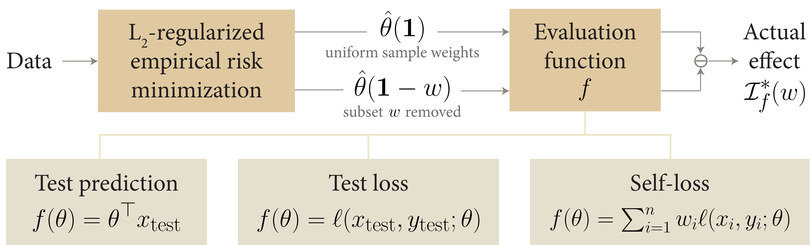
On the Accuracy of Influence Functions for Measuring Group Effects

Stanford University

Pang Wei Koh* Kai-Siang Ang* Hubert H. K. Teo* Percy Liang

OVERVIEW

Goal. Quantify how a model's predictions depend on its training data by measuring the effect of removing training points and retraining it.



Prior work. Influence functions accurately estimate the effect of removing **individual** training points via a 1st-order Taylor approximation.[1, 2]

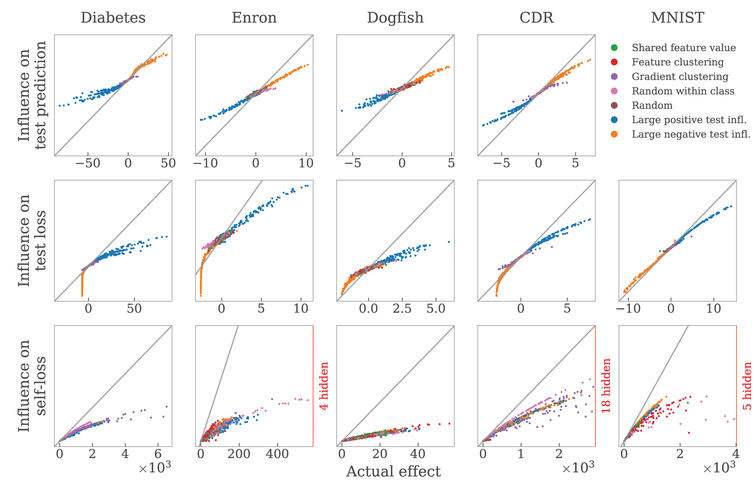
$$\textbf{Actual effect} \quad \mathcal{I}_f^*(w) = f\big(\hat{\theta}(\mathbf{1})\big) - f\big(\hat{\theta}(\mathbf{1}-w)\big)$$

$$\textbf{Influence} \quad \mathcal{I}_f(w) \triangleq \frac{d}{dt} f\big(\hat{\theta}(\mathbf{1}-tw)\big)\Big|_{t=0} = \nabla_\theta f\big(\hat{\theta}(\mathbf{1})\big)^\top H_{\lambda,\mathbf{1}}^{-1} g_{\mathbf{1}}(w)$$

Our work. The effect of removing **large groups** of training points is surprisingly additive and well-approximated by influence functions.

EXPERIMENTS

Across 5 datasets, the influence of removing a **group** of training data was highly **correlated** with its actual effect, even when these groups were **large** (up to 25% of the dataset) and coherent (e.g. clusters). Additionally, the influence tends to **underestimate** the actual effect.



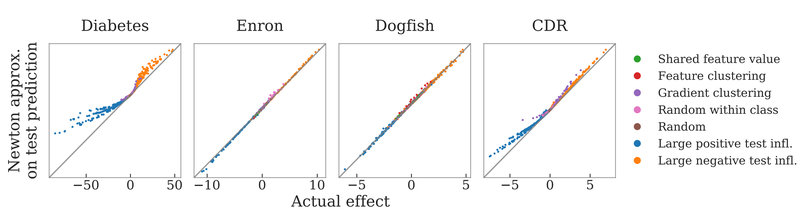
Influences vs. actual effects of coherent groups ranging from 0.25% to 25% in size. Each point corresponds to a group, and its color reflects how that group was constructed.

ANALYSIS

We analyse the accuracy of influence functions for large changes using the one-step Newton approximation as an intermediate. This approximation is surprisingly accurate, but slower to compute.

$$\textbf{Estimated param. change} \quad \Delta\theta_{\text{Nt}} = \big(H_{\lambda,\mathbf{1}}(\mathbf{1}-w)\big)^{-1} g_{\mathbf{1}}(w)$$

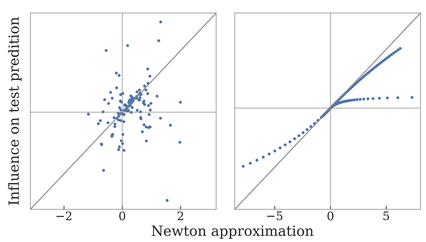
$$\textbf{Newton approx.} \quad \mathcal{I}_f^{\text{Nt}}(w) \triangleq f\big(\hat{\theta}(\mathbf{1}) + \Delta\theta_{\text{Nt}}\big) - f\big(\hat{\theta}(\mathbf{1})\big)$$



In some restricted settings (self-loss and removing copies of a single training point), we prove an approximate cone constraint that captures both the **correlation** and **underestimation** observed empirically.

$$\mathcal{I}_f(w) \lessapprox \mathcal{I}_f^{\text{Nt}}(w) \lessapprox (1 + C_{\max})\mathcal{I}_f(w)$$

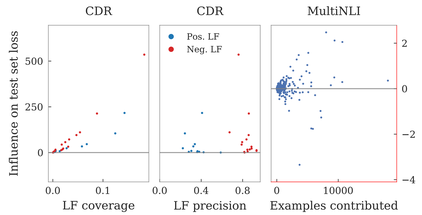
This does not apply in all settings. For the influence on the test prediction, we show that correlation and underestimation need not hold. We synthesize datasets with subsets of training points that violate the cone constraint (right).



APPLICATIONS

We used influence functions to study group effects in two case studies where groups occur naturally.

Data programming. In a dataset of chemical-disease relations[3], we associated each labeling function (LF) with the points it labeled. The influence of an LF on the test loss was associated with its coverage but not its precision, suggesting that practitioners should focus on constructing LFs with high coverage.



Credit attribution. In the MultiNLI dataset[4], we associated each crowdworker with the set of examples they created. The number of these examples was uncorrelated with their influence, suggesting that practitioners should focus on eliciting high-quality examples over increasing quantity.

DISCUSSION

We showed empirically that the influences of groups of points are highly correlated with, and consistently underestimate, their actual effects. Our analysis shows that this need not arise in general, suggesting that the striking correlation that we observe is driven by specific properties of real datasets.

REFERENCES

[1] F. R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

[2] R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A Swiss Army infinitesimal jackknife. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1139–1147, 019.

[3] C. Wei, Y. Peng, R. Leaman, A. P. Davis, C. J. Mattingly, J. Li, T. C. Wiegers, and Z. Lu. Overview of the BioCreative V chemical disease relation (cdr) task. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 154–166, 2015.

[4] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Association for Computational Linguistics (ACL)*, pages 1112–1122, 2018.