We thank the reviewers for their constructive feedback, especially for Rev#2 and #4 for suggesting related studies. We will improve the readability of the paper, and enrich the discussion with suggested related studies. Please find below the answer to the reviews (apologize it is not exhaustive due to space limitation).

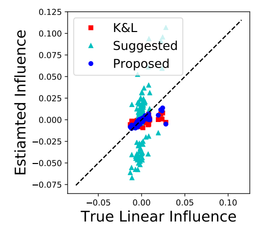

Figure 1: DNN: Adult

Rev#1: mere sum over the gradients The suggested estimator is expressed as $\Delta\theta_{-j} = \sum_{k=1}^{K} \frac{\eta_{\pi_k(j)}}{|S_{\pi_k(j)}|} g(z_j; \theta^{[\pi_k(j)]})$. Figure 1 shows that the suggested estimator did not work well for the adult dataset with DNN in the experiment in Sec7.1, which suggests that the use of Hessian in the proposed estimator is essential. We will add this result in the future version. Also, please note that we store the model parameter $\theta^{[t]}$ but not the gradients in each step of SGD. The stored parameter is first loaded to the model, and then the gradient for each instance is computed.

Rev#1: fair comparison? [Koh & Liang] mentioned that their estimator (with the diagonal regularization) is effective for DNNs. The comparison is therefore essential to show that our estimator is more suitable for DNNs.

Rev#2: human-in-the-loop? We are grateful for a suggestion for clarifying our contribution. Our method is automatic and does not require user intervention. We will update the abstract/introduction to be aligned with the proposed method. We mentioned the user intervention because that is a common way the most data scientists do for data cleansing.

Rev#2: relation to noise tolerance, label cleansing, anomaly detection We appreciate you for raising so many related research topics. Anomaly detection looks for instances away from the data distribution, however, it is not guaranteed that such instances are influential to the model performance. The proposed method directly looks for instances that minimize the validation loss. The results in Sec7.2 confirm that this direct approach is more effective. The noise tolerant learning and label cleansing will be interesting related topics. We will cite references and discuss them as related studies. The difference from our study is that these studies assume that the label noise is an only issue. However, as Figs 11 and 12 show, the model performance depends not only on label noises but atypical inputs also. For example, in Fig11, we can find several atypical instances that even human cannot label them confidently. These atypical instances should be removed from the training rather than fixing the labels because we cannot put correct labels to them.

Rev#2: how others have handled non-convexity To our knowledge, none of the raised studies can handle non-conveixty. Capability of handling non-convexity is therefore our notable contribution. [Koh & Liang; Zhang, et al.] assume convexity explicitly, while [Khanna, et al] does not (but it is implicitly assumed in Proposition 3). The active learning studies [Settles, Craven & Ray], [Cai, et al.], [Cai, et al.] are evaluated only on convex problems such as SVM and logistic regression. The theoretical analysis given by [Cai, et al.] is limited to these convex problems also.

Rev#2: clear contrast with [Koh & Liang] and works that follow We adopted [Koh & Liang] as the baseline and excluded [Zhang, et al.] and [Khanna, et al] because (i) [Zhang, et al.] is devoted for label collection, which is different from our goal (i.e. removing harmful instances); (ii) The method of [Khanna, et al] is computationally very expensive, which requires computing the inner product $\langle \nabla_\theta \ell(z_i;\theta), \nabla_\theta \ell(z'_j;\theta) \rangle$ for all the training instances z_i and for all the validation instances z'_j (the time complexity is at most $O(N^2)$). We will clarify this point in the future version.

Rev#3: DNN experiments are understandably bad compared to logreg For the non-convex case, as Theorem 6 shows, the estimation error can grow as the SGD steps T gets large. In our preliminary experiment, we observed this holds true in practice. The less accurate estimation compared to logreg in Fig1 and Tab1 is therefore a natural consequence. Constructing a more accurate estimator would be an important future direction.

Rev#4: overfitting problem This is a consequence from the definition of SGD-influence. SGD-influence (as well as ordinary influence) considers removing only one instance, and ignores the higher-order interactions between the instances. Extending the method for multiple instances is an important future direction. We will clarify this point.

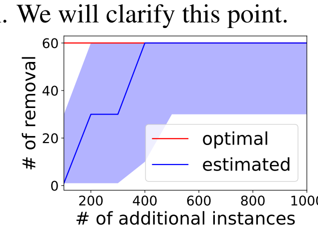

Figure 2: Removal on MNIST

Rev#4: how much data to be removed A reasonable solution would be selecting the number of removal that minimizes the error on an additional set. Note that this additional set is used only for determining a single scalar, and we do not need many instances for this purpose. In both MNIST and CIFAR10, we observed that preparing only 500 instances would be sufficient. See Figure 2 for the case of MNIST (shade = variation among random repetitions). We will add this result in the future version.

Rev#4: related works [1, 2, 3] We appreciate you for raising related studies. Unfortunately, the reference [3] was missing in the review, and we therefore checked only [1] and [2]. First of all, we are happy to find out that several different approaches are studied. We will include them as related works in the future version. The advantage of our study is in theories of the estimation error. We believe establishing solid theoretical foundation is essential to move the entire field forward. We hope our study to be a first step towards establishing further principled and sophisticated algorithms for automatic data cleansing in the future.