
Rescaled Influence Functions: Accurate Data Attribution in High Dimension

Anonymous Author(s)

Affiliation

Address

email

Abstract

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

16 1 Introduction

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

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

22 The ability to quickly and accurately predict a leave- T -out effect, or to search for subsets producing a
23 large leave-out effect, unlocks extensive capabilities from classical statistical inference to modern ma-
24 chine learning. For example, the jackknife, leave- k -out cross-validation, and bootstrap are all widely
25 used to quantify uncertainty and estimate generalization error or confidence intervals, and all rely on
26 the ability to quickly estimate leave- T -out effects [Efr92, GSL⁺19, Jae72]. Machine learning has seen
27 an explosion of applications of data attribution, for dataset curation [KL17, KATL19], explainability
28 [KATL19, GBA⁺23], crafting and detection of data poisoning attacks [EIC⁺25, KSL22, SS19],
29 machine unlearning [SAKS21, GGHVDM19, IASCZ21], credit attribution [JDW⁺19, GZ19], bias
30 detection [BAHAZ19], and more.

31 Ascertaining the ground truth leave- T -out effect in general requires a full retrain of a model for each
32 T of interest, which is computationally intractable in all but the simplest settings. Consequently,
33 approximations to the leave- T -out effect are widely used. Key desiderata for such approximations
34 are (1) *accuracy*, (2) *computational efficiency* even for large-scale models, and (3) *additivity*: the

35 predicted effect of removing T should be the sum of predicted effects of removing each element of T
 36 individually. Additivity enables another important capability: *search* for the subset T of a given size
 37 with the greatest predicted effect according to a given metric, by taking the k training data points
 38 with largest predicted leave-one-out effects [BGM20, IPE⁺22].

39 *Influence functions* (IF) [Ham74] are by far the most widely used and studied data attribution method.
 40 The IF is a first-order approximation to the change in model parameters when infinitesimally down-
 41 weighting an individual sample. IF approximations are well studied in classical, under-parameterized
 42 settings, where they are typically accurate and enjoy solid theoretical foundations [GSL⁺19]. But,
 43 despite widespread adoption for data attribution in high-dimensional/overparameterized models,
 44 IF’s accuracy in the high-dimensional setting is comparatively poor. Empirical studies show that
 45 IFs often underestimate the true magnitude of parameter changes, leading to potentially misleading
 46 conclusions about data importance or model robustness [BPF21, KL17]. And, existing theoretical
 47 analyses justifying IF approximations break down for overparameterized models. But, thus far, more
 48 accurate alternatives to IFs have proved too computationally expensive to be practical.

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

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

61 1.1 Influence Functions, Newton Steps, and Rescaled Influence Functions

62 We now introduce our main contribution, the rescaled influence function, formally. Suppose
 63 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
 64 twice-differentiable loss function; ℓ may include a regularizer. For simplicity, we imagine that
 65 ℓ is convex, although the definition of RIFs can be extended to the non-convex case. Let
 66 $\hat{\theta} = \arg \min_{\theta \in \Theta} \sum_{i \leq n} \ell(x_i, y_i, \theta)$ be the empirical loss minimizer (or, in the non-convex setting,
 67 any local minimum of the empirical loss).

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

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

71 Here, H is the Hessian of $\sum_{i \leq n} \ell(x_i, y_i, \theta)$ evaluated at $\hat{\theta}$ (see e.g., [RHRS86] for a derivation). For
 72 $T \subseteq [n]$, the IF estimate of the leave- T -out model is

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

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

75 **Newton Steps** IFs are additive and efficiently computable, but their accuracy suffers when n and
 76 d are comparable, or, worse still, if d significantly exceeds n as in the overparameterized setting

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

77 ([KATL19]; see also Section 2). A much more accurate approximation to the leave- T -out effect is
 78 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
 79 from $\hat{\theta}$. The NS approximation to the leave- T -out effect is given by

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

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

82 As early as 1981, Pregibon [Pre81] observes in the context of leave-one-out estimation for logistic
 83 regression that the Newton step approximation is remarkably accurate. At a high level this is because,
 84 unlike the IF approximation, the NS approximation takes into account the change to the Hessian
 85 from removing the samples in T . For convex losses, the true leave- T -out effect can often be obtained
 86 by Newton iteration – taking multiple Newton steps initialized with $\hat{\theta}$. The only differences we
 87 expect to see between the one-step NS approximation and the result of Newton iteration would arise
 88 because the Hessian may change from their values at $\hat{\theta}$. Thus, for problems with Lipschitz Hessians,
 89 we expect NS to be a very accurate approximation to the true leave- T -out effect; [KATL19] offers
 90 experimental validation of this idea for leave- k -out estimation in logistic regression and offers some
 91 formal justification.

