

## A NAIVE SYMMETRIC MAX-SCALED BLOCK-WISE QUANTIZATION

In this section, we rigorously define the idea of *naive symmetric max-scaled block-wise quantization* as defined in [Gholami et al. \(2021\)](#).

### A.1 BLOCK-WISE

Suppose we divide the weights into blocks (groups)  $\{\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(B)}\}$ , defined by a partition  $\{P_1, \dots, P_B\}$  of  $[d]$ . Typically, blocks are collections of rows or columns. That is, in order for a quantization scheme to be block-wise, we require that  $Q_i = Q_j$  and  $D_i = D_j$  for all  $i, j \in P_k$ , where  $k \in [B]$ . Intuitively, this requires that the dequantization functions be the same for all weights in a block. This is desirable for hardware purposes, since nearby elements will share the same dequantization operations. If  $B = 1$ , we call this a *tensor-wise* quantization scheme.

### A.2 MAX-SCALED

We define a quantization scheme to be *max-scaled* with respect to blocks  $\{\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(B)}\}$  iff the following conditions hold.

1.  $\Theta = \mathbb{R}^B$
2. For all  $k \in [B]$  and  $i \in P_k$ ,  $\Phi$  is defined by  $(\Phi(\mathbf{w}))_i = \max_{j \in P_k} |\mathbf{w}_j|$ .
3. For all  $k \in [B]$  and  $i \in P_k$ , we require that  $D_i$  is defined as  $(q, \theta) \mapsto q\theta_k$ .

By the last condition, we can see that max-scaled quantization schemes are also block-wise. We restrict the parameters to be the maximum elements of each block. Then, we require that the dequantization simply scale this maximum. In other words, the quantization grid (range of  $D_i$ ) is parameterized only by the maximum magnitude for an element in a block. The grid scales with the maximum magnitude element. Removing the second condition will give us a *scaled* quantization scheme.

### A.3 NAIVE

Finally, we define a naive quantization scheme. A quantization scheme is *naive* iff

$$Q_i(\mathbf{w}, \theta) = \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \theta)|$$

for all  $\mathbf{w} \in \mathbb{R}^d$  and  $\theta \in \Theta$ . In particular, this quantizes  $\mathbf{w}_i$  to the nearest point in the quantization grid. This is also known as a rounding-to-nearest scheme. Note that under this naive assumption,  $Q_i$  is only a function of  $\mathbf{w}_i$ . In particular, there exists a function  $\tilde{Q}_i : \mathbb{R} \times \Theta \rightarrow \mathcal{Q}_i$  such that

$$\tilde{Q}_i(\mathbf{w}_i, \theta) = Q_i(\mathbf{w}, \theta)$$

for all  $\mathbf{w} \in \mathbb{R}^d$  and  $\theta \in \Theta$ .

### A.4 SYMMETRIC

We call a quantization scheme *symmetric* iff the following three conditions hold.

1. For all  $i \in [d]$ ,  $0 \in \mathcal{Q}_i$ .
2. For all  $i \in [d]$ ,  $\mathcal{Q}_i = -\mathcal{Q}_i$ .
3. For all  $i \in [d]$ ,  $D_i(q, \theta) = -D_i(-q, \theta)$ .

Intuitively, the quantization grid is symmetric iff the set of levels is symmetric about zero, and the corresponding dequantization function preserves this symmetry.

## A.5 ORDER-PRESERVING PROPERTY

Now, we illuminate a key property of naive scaled block-wise quantization, which we will call the *order-preserving property*.

**Theorem 5.** *Let  $i, j \in P_k$  for some block  $k$ . Furthermore, suppose  $\bar{\mathbf{Q}}$  is a naive scaled block-wise quantization scheme. Then, if  $\mathbf{w}_i \leq \mathbf{w}_j$ , we then claim that  $\bar{\mathbf{Q}}(\mathbf{w})_i \leq \bar{\mathbf{Q}}(\mathbf{w})_j$ .*

*Proof.* We want to show that

$$\bar{\mathbf{Q}}(\mathbf{w})_i \leq \bar{\mathbf{Q}}(\mathbf{w})_j.$$

By definition, we have the following.

$$\begin{aligned} \bar{\mathbf{Q}}(\mathbf{w})_i &\leq \bar{\mathbf{Q}}(\mathbf{w})_j \\ \mathbf{D}(\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))_i &\leq \mathbf{D}(\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))_j \\ D_i(\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w}))_i, \Phi(\mathbf{w})) &\leq D_j(\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w}))_j, \Phi(\mathbf{w})) \\ D_i(Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w})) &\leq D_j(Q_j(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w})) \end{aligned}$$

First, note that  $D_i = D_j$  since  $i$  and  $j$  are in the same partition by the block-wise assumption. Thus, it suffices to show that  $D_i(Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w})) \leq D_j(Q_j(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))$ . Next, note the following by the scaled assumption.

$$D_i(Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w})) = Q_i(\mathbf{w}, \Phi(\mathbf{w})) \Phi(\mathbf{w})_k$$

Similarly, note that

$$D_j(Q_j(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w})) = Q_j(\mathbf{w}, \Phi(\mathbf{w})) \Phi(\mathbf{w})_k.$$

Since  $\max_{\ell \in P_k} |\mathbf{w}_\ell| \geq 0$ , it suffices to show that

$$Q_i(\mathbf{w}, \Phi(\mathbf{w})) \leq Q_j(\mathbf{w}, \Phi(\mathbf{w})).$$

Finally, note that by the naive property, we have that

$$\begin{aligned} Q_i(\mathbf{w}, \Phi(\mathbf{w})) &= \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \Phi(\mathbf{w}))| \\ Q_j(\mathbf{w}, \Phi(\mathbf{w})) &= \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_j - D_i(a, \Phi(\mathbf{w}))|. \end{aligned}$$

Suppose towards a contradiction that  $Q_j(\mathbf{w}, \Phi(\mathbf{w})) < Q_i(\mathbf{w}, \Phi(\mathbf{w}))$ . Then, we have that

$$\arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_j - D_i(a, \Phi(\mathbf{w}))| < \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \Phi(\mathbf{w}))|.$$

Let  $a_i = \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \Phi(\mathbf{w}))|$  and  $a_j = \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_j - D_i(a, \Phi(\mathbf{w}))|$ . Furthermore, we have that

$$D_i(a, \Phi(\mathbf{w})) = a \Phi(\mathbf{w})_k.$$

Let  $s_k = a \Phi(\mathbf{w})_k$ . We claim that

$$|\mathbf{w}_j - a_j s_k| \geq |\mathbf{w}_j - a_i s_k|.$$

We have that  $a_j < a_i$  (by assumption),  $\mathbf{w}_i < \mathbf{w}_j$ , and  $|\mathbf{w}_i - a_i s_k| \leq |\mathbf{w}_i - a_j s_k|$ . We have the following.

