Supplementary Proofs for Rounding Error Analysis of 1D Weak Form Kernel

Huiyu Xie

Apr 14, 2024

The goal of this article is to provide supplementary proofs for the induction from (14) to (15) in the rounding error analysis of the 1D weak form kernel. It can be treated as a continuation of the previous work on this topic.

To make life easier, we simplify the notations used in the previous work. It is easy to see that (13) is equivalent to

$$|\boldsymbol{y} - \hat{\boldsymbol{y}}| \leq \gamma |B||\boldsymbol{x}| \tag{16}$$

and (14) is equivalent to

$$|A - \hat{A}| \leq \gamma |B||C| \tag{17}$$

and (15) is equivalent to

$$\|A - \hat{A}\|_p \leq \gamma \|B\|_p \|C\|_p, \quad p = 1, \infty, F \tag{18}$$

where we assume $A \in \mathbb{R}^{m \times k}$, $B \in \mathbb{R}^{m \times n}$, and $C \in \mathbb{R}^{n \times k}$ (thus $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$). Also note that the absolute value operation is applied componentwise for all of the above. In order to achieve the induction from (17) to (18), we have to first obtain some auxiliary results (Proposition 1) based on (16).

Note that the cases involving the infinity norm and 2-norm in Proposition 1 can be omitted when reading as they are included for fun of proof.

Proposition 1. *Let $\gamma \leq 0$, $\boldsymbol{x} \in \mathbb{R}^n$, $B \in \mathbb{R}^{m \times n}$, $\boldsymbol{y}$ and $\hat{\boldsymbol{y}} \in \mathbb{R}^m$. If*

$$|\boldsymbol{y} - \hat{\boldsymbol{y}}| \leq \gamma |B||\boldsymbol{x}| \tag{16}$$

holds, then

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_p \leq \gamma \|B\|_p \|\boldsymbol{x}\|_p, \quad p = 1, \infty \tag{19}$$

and for the 2-norm

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 \leq min(m, n)^{1/2} \gamma \|B\|_2 \|\boldsymbol{x}\|_2 \tag{20}$$

Proof. First consider the situation of $p = 1$. Based on (16), we have

$$|y_i| \leq \gamma \sum_{j=1}^{n} (|b_{ij}||x_j|), \quad i = 1, ..., m$$

then we can further have

$$\begin{aligned}\sum_{i=1}^{m} |y_i| &\leq \gamma \sum_{i=1}^{m} \sum_{j=1}^{n} (|b_{ij}||x_j|) = \gamma \sum_{j=1}^{n} \left(|x_j| \sum_{i=1}^{m} |b_{ij}| \right) \\ &\leq \gamma \sum_{j=1}^{n} \left(|x_j| \max_{1 \leq j \leq n} \sum_{i=1}^{m} |b_{ij}| \right) = \gamma \left(\max_{1 \leq j \leq n} \sum_{i=1}^{m} |b_{ij}| \right) \sum_{j=1}^{n} |x_j|\end{aligned}$$

By the definition of 1-norm, we have

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1 \leq \gamma \|B\|_1 \|\boldsymbol{x}\|_1$$

So the situation of $p = 1$ is proved, and then we go for the case of $p = \infty$. Again, based on (16), we can have

$$\begin{aligned}\max_{1 \leq i \leq m} |y_i| &\leq \max_{1 \leq i \leq m} \gamma \sum_{j=1}^{n} (|b_{ij}||x_j|) = \gamma \max_{1 \leq i \leq m} \sum_{j=1}^{n} (|b_{ij}||x_j|) \\ &\leq \gamma \max_{1 \leq i \leq m} \sum_{j=1}^{n} \left(|b_{ij}| \max_{1 \leq j \leq n} |x_j| \right) = \gamma \left(\max_{1 \leq i \leq m} \sum_{j=1}^{n} |b_{ij}| \right) \max_{1 \leq j \leq n} |x_j|\end{aligned}$$

By the definition of infinity norm, we have

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_\infty \leq \gamma \|B\|_\infty \|\boldsymbol{x}\|_\infty$$

So the situation of $p = \infty$ is proved, and then we go for the case of $p = 2$. In order to prove this, we have to mention another lemma first.

