Dear reviewers, thank you for your time to thoroughly read and review our paper. For your convenience, we briefly summarize our contributions again below before addressing some of your specific questions and criticisms.

You said "the addressed problem is relevant and timely" (R4), "useful in practice with many applications" (R1), with the paper being "focused and well-written" (R3), having "contributions which are relevant (for a part of) the NeurIPS community" (R2) and "interesting for many communities" (R1). You further acknowledged the following contributions:

- Demonstrating that a uniform noise (UN) channel can be implemented at test time (R1, R2, R3)
- Eliminating the train-test mismatch while maintaining a differentiable loss function (R1, R2, R3)
- Bridging compression with and without quantization by showing that quantization is a limiting case of a soft quantizer applied to the uniform noise channel (R2, R3)
- Reducing the variance of gradients by analytically integrating out the noise (R1, R2)

We want to highlight these additional contributions which you may have missed:

- *Proving that the general problem of communicating a sample from an arbitrary distribution is computationally hard* (Lemma 1, l.105). This has practical implications and makes the special case of uniform noise particularly interesting. Beyond compression, our lemma shows that sampling from a distribution is generally hard even with access to samples from a related distribution.
- Quantifying for the first time the gap between train and test losses in the approach of Ballé et al. (see below)

The proposed approach only marginally improves PSNR for hyperprior models. (R1, R2, R3) While the benefits of our approach in terms of PSNR are modest for the complex hyperprior model, we observe significant improvements for the linear model (Figure 2A). Note that in compression *lightweight models are very relevant in practice*.

Secondly, *our empirical results allow us for the first time to quantify the gap between training and test losses*. Until now, nobody knew to which extent a train-test mismatch hurts (or helps) performance. That is, our results are interesting from a practical point of view even without any immediate improvements in performance.

Finally, *universal quantization has the potential to lead to much bigger improvements in the future*. While MSE and PSNR are still the most widely used metrics, they are also flawed. Adversarial losses and advanced perceptual metrics will be more sensitive to the perceptual differences between quantization and noise. However, as these metrics are still an active area of research and can be difficult to train and evaluate, we believe it is right to focus on PSNR first.

Why did you only evaluate on the Kodak dataset? (R1, R2, R4) Unlike other tasks, compression results tend to generalize well across natural image datasets (e.g., Agustsson et al., 2017). This explains why evaluations on a single dataset, namely Kodak, are common practice (e.g., Theis et al., 2016; Ballé et al., 2018; Choi et al., 2019). We therefore decided to use the limited space to explore our conceptual contributions instead of providing results on other datasets.

Can we fix the train-test mismatch by applying soft-rounding to hard quantization during both training and testing? (R1) Combining hard quantization with soft-rounding would be equivalent to hard quantization, $s(\lfloor s(y) \rceil) = \lfloor y \rceil$, and thus would not be differentiable. Using soft-rounding without noise during training (Agustsson et al., 2017) would not create a bottleneck (since soft-rounding s_α is invertible for any finite α). Agustsson et al. (2017) point out that "choosing the annealing schedule is crucial" (p.6) as annealing needs to be fast enough to avoid inversion, which then causes large gradients. We do not suffer from these problems due to the noise and our proposed variance reduction.

What is Figure 1B trying to say? (R1) Figure 1B shows an example where the derivatives of $h(y)$ can vary wildly for small changes in y. Using $h'(y + u)$ with sampled noise u during training thus leads to gradients with high variance. On other hand, the expected derivative $\mathbb{E}_U[h'(y + U)]$ varies smoothly and thus leads to gradients of lower variance.

Why does taking the expectation reduce variance? (R1) Taking the expectation to reduce variance is a commonly used trick which can be motivated by the *law of total variance*, $\mathrm{Var}[\mathbb{E}[\Delta \mid U]] = \mathrm{Var}[\Delta] - \mathbb{E}[\mathrm{Var}[\Delta \mid U]] \leq \mathrm{Var}[\Delta]$.

What do you mean with the "single coefficient which is always zero"? (R2) The encoder outputs $\mathbf{y} = f(\mathbf{x})$. If $y_i = 0$ and the same noise is used everywhere (Choi et al., 2019), $\mathbf{z} = \mathbf{y} + u$, then the decoder could recover $\mathbf{y} = \mathbf{z} - z_i$.

What is the added training complexity of the proposed method? (R4) For the hyperprior model (UN + UQ + SR) we observe an increase in training times of 2-4% with variance reduction and less than 1% without variance reduction.

Is the source code provided? (R4) Source code will be made available.

Related work on soft-rounding. (R2) We'll add references to the mentioned work on soft-rounding.

Do you mean the prior should be flat in the interval at l. 168-169? (R2) That is correct, we will clarify the text.

Where exactly is the universal quantization applied with soft-rounding? (R2) We apply it to the result of soft-rounding, $s(y)$, i.e. $s(y) + u$ during training and universal quantization $\lfloor s(y) - u \rceil + u$ at test-time. See also l.254.