$$\begin{aligned}
|\mathbf{w}_i - a_j s_k| &\geq |\mathbf{w}_i - a_i s_k| \\
(\mathbf{w}_i - a_j s_k)^2 &\geq (\mathbf{w}_i - a_i s_k)^2 \\
\mathbf{w}_i^2 - 2\mathbf{w}_i a_j s_k + a_j^2 s_k^2 &\geq \mathbf{w}_i^2 - 2\mathbf{w}_i a_i s_k + a_i^2 s_k^2 \\
-2\mathbf{w}_i a_j s_k + a_j^2 s_k^2 &\geq -2\mathbf{w}_i a_i s_k + a_i^2 s_k^2 \\
2\mathbf{w}_i a_i s_k - 2\mathbf{w}_i a_j s_k &\geq a_i^2 s_k^2 - a_j^2 s_k^2 \\
2\mathbf{w}_i (a_i s_k - a_j s_k) &\geq (a_i s_k + a_j s_k) (a_i s_k - a_j s_k) \\
2\mathbf{w}_i &\geq (a_j s_k + a_i s_k) & (a_j < a_i) \\
\mathbf{w}_i &\geq \frac{a_i s_k + a_j s_k}{2} \\
\mathbf{w}_j &\geq \frac{a_i s_k + a_j s_k}{2} & (\mathbf{w}_i < \mathbf{w}_j) \\
2\mathbf{w}_j &\geq a_i s_k + a_j s_k \\
2\mathbf{w}_j (a_i s_k - a_j s_k) &\geq (a_i s_k + a_j s_k) (a_i s_k - a_j s_k) \\
(\mathbf{w}_j - a_j s_k)^2 &\geq (\mathbf{w}_j - a_i s_k)^2 \\
|\mathbf{w}_j - a_j s_k| &\geq |\mathbf{w}_j - a_i s_k|
\end{aligned}$$

Thus, we have a contradiction, since we assumed  $a_j = \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_j - D_i(a, \Phi(\mathbf{w}))|$ .  $\square$

We also have the following corollary that is a natural extension of Theorem 5 adding the stronger symmetric property.

**Corollary 6.** *Let  $i, j \in P_k$  for some block  $k$ . Furthermore, suppose  $\bar{\mathbf{Q}}$  is a naive symmetric scaled block-wise quantization scheme. Then, if  $|\mathbf{w}_i| \leq |\mathbf{w}_j|$ , then*

$$|\bar{\mathbf{Q}}(\mathbf{w})_i| \leq |\bar{\mathbf{Q}}(\mathbf{w})_j|.$$

*Proof.* Before, we prove the corollary, we prove the following lemmas.

**Lemma 7.** *Suppose  $\bar{\mathbf{Q}}$  is a naive symmetric scaled block-wise quantization scheme. Then, if  $\mathbf{w}_i = 0$ , then  $\bar{\mathbf{Q}}(\mathbf{w})_i = 0$ .*

*Proof.* Since  $\bar{\mathbf{Q}}$  is symmetric, we have that  $0 \in \mathcal{Q}_i$  for all  $i \in [d]$ . Furthermore, since  $\bar{\mathbf{Q}}$  is scaled, we have that  $D_i(0, \theta) = 0$  for all  $i \in [d]$  and  $\theta \in \Theta$ . Therefore, we have that  $Q_i(\mathbf{w}, \theta) = 0$  for all  $i \in [d]$  and  $\theta \in \Theta$  by the naive assumption. Thus, we have that  $\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w}))_i = 0$  and  $\bar{\mathbf{Q}}(\mathbf{w})_i = \mathbf{D}(\mathbf{Q}(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))_i = 0$ .  $\square$

**Lemma 8.** *Suppose  $\bar{\mathbf{Q}}$  is a naive symmetric scaled block-wise quantization scheme. Construct  $\tilde{\mathbf{w}} \in \mathbb{R}^d$ . Let  $\tilde{\mathbf{w}}_i = -\mathbf{w}_i$  and  $\tilde{\mathbf{w}}_j = \mathbf{w}_j$  for all  $i \neq j$ . Then,*

$$|\bar{\mathbf{Q}}(\mathbf{w})_k| = |\bar{\mathbf{Q}}(\tilde{\mathbf{w}})_k|$$

for all  $k \in [d]$ .

*Proof.* We have that  $\Phi(\tilde{\mathbf{w}}) = \Phi(\mathbf{w})$ . First, consider  $k \neq i$ . Then, we have that  $\bar{\mathbf{Q}}(\mathbf{w})_k = \bar{\mathbf{Q}}(\tilde{\mathbf{w}})_k$ , so  $|\bar{\mathbf{Q}}(\mathbf{w})_k| = |\bar{\mathbf{Q}}(\tilde{\mathbf{w}})_k|$ , since  $Q_i$  and  $D_i$  are only functions of  $\mathbf{w}_i$  and  $\Phi(\mathbf{w})$  by our naive and scaled assumption.

Now, consider  $k = i$ . We have the following.

$$\begin{aligned}
|\bar{\mathbf{Q}}(\tilde{\mathbf{w}})_i| &= |\mathbf{D}(\mathbf{Q}(\tilde{\mathbf{w}}, \Phi(\tilde{\mathbf{w}})), \Phi(\tilde{\mathbf{w}}))_i| \\
&= |\mathbf{D}(\mathbf{Q}(\tilde{\mathbf{w}}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))_i| & (\Phi(\mathbf{w}) = \Phi(\tilde{\mathbf{w}})) \\
&= |D_i(\mathbf{Q}(\tilde{\mathbf{w}}, \Phi(\mathbf{w}))_i, \Phi(\mathbf{w}))| \\
&= |D_i(Q_i(\tilde{\mathbf{w}}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))|
\end{aligned}$$

By the naive assumption, we have that

$$Q_i(\mathbf{w}, \theta) = \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \theta)|$$

and

$$Q_i(\tilde{\mathbf{w}}, \theta) = \arg \min_{a \in \mathcal{Q}_i} |-\mathbf{w}_i - D_i(a, \theta)|.$$

Then, we have the following.

$$\begin{aligned} Q_i(\tilde{\mathbf{w}}, \theta) &= \arg \min_{a \in \mathcal{Q}_i} |-\mathbf{w}_i - D_i(a, \theta)| \\ &= \arg \min_{-a \in \mathcal{Q}_i} |-\mathbf{w}_i - D_i(-a, \theta)| && (\mathcal{Q}_i = -\mathcal{Q}_i) \\ &= \arg \min_{-a \in \mathcal{Q}_i} |-\mathbf{w}_i + D_i(a, \theta)| && (D_i(q, \theta) = -D_i(-q, \theta)) \\ &= \arg \min_{-a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \theta)| \\ &= -Q_i(\mathbf{w}, \theta) \end{aligned}$$

Thus, we have the following.

$$\begin{aligned} |\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i| &= |D_i(-Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))| \\ &= |-D_i(Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))| && (D_i(q, \theta) = -D_i(-q, \theta)) \\ &= |D_i(Q_i(\mathbf{w}, \Phi(\mathbf{w})), \Phi(\mathbf{w}))| \\ &= |\overline{\mathbf{Q}}(\mathbf{w})_i| \end{aligned}$$

□

There are four cases to consider. First, consider the case when  $0 \leq \mathbf{w}_i < \mathbf{w}_j$ . Then, by Theorem 5 and Lemma 7, we have that  $0 \leq \overline{\mathbf{Q}}(\mathbf{w})_i \leq \overline{\mathbf{Q}}(\mathbf{w})_j$ . Thus,  $|\overline{\mathbf{Q}}(\mathbf{w})_i| \leq |\overline{\mathbf{Q}}(\mathbf{w})_j|$ .