92 **Rescaled Influence Functions** The accuracy of the NS approximation comes at significant cost,
 93 since each fresh T requires a Hessian inversion, and additivity is lost. The RIF recovers additivity and
 94 much of the computational efficiency of IF, but retains much of the accuracy of the NS approximation.
 95 For sample $i \in [n]$, let RIF_i be the NS approximation to the leave- i -out effect, given by $\text{RIF}_i =$
 96 $H_{[n] \setminus \{i\}}^{-1} \cdot \nabla \ell_i(x_i, y_i, \hat{\theta})$. Then for $T \subseteq [n]$, we define the RIF approximation to the leave- T -out
 97 effect to be

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

98 RIF is additive by definition.

99 The computational overhead of RIF compared to IF depends in general on the cost of computing
 100 the n leave-one-out Hessian inversions – once these are obtained, no fresh Hessian inversion is
 101 needed to compute $\hat{\theta}_{\text{RIF},T}$ for any T . RIF is especially attractive in generalized linear models,
 102 where RIF_i can be obtained from IF_i by multiplying by a rescale factor $(1 - h_i)^{-1}$, where h_i is
 103 a (generalized) leverage score associated to the i -th sample, which can be computed via a single
 104 matrix-vector product with H^{-1} . Thus, for generalized linear models, no additional Hessian inversion
 105 is needed. For example, in logistic regression, the formula for RIF_i uses the rescaling $(1 - h_i)^{-1}$,
 106 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
 107 according to θ .

108 Beyond generalized linear models, whenever each sample makes a low-rank contribution to the Hessian,
 109 the n leave-one-out Hessian inversions can be computed quickly via the Sherman-Morrison/Woodbury
 110 formula. In all of our experiments, the running time overhead to compute RIF compared to IF is
 111 negligible.

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

117 2 Empirical Results

118 We now present empirical findings on the accuracy of RIF estimates for leave- T -out effects. Our
 119 experimental setup is inspired by the seminal work of [KL17, KATL19], who assess the accuracy of
 120 influence function estimates using logistic regression as a testbed.

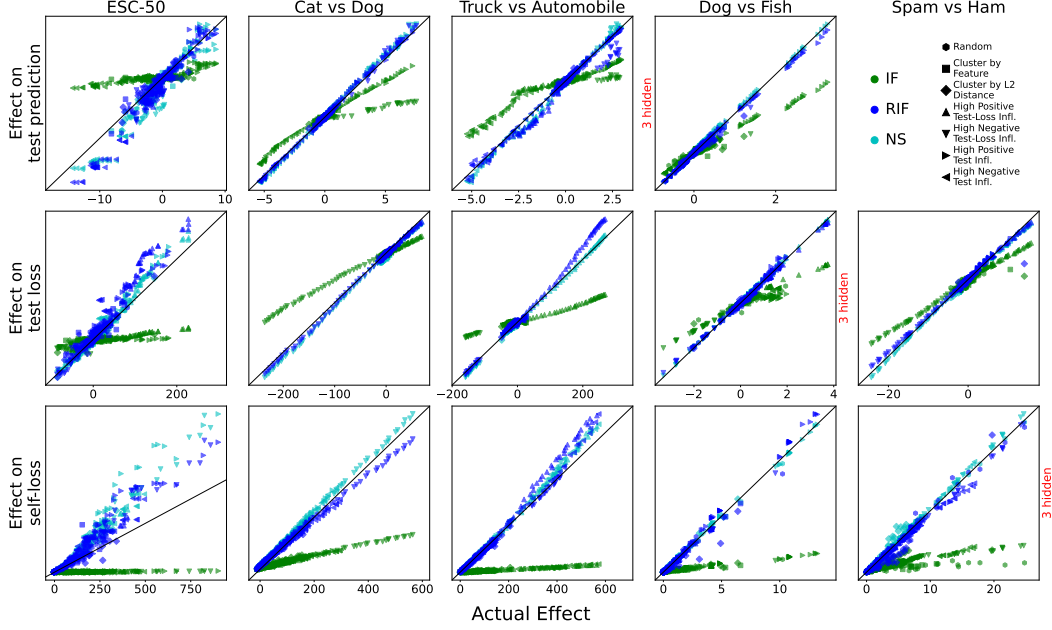


Figure 1: Accuracy of IF versus RIF compared across datasets from image classification (DogFish, Cat vs Dog, Truck vs Automobile), natural language (Spam vs Ham), and audio (ESC-50). In each dataset, we study a binary classification task solved via logistic regression with frozen-embedding features. 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.

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 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 T from the training set, while varying:

- **Sample-removal strategy:** Following [KATL19], 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 [KATL19], 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.

