How to Parameterize Asymmetric Quantization Ranges for Quantization-Aware Training

Jaeseong You, Minseop Park, Kyunggeun Lee, Seokjun An, Chirag Patel, & Markus Nagel
Qualcomm AI Research [*]
{jaeseong,minspark,kyunggeu,seokan,cpatel,markusn}@qti.qualcomm.com

Abstract

This paper investigates three different parameterizations of asymmetric uniform quantization for quantization-aware training: (1) scale and offset, (2) minimum and maximum, and (3) beta and gamma. We perform a comprehensive comparative analysis of these parameterizations' influence on quantization-aware training, using both controlled experiments and real-world large language models. Our particular focus is on their changing behavior in response to critical training hyperparameters, bit width and learning rate. Based on our investigation, we propose best practices to stabilize and accelerate quantization-aware training with learnable asymmetric quantization ranges.

1 Introduction

In settings with limited low-resources, such as on-device applications or in developing countries, model efficiency is critical. Quantization serves as a practical and effective solution to this end (Kuzmin et al., 2023). In the field of deep learning, quantization refers to the method of mapping floating-point values (i.e., model weights or intermediate activations) to lower-bit integers. The benefits are two-fold: it reduces memory footprint and accelerates computation. The demand for quantization has increased as neural networks have grown in size to achieve state-of-the-art performance. Large language models (LLMs) have been a driving force behind this trend in recent years (Brown et al., 2020; Zhang et al., 2022; Chowdhery et al., 2023; Luccioni et al., 2023; Hoffmann et al., 2022; Touvron et al., 2023), and similar patterns are also evident across various domains (OpenAI, 2023; Dehghani et al., 2023; Chu et al., 2023).

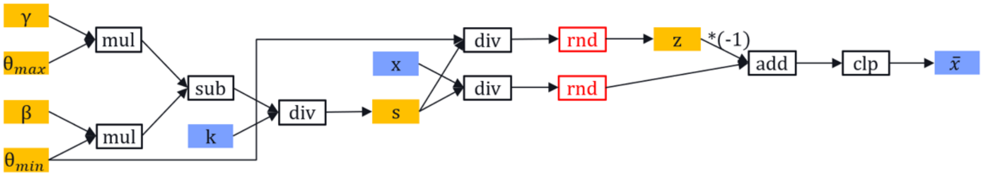

Figure 1: Computational graph of asymmetric quantization.

Asymmetric uniform quantization and dequantization are defined as follows:

$$\begin{aligned}
&\bar{x} = Q(x, s, z, k) = \text{clip}\left(\left\lfloor \frac{x}{s} \right\rceil - \lfloor z \rceil, 0, k\right), \\
&\hat{x} = DQ(\bar{x}, s, z) = s(\bar{x} + \lfloor z \rceil), \qquad\qquad (1) \\
&\text{where} \quad k = 2^b - 1, \quad s = \frac{\theta_{max} - \theta_{min}}{k}, \quad z = \frac{\theta_{min}}{s}.
\end{aligned}$$

Here, θ_{min} and θ_{max} are typically initialized to the minimum and maximum values of the input data x, and b is the target bit width. While in quantization-aware-training (QAT) with learnable asymmetric quantization ranges, the standard practice is to learn s and z (Bhalgat et al., 2020), one can

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

opt to set other pairs of parameters as learnable, rather than s and z (denoted as *scale/offset* hereafter). The yellow boxes in Figure 1 illustrate that these learnable candidates could be either θ_{min} and θ_{max} (denoted as *min/max* hereafter) or β and γ (denoted as *beta/gamma* hereafter) as well. In this paper, we (1) demonstrate that the learning patterns of these asymmetric parameterizations can be different from one another during QAT, (2) provide a comparative analysis of their differences, and (3) propose best practices to stabilize and accelerate QAT.

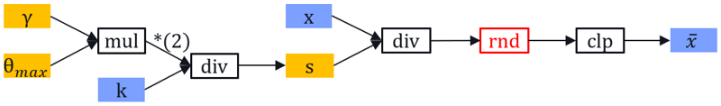

Figure 2: Computational graph of symmetric quantization.

The exploration of differences between these parameterizations is timely, as an increasing number of studies are focusing on learning-based optimization of quantization ranges, especially for extreme low-bit quantization of LLMs. These efforts include not only conventional QAT approaches (Liu et al., 2023b; Kim et al., 2023b; Wu et al., 2023), but also quasi-QAT methods based on local—block-wise or layer-wise—optimization (Lin et al., 2023; Shao et al., 2023; Ding et al., 2023). Our study has the potential to offer insights and benefits in both contexts.

2 Related Works

There are two uniform quantization schema that are widely employed: symmetric (depicted in Figure 2) and asymmetric (depicted in Figure 1). The quantization method can also be largely categorized into two: Post-training quantization (PTQ) and QAT. PTQ obtains effective quantization ranges with no (or minimal) modification of model weights (Gong et al., 2018; Banner et al., 2019). On the other hand, QAT learns model weights with the quantization effect taken into account. This is commonly achieved by using the straight-through estimator for the non-differentiable rounding operation (Bengio et al., 2013).

