Proof of LQER with an Assumption to be Verified

Background

- Input vector of linear layer: $\mathbf{x} = [x_1, x_2, ..., x_n]$, where n is the hidden size.
- linear layer: $\mathbf{y} = \mathbf{x}W$
- Approximated linear layer: $\tilde{\mathbf{y}} = \mathbf{x}\left(\widetilde{W} + C\right)$, where C is a low-rank approximation of quantization error, and $\widetilde{W} = \text{quantize}(W)$

Geroge's Proof with Assumption

We target minimizing the expectation of L2 norm for output vectors, *i.e.* , $\min \mathrm{E}_{\mathbf{y}}\{\| \tilde{\mathbf{y}} - \mathbf{y} \|_2^2\}$.

We have

$$\mathrm{E}_{\mathbf{y}}\{\| \tilde{\mathbf{y}} - \mathbf{y} \|_2^2\} = \mathrm{E}_{\mathbf{x}}\left\{\| \mathbf{x}\left(\widetilde{W} - W + C\right) \|_2^2\right\} \tag{1}$$

where $\mathrm{E}\{\cdot\}$ stands for expectation.

Let $M = \widetilde{W} - W + C = \begin{bmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_n} \end{bmatrix}$ and put it into RHS of (1).

$$\begin{aligned}
\mathrm{E}_{\mathbf{y}}\{\|\tilde{\mathbf{y}} - \mathbf{y}\|\} &= \mathrm{E}_{\mathbf{x}}\{\|\mathbf{x}M\|_2^2\} \\
&= \mathrm{E}_{\mathbf{x}}\left\{\| [x_1, x_2, ..., x_n] \begin{bmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_n} \end{bmatrix} \|_2^2\right\} \\
&= \mathrm{E}_{\mathbf{x}}\left\{\| \sum_{i=1}^{n} x_i \mathbf{m_i} \|_2^2\right\} \\
&= \mathrm{E}_{\mathbf{x}}\left\{\left(\sum_{i=1}^{n} x_i \mathbf{m_i}\right)\left(\sum_{j=1}^{n} x_j \mathbf{m_j^T}\right)\right\} \\
&= \mathrm{E}_{\mathbf{x}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \mathbf{m_i}\mathbf{m_j^T}\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}_{\mathbf{x}}\left\{x_i x_j \mathbf{m_i}\mathbf{m_j^T}\right\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{E}_{\mathbf{x}}\{x_i x_j\}\mathbf{m_i}\mathbf{m_j^T}
\end{aligned} \tag{2}$$

Assumption:

$$\mathrm{E}\{x_i x_j\} = 0 \text{ for } i \neq j \tag{3}$$

Then the RHS of (2) becomes

$$\mathrm{E}_{\mathbf{y}}\{\|\tilde{\mathbf{y}}-\mathbf{y}\|\} = \sum_{i=1}^{n} \mathrm{E}\{x_i^2\}\ \mathbf{m_i}\mathbf{m_i^T} \tag{4}$$

💡 One interpretation of (3) is that the i-th activation dim x_i is zero-mean, and independent from the j-th dim ($i \neq j$).

If we assign diagonal matrix $S = \mathrm{diag}\left(\sqrt{\mathrm{E}\{x_1^2\}}, \sqrt{\mathrm{E}\{x_2^2\}}, ..., \sqrt{\mathrm{E}\{x_n^2\}}\right)$, the LHS of (4) becomes

$$\begin{aligned} \mathrm{E}_{\mathbf{y}}\{\|\tilde{\mathbf{y}}-\mathbf{y}\|\} &= \mathrm{Trace}(SMM^TS^T) \\ &= \|\ SM\ \|_F^2 \end{aligned} \tag{5}$$

where $\|\cdot\|_F^2$ denotes Frobenius norm.

