We thank the reviewers for the time they have spent on this work and their valuable comments. We will try our best to make the suggested improvements in the final version. Below please find the responses to the main raised comments.

•**R1: Section 5 seems disconnected.** The results of Section 5 are based on a direct application of "conditioning" (Lemma 3), and thus a necessary element of our work. Note that Lemma 3 is a quite general result which is applicable to any concave information-theoretic bound. Its application along with Lemma 2 produces the results of Section 4. But it is not restricted to this case (the title of Section 4 will be changed to eliminate this confusion). To demonstrate this, we decided to apply the technique on the chaining bound of [2] which might seem quite different. Note that the bound of Eq. (21) is a concave function of $I(\tilde{W}_k; S), \forall k$ and this is enough to apply conditioning (as was done in Eq. 26). Another application of conditioning, omitted due to space constraints, is a generalization of Lemma 1 which uses Cumulant Generating Functions (CGF) and the square root function is replaced by another concave function (the inverse of Legendre dual of CGF in case it exists) and conditioning technique still applies.

•**R1: Inequality F.3.** Thank you for this careful observation. This inequality is correct under the assumption that W is a deterministic function of $X_{\mathcal{W}}$ (e.g. a deterministic ERM); i.e., we have $\tilde{W}_{k-1} \perp\!\!\!\perp \tilde{W}_k | W \Rightarrow \tilde{W}_{k-1} \perp\!\!\!\perp \tilde{W}_k | X_{\mathcal{W}}$. This assumption will be explicitly stated in the theorem and the text will be updated adequately. The rest of the reasoning of Section 5 for proving the generalization bound remains unchanged (actually the term "rate-distortion" was used taking the deterministic setting in mind, as stated in Line 199). Note that for non-deterministic algorithms, $I(\tilde{W}_{k-1}, \tilde{W}_k; X_{\mathcal{W}})$ in Eq. (F.2) can be bounded by $I(\tilde{W}_{k-1}, \tilde{W}_k; W)$ which yields similar results for $I(\tilde{W}_k; W)$ instead of $I(\tilde{W}_k; S)$.

•**R1: Theorem 4.** This theorem is a regathering of simple information-theoretic inequalities (such as the mentioned corollary of Cover) to summarize some conditions on the graphical model which can result in conditional mutual information being less than its unconditional counterpart. We will emphasize on this and add the related reference.

•**R1: Beyond uniform bounds.** The main purpose of Section 5 is to show that by utilizing the conditioning technique it is possible to close the gap between the classic generalization bounds of VC-theory and information-theoretic bounds. Though the final results of that section do not provide new bounds, it shows that the information-theoretic approach is at least as strong as the VC-theory (a question which was frequently raised in this line of work and partially addressed in [15]). It is worth mentioning that we are working on more advanced usages of these techniques for future work.

•**R1: Literature review.** We will elaborate more about the mentined paragraphs. We will also include discussions about Audibert and Bousquet papers and some discussions about PAC-Bayesian bounds in general.

•**R1 & R2: Relation with Steinke and Zakynthinou [15].** We think an approach similar to the one discussed in Section 5 can be used to answer the first conjecture presented by that paper (as this is another track we are following, we should add that the connection is not trivial). But, here we want to clarify the differences in our problem settings. Their work on VC-dimension and the conjectures therein are about proving the "existence of an ERM algorithm" which has a small CMI. On the other hand, they also show that CMI can actually be quite large for some other ERM algorithms and it is not controlled by d_{vc} in general (not even a bound with $\log(n)$ factor exists). Thus, there are some ERM algorithms whose generalization cannot be explained by CMI, which shows that fundamentally CMI is not as expressive as the standard uniform convergence. In their first conjecture (the one related to our work in Section 5), they even had to go beyond ERM algorithms and allow ϵ empirical risk. Our work does not have these limitations. To do that, we had to go beyond regular CMI and also incorporate processing (Section 4.2) and chaining (Section 5) techniques.

•**R2: Extending table.** As suggested, we will try to include the results of Negrea et al. [10] in Table 1 (it is an extension of the second row which in this case a group of indices are conditioned) and also a column for references.

•**R3: Demonstrating that bounds are tightened.** We acknowledge that the discussion on this matter is better to be clarified in the final version. But, these techniques can indeed tighten the bounds: 1) in Section 4 it is demonstrated that the results of [5] and [8] on tightening the mutual information bounds can be explained using the proposed framework in a straightforward manner. There are some theoretical and empirical discussions on these papers demonstrating that these bounds can be (much) tighter. These were not included in the text due to space constraints. We will add some elaboration on this. 2) basic mutual information bounds can be very large even for simple hypothesis sets (as demonstrated in [4]). This problem was partially addressed in [15] and was followed in the current work by proving optimal bounds comparable to VC-theory (both of these approaches use the conditioning Lemma at their core). This is another application of using these techniques to improve the bounds.

•**R3 & R4: Concrete and recent practical applications.** We can use the extra page to discuss some recent applications of the bounds in practice (in particular more discussion on SGLD which was suggested by reviewers 4). We believe this discussion will demonstrate how such tightened bounds can be used in practice to analyze neural networks.

•**R4: Mathematical novelty.** It is hard to claim true mathematical "novelty", but here we summarize some of our contributions: 1) Extending Lemma 1 to stochastic mappings and also providing general formulations for tail bounds, 2) formulation of conditioning Lemma as a general technique explaining a variety of previous results and applicable to new situations, 3) characterization of the chaining mutual information bound as a (variant of) rate-distortion problem, 4) combining chaining and conditioning to close the gap between information-theoretic and uniform convergence bounds.

•**R4: More complex graphical models.** The studied graphical models already have a variety of applications, as discussed in Sections 4 and 5. But, we agree that much is remained to be explored in future work.