Author Response

We thank the reviewers for their valuable feedback.

R1: Comparison to "training with noisy labels".
R3: Comparison to "loss based selection method" from [DataLens IJCNN 20].

We were simply following an evaluation technique proposed by the two previous papers (influence functions, representer) on the topic. In this sense, identifying mislabeled examples using self-influence is simply a way to compare *influence techniques*. We do not claim to be the best way to fix or work with mislabelled data.

R1: Other than fixing wrong labels, the influence measured by the current method is not very easily assessed.

It is more challenging empirically to evaluate influence techniques (which *depend on how the model operates) in comparison to prediction/classification problems (where ground truth is specified independent of the model). Besides the "fixing labels" eval, we also provide conceptual arguments in favor or our method (Appendix: Section A), and comparative visual results on CIFAR (Appendix: Figure 6) and MNIST (Appendix: Figure 9).

R2: The experimental setup involves introducing an artificial percent of mislabeled samples. Is the performance of the method influenced by choosing a different percent?
R3: I expect TrackIn to perform poorly when we increase the mislabelled data.

Yes, one would expect *any self-influence based technique to perform poorly when the fraction of mislabelled data is high (say >30%). But this does not imply that TrackIn would do worse than representer or influence functions.

That said, we picked what we thought was a practically reasonable rate of mislabeling.

R2: In a non-toy dataset or in one with less wrong labels, it would be difficult to use this solution to cherry pick by hand mislabeled samples. Reporting also the precision here would be helpful to know where we are from this perspective.

The goal of the evaluation with a fixed percentage of mislabeled examples is to compare with prior works which also use the same metric. The trend should be the same regardless of precision or recall. We agree that reporting precision would be helpful in a "non-toy" dataset with less wrong labels and we will make this point in our next revision.

R2: How do the authors explain the difference between the classes ratio for mislabeled examples in different checkpoints?... at the end of training all classes have a similar number of mislabeled examples in top 10.

During the training process, the decrease in loss for each class (averaged over instances of the class) is not uniform. Frogs and Deers converge pretty early, and then Trucks. Therefore, for earlier checkpoints, the self-influence technique is more effective on these classes. In the final checkpoint, the model has converged to 99% accuracy, i.e., it is doing well on all classes, consequently, the performance of the self-influence technique is similar across classes.

R3: Checkpoint ensembling is a widely used technique One can argue that influence functions can also benefit from the checkpoint ensembling. Also, the paper should cite prior work related to checkpoint ensembling as a motivation for picking multiple checkpoints.

Notice that for us checkpoint ensembling *arises* from trying to practically implement the mathematical form of Idealized TrackIn (Lemma 3.1); in this sense our motivation for using checkpoints is perhaps different. We will definitely cite the suggested literature to point out the resemblance.

While other influence techniques may also benefit from checkpoint ensembling, they remain harder to implement than TrackIn.