RESCALED INFLUENCE FUNCTIONS: ACCURATE DATA ATTRIBUTION IN HIGH DIMENSION

Ittai Rubinstein
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
ittair@mit.edu

Samuel B. Hopkins
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
samhop@mit.edu

ABSTRACT

How does the training data affect a model's behavior? This is the question we seek to answer with *data attribution*. The leading practical approaches to data attribution are based on *influence functions* (IF). IFs utilize a first-order Taylor approximation to efficiently predict the effect of removing a set of samples from the training set without retraining the model, and are used in a wide variety of machine learning applications. However, especially in the high-dimensional regime (# params $\geq \Omega($# samples$)$), they are often imprecise and tend to underestimate the effect of sample removals, even for simple models such as logistic regression. We present *rescaled influence functions* (RIF) – a tool for data attribution which can be used as a drop-in replacement for influence functions, with little computational overhead but significant improvement in accuracy. We compare IF and RIF on a range of real-world datasets, showing that RIFs offer significantly better predictions in practice, and present a theoretical analysis explaining this improvement. Finally, we present a simple class of data poisoning attacks that would fool IF-based detections but would be detected by RIF.

1 INTRODUCTION

Data attribution aims to explain the behavior of a machine learning model in terms of its training data. If θ is a model trained on a dataset $\{(x_i, y_i)\}_{i\in[n]}$, the fundamental algorithmic task in data attribution is to answer the question:

Leave-T-Out Effect: *How would θ have been different if some subset $T \subseteq [n]$ of the training set had been missing?*

The ability to quickly and accurately predict a leave-T-out (LTO) effect, or to search for subsets producing a large leave-out effect, unlocks extensive capabilities from classical statistical inference to modern machine learning. For example, the jackknife, leave-T-out cross-validation, and bootstrap are all widely used to quantify uncertainty and estimate generalization error or confidence intervals, and all rely on the ability to quickly estimate LTO effects Efron (1992); Giordano et al. (2019b); Jaeckel (1972). Machine learning has seen an explosion of applications of data attribution, for dataset curation Koh & Liang (2017); Koh et al. (2019), explainability Koh et al. (2019); Grosse et al. (2023), crafting and detection of data poisoning attacks Engstrom et al. (2025); Koh et al. (2022); Schulam & Saria (2019), machine unlearning Sekhari et al. (2021); Guo et al. (2019); Izzo et al. (2021), credit attribution Jia et al. (2019); Ghorbani & Zou (2019), bias detection Brunet et al. (2019), and more.

Ascertaining the ground truth leave-T-out effect in general requires a full retrain of a model for each T of interest, which is computationally intractable in all but the simplest settings. Consequently, approximations to the leave-T-out effect are widely used. Key desiderata for such approximations are (1) *accuracy*, (2) *computational efficiency* even for large-scale models, and (3) *additivity*: the predicted effect of removing T should be the sum of predicted effects of removing each element of T individually. Additivity enables another important capability: *search* for the subset T of a given size with the greatest predicted effect according to a given metric, by taking the k training data points

with largest predicted leave-one-out (LOO) effects Broderick et al. (2020); Ilyas et al. (2022); Huang et al. (2024).

Influence functions (IF) Hampel (1974) are by far the most widely used and studied data attribution method. The IF is a first-order approximation to the change in model parameters when infinitesimally down-weighting an individual sample. IF approximations are well studied in classical, under-parameterized settings, where they are typically accurate and enjoy solid theoretical foundations Giordano et al. (2019b). But, despite widespread adoption for data attribution in high-dimensional/overparameterized models, IF's accuracy in the high-dimensional setting is comparatively poor. Empirical studies show that IFs often underestimate the true magnitude of parameter changes, leading to potentially misleading conclusions about data importance or model robustness Basu et al. (2021); Koh & Liang (2017). And, existing theoretical analyses justifying IF approximations break down for overparametrized models. But, thus far, more accurate alternatives to IFs have proved too computationally expensive to be practical.

We study a simple and fast-to-compute modification of the influence function, which we term the *rescaled influence function* (RIF). RIFs improve accuracy by incorporating a limited amount of higher-order information about the change in model parameters from sample removal, but retain the additivity and in many settings also the computational efficiency of IFs. We show via experiments and theoretical analysis that RIFs are accurate for data attribution in overparameterized models where IFs struggle. Like IFs, RIFs are model and task agnostic, meaning that they can be applied to any empirical risk minimization-based training method with smooth losses, and they can estimate the leave-T-out effect according to any (smooth) measure of change to model parameters. We therefore advocate using RIFs as a drop-in replacement for IFs across data attribution applications.

Organization In Section 1.1, we introduce RIFs formally. Section 2 presents our experimental results, and Section 3 presents our theoretical analysis of RIF. We discuss context and conclusions in Sections 4 and 5

1.1 Influence Functions, Newton Steps, and Rescaled Influence Functions

We now introduce the rescaled influence function formally. Suppose that $\{(x_i, y_i)\}_{i\in[n]}$ is a training data set, $\Theta \subseteq \mathbb{R}^d$ is a class of models, and $\ell(x, y, \theta)$ is a twice-differentiable loss function; ℓ may include a regularizer. For simplicity, we imagine that ℓ is convex, although the definition of RIFs can be extended to the non-convex case. Let $\hat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i\le n} \ell(x_i, y_i, \theta)$ be the empirical loss minimizer (or, in the non-convex setting, any local minimum of the empirical loss).

Influence Functions The influence function $\text{IF}_i \in \mathbb{R}^d$ associated to the i-th training sample is a first-order estimate of the effect of dropping that sample.[1] Introducing a weight $w_i \in [0, 1]$ associated to each sample i and allowing $\hat{\theta}$ to depend on w via $\hat{\theta}(w) = \arg\min_{\theta\in\Theta} \sum_{i\le n} w_i \cdot \ell(x_i, y_i, \theta)$,

$$\text{IF}_i = -\left[\frac{d}{dw_i} \cdot \hat{\theta}(w)\right] \Big|_{w=\mathbf{1}} = H^{-1} \cdot \nabla\ell(x_i, y_i, \hat{\theta}) \,.$$

Here, H is the Hessian of $\sum_{i\le n} \ell(x_i, y_i, \theta)$ evaluated at $\hat{\theta}$ (see e.g., Rousseeuw et al. (1986) for a derivation). For $T \subseteq [n]$, the IF estimate of the leave-T-out model is

$$\hat{\theta}_{\text{IF},T} = \hat{\theta} + \sum_{i\in T} \text{IF}_i \,.$$

We can obtain all the single-sample IF estimates IF_i at the cost of a single Hessian inversion and n gradient computations, which then suffice to obtain $\hat{\theta}_{\text{IF},T}$ for any T via additivity.

Newton Steps IFs are additive and efficiently computable, but their accuracy suffers when n and d are comparable, or, worse still, if d significantly exceeds n as in the overparameterized setting (Koh et al. (2019); see also Section 2). A much more accurate approximation to the leave-T-out effect

[1] Some treatments replace dropping with up-weighting, with a resulting difference of sign compared to our convention.

is given by taking a single Newton step (NS) to optimize the leave-T-out loss $\sum_{i \notin T} \ell(x_i, y_i, \theta)$, starting from $\hat{\theta}$. The NS approximation to the leave-T-out effect is given by

$$\hat{\theta}_{\text{NS},T} = \hat{\theta} - H_{[n]\setminus T}^{-1} \left(\sum_{i \notin T} \nabla \ell(x_i, y_i, \hat{\theta}) \right) = \hat{\theta} + H_{[n]\setminus T}^{-1} \left(\sum_{i \in T} \nabla \ell(x_i, y_i, \hat{\theta}) \right) .$$

Here, $H_{[n]\setminus T}$ is the Hessian of the leave-T-out loss, evaluated at $\hat{\theta}$, and the second equality follows from the fact that θ is a local optimum of ℓ.

As early as 1981, Pregibon Pregibon (1981) observes in the context of leave-one-out estimation for logistic regression that the Newton step approximation is remarkably accurate. At a high level this is because, unlike the IF approximation, the NS approximation takes into account the change to the Hessian from removing the samples in T. For convex losses, the true leave-T-out effect can often be obtained by Newton iteration – taking multiple Newton steps initialized with $\hat{\theta}$. The only differences we expect to see between the one-step NS approximation and the result of Newton iteration would arise because the Hessian may change from its value at $\hat{\theta}$. Thus, for problems with Lipschitz Hessians, we expect NS to be a very accurate approximation to the true leave-T-out effect; Koh et al. (2019) offers experimental validation of this idea for leave-k-out estimation in logistic regression, and some formal justification.

Rescaled Influence Functions The accuracy of the NS approximation comes at significant cost, since each fresh T requires a Hessian inversion, and additivity is lost. The RIF recovers additivity and much of the computational efficiency of IF, but retains much of the accuracy of the NS approximation. For sample $i \in [n]$, let RIF_i be the NS approximation to the leave-i-out effect, given by $\text{RIF}_i = H_{[n]\setminus\{i\}}^{-1} \cdot \nabla \ell_i(x_i, y_i, \hat{\theta})$. Then for $T \subseteq [n]$, we define the RIF approximation to the leave-T-out effect to be

$$\hat{\theta}_{\text{RIF},T} = \hat{\theta} + \sum_{i \in T} \text{RIF}_i \, .$$

RIF is additive by definition.

The computational overhead of RIF compared to IF depends in general on the cost of computing the n leave-one-out Hessian inversions – once these are obtained, no fresh Hessian inversion is needed to compute $\hat{\theta}_{\text{RIF},T}$ for any T. RIF is especially attractive in generalized linear models and neural networks with a ReLU activation function, where RIF_i can be obtained from IF_i by multiplying by a rescale factor $(1 - h_i)^{-1}$, where h_i is a (generalized) leverage score associated to the i-th sample, which can be computed via a single matrix-vector product with H^{-1}. Thus, for generalized linear models, no additional Hessian inversion is needed. For example, in logistic regression, the formula for RIF_i uses the rescaling $(1 - h_i)^{-1}$, where $h_i = \hat{y}_i(1 - \hat{y}_i) \cdot x_i^\top H^{-1} x_i$; here $\hat{y}_i \in [0, 1]$ is the logistic predicted label of the i-th sample according to $\hat{\theta}$.

Beyond generalized linear models and ReLU neural networks, whenever each sample makes a low-rank contribution the Hessian, the n leave-one-out Hessian inversions can be computed quickly via the Sherman-Morrison/Woodbury formula. In all of our experiments, the running time overhead to compute RIF is negligible (see Table 2).

In underparameterized settings, it is reasonable to expect that removing a single sample has a negligible effect on the Hessian, and so $\text{IF}_i \approx \text{RIF}_i$. But for high-dimensional or overparameterized models, a single sample removal can have a significant effect on the Hessian. Our experiments and theory demonstrate the significant accuracy improvement of RIF compared to IF in high-dimensional and overparameterized models.