Next, consider the case where  $\mathbf{w}_j \leq \mathbf{w}_i \leq 0$ . Again, by Theorem 5 and Lemma 7, we have that  $\overline{\mathbf{Q}}(\mathbf{w}_j) \leq \overline{\mathbf{Q}}(\mathbf{w}_i) \leq 0$ . Thus,  $|\overline{\mathbf{Q}}(\mathbf{w}_i)| \leq |\overline{\mathbf{Q}}(\mathbf{w}_j)|$ .

Next, consider the case where  $\mathbf{w}_i \leq 0 \leq \mathbf{w}_j$ . Then, construct  $\tilde{\mathbf{w}}$  as described in Lemma 8. We have that  $0 \leq -\mathbf{w}_i \leq \mathbf{w}_j$ , so  $0 \leq \tilde{\mathbf{w}}_i \leq \tilde{\mathbf{w}}_j$ . Then, by the first case and Lemma 8, we have that

$$|\overline{\mathbf{Q}}(\mathbf{w})_i| = |\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i| < |\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_j| = |\overline{\mathbf{Q}}(\mathbf{w})_j|.$$

Finally, consider the case where  $\mathbf{w}_j \leq 0 \leq \mathbf{w}_i$ . Then, construct  $\tilde{\mathbf{w}}$  as described in Lemma 8. We have that  $\mathbf{w}_j \leq -\mathbf{w}_i \leq 0$ , so  $\tilde{\mathbf{w}}_j \leq \tilde{\mathbf{w}}_i \leq 0$ . Then, by the second case and Lemma 8, we have that

$$|\overline{\mathbf{Q}}(\mathbf{w})_i| = |\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i| < |\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_j| = |\overline{\mathbf{Q}}(\mathbf{w})_j|.$$

□

## A.6 ALTERNATIVE QUANTIZATION SCHEMES

However, we should note that most modern quantization schemes are not naive symmetric max-scaled block-wise quantization schemes. AdaRound (Nagel et al., 2020) was an early alternative. In particular, AdaRound violates the naive assumption, since it does not employ rounding-to-nearest. Instead, it solves a binary optimization problem to decide whether to round  $\mathbf{w}_i$  “up” or “down”. In particular,

$$Q_i(\mathbf{w}, \theta) \neq \arg \min_{a \in \mathcal{Q}_i} |\mathbf{w}_i - D_i(a, \theta)|.$$

This may round in the opposite direction of the naive assumption and violates the order-preserving property. In particular, since we decide whether to round up or round down on each weight, it is possible that  $\mathbf{w}_i$  rounds up and  $\mathbf{w}_j$  rounds down, even if  $\mathbf{w}_i < \mathbf{w}_j$ .

Recently, Activation-Aware Quantization (AWQ) (Lin et al., 2024) has emerged as a popular quantization scheme for LLMs. AWQ also violates the naive assumption. We formalize the AWQ quantization scheme presented in Lin et al. (2024) and show that it is not naive. We start with a naive symmetric max-scaled block-wise quantization scheme  $\bar{\mathbf{Q}}$ . Then, we will construct the AWQ quantization scheme,  $\tilde{\mathbf{Q}}$ .

Define a new  $\tilde{\Phi}(\mathbf{w}) = \begin{bmatrix} \Phi(\mathbf{w}) \\ \mathbf{r} \end{bmatrix}$ , where  $\mathbf{r} \in \mathbb{R}^{d'}$ .  $\mathbf{r}$  is defined in Lin et al. (2024). Then, we define

$$\tilde{Q}_i(\mathbf{w}, \theta) = Q_i(\text{diag}(\mathbf{r}) \mathbf{w}, \mathbf{s}),$$

where  $\theta = \begin{bmatrix} \mathbf{s} \\ \mathbf{r} \end{bmatrix}$ . Then, we define

$$\tilde{D}_i(q, \theta) = D_i(q, \theta).$$

AWQ’s quantization strategy is unique in that there are two different scaling factors,  $\mathbf{r}$  and  $\mathbf{s}$ . A group of weights is still on the same quantization grid, but each of the rows will have different scaling factors applied to it. Therefore, even if rounding-to-nearest is used for the original quantization grid, the additional scaling factor may break the order-preserving property. In particular, let  $\mathbf{w}_i < \mathbf{w}_j$ . Suppose that these weights are in the same group, but in different rows with scaling factors  $r_i$  and  $r_j$  respectively. Then, let

$$\left\lfloor \frac{r_i \mathbf{w}_i}{s} \right\rfloor = \left\lfloor \frac{r_j \mathbf{w}_j}{s} \right\rfloor,$$

where  $s$  is the scaling factor for the group. Then, it is possible that

$$\frac{s}{r_i} \left\lfloor \frac{r_i \mathbf{w}_i}{s} \right\rfloor > \frac{s}{r_j} \left\lfloor \frac{r_j \mathbf{w}_j}{s} \right\rfloor.$$

In particular, let  $r_i = 2$ ,  $\mathbf{w}_i = 0.4$ ,  $r_j = 1$ ,  $\mathbf{w}_j = 0.45$ , and  $s = 1$ . Then, we have

$$\frac{s}{r_i} \left\lfloor \frac{r_i \mathbf{w}_i}{s} \right\rfloor = \frac{1}{2}$$

and

$$\frac{s}{r_j} \left\lfloor \frac{r_j \mathbf{w}_j}{s} \right\rfloor = 0,$$

assuming  $Q_i = Q_j = \mathbb{Z}$ . Thus, AWQ is not a naive symmetric max-scaled block-wise quantization scheme.

## B MAGNITUDE-BASED BLOCK-WISE SPARSIFICATION

We now rigorously define *magnitude-based block-wise sparsification* as defined in Hoefler et al. (2021).