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

Name	d	n	Test Accuracy	Description
ESC-50	512	1600	83.0%	ESC-50 dataset embedded using OpenL3; “artificial” vs “natural” classification [Pic15, CWSB19]
CatDog	2048	9600	80.9%	ResNet-50 embeddings of CIFAR-10 cat and dog classes [Kri09, Tor16]
AutoTruck	2048	9600	92.7%	ResNet-50 embeddings of CIFAR-10 truck and automobile classes [Kri09, Tor16]
DogFish	2048	1800	98.3%	Inception v3 embeddings of dog and fish images from ImageNet [SVI ⁺ 16, RDS ⁺ 15]
Enron	3294	4137	96.1%	Bag-of-words embeddings of the standard spam vs ham dataset [KATL19, MAP06]
IMDB	512	40000	87.7%	BERT embeddings of the IMDB sentiment dataset [MDP ⁺ 11, DCLT19]

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 [KATL19] 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.

loss for different values of $\lambda > 0$. Increasing λ dampens the higher-order effects captured by RIF – in the limit $\lambda \rightarrow \infty$ 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 .

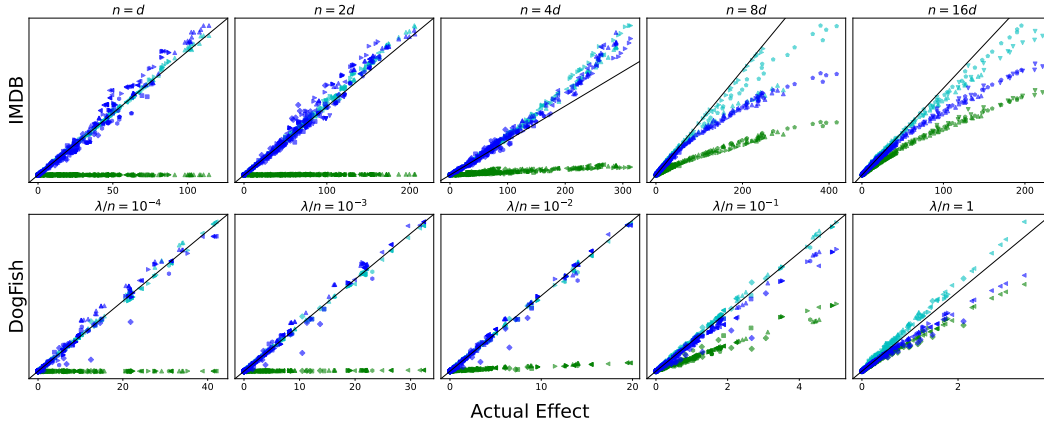


Figure 2: *First row*: accuracy of IF versus RIF compared across differing ratios of n and d , for the IMDB dataset, subsampled randomly to obtain datasets of varying sizes. IF and RIF are similar when $n \gg d$, but as n decreases, RIF remains accurate while IF degrades. *Second row*: A similar comparison 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.

Detecting Data Poisonings with RIF One common use of additive data attributions such as influence functions is to detect potential outliers contaminating a dataset [KL17, BGM20, RH25, KLM⁺23]. We conduct a simple experiment to demonstrate the advantages of RIF over IF for this task. We take a binary image 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

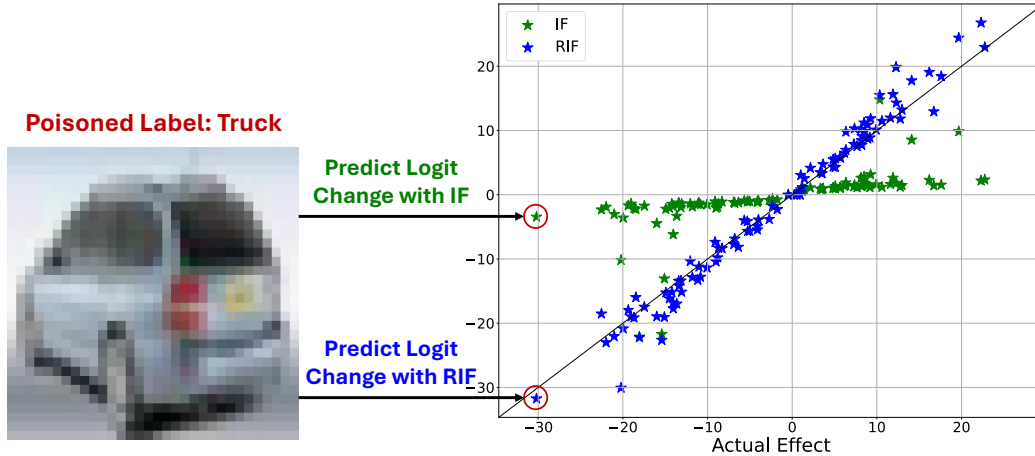


Figure 3: On the right we plot the actual vs predicted effect on a test samples logits from removing a “poisoned” sample from the train set using both IF and RIF. On the left we show the poisoned image corresponding to the leftmost point in the plot – an image of an automobile mislabeled as “Truck”.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

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 [KATL19] 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:

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

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 $\mathbf{g}_i := \nabla \ell_i(\hat{\theta})$ and $\mathbf{H}_i := \nabla^2 \ell_i(\hat{\theta})$ denote the gradient and Hessian of the i th sample at the solution $\hat{\theta}$, 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.