In QAT with range-leaning, the quantization ranges themselves are learned, either independently with the model weights frozen (Kim et al., 2023a) or jointly with the model weights (Esser et al., 2020). The idea of learning quantization ranges was initially introduced in symmetric form (Choi et al., 2018; Esser et al., 2020). The concept of learnable range was then extended to asymmetric quantization. Bhalgat et al. (2020) utilized the *scale/offset*, while Siddegowda et al. (2022) adopted the *min/max*. Furthermore, Shao et al. (2023) introduced a novel *beta/gamma* parameterization derived from *min/max*. For a more in-depth understanding of quantization fundamentals, please refer to Krishnamoorthi (2018); Nagel et al. (2021); Siddegowda et al. (2022).

3 Comparative Analysis of Asymmetric QAT Parameterizations

In symmetric quantization, all the learnable range parameters, s, θ_{max} and γ, depend linearly on one another, as shown in Figure 2. This results in gradients that are identical except for scaling factors (see Table 3 in the Appendix). However, in asymmetric quantization, the two range parameters are mutually dependent as in Figure 1, resulting in complex gradients as in Table 1. See A. 2 in the Appendix why they can lead to different solutions after training.

***scale/offset* vs. *min/max*.** Given the different QAT behaviors exhibited by the three parameterizations, the question naturally arises: which one should we use? Let us first compare *scale/offset* and *min/max*. One potential problem with *scale/offset* is that s and z reside in different spaces, forming an inverse relation to one another as in equation 1. Assigning identical learning rates to them would thus not be sensible, and it is unclear how to appropriately assign different rates (see the Appendix for three possible options). Another issue arises becasuse the gradients of s and z do not incorporate k, which means they cannot properly respond to changes in bit width. On the other hand, the gradients of θ_{min} and θ_{max} incorporate k as in Table 1, reducing bit-width sensitivity.

An additional interesting observation about *scale/offset* is that it is prone to error in situations where one of θ_{min} and θ_{max} is on its optimal point and the other is not. Once one quantization encoding

	$n < x < p$	$x < n$	$x > p$
$\frac{d\hat{x}}{ds}$	$\lfloor \frac{x}{s} \rceil - \frac{x}{s}$	n	p
$\frac{d\hat{x}}{dz}$	0	1	1
$\frac{d\hat{x}}{d\theta_{min}}$	$-\frac{1}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$-\frac{n}{k} + \frac{\lfloor z \rceil - z}{k} + 1$	$\frac{\lfloor z \rceil - z}{k}$
$\frac{d\hat{x}}{d\theta_{max}}$	$\frac{1}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$\frac{n}{k} - \frac{\lfloor z \rceil - z}{k}$	$-\frac{\lfloor z \rceil - z}{k} + 1$
$\frac{d\hat{x}}{d\beta}$	$-\theta_{min}\frac{1}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$\theta_{min}(-\frac{n}{k} + \frac{\lfloor z \rceil - z}{k} + 1)$	$\theta_{min}\frac{\lfloor z \rceil - z}{k}$
$\frac{d\hat{x}}{d\gamma}$	$\theta_{max}\frac{1}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$\theta_{max}(\frac{n}{k} - \frac{\lfloor z \rceil - z}{k})$	$\theta_{max}(-\frac{\lfloor z \rceil - z}{k} + 1)$

Table 1: Gradients of asymmetric quantization ranges.

reaches a local minimum, oscillation starts due to the push-and-pull between the clipping error and the quantization error. This could cause unwanted irregularities on the other encoding that has not yet converged. A good example is ReLU. While *min/max* can simply fixate θ_{min} at 0 and learn only θ_{max}, *scale/offset* is required to move both s and z simultaneously at all time, which makes it more vulnerable to unstable oscillation (see Figure 8 in the Appendix).

To confirm whether the aforementioned issues indeed impede the QAT performance of *scale/offset*, we perform a controlled toy experiment. We quantize a tensor of 10,000 values that follow a normal distribution. To examine bit-width sensitivity, we try low bit (3 bit) and high bit (10 bit). We also compare learning rates of 1e-2 and 5e-3, thereby ablating the impact of learning rate. The quantization range is learned to minimize the mean-squared-error (MSE) between the original tensor and the quantized-dequantized tensor:

$$\underset{enc_a, enc_b}{\arg\min} \frac{1}{N} \sum_{i}^{N} (DQ(Q(x_i, enc_a, enc_b, k), enc_a, enc_b) - x_i)^2. \tag{2}$$

Here, enc_a and enc_b are the learned quantization encodings (i.e. s and z or θ_{min} and θ_{max}). We use the Adam optimizer with no weight decay (Kingma & Ba, 2015). The initial encoding θ^0_{min} is set to the minimum value of the tensor while θ^0_{max} is set to three times larger than the maximum value of the tensor. This is done to make the task sufficiently challenging by giving the quantizer longer asymmetric distances to manage. As observed in Figure 3, *scale/offset* responds sensitively to the learning rate and fails to converge in the high-bit case. On the other hand, *min/max* converges consistently in all scenarios.

