We thank all reviewers for their time and valuable comments. Below we address each review in turn.

Reviewer 1 : We incorporated your feedback on the presentation and added more explanation within the proofs.
Significance of the results in Section 2: The purpose of Section 2 is to unify the two main information-theoretic approaches for studying generalization. [21] gave instances where the gap between CMI / IOMI was large, but did not provide a general result. Thm. 2.1 shows $\mathrm{CMI}^k_{\mathcal{D}}(\mathcal{A})$ is a tighter measure of dependence than $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A})$ for *any learning algorithm* and *any data distribution*. This is an important, novel inequality, even if it admits a straightforward proof. We also address the role of the size of the super-sample in CMI. In [21], CMI is defined using a super-sample of size $2n$ ($k = 2$) only. Thm. 2.2 addresses this by showing $\mathrm{CMI}^k_{\mathcal{D}}(\mathcal{A})$ and $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A})$ agree in the limit as $k \to \infty$. We conjecture the finiteness assumption is an artifact of our analysis, though we don't have a more general proof.
Novelty of the proof techniques: Assuming this refers to Section 2 (as our new bounds for Langevin dynamics represent such a clear and material advance), we believe our results in Section 2 are essential for understanding the relationship between CMI and IOMI, irrespective of the novelty of the proof techniques. That said, we think there are several fun and subtle arguments buried in the proofs of Section 2.

Reviewer 2 : Insights from the LD bound: A prevailing method for analyzing the generalization error for iterative algorithms is via the chain rule for KL, using priors for the joint distribution of weight vectors that are Markov, i.e., given the tth weight, the $(t+1)$th weight is conditionally independent from the trajectory so far. Existing results using this approach accumulate a "penalty" for each step. In [14, 15, 16] the penalty terms are (resp.) the squared norm of the gradients, the trace of the gradient covariance, and the squared Lipschitz constant. The penalty term in our paper is the squared norm of "two-sample incoherence", defined as the squared norm of the difference between the gradient of a randomly selected training point and the held-out point. However, the use of chain rule along with existing "Markovian" priors introduces a source of looseness, i.e., the accumulating penalty may diverge to $+\infty$ yielding vacuous bounds (as seen in Fig. 1). The distinguishing feature of our data-dependent CMI analysis is that the penalty terms get "filtered" by the online hypothesis test via our *non*-Markovian prior, i.e., our prediction for $t+1$ depends on whole trajectory. When the true index can be inferred from the prev. weights, then the penalty essentially stops accumulating. For instance it can be seen in the middle column of Fig. 1 that the penalty term of our paper is close to or larger than [15]'s. Nevertheless, the online hypothesis test discounts our penalties, yielding non-vacuous bounds where earlier bounds fails to.
Designing better algorithms: We're aware of work using PAC-Bayes bounds as training objectives. A similar approach based on $\mathrm{CMI}^k_{\mathcal{D}}(\mathcal{A})$ is an avenue for future work, though use of data-dependent priors presents new challenges.

Reviewer 3 : We thank the reviewer for their exceptionally in-depth review. As the reviewer engaged with the material at a high level, we believe it would be reasonable for Rev. 3 to increase their confidence score.
8. Regarding the first step in Eq. 9, the answer to your question is "yes". We will clarify this in the paper.
9.–12., 17. We included more intuition and explanations, and adopted your suggestions regarding the presentation.
14.(1) The prior in [15] doesn't exploit the optimization trajectory; it's a Markov process. Our prior is not Markov; it "learns" the identity of the held out data point from the history. This is an important difference between the bounds.
14.(2) Recall the variational representation $I(X;Y) = \inf_P \mathbb{E}[\mathrm{KL}(Q(X)\|P)]$, where the infimum achieved by $P = \mathbb{E}[Q(X)]$. If we always use the "optimal" prior in Eq. 3, we have "KL-based" bounds are tightest when $m = 1$. However, since the priors for different values of m are probability kernels with different input spaces (J is a subset of size m), we cannot compare the bound for a fixed prior as m varies as it is impossible to have the same prior for different m. Also note that to compute optimal prior we would need to know the data distribution.
15. For $m > 1$, the problem could either be reduced to a single selection between 2^m different alternatives, or to m separate binary tests and a union bound. We consider this to be potential future work.
16. Fig. 2 in App. G plots the test and training error. We will move the plots to the body in the final version.

Reviewer 4 : 1. As stated in Sec. 1, our focus is on $[0,1]$-valued loss functions. In Thm. 1.1, we translate the sub-Gaussian results from [18] and [24] to bounded loss. Thm. 1.3, however, is taken from [21], which also does not provide generalization bounds for Sub-Gaussian losses. While extensions to sub-Gaussian losses are often trivial, extensions of the bounds in Section 3 are not immediate due to our particular use of boundedness. 2. Thm B.1 shows that asymptotically, as the super-sample size increases, the leading coefficients can be taken to be the same. Also, the $\mathrm{CMI}^k_{\mathcal{D}}(\mathcal{A})$ and $\mathrm{IOMI}_{\mathcal{D}}(\mathcal{A})$ play the same role in their respective bounds, capturing the dependence of the bound on the complexity of the learned hypothesis, and it's dependence on the training data relative to the sample size.
3. Please see response to Reviewer 2's second question. 4. Perhaps this is a misunderstanding, but, in Sec. 4, we study the generalization error of the Langevin dynamics algorithm as an example of noisy, iterative algorithms using the bounds proposed in Section 3. We apply our bounds to large, overparametrized neural networks. We estimate an upper bound on the mutual information via the KL between two Gaussians, to sidestep computational intractability of MI. We use a data-dependent technique to estimate the mutual information in order to obtain tighter and easily simulated bounds. In Sec. 4 we evaluate our proposed bound numerically and find that it is non-vacuous across various neural network architectures and datasets where # networks parameters $\gg$ # training points. Finally, for the case of Lipschitz loss function, we show in Remark 4.6 that our bound does not depend on the number of parameters.