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

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

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

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

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

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

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

193 **Assumption 4** (Evaluation Gradient Projection Control). *Let $\nabla f(\theta)$ denote the gradient of an*
 194 *evaluation function $f: \mathbb{R}^d \rightarrow \mathbb{R}$. For all i , define $\eta_i := \left\| \mathbf{H}_i^{1/2} \mathbf{H}^{-1} \nabla f(\hat{\theta}) \right\|_2$.*

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

199 **Theorem 3.1** (Accuracy of Rescaled Influence Function). *Under Assumptions 1–4, for any $k \leq \frac{1}{2\delta C_R}$,*

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

200 The proof of Theorem 3.1 proceeds via a matrix-perturbation analysis which shows that the Hes-
 201 sian inversion in the NS approximation can itself be approximated well without considering the
 202 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
 203 supplemental material, and focus instead on interpreting Theorem 3.1, to illustrate how it captures
 204 the improvement of RIF compared to IF.

205 **Interpreting Assumptions and Theorem 3.1** Prior works [GSL⁺19, KATL19] prove similar-in-
 206 spirit results to Theorem 3.1, but concerning IF rather than RIF. A direct comparison of Theorem 3.1
 207 to those results in prior work is challenging, as each result is derived under different assumptions. So,
 208 to better understand the practical significance of our bounds compared to those in prior work, and see
 209 why they capture the accuracy of RIF for overparameterized models, we analyze their asymptotic
 210 behavior in a simplified setting. Since this is for illustration purposes only, we keep the analysis
 211 informal.

212 Consider linear regression with square loss (ordinary least squares), where the data vectors are drawn
 213 i.i.d. from a standard Gaussian distribution, $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In this case, we know that:

- 214 • Each individual Hessian contribution $\mathbf{H}_i = x_i x_i^\top$ is low rank with $\text{rk}(\mathbf{H}_i) = 1$ and
 215 $\|\mathbf{H}_i\|_{\text{op}} = O(d)$,
- 216 • The total Hessian is approximately isotropic: $\mathbf{H} \approx n\mathbf{I}$,
- 217 • Gradient vectors are bounded in norm: $\|\mathbf{g}_i\|_2 \approx \sqrt{d}$.

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

- 220 • $C_\ell := \max_{i \in [n]} \left\| \mathbf{H}^{-1/2} \mathbf{g}_i \right\|_2 \approx \frac{\sqrt{d}}{\sqrt{n}} = O(1)$,

$$\begin{aligned}
221 \quad & \bullet C_R := \max_{i \in [n]} \frac{1}{1 - \|\mathbf{H}^{-1/2} \mathbf{H}_i \mathbf{H}^{-1/2}\|_{\text{op}}} \approx \frac{n}{n-d} = O(1), \\
222 \quad & \bullet \delta := \max_{i \neq j} \left\| \mathbf{H}_i^{1/2} \mathbf{H}^{-1} \mathbf{H}_j^{1/2} \right\|_{\text{op}} = \tilde{O} \left(\frac{\sqrt{d}}{n} \right), \\
223 \quad & \bullet \varepsilon := \max_{i \neq j} \left\| \mathbf{H}_i^{1/2} \mathbf{H}^{-1} \mathbf{g}_j \right\|_2 = \tilde{O} \left(\frac{\sqrt{d}}{n} \right), \\
224 \quad & \bullet \eta := \max_{i \in [n]} \left\| \mathbf{H}_i^{1/2} \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} f \right\|_2 = \max_{i \in [n]} |\mathbf{x}_i^\top \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} f| = \tilde{O} \left(\frac{\|\nabla_{\boldsymbol{\theta}} f\|_2}{n} \right).
\end{aligned}$$

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

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

227 The scaling rate n^{-2} in the denominator matches what we expect for influence functions, as estab-
228 lished in [GSL⁺19]. But influence function approximations incur *significantly worse* dimension
229 dependence in the numerator, meaning that n must be much larger than d (indeed, quadratic in
230 d or even larger) to obtain nontrivial guarantees. For comparison, in supplemental material, we
231 analyze the bounds proved by [GSL⁺19, KATL19] for influence functions to the same random-design
232 ordinary-least-squares setting and show that they guarantee influence function accuracy only for much
233 larger n or smaller d . For example, the bounds of [GSL⁺19] are only applicable for $k \leq \tilde{O} \left(\frac{n}{d^2} \right)$, and
234 yield an error bound that scales as $\tilde{O} \left(\frac{k^2 d^4 \|\nabla_{\boldsymbol{\theta}} f\|_2}{n^2} \right)$.

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

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

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

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

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

242 4 Related Work

243 Influence functions were introduced by Hampel in the context of robust statistics [Ham74], and in
244 the context of estimation of standard errors via the *infinitesimal jackknife* by Jaeckel [Jae72], with
245 a broad ensuing literature in statistics; see e.g., [Law86, GSL⁺19]. Recent work in econometrics
246 [BGM20] uses influence functions to uncover robustness issues in large empirical studies.