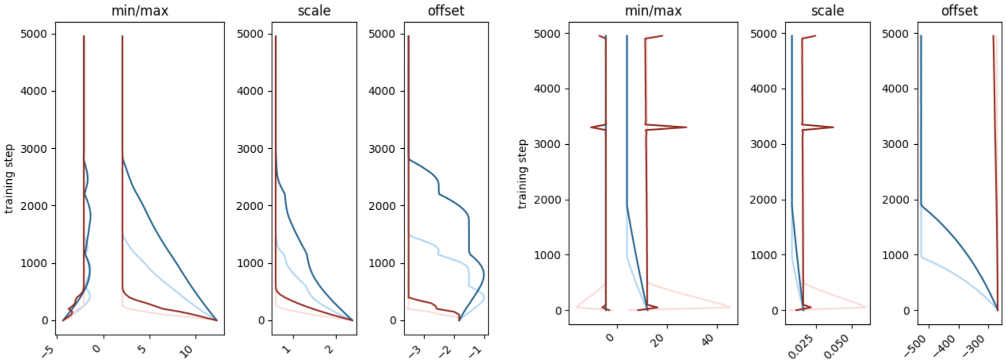

Figure 3: Learnable ranges of *scale/offset* and *min/max* (x-axis) changing over 5k steps of QAT (y-axis). *scale/offset* and *min/max* are respectively color-coded as red and blue, and lighter shades correspond to a learning rate of 1e-2 (darker shades to that of 1e-3). The left subfigure represents 3-bit quantization (10-bit on the right). Although we experimented with 16 bit as well, *scale/offset* resulted in excessively large values that could not be effectively visualized.

Extending the comparison between *scale/offset* and *min/max* to a real-life scenario, we perform QAT of GPT2-small on WikiText-2 (Merity et al., 2017), as shown in Figure 4. All the weights are quantized to symmetric 4-bit integers, and all the activations are quantized to asymmetric 12-bit integers. The only exception is the layernorm weights, which follow the quantization scheme of the activations. The quantization ranges for both the weights and the activations are learned using a batch size of 8, while the model weights remain frozen. This experiment reaffirms the instability of the *scale/offset* method. In contrast, *min/max* reduces the cross-entropy loss consistently, irrespective of the different learning rates. We repeat the same experiment across GPT2 and OPT of different sizes as in Table 2, observing similar patterns.

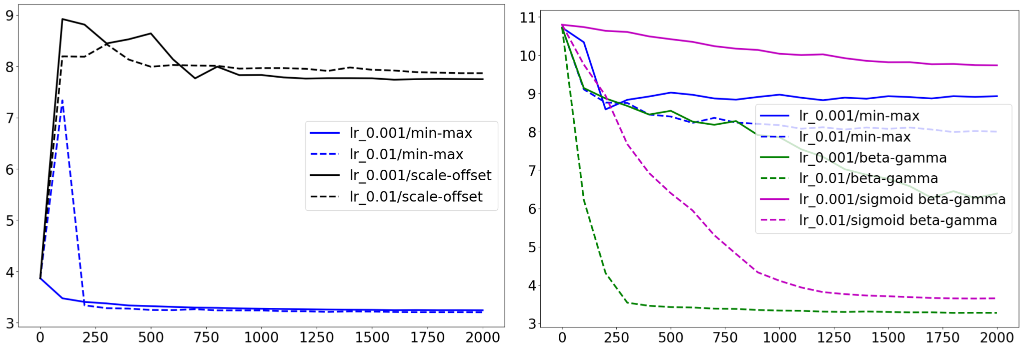

Figure 4: Cross-entropy loss of GPT2-small QAT (y-axis) over 2k training steps (x-axis). Left depicts QAT based on *min/max* and *scale/offset*. Right depicts QAT based on *min/max* and *beta/gamma* (with and without sigmoid).

Given the apparent flaws of *scale/offset*, one might find it puzzling how it has become the de-facto standard for QAT parameterization. Firstly, many QAT studies employ a symmetric quantization scheme (Esser et al., 2020; Choi et al., 2018; He et al., 2023; Ding et al., 2023), which is free from the instability of asymmetric *scale/offset*. Secondly, in LLM quantization, it is often the case that only weights are quantized (Frantar et al., 2022; 2023; Shao et al., 2023; Ding et al., 2023). For weight quantization, granularity is usually per-channel (as opposed to per-tensor activation quantization), and distributions tend to be symmetric with much regularized ranges compared to those of activations. Under such conditions, we find that QAT converges well regardless of parameterizations (see Figure 9 in the Appendix).

min/max* vs. *beta/gamma. Given its greater robustness to different bit widths/learning rates and its independent control over each of the quantization encodings, is *min/max* the preferred parameterization? However, one caveat with *min/max* is its slow convergence when quantization ranges must traverse large distances to reach their minima. This limitation has critical implications in practice, as studies have observed that some activations of LLM contain extremely large values (Xiao et al., 2023; Liu et al., 2023a).

