We would like to thank all three reviewers for their constructive assessment of our work.

**Reviewer 1**

*Importance of the experimental results:* Non-vacuous generalization bounds are arguably desirable for delicate machine learning applications. Since such results are actually rare for neural networks, we consider important to show empirically that our approach leads to sound performances. We believe that this can encourage others to develop similar methods that will eventually lead to higher practical impact, following this work.

*Error bars:* The reviewer concern is *partially* addressed in Section B.3 of the appendix, through the Monte Carlo sampling size effect study. Indeed, Figure 5 displays the test error — with error bars — for PBGNet and PBGNet$_\ell$ when varying the sample size. Note that each result is obtained by averaging over 20 repetitions of the learning procedure, each of them executed on different (random) train/test/valid dataset splits, and the stochastic gradient descent is initialized with different random weights. That being said, we undertake to push forward the variance analysis for the final version of the paper, as detailed in Reviewer 3 *Experiments* section.

**Reviewer 2**

*Choice of the activation function*: A variety of activation functions exists in the literature and we obviously do not focus on the most common one. Nevertheless, the sign activation function is used in the binary networks cited in the paper, notably to reduce the memory footprint of such networks which could be embedded on small devices (e.g., Bengio, 2009). In our context, the sign activation is crucial to apply the mathematical trick, and express the predicted outcome as the $\mathrm{erf}$ function. This gives a principled way of training the network and derive the PAC-Bayes generalization bound.

*Choice of prior*: The bound holds regardless of the choice of the prior $\mu$. In our experiments, we centered the prior on the SGD initialisation weights (as in Dziugaite and Roy, 2017), which corresponds to the real-life scenario where one does not have prior knowledge about the task at hand. The PBGNet$_{\mathrm{pre}}$ variant of our algorithm opens the way for using a prior from a precedent learning task, as the transfer learning scenario mentioned in the paper (see Lines 219-222).

*Mismatch between theory and experiments*: We worked hard to provide a rigorous and honest empirical study of our theoretical analysis strengths and weaknesses. It is certainly disappointing that there exist good predictors with trivial bounds (as mentioned by Rev. 3), but we still managed to obtain very tight bounds for more than decent predictors. Note also that we compared to $\tanh$ networks as this activation is similar to the $\mathrm{erf}$ function derived from our analysis. It allows us to use the same optimisation scheme and hyperparameter grid to compare the methods on equivalent basis.

**Reviewer 3**

*Improvement due to the binary activations:* Relying on binary activation function allows us to express a close-formed solution for the PAC-Bayes bound, without other assumptions than the *iid* one. Another appeal of our bound: it relies more on the network architecture (*e.g.,* $d_k$, the layer width of each layer $k$, appears in Eq. 16) than previous results (in the seminal work of Dziugaite and Roy, 2017, a posterior distribution is used over a set of a neural network weights, without taking into account any other architecture specific information).

*Bounds for (a unique) binary activated networks:* The reviewer is right to say that our analysis does not provide guarantees for a single deterministic binary activated (BAM) network, but for a continuous aggregation of such BAM networks. We clearly mention this in the introduction (Lines 26-27), but we agree that Line 16 may be ambiguous and we will rephrase it. That being said, several points are worth mentioning: (i) Even if computationally expensive, our predictor closed-form expression is deterministic; (ii) The prediction using Monte Carlo sampling empirically shows a small standard deviation consistently below $10^{-3}$, as discussed in the appendix (see Lines 432-448); (iii) In our experiments, we observed that predicting with the single *Maximum-A-Posteriori* BAM network generally gives results remarkably close to the aggregated PBGNet predictor (hinting that the posterior may be quite peaked). Recall that this unique BAM network represents exactly the same prediction function as its mapped tree predictor. This preliminary observation suggests that the bound can provide an appropriate guide to train a BAM network.

*Link with similar optimization procedures:* There is definitely a strong connection between our optimization procedure and the REINFORCE method. We greatly thank the reviewer for pointing us the relevant literature; we were genuinely not aware of it. Indeed, we will rewrite parts of Section 3 to highlight these connections, and express our sampling scheme as a particular case of a general technique rather than a new one. As a matter of fact, we think this will enrich the paper, as our PAC-Bayesian analysis support existing strategies to train non-differentiable neural networks. Moreover, the connection with Variational Bayes (for a fixed $C$) is also very relevant, and will be mentioned in the paper.

*Experiments:* We share the reviewer concerns about the lack of standard deviation analysis in Table 1. Given our computing resources, we cannot provide these results in this rebuttal. Nevertheless, we commit to produce for the final version of the paper a thorough stability analysis with 20 different random train/test/valid splits for all six datasets and five considered models of Table 1, for a slightly reduced hyperparameters search grid. Furthermore, we will add a study of the training sample size effect on PBGNet$_\ell$ and PBGNet with fixed parameters for the biggest dataset (mnistLH).