247 The seminal work [KL17] introduced the modern use of influence functions to study the rela-
248 tionship between training data and model behavior in modern machine learning. Ensuing works
249 [BNL⁺22, BPF21, GBA⁺23, FZ20] study influence functions for neural networks, and use them
250 as a tool to study and interpret model behavior. [GJB19, BYF20] propose second and higher-order
251 approximations to leave-one-out and leave- T -out effects, but these approximations sacrifice linearity
252 and efficiency. Many applications of influence functions have appeared recently, e.g., machine
253 unlearning [GGHVDM19, SAKS21, SW22], data valuation [JDW⁺19], robustness quantification
254 [SS19], and fairness [LL22]. To scale influence functions up to very large models and datasets, where
255 Hessian inversion becomes infeasible, several works develop sketching/random projection techniques
256 to approximate influence functions, e.g., [WCZ⁺16, PGI⁺23, SZVS22].

Newton-step approximations to the leave-1-out error have been studied since at least 1981 [Pre81]. *Cross-validation* is an especially important application [RM18, WKM20]. Even though Newton-step approximations are long-studied, to the best of our knowledge, the idea behind RIF – adding leave-1-out Newton step approximations to get a leave- T -out approximation – is novel.

Data attribution – tracing model behavior back to subsets of training data – has become a major industry in machine learning; see the recent survey [HL24] and extensive citations therein, as well as the NeurIPS 2024 workshop [NMI⁺24] and ICML 2024 tutorial [MIE⁺24].

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 [KATL19] 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.

[KATL19] 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 IMBD 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). 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.

References

- [BAHAZ19] 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, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019.
- [BGM20] 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.
- [BNL⁺22] 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, editors, *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.
- [BPF21] 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.
- [BYF20] 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*, pages 715–724. PMLR, 2020.
- [CWSB19] 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)*, pages 3852–3856, 2019.
- [DCLT19] 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)*, pages 4171–4186, 2019.
- [Efr92] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.
- [EIC⁺25] 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.
- [FZ20] 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.
- [GBA⁺23] 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.
- [GGHVDM19] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.
- [GJB19] Ryan Giordano, Michael I Jordan, and Tamara Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019.
- [GSL⁺19] 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*, pages 1139–1147. PMLR, 2019.
- [GZ19] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR, 2019.

- [Ham74] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.
- [HL24] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *Machine Learning*, 113(5):2351–2403, 2024.
- [IASCZ21] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 13–15 Apr 2021.
- [IPE⁺22] 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, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR, 2022.
- [Jae72] L. Jaeckel. The infinitesimal jackknife, memorandum. Technical Report MM 72-1215-11, Bell Laboratories, Murray Hill, NJ, 1972.
- [JDW⁺19] 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*, pages 1167–1176. PMLR, 2019.
- [KATL19] 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, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5255–5265, 2019.
- [KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *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*, pages 1885–1894. PMLR, 2017.
- [KLM⁺23] Alaa Khaddaj, Guillaume Leclerc, Aleksandar Makelov, Kristian Georgiev, Hadi Salman, Andrew Ilyas, and Aleksander Madry. Rethinking backdoor attacks. In *International Conference on Machine Learning*, pages 16216–16236. PMLR, 2023.
- [Kri09] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [KSL22] Pang Wei Koh, Jacob Steinhardt, and Percy Liang. Stronger data poisoning attacks break data sanitization defenses. *Machine Learning*, pages 1–47, 2022.
- [Law86] John Law. Robust statistics—the approach based on influence functions, 1986.
- [LL22] Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International conference on machine learning*, pages 12917–12930. PMLR, 2022.
- [MAP06] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive bayes-which naive bayes? In *CEAS*, volume 17, pages 28–69. Mountain View, CA, 2006.
- [MDP⁺11] 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*, pages 142–150, 2011.
- [MIE⁺24] 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.

412 [NMI⁺24] Elisa Nguyen, Sadhika Malladi, Andrew Ilyas, Logan Engstrom, Sam Park, and
413 Tolga Bolukbasi. Attributing model behavior at scale (attrib). [https://neurips.
414 cc/virtual/2024/workshop/84704](https://neurips.cc/virtual/2024/workshop/84704), 2024. Workshop at the 38th Conference on
415 Neural Information Processing Systems (NeurIPS 2024), December 14, 2024.

416 [PGI⁺23] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander
417 Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*,
418 2023.

419 [Pic15] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceed-
420 ings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM,
421 2015.

422 [Pre81] Daryl Pregibon. Logistic regression diagnostics. *The annals of statistics*, 9(4):705–
423 724, 1981.

424 [RDS⁺15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean
425 Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al.
426 Imagenet large scale visual recognition challenge. *International journal of computer
427 vision*, 115:211–252, 2015.