beta/gamma effectively overcomes this difficulty. The idea is simple. Instead of learning θ_{min} and θ_{max} themselves, new parameters β and γ are introduced to scale θ_{min} and θ_{max}:

$$s = \frac{\gamma\theta_{max} - \beta\theta_{min}}{k} \text{ or } \frac{\sigma(\gamma)\theta_{max} - \sigma(\beta)\theta_{min}}{k}, \quad z = \frac{\beta\theta_{min}}{s} \text{ or } \frac{\sigma(\beta)\theta_{min}}{s}. \tag{3}$$

In Figure 5, we quantize a normal distribution with a standard deviation of 50 using both *min/max* and *beta/gamma*. It is evident that *beta/gamma* converges quickly, in stark contrast to *min/max*. This is because *beta/gamma* utilizes $|\theta_{min}|$ and $|\theta_{max}|$ (i.e. to scale the gradients of β and γ, as shown in Table 1. In other words, it scales the gradients of the quantization ranges proportionally to the expected distances they need to travel (i.e. by $|\min(x_t)|$ and $|\max(x_t)|$).

As an astute reader may have noticed, *beta/gamma* is highly similar to *min/max* whose learning rates are scaled by $|\theta^0_{min}|$ and $|\theta^0_{max}|$ (denoted as *min/max+* hereafter). Their similarity is experimentally verified in Figure 5; notice how the blue dashed line (*min/max+*) overlaps perfectly with the green solid line (*beta/gamma* without sigmoid). They are, however, not equivalent in all cases. (1) *beta/gamma* can dynamically set θ_{min} and θ_{max} to the true minimum/maximum values of x. Such dynamism cannot be readily attained in *min/max+*. (2) *beta/gamma* enables per-channel scaling of gradients by having θ_{min} and θ_{max} in vector forms. On the other hand, *min/max+* requires

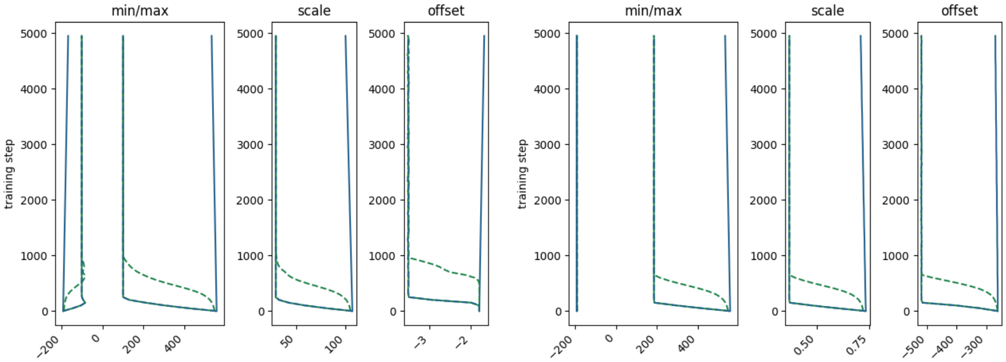

Figure 5: Learnable ranges of *min/max* and *beta/gamma* changing over the course of QAT. *beta/gamma* is color-coded in green (*min/max* in blue). *min/max*+ and sigmoid-applied *beta/gamma* are depicted with dashed lines. The other details of the experiment are identical to those in Figure 3 except that we have omitted the case of $lr = 1e - 2$ for visual clarity.

those values to be passed as scalars to the optimizer outside the model. (3) Finally, Shao et al. (2023) apply a sigmoid function to β and γ as in equation 3. Such additional treatment on the quantization encodings further differentiates *beta/gamma* from *min/max+*.

Let us examine these three differences. The per-channel granularity from having β, γ, θ_{min}, and θ_{max} as model states is a clear advantage. The benefits of dynamic θ_{min} and θ_{max} are also evident, following the same logic as in dynamic versus static quantization. The sigmoid function on β and γ is, however, a double-edged sword. It stabilizes the training process, but at the cost of constraining the quantization range not to expand beyond its initial value and of slowing down the training process by compressing β and γ. We test the impact of the sigmoid function in *beta/gamma* with the controlled toy example (Figure 5) and the LLM QAT (the right subfigure of Figure 4). In both cases, the sigmoid-free approach converges more quickly, and in the LLM experiment, it finds a lower minimum.

	FP	*scale/offset*		*min/max*		*beta/gamma*	
		1e-2	1e-3	1e-2	1e-3	1e-2	1e-3
GPT2-small	30.0	2349.1	50256.8	28.6	28.5	**25.6**	27.0
GPT2-XL	18.4	1712.5	266.6	17.5	16.6	**15.5**	15.8
OPT-125M	31.8	4825.7	746.1	2036.1	52.5	**30.6**	31.6
OPT-1.3B	16.8	3192.3	16.9	17.1	16.6	16.5	**14.9**

Table 2: Perplexity results of LLM QAT with learned asymmetric ranges, organized by model, learning rate, and parameterization. The context length is 1024, with the exception of GPT2-XL, for which a context length of 768 is used. *beta/gamma* is sigmoid-free.