### B.1 BLOCK-WISE

We can define a “block-wise” sparsification scheme is as follows. Let  $\{M_i\}_{i=1}^d$  consists of a set of score (or saliency) functions  $M_i : \mathbb{R}^d \rightarrow \mathbb{R}$  for each weight  $\mathbf{w}_i$ . The intuition is that we can calculate the score for each weight  $\mathbf{w}_i$  and prune those that have the smallest score with some blocks. Again, like block-wise quantization, divide the weights into blocks  $\{\mathbf{w}_{(1)}, \dots, \mathbf{w}_{(B)}\}$ , defined by a partition  $\{P_1, \dots, P_B\}$  of  $[d]$ . Like in quantization, if  $B = 1$ , we call this a *tensor-wise* sparsification scheme.

Now, consider the set  $T_k = \{M_i(\mathbf{w}) : i \in P_k\}$  for all  $k \in [B]$ . Then, to achieve  $p\%$  sparsity for  $P_k$ , we select some  $t_k \geq p\% \cdot |P_k|$ . Let  $T_k^{(t_k)}$  be the  $t_k$ th element when  $T_k$  is sorted by magnitude. Then, for all  $M_i(\mathbf{w}) \leq T_k^{(t_k)}$ , we want to map the weight  $\mathbf{w}_i$  to 0. Thus, for  $i \in P_k$ , we can define  $S_i : \mathbb{R}^d \rightarrow \{0, 1\}$  such that

$$S_i(\mathbf{w}) = \begin{cases} 0 & M_i(\mathbf{w}) \leq T_k^{(t_k)} \\ 1 & \text{otherwise.} \end{cases}.$$

We have that this is a valid  $p\%$  sparsification scheme, since

$$|\{S_i(\mathbf{w}) = 0 \mid i \in [d]\}| = \sum_{k=1}^B \left| \left\{ M_i(\mathbf{w}) \leq T_k^{(t_k)} \mid i \in P_k \right\} \right| \geq \sum_{k=1}^B p\% \cdot |P_k| \geq p\% \cdot d.$$

## B.2 MAGNITUDE-BASED

As the name suggests, the magnitude-based sparsification scheme uses weight magnitude as a metric of saliency. Using a block-wise sparsification scheme, we can simply define  $M_i : \mathbb{R}^d \rightarrow \mathbb{R}$  as  $\mathbf{w} \mapsto |\mathbf{w}_i|$ . Intuitively, we set a block-wise threshold and prune all the weights with magnitude less than that threshold in that block. It is known that under the assumption that  $\mathbf{H}_L(\mathbf{w}) = \mathbb{I}_{d \times d}$  that the magnitude-based sparsification scheme is optimal.

## B.3 ALTERNATIVE SPARSIFICATION SCHEMES

As mentioned above, the magnitude-based sparsification scheme relies on the assumption that the Hessian is the identity matrix. Thus, more sophisticated sparsification schemes have been developed. For example, Optimal Brain Damage (OBD) (LeCun et al., 1989), assumes that the Hessian is diagonal, defining the score function

$$M_i(\mathbf{w}) \triangleq \frac{1}{2} h_{ii} \mathbf{w}_i^2,$$

where  $h_{ii} = \frac{\partial^2 L}{\partial \mathbf{w}_i^2}$  is the  $i$ th diagonal element of the Hessian matrix. Optimal Brain Surgeon (OBS) (Hassibi & Stork, 1992) and Optimal Brian Compression (OBC) (Frantar et al., 2023b) are two variations of this. Again, these non-examples motivate the need for a more general framework for analyzing these problems.

## C PROOF OF THEOREM 1

*Proof.* We first prove the desired result for finite  $p$ . We have that

$$\begin{aligned} \|\mathbf{w} - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))\|_p &= \left( \sum_{i=1}^d |\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i|^p \right)^{\frac{1}{p}} \\ \|\mathbf{w} - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))\|_p &= \left( \sum_{i=1}^d |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i|^p \right)^{\frac{1}{p}}. \end{aligned}$$

First, we consider the degenerate case where all weights are pruned. Then, we have that  $\overline{\mathbf{S}}(\mathbf{w}) = \mathbf{0}$ . By Lemma 7, we have that  $\overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w})) = \overline{\mathbf{Q}}(\mathbf{0}) = \mathbf{0}$ . Now, we show that  $\overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w})) = \mathbf{0}$ . For  $\overline{\mathbf{S}}$  applied to  $\mathbf{w}$ , let the threshold be  $|\mathbf{w}_t|$  for some  $t \in [d]$ . Then, we have that

$$|\{\mathbf{w}_i| \leq |\mathbf{w}_t| : i \in [d]\}| = |\{|\overline{\mathbf{Q}}(\mathbf{w})_i| \leq |\overline{\mathbf{Q}}(\mathbf{w})_t| : i \in [d]\}|$$

by Corollary 6. Thus, we have that  $\overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w})) = \mathbf{0}$  as desired.

Now, consider the case where there exists an element that is not pruned. First, we have that  $\Phi(\mathbf{w}) = \Phi(\overline{\mathbf{S}}(\mathbf{w}))$ , since the element with the largest magnitude is not pruned. Then, for  $\overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i$ , we have that

$$\overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i = \overline{\mathbf{Q}}(\mathbf{w})_i$$

if  $\mathbf{w}_i$  is not pruned. This follows from the fact that  $Q_i$  and  $D_i$  are only functions of  $\mathbf{w}_i$  and  $\Phi(\mathbf{w})$  by our naive and max-scaled assumption. If  $\mathbf{w}_i$  is pruned, then we have that  $\overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i = 0$ , since  $0 \in \mathcal{Q}_i$  for all  $i \in [d]$ . Thus, we have that

$$|\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i| = \begin{cases} |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i| & \text{if } \mathbf{w}_i \text{ is not pruned} \\ |\mathbf{w}_i - \overline{\mathbf{S}}(\mathbf{w})_i| & \text{otherwise.} \end{cases}$$

Using this, we will prove the statement

$$\sum_{i \in T} |\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i|^p \leq \sum_{i \in T} |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i|^p$$

inductively on the size of set  $T \subseteq [d]$ . Let  $n = |T|$ .

First, we consider the base case where  $n = 0$ . Then, we have that the statement is trivially true. Now, we show the inductive step. Suppose that the statement is true for some  $n = k$ . Then, we will show that the statement is true for  $n = k + 1$ . Consider some set  $T$  with cardinality  $k + 1$ .