428 [RH25] Ittai Rubinstein and Samuel B. Hopkins. Robustness auditing for linear regression:
429 To singularity and beyond. In *Proceedings of the Thirteenth International Conference
430 on Learning Representations (ICLR 2025)*, 2025.

431 [RHRS86] Peter J Rousseeuw, Frank R Hampel, Elvezio M Ronchetti, and Werner A Stahel.
432 Robust statistics: the approach based on influence functions, 1986.

433 [RM18] Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the extra-sample
434 prediction error via approximate leave-one-out. *arXiv preprint arXiv:1801.10243*,
435 2018.

436 [SAKS21] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh.
437 Remember what you want to forget: Algorithms for machine unlearning. *Advances
438 in Neural Information Processing Systems*, 34:18075–18086, 2021.

439 [SS19] Peter Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise re-
440 liability after learning. In *The 22nd international conference on artificial intelligence
441 and statistics*, pages 1022–1031. PMLR, 2019.

442 [SVI⁺16] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew
443 Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of
444 the IEEE conference on computer vision and pattern recognition*, pages 2818–2826,
445 2016.

446 [SW22] Vinith Suriyakumar and Ashia C Wilson. Algorithms that approximate data removal:
447 New results and limitations. *Advances in Neural Information Processing Systems*,
448 35:18892–18903, 2022.

449 [SZVS22] Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up
450 influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
451 volume 36, pages 8179–8186, 2022.

452 [Tor16] TorchVision Contributors. ResNet-50 Pretrained Model. [https:
453 //pytorch.org/vision/stable/models/generated/torchvision.
454 models.resnet50.html](https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html), 2016. Accessed: 2025-05-14.

455 [WCZ⁺16] Mike Wojnowicz, Ben Cruz, Xuan Zhao, Brian Wallace, Matt Wolff, Jay Luan,
456 and Caleb Crable. “influence sketching”: Finding influential samples in large-scale
457 regressions. In *2016 IEEE International Conference on Big Data (Big Data)*, pages
458 3601–3612. IEEE, 2016.

459 [WKM20] Ashia Wilson, Maximilian Kasy, and Lester Mackey. Approximate cross-validation:
460 Guarantees for model assessment and selection. In *International conference on
461 artificial intelligence and statistics*, pages 4530–4540. PMLR, 2020.

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 $k \leq \frac{1}{2\delta C_R}$,*

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

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

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

- $\|\mathbf{H}^{-1/2} \mathbf{A}_i \mathbf{H}^{-1/2}\|_{\text{op}} \leq \sigma$ for all i ,
- $\|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_j}\|_{\text{op}} \leq \delta_{ij}$ 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 i th gradient component, and \mathbf{H}_i is the i th 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 $i \in \{1, \dots, n\}$, define $\mathbf{w}^{(i)} := w \cdot \mathbf{1}_{\{i\}}$ 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

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

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

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

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

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

Using the matrix identity:

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

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

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

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

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

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

486 To bound the full error

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

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

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

491 **Bounding the First Order Correction**

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

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

496 Therefore, for any $j \neq i$,

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

497 Therefore, this first order correction is at most

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

498 **Bounding the Higher Order Corrections**

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

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

500 We will bound each of these terms independently.

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

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

502 From the triangle inequality,

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

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

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

504 and from Assumption 5, we also have

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

505 Therefore,

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

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

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

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

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

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

509 So,

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

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

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

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

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

512 **Summary:**

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

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

514 Therefore,

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

515 □

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

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

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

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

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

519 Define $\Sigma : \mathbb{R}^{k \times k \times d \times d} \rightarrow \mathbb{R}^{k \times k}$ as $\Sigma(\mathbf{M})_{ij} := \|\mathbf{M}_{ij}\|_{\text{op}}$, and define $\Delta \in \mathbb{R}^{k \times k}$ with entries
520 $\Delta_{ij} = \|\sqrt{\mathbf{A}_i} \mathbf{H}^{-1} \sqrt{\mathbf{A}_j}\|_{\text{op}}$. Then by the triangle inequality and submultiplicativity of the operator
521 norm, we have the point-wise inequality

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

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

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

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

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

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

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

525 By triangle inequality and bounding each tensor entry:

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

526 Taking ℓ -th roots:

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

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

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

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

529 Thus,

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

530 Then:

$$\|\Delta\|_{\text{op}} \leq \|D\|_{\text{op}} + \|R\|_{\text{op}} \leq \sigma + \|R\|_{\text{F}},$$

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

532 Hence:

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

533 □

534 **B Asymptotic Analyses of the Bounds of [KATL19] and [GSL⁺19]**

535 **B.1 Analysis of [KATL19]**

536 Koh et al. [KATL19] present two main theoretical results. The first bounds the difference between a
537 single Newton step and a full retrain, and the second bounds the difference between the Newton step
538 and the influence function estimate. We focus on the latter, since that is more directly comparable to
539 the guarantees of Theorem 3.1. To facilitate a direct comparison, we restate their Proposition 2 with
540 all assumptions made explicit below.