4 Conclusion

Range-learning QAT is inherently unstable as it governs the rounding up/down of numerous elements by modifying a pair of quantization encodings. Adding to the complexity is our limited understanding of the impact of various parameterizations. In our efforts to stabilize and accelerate this challenging QAT process, we have made the following contributions: (1) We experimentally demonstrated that different asymmetric quantization parametrizations can behave differently during QAT. (2) We conducted a comparative analysis between *scale/offset* and *min/max*, demonstrating the favorable properties of the latter in terms of bit-width/learning-rate sensitivity and independent control of two quantization encodings. (3) We conducted a comparative analysis between *min/max* and *beta/gamma*, proposing their respective best QAT practices: *min/max* with adjusted learning rates and sigmoid-free *beta/gamma*.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0a62e133894cdce435bcb4a5df1db2d-Paper.pdf.

Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013. URL http://arxiv.org/abs/1308.3432.

Yash Bhalgat, Jinwon Lee, Markus Nagel, Tijmen Blankevoort, and Nojun Kwak. Lsq+: Improving low-bit quantization through learnable offsets and better initialization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: parameterized clipping activation for quantized neural networks. *CoRR*, abs/1805.06085, 2018. URL http://arxiv.org/abs/1805.06085.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919, 2023. doi: 10.48550/ARXIV.2311.07919. URL https://doi.org/10.48550/arXiv.2311.07919.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters. In Andreas Krause, Emma

Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7480–7512. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/dehghani23a.html.

Xin Ding, Xiaoyu Liu, Yun Zhang, Zhijun Tu, Wei Li, Jie Hu, Hanting Chen, Yehui Tang, Zhiwei Xiong, Baoqun Yin, and Yunhe Wang. CBQ: cross-block quantization for large language models. *CoRR*, abs/2312.07950, 2023. doi: 10.48550/ARXIV.2312.07950. URL https://doi.org/10.48550/arXiv.2312.07950.

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S. Modha. Learned step size quantization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkgO66VKDS.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: accurate post-training quantization for generative pre-trained transformers. *CoRR*, abs/2210.17323, 2022. doi: 10.48550/ARXIV.2210.17323. URL https://doi.org/10.48550/arXiv.2210.17323.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=tcbBPnfwxS.

Jiong Gong, Haihao Shen, Guoming Zhang, Xiaoli Liu, Shane Li, Ge Jin, Niharika Maheshwari, Evarist Fomenko, and Eden Segal. Highly efficient 8-bit low precision inference of convolutional neural networks with intelcaffe. In Luis Ceze, Natalie D. Enright Jerger, Babak Falsafi, Grigori Fursin, Anton Lokhmotov, Thierry Moreau, Adrian Sampson, and Phillip Stanley-Marbell (eds.), *Proceedings of the 1st on Reproducible Quality-Efficient Systems Tournament on Co-designing Pareto-efficient Deep Learning, ReQuEST@ASPLOS 2018, Williamsburg, VA, USA, March 24, 2018*, pp. 2. ACM, 2018. doi: 10.1145/3229762.3229763. URL https://doi.org/10.1145/3229762.3229763.

Yefei He, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Efficientdm: Efficient quantization-aware fine-tuning of low-bit diffusion models. *CoRR*, abs/2310.03270, 2023. doi: 10.48550/ARXIV.2310.03270. URL https://doi.org/10.48550/arXiv.2310.03270.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *CoRR*, abs/2203.15556, 2022. doi: 10.48550/ARXIV.2203.15556. URL https://doi.org/10.48550/arXiv.2203.15556.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *CoRR*, abs/2305.14152, 2023a. doi: 10.48550/ARXIV.2305.14152. URL https://doi.org/10.48550/arXiv.2305.14152.

Minsoo Kim, Sihwa Lee, Janghwan Lee, Sukjin Hong, Du-Seong Chang, Wonyong Sung, and Jungwook Choi. Token-scaled logit distillation for ternary weight generative language models. *CoRR*, abs/2308.06744, 2023b. doi: 10.48550/ARXIV.2308.06744. URL https://doi.org/10.48550/arXiv.2308.06744.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *CoRR*, abs/1806.08342, 2018. URL http://arxiv.org/abs/1806.08342.

Andrey Kuzmin, Markus Nagel, Mart van Baalen, Arash Behboodi, and Tijmen Blankevoort. Pruning vs quantization: Which is better? *CoRR*, abs/2307.02973, 2023. doi: 10.48550/ARXIV.2307.02973. URL https://doi.org/10.48550/arXiv.2307.02973.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. AWQ: activation-aware weight quantization for LLM compression and acceleration. *CoRR*, abs/2306.00978, 2023. doi: 10.48550/ARXIV.2306.00978. URL https://doi.org/10.48550/arXiv.2306.00978.

