1. All reviewers. We acknowledge the paper uses some unusual notation and, in parts, assumes background in information theory. We now give all standard definitions. We have corrected typos and grammatical errors.

2. Reviewer #1 asks how the mutual information bounds are used and why we seem to focus on KL bounds, which resemble PAC-Bayes bounds. Our SGLD bounds are built using Theorem 2.2 (mutual inf. based) and Theorem 2.3 (KL based). Note that the mutual information in Thm. 2.2 involves unknown distributions and cannot be computed directly. We get explicit bounds on mutual informations via the known identity $I(X;Y) = \inf_P \mathbb{E}[\mathrm{KL}(Q(X)\|P)]$, where $Q(X)$ is the conditional law of Y given X, and P ranges over all distributions on the same space as Y. (The Ps are called "priors" in the PAC-Bayes literature.) Eqs. (8),(9) exploit this identity. We now highlight this identity and explain its role clearly. One of our key contributions is inventing data-dependent priors (i.e., the prior depends on S_J) that yield explicit bounds that are much more adapted to the true unknown distribution, and thus yield much tighter bounds. These same data-dependent priors appear in Thm. 2.3.

3. Reviewer #2 points out some missing related work. First, we acknowledge and thank the reviewer for their in-depth review. Our original related work section focused on immediate predecessors of our results. We agree that it makes sense to provide context by citing a broader range of related work. We will add references to work on PAC-Bayes ([Rivasplata et al., 2018], [Dziugaite & Roy, 2018], etc.), on information theoretic bounds ([Russo & Zou, 2016], [Raginsky et al., 2016], etc.) and on the Langevin algorithm ([Raginsky et al., 2017], etc.). Reviewer #2 also points out that Russo and Zou provide a correct proof for Xu–Raginsky; we now cite them for this as well. If the reviewers believe there are other articles we should cite, we would be grateful if they would could bring those to our attention.

4. Reviewer #2 points out that in Equation (31) it is not clear that the RHS is constant in J. Note that S_J^c is a random variable that is a $(n-m)$-tuple. The statement is true since the mutual information $I(W;S_J^c)$ is a number that only depends on joint law of (W, S_J^c), not on the values of W or S or J in a particular realization.

5. Reviewer #2 requests that we compare actual generalization bounds as opposed to "key quantities". Reviewer #1 also remarks that our figure labels are not informative. We note that comparing actual generalisation bounds is made cumbersome due to the non-linearity of our bound. In particular, the form $\mathbb{E}[\sqrt{\dots}]$ can be upper bounded by $\sqrt{\mathbb{E}\dots}$ using Jensen's inequality, but this change makes a material difference in the quality of the bound. Moving as many expectation outside of the $\sqrt{\cdot}$ as possible is one of the key advantages of our work over other related work. For this reason we plotted the *sum of the gradient covariances* and *sum of gradient norms* against epoch number. We clarify our numerical results and properly label our figures in the final version. To provide evidence that our method does improve over other work and yield a non-vacuous bound, we provide Monte Carlo estimates of our bound (Eq. (22)) and that of [Mou et al., 2017] (their Eq. (69)) in the table below. The bounds are dependent on architecture, data

	MNIST with MLP			**MNIST with CNN**		
	Epoch 1	Epoch 2	Epoch 3	Epoch 1	Epoch 2	Epoch 3
Training Classification Error	$25.52 \pm 0.08\%$	$16.17 \pm 0.04\%$	$12.38 \pm 0.02\%$	$21.89 \pm 0.21\%$	$14.07 \pm 0.14\%$	$10.78 \pm 0.10\%$
Test Classification Error	$25.57 \pm 0.06\%$	$16.29 \pm 0.04\%$	$12.45 \pm 0.02\%$	$22.93 \pm 0.20\%$	$14.72 \pm 0.14\%$	$11.24 \pm 0.09\%$
Generalization Gap (Mou et al.)	$33.8 \pm 1.4\%$	$76.0 \pm 3.0\%$	$139.4 \pm 5.9\%$	$46.5 \pm 2.2\%$	$78.6 \pm 3.0\%$	$130.6 \pm 4.6\%$
Generalization Gap (Our Bound)	$10.0 \pm 1.6\%$	$20.5 \pm 4.0\%$	$29.0 \pm 6.7\%$	$15.3 \pm 2.8\%$	$25.8 \pm 4.4\%$	$49.2 \pm 10.4\%$

distribution, and hyperparameters. We did *not* attempt to tune the hyperparameters to make the predictive performance or generalization bounds better. If the reviewers think it's worthwhile, we can add these results to the final version of the paper. However, we believe our current experiments are also sufficient and demonstrate the theoretical advance. Note that, in Appendix G, we already give an example where the improvement of our bound over related work can be made arbitrarily large when the data distribution is sufficiently heavy tailed, due to the order of $\mathbb{E}$ and $\sqrt{\cdot}$ yielding $\mathbb{E}\,|Z| \ll \sqrt{\mathbb{E}Z^2}$.

6. Reviewer #3 suggests that we state the results for LD and SGLD as formal theorems. First, we acknowledge and thank the reviewer for their in-depth review. We like this idea, thank you.

7. Reviewer #3 suggests that we should provide more intuition regarding the roles of the subsample J and the auxiliary random variable X. The utility in keeping track of these quantities as analytical tools is one of the main contributions of our work. The subsample, J, is the key to getting a data-dependent bound – in order to compare the outcome of an algorithm to that of a similar algorithm run on a subset of the data we need to select such a subset, and the randomness in the selection of the subset allows us to turn this comparison into a generalization bound. In the case of SGLD, the auxiliary variable will represent the order in which the indices of our data points appear in the minibatches. In general, the auxiliary variable collects together all nuisance variables we wish to couple.

8. Reviewer #3 intuits that for SGLD the generalization bound should be of the same order as that for LD, while our work presents a seemingly slower rate. The same lower order rate for SGLD may be found in related work, such as the PAC-Bayesian bound of [Mou et al., 2018]. Proving a generalization bound of order $O(n^{-1})$ for SGLD without unrealistically strong stability assumptions is an open problem. If we consider a constant number of iterations (i.e., less than one epoch of SGLD), our theorems can yield a $O(n^{-1})$ rate, but with worse apparent dependence on the number of iterations. We did not include this result in the submitted version as we believe results for ≥ 1 epoch are of more interest, however some stability based work essentially depends on considering only the fractional epoch regime. If the reviewer thinks that it is worthwhile, we can expand on these issues in the final version.