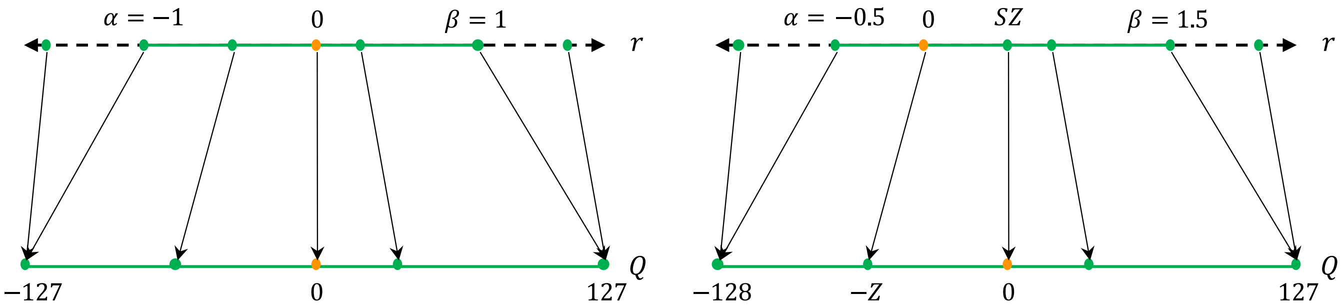

Figure 2: *Illustration of symmetric quantization and asymmetric quantization. Symmetric quantization with restricted range maps real values to [-127, 127], and full range maps to [-128, 127] for 8-bit quantization.*

function takes real values in floating point, and it maps them to a lower precision range, as illustrated in Figure 1. A popular choice for a quantization function is as follows:

$$Q(r) = \text{Int}\big(r/S\big) - Z, \tag{2}$$

where Q is the quantization operator, r is a real valued input (activation or weight), S is a real valued scaling factor, and Z is an integer zero point. Furthermore, the Int function maps a real value to an integer value through a rounding operation (e.g., round to nearest and truncation). In essence, this function is a mapping from real values r to some integer values. This method of quantization is also known as *uniform quantization*, as the resulting quantized values (aka quantization levels) are uniformly spaced (Figure 1, left). There are also *non-uniform quantization* methods whose quantized values are not necessarily uniformly spaced (Figure 1, right), and these methods will be discussed in more detail in Section III-F. It is possible to recover real values r from the quantized values $Q(r)$ through an operation that is often referred to as *dequantization*:

$$\tilde{r} = S(Q(r) + Z). \tag{3}$$

Note that the recovered real values $\tilde{r}$ will not exactly match r due to the rounding operation.

C. Symmetric and Asymmetric Quantization

One important factor in uniform quantization is the choice of the scaling factor S in Eq. 2. This scaling factor essentially divides a given range of real values r into a number of partitions (as discussed in [113, 133]):

$$S = \frac{\beta - \alpha}{2^b - 1}, \tag{4}$$

where $[\alpha, \beta]$ denotes the clipping range, a bounded range that we are clipping the real values with, and b is the quantization bit width. Therefore, in order for the scaling factor to be defined, the clipping range $[\alpha, \beta]$ should first be determined. The process of choosing the clipping range is often referred to as *calibration*. A straightforward choice is to use the min/max of the signal for the clipping range, i.e., $\alpha = r_{min}$, and $\beta = r_{max}$. This approach is an *asymmetric quantization* scheme, since the clipping range is not necessarily symmetric with respect to the origin, i.e., $-\alpha \neq \beta$, as illustrated in Figure 2 (Right). It is also possible to use a *symmetric quantization* scheme by choosing a symmetric clipping range of $\alpha = -\beta$. A popular choice is to choose these based on the min/max values of the signal: $-\alpha = \beta = \max(|r_{max}|, |r_{min}|)$. Asymmetric quantization often results in a tighter clipping range as compared to symmetric quantization. This is especially important when the target weights or activations are imbalanced, e.g., the activation after ReLU that always has non-negative values. Using symmetric quantization, however, simplifies the quantization function in Eq. 2 by replacing the zero point with $Z = 0$:

$$Q(r) = \text{Int}\left(\frac{r}{S}\right). \tag{5}$$

Here, there are two choices for the scaling factor. In "full range" symmetric quantization S is chosen as $\frac{2max(|r|)}{2^n-1}$ (with floor rounding mode), to use the full INT8 range of [-128,127]. However, in "restricted range" S is chosen as $\frac{max(|r|)}{2^{n-1}-1}$, which only uses the range of [-127,127]. As expected, the full range approach is more accurate. Symmetric quantization is widely adopted in practice for quantizing weights because zeroing out the zero point can lead to reduction in computational cost during inference [255], and also makes the implementation more straightforward. However, note that for activation the cross terms occupying due to the offset in the asymmetric activations are a static data independent term