Jing Liu, Ruihao Gong, Xiuying Wei, Zhiwei Dong, Jianfei Cai, and Bohan Zhuang. QLLM: accurate and efficient low-bitwidth quantization for large language models. *CoRR*, abs/2310.08041, 2023a. doi: 10.48550/ARXIV.2310.08041. URL https://doi.org/10.48550/arXiv.2310.08041.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: data-free quantization aware training for large language models. *CoRR*, abs/2305.17888, 2023b. doi: 10.48550/ARXIV.2305.17888. URL https://doi.org/10.48550/arXiv.2305.17888.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of Machine Learning Research*, 24:253:1–253:15, 2023. URL http://jmlr.org/papers/v24/23-0069.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *CoRR*, abs/2106.08295, 2021. URL https://arxiv.org/abs/2106.08295.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *CoRR*, abs/2308.13137, 2023. doi: 10.48550/ARXIV.2308.13137. URL https://doi.org/10.48550/arXiv.2308.13137.

Sangeetha Siddegowda, Marios Fournarakis, Markus Nagel, Tijmen Blankevoort, Chirag Patel, and Abhijit Khobare. Neural network quantization with AI model efficiency toolkit (AIMET). *CoRR*, abs/2201.08442, 2022. URL https://arxiv.org/abs/2201.08442.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Xiaoxia Wu, Cheng Li, Reza Yazdani Aminabadi, Zhewei Yao, and Yuxiong He. Understanding int4 quantization for language models: Latency speedup, composability, and failure cases. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37524–37539. PMLR, 2023. URL https://proceedings.mlr.press/v202/wu23k.html.

Guangxuan Xiao, Ji Lin, Mickaël Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/ARXIV.2205.01068. URL `https://doi.org/10.48550/arXiv.2205.01068`.

A Appendix

A.1 Appropriate learning rates for *scale/offset*

Given that s and z exist in different spaces, it becomes necessary for QAT to adjust their gradients accordingly. A straightforward approach involves scaling them based on the absolute values of their corresponding parameters (denoted as *naive* hereafter). For a more sophisticated method, we trace the implicit θ_{min} and θ_{max} to determine the amount of updates for s and z. Let us first investigate the learning rate for s. The one-step updates of θ_{min} and θ_{max} with Adam optimizer are as follows:

$$\begin{aligned}
s^{(t+1)} &= s^{(t)} - \eta \frac{\mathbb{E}[u_s]}{\sqrt{\mathbb{E}[u_s^2]}} \\
u_s^{adam} &= \frac{\mathbb{E}[u_s]}{\sqrt{\mathbb{E}[u_s^2]}} \\
s^{(t+1)} &= s^{(t)} - \eta u_s^{adam} \\
\theta_{min}^{(t+1)} &= \theta_{min}^{(t)} - \eta \frac{\mathbb{E}[-\frac{1}{p}u_s]}{\sqrt{\mathbb{E}[(-\frac{1}{p}u_s)^2]}} \\
\theta_{min}^{(t+1)} &= \theta_{min}^{(t)} + \eta u_s^{adam} \\
\theta_{max}^{(t+1)} &= \theta_{max}^{(t)} - \eta u_s^{adam} \quad \text{(by the same logic).}
\end{aligned} \tag{4}$$

The update of s under *min/max* can then be expressed as:

$$\begin{aligned}
s'^{(t+1)} &= \frac{(\theta_{max}^{(t)} - \eta u_s^{adam}) - (\theta_{min}^{(t)} + \eta u_s^{adam})}{k} \\
&= \frac{(\theta_{max}^{(t)} - \theta_{min}^{(t)})}{k} - \eta \frac{2}{k} u_s^{adam} \\
&= s'^{(t)} - \eta \frac{2}{k} u_s^{adam}.
\end{aligned} \tag{5}$$

We can similarly derive a scaling factor in the case of Stochastic Gradient Descent (SGD) optimizer:

$$\begin{aligned}
s'^{(t+1)} &= \frac{(\theta_{max}^{(t)} - \eta \frac{1}{k} u_s^{sgd}) - (\theta_{min}^{(t)} + \eta \frac{1}{k} u_s^{sgd})}{k} \\
&= \frac{(\theta_{max}^{(t)} - \theta_{min}^{(t)})}{k} - \eta \frac{2}{k^2} u_s^{sgd} \\
&= s'^{(t)} - \eta \frac{2}{k^2} u_s^{sgd}
\end{aligned} \tag{6}$$

To summarize, in the case of scale, we can scale the update of s under *scale/offset* by $\frac{2}{k}$ to emulate the update of the derived s under *min/max* for Adam optimizer. The matter is, however, not straightforward for offset since the derivative of offset with respect to θ_{min} (and θ_{max}) is once again complicatedly dependent on θ_{min} and θ_{max}:

$$z'^{(t+1)} = k \frac{\theta_{min} - u_{min}^{adam}}{(\theta_{max} - u_{max}^{adam}) - (\theta_{min} - u_{min}^{adam})} \tag{7}$$

One practical alternative is to use the relationship between scale and offset as defined in equation 1, based on which one can scale gradient to offset as follows:

$$\frac{dL}{dz} = \frac{dL}{d\theta_{min}} \frac{1}{s} \tag{8}$$

The proposed scaling hold regardless of optimizers, given the premise that the relationship between z and θ_{min} in equation 1 should be maintained throughout QAT. We denote this particular scaling of learning rates for s and z as *sophisticated* hereafter.

