We thank the reviewers for their valuable feedback. We address your concerns as outlined below.

Reviewer #1

- *Claims about SOTA performance*. We toned down the discussion of our experimental results. We now claim "competitive performance", and rely on the simplicity of SQR/OC to establish them as "new and strong baselines".
- *Performance of SQR*. The confidence intervals of SQR are slightly better calibrated than ConditionalGaussian, while being narrower. However, we agree that SQR shines when dealing with multi-modal uncertainties. We now illustrate this on real data from the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.
- *Performance of OCs*. We now report the average accuracy of out-of-distribution detection across the four presented datasets for all methods. OCs are the best performing with an average accuracy of 90%. The second best performance is obtained by the entropy estimator, at 87%.
- *Performance of OCs across datasets*. We believe that OOD must be powerful enough as to reject images with the same statistics as the training data, but different classes. We now also report an experiment where we measure OOD performance across datasets. OCs remain the best performing method. For example, OCs are able to use an MNIST network to reject 96% of Fashion-MNIST examples, followed by ODIN at 85%.
- *Sensitivity wrt λ*. We now report an ablation study showing that OCs are stable with respect to the choice of λ. Our ablation shows that any value greater than 10 achieves competitive results. Looking at a hard-constrained version of OCs is an exciting direction for future work.
- *Neural network architectures and dropout cross-validation*. We now specify all the neural network architectures in the Appendix (for SQR, two layers of 64 ReLU units). We now cross-validate the value of dropout for all experiments in $\{0.1, 0.25, 0.5, 0.75\}$, obtaining very similar results to the ones reported in our submission.
- *Loss choice for OCs*. We now emphasize the text (L133-135) explaining that we use the same loss as the one employed to train the entire network. In this case, the cross-entropy loss.
- *Report MSE in SQR experiments.* The only method trained to optimize MSE is ConditionalGaussian. SQR (as well as the other quantile-based methods) optimizes MAE (at the median). If minimizing MSE is important for the task, we propose having an extra output in the neural network dedicated to that issue.

Reviewer #2

- *Benefits and performance of the proposed methods*. Please see the first four items in the response to Reviewer #1.
- *Sample efficiency*. We now leverage results from the literature on quantile regression and principal component analysis to discuss the (optimal) sample efficiency of SQR and OC.
- *Comparison to sampling-free methods*. We now discuss (Hwang et al., 2018; Le et al., 2018; Feng et al., 2019).
- *Test SQR on ImageNET*. We now refrain from talking about classification in the SQR section, and focus the application of SQR to regression problems. Therefore, SQR is not applicable to ImageNET. However, please note that SQR can be implemented as an extra output neuron, therefore not impacting the original accuracy of the network.
- *Linearity assumption on OCs*. We now emphasize the text (L130-132) describing how to construct non-linear OCs. In any case, restricting our experiments to linear OCs is justified by the manifold hypothesis (Bengio et al., 2013), which states that deep learning models attempt to arrange the data manifold to linearly separate the examples.
- *Choice of l_c*. We now emphasize the text (L134-135) discussing the choice of ℓ, and added a small experiment in the Appendix that compares the cross-entropy (superior) to the MSE loss.

Reviewer #4

- *Theorem 1 and causality experiment*. We have now moved Theorem 1 to the Appendix, and the causality experiment to the main text.
- *SQR and classification*. We now refrain from applying SQR to classification, and focus on regression problems.
- *Behaviour of $\|C^\top x\|$*. We thank the reviewer for pointing this out this subtle mistake. We have removed the explanatory paragraph (L165-167). Now we rely on Theorem 1 to describe (in a more precise manner) the behaviour of $\|C^\top x'\|$ as x' moves far away from the nullspace of the training data.
- *Missing references*. We now discuss (Geifman and El-Yaniv; Chen et al) in our related work section.