We thank the reviewers for their feedback and their time. (References here point to the submitted file's bibliography.)

Novelty of the work. We believe that R2 and R3 omitted key contributions that previous work has been unable to achieve: (i) our local reparameterization allows us to scale correlated Gaussian posteriors beyond what was thought to be possible previously, (ii) we show that a variational posterior 'LR + isotropic diag' outperforms existing VI methods for BNNs in most experiments, (iii) we show that correlated Gaussian posteriors offer a systematic solution to the underfitting of MF-VI, but this is limited only to small networks due to computational constraints, (iv) we identify problems with algorithms derived in previously published works [29, 32, 39, 41]. As R1 says, we believe these contributions to be significant for the community, and hope they will catalyze research on reparametrizing BNNs.

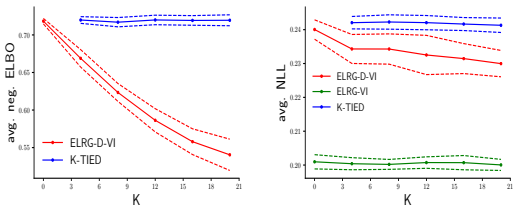
R2: Predictive performance benefit of ELRG-VI vs original low-rank parametrization. The original, naive, low-rank parametrization [41, 32, 39] is computationally expensive. This then requires sharing variational samples among inputs, leading to worse predictive performance (see Fig 4 (left)). Other benefits of separate variational samples are extensively studied in the literature (see references in Section 2) this was the primary motivation for work [19, 43].

R2: "MAP can give more accurate predictions". Although this is true, our results show that MAP has bad uncertainty calibration (as R4 says). We show problems with MAP estimation with calibration curves (Figure 3) and an out-of-distribution experiment. Losing slightly sometimes in accuracy but gaining a much better predictive distribution is often key for strong performance on real world tasks [4, 17]. Specifically, [11] shows overfitting to data yields better accuracy.

R2 and R4: Why does ELRG-VI have a better predictive performance than ELRG-D-VI, with a more expressive non-isotropic diagonal? We believe this confusing result arises due to a poor choice of prior over weights (see "Bayesian Deep Learning and a Probabilistic Perspective of Generalization", section 9.1), which becomes more apparent for large NNs. The general unintuitive problem that more expressive posteriors can yield worse predictions is an open research problem. For MF-VI, many units' variances converge to values close to prior variance, as shown in Figure 1 (densities), which is causing underfitting (see Appendix A7 for reasons). Adding low-rank (LR) terms mitigates this, but only for small networks. For larger networks, as K is small compared to matrix size, the described mechanism is negligible, and we need to resort to using a simpler posterior where units cannot converge to the prior. R2 also asked why LR + isotropic diagonal is a better choice for specifically CNNs. As per line 130, for CNNs, LRT cannot be applied [43], but the LR component can still be reparametrized. In practice, many authors still use LRT for CNNs, claiming this is an approximation. Using LR + isotropic diagonal with a tiny diagonal scale reduces the effect of this approximation as the problematic diagonal term is very small (see Appendix A.3). We will improve this writing in the paper.

R3 (and R1): Comparison to Swiatkowski et al., 2020, "k-tied". We provide a comparison on a toy experiment here (see Figure) and will provide a more extensive comparison for the camera-ready version. However, we feel one should be careful about excessively focussing on the conclusions from such a comparison.K-tied approach lowers the complexity of the posterior, whereas our paper adds low-rank terms to increase the expressiveness. In theory our approach should yield better predictions than k-tied if ELBO is correlated with predictive performance. If this is not happening, it is likely we face issues with problem formulation, possibly due to a poor prior distribution (see discussion above for details). In terms of ELBO values, ELRG-D-VI $\geq$ MF-VI $\geq$ k-tied, assuming we find the global maximizer. As R1 says, it is hard to compare with ELRG-VI's ELBO value. We also note that k-tied + our approach will yield higher ELBO than k-tied alone. Here, we test on the same setup for Figure 1 in the main paper. We find that k-tied results in higher ELBO than ELRG-VI (ELRG-VI value worse by a factor of 10 than other algorithms, hence not plotted), but ELRG-VI achieves the best predictive performance. For small NNs, we observe that k-tied performance matches standard MF-VI and does not change with K. Also, we compare (Figure 2b) to another work [42] by increasing the complexity of WN posterior discussed there.

R3: Questions about the out-of-distribution (OOD) experiment. R3 has misunderstood the context of this experiment. We specifically refer to confidence for **test** and **train** data (not OOD data) corresponding to the dataset used to train the model. Lack of confidence for test/train data means poor predictive performance (such as with MF-VI). As R3 notes, we then expect high uncertainty for OOD data. MF-VI gives high uncertainty both on test data and OOD data but offers weak predictions due to underfitting. We will clarify these points in the text. The OOD experiment closely follows the related paper [33] and matches their conclusions.



R4: ELRG-VI vs MF-VI predictive uncertainty. From our experiments, it is visible MF-VI yields poor predictions and predictive uncertainty compared to ELRG-VI (Table 3 and Fig 4). We will improve the discussion of this point.

R4: Shape and convergence of curves in Figure 2. In Fig 2, the experiments for $\approx$400K optimization steps. This number is a highly excessive number of updates for MNIST, many more than other papers ($\approx$100K more than Bayes by Backprop). Learning curves will change insignificantly if optimization is run for longer. R4 asked about the shape of curves in Figure 2 (a,b). The momentum in the ADAM optimizer causes the bump (please note the curves show *validation* performance, not the optimized quantity). R4 also asked about the equivalence of approaches without LR in Figure 2(a). It turns out that isotropic diagonal has slightly better performance, in line with conclusions in [42].