We note that the idea of summing over estimates of leave-one-out effects to estimate the leave-T-out effect is not new, and has been a central component of many previous data models Ilyas et al. (2022). In their seminal TRAK paper, Park et al. separately consider both the idea of combining LOO effects additively Park et al. (2023a)[Definition 2.3] and the idea of using a Newton step to estimate LOO effects of a logistic regression Park et al. (2023a)[Definition 3.1] but do not explicitly combine the rescaling effect in their estimator except to note that the rescaling correction has little to no effect in their setting.

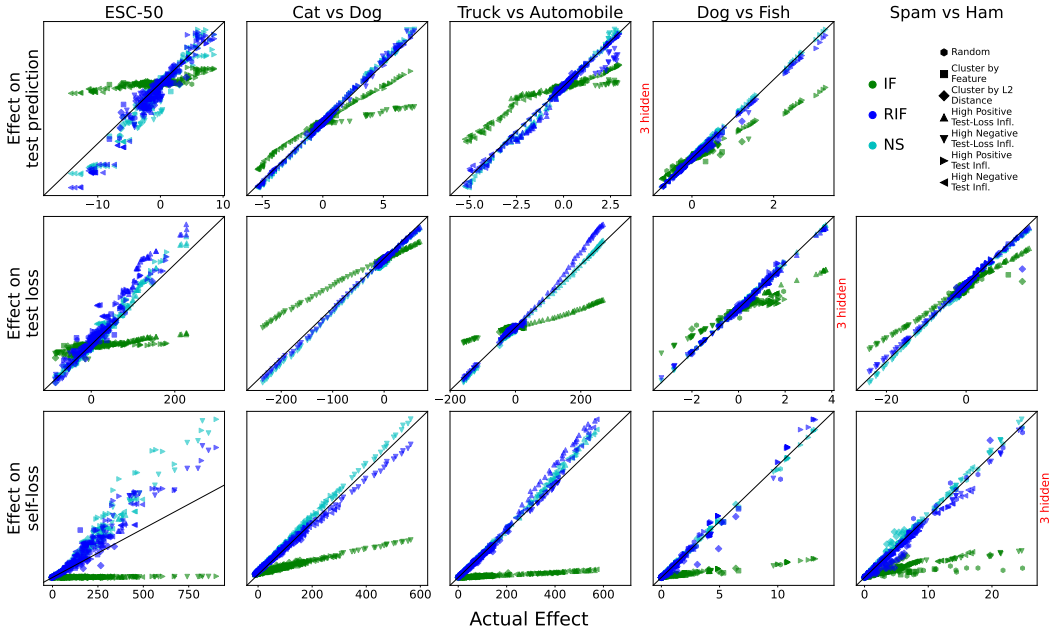

Figure 1: Accuracy of IF versus RIF compared across datasets from image classification (DogFish, Cat vs Dog, Truck vs Automobile), natural language (Spam vs Ham), and audio (ESC-50). In each dataset, we study a binary classification task solved via logistic regression with frozen-embedding features. Each point represents a single choice of subset T. The horizontal axis represents ground truth leave-T-out effect as measured by changes to test predictions, test losses, and self-loss, computed via refitting the logistic model. The vertical axis represents the prediction of this effect made by IF/RIF/NS. A perfectly accurate prediction falls along the black diagonal line. In essentially every case, the RIF prediction falls nicely along this "ground truth" line, agreeing with the NS prediction, while IF typically underestimates the leave-T-out effect.

A similar approach that has been the focus of recent research is the Additive-One-Exact data model, which estimates the LTO effect by summing over the *exact* LOO effects. This data model was introduced by Kuschnig et al. Kuschnig et al. (2021) and further analyzed by Hu et al. Hu et al. (2024) and Huang et al. Huang et al. (2024). Both Kuschnig et al. and Huang et al. study the accuracy of this method for identifying sets of highly influential samples in ordinary least squares (OLS) regressions. Moreover, Huang et al. also note that because a single Newton step is equivalent to a full retrain for the case of OLS, a natural extension of the Additive-One-Exact data model is to sum over the single-Newton step attributions of the individual samples Huang et al. (2024)[Appendix C.2]. But, to the best of our knowledge, no prior work offers quantitative experimental or theoretical comparisons between RIF and other data attribution methods in the high-dimensional settings where the differences we study emerge, or beyond the case of OLS where it is equivalent to the Additive-One-Exact model. [2]

2 Empirical Results

We now present empirical findings on the accuracy of RIF estimates for leave-T-out effects[3]. Our experimental setup is inspired by the seminal work of Koh & Liang (2017); Koh et al. (2019), who evaluate the accuracy of influence function estimates using logistic regression as a testbed.

We compare IF, NS, and RIF estimates across the first five datasets in Table 1, spanning vision, NLP, and audio classification tasks. Each dataset is processed using a domain-specific embedding, and

[2]We are grateful to Tamara Broderick and Jenny Huang for making us aware of these prior works via personal communication.

[3]An implementation of our experimental design is available at https://github.com/ittai-rubinstein/rescaled-influence-functions.

we train a logistic regression model to solve a binary classification task on the embedded data. We compare the actual vs predicted effect of removing a given set of samples T from the training set, while varying:

- **Sample-removal strategy:** Following Koh et al. (2019), we evaluate both random subsets and more structured sets of training points, selected using heuristics such as clustering by a random feature or by Euclidean distance in feature space.
- **Accuracy metric:** As in Koh et al. (2019), we assess accuracy by comparing predicted and actual changes in three scalar quantities when a set T is removed: (1) the total predicted probability for a target class over a subset of test samples, (2) the total test loss on this subset, and (3) the loss on the training set including the removed samples ("self-loss"). The test subset is selected to include a balanced mix of high-loss and randomly chosen test points.
- **Size of removed subset:** We consider values of $|T|$ ranging from 0.1% to 5% of the training set.

We illustrate our main findings in Figure 1. Across every dataset, fraction of sample removals, and accuracy metric, we find that RIF significantly outperforms IF. For more details on our experimental setup, see the supplemental material.

Table 1: Summary of datasets used in our experiments. Each dataset involves a binary classification task which we solve using a regularized logistic regression with mild L_2 regularization. We include both datasets used in the Koh et al. (2019) benchmark (DogFish and Enron), as well as several new datasets spanning a wide range of domains, including vision, natural language processing and audio. For more details about these datasets, see supplementary material.

Name	d	n	Test Accuracy	Description
ESC-50	512	1600	83.0%	ESC-50 dataset embedded using OpenL3; "artificial" vs "natural" classification Piczak (2015); Cramer et al. (2019)
CatDog	2048	9600	80.9%	ResNet-50 embeddings of CIFAR-10 cat and dog classes Krizhevsky (2009); TorchVision Contributors (2016)
AutoTruck	2048	9600	92.7%	ResNet-50 embeddings of CIFAR-10 truck and automobile classes Krizhevsky (2009); TorchVision Contributors (2016)
DogFish	2048	1800	98.3%	Inception v3 embeddings of dog and fish images from ImageNet Szegedy et al. (2016); Russakovsky et al. (2015)
Enron	3294	4137	96.1%	Bag-of-words embeddings of the standard spam vs ham dataset Koh et al. (2019); Metsis et al. (2006)
IMDB	512	40000	87.7%	BERT embeddings of the IMDB sentiment dataset Maas et al. (2011); Devlin et al. (2019)

Tradeoff: Dimension and Regularization As the number of samples n decreases compared to the model dimension d, we expect the higher-order effect captured by RIF to be stronger. Figure 2 shows this tradeoff, comparing the IF and RIF accuracy while varying the ratio of n and d by sub-sampling a fixed dataset. A similar tradeoff appears when we add an L_2 regularization term of $\frac{1}{2}\lambda\|\theta\|^2$ to the loss for different values of $\lambda > 0$. Increasing λ dampens the higher-order effects captured by RIF – in the limit $\lambda \to \infty$ the Hessian does not vary as samples are removed. In Figure 2 we illustrate this tradeoff by varying λ for a fixed dataset (DogFish), observing that IF and RIF agree for large λ but not for small λ.

Detecting Data Poisonings with RIF One common use of additive data attributions such as influence functions is to detect potential outliers contaminating a dataset Koh & Liang (2017); Broderick et al. (2020); Rubinstein & Hopkins (2025); Khaddaj et al. (2023). We conduct a simple experiment to demonstrate the advantages of RIF over IF for this task. We take a binary image

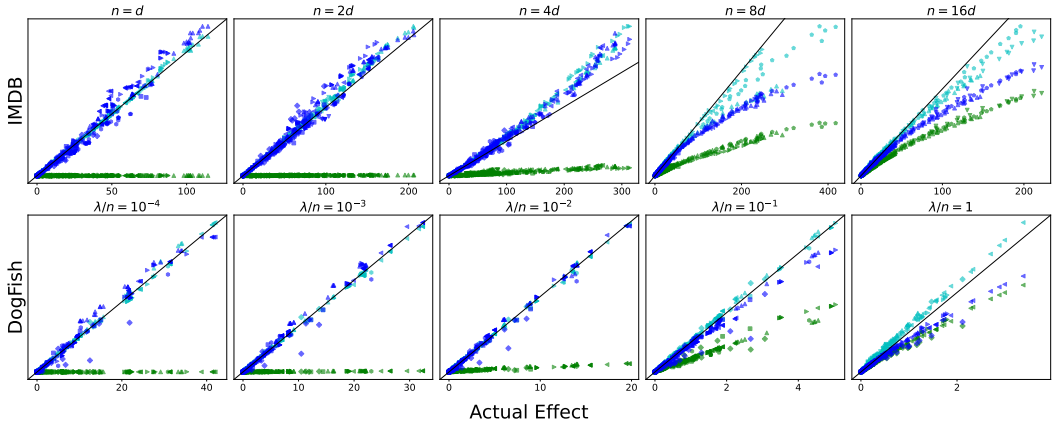

Figure 2: *First row:* accuracy of IF versus RIF compared across differing ratios of n and d, for the IMDB dataset, subsampled randomly to obtain datasets of varying sizes. IF and RIF are similar when $n \gg d$, but as n decreases, RIF remains accurate while IF degrades. *Second row:* A similar comparison for the overparameterized DogFish dataset, where we vary the regularization strength λ. IF becomes accurate only under strong regularization, while RIF remains robust across settings. In all plots, we compare the predicted versus actual values of the self-loss metric. Blue points show the RIF estimate, green points the IF estimate, and cyan points the Newton step. Point shapes indicate different strategies for selecting training samples to remove, as in Figure 1.