First, consider the simpler case. In particular, consider the case where the same elements are pruned before and after quantization. Then, the elements that  $\overline{\mathbf{S}}$  prunes remains the same regardless if it is applied to  $\mathbf{w}$  or  $\overline{\mathbf{Q}}(\mathbf{w})$ . First, suppose that  $\mathbf{w}_i$  is pruned. Then, we have that

$$|\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i| = |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i| = |\mathbf{w}_i|.$$

Suppose  $\mathbf{w}_i$  is not pruned. Then,

$$|\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i| = |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i| = |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|.$$

This is true for all  $i \in [d]$ . Therefore, we have that

$$\|\mathbf{w} - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))\|_p = \|\mathbf{w} - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))\|_p.$$

Now, consider the case where different elements are pruned before and after quantization. This can only occur if there exists  $i, j \in T$  such that  $|\mathbf{w}_i| < |\mathbf{w}_j|$ , but  $|\overline{\mathbf{Q}}(\mathbf{w})_i| = |\overline{\mathbf{Q}}(\mathbf{w})_j|$ . Furthermore, suppose the  $j$ th element is pruned after quantization but not pruned before quantization, while the  $i$ th element is pruned before quantization and not after. Then, we have that

$$\begin{aligned} |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i| &= |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i| \\ |\mathbf{w}_j - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_j| &= |\mathbf{w}_j| \end{aligned}$$

Our goal will be to show that

$$|\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i|^p + |\mathbf{w}_j - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_j|^p \leq |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i|^p + |\mathbf{w}_j - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_j|^p.$$

Then, by the inductive hypothesis, we will have that

$$\sum_{i \in T \setminus \{i, j\}} |\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i|^p \leq \sum_{i \in T \setminus \{i, j\}} |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i|^p.$$

Thus, we will have that

$$\sum_{i \in T} |\mathbf{w}_i - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))_i|^p \leq \sum_{i \in T} |\mathbf{w}_i - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))_i|^p.$$

We assume that  $\mathbf{w}_i$  was pruned before quantization and not after. Thus, we have the desired inequality is

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_j|^p \leq |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|^p + |\mathbf{w}_j|^p.$$

First, suppose that  $\mathbf{w}_i$  and  $\mathbf{w}_j$  are opposite signs. Construct  $\tilde{\mathbf{w}}$  as described in Lemma 8. We have that  $\tilde{\mathbf{w}}_i = -\mathbf{w}_i$  and  $\tilde{\mathbf{w}}_j = \mathbf{w}_j$ . Furthermore, by Lemma 8, we have that  $\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i = -\overline{\mathbf{Q}}(\mathbf{w})_i$  and  $\overline{\mathbf{Q}}(\tilde{\mathbf{w}})_j = \overline{\mathbf{Q}}(\mathbf{w})_j$ . Thus, we have that  $|\tilde{\mathbf{w}}_i| = |\mathbf{w}_i|$ ,  $|\tilde{\mathbf{w}}_j| = |\mathbf{w}_j|$ ,  $|\tilde{\mathbf{w}}_i - \overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i| = |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|$ ,  $|\tilde{\mathbf{w}}_j - \overline{\mathbf{Q}}(\tilde{\mathbf{w}})_j| = |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_j|$ . Therefore, we can equivalently show that

$$|\tilde{\mathbf{w}}_i|^p + |\tilde{\mathbf{w}}_j - \overline{\mathbf{Q}}(\tilde{\mathbf{w}})_j|^p \leq |\tilde{\mathbf{w}}_i - \overline{\mathbf{Q}}(\tilde{\mathbf{w}})_i|^p + |\tilde{\mathbf{w}}_j|^p.$$

Thus, it suffices to consider when  $\mathbf{w}_i$  and  $\mathbf{w}_j$  share the same sign.

Assume that  $\mathbf{w}_i$  and  $\mathbf{w}_j$  are the same sign. Without loss of generality, assume that  $\mathbf{w}_i, \mathbf{w}_j \geq 0$ . Then, by Theorem 5 and Lemma 7 we have that  $\bar{\mathbf{Q}}(\mathbf{w})_i = \bar{\mathbf{Q}}(\mathbf{w})_j \geq 0$ . Then, we consider the following three cases:  $|\mathbf{w}_i| \leq |\bar{\mathbf{Q}}(\mathbf{w})_i| \leq |\mathbf{w}_j|$ ,  $|\mathbf{w}_i| \leq |\mathbf{w}_j| \leq |\bar{\mathbf{Q}}(\mathbf{w})_i|$ , and  $|\bar{\mathbf{Q}}(\mathbf{w})_i| \leq |\mathbf{w}_i| \leq |\mathbf{w}_j|$ .

**Case 1:**  $|\mathbf{w}_i| \leq |\bar{\mathbf{Q}}(\mathbf{w})_i| \leq |\mathbf{w}_j|$ . We want to show that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i|^p + |\mathbf{w}_j|^p.$$

We claim that  $|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p$  is decreasing on  $\bar{\mathbf{Q}}(\mathbf{w})_i$  assuming that  $\mathbf{w}_i \leq \bar{\mathbf{Q}}(\mathbf{w})_i \leq \mathbf{w}_j$ . We have that  $|\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i| = \mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i$ . Thus, we have that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p = |\mathbf{w}_i|^p + (\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i)^p,$$

which is clearly decreasing on  $\bar{\mathbf{Q}}(\mathbf{w})_i$ . Thus, we have that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_i|^p + |\mathbf{w}_j - \mathbf{w}_i|^p.$$

Then, we claim that  $|\mathbf{w}_i|^p + |\mathbf{w}_j - \mathbf{w}_i|^p \leq |\mathbf{w}_j|^p$ . We will equivalently show that

$$|\mathbf{w}_i|^p \leq |\mathbf{w}_j|^p - |\mathbf{w}_j - \mathbf{w}_i|^p.$$

We have that

$$|\mathbf{w}_i|^p = \mathbf{w}_i^p = \int_0^{\mathbf{w}_i} px^{p-1} dx$$

and

$$|\mathbf{w}_j|^p - |\mathbf{w}_j - \mathbf{w}_i|^p = \int_0^{\mathbf{w}_i} p(\mathbf{w}_j - \mathbf{w}_i + x)^{p-1} dx.$$

We have that  $\mathbf{w}_i < \mathbf{w}_j$ , so  $px^{p-1} < p(\mathbf{w}_j - \mathbf{w}_i + x)^{p-1}$  for  $x \in [0, \mathbf{w}_i]$ . Thus, we have that

$$\int_0^{\mathbf{w}_i} px^{p-1} dx \leq \int_0^{\mathbf{w}_i} p(\mathbf{w}_j - \mathbf{w}_i + x)^{p-1} dx$$

and

$$|\mathbf{w}_i|^p \leq |\mathbf{w}_j|^p - |\mathbf{w}_j - \mathbf{w}_i|^p.$$

Thus, we have that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_i|^p + |\mathbf{w}_j - \mathbf{w}_i|^p \leq |\mathbf{w}_j|^p.$$

Finally, we have that

$$|\mathbf{w}_j|^p \leq |\mathbf{w}_j|^p + |\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i|^p,$$

since  $|\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i| \geq 0$ . Therefore, we have proved the desired inequality.

