We thank all the reviewers for their helpful feedback, and for being unanimously positive about the submission: **R1:** *"The authors provided strong reasoning behind why a uniform shape is beneficial"*; **R2:** *"The paper is easy to follow"*; **R3:** *"Authors did enough experiments on different data sets and different neural networks"*; **R4:** *"The authors have presented an excellent analysis where they theoretically show how uniformly distributed data suffers less quantization noise"*. Below we address the main suggestions for improvements. Following Reviewer 1 suggestions, we ran two additional experiments that will be included in the final version along with accompanying code. If we address the reviewers' comments, we kindly ask that they adjust their scores to reflect their favorable opinion.

R1: *"There is little explanation about the impact of Kurtois to the activation quantization."* — Activations are typically less sensitive to quantization for vision workloads. Still, for the Neural Collaborative Filtering (NCF) model, we did some tests for activations with favorable results in all configurations: 63.07% vs. 62.09% (4 bits); 62.35% vs. 59.03% (3 bits); 56.06% vs. 36.46% (2 bits).

"...a solution that can easily modify the step size to become a power of two would be very desirable."— We conducted some new tests on ImageNet. When rounded to nearest power-of-two and in the case of 4-bit quantization, our method improves from 61.4% to 66.2%, and from 63.6% to 74.2% for ResNet18 and ResNet50, respectively. This becomes even more pronounced for 3-bit quantization going from 37.5% to 55.8%, and from 53.2% to 71.6% for ResNet18 and ResNet50, respectively.

"Is there any particular reason of choosing Kurtosis over other statistical measure, such as coefficient of variation?" — Kurtosis is a differentiable measure we can optimize to re-shape tensors into a uniform-like distribution. Entropy maximization is another option for data uniformization, which didn't work better than the Kurtisis measure but was more complicated to use in practice.

R2: *"In table 1, it can be observed that from 4-bit quantization to 3-bit quantization, the performance drops a lot. Can the authors provide any explanation about this?"* — Indeed, post-training quantization (PTQ) methods obtain mild degradation up to 4-bit quantization, which increases rapidly below that point. Our work pushes this boundary further offering a better trade-off for PTQ-based methods.

R3 *"No experimental parameter settings are provided, and no comprehensive comparison with the latest SOTA method is provided in the paper."* — A detailed description of all experimental parameter settings is provided in the appendix (titled "hyperparameters to reproduce results") as well as a documented code. We compare against the results recently reported by Alizadeh et al. [2020] and demonstrate the improved robustness on two additional SOTA methods ([Nahshan et al., 2019] & [Esser et al., 2019]).

"I don't get the claim of the title of this paper "One model to rule them all"" — We store a single set of weights ("one model") that can be applied with a large number of data-types ("to rule them all"). *"Almost all quantization approach can be applied to different applications as long as enough data were provided given the context of DNN quantization"* — Correct, but current methods do not show robustness when quantized to bit-widths other than the one they were trained for. In contrast, we allow for a single model to operate at various quantization levels (e.g., employ a 4-bit variant of the model when the battery is below 20% but the full precision one when the battery is over 80%.).

"Second, the comparasion between KURE and the baseline model could be biased in Table 1. Since applying regularization, e.g. L2 regularization, is standard procedure in training DNN, it is unfair to compare with baseline approaches with no regularization in the experiments." — "no regularization" means "no kurtosis regularization", but L2-regularization still applies. We will clarify that in the final version of the paper.

Additional comments: *"1. Line 114 proposes that the uniform distribution is more robust to the quantization process than the Gaussian distribution. However, formula (6) is derived based on ... a uniform distribution. "* — Equation 6 is only used to show that an optimal quantizer has a sensitivity of $\frac{\varepsilon^2}{4}$ for uniform inputs. Lines 124-132 prove that this sensitivity is larger than $\frac{\varepsilon^2}{4}$ for *any* quantizer with Gaussian inputs; *"2. Only the picture b in Figure 1 is mentioned"* — Thanks, we will include a reference to Fig. 1(a). *"3. The meaning of W_i in line 162 is not stated."* — W_i denotes the weight tensor of the i-th layer. *"4. What is Kurt's formula in Eq 15?"* — The Kurtosis formula is defined in Equation 13. *"5. What is the difference between L1 Regularization and L1 Regulation (λ = 0.05) in Table 2"* — In one configuration λ is found through a grid-search, and in the other it is set to $\lambda = 0.05$.

R4: *"Does the kurtosis regularization impact the attainable loss in the original objective."* — FP accuracies before and after applying Kurtosis regularization are almost identical (See Table 1).

"A lot of works are now looking at other number formats... I do believe KURE can still be applied, but with an identification of another target kurtosis."— Thanks, we are currently working on other kurtosis targets to shape tensor distributions and make them more suitable for FP and binary quantizations.