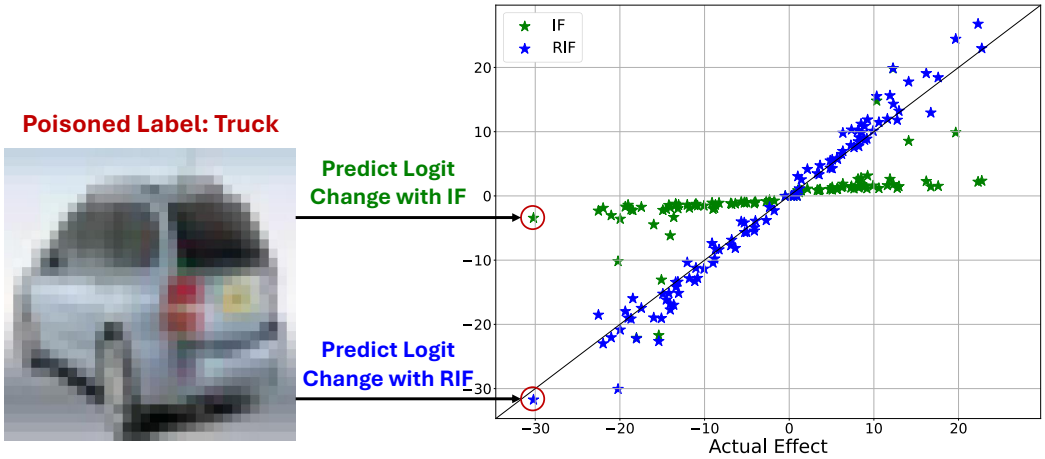

Figure 3: On the right we plot the actual vs predicted effect on a test samples logits from removing a "poisoned" sample from the train set using both IF and RIF. On the left we show the poisoned image corresponding to the leftmost point in the plot – an image of an automobile mislabeled as "Truck". RIF predictions (blue) align much more closely with the actual effects, while IF predictions (green) tend to underestimate these effects.

classification problem (Truck vs Automobile), add an incorrectly-labeled test sample to the training set, and train a logistic regression model on the resulting poisoned dataset. We then compare the accuracy of IF and RIF estimates of the effect that removing the poisoned sample would have on the model's prediction for that test sample. RIF significantly outperforms IF. See Figure 3.

3 Theoretical Results

We turn to a theoretical explanation of the effectiveness of RIF to estimate leave-T-out effects in high dimensions. Prior work Koh et al. (2019) shows that under reasonable assumptions, the NS

approximation provides a very accurate approximation of the true leave-T-out effect; this is also easily visible in the experiments we reproduced above. Importantly, the NS approximation remains accurate even when the IF estimate is poor. Motivated by this, we focus our analysis on the gap between our RIF estimate and the NS estimate. This leads to a comparatively simple theorem statement, avoiding too many assumptions.

Our setting is as follows. We assume that a model is trained via minimization of a convex empirical risk of the form:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})\,.$$

We think of each ℓ_i as a per-sample loss from the i-th sample in an underlying training set, although we do not actually need to assume such a training set underlies the optimization problem. Let $\mathbf{g}_i := \nabla\ell_i(\hat{\boldsymbol{\theta}})$ and $\mathbf{H}_i := \nabla^2\ell_i(\hat{\boldsymbol{\theta}})$ denote the gradient and Hessian of the ith sample at the solution $\hat{\boldsymbol{\theta}}$, and define the total Hessian $\mathbf{H} := \sum_{i=1}^{n}\mathbf{H}_i$.

We make the following set of assumptions on the loss functions. Most of the assumptions are parameterized quantitatively, and our final theorem bounding the quality of the RIF approximation depends on these parameters. Crucially, these assumptions allow for $n \approx d$ (or even $n \ll d$, if regularization is added), so that our main theorem captures how RIF remains accurate for high-dimensional barely-underparameterized or even overparameterized models. We discuss after our main theorem statement how to interpret these assumptions quantitatively.

Assumption 1 (Positive Semidefiniteness/Convexity)**.** *We assume that each $\mathbf{H}_i$ is positive semidefinite, or equivalently, that ℓ_i is convex.*

The next two assumptions are the key quantitative ones. We offer some discussion now and more after we state our main theorem.

Assumption 2 (No Single-Sample Gradient or Hessian Too Large)**.** *For all $i \in \{1,\dots,n\}$, we assume*

$$\left\|\mathbf{H}^{-1/2}\mathbf{g}_i\right\|_2 \le C_\ell \;\text{ and }\; \left\|\mathbf{H}^{-1/2}\mathbf{H}_i\mathbf{H}^{-1/2}\right\|_{\text{op}} \le 1 - \frac{1}{C_R}\,,$$

for some $C_\ell, C_R > 0$. Here $\|\cdot\|_{\text{op}}$ is the operator norm/maximum singular value.

The second clause of Assumption 2 can be rewritten as $\mathbf{H}_i \preceq (1 - C_R^{-1})\sum_{j\neq i}\mathbf{H}_j$. This just captures that no single-sample Hessian $\mathbf{H}_i$ is too much larger in any direction than the sum of all the others. This is the key condition allowing for large dimension d: even if $n \approx d$, this condition can be satisfied (and indeed will be satisfied for, e.g., random low-rank $\mathbf{H}_i$) without taking $C_R = \omega(1)$.

Assumption 3 (Cross-Sample Incoherence)**.** *For some $\varepsilon, \delta > 0$, and for all $i \neq j$, $\left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\mathbf{H}_j^{1/2}\right\|_{\text{op}} \le \delta$ and $\left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\mathbf{g}_j\right\|_2 \le \varepsilon$.*

We expect ε, δ to be small because in high dimensions gradients and Hessians of distinct samples are likely to point in close-to-orthogonal directions. We carry this intuition out in more detail below.

Ultimately, we use IF/RIF/NS to estimate the change to $f(\hat{\boldsymbol{\theta}})$ for some *evaluation function* f. For instance, in our experiments, f is typically test loss or a test prediction. To show that the RIF and NS estimates are close, we require our evaluation function f to have bounded gradients:

Assumption 4 (Evaluation Gradient Projection Control)**.** *Let $\nabla f(\boldsymbol{\theta})$ denote the gradient of an* evaluation function $f\colon \mathbb{R}^d \to \mathbb{R}$. *For all i, $\left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\nabla f(\hat{\boldsymbol{\theta}})\right\|_2 \le \eta$ for some $\eta > 0$.*

Let $\mathbf{w} \in [0,1]^n$ be a weight change vector. We study the NS and RIF approximations to the optimum of the weighted loss $\sum_{i\le n} w_i\ell_i(\boldsymbol{\theta})$. (So, to capture leave-$T$-out, we set $w_i = 0$ for $i \in T$ and otherwise $w_i = 1$.) We define $\hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}$ and $\hat{\boldsymbol{\theta}}_{\text{NS},\mathbf{w}}$ analogously to $\hat{\boldsymbol{\theta}}_{\text{RIF},T}, \hat{\boldsymbol{\theta}}_{\text{NS},T}$, respectively. We are now ready to state our main theorem:

Theorem 3.1 (Accuracy of Rescaled Influence Function)**.** *Under Assumptions 1–4, for any $k = \|\mathbf{w}\|_1 \le \frac{1}{2\delta C_R}$,*

$$|\langle\nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{NS,\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}\rangle| \le k^2\eta\,(1 + 2C_R)\,(\varepsilon + C_R C_\ell\delta)$$

The proof of Theorem 3.1 proceeds via a matrix-perturbation analysis which shows that the Hessian inversion in the NS approximation can itself be approximated well without considering the contributions to the inverse from $\nabla^2\ell_i$'s interaction with $\nabla^2\ell_j$ when $i \neq j$. We defer the proof to supplemental material, and focus instead on interpreting Theorem 3.1, to illustrate how it captures the improvement of RIF compared to IF.

Interpreting Assumptions and Theorem 3.1 Prior works Giordano et al. (2019b); Koh et al. (2019) prove similar-in-spirit results to Theorem 3.1, but concerning IF rather than RIF. A direct comparison of Theorem 3.1 to those results in prior work is challenging, as each result is derived under different assumptions. So, to better understand the practical significance of our bounds compared to those in prior work, and see why they capture the accuracy of RIF for overparameterized models, we analyze their asymptotic behavior in a simplified setting. Since this is for illustration purposes only, we keep the analysis informal.

Consider linear regression with square loss (ordinary least squares), where the data vectors are drawn i.i.d. from a standard Gaussian distribution, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. And suppose $n \geq (1 + \Omega(1))d$, i.e., n and d are comparable. In this case, we know that:

- Each individual Hessian contribution $\mathbf{H}_i = x_i x_i^\top$ is low rank with $\mathrm{rk}(\mathbf{H}_i) = 1$ and $\|\mathbf{H}_i\|_{\mathrm{op}} = O(d)$,
- The total Hessian is approximately isotropic: $\mathbf{H} \approx n\mathbf{I}$,
- Gradient vectors are bounded in norm: $\|\mathbf{g}_i\|_2 \approx \sqrt{d}$.

We can apply the heuristic that random vectors $u, v \in \mathbb{R}^d$ are likely to have $|\langle u, v\rangle| \approx \|u\|\|v\|/\sqrt{d}$, and so long as $n \geq (1 + \Omega(1))d$, we expect the key variables in Theorem 3.1 to scale as:

- $C_\ell := \max_{i\in[n]} \left\|\mathbf{H}^{-1/2}\mathbf{g}_i\right\|_2 \approx \frac{\sqrt{d}}{\sqrt{n}} = O(1)$,
- $C_R := \max_{i\in[n]} \frac{1}{1-\left\|\mathbf{H}^{-1/2}\mathbf{H}_i\mathbf{H}^{-1/2}\right\|_{\mathrm{op}}} \approx \frac{n}{n-d} = O(1)$,
- $\delta := \max_{i\neq j} \left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\mathbf{H}_j^{1/2}\right\|_{\mathrm{op}} = \widetilde{O}\left(\frac{\sqrt{d}}{n}\right)$,
- $\varepsilon := \max_{i\neq j} \left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\mathbf{g}_j\right\|_2 = \widetilde{O}\left(\frac{\sqrt{d}}{n}\right)$,
- $\eta := \max_{i\in[n]} \left\|\mathbf{H}_i^{1/2}\mathbf{H}^{-1}\nabla_{\boldsymbol{\theta}} f\right\|_2 = \max_{i\in[n]} \left|\mathbf{x}_i^\intercal\mathbf{H}^{-1}\nabla_{\boldsymbol{\theta}} f\right| = \widetilde{O}\left(\frac{\|\nabla_{\boldsymbol{\theta}} f\|_2}{n}\right)$.

