We thank our reviewers very much for their time and constructive feedback! Responses are interleaved with the reviews and highlighted in red.

Reviewer 1

[…] I agree the factorized Gaussian posterior is used in some work. But this assumption itself introduces model error in most applications. From an applied point of view, adopting this method is totally fine. However, in the context of quantifying the mutual information and other types of information, should a more careful analysis by taking into account both the model error and useful information be implemented? There is no discussion about the potential model error here. […]

Any Bayesian method is rooted in the assumption that the model class captures the data generating distribution. We did not discuss model error in the paper since this assumption is very common, and it is hard to make any predictions without such assumptions about the data distribution.

Reviewer 2

[…] My understanding is that this conclusion (as it appears in the abstract) is not supported by the results in the paper. I am not saying that the conclusion is erroneous, just that the presented results (theoretical and numerical) do not seem enough to deduce it. The main reason is that the maximum capacity does appear to control the generalization error but the mutual information is only a lower bound for the capacity. What this means is that bounding the mutual information does not control the capacity, thus there is no control of the generalization error. […]

We believe that there is a misunderstanding. We have perfect control over the capacity and in fact, we set it to various values between 3 and 10 bits per dimension in the first experiment (see figure 3). This is done by choosing appropriate values for the prior and noise variance, resulting in capacities of our choice according to eq. 8. The data processing inequality (eq. 7) ensures that this capacity is an upper bound of the mutual information between data and parameters. Under the stated assumptions, bounding $I(D, \theta)$ leads to a bounded generalization error.

1) l.5 “…even when the KL-divergence in the objective is rescaled.” The authors refer to multiplying the KL-divergence term with the factor $\beta$ for the case of learned-variance Gaussian mean field inference. I do not know if this qualifies as a rescaling. It is an embedding of the objective function in a larger family of functions.

Correct, our analysis applies to this generalized class of objective functions. We did not intend to suggest that the family of functions remains the same. We updated the abstract for clarity.

2) l.54 “…we define the noisy model $p'(\theta, \theta^*; D) = p'(\theta)p'(\theta^*|\theta)p'(D|\theta^*)$...”. Do the authors actually define the noisy model to be like that or they deduce it as a possible expression for the joint
p'(θ,θ̃,D)? I am asking because I could not get the expression p'(θ)p'(θ|θ)p'(D|θ̃) when I start from p'(θ,θ̃,D). Is there some extra assumption that allows this that I am missing?

Correct, we define the noisy model p’ to be like that. The aim of this definition is to match the original objective.

Is this related to the assumption that the noise distribution is the same as the inference distribution as stated in l.51?

Correct, this is related. We just define the noise distribution p'(θ|θ) so that this assumption is met.

3) l. 85 “Yet, we believe that those bounds motivate the expectation that variational inference aids generalization by limiting the learned information.” There is a gap between the sentiment of this sentence (which is just expectation) and the certainty of the conclusion in the abstract that limiting the learned information does aid generalization. I have made that point also above.

This is a good point. We updated the abstract to reflect this.

4) l.166-167 “...applying Gaussian mean field inference of fixed scale to the model parameters...” What does fixed scale mean?

There is a model choice that determines whether we define a (fixed) noise standard deviation or whether it is inferred (learned). We call the first case “fixed scale”, the latter “learned scale”.

5) l.261 “Our experiments confirm the relevance of these other factors:”. I think that the authors mean “...the relevance of the following factors:”. That is the only way I can make sense of this sentence. There is a related ambiguity which comes at l.268 “This observed dependence on other factors...”. I suppose that by other factors the authors mean the factors in the previous paragraph but it is not clear.

We refer to the previous paragraph. We updated the text for more clarity.

6) l.268 “This observed dependence on other factors suggests that quantifying mutual information I_t(θ, D) of the actual distribution created by the learning dynamics might be a promising approach to explain why neural networks often generalize well on their own.” I suppose that the authors mean that since these factors (mentioned in the previous paragraph of the manuscript) do seem to affect generalization, quantifying the mutual information I_t(θ,D) of the actual distribution that is due to the learning dynamics could provide some indication, a link, between mutual information and generalization. However, I am not sure that this is what the authors mean. Please explain.

Correct, that is exactly what we mean! This link might explain why neural networks generalize well on their own, i. e. without explicit variational inference. We updated the formulation to make this more clear.
7) l.272 “On the other hand, the architecture choice also had an influence on generalization, which is expected by our theory since we only formulate a bound on mutual information that is completely agnostic to the actual model choice.” Actually, since the theory is agnostic to the actual model choice, then it cannot anticipate the influence of the architecture. On the contrary, being agnostic to the actual model choice means that it treats all model choices equally. The authors should change this sentence to reflect that.

We agree and updated our work to reflect this.

8) l.281 “We have explained the regularizing effects observed in Gaussian mean field approaches from an information-theoretic perspective.” This goes back to the conclusion appearing in the abstract. I do not think that the authors have fully explained the regularizing effects in Gaussian mean field approaches.

We agree and updated the formulation.