The manuscript “Gaussian Mean Field Regularizes by Limiting Learned Information” by Kunze, Kirsch, Ritter and Barber uses the concept of mutual information between learned parameters and data to examine the generalization ability of a neural network. In particular, the manuscript aims to explain the empirically observed superior performance of variational inference with a Gaussian approximate posterior compared to more flexible posteriors. The authors establish a maximum capacity which bounds from above the mutual information and also appears to control the generalization error (in experiments). Based on that, the authors conclude that bounding the mutual information between the parameters and data efficiently regularizes neural networks on both supervised and unsupervised tasks.

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.

Of course, there is a mathematical reason why this gap appears. There is recent work which the authors cite that relates the generalization error of neural networks to the mutual information between parameters and data. But this work assumes that inference is performed exactly. On the other hand, the authors use variational inference since exact inference is intractable. The result of using variational inference is that the authors can only express the mutual information as a lower bound of the capacity and thus there is no direct link to the generalization error.

The authors acknowledge this situation in Section 2.2 (Generalization Error vs. Limited Information). Therefore they suggest that the gap may be closed i) by performing more accurate inference in the noisy model or ii) by taking the dynamics of the training algorithm into account when bounding mutual infor-
mation. Suggestion i) is not straightforward since using more flexible variational distributions leads to a difficult estimation problem for the mutual information (see Section 2.5). How suggestion ii) will work is not clear to me even though I did read Section 5.2 repeatedly (see also point 6) below).

The manuscript is relatively well-written but there are parts that remained mystifying to me even though I read them repeatedly. I have made a list of those parts below. I think that some of these parts remained cryptic to me because I could not see the causal connection between the theoretical/numerical results and the conclusion of the manuscript.

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.

2) l.54 “..we define the noisy model $p'(\theta, \hat{\theta}, D) = p'(\theta)p'(\hat{\theta} | \theta)p'(D | \hat{\theta})...”. Do the authors actually define the noisy model to be like that or they deduce it as a possible expression for the joint $p'(\theta, \hat{\theta}, D)$? I am asking because I could not get the expression $p'(\theta)p'(\hat{\theta} | \theta)p'(D | \hat{\theta})$ when I start from $p'(\theta, \hat{\theta}, D)$. Is there some extra assumption that allows this that I am missing? Is this related to the assumption that the noise distribution is the same as the inference distribution as stated in l.51?

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.

4) l.166-167 “...applying Gaussian mean field inference of fixed scale to the model parameters....” What does fixed scale mean? Does it refer to the parameter $\beta$ somehow?

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.

6) l.268 “This observed dependence on other factors suggests that quantifying mutual information $I_t(\theta, D)$ 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 $I_t(\theta, D)$ 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.

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.

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.

Based on the above, before I recommend the manuscript for publication, I have to ask the authors to address the concerns above and change accordingly the conclusion of the manuscript or else provide convincing evidence for the current conclusion.