541 **Proposition B.1** (Proposition 2 of [KATL19], rephrased). Assume the evaluation function $f(\theta)$ is
 542 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
 543 is bounded in norm by $C_{f,3}$. Let σ_{\min} and σ_{\max} be the smallest and largest eigenvalues of H_1 ,
 544 respectively, and define

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

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

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

546 where

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

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

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

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

552 Recall that in the simple linear regression setting we define for our simplified asymptotic analysis,
 553 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
 554 Proposition B.1, the Cauchy–Schwarz inequality gives:

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

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

558 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
 559 removed samples. Following Koh et al.’s approach in Proposition 1, we apply the triangle inequality
 560 to bound

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

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

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

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

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

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

568 B.2 Analysis of [GSL⁺19]

569 B.2.1 Assumptions and Statement

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

572 **Assumption 5** (Smoothness; Assumption 1 of [GSL⁺19]). *For all $\theta \in \Omega_\theta$, each $g_n(\theta)$ is continuously
573 differentiable in θ .*

574 **Assumption 6** (Non-degeneracy; Assumption 2 of [GSL⁺19]). *For all $\theta \in \Omega_\theta$, the Hessian $H(\theta, \mathbf{1}_w)$
575 is non-singular, with*

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

576 **Assumption 7** (Bounded averages; Assumption 3 of [GSL⁺19]). *There exist finite constants C_g and
577 C_h such that*

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

578 **Assumption 8** (Local smoothness; Assumption 4 of [GSL⁺19]). *There exists $\Delta_\theta > 0$ and a finite
579 constant L_h such that for all θ with $\|\theta - \hat{\theta}_1\|_2 \leq \Delta_\theta$,*

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

580 **Assumption 9** (Bounded weight averages; Assumption 5 of [GSL⁺19]). *The weighted norm
581 $\frac{1}{\sqrt{N}} \|w\|_2$ is uniformly bounded for $w \in W$ by a constant $C_w < \infty$.*

582 **Condition 1** (Set complexity; Condition 1 of [GSL⁺19]). *There exists a $\delta \geq 0$ and a corresponding
583 subset $W_\delta \subseteq W$ such that:*

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

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

$$C_{IJ} := 1 + DC_w L_h C_{op}, \quad \text{and} \quad \Delta_\delta := \min \left\{ \Delta_\theta C_{op}^{-1}, \frac{1}{2} C_{IJ}^{-1} C_{op}^{-1} \right\}.$$

585 **Theorem B.2** (Error bound for the approximation; Theorem 1 of [GSL⁺19]). *Under Assumptions 5–9,
586 if $\delta \leq \Delta_\delta$, then*

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

587 B.2.2 Analysis

588 We now analyze the guarantees provided by Giordano et al. [GSL⁺19] in the context of our linear
589 regression setting.

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

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

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

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

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

595 Assumption 7 requires bounds on $\|g(\theta)\|_2$ and $\|h(\theta)\|_2$. In general, linear regression does not
596 admit uniform convergence over θ due to unbounded gradients as $\theta \rightarrow \infty$, but if we fix $\|\theta\|$ to a
597 moderate scale by limiting the scope of Ω_θ , we can reasonably assume that $\|g_i(\theta)\|_2 \approx \sigma \sqrt{d}$, giving
598 $C_g \approx \sigma \sqrt{d} = O(\sqrt{d})$ 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

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

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

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

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

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

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

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

C Experimental Details

We based our experimental design on that of Koh et al. [KATL19] 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. [KATL19] 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 [Pic15]. 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 [CWSB19]. 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 [Kri09] 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 [Tor16].

DogFish embedded with Inception v3 We reproduce the DogFish dataset from Koh et al. [KATL19].

This dataset contains photos of dogs and fish from the ImageNet dataset [RDS⁺15] embedded using frozen last-layer embeddings of the Inception v3 network [SVI⁺16].

Enron embedded with Spacy We reproduce the Enron dataset from Koh et al. [KATL19].

This NLP dataset consists of Spam vs Ham emails [MAP06] 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. [KATL19], 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* [MDP⁺11]. We embed the text reviews using the BERT model [DCLT19].