**Case 2:**  $|\mathbf{w}_i| \leq |\mathbf{w}_j| \leq |\bar{\mathbf{Q}}(\mathbf{w})_i|$ . We want to show that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i|^p + |\mathbf{w}_j|^p.$$

Rearranging, we have that

$$|\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i|^p - |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p \geq |\mathbf{w}_i|^p - |\mathbf{w}_j|^p.$$

First, we claim that  $|\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i|^p - |\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i|^p$  is increasing in  $\bar{\mathbf{Q}}(\mathbf{w})_i$ . Since  $|\mathbf{w}_i| \leq |\mathbf{w}_j| \leq |\bar{\mathbf{Q}}(\mathbf{w})_i|$  and we are assuming that  $\mathbf{w}_i, \mathbf{w}_j, \bar{\mathbf{Q}}(\mathbf{w})_i \geq 0$ , then we have that

$$|\mathbf{w}_j - \bar{\mathbf{Q}}(\mathbf{w})_i| = \bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_j$$

and

$$|\mathbf{w}_i - \bar{\mathbf{Q}}(\mathbf{w})_i| = \bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_i.$$

Thus, it suffices to show that  $(\bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_i)^p - (\bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_j)^p$  is increasing in  $\bar{\mathbf{Q}}(\mathbf{w})_i$ . We have that

$$(\bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_i)^p - (\bar{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_j)^p = \int_{-\mathbf{w}_j}^{-\mathbf{w}_i} p(\bar{\mathbf{Q}}(\mathbf{w})_i + x)^{p-1} dx.$$



We have that  $p(\overline{\mathbf{Q}}(\mathbf{w})_i + x)^{p-1}$  is increasing in  $\overline{\mathbf{Q}}(\mathbf{w})_i$ , and  $-\mathbf{w}_j < -\mathbf{w}_i$  by assumption. Therefore,  $(\overline{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_i)^p - (\overline{\mathbf{Q}}(\mathbf{w})_i - \mathbf{w}_j)^p$  is increasing in  $\overline{\mathbf{Q}}(\mathbf{w})_i$ . Therefore, we have that

$$|\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|^p - |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_i|^p \geq |\mathbf{w}_i - \mathbf{w}_j|^p - |\mathbf{w}_j - \mathbf{w}_j|^p = |\mathbf{w}_i - \mathbf{w}_j|^p > 0.$$

Furthermore, we have that  $|\mathbf{w}_i|^p - |\mathbf{w}_j|^p < 0$ , since  $0 \leq \mathbf{w}_i < \mathbf{w}_j$ . Therefore, the desired inequality is shown.

**Case 3:**  $|\overline{\mathbf{Q}}(\mathbf{w})_i| \leq |\mathbf{w}_i| \leq |\mathbf{w}_j|$ . Again, we want to show that

$$|\mathbf{w}_i|^p + |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|^p + |\mathbf{w}_j|^p.$$

Rearranging, we have that

$$|\mathbf{w}_i|^p - |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|^p \leq |\mathbf{w}_j|^p - |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_i|^p.$$

Then, we have that

$$|\mathbf{w}_i|^p - |\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i|^p = \int_{-\overline{\mathbf{Q}}(\mathbf{w})_i}^0 p(\mathbf{w}_j + x)^{p-1} dx$$

and

$$|\mathbf{w}_j|^p - |\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_i|^p = \int_{-\overline{\mathbf{Q}}(\mathbf{w})_i}^0 p(\mathbf{w}_i + x)^{p-1} dx.$$

We have that  $-\overline{\mathbf{Q}}(\mathbf{w})_i < 0$  and  $\mathbf{w}_i < \mathbf{w}_j$ . Thus,

$$\int_{-\overline{\mathbf{Q}}(\mathbf{w})_i}^0 p(\mathbf{w}_j + x)^{p-1} dx < \int_{-\overline{\mathbf{Q}}(\mathbf{w})_i}^0 p(\mathbf{w}_i + x)^{p-1} dx.$$

Therefore, we have the inequality as desired.

This proves the statement for finite  $p$ . Note that the  $L_\infty$  norm is the limit of the  $L_p$  norm as  $p \rightarrow \infty$ . Thus, using the proven inequality and taking the limit as  $p \rightarrow \infty$ , this proves the corresponding result for the  $L_\infty$  norm.  $\square$

## D COUNTEREXAMPLE FOR NON-NAIVE QUANTIZATION SCHEMES

Theorem 1 no longer holds under non-naive quantization schemes. We provide a counterexample. Let

$$\mathbf{w} = \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}.$$

Let  $Q_i = \frac{3}{2}\mathbb{Z}$  for all  $i \in [d]$ . Furthermore, let  $\Phi, D_i$  be defined the same way as in a naive symmetric max-scaled block-wise quantization scheme. However, we modify  $Q_i$  slightly. In particular, like AdaRound, we choose a random direction to round. Thus, we have that

$$\Phi(\mathbf{w}) = [0.9].$$

Then, suppose that

$$\mathbf{Q}(\mathbf{w}) = \begin{bmatrix} 0 \\ \frac{3}{2} \end{bmatrix}.$$

Then, we have that

$$\overline{\mathbf{Q}}(\mathbf{w}) = \begin{bmatrix} 0 \\ 1.35 \end{bmatrix}.$$

Now, suppose  $\overline{\mathbf{S}}$  is a 50% magnitude-based block-wise sparsification scheme. Then, we have that

$$\overline{\mathbf{S}}(\mathbf{w}) = \begin{bmatrix} 0 \\ 1.8 \end{bmatrix}.$$

Furthermore, let

$$\mathbf{Q}(\overline{\mathbf{S}}(\mathbf{w})) = \overline{\mathbf{Q}}\left(\begin{bmatrix} 0 \\ 1.8 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Here, we assume that we randomly choose to round down from 0.9 to 0, instead of round up to 1.35.

Then, we have that

$$\|\mathbf{w} - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))\|_1 = \left\| \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix} \right\|_1 = 1.$$

We also have that

$$\|\mathbf{w} - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))\|_1 = \left\| \begin{bmatrix} 0.1 \\ -0.45 \end{bmatrix} \right\|_1 = 0.55,$$

since

$$\overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w})) = \begin{bmatrix} 0 \\ 1.35 \end{bmatrix}.$$

Therefore, it is also not necessarily true that

$$\|\mathbf{w} - \overline{\mathbf{Q}}(\overline{\mathbf{S}}(\mathbf{w}))\|_1 \leq \|\mathbf{w} - \overline{\mathbf{S}}(\overline{\mathbf{Q}}(\mathbf{w}))\|_1.$$

## E PROOF OF THEOREM 3

*Proof.* We use induction on  $n$ , defined as the number of entries that are sparsified in exactly one of  $\mathbf{Q} \rightarrow \mathbf{S}$  and  $\mathbf{S} \rightarrow \mathbf{Q}$ . The base case,  $n = 0$ , is trivial, since the weights will be the same. Otherwise, consider any indices  $i, j$  that are pruned differently by the two procedures. Since we only care about the magnitudes of the weights, like in the proof for Theorem 1, assume that  $\mathbf{w}_i \geq 0, \mathbf{w}_j \geq 0$ . Without loss of generality, we further assume that  $\mathbf{w}_i \leq \mathbf{w}_j$ . We know that  $\overline{\mathbf{Q}}(\mathbf{w})_i \leq \overline{\mathbf{Q}}(\mathbf{w})_j$  by Theorem 5. Let  $h_{ii}$  be the  $i$ th diagonal element of  $\mathbf{H}_L(\mathbf{w})$ .

