \underset{\mathrm{d}Z}{\arg\min} \left\langle G, X \, \mathrm{d}Z X^\top \right\rangle_F \text{ s.t. } \left\|X \, \mathrm{d}Z X^\top\right\|_F \leq \gamma \\ &= -\underset{\mathrm{d}Z}{\arg\max} \left\langle G, X \, \mathrm{d}Z X^\top \right\rangle_F \text{ s.t. } \left\|X \, \mathrm{d}Z X^\top\right\|_F \leq \gamma \\ &= -\underset{\mathrm{d}Z}{\arg\max} \left\langle G, X \, \mathrm{d}Z X^\top \right\rangle_F \text{ s.t. } \left\|X \, \mathrm{d}Z X^\top\right\|_F = \gamma \ (*) \\ &= -\gamma \underset{\mathrm{d}Z}{\arg\max} \left\langle G, X \, \mathrm{d}Z X^\top \right\rangle_F \text{ s.t. } \left\|X \, \mathrm{d}Z X^\top\right\|_F = 1 \\ &= -\frac{\gamma}{\alpha} \underset{\mathrm{d}Z}{\arg\min} \left\|G - X \, \mathrm{d}Z X^\top\right\|_F^2 \end{aligned}$$

$(*)$ We make the hypothesis that we can always find a $\left\langle G, X \, \mathrm{d}Z X^\top \right\rangle_F < 0$.

🔥 Justification for α, $\alpha = \|?\|_F$

Let $\lambda := \frac{\gamma}{\alpha}$, $P := \arg\min_{\mathrm{d}Z} \left\|G - X \, \mathrm{d}Z X^\top\right\|_F^2$.

We will first search P, then λ with a line search.

2.1 Find P

Let

$$f(P) = \left\|G - X P X^\top\right\|_F^2.$$

$P \mapsto X P X^\top$ is linear, $P \mapsto G - X P X^\top$ is affine, and $A \mapsto \|A\|_F^2$ is convex. f is a composition of those functions, hence is convex.

We have

$$\nabla_P f = -2 X^\top (G - X P X^\top) X$$

so

$$\nabla_P f = 0 \Longleftrightarrow X^\top X P X^\top X = X^\top G X$$

2.1.1 X full column rank (🔥 true in practice?)

Under the hypothesis that X has full column (d_e) rank, $X^\top X$ is invertible, we have the pseudoinverse

$$X^+ = \left(X^\top X\right)^{-1} X^\top$$

and the solution

$$\begin{aligned} P^\star &= \left(X^\top X\right)^{-1} X^\top G X \left(X^\top X\right)^{-1} \\ &= \left(X^\top X\right)^{+} X^\top G X \left(X^\top X\right)^{+} \\ &= X^+ G \left(X^+\right)^\top \end{aligned}$$

🔥 Which formula to implement for $P^\star$? Numerical stability?

2.1.2 $X^\top X$ not invertible

Let

$$\mathcal{A} := P \mapsto X^\top X P X^\top X$$

If $X^\top X$ is not invertible, $X^\top X$ is not injective, and under the hypothesis that $X^\top X \neq 0$,

$$X^\top X N X^\top X = 0 \iff X N X^\top = 0$$

hence

$$\ker(\mathcal{A}) = \left\{N \in \mathbb{R}^{d_e \times d_e} \mid X N X^\top = 0\right\}$$

Hence any solution P_0 of $X^\top X P X^\top X = X^\top G X$ can be changed by $N \in \ker(\mathcal{A})$.

$$P = P_0 + N \Rightarrow X^\top X (P_0 + N) X^\top X = X^\top X P_0 X^\top X$$

We can always take $N = 0$.

🔥 Can we take $P_0 = X^+ G \left(X^+\right)^\top$?

2.2 Batch

To account for the batch, there are several ways to average:

$$\mathbb{E}_X\left[\left(X^\top X\right)^{+} X^\top G X \left(X^\top X\right)^{+}\right] = \mathbb{E}_X\left[X^+ G \left(X^+\right)^\top\right]$$

$$\mathbb{E}_X\left[X^\top X\right]^{+} \mathbb{E}_X\left[X^\top G X\right] \mathbb{E}_X\left[X^\top X\right]^{+}$$

$$\mathbb{E}_X[X]^{+} \mathbb{E}_X[G] \left(\mathbb{E}_X[X]^{+}\right)^\top$$

$$\mathbb{E}_X\left[X^\top X\right]^{-1} \mathbb{E}_X\left[X^\top G X\right] \mathbb{E}_X\left[X^\top X\right]^{-1}$$

🔥 Which one to take?

When $d_s \gg d_e$, we see with experiments:

$$\mathbb{E}_X\left[X^\top X\right]^{+} \mathbb{E}_X\left[X^\top G X\right] \mathbb{E}_X\left[X^\top X\right]^{+} \to \mathbb{E}_X\left[\left(X^\top X\right)^{+} X^\top G X \left(X^\top X\right)^{+}\right]$$

With the left member being cheaper to compute

🔥 Test full pinv VS (if possible inv else pinv)

2.3 Line search

Do a line search to find λ

2.3.1 "Normal" way

Line search on

$$\mathcal{L}\Big(X\ \mathrm{SVD}_{d_k+p}\Big(W_Q^t W_K^{t^\top} - \lambda P\Big)X^\top\Big)$$

2.3.2 Testing fast way

Lose the rank constraint, but lose the SVD so faster. Get a good approximation of λ ?

$$\mathcal{L}\Big(X\Big(W_Q^t W_K^{t^\top} - \lambda P\Big)X^\top\Big) = \mathcal{L}\Big(XW_Q^t W_K^{t^\top} X^\top - \lambda XPX^\top\Big)$$