Under these conditions, Theorem 3.1 guarantees that for any set of at most $k \leq k_{\text{threshold}} = \widetilde{\Omega}\left(\frac{n}{\sqrt{d}}\right)$ removed samples, the discrepancy between the RIF and Newton step estimates is bounded by:

$$|\langle \nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{\text{NS},\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}\rangle| \leq k^2\eta\,(1 + 2C_R)\,(\varepsilon + C_R C_\ell \delta) = \widetilde{O}\left(\frac{k^2\sqrt{d}\,\|\nabla_{\boldsymbol{\theta}} f\|_2}{n^2}\right).$$

The scaling rate n^{-2} in the denominator matches what we expect for influence functions, as established in Giordano et al. (2019b). But influence function approximations incur *significantly worse* dimension dependence in the numerator, meaning that n must be much larger than d (indeed, quadratic in d or even larger) to obtain nontrivial guarantees. For comparison, in supplemental material, we analyze the bounds proved by Giordano et al. (2019b); Koh et al. (2019) for influence functions to the same random-design ordinary-least-squares setting and show that they guarantee influence function accuracy only for much larger n or smaller d. For example, the bounds of Giordano et al. (2019b) are only applicable for $k \leq \widetilde{O}\left(\frac{n}{d^2}\right)$, and yield an error bound that scales as $\widetilde{O}\left(\frac{k^2 d^4 \|\nabla_{\boldsymbol{\theta}} f\|_2}{n^2}\right)$.

Finally, to assess the tightness of our result relative to the RIF magnitude itself, we note that under the same random-design least-squares setup and the same heuristics about inner products of high-dimensional random vectors, the RIF estimate for the removal of the top-k most influential samples scales as

$$\max\left\{|\langle \nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}\rangle| \;:\; \|\mathbf{w}\|_1 = k\right\} = \Omega\left(\frac{k\,\|\nabla_{\boldsymbol{\theta}} f\|_2}{n}\right).$$

Hence, the ratio of the RIF estimate ("signal") to the RIF–NS error ("noise") is

$$\text{SNR} := \frac{\max\left\{|\langle \nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}\rangle| \,:\, \|\mathbf{w}\|_1 = k\right\}}{\max\left\{|\langle \nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{\text{NS},\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}}\rangle| \,:\, \|\mathbf{w}\|_1 = k\right\}} = \widetilde{\Omega}\left(\frac{n}{k\sqrt{d}}\right).$$

This implies that RIF provides a good relative-error approximation to NS even in high dimensions, provided $k \ll \frac{n}{\sqrt{d}}$.

4 Related Work

Influence functions were introduced by Hampel in the context of robust statistics Hampel (1974), and in the context of estimation of standard errors via the *infinitesimal jackknife* by Jaeckel Jaeckel (1972), with a broad ensuing literature in statistics; see e.g., Law (1986); Giordano et al. (2019b). Recent work in econometrics Broderick et al. (2020) uses influence functions to uncover robustness issues in large empirical studies.

The seminal work Koh & Liang (2017) introduced the modern use of influence functions to study the relationship between training data and model behavior in modern machine learning. Ensuing works Bae et al. (2022); Basu et al. (2021); Grosse et al. (2023); Feldman & Zhang (2020) study influence functions for neural networks, and use them as a tool to study and interpret model behavior. Giordano et al. (2019a); Basu et al. (2020) propose second and higher-order approximations to leave-one-out and leave-T-out effects, but these approximations sacrifice linearity and efficiency. Many applications of influence functions have appeared recently, e.g., machine unlearning Guo et al. (2019); Sekhari et al. (2021); Suriyakumar & Wilson (2022), data valuation Jia et al. (2019), robustness quantification Schulam & Saria (2019), and fairness Li & Liu (2022). To scale influence functions up to very large models and datasets, where Hessian inversion becomes infeasible, several works develop sketching/random projection techniques to approximate influence functions, e.g., Wojnowicz et al. (2016); Park et al. (2023b); Schioppa et al. (2022).

Data attribution – tracing model behavior back to subsets of training data – has become a major industry in machine learning; see the recent survey Hammoudeh & Lowd (2024) and extensive citations therein, as well as the NeurIPS 2024 workshop Nguyen et al. (2024) and ICML 2024 tutorial Madry et al. (2024).

Newton-step approximations to the leave-1-out error have been studied since at least 1981 Pregibon (1981). *Cross-validation* is an especially important application Rad & Maleki (2018); Wilson et al. (2020). Additionally, several recent works consider data models that additively combine estimates of leave-one-out effects to compute a leave-T-out effect Kuschnig et al. (2021); Ilyas et al. (2022); Park et al. (2023a); Hu et al. (2024); Huang et al. (2024). However, to the best of our knowledge no previous work provides an empirical or theoretical evaluation of the RIF method beyond low-dimensional least-squares regression.

5 Discussion and Conclusion

IFs and Importance-Ordering: Revisiting the Common Wisdom Common wisdom regarding IF approximations to leave-T-out effects for high-dimensional models holds that the approximations typically *underestimate* the true leave-T-out effect, but that there is a strong correlation between the influence-function approximation to the leave-T-out effects and the true leave-T-out effects, especially measured in terms of the *ordering* of subsets based on their predicted/actual leave-T-out effect. The seminal Koh et al. (2019) even phrases this as an outstanding open question, writing that their work "opens up the intriguing question of why we observe [correlation and underestimation] across a wide range of empirical settings".

Our work sheds significant light on this question. First of all, it explains why we see such correlation in a great many cases – if most samples have a similar "rescale factor" relating IF and RIF (which we would expect to happen for e.g., random data), this induces a linear relationship between RIF and IF estimates. Since RIF is an excellent approximation to the true leave-T-out effect, this explains the correlation between IF and the ground truth, and explains why IF typically underestimates the truth – the rescale factors are always larger than 1.

Koh et al. (2019) also note that this IF/ground-truth correlation phenomenon need not be universal, and indeed we observe several experiments where it does not hold. For instance, in the first row of Figure 1, in the Cat vs Dog dataset, we see a dramatically non-linear and even non-monotone relationship between IF and ground truth, since different subset-selection strategies yield very different relationships between IF and ground truth. Even the ordering of subsets by IF-predicted effect is not accurate in this example, but RIF remains accurate.

Limitations Although much more accurate than IFs, RIFs are still imperfect predictors of ground-truth – see e.g., the ESC-50 dataset in Figure 1 or the rightmost variants of the IMDB dataset in Figure 2. We expect high-dimensional logistic regression to be a good "model organism" for high-dimensional machine learning, so our experiments are limited to that setting. RIF also still requires inverting the Hessian; as discussed in related work for very large-scale models this can be computationally infeasible, and approximate techniques are required. While we show that RIFs are preferable to IFs for detecting certain simple data-poisoning attacks, we do not expect that RIFs are a secure general defense against data poisoning.

Conclusion We show that RIFs are an appealing drop-in replacement for IFs, with little computational overhead in generalized linear models (or whenever individual training samples contribute low-rank terms to the Hessian), but dramatically improved accuracy. Both experiments and theory support this conclusion. Furthermore, the fact that RIFs and IFs differ by a per-sample scaling factor helps to resolve an open question from prior work, showing that the correlation between IF and ground truth leave-T-out occurs when the per-sample scalings all (approximately) agree.

Compute Resources

All experiments were conducted on a server equipped with 64GB RAM, 2 IBM POWER9 CPU cores, and 4 NVIDIA Tesla V100 SXM2 GPUs (each with 32GB memory).

Table 2 details the computational cost of training the base models and computing their IF and RIF data attribution. Another major computational overhead was in retraining the model to obtain ground-truth values for the retrain effect. Despite this, compute resources were not a bottleneck for our work. The total wall-clock time for all experiments reported in the paper was under 100 hours.

Table 2: Comparison of runtime components across datasets. The **Rescaling** step consistently added negligible overhead across all experiments.

Dataset	Training	Hessian	Inversion	Influence	Rescaling
ESC50	1.8 s	0.056 s	0.0005 s	0.051 s	0.0033 s (0.2%)
CatDog	76 s	4.9 s	0.010 s	4.8 s	0.087 s (0.1%)
AutoTruck	48 s	4.9 s	0.0094 s	4.8 s	0.087 s (0.2%)
DogFish	0.43 s	0.92 s	0.0095 s	0.89 s	0.015 s (0.7%)
Enron	6.7 s	15 s	0.065 s	15 s	0.095 s (0.3%)
IMDB (n=16d)	20 s	0.92 s	0.0012 s	0.87 s	0.044 s (0.2%)

Acknowledgments

We used large language models (LLMs) to assist with implementing some of the numerical experiments, writing initial drafts of some of the proofs in this paper, and for LaTeX related formatting (e.g., converting the paper to the DATA-FM format). All LLM-generated content was reviewed and edited by human authors, and the authors take full responsibility for the content and originality of this submission.

We thank Jenny Y. Huang, David R. Burt, Yunyi Shen, Tin D. Nguyen, Vishwak Srinivasan, Tamara Broderick, Yuzheng Hu, and Jiaqi Ma for helpful conversations and correspondence.

This work was supported by NSF Award No. 2238080 and CSAIL Alliances. Ittai Rubinstein was additionally supported by the MIT EECS MathWorks Fellowship.

REFERENCES

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/7234e0c36fdbcb23e7bd56b68838999b-Abstract-Conference.html.

Samyadeep Basu, Xuchen You, and Soheil Feizi. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 715–724. PMLR, 2020. URL http://proceedings.mlr.press/v119/basu20b.html.

Samyadeep Basu, Phillip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=xHKVVHGDOEk.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*, 2020.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 803–811. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brunet19a.html.

Aurora Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen and learn more: Design choices for deep audio embeddings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856, 2019. doi: 10.1109/ICASSP.2019.8683658. URL https://github.com/marl/openl3.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Logan Engstrom, Andrew Ilyas, Benjamin Chen, Axel Feldmann, William Moses, and Aleksander Madry. Optimizing ml training with metagradient descent. *arXiv preprint arXiv:2503.13751*, 2025.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Ryan Giordano, Michael I Jordan, and Tamara Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1139–1147. PMLR, 2019b.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Yuzheng Hu, Pingbang Hu, Han Zhao, and Jiaqi Ma. Most influential subset selection: Challenges, promises, and beyond. *Advances in Neural Information Processing Systems*, 37:119778–119810, 2024.

Jenny Y Huang, David R Burt, Tin D Nguyen, Yunyi Shen, and Tamara Broderick. Approximations to worst-case data dropping: unmasking failure modes, 2024. Version 5, uploaded 30 May 2025.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587. PMLR, 2022. URL https://proceedings.mlr.press/v162/ilyas22a.html.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2008–2016. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/izzo21a.html.

L. Jaeckel. The infinitesimal jackknife, memorandum. Technical Report MM 72-1215-11, Bell Laboratories, Murray Hill, NJ, 1972.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019.

Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In *International Conference on Machine Learning*, pp. 16216–16236. PMLR, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5255–5265, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/a78482ce76496fcf49085f2190e675b4-Abstract.html.

Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, pp. 1–47, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Nikolas Kuschnig, Gregor Zens, and Jesús Crespo Cuaresma. Hidden in plain sight: Influential sets in linear models. Technical report, CESifo Working Paper, 2021.

