We want to express our gratitude to be part of a community that is so generous in offering such insightful and thorough reviews. We were glad that reviewers thought that the paper was “extremely well written” with analysis that is “great and easy to follow” and a “strong step towards seriously understanding why MF assumptions are fine in practice with deep networks” [**R1**], “timely and interesting” [**R2**], with experiments that are “genuinely interesting to think about, as well as genuinely novel” [**R3**], and “significant and novel” [**R4**].

[**R1**,**R3**,**R4**] **Motivation:** **R1** felt MFVI already shows “strong performance”, excepting optimization problems, and **R4** agreed its use was “common practice”. In contrast, **R3** was unsure anyone “would want to use MFVI”. 1) The disagreement between reviewers shows how timely and relevant our work is. 2) This highlights how difficult it is to pin down ‘what everyone believes’. 3) To substantiate our view of what many published authors believe, our work cites 12 papers in §2 that say something like e.g. “a fully factorized posterior... is very restricting” [Louizos and Welling 2016.]. 4) We agree about optimization issues, and discuss this on L109-110 as well as in S7. 5) We like the reframing **R1** suggested, that our paper justifies common practice in addition to challenging a mistaken belief, and will add this. **But whatever the reviewers personally believe, many publish based on the beliefs we critique, motivating our work.**

[**R2**,**R3**,**R4**] **Figure 2—in-between uncertainty**: This *toy* fig. shows that in-between uncertainty is possible in deeper networks; larger-scale exps. should carry much more weight. But below we show that for 3-layer MFVI the uncertainty is worse but of the same order-of-magnitude as an HMC baseline, while in 1-layer MFVI fails to capture uncertainty.

[**R2**,**R4**] **Special cases:** [**R2**,**R4**] **ReLUs**: Thm. 2 applies to all piecewise linear (PWL) activations (L160). This includes ReLUs and many variants, which are very widely used. Many more theoretical papers focus on PWL, and other nonlinearities can be approximated as PWL. [**R2**] **Gaussians**: in fact, lemmas 1 and 2 and Thm. 2 apply to any distribution with finite first- and second- moments, which we will clarify. Also, fully-factored Gaussians are very commonly used. [**R4**] **MVG**: We give the *general* form of the covariance (L125, D.1 eq. 15) and show equivalence to the commonly used MVG distribution only as an intuitive lower-bound on the expressiveness of the covariance.

[**R2**,**R3**] **“Depth induces correlation is well-known”**: **R3** cites Hinton & van Camp 1993 and **R2** offers no citations. H&vC *hypothesise* MF networks could learn complicated posterior distributions over functions, a two sentence comment in their discussion (which we will cite). We *characterise* the form the correspondence takes and compare to the popular MVG approximation. Moreover, many papers are motivated by the opposite view, see above.

[**R1**] **Completeness of Thm. 2**: This is a subtle point which we will make clearer by making Thm. 2 more mathematically precise. Thm. 2 is an existence proof (“can have non-zero covariance” L185) so proof by construction is valid (L808). But it does not entail the WDH. We then make the admittedly informal argument based on experiment, which you criticize, that the WDH represents a typical case. We can further develop the discussion of this in L189-191.

[**R1**] **Does full-covariance ever help?**: 1-layer FC beats MF. See below HMC or UCI in Louizos & Welling [2016].

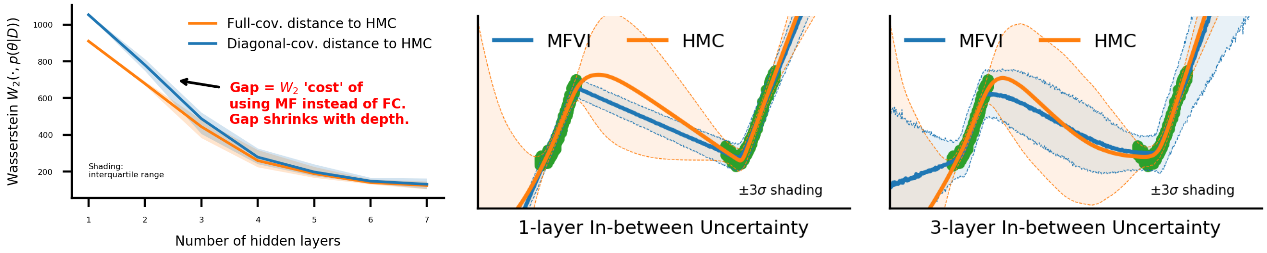
[**R2**] **Correlations v. structure**: All exps. compare same model w/ and w/o correlations, which disambiguates this.

[**R3**] **MFVI over-confident?**: 1) Thanks for Wasserstein suggestion. See LHS fig. below comparing W_2 distance of FC and MF Gaussian to HMC samples. The gap between W_2 lines is the distance cost imposed by using MF instead of FC. The gap shrinks with depth, and is negligible after 4 layers, supporting our argument. **This directly addresses your point and might justify an increased score**. 2) To clarify: our scope is defending MF assumption specifically, we agree that VI and unimodal apprx. have problems which are not in our scope. 3) Table 2 ECE: As you say, for SWAG, MF ECE is worse, but VOGN v. Noisy K-FAC tells opposite story. This makes sense if some but not all modes are apprx. MF. VI finds the MF modes but SWAG training is indifferent to this (L310-313). Also ECE differences are small.

[**R3**] **Non-Bayesian baseline**: Fascinating point: note deterministic baseline has no correlations! Success of deterministic NNs supports our argument. We will add this. Note also a general defence of Bayesian methods is out of our scope.

[**R3**] **Novelty**: Only Thm. 3 draws from Foong et al. 2020, which we are very explicit about, while the rest of our results were developed in parallel. Thm. 3 is only half a page of our paper, which makes it hard to see how this result is “overblown”. We will add discussion of Cherief-Abdellatif [2020], but observe (as you highlight) the setting considered has significant differences. We feel that while you acknowledge the novelty of our experimental contribution, you might also ascribe too much prescience to H&vC (see above) and have neglected key contributions (Thm.s 1 and 2).

[**R4**] **Exp. critique & Other**: 1) 1-layer HMC is below center. 2) Fig 3 sadly must be small (see C.3) but it does show MF beat FC (maybe unclear in black-and-white, we will make fig. less color-dependent). 3) We discuss KL not going to zero in L268. But also see Wasserstein fig. below LHS, where gap nearly disappears. 4) We agree deciding sufficient width is an unsolved problem and list this and other problems in L214-221. 5) Thanks for the detailed further comments.



1-layer In-between Uncertainty

3-layer In-between Uncertainty