Given (5), our target now is

$$\begin{aligned} \min \mathrm{E}_{\mathbf{y}}\{\|\ \tilde{\mathbf{y}}-\mathbf{y}\ \|_2^2\} &= \min\|\ SM\ \|_F^2 \\ &= \min\|\ S\left(\widetilde{W}-W+C\right)\ \|_F^2 \end{aligned} \tag{6}$$

If we assign $A = S\left(W-\widetilde{W}\right)$ and $\tilde{A} = SC$, the RHS of (6) becomes

$$\min \mathrm{E}_{\mathbf{y}}\{\|\ \tilde{\mathbf{y}}-\mathbf{y}\ \|_2^2\} = \min\|\ \tilde{A}-A\ \|_F^2 \tag{7}$$

According to Eckart–Young theorem, the best rank k approximation to A (noted as $\tilde{A}_k$) is

$$\tilde{A}_k = U_k\Sigma_kV_k^T \tag{8}$$

where U_k, Σ_k, V_k is the rank-k SVD of A, *i.e.*, $A = S\left(W-\widetilde{W}\right) = U\Sigma V^T$.

Therefore, the closed-form solution of C based on assumption (3) is

$$C = S^{-1}\tilde{A}_k = S^{-1}U_k\Sigma_kV_k^T \tag{9}$$

For implementation, we assign two low-rank matrices $S^{-1}U_k$ and $\Sigma_kV_k^T$ to save FLOPs.

An Example of Statistic Profile

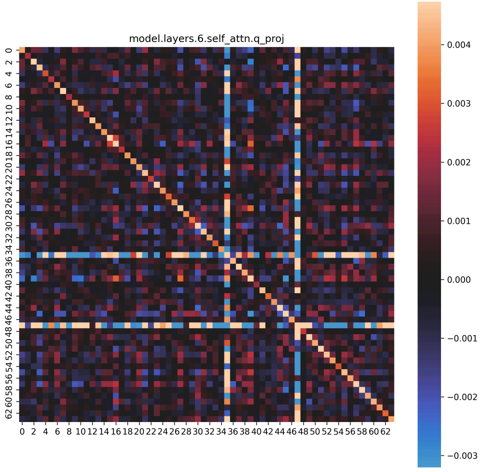
We can calculate the auto-correlation matrix for $\mathbf{x} = [x_1, x_2, ..., x_n]$:

$$R_{\mathbf{xx}} = \begin{pmatrix} \mathrm{E}(x_1x_1) & \mathrm{E}(x_1x_2) & ... & \mathrm{E}(x_1x_n) \\ \mathrm{E}(x_2x_1) & \mathrm{E}(x_2x_2) & ... & \mathrm{E}(x_2x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(x_nx_1) & \mathrm{E}(x_nx_2) & ... & \mathrm{E}(x_nx_n) \end{pmatrix} \tag{10}$$

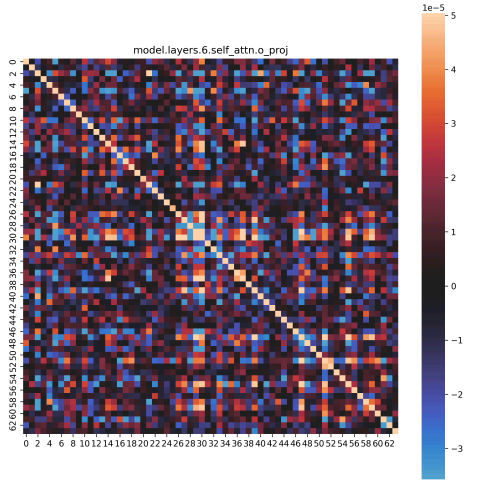
where $\mathrm{E}\left(x_ix_j\right)$ is the correlation between $\mathbf{x}$'s i-th dim and j-th dim.

Figure 1 is the profiled first 64 dim of R_{xx} for the linear layers of 6-th decoder in TinyLlama-1.1B. Color black mean 0 value.