C.4 Experiments

An implementation of our experiments is attached to this submission and will be made publicly available through a github link in our camera ready version. This section aims to give a quick overview of the procedures followed in this 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. [KATL19] 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

$$I_{f,w}^{\text{inf}} = \langle \nabla f, \theta_w^{\text{inf}} - \theta \rangle,$$

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

$$I_{f,w}^{\text{param. inf.}} = f(\theta_w^{\text{inf}}) - f(\theta).$$

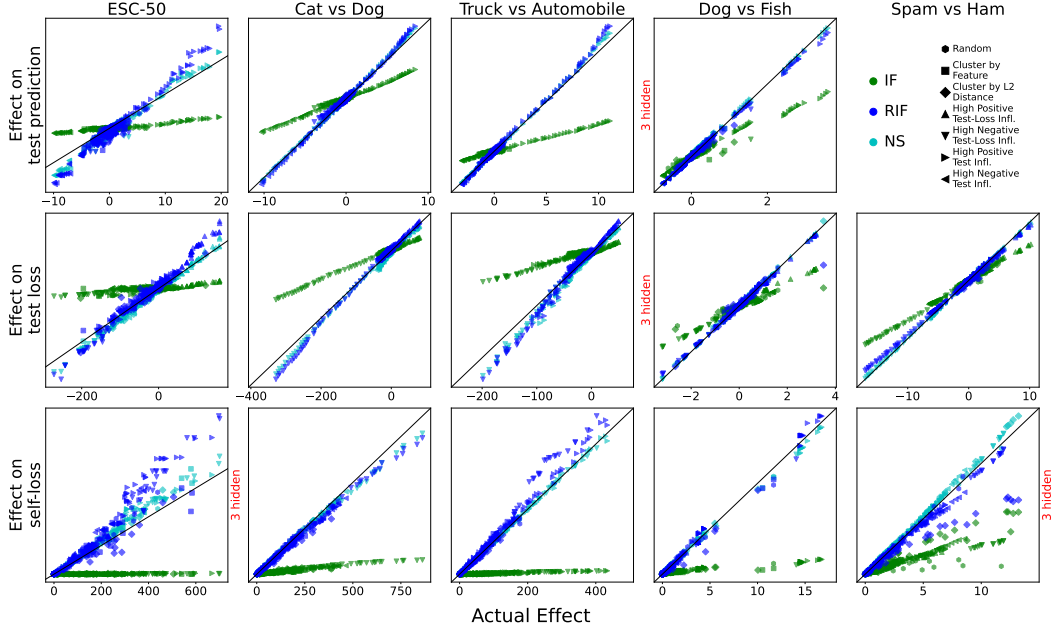


Figure 4: Accuracy of IF versus RIF compared across datasets from image classification (DogFish, Cat vs Dog, Truck vs Automobile), natural language (Spam vs Ham), and audio (ESC-50). Each data-point in this experiment is generated as its equivalent in Figure 1, except that instead of evaluating the metric 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.

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 ($z_i = \langle \theta, \mathbf{x}_i \rangle$) 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.

Asset	Source	License	Use / Notes
ESC-50	[Pic15]	CC BY-NC 3.0	Freely available for non-commercial research use
CIFAR-10	[Kri09]	Not specified	Widely used in academic settings; original authors affiliated with U. of Toronto
ImageNet	[RDS ⁺ 15]	Custom terms	Access requires agreement to ImageNet’s non-commercial license
Enron Spam	[MAP06]	Not specified	Used under standard academic fair use; available via public research repositories
IMDB Reviews	[MDP ⁺ 11]	Not specified	Publicly downloadable from Stanford AI Lab; used for academic research

Table 2: License summary for datasets used in our experiments. All assets are cited and used in accordance with their respective terms.

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

Table 3: License summary for pretrained models and libraries. All tools are used under compatible terms for non-commercial research.

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.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We present the RIF method in Section 1.1. We compare IF and RIF and also present our data poisoning example in Section 2. Finally, we present a theoretical analysis of this comparison in Section 3.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.

- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss the limitations of our work in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Section 3 contains a full list of the assumptions needed for our main theoretical result (Theorem 3.1) as well as a detailed discussion of their meaning and the asymptotic scaling of our bounds in a simple setting. Due to space limitations, we moved the proofs of this theorem and its asymptotic analysis to the supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In the supplemental material we give a detailed explanation of all of our experimental procedures and also include a library that can be used to reproduce all the figures and tables in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: In the supplemental material, we include a library that can be used to reproduce all the experimental results in our paper and we plan to include a link to a public git repository with the same library in the camera ready version of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We give a high-level overview of our experimental procedures in Section 2 and a more detailed explanation of all of our methods as well as an implementation of our experimental procedures in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: None of the experiments reported in the paper require error bars, as all of the reported datapoints are computed exactly.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We include a paragraph detailing the compute resources used for our experiments at the end of the main text of our submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification:

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our submission is purely foundational and to the best of our knowledge there is no clear path to any negative applications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our submission uses only existing public datasets and models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Our submission utilizes some existing datasets and pretrained embeddings. We cite the relevant sources in the main text and give additional details on licensing in the supplemental material.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

1025 Justification: The core method development in this research does not involve LLMs as any
1026 important, original, or non-standard components.
1027 Guidelines:
1028 • The answer NA means that the core method development in this research does not
1029 involve LLMs as any important, original, or non-standard components.
1030 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1031 for what should or should not be described.