We thank the reviewers for the useful feedback. We will add the accidentally missing legend to Fig. 2. (orange line is the accuracy after removal of memorized examples and green line is accuracy after removal of the same number of randomly chosen examples). Responses to specific comments are below:

R1

more experiments to strengthen the validation of the theory ▷ We welcome suggestions of other experiments. The correlation between memorization and influence is addressed by our experiments in Fig 2 since there we measure the cumulative influence of examples with mem. value above some threshold. In particular, it implies that memorization and overall influence are positively correlated.

R2

the author propose a "closely-related" statistic, called infl_m, by keeping m random subsets instead of removing one example ▷ This is not an accurate description of the estimator. We estimate the effect of removal of a single example from a random subsample of the original dataset.
some explanations missing: relationship between infl and infl_m, and between stddev and time complexity ▷ The relationship is explained in Lines 86-89. infl_m is not equal to leave-one-out influence but the relationship between them is that of using a smaller random subset instead of the entire dataset and then taking the expectation. Note that this is also the classical jackknifing approach in statistics. The relationship between std. dev and time complexity of estimating the infl_m for all training examples at the same time (which is what we need) is stated in Lemma 2.1.
need rudimentary experiment to show effectiveness of the proposed method, compared with naive monte carlo estimation ▷ Our estimator is formally equivalent to the naive way to do it. The point of our algorithm is that it estimates all the values at the same time.
paper reorganization and title change ▷ Our primary contribution is validating and clarifying an explanation for a fundamental phenomenon in machine learning. The title and organization are aimed at making this clear. This suggestion appears to be based on a different view of our contributions with which we respectfully disagree. However we welcome and will definitely consider concrete suggestions about the title and organization.

R3

clarification on the long tail theory. Assume we define a more general measure. . . ▷ We do not see how the proposed definition captures the intuitive notion of memorization since the value is large even if a single example out of the k that were removed is not fit by the model. More generally, inference based on sets of examples is what distinguishes the traditional view of learning from memorization.
Near duplicated examples are dataset artifacts, not demonstration of long tail ▷ Indeed very high-influence pairs usually come from artifacts of data collection. However a large fraction of high-influence pairs that have somewhat lower values (in the 0.15-0.3) range do **not** look like such artifacts. So while memorization may be unnaturally important for CIFAR and ImageNet due to these artifacts it would still be important without them.
Extend the analysis in pp.7 to last 2, 3, or more layers? ▷ Thank you for the interesting suggestions. We are definitely planning additional experiments related to this work (and hope that others will do them too).
why $m = 0.7n$ not $0.5n$? ▷ Larger m makes the value closer to the original leave-one-out estimator and better at estimating marginal utility since the models become closer to the one computed on the entire dataset. m=0.7n is both quite close in accuracy to full-dataset models (unlike 0.5n) and is sufficiently efficient (efficiency drops linearly as fraction approaches 1).

R4

The only concern I have is that the applications... ▷ First, by far our main goal is understanding of memorization, a fundamental question about ML that has been puzzling the research community since the "Understanding Deep Learning ..." work Zhang et. al. The development of the influence estimator is just a potential bonus and thus we do not provide a detailed comparison with existing methods. In terms of efficiency note that our method simultaneously estimates the influence of all training examples on all datapoints. We are not aware of any method that can do that more efficiently and provide results of comparable quality. That said, we agree that efficiency is a concern for these applications. We believe that it is possible to develop more efficient estimators of comparable accuracy but leave it for future work. To stimulate this work we have already made the values of our estimator on CIFAR-100 and ImageNet publicly available.
Randomness from mini-batch ordering [Toneva et al, ICLR2019] ▷ The definition of influence/memorization contains expectation over the randomness of the algorithm. So our estimator measures expected memorization over all possible choices of minibatches. Also note that despite the use of a related "forgetting" word, the notion is completely unrelated to memorization that we study. We will clarify that in the related work section.