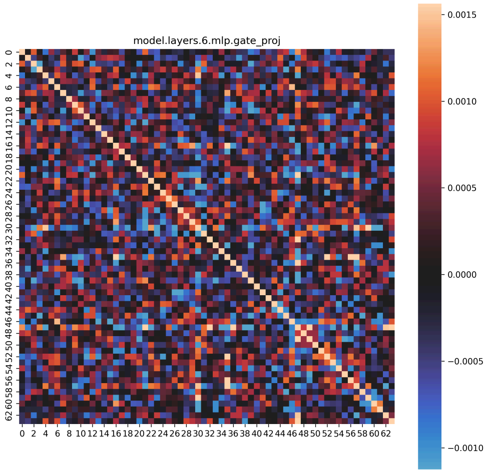
It is observed that our assumption is reasonable but looks a bit strong (?).



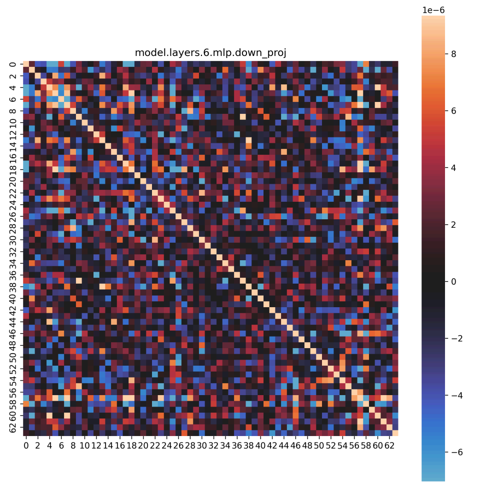
(a) The input of q/k/v_proj layer of 6-th decoder



(b) The input of o_proj layer of 6-th decoder



(c) The input of gate/up_proj layer of 6-th decoder



(d) The input of down_proj layer of 6-th decoder

What if the Assumption does not hold?

George hypothesizes that if the assumption (3) does not hold, the closed-form of S might be $S = R_{\mathbf{xx}}^{\frac{1}{2}}$, where $R_{\mathbf{xx}} = \mathrm{E}\{\mathbf{x}^{\mathbf{T}}\mathbf{x}\}$. Then (2) may become $\| R_{\mathbf{xx}}^{\frac{1}{2}} \odot M \|_F^2$ where $\odot$ is elementwise-multiply.

Cheng's Try without the Assumption

Cheng: I tried the following to push the proof.

If we continue at ⚓ in (2), we have:

$$
\begin{aligned}
\mathrm{E}_{\mathbf{y}}\{\|\tilde{\mathbf{y}}-\mathbf{y}\|\} &= \mathrm{E}_{\mathbf{x}}\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \mathbf{m_i} \mathbf{m_j^T}\right\} \\
&= \mathrm{E}_{\mathbf{x}}\left\{\mathbf{e}\begin{pmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_n \\ x_2x_1 & x_2x_2 & \dots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \dots & x_nx_n \end{pmatrix} \odot \begin{pmatrix} \mathbf{m_1m_1^T} & \mathbf{m_1m_2^T} & \dots & \mathbf{m_1m_n^T} \\ \mathbf{m_2m_1^T} & \mathbf{m_2m_2^T} & \dots & \mathbf{m_2m_n^T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_nm_1^T} & \mathbf{m_nm_2^T} & \dots & \mathbf{m_nm_n^T} \end{pmatrix} \mathbf{e^T}\right\} \quad ⚓ \\
&= \mathrm{E}_{\mathbf{x}}\{\mathbf{e}\cdot((\mathbf{x^T x})\odot(MM^T))\cdot\mathbf{e^T}\}
\end{aligned}
\tag{11}
$$

where $\mathbf{e}=\underbrace{[1,1,...,1]}_{\mathrm{n}}$ is a row vector of n ones, and $\odot$ is elementwise multiply. ⚓ in (11) is based on this StackExchange page.