John Law. Robust statistics—the approach based on influence functions, 1986.

Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International conference on machine learning*, pp. 12917–12930. PMLR, 2022.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, 2011. URL https://ai.stanford.edu/~amaas/data/sentiment/.

Aleksander Madry, Andrew Ilyas, Logan Engstrom, Sung Min (Sam) Park, and Kristian Georgiev. Data attribution at scale. https://icml.cc/virtual/2024/tutorial/35228, 2024. Tutorial presented at the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, July 22, 2024.

Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes? In *CEAS*, volume 17, pp. 28–69. Mountain View, CA, 2006.

Elisa Nguyen, Sadhika Malladi, Andrew Ilyas, Logan Engstrom, Sam Park, and Tolga Bolukbasi. Attributing model behavior at scale (attrib). https://neurips.cc/virtual/2024/workshop/84704, 2024. Workshop at the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), December 14, 2024.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: attributing model behavior at scale. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023a. URL https://proceedings.mlr.press/v202/park23c.html.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023b.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM, 2015. doi: 10.1145/2733373.2806390. URL https://dl.acm.org/doi/10.1145/2733373.2806390.

Daryl Pregibon. Logistic regression diagnostics. *The annals of statistics*, 9(4):705–724, 1981.

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv preprint arXiv:1801.10243*, 2018.

Peter J Rousseeuw, Frank R Hampel, Elvezio M Ronchetti, and Werner A Stahel. Robust statistics: the approach based on influence functions, 1986.

Ittai Rubinstein and Samuel B Hopkins. Robustness auditing for linear regression: To singularity and beyond. In *The Thirteenth International Conference on Learning Representations*, 2025.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8179–8186, 2022.

Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1022–1031. PMLR, 2019.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Vinith Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

TorchVision Contributors. ResNet-50 Pretrained Model. https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html, 2016. Accessed: 2025-05-14.

Ashia Wilson, Maximilian Kasy, and Lester Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International conference on artificial intelligence and statistics*, pp. 4530–4540. PMLR, 2020.

Mike Wojnowicz, Ben Cruz, Xuan Zhao, Brian Wallace, Matt Wolff, Jay Luan, and Caleb Crable. "influence sketching": Finding influential samples in large-scale regressions. In *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3601–3612. IEEE, 2016.

A Proof of Theorem 3.1

Recall our main theoretical result from Section 3:

Theorem A.1 (Theorem 3.1 (restated))**.** *Under Assumptions 1–4, for any* $k \leq \frac{1}{2\delta C_R}$,

$$|\langle \nabla f(\hat{\boldsymbol{\theta}}), \hat{\boldsymbol{\theta}}_{NS,\mathbf{w}} - \hat{\boldsymbol{\theta}}_{\text{RIF},\mathbf{w}} \rangle| \leq k^2 \eta \left(1 + 2C_R\right)\left(\varepsilon + C_R C_\ell \delta\right)$$

Before delving into the proof of Theorem 3.1, we introduce a useful technical lemma:

Lemma A.2. *Let* $\mathbf{A}_1, \ldots, \mathbf{A}_k \in \mathbb{R}^{d\times d}$ *and let* $\mathbf{H} \in \mathbb{R}^{d\times d}$ *be positive semidefinite. Suppose:*

- $\left\|\mathbf{H}^{-1/2}\mathbf{A}_i\mathbf{H}^{-1/2}\right\|_{\text{op}} \leq \sigma$ *for all* i,
- $\left\|\sqrt{\mathbf{A}_i}\mathbf{H}^{-1}\sqrt{\mathbf{A}_j}\right\|_{\text{op}} \leq \delta_{ij}$ *for all* $i \neq j$.

Then,

$$\left\|\sum_{i=1}^{k} \mathbf{H}^{-1/2}\mathbf{A}_i\mathbf{H}^{-1/2}\right\|_{\text{op}} \leq \sigma + \sqrt{\sum_{i\neq j} \delta_{ij}^2}.$$

Proof of Theorem 3.1. We begin by analyzing the difference between the Newton step and the rescaled influence function (RIF) approximation.

Recall that the Newton step is defined as:

$$\text{Newton Step} = (\nabla f)^\top \left(\mathbf{H} - \sum_{j=1}^{n} w_j \mathbf{H}_j\right)^{-1} \sum_{i=1}^{n} w_i \mathbf{g}_i,$$

where each $\mathbf{g}_i \in \mathbb{R}^d$ is the ith gradient component, and $\mathbf{H}_i$ is the ith contribution to the Hessian. Define the weighted Hessian:

$$\mathbf{H}_{\mathbf{w}} := \mathbf{H} - \sum_{j=1}^{n} w_j \mathbf{H}_j.$$

For each $i \in \{1, \ldots, n\}$, define $\mathbf{w}^{(i)} := w \cdot \mathbb{1}_{\{i\}}$ to isolate the i-th coordinate. The RIF estimator is given by:

$$\text{RIF}_i = \sum_{i=1}^{n} (\nabla f)^\top \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} w_i \mathbf{g}_i \,.$$

Our goal is to bound the difference between the Newton step and RIF estimators and we do this by bounding the contribution of each individual sample. That is, for each $i \in [n]$, we will try to bound

$$(\nabla f)^\top \left(\mathbf{H}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}\right) \mathbf{g}_i \,.$$

To do so, we begin by expressing each matrix in terms of $\mathbf{H}$ and its perturbations. Observe:

$$\mathbf{H}_{\mathbf{w}} = \mathbf{H}^{1/2} \left(I - \mathbf{G}_{\mathbf{w}}\right) \mathbf{H}^{1/2}, \quad \text{where } \mathbf{G}_{\mathbf{w}} := \sum_j \mathbf{H}^{-1/2} w_j \mathbf{H}_j \mathbf{H}^{-1/2}.$$

Moreover, we define $\mathbf{R} := \left(I - \mathbf{G}_{\mathbf{w}^{(i)}}\right)^{-1}$, where $\mathbf{G}_{\mathbf{w}^{(i)}} = \mathbf{H}^{-1/2} w_i \mathbf{H}_i \mathbf{H}^{-1/2}$. We have

$$\mathbf{H}_{\mathbf{w}^{(i)}} = \mathbf{H}^{1/2} \left(I - \mathbf{G}_{\mathbf{w}^{(i)}}\right) \mathbf{H}^{1/2}.$$

Using the matrix identity:

$$(A - B)^{-1} = A^{-1} + (A - B)^{-1} B A^{-1},$$

with $A = \mathbf{H}_{\mathbf{w}^{(i)}}$, $B = \mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}$, we obtain:

$$\mathbf{H}_{\mathbf{w}}^{-1} = \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} + \mathbf{H}_{\mathbf{w}}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}.$$

We now expand the correction term on the right-hand side further by applying the same identity again, this time expanding $\mathbf{H}_{\mathbf{w}} = \mathbf{H} - (\mathbf{H} - \mathbf{H}_{\mathbf{w}})$,

$$\mathbf{H}_{\mathbf{w}}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} = \mathbf{H}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} + \mathbf{H}^{-1} \left(\mathbf{H} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1},$$

where the second term reflects higher-order correction contributions due to recursive matrix inversion.

To bound the full error

$$\begin{aligned}(\nabla f)^\top \left(\mathbf{H}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}\right) \mathbf{g}_i = &(\nabla f)^\top \mathbf{H}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i + \\ &+ (\nabla f)^\top \mathbf{H}^{-1} \left(\mathbf{H} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i \,.\end{aligned}$$

It suffices to control the size of each of these terms separately. In other words, we will proceed to bound:

1. The first order correction $(\nabla f)^\top \mathbf{H}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i$,
2. The higher order terms$(\nabla f)^\top \mathbf{H}^{-1} \left(\mathbf{H} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}}^{-1} \left(\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_{\mathbf{w}}\right) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i$.

Bounding the First Order Correction

To bound the first order correction, we use the same formula above to split $\mathbf{H}_{\mathbf{w}^{(i)}}^{-1}$ into a leading term and higher order terms. The goal of this separation is to show that this update to the Hessian does not rotate too much of the weight of $\mathbf{g}_i$ onto the eigenspace of $\mathbf{H}_j$ for any $j \neq i$

We have $\mathbf{H}_{\mathbf{w}^{(i)}}^{-1} = \mathbf{H}^{-1} + \mathbf{H}^{-1} w_i \mathbf{H}_i \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}$.

Therefore, for any $j \neq i$,

$$\left\| \mathbf{H}_j^{1/2} \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i \right\|_2 \leq \underbrace{\left\| \mathbf{H}_j^{1/2} \mathbf{H}^{-1} \mathbf{g}_i \right\|_2}_{\leq \varepsilon} + \underbrace{\left\| w_i \mathbf{H}_j^{1/2} \mathbf{H}^{-1} \mathbf{H}_i \mathbf{H}^{-1/2} \mathbf{R} \mathbf{H}^{-1/2} \mathbf{g}_i \right\|_2}_{\leq |w_i| \delta C_R C_\ell \leq \delta C_R C_\ell} \leq \varepsilon + \delta C_R C_\ell$$

Therefore, this first order correction is at most

$$\sum_{j \neq i} w_j \nabla f^\intercal \mathbf{H}^{-1} \mathbf{H}_j \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i \leq \sum_{j \neq i} w_j \underbrace{\left\| \nabla f^\intercal \mathbf{H}^{-1} \mathbf{H}_j^{1/2} \right\|_2}_{\leq \eta} \underbrace{\left\| \mathbf{H}_j^{1/2} \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i \right\|_2}_{\leq \varepsilon + C_R C_\ell \delta} \leq k\eta \left(\varepsilon + C_R C_\ell \delta\right)$$

Bounding the Higher Order Corrections

We next bound the second (higher-order) term using the Cauchy-Schwarz inequality.

$$\left|(\nabla f)^\top \mathbf{H}^{-1} (\mathbf{H} - \mathbf{H}_\mathbf{w}) \mathbf{H}_\mathbf{w}^{-1} (\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_\mathbf{w}) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i\right| \leq$$
$$\leq \left\|(\nabla f)^\top \mathbf{H}^{-1} (\mathbf{H} - \mathbf{H}_\mathbf{w}) \mathbf{H}^{-1/2}\right\|_2 \times \left\|(\mathbf{I} - \mathbf{G}_\mathbf{w})^{-1}\right\|_{\text{op}} \times \left\|\mathbf{H}^{-1/2} (\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_\mathbf{w}) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i\right\|_2$$

We will bound each of these terms independently.