Besides *naive* and *sophisticated*, we can additionally devise a new parameterization that takes out the bit-width component k out of the learnable parameters:

$$s' = \theta_{max} - \theta_{min}, \quad z' = \frac{\theta_{min}}{\theta_{max} - \theta_{min}}. \tag{9}$$

The scale and offset variables can then be trivially retrieved from s' and z' such that $s = \frac{1}{k}s'$ and $z = kz'$. With the k taken out, s' and z' are now located in the same space, making any learning rate adjustment unnecessary (denoted as *kscale/koffset* hereafter).

We perform the same experiment of quantizing a normal distribution as in Figure 6 with the three aforementioned methods: *naive*, *sophisticated*, and *kscale/koffset*. The results are illustrated in Figure 6. While all the alternatives to *scale/offset* show better performance than the vanilla *scale/offset* in 10-bit quantization, none shows the stability of *min/max* and *beta/gamma*.

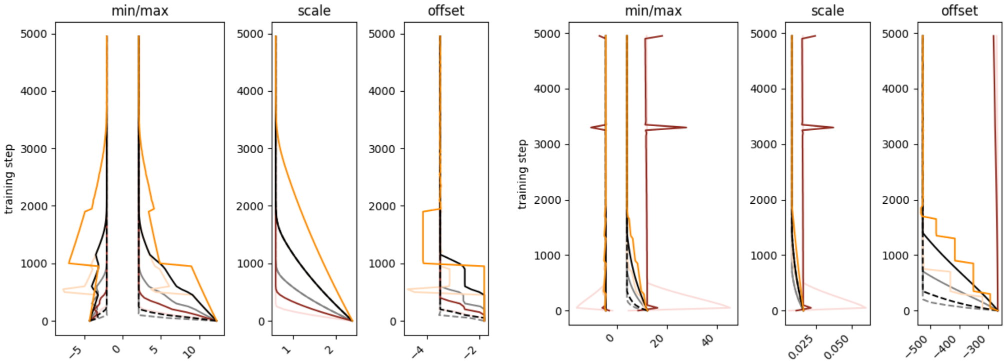

Figure 6: Learnable ranges of (1) *scale/offset* (red), (2) *kscale/koffset* (orange), (3) *naive* (black), and (4) *sophisticated* (gray) changing over the course of QAT. The other details of the experiment are identical to those in Figure 3

A.2 Fundamental difference between *scale/offset* and *min/max*

From equation 5 and equation 7, we observe that the relationship between $s'^{(t)}$ and $s'^{(t+1)}$ is not the same as the relationship between $z'^{(t)}$ and $z'^{(t+1)}$. In other words, even if we make the update of the derived scale under *min/max* and the update of the scale under *scale/offset* identical via linear scaling, the updates of the offset will be different. After an indefinite number of updates, given $x = 0$, quantization/dequantization can thus result in different answers for *scale/offset* and *min/max* due to their discrepancy in z. This is one example that evinces *scale/offset* and *min/max* are not one and the same. There exists no straightforward linear transformation that ensures both scale and offset undergo identical updates under *min/max* and *scale/offset* across their entire domain.

A.3 Symmetric quantization

The symmetric correspondence of equation 1 is defined as follows:

$$\begin{aligned} &\bar{x} = Q(x) = \text{clip}(\lfloor \frac{x}{s} \rceil, -\frac{k-1}{2}, \frac{k-1}{2}), \\ &\hat{x} = DQ(\bar{x}) = s(\bar{x}), \\ &\text{where} \quad k = 2^b - 1, \quad \theta_{max} = \max(|x|), \quad s = \frac{2 * \theta_{max}}{k}. \end{aligned} \tag{10}$$

One of s, θ_{max}, and γ can be designated as a learnable parameter, as illustrated in Figure 2. Given the gradients of these parameters appropriately scaled as in Table 3, it is evident that they would behave identically during QAT. See Figure 7 for experimental verification, in which all the lines perfectly overlap.

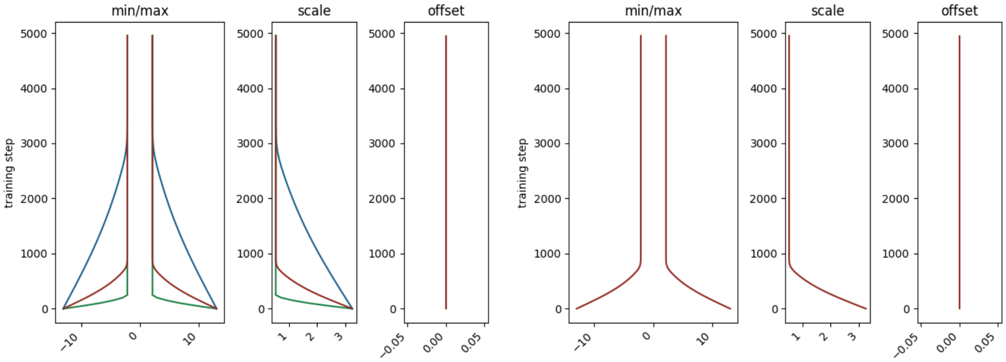

Figure 7: Learnable ranges of *scale/offset* (red), *min/max* (blue), and *beta/gamma* (green), changing over the course of symmetric 3-bit QAT. On the left, all parameterizations receive the same learning rate of 5e-3. On the right, the learning rates are appropriately scaled.