Lemma 2. *Let $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$, then*

$$\|A\boldsymbol{x}\|_2 \leq \|A\|_2 \|\boldsymbol{x}\|_2 \tag{21}$$

$$\|A\|_2 \leq \||A|\|_2 \leq \sqrt{rank(A)} \|A\|_2 \tag{22}$$

The proof for Lemma 2 is skipped here as it is a widely known result and can be easily proved. Based on (16), we have

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 = \||\boldsymbol{y} - \hat{\boldsymbol{y}}|\|_2 \leq \|\gamma |B||\boldsymbol{x}|\|_2 = \gamma \||B||\boldsymbol{x}|\|_2$$

and by (21) and (22), we can further have

$$\begin{aligned}\||B||\boldsymbol{x}|\|_2 \leq \||B|\|_2 \|\boldsymbol{x}\|_2 &\leq \sqrt{\text{rank}(B)} \|B\|_2 \|\boldsymbol{x}\|_2 \\ &\leq \min(m, n)^{1/2} \|B\|_2 \|\boldsymbol{x}\|_2\end{aligned}$$

Thus we also prove that

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2 \leq \min(m,n)^{1/2}\gamma\|B\|_2\|\boldsymbol{x}\|_2$$

□

Now it is the right time to go for the real induction form (14) to (15) using Proposition 1. For convenience, we form (17) and (18) as Proposition 2.

Proposition 2. *Let $\gamma \leq 0$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{n\times p}$, and $A \in \mathbb{R}^{m\times p}$. If*

$$|A - \hat{A}| \leq \gamma|B||C| \tag{17}$$

holds, then

$$\|A - \hat{A}\|_p \leq \gamma\|B\|_p\|C\|_p, \quad p = 1, \infty, F \tag{18}$$

Proof. First consider the situation of $p = 1$. Based on (17), we have

$$|\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j| \leq \gamma|B||\boldsymbol{c}_j|, \quad j = 1, ..., p$$

and by Proposition 1, we further have

$$\|\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j\|_1 \leq \gamma\|B\|_1\|\boldsymbol{c}_j\|_1, \quad j = 1, ..., p$$

then we can can get

$$\max_{1\leq j\leq p}\|\boldsymbol{a}_j - \hat{\boldsymbol{a}}_j\|_1 \leq \max_{1\leq j\leq p}\gamma\|B\|_1\|\boldsymbol{c}_j\|_1 = \gamma\|B\|_1\left(\max_{1\leq j\leq p}\|\boldsymbol{c}_j\|_1\right)$$

which is equivalent to

$$\|A - \hat{A}\|_1 \leq \gamma\|B\|_1\|C\|_1$$

So the situation of $p = 1$ is proved, and then we go for the case of $p = \infty$. Suppose there exists a constant $k \in \{1, ..., m\}$ such that

$$\|A - \hat{A}\|_\infty = \sum_{j=1}^{p}|a_{kj}|$$

then it is easy to get

$$\begin{aligned}\sum_{j=1}^{p}|a_{kj}| &= \sum_{j=1}^{n}|b_{kj}|\left(\sum_{i=1}^{p}|c_{ji}|\right)\\ &\leq \left(\max_{1\leq i\leq m}\sum_{j=1}^{n}|b_{ij}|\right)\left(\max_{1\leq j\leq n}\sum_{i=1}^{p}|c_{ji}|\right)\end{aligned}$$

which is equivalent to

$$\|A - \hat{A}\|_\infty \leq \gamma\|B\|_\infty\|C\|_\infty$$

So the situation of $p = \infty$ is proved and finally we go for the case of $p = F$. Again, in order to prove this case, we have to mention another lemma first.

Lemma 3. *Let $A, B \in \mathbb{R}^{m\times n}$, then*

$$\|AB\|_F \le \|A\|_F \|B\|_F \tag{23}$$
$$\|A\|_F \le \|B\|_F \text{ if } \|\boldsymbol{a}_j\|_2 \le \|\boldsymbol{b}_j\|_2 \text{ for } j = 1, \ldots, n \tag{24}$$

The proof for Lemma 3 is skipped here as it is a widely known result and can be easily proved. Based on (17), it is easy to have

$$\|(A - \hat{A})_{\cdot j}\|_2 \le \|(\gamma|B||C|)_{\cdot j}\|_2, \quad j = 1, ..., p$$

and then by (23) and (24) we have

$$\begin{aligned} \|A - \hat{A}\|_F \le \|\gamma|B||C|\|_F &= \gamma\||B||C|\|_F \\ &\le \gamma\|B\|_F\|C\|_F \end{aligned}$$

So the situation of $p = F$ is also proved. □

Thus we complete the supplementary proofs for rounding error analysis of 1D weak form kernel.