The right-most multiplicand is bounded using the analysis of the first order term

$$\left\|\mathbf{H}^{-1/2} (\mathbf{H}_{\mathbf{w}^{(i)}} - \mathbf{H}_\mathbf{w}) \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i\right\|_2 \leq \sum_{j \neq i} \left\|\mathbf{H}^{-1/2} w_j \mathbf{H}_j^{1/2} \mathbf{H}_j^{1/2} \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i\right\|_2 \leq$$
$$\leq \sum_{j \neq i} \left\|w_j \mathbf{H}_j^{1/2} \mathbf{H}_{\mathbf{w}^{(i)}}^{-1} \mathbf{g}_i\right\|_2 \leq k (\varepsilon + C_R C_\ell \delta)$$

From the triangle inequality,

$$\left\|\sum_{j \neq i} \nabla f^\top \mathbf{H}^{-1} \mathbf{H}_j \mathbf{H}^{-1/2}\right\| \leq \sum_{j \neq i} |w_j| \cdot \left\|\nabla f^\top \mathbf{H}^{-1} \mathbf{H}_j^{1/2}\right\| \cdot \left\|\mathbf{H}_j^{1/2} \mathbf{H}^{-1/2}\right\|_{\text{op}}.$$

Using the assumption $\left\|\mathbf{H}^{-1/2} \mathbf{H}_j \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq 1$, it follows that

$$\left\|\mathbf{H}_j^{1/2} \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq 1,$$

and from Assumption 4, we also have

$$\left\|\nabla f^\top \mathbf{H}^{-1} \mathbf{H}_j^{1/2}\right\| \leq \eta.$$

Therefore,

$$\left\|\sum_{j \neq i} \nabla f^\top \mathbf{H}^{-1} \mathbf{H}_j \mathbf{H}^{-1/2}\right\| \leq \eta \sum_{j \neq i} |w_j| \leq \eta \|w\|_1 = \eta k.$$

Next, define $\mathbf{A}_j = w_j \mathbf{H}^{-1/2} \mathbf{H}_j \mathbf{H}^{-1/2}$. Then for all j,

$$\left\|\mathbf{H}^{-1/2} \mathbf{A}_j \mathbf{H}^{-1/2}\right\|_{\text{op}} = |w_j| \cdot \left\|\mathbf{H}^{-1/2} \mathbf{H}_j \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq 1 - \frac{1}{C_R},$$

since $\|w\|_\infty \leq 1$ and by Assumption 2 $\left\|\mathbf{H}^{-1/2} \mathbf{H}_j \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq 1 - \frac{1}{C_R}$.

Moreover, for all $i \neq j$, we have

$$\left\|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_j}\right\|_{\text{op}} \leq \sqrt{|w_i|} \cdot \sqrt{|w_j|} \cdot \delta_{ij}.$$

So,

$$\sum_{i \neq j} \left\|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_j}\right\|_{\text{op}}^2 \leq \sum_{i \neq j} |w_i| |w_j| \delta_{ij}^2 \leq (\|w\|_1)^2 \cdot \delta^2 = k^2 \delta^2.$$

Applying Lemma A.2 to the collection $\{\mathbf{A}_j\}$, we conclude that

$$\|\mathbf{G}_\mathbf{w}\|_{\text{op}} \leq 1 - \frac{1}{C_R} + k\delta.$$

For any $k < \frac{1}{2\delta C_R}$, it follows that $I - \mathbf{G}_\mathbf{w}$ is PSD and $\|\mathbf{G}_\mathbf{w}\|_{\text{op}} < 1$, so we have

$$\left\|(I - \mathbf{G}_\mathbf{w})^{-1}\right\|_{\text{op}} \leq \frac{1}{\frac{1}{C_R} - k\delta} \leq 2C_R\,.$$

Summary:

So far, we have show that for all $i \in [n]$,

$$\left|(\nabla f)^\top \left(\mathbf{H}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}\right) \mathbf{g}_i\right| \leq \eta k \left(\varepsilon + C_R C_\ell \delta\right) + \eta k \times 2C_R \times \left(\varepsilon + C_R C_\ell \delta\right) .$$

Therefore,

$$|\text{Newton Step} - \text{RIF}| = \left|\sum_{i=1}^{n} w_i \left(\nabla f\right)^\top \left(\mathbf{H}_{\mathbf{w}}^{-1} - \mathbf{H}_{\mathbf{w}^{(i)}}^{-1}\right) \mathbf{g}_i\right| \leq k^2 \eta \left(1 + 2C_R\right) \left(\varepsilon + C_R C_\ell \delta\right)$$

□

Proof of Lemma A.2. We define the linear operator $C : \mathbb{R}^{k\times k\times d\times d} \to \mathbb{R}^{d\times d}$ to be

$$C(\mathbf{M}) := \sum_{i,j} \mathbf{H}^{-1/2} \sqrt{\mathbf{A}_i}\, \mathbf{M}_{ij}\, \sqrt{\mathbf{A}_j} \mathbf{H}^{-1/2},$$

where $\mathbf{M} \in \mathbb{R}^{k\times k\times d\times d}$ is a rank-4 tensor with $\mathbf{M}_{ij} \in \mathbb{R}^{d\times d}$.

For tensors $\mathbf{M}$, $\mathbf{N}$, define their contraction:

$$C(\mathbf{M})C(\mathbf{N}) = C(\mathbf{L}), \quad \text{where } \mathbf{L}_{ij} = \sum_{q,r} \mathbf{M}_{iq} \cdot \sqrt{\mathbf{A}_q} \mathbf{H}^{-1} \sqrt{\mathbf{A}_r} \cdot \mathbf{N}_{rj}.$$

Define $\Sigma : \mathbb{R}^{k\times k\times d\times d} \to \mathbb{R}^{k\times k}$ as $\Sigma(\mathbf{M})_{ij} := \|\mathbf{M}_{ij}\|_{\text{op}}$, and define $\mathbf{\Delta} \in \mathbb{R}^{k\times k}$ with entries $\mathbf{\Delta}_{ij} = \left\|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_j}\right\|_{\text{op}}$. Then by the triangle inequality and submultiplicativity of the operator norm, we have the point-wise inequality

$$\Sigma(\mathbf{L}) \leq \Sigma(\mathbf{M}) \cdot \mathbf{\Delta} \cdot \Sigma(\mathbf{N}).$$

Applying this iteratively for a sequence $\mathbf{M}_1, \ldots, \mathbf{M}_\ell$, we obtain:

$$\Sigma(\mathbf{N}) \leq \Sigma(\mathbf{M}_1) \cdot \mathbf{\Delta} \cdot \Sigma(\mathbf{M}_2) \cdot \mathbf{\Delta} \cdots \mathbf{\Delta} \cdot \Sigma(\mathbf{M}_\ell).$$

Now consider the identity tensor $\mathbf{M}$ with $\mathbf{M}_{ii} = I_d$ and $\mathbf{M}_{ij} = 0$ for $i \neq j$. Then:

$$C(\mathbf{M}) = \sum_i \mathbf{H}^{-1/2} \sqrt{\mathbf{A}_i} I_d \sqrt{\mathbf{A}_i} \mathbf{H}^{-1/2} = \sum_i \mathbf{H}^{-1/2} \mathbf{A}_i \mathbf{H}^{-1/2}.$$

Let $C := C(\mathbf{M})$. Then:

$$C^\ell = C(\mathbf{M})^\ell = C(\mathbf{N}), \quad \text{with } \Sigma(\mathbf{N}) \leq \mathbf{\Delta}^\ell.$$

By triangle inequality and bounding each tensor entry:

$$\left\|C^\ell\right\|_{\text{op}} \leq k^2 d^2 \cdot \max_i \left\|\mathbf{H}^{-1/2} \mathbf{A}_i^{1/2}\right\|_{\text{op}}^2 \cdot \left\|\mathbf{\Delta}^\ell\right\|_{\text{op}} \leq k^2 d^2 \sigma \cdot \|\mathbf{\Delta}\|_{\text{op}}^\ell .$$

Taking ℓ-th roots:

$$\|C\|_{\text{op}} \leq (k^2 d^2 \sigma)^{1/\ell} \cdot \|\mathbf{\Delta}\|_{\text{op}} .$$

Letting $\ell \to \infty$, the prefactor tends to 1, giving:

$$\left\|\sum_i \mathbf{H}^{-1/2} \mathbf{A}_i \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq \|\mathbf{\Delta}\|_{\text{op}} .$$

Now bound $\|\mathbf{\Delta}\|_{\text{op}}$. Each diagonal entry $\mathbf{\Delta}_{ii} = \left\|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_i}\right\|_{\text{op}} = \left\|\mathbf{H}^{-1/2} \mathbf{A}_i \mathbf{H}^{-1/2}\right\|_{\text{op}} \leq \sigma$. Thus,

$$\mathbf{\Delta} = D + R, \quad \text{with } D = \text{diag}(\left\|\mathbf{H}^{-1/2} \mathbf{A}_1 \mathbf{H}^{-1/2}\right\|_{\text{op}}, \ldots), \quad \|D\|_{\text{op}} \leq \sigma.$$

Then:

$$\|\mathbf{\Delta}\|_{\mathrm{op}} \le \|D\|_{\mathrm{op}} + \|R\|_{\mathrm{op}} \le \sigma + \|R\|_{\mathrm{F}},$$

where R is the off-diagonal part of $\mathbf{\Delta}$ and $\|R\|_{\mathrm{F}}^2 = \sum_{i \neq j} \delta_{ij}^2$.

Hence:

$$\left\| \sum_{i=1}^{k} \mathbf{H}^{-1/2} \mathbf{A}_i \mathbf{H}^{-1/2} \right\|_{\mathrm{op}} \le \sigma + \sqrt{\sum_{i \neq j} \delta_{ij}^2}.$$

□

B Asymptotic Analyses of the Bounds of Koh et al. (2019) and Giordano et al. (2019b)

B.1 Analysis of Koh et al. (2019)

Koh et al. Koh et al. (2019) present two main theoretical results. The first bounds the difference between a single Newton step and a full retrain, and the second bounds the difference between the Newton step and the influence function estimate. We focus on the latter, since that is more directly comparable to the guarantees of Theorem 3.1. To facilitate a direct comparison, we restate their Proposition 2 with all assumptions made explicit below.