Using the following property of Hadamard product

- For vectors $\mathbf{x}$ and $\mathbf{y}$, and corresponding diagonal matrices $D_{\mathbf{x}}$ and $D_{\mathbf{y}}$ with these vectors as their main diagonals, the following identity holds:[1]:479

  $$\mathbf{x}^*(A\circ B)\mathbf{y}=\mathrm{tr}\left(D_{\mathbf{x}}^* A D_{\mathbf{y}} B^{\mathsf{T}}\right),$$

  where $\mathbf{x}^*$ denotes the conjugate transpose of $\mathbf{x}$. In particular, using vectors of ones, this shows that the sum of all elements in the Hadamard product is the trace of AB^{T} where superscript T denotes the matrix transpose, that is, $\mathrm{tr}\left(AB^{\mathsf{T}}\right)=\mathbf{1}^{\mathsf{T}}\left(A\odot B\right)\mathbf{1}$. A related result for square A and B, is that the row-sums of their Hadamard product are the diagonal elements of AB^{T}:[8]

  $$\sum_i (A\circ B)_{ij}=\left(B^{\mathsf{T}}A\right)_{jj}=\left(AB^{\mathsf{T}}\right)_{ii}.$$

  Similarly,

  $$(\mathbf{y}\mathbf{x}^*)\odot A=D_{\mathbf{y}} A D_{\mathbf{x}}^*$$

  Furthermore, a Hadamard matrix-vector product can be expressed as:

  $$(A\odot B)\mathbf{y}=\mathrm{diag}\left(AD_{\mathbf{y}}B^{\mathsf{T}}\right)$$

  where $\mathrm{diag}(M)$ is the vector formed from the diagonals of matrix M.

The LHS of (11) becomes

$$
\begin{aligned}
\mathrm{E}_{\mathbf{y}}\{\|\tilde{\mathbf{y}}-\mathbf{y}\| &= \mathrm{E}_{\mathbf{x}}\{\mathrm{Trace}(\mathbf{x^T x}MM^T)\} \\
&= \mathrm{Trace}(\mathrm{E}_{\mathbf{x}}\{\mathbf{x^T x}MM^T\}) \quad 🚀 \\
&= \mathrm{Trace}(\mathrm{E}_{\mathbf{x}}\{\mathbf{x^T x}\}MM^T) \\
&= \mathrm{Trace}(R_{\mathbf{xx}}MM^T) \\
&= \mathrm{Trace}\left(R_{\mathbf{xx}}^{\frac{1}{2}}MM^TR_{\mathbf{xx}}^{\frac{1}{2}}\right) \quad ⚓ \\
&= \mathrm{Trace}\left(R_{\mathbf{xx}}^{\frac{1}{2}}MM^T\left(R_{\mathbf{xx}}^{\frac{1}{2}}\right)^T\right) \\
&= \|\ R_{\mathbf{xx}}^{\frac{1}{2}}M\ \|_F^2
\end{aligned}
\tag{12}
$$

🚀 in (12) is based on this page.

⚓ in (12) is based on the following clues:

- R_{xx} is symmetric and positive-semidefinite. Refer to wiki page.
- MM^T is positive-semidefinites. Refer to StackExchange page
- For two positive-semidefinite matrices A and B, we have $\mathrm{Trace}(AB)=\mathrm{Trace}\left(A^{\frac{1}{2}}BA^{\frac{1}{2}}\right)$: refer to wiki page

Trace [edit]

The diagonal entries m_{ii} of a positive-semidefinite matrix are real and non-negative. As a consequence the trace, $\operatorname{tr}(M) \geq 0$. Furthermore,[13] since every principal sub-matrix (in particular, 2-by-2) is positive semidefinite,

$$|m_{ij}| \leq \sqrt{m_{ii} m_{jj}} \quad \forall i, j$$

and thus, when $n \geq 1$,

$$\max_{i,j} |m_{ij}| \leq \max_i m_{ii}$$

An $n \times n$ Hermitian matrix M is positive definite if it satisfies the following trace inequalities:[14]

$$\operatorname{tr}(M) > 0 \quad \text{and} \quad \frac{(\operatorname{tr}(M))^2}{\operatorname{tr}(M^2)} > n - 1.$$

Another important result is that for any M and N positive-semidefinite matrices, $\operatorname{tr}(MN) \geq 0$. $\boxed{\text{This follows by writing } \operatorname{tr}(MN) = \operatorname{tr}(M^{\frac{1}{2}} N M^{\frac{1}{2}}).}$ The matrix $M^{\frac{1}{2}} N M^{\frac{1}{2}}$ is positive-semidefinite and thus has non-negative eigenvalues, whose sum, the trace, is therefore also non-negative.