Note that we cannot have either of  $\mathbf{w}_i$  equal to zero, since the only way  $\mathbf{Q} \rightarrow \mathbf{S}$  and  $\mathbf{S} \rightarrow \mathbf{Q}$  could prune different weights would be for  $\overline{\mathbf{Q}}(\mathbf{w}_j) = 0$ , but this would imply that elements  $i$  and  $j$  of  $\mathbf{w}_{\mathbf{Q} \rightarrow \mathbf{S}}$  and  $\mathbf{w}_{\mathbf{S} \rightarrow \mathbf{Q}}$  are equal, which is a contradiction. Thus, we can restrict without loss of generality to the case  $\mathbf{w}_i > 0, \mathbf{w}_j > 0$ .

Then there are two cases.

**Case 1:**  $h_{ii}\mathbf{w}_i^2 \geq h_{jj}\mathbf{w}_j^2$ . Since the two procedures prune different weights, this implies  $h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 < h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2$ . This implies  $\mathbf{S} \rightarrow \mathbf{Q}$  will prune  $\mathbf{w}_j$ , while  $\mathbf{Q} \rightarrow \mathbf{S}$  will prune  $\mathbf{w}_i$ . We wish to show

$$L(\hat{\mathbf{w}}_{\mathbf{S} \rightarrow \mathbf{Q}}) \leq L(\hat{\mathbf{w}}_{\mathbf{Q} \rightarrow \mathbf{S}}) + \mathcal{O}(\|\hat{\mathbf{w}}_{\mathbf{Q} \rightarrow \mathbf{S}} - \mathbf{w}\|^3),$$

Note that when using a second-order Taylor expansion,  $\mathcal{O}(\|\hat{\mathbf{w}}_{\mathbf{S} \rightarrow \mathbf{Q}} - \mathbf{w}\|^3) \subseteq \mathcal{O}(\|\hat{\mathbf{w}}_{\mathbf{Q} \rightarrow \mathbf{S}} - \mathbf{w}\|^3)$  by Theorem 1. Then using the inductive hypothesis, it suffices to show the terms indexed by  $i$  and  $j$  in the loss satisfy the desired inequality, so it suffices to show that

$$h_{ii}(\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i)^2 + h_{jj}\mathbf{w}_j^2 \leq h_{ii}\mathbf{w}_i^2 + h_{jj}(\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_j)^2$$

which can be written as

$$2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) + (h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2) \leq 0.$$

We already know the second term  $h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2 < 0$ . Then, note that we have

$$\frac{\mathbf{w}_j}{\mathbf{w}_i} \leq \frac{\sqrt{h_{ii}}}{\sqrt{h_{jj}}}.$$

For a quantization grid of step size  $\delta$  with  $0 < \delta < 2\mathbf{w}_i$ ,

$$\frac{\overline{\mathbf{Q}}(\mathbf{w})_j}{\overline{\mathbf{Q}}(\mathbf{w})_i} \leq \frac{\mathbf{w}_j + \delta/2}{\mathbf{w}_i - \delta/2} \leq \frac{\mathbf{w}_j}{\mathbf{w}_i} \cdot \frac{1 + \frac{\delta}{2\mathbf{w}_j}}{1 - \frac{\delta}{2\mathbf{w}_i}} \leq \frac{\sqrt{h_{ii}}}{\sqrt{h_{jj}}} \cdot \frac{1 + \frac{\delta}{2\mathbf{w}_j}}{1 - \frac{\delta}{2\mathbf{w}_i}}.$$

Let  $c_1(\delta) = (1 + \frac{\delta}{2\mathbf{w}_j})/(1 - \frac{\delta}{2\mathbf{w}_i})$ , where we know that as  $\lim_{\delta \rightarrow 0} c_1(\delta) = 1$ . Combining the two inequalities, this implies that

$$\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j \leq \frac{h_{ii}}{h_{jj}}c_1(\delta)\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i,$$

so that

$$2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) \leq 2h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i(c_1(\delta) - 1),$$

which approaches zero as  $\delta \rightarrow 0$ . Since we know that  $h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2 < 0$ , this implies there exists some  $\varepsilon_{i,j,1} > 0$  such that  $\delta < \varepsilon_{i,j,1}$  and  $\sqrt{h_{ii}}\mathbf{w}_i \geq \sqrt{h_{jj}}\mathbf{w}_j$  imply  $2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) + (h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2) \leq 0$ .

**Case 2:**  $h_{ii}\mathbf{w}_i^2 \leq h_{jj}\mathbf{w}_j^2$ . Since the two procedures prune different weights, this implies  $h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 \geq h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2$ . This implies  $\mathbf{S} \rightarrow \mathbf{Q}$  will prune  $\mathbf{w}_i$ , while  $\mathbf{Q} \rightarrow \mathbf{S}$  will prune  $\mathbf{w}_j$ . Hence, since only two terms in the loss function are differing, we can perform similar steps to case 1. By the inductive hypothesis and a second-order Taylor expansion, it suffices to show that

$$h_{ii}\mathbf{w}_i^2 + h_{jj}(\mathbf{w}_j - \overline{\mathbf{Q}}(\mathbf{w})_j)^2 \leq h_{ii}(\mathbf{w}_i - \overline{\mathbf{Q}}(\mathbf{w})_i)^2 + h_{jj}\mathbf{w}_j^2$$

which reduces to

$$2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) + (h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2) \geq 0$$

We already know the second term  $h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2 \geq 0$ . For a quantization grid of step size  $\delta$ , when  $0 < \delta < 2\mathbf{w}_j$ ,

$$\frac{\overline{\mathbf{Q}}(\mathbf{w})_j}{\overline{\mathbf{Q}}(\mathbf{w})_i} \geq \frac{\mathbf{w}_j - \delta/2}{\mathbf{w}_i + \delta/2} = \frac{\mathbf{w}_j}{\mathbf{w}_i} \cdot \frac{1 - \frac{\delta}{2\mathbf{w}_j}}{1 + \frac{\delta}{2\mathbf{w}_i}} \geq \frac{\sqrt{h_{ii}}}{\sqrt{h_{jj}}} \cdot \frac{1 - \frac{\delta}{2\mathbf{w}_j}}{1 + \frac{\delta}{2\mathbf{w}_i}},$$

where we define  $c_2(\delta) = (1 - \frac{\delta}{2\mathbf{w}_j})/(1 + \frac{\delta}{2\mathbf{w}_i})$ . Combining this inequality with the fact that  $h_{jj}\mathbf{w}_j^2 \geq h_{ii}\mathbf{w}_i^2$ , we see that

$$\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j \geq \frac{h_{ii}}{h_{jj}}c_2(\delta)\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i,$$

so that

$$2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) \geq 2h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i(c_w(\delta) - 1),$$

which approaches zero as  $\delta \rightarrow 0$ . Thus, this shows that this implies there exists some  $\varepsilon_{i,j,2} > 0$  such that  $\delta < \varepsilon_{i,j,2}$  and  $\sqrt{h_{ii}}\mathbf{w}_i \leq \sqrt{h_{jj}}\mathbf{w}_j$  imply  $2(h_{jj}\mathbf{w}_j\overline{\mathbf{Q}}(\mathbf{w})_j - h_{ii}\mathbf{w}_i\overline{\mathbf{Q}}(\mathbf{w})_i) + (h_{ii}\overline{\mathbf{Q}}(\mathbf{w})_i^2 - h_{jj}\overline{\mathbf{Q}}(\mathbf{w})_j^2) \geq 0$ .

