We thank the reviewers for their feedback. All reviewers found the exposition clear and most agreed the results were novel and of interest to the Bayesian machine learning community. As suggested, we will add an expanded broader impact statement, examining the role of uncertainty quantification in high-risk and scientific applications.

R1. Potential extension to classification: Thms 1-3 apply to the variance of the logits in classification. However, an extension of Thms 1-2 to classifier uncertainty is not straightforward since that also depends on the logit means.
R1. Justification of the chosen prior: HMC and the limiting GP provide good in-between uncertainty (Fig 3) and active learning performance (Sec 5, Table 1). This means the chosen prior encodes reasonable assumptions for the low dimensional regression tasks we consider, and the poor performance of MFVI/MCDO is due to bad inference. In contrast, some recent work examining BNN priors, e.g. Wenzel et al. 2020, considers image classification with Bayesian ResNets, which is a different probabilistic model and a different task.

R2. Clarification on distinction between model and inference: We will move Fig 9 into the text to illustrate Thm 3, and emphasise in the introduction that Criterion 1 *is* satisfied for deep BNNs for MFVI and MCDO posteriors (lines 168-170, 264). We may be using the term "model" differently from R2: we mean "probabilistic model" i.e. a prior and likelihood. We therefore see Thm 3 as a result about inference rather than a model (since it concerns the form of an approximate posterior, rather than the exact posterior) — we will clarify this.
R2. Clarifying limitations of empirical evidence: We will state on lines 51-53 that our experiments focus on the small-data regime and low-dimensional regression, where comparison with exact inference is easier to perform. Although previous authors have obtained good empirical results on downstream tasks (as mentioned on line 23), previous work does not generally focus on how well the approximate predictive resembles the exact one, as we do (lines 296-298).
R2. Potential sub-optimality of hyperparameters in active learning experiments (Table 1): Following the review, we performed some manual hyperparameter tuning for the prior & dropout rate for MCDO Random (validating on the test set). This brought 4HL RMSE from 0.443 ± 0.01 to 0.387 ± 0.02, but this result is still worse than the 1HL case. More extensive search may be able to improve this further, but extensive hyperparameter search is generally impractical in online active learning. Our main goal in Sec 5 was to evaluate the quality of approximate inference compared to the limiting GP (which is closer to exact inference), rather than to improve active learning performance in general. The GP performs significantly better than random selection for all depths, meaning that in the small data regime, the benefit of this Bayesian prior combined with accurate inference is clear. In comparison, the poor results of MFVI & MCDO suggest the worse performance is mainly due to the bad approximate inference in deep BNNs (lines 257-259).

R3. Potential directions on improving uncertainty quantification (UQ): For 1HL BNNs new objective functions with mean-field Gaussian families will not solve issues regarding UQ (Thms 1,2). For deep BNNs Thm 3 tells us that MFVI/MCDO are sufficiently expressive for many tasks that rely on UQ, so improved objective functions may lead to improvements (e.g. Sun et al. 2019 [37], Fig 1).

R4."Lack of novelty": We respectfully disagree. To the best of our knowledge, Thms 1-3 are the first theoretical results on the quality of BNN approximate inference in terms of *estimating function space uncertainty*. The derivations are non-trivial, and the results apply *regardless* of the inference algorithm (not just VI, see lines 73-75, 118 & 130-131). This includes methods which are usually not expected to be over-confident, e.g. EP and Rényi VI, as long as factorised Gaussians/dropout distributions are used as approximate posteriors.
R4. Parameter vs function space: The over-confidence of VI in *parameter space* is well-known. However, it is not obvious how this translates into *function space*. Our results focus on uncertainty in function space (lines 35-36, 144-145). To illustrate this difference we run mean-field VI for a simple Bayesian linear regression (BLR) model with RBF features, defined by $y(x) = \sum_{i=-10}^{10} w_i \psi_i(x)$, $\psi_i(x) = \exp(-(x-i)^2)$, with $\mathcal{N}(0,1)$ priors on w_i. Even though MFVI is over-confident in parameter space (Fig 1 bottom), it still shows significant in-between uncertainty in function space (Fig 1 top). This shows that weight-space intuition does not necessarily translate to function space. The fact that 1HL MFVI & MCDO BNNs, usually thought of as more flexible than BLR, *cannot* represent this kind of uncertainty is non-obvious, and this is pointed out by our contributions.

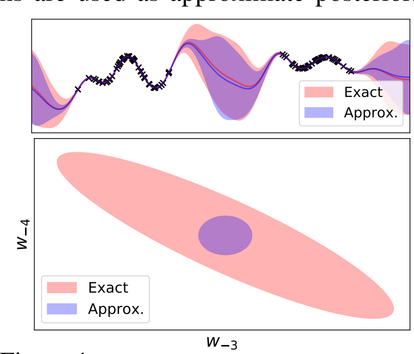

Figure 1: BLR with RBF features & MFVI.

R4. Why we minimise squared error in Fig. 2: Thms 1 & 2 show that no factorised Gaussian or dropout posterior can give in-between uncertainty for 1HL ReLU BNNs, regardless of objective. To verify this, we find the closest approximation within the approximating family to a desired target (with in-between uncertainty) by directly minimising the squared error between the variance functions. The symmetry arises because we chose a symmetric target, not because of the squared error loss. We also optimise the ELBO (Fig 3c,d, Fig 4).
R4. Extension to other non-linearities: Thms 1 & 2 apply to ReLU non-linearities. Empirically, we observed that Tanh BNNs also struggle with in-between uncertainty, but we currently do not have a theoretical proof of this. Thm 3 can likely be extended to other non-linearities and we leave this to future work.