Proposition B.1 (Proposition 2 of Koh et al. (2019), rephrased)**.** *Assume the evaluation function $f(\theta)$ is C_f-Lipschitz, the Hessian $\nabla_\theta^2 \ell(x, y, \theta)$ is C_H-Lipschitz, and the third derivative of $f(\theta)$ exists and is bounded in norm by $C_{f,3}$. Let $\sigma_{\min}$ and $\sigma_{\max}$ be the smallest and largest eigenvalues of H_1, respectively, and define*

$$C_\ell \triangleq \max_{1 \le i \le n} \left\| \nabla_\theta \ell(x_i, y_i; \hat{\theta}(1)) \right\|_2.$$

Then the Newton-influence error $\mathrm{Err}_{\text{Nt-inf}}(w)$ *is*

$$\mathrm{Err}_{\text{Nt-inf}}(w) = \nabla_\theta f(\hat{\theta}(1))^\top \mathbf{H}_{\lambda,1}^{-1/2} \mathbf{D}(w) \mathbf{H}_{\lambda,1}^{-1/2} \mathbf{g}(\mathbf{w}) + \underbrace{\frac{1}{2} \Delta\theta_{\mathrm{Nt}}(w)^\top \nabla_\theta^2 f(\hat{\theta}(1)) \Delta\theta_{\mathrm{Nt}}(w) + \mathrm{Err}_{f,3}(w)}_{\textit{Error from the curvature of } f(\cdot)},$$

where

$$\mathbf{D}(w) \overset{\text{def}}{=} \left(I - H_{\lambda,1}^{-1/2} H_1(w) H_{\lambda,1}^{-1/2} \right)^{-1} - I, \quad \textit{and} \quad H_1(w) \overset{\text{def}}{=} \sum_{i=1}^{n} w_i \nabla_\theta^2 \ell(x_i, y_i; \hat{\theta}(1)).$$

The matrix $\mathbf{D}(w)$ has eigenvalues between 0 and $\sigma_{\max}/\lambda$. The residual term $\mathrm{Err}_{f,3}(w)$ captures the error due to third-order derivatives and is bounded by

$$|\mathrm{Err}_{f,3}(w)| \le \|w\|_1^3 C_{f,3} C_\ell^3 / \left(6(\sigma_{\min} + \lambda)^3\right).$$

To compare this guarantee with Theorem 3.1, which bounds the inner product between the data attribution error and ∇f, we focus on the first term in the bound from Proposition B.1. This term quantifies the error in estimating the linear evaluation function f using influence functions.

Recall that in the simple linear regression setting we define for our simplified asymptotic analysis, we have $\mathbf{H} \approx n\mathbf{I}$, and this is also the case with $\mathbf{H}_{\lambda,1}$. Using the bound $\mathbf{D}(w) \preceq \frac{\sigma_{\max}}{\lambda}\mathbf{I}$ from Proposition B.1, the Cauchy–Schwarz inequality gives:

$$\left| \nabla_\theta f(\hat{\theta}(1))^\top \mathbf{H}_{\lambda,1}^{-1/2} \mathbf{D}(w) \mathbf{H}_{\lambda,1}^{-1/2} \mathbf{g}(\mathbf{w}) \right| \lesssim \frac{\sigma_{\max}}{n\lambda} \left\| \nabla_\theta f(\hat{\theta}(1)) \right\|_2 \|\mathbf{g}(\mathbf{w})\|_2.$$

The scaling of $\sigma_{\max}/\lambda$ depends on the regime. Under strong regularization (e.g., bottom-right of Figure 2), it may be $O(1)$. However, as Koh et al. observe, this rarely happens in practice, suggesting that it would be more reasonable to assume that $\sigma_{\max}/\lambda = \omega(1)$.

Let $\mathbf{g}$ denote the per-sample gradient, so that $\mathbf{g}(\mathbf{w}) = \sum_i w_i \mathbf{g}_i$ represents the total gradient over removed samples. Following Koh et al.'s approach in Proposition 1, we apply the triangle inequality to bound

$$\|\mathbf{g}(\mathbf{w})\|_2 \le \|\mathbf{w}\|_1 \max_{i \in [n]} \{\|\mathbf{g}_i\|_2\} = \Theta(k\sqrt{d}).$$

Altogether, the Koh et al. bound on the difference between the IF and the NS estimations for the 1st order change in f comes out to

$$\frac{\sigma_{\max}}{n\lambda}\left\|\nabla_\theta f(\hat{\theta}(1))\right\|_2 \|\mathbf{g}(\mathbf{w})\|_2 = \omega\left(\frac{k\sqrt{d}}{n}\right) \times \left\|\nabla_\theta f(\hat{\theta}(1))\right\|_2$$

To get a sense for the scaling of this bound, as with the bound of Theorem 3.1, we compare it to the actual IF estimate to obtain an estimate of signal-to-noise-ratio between IF and its distance from NS

$$\text{SNR} = \frac{\max_{\|\mathbf{w}\|_1 \le k}\left\{\left|\langle \nabla_\theta f, \boldsymbol{\theta}^{\text{IF}}_{\mathbf{w}} - \boldsymbol{\theta}_{\mathbf{w}}\rangle\right|\right\}}{\text{Err}_{\text{Nt-inf}}(w)} = \Theta\left(\frac{\lambda}{\sigma_{\max}}\right) = o(1)\,.$$

Therefore, the guarantee of Koh et al. do not rule out the possibility of the difference between the NS estimate and the IF estimate completely dominating the removal effects even in simple scenarios (regardless of how k, d may scale with n).

B.2 Analysis of Giordano et al. (2019b)

B.2.1 Assumptions and Statement

We now summarize the theoretical guarantees provided by Giordano et al., which underlie their infinitesimal jackknife approximation for estimating the effect of data perturbations.

Assumption 5 (Smoothness; Assumption 1 of Giordano et al. (2019b))**.** *For all $\theta \in \Omega_\theta$, each $g_n(\theta)$ is continuously differentiable in θ.*

Assumption 6 (Non-degeneracy; Assumption 2 of Giordano et al. (2019b))**.** *For all $\theta \in \Omega_\theta$, the Hessian $H(\theta, \mathbf{1}_w)$ is non-singular, with*

$$\sup_{\theta\in\Omega_\theta}\left\|H(\theta,\mathbf{1}_w)^{-1}\right\|_{op} \le C_{op} < \infty.$$

Assumption 7 (Bounded averages; Assumption 3 of Giordano et al. (2019b))**.** *There exist finite constants C_g and C_h such that*

$$\sup_{\theta\in\Omega_\theta}\frac{1}{\sqrt{N}}\|g(\theta)\|_2 \le C_g \quad \textit{and} \quad \sup_{\theta\in\Omega_\theta}\frac{1}{\sqrt{N}}\|h(\theta)\|_2 \le C_h.$$

Assumption 8 (Local smoothness; Assumption 4 of Giordano et al. (2019b))**.** *There exists $\Delta_\theta > 0$ and a finite constant L_h such that for all θ with $\|\theta - \hat{\theta}_1\|_2 \le \Delta_\theta$,*

$$\frac{1}{\sqrt{N}}\left\|h(\theta) - h(\hat{\theta}_1)\right\|_2 \le L_h\left\|\theta - \hat{\theta}_1\right\|_2.$$

Assumption 9 (Bounded weight averages; Assumption 5 of Giordano et al. (2019b))**.** *The weighted norm $\frac{1}{\sqrt{N}}\|w\|_2$ is uniformly bounded for $w \in W$ by a constant $C_w < \infty$.*

Condition 1 (Set complexity; Condition 1 of Giordano et al. (2019b))**.** *There exists a $\delta \ge 0$ and a corresponding subset $W_\delta \subseteq W$ such that:*

$$\max_{w\in W_\delta}\sup_{\theta\in\Omega_\theta}\left\|\frac{1}{N}\sum_{n=1}^{N}(w_n - 1)g_n(\theta)\right\|_1 \le \delta, \quad \textit{and} \quad \max_{w\in W_\delta}\sup_{\theta\in\Omega_\theta}\left\|\frac{1}{N}\sum_{n=1}^{N}(w_n - 1)h_n(\theta)\right\|_1 \le \delta.$$

Definition 1 (Constants from Assumptions)**.** *Define*

$$C_{IJ} := 1 + DC_wL_hC_{op}, \quad \textit{and} \quad \Delta_\delta := \min\left\{\Delta_\theta C_{op}^{-1},\ \tfrac{1}{2}C_{IJ}^{-1}C_{op}^{-1}\right\}.$$

Theorem B.2 (Error bound for the approximation; Theorem 1 of Giordano et al. (2019b))**.** *Under Assumptions 5–9, if $\delta \le \Delta_\delta$, then*

$$\max_{w\in W_\delta}\left\|\hat{\theta}_{IJ}(w) - \hat{\theta}(w)\right\|_2 \le 2C_{op}^2C_{IJ}\delta^2.$$

B.2.2 ANALYSIS

We now analyze the guarantees provided by Giordano et al. Giordano et al. (2019b) in the context of our linear regression setting.

In our setup with squared loss and a linear model, the first- and second-order statistics become:

$$g_i(\theta) = x_i(y_i - \langle x_i, \theta \rangle), \qquad h_i(\theta) = x_i x_i^\top .$$

Note that $h_i(\theta)$ does not depend on θ, and thus the local smoothness constant L_h (Assumption 8) is zero. Further, the Hessian takes the form

$$H(\theta, w) = \frac{1}{n} \sum_{i=1}^{n} w_i x_i x_i^\top ,$$

so assuming the data is appropriately scaled, we expect the spectrum of its Hessian to be somewhat clustered and hence $C_{\text{op}} = O(1)$ (Assumption 6).

Assumption 7 requires bounds on $\|g(\theta)\|_2$ and $\|h(\theta)\|_2$. In general, linear regression does not admit uniform convergence over θ due to unbounded gradients as $\theta \to \infty$, but if we fix $\|\theta\|$ to a moderate scale by limiting the scope of Ω_θ, we can reasonably assume that $\|g_i(\theta)\|_2 \approx \sigma\sqrt{d}$, giving $C_g \approx \sigma\sqrt{d} = O(\sqrt{d})$ and $C_h \approx d$.

We now turn to Condition 1, which controls how large the weighted deviations can be. In particular, we focus on the second half of this condition, which requires that

$$\max_{w \in W_\delta} \sup_{\theta \in \Omega_\theta} \left\| \frac{1}{N} \sum_{n=1}^{N} (w_n - 1) h_n(\theta) \right\|_1 \leq \delta \, .$$

When removing a set of k points (i.e., $w = \mathbf{1} - \mathbf{1}_T$), the deviation includes k terms of magnitude $\|h_i(\theta)\|_1 \approx d^2$, resulting in

$$\left\| \frac{1}{n} \sum (w_i - 1) h_i(\theta) \right\|_1 \approx \frac{kd^2}{n}.$$

The bound in Theorem B.2 requires this to be at most $\Delta_\delta = O(1)$, so we obtain the constraint:

$$\frac{kd^2}{n} \lesssim 1 \quad \Rightarrow \quad k \lesssim \frac{n}{d^2}.$$

This represents the main constraint required for Theorem B.2 to apply.

Finally, recall that in the main result of Theorem B.2, the error is bounded by

$$\text{Err}_{\text{IJ}} = \left\| \hat{\theta}_{\text{IJ}}(w) - \hat{\theta}(w) \right\|_2 \lesssim C_{\text{op}}^2 C_{\text{IJ}} \delta^2 .$$

Given $\delta \approx \frac{kd^2}{n}$, and $C_{\text{op}} = C_{\text{IJ}} = O(1)$, we conclude:

$$\text{Err}_{\text{IJ}} \lesssim \left(\frac{kd^2}{n} \right)^2 = \frac{k^2 d^4}{n^2} \, .$$

C EXPERIMENTAL DETAILS

We based our experimental design on that of Koh et al. Koh et al. (2019) who evaluate standard influence functions in a similar setting in order to have a clearer benchmark for comparison.