This completes the induction, so we have shown that

$$L(\hat{\mathbf{w}}_{\mathbf{S} \rightarrow \mathbf{Q}}) \leq L(\hat{\mathbf{w}}_{\mathbf{Q} \rightarrow \mathbf{S}}) + \mathcal{O}(\|\hat{\mathbf{w}}_{\mathbf{Q} \rightarrow \mathbf{S}} - \mathbf{w}\|^3).$$

Since there are finitely many inductive steps, this bound holds for some  $\varepsilon > 0$ . For example, we can conservatively take  $\varepsilon = \min\{\varepsilon_{i,j,k}\}_{k \in \{1,2\}, i \neq j}$ .  $\square$

## F PROOF OF THEOREM 4

*Proof.* Suppose we have a quantization scheme followed by a sparsification scheme that introduces a  $\Delta\mathbf{w}_{\mathbf{Q}} + \Delta\mathbf{w}_{\mathbf{S}|\mathbf{Q}}$  change to the weights. Then, we had the following, assuming that  $\mathbf{w}$  was a local optimum.

$$\begin{aligned} & L(\mathbf{w} + \Delta\mathbf{w}_{\mathbf{Q}} + \Delta\mathbf{w}_{\mathbf{S}|\mathbf{Q}}) - L(\mathbf{w}) \\ &= \frac{1}{2}(\Delta\mathbf{w}_{\mathbf{Q}} + \Delta\mathbf{w}_{\mathbf{S}|\mathbf{Q}})^\top \mathbf{H}_L(\mathbf{w})(\Delta\mathbf{w}_{\mathbf{Q}} + \Delta\mathbf{w}_{\mathbf{S}|\mathbf{Q}}) + \mathcal{O}(\|\Delta\mathbf{w}_{\mathbf{Q}} + \Delta\mathbf{w}_{\mathbf{S}|\mathbf{Q}}\|^3) \end{aligned}$$

Similarly, we assume that we have a sparsification scheme followed by a quantization scheme that introduces  $\Delta\mathbf{w}_{\mathbf{S}} + \Delta\mathbf{w}_{\mathbf{Q}|\mathbf{S}}$  change to the weights. Then, we had the following, assuming that  $\mathbf{w}$  was a local optimum.

$$\begin{aligned} & L(\mathbf{w} + \Delta\mathbf{w}_{\mathbf{S}} + \Delta\mathbf{w}_{\mathbf{Q}|\mathbf{S}}) - L(\mathbf{w}) \\ &= \frac{1}{2}(\Delta\mathbf{w}_{\mathbf{S}} + \Delta\mathbf{w}_{\mathbf{Q}|\mathbf{S}})^\top \mathbf{H}_L(\mathbf{w})(\Delta\mathbf{w}_{\mathbf{S}} + \Delta\mathbf{w}_{\mathbf{Q}|\mathbf{S}}) + \mathcal{O}(\|\Delta\mathbf{w}_{\mathbf{S}} + \Delta\mathbf{w}_{\mathbf{Q}|\mathbf{S}}\|^3) \end{aligned}$$

Using the fact that  $\varepsilon_S = \Delta \mathbf{w}_{S|Q} - \Delta \mathbf{w}_S$ , we get the following for the change in loss under  $Q \rightarrow S$ .

$$\begin{aligned}
& L(\mathbf{w} + \Delta \mathbf{w}_Q + \Delta \mathbf{w}_{S|Q}) - L(\mathbf{w}) \\
&= \frac{1}{2} (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S + \varepsilon_S)^\top \mathbf{H}_L(\mathbf{w}) (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S + \varepsilon_S) + \mathcal{O}(\|\Delta \mathbf{w}_Q + \Delta \mathbf{w}_{S|Q}\|^3) \\
&= \frac{1}{2} (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S)^\top \mathbf{H}_L(\mathbf{w}) (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S) \\
&\quad + \varepsilon_S^\top \mathbf{H}_L(\mathbf{w}) (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S) \\
&\quad + \frac{1}{2} \varepsilon_S^\top \mathbf{H}_L(\mathbf{w}) \varepsilon_S + \mathcal{O}(\|\Delta \mathbf{w}_Q + \Delta \mathbf{w}_{S|Q}\|^3)
\end{aligned}$$

We have the following for  $S \rightarrow Q$  under the assumption that  $\Delta \mathbf{w}_{Q|S} = \Delta \mathbf{w}_Q$ .

$$\begin{aligned}
& L(\mathbf{w} + \Delta \mathbf{w}_S + \Delta \mathbf{w}_{Q|S}) - L(\mathbf{w}) \\
&= \frac{1}{2} (\Delta \mathbf{w}_S + \Delta \mathbf{w}_Q)^\top \mathbf{H}_L(\mathbf{w}) (\Delta \mathbf{w}_S + \Delta \mathbf{w}_Q) + \mathcal{O}(\|\Delta \mathbf{w}_S + \Delta \mathbf{w}_{Q|S}\|^3)
\end{aligned}$$

Now, we take the difference between the change in loss for  $Q \rightarrow S$  and  $S \rightarrow Q$ .

$$\begin{aligned}
& L(\mathbf{w} + \Delta \mathbf{w}_Q + \Delta \mathbf{w}_{S|Q}) - L(\mathbf{w} + \Delta \mathbf{w}_S + \Delta \mathbf{w}_{Q|S}) \\
&= \varepsilon_S^\top \mathbf{H}_L(\mathbf{w}) (\Delta \mathbf{w}_Q + \Delta \mathbf{w}_S) + \frac{1}{2} \varepsilon_S^\top \mathbf{H}_L(\mathbf{w}) \varepsilon_S + \mathcal{O}(\|\Delta \mathbf{w}_{Q \rightarrow S}\|^3) + \mathcal{O}(\|\Delta \mathbf{w}_{S \rightarrow Q}\|^3)
\end{aligned}$$

□