- $R_{\mathbf{xx}}$ has a symmetric and positive semidefinite square root: $R_{\mathbf{xx}} = R_{\mathbf{xx}}^{\frac{1}{2}} R_{\mathbf{xx}}^{\frac{1}{2}}$. Refer to wiki page

Positive semidefinite matrices [edit]

See also: Positive definite matrix § Decomposition

A symmetric real *n* × *n* matrix is called positive semidefinite if $x^\mathsf{T} A x \geq 0$ for all $x \in \mathbb{R}^n$ (here x^T denotes the transpose, changing a column vector x into a row vector). A square real matrix is positive semidefinite if and only if $A = B^\mathsf{T} B$ for some matrix B. There can be many different such matrices B. A positive semidefinite matrix A can also have many matrices B such that $A = BB$. However, A always has precisely one square root B that is positive semidefinite and symmetric. In particular, since B is required to be symmetric, $B = B^\mathsf{T}$, so the two conditions $A = BB$ or $A = B^\mathsf{T} B$ are equivalent.

Now our target becomes

$$\begin{aligned}\min \mathrm{E}_{\mathbf{y}}\{\| \tilde{\mathbf{y}} - \mathbf{y} \|_2^2\} &= \min \| R_{\mathbf{xx}}^{\frac{1}{2}} M \|_F^2 \\ &= \min \| R_{\mathbf{xx}}^{\frac{1}{2}} \left(\widetilde{W} - W + C \right) \|_F^2\end{aligned} \tag{13}$$

If we assign $A = R_{\mathbf{xx}}^{\frac{1}{2}} \left(W - \widetilde{W} \right)$ and $\tilde{A} = R_{\mathbf{xx}}^{\frac{1}{2}} C$, the RHS of (13) becomes

$$\min \mathrm{E}_{\mathbf{y}}\{\| \tilde{\mathbf{y}} - \mathbf{y} \|_2^2\} = \min \| \tilde{A} - A \|_F^2 \tag{14}$$

According to Eckart–Young theorem, the best rank k approximation to A (noted as $\tilde{A}_k$) is

$$\tilde{A}_k = U_k \Sigma_k V_k^T \tag{15}$$

Therefore, the closed-form solution of C without the assumption is

$$C = \left(R_{\mathbf{xx}}^{\frac{1}{2}} \right)^{-1} \tilde{A}_k = \left(R_{\mathbf{xx}}^{\frac{1}{2}} \right)^{-1} U_k \Sigma_k V_k^T \tag{16}$$

For implementation, we assign two low-rank matrices $\left(\left(R_{\mathbf{xx}}^{\frac{1}{2}} \right)^{-1} U_k \right)$ and $\left(\Sigma_k V_k^T \right)$ to save FLOPs.

❓ A concern here is if $R_{\mathbf{xx}}^{\frac{1}{2}}$ is always invertible? We assume in practice it is inveritble for a trained network, but we need a backup plan if it is not (Probably add perturbation to $R_{\mathbf{xx}}^{\frac{1}{2}}$, or fall back to the diagonal S case)

Numerical Stability of Matrix Square Root and SVD

- The numerical stability of this algorithm is not ensured and has not been investigated.
- Current empirical implementation uses use FP32 to compute $R_{\mathbf{xx}}$, use FP64 to accumulated $R_{\mathbf{xx}}$, and use FP32 to solve the matrix square root and SVD. The evaluation of the quantized network may use FP32, or BF16 to align with MXINT's 8-bit exponent.