	$n < x < p$	$x < n$	$x > p$
$\frac{d\hat{x}}{ds}$	$\lfloor \frac{x}{s} \rceil - \frac{x}{s}$	n	p
$\frac{d\hat{x}}{d\theta_{max}}$	$\frac{2}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$\frac{2}{k}n$	$\frac{2}{k}p$
$\frac{d\hat{x}}{d\gamma}$	$\theta_{max}\frac{2}{k}(\lfloor \frac{x}{s} \rceil - \frac{x}{s})$	$\theta_{max}\frac{2}{k}n$	$\theta_{max}\frac{2}{k}p$

Table 3: Gradients of symmetric quantization ranges.

A.4 ReLU case

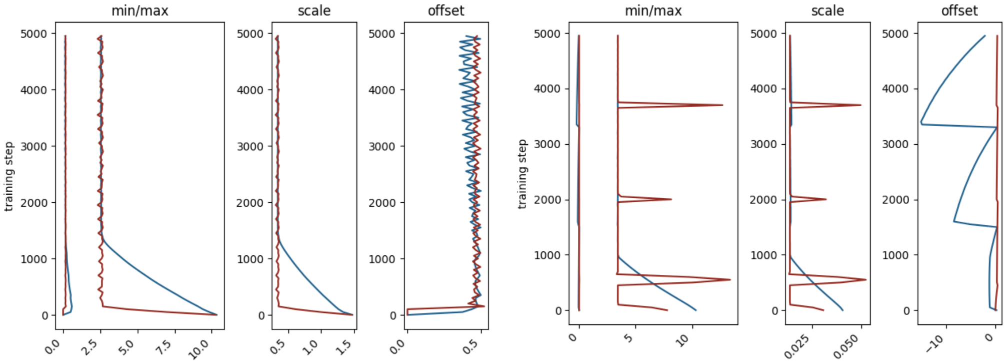

Figure 8: Learnable ranges of *scale/offset* and *min/max* changing over the course of QAT. The details of the experiment are identical to those in Figure 5, except that the right subfigure involves 8-bit quantization instead of 10-bit. This adjustment was made because 10-bit quantization results in values that are too large to be effectively visualized.

As discussed in the main body of this work, *scale/offset* is particularly unstable when one of θ_{min} and θ_{max} has already converged to its optimum and the other is still moving. This is typical of an activation after ReLU where θ_{min} is likely to be placed on the near-optimal position 0.0 from the beginning. We perform QAT on a ReLU-applied normal distribution in Figure 8, in which we observe severe instabilities for *scale/offset*.

A.5 Normal Quantization

As discussed in the main body of the paper, it might seem puzzling that there are numerous successful *scale/offset* cases for QAT with learned asymmetric ranges, despite the apparent risks. To

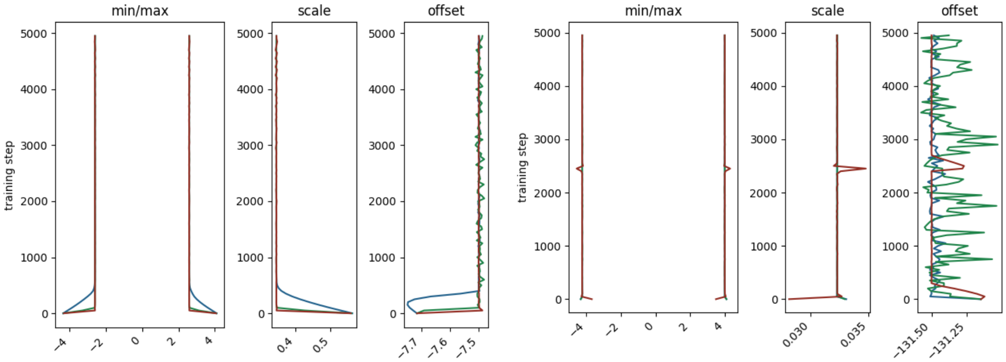

Figure 9: Learnable ranges of *scale/offset* (red) and *min/max* (blue) changing over the course of QAT. The details of the experiment are identical to those in Figure 3, except for the initial starting points: $\theta^0_{min} = \min(x)$ and $\theta^0_{max} = \max(x)$.

investigate whether *scale/offset* can still converge successfully under less extreme conditions, we conduct an experiment, as depicted in Figure 9, keeping the experimental setup consistent with that in Figure 3. However, we quantize the tensor to 4 bits and 8 bits (rather than 3 bits and 10 bits) and set θ_{min} and θ_{max} to $\min(x)$ and $\max(x)$ (instead of $\min(x)$ and $3 * \max(x)$), to alleviate the difficulty of the task. The results indicate that θ_{min} and θ_{max} of all parameterizations converge to the identical positions.