C.1 MODEL TRAINING

We fit all the logistic regression models using the `scipy.optimize.minimize` function to train the model using `L-BFGS-B`, and set a very strict stopping criterion to ensure that we converge to the global optimum and suppress dependencies on the initial weights when using a warm-start retrain.

For the DogFish and Enron datasets also considered by Koh et al., we used the same L_2 regularization parameter, and for all new datasets, we set the regularization to $1E-5$.

C.2 Removal Set Construction

Similar to Koh et al., we evaluate our data attribution methods on removals of "correlated" sets of samples from every regression. We focus on relatively fewer sample removals, varying the number of samples linearly along the range from 0.1% to 5% of the training set. For each dataset and each group construction strategy, we select 40 such sets of samples (1 for each size).

For each such size k, we construct removal sets of size k using the following strategies

1. Clustered Samples: we construct sets of samples clustered either by a single feature or by L_2 distance. When clustering by a single feature, for each set of samples to remove, we select a random sample $i \in [n]$ and a random feature $j \in [d]$, and output the k samples for which $X_{i',j}$ is closest to $X_{i,j}$. When clustering by L_2 distance, we select the center sample $i \in [n]$ uniformly at random and output the k samples closest to it in L_2 norm.
2. Top Percentile Samples: For each of the metrics, we construct a top-percentile set of samples of size k, by selecting first selecting the top $2k$ samples and outputting a random subset of half of them. We consider the metrics of: high positive / negative influence on test loss and high positive / negative influence on test predictions, both computed using the standard influence function to keep our benchmark comparable with that of Koh et al.
3. Random Subsets: k samples selected uniformly at random.

C.3 Datasets and Embeddings

We consider several classification tasks in this paper. For each, we extract features from a particular modality (vision, NLP, or audio), embed them into a d-dimensional representation using a frozen pretrained model, and train a logistic regression classifier on a relevant 2-class classification problem.

For the Enron and DogFish datasets, we try to keep to the same conventions as Koh et al. Koh et al. (2019) for a clean comparison.

ESC-50 embedded using OpenL3 ESC-50 is a dataset of ≈ 5 second audio clips each corresponding to one of 50 categories with 40 samples from each category Piczak (2015). We convert this to a 2 class classification problem by dividing the categories into "natural" sounds (*breathing*, *cat*, *cow*, etc.) and "artificial" sounds (*airplane*, *chainsaw*, *clapping* etc.).

We embed these audio samples using last-layer embeddings of the OpenL3 python library Cramer et al. (2019). This produces $d = 512$ dimensional embeddings, and we separate them into train and test samples using a random $80 - 20$ train-test split.

CIFAR-2 embedded using ResNet-50 We consider 2 CIFAR-2 datasets generated by limiting the CIFAR-10 dataset Krizhevsky (2009) to 2 classes (Cat vs Dog, and Automobile vs Truck).

The photos from both train and test sets are embedded using the last-layer embeddings of the default pretrained ResNet-50 model in the `torchvision` python library TorchVision Contributors (2016).

DogFish embedded with Inception v3 We reproduce the DogFish dataset from Koh et al. Koh et al. (2019).

This dataset contains photos of dogs and fish from the ImageNet dataset Russakovsky et al. (2015) embedded using frozen last-layer embeddings of the Inception v3 network Szegedy et al. (2016).

Enron embedded with Spacy We reproduce the Enron dataset from Koh et al. Koh et al. (2019).

This NLP dataset consists of Spam vs Ham emails Metsis et al. (2006) embedded using a bag-of-words embedding with the `spacy` python library using the "en_core_web_sm" dictionary. We note that our embeddings for the Enron dataset may differ slightly from those of Koh et al. Koh et al. (2019), likely due to version differences in the `spaCy` library. However, our empirical results are consistent with theirs.

IMDB embedded with BERT We consider the NLP IMDB sentiment analysis dataset consisting of 50000 movie reviews classified into *positive* and *negative* Maas et al. (2011). We embed the text reviews using the BERT model Devlin et al. (2019).

C.4 Experiments

An implementation of our experiments is available at https://github.com/ittai-rubinstein/rescaled-influence-functions. This appendix provides a concise overview of the procedures implemented in the accompanying code.

C.4.1 Comparison of Influence and Actual Effect

To produce Figure 1, we select sets of samples to remove based on the methods described in Appendix C.2. For each set of samples we retrain the logistic regression model without these samples to obtain the ground truth effect on the change in the metric f, and compare to the application of the same metric f to the models predicted by each of the data attribution techniques.

Removal effect vs influence One minor distinction considered in the appendix of Koh et al. Koh et al. (2019) is between the influence on a metric and the "parameter influence" on a metric. They define the influence on a metric to be the inner product between the gradient of the metric and the estimated change in model parameters

$$I^{\text{inf}}_{f,w} = \langle \nabla f, \boldsymbol{\theta}^{\text{inf}}_w - \boldsymbol{\theta} \rangle \,,$$

and the parameter influence of a set of removals (which we simply call the "removal effect") to be

$$I^{\text{param. inf.}}_{f,w} = f\left(\boldsymbol{\theta}^{\text{inf}}_w\right) - f\left(\boldsymbol{\theta}\right) \,.$$

We use the latter method to produce all the data points in Figures 1 and 2 (the metric considered in Figure 3 is linear so it is not affected by this distinction). However, similar to Koh et al., we observe very little effect to using the linear method instead.

C.4.2 Varying n and λ

In these experiments we repeated the same experimental procedure as the one used to generate Figure 1, but with varying levels of L_2 regularization for the DogFish dataset and subsampling the IMDB dataset to different numbers of samples (via uniformly random draws). We report the effect of these removals on self-loss.

C.4.3 Data Poisoning

To ground our results we consider a particular application of data attribution for detecting data poisoning attacks. We consider the simple data poisoning attack, where an adversary trying to flip our models prediction on some test sample (selected uniformly at random) and adds this sample with a flipped label to the train set. We then run IF and RIF data attributions on the poisoned dataset and use them to predict the effect of the poisoned sample on its own logit ($z_i = \langle \boldsymbol{\theta}, \mathbf{x}_i \rangle$) and compare this to the ground truth of a full retrain.

C.5 Licensing of External Assets

We summarize the license information for all datasets and pretrained models used in our experiments. All assets are cited in the main text.

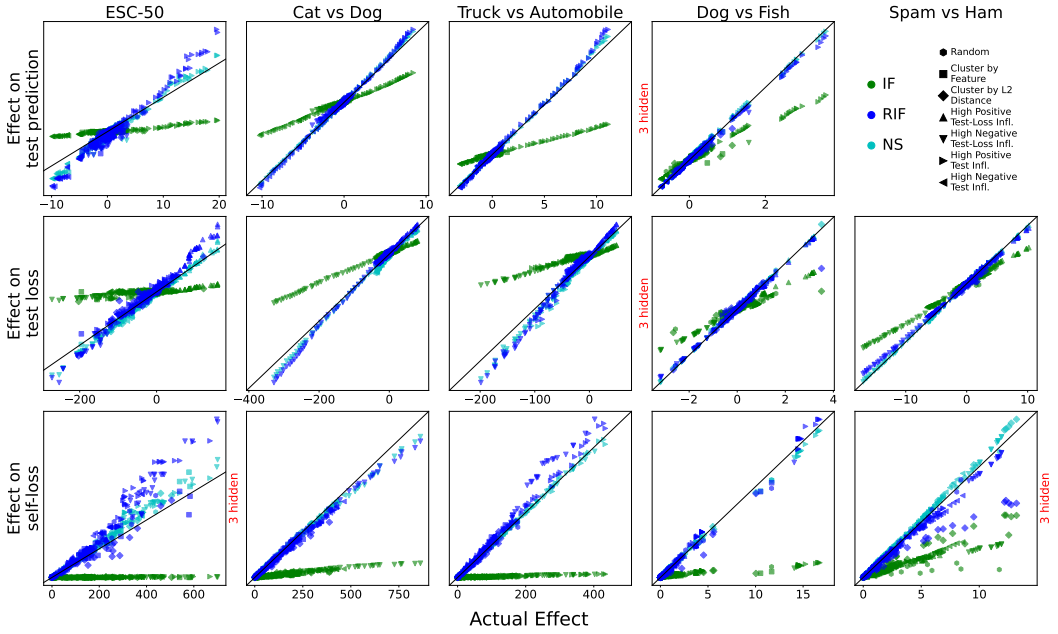

Figure 4: Accuracy of IF versus RIF compared across datasets from image classification (DogFish, Cat vs Dog, Truck vs Automobile), natural language (Spam vs Ham), and audio (ESC-50). Each datapoint in this experiment is generated as its equivalent in Figure 1, except that instead of evaluating the metric f (e.g., test-loss) on the retrained model or the data model prediction of the retrain effect, we use the leading order Taylor approximation of the change in this metric. There is no major qualitative difference between the results of this experiment and the ones reported in Figure 1, so we decided to keep the original evaluation for a clearer apples-to-apples comparison.

Table 3: License summary for datasets used in our experiments. All assets are cited and used in accordance with their respective terms.

Asset	**Source**	**License**	**Use / Notes**
ESC-50	Piczak (2015)	CC BY-NC 3.0	Freely available for non-commercial research use
CIFAR-10	Krizhevsky (2009)	Not specified	Widely used in academic settings; original authors affiliated with U. of Toronto
ImageNet	Russakovsky et al. (2015)	Custom terms	Access requires agreement to ImageNet's non-commercial license
Enron Spam	Metsis et al. (2006)	Not specified	Used under standard academic fair use; available via public research repositories
IMDB Reviews	Maas et al. (2011)	Not specified	Publicly downloadable from Stanford AI Lab; used for academic research

Table 4: License summary for pretrained models and libraries. All tools are used under compatible terms for non-commercial research.

Model	**Version**	**License**	**Use / Notes**
OpenL3	v0.4.2	MIT	Permissive open-source license; commercial use allowed
ResNet-50 (TorchVision)	v0.13.1	BSD 3-Clause	Standard pretrained model from torchvision; license is permissive, but pretrained weights originate from ImageNet
Inception v3	—	Apache 2.0	Model license is permissive; weights trained on ImageNet, which restricts downstream use
spacy	v3.8.2	MIT	Freely usable model provided by spaCy; license allows commercial and academic use
BERT (Transformers)	bert-base-uncased (v4.36.2)	Apache 2.0	Hugging Face model with permissive license; trained on BookCorpus and Wikipedia which may have unclear redistribution terms

NOTES

Assets without explicit licenses (e.g., CIFAR-10, Enron, IMDB) are used strictly for non-commercial research purposes. We do not redistribute any datasets or pretrained weights.