We thank all the reviewers for their insightful comments and suggestions.

Reviewer 1:

(1) Prior Literature (A)-(C): Thanks for the suggestions. We agree and will integrate them into the paper.

(2) Separate assumption latex environment: We fully agree. Thanks for the suggestion.

(3) Diversity: This is a good point. We will add a paragraph.

(4) Prior: It was not our intention to claim that this is a **novel** prior family. We will rewrite that.

(5) Discretized Prior: We fully agree with you that there are many arguments against the use of a discretized prior. It is not a natural choice. We do not advocate for this kind of priors, we use them because they allow us to characterize ensembles as variational methods which approximate the posterior distribution with a finite set of particles (instead of a continuous density belonging to some parametric family, as it is usually the case). One could, for example, use a mixture of Gaussians, which (in my opinion) will be more powerful and will also capture multimodality, but their treatment will be much more involved. But this characterization, which has been previously mentioned in [53], helps us to establish a direct link between our theoretical analysis and ensemble learning algorithms. So, the main purpose of Section 6.3 (and the use of the discretized prior) is to show that this work provides a candidate theoretical tool to study the generalization capacity of ensembles-like algorithms. In any case, we will add a caveat about the use of these priors and the reasons because we introduce them.

Bernstein/Hoeffding type assumptions: Theorem 1 was initially presented in [2] and was restricted to "losses/likelihoods (and priors) under which the usual Bernstein/Hoeffding type assumptions hold". But later, [17,45] showed that Theorem 1 was also valid under general unbounded losses.

Reviewer 2:

Thanks for enumerating the typos. We plan to send the paper to a native English speaker.

The paper focuses on density estimation because it is easier to present (and the notation is a bit simpler), but it readily applies to conditional densities too. Experiments were performed in supervised problems involving Bayesian neural networks because these are the settings where the community is currently paying more attention.

Reviewer 3:

Computational Complexity: Thanks for the point. You are right and we haven't discussed it in the paper. But we are glad to introduce the following discussion either in the main body or in the appendix. From our theoretical analysis, we derive a new **loss** function, introduced in Equation (4). The computation of the gradient of this loss function is more complex than the standard variational loss function (see Equation (2)), but not too much. Looking at Equation (C.12), we can see that it is a constant factor 2 more complex. However, it is not clear if the optimization of this new loss function is or not more involved than the optimization of the standard variational loss function (Equation (2)), this is going to depend on the loss landscape and the noise of the gradients. More research in this direction is needed in order to properly answer this question.

Evaluation Simpler Models: This is also a good point. This is mostly a theoretical paper, but we decided to illustrate this approach in a *modern* or *hot* problem to highlight that this theoretical analysis could be potentially very useful (and it is not just a theoretical curiosity). In any case, Figures C5-C9 in the appendix further illustrate this approach on simpler models. We are currently working on an evaluation based on a mixture of Gaussians where the number of components is misspecified, and we plan to include it in the appendix (if the paper is accepted). And, of course, future works will focus on extensive empirical evaluations of this approach.

Reviewer 4:

Comparisions with stronger baselines: Thanks for the reference, we were not aware of SWAG, and we will try to include it in the experimental evaluation, we think it is a fair comparison. We use standard mean-field variational because is simple and widely used in this kind of problems. In this case, the mean-field approach defines the solution space. So, the variational approach gets the minimum of Equation (2) within this solution space. We think it is fair to compare with the minimum of Equation (4) within the same solution space. Because, in this case, we can clearly evaluate which is the effect of considering the new presented loss function. Richer solution spaces could be considered too (e.g. "The k-tied Normal Distribution" Swiatkowski et al (2020)). But the fair comparison will be to evaluate both the minimum of Equation (2) versus the minimum of Equation (4) within the same solution space. Moreover, one could eventually think if a counter-part version of SWAG could be defined using the insights derived from this paper. In any case, we agree that an extensive empirical evaluation of this approach is needed to see whether this research direction is able to provide new SOTA results.