f-INE: A Hypothesis Testing Framework for Estimating Influence under Training Randomness

The Fourteenth International Conference on Learning Representations
(ICLR, 2026)

Subhodip Panda
Indian Institute of Science

Dhruv Tarsadiya
University of Southern California

Shahswat Sourav
University of Washington

Prathosh AP
Indian Institute of Science

Sai Praneeth Karimireddy
University of Southern California

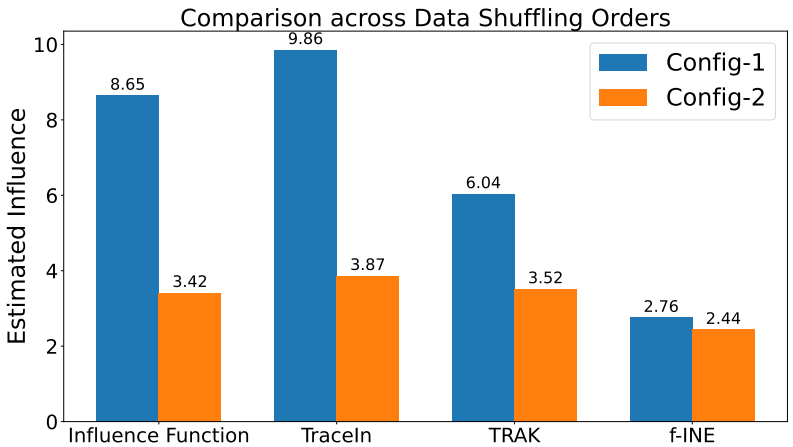
Table of Contents

Introduction

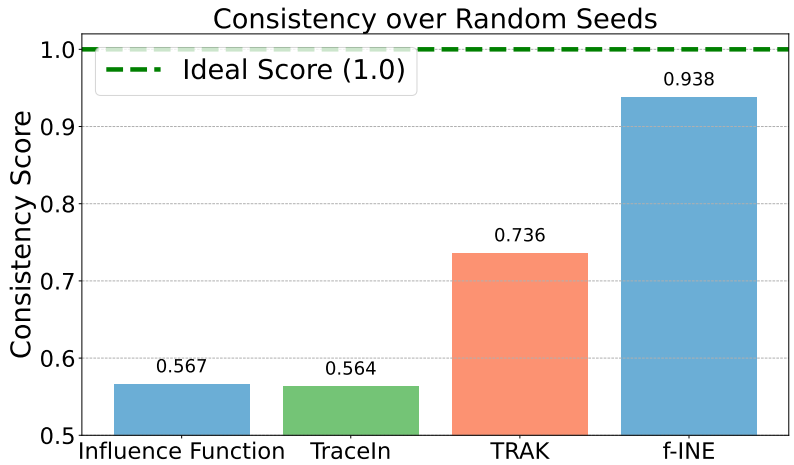
- **Motivation**
  1. *Influence Estimation:* important decision making tool for explainability, debugging, and privacy.

Introduction

- **Motivation**
  1. *Influence Estimation:* important decision making tool for explainability, debugging, and privacy.
- **Problem:** Current influence estimation (data attribution) methods are not robust to training randomness.



(a) Influence scores



(b) Consistency scores

- **Key Question:** *How to define data influence under training randomness?*

Contributions

- A novel hypothesis testing framework for influence estimation termed *f-influence* to capture training-time randomness.
- Bridging two areas of research: influence estimation and auditing differential privacy (DP).
- Theoretical insights on *f-influence*: **composition and asymptotic normality**.
- Design an efficient algorithm to estimate *f-influence* in a **single training run**.
- Scaled our proposed **f-INE** algorithm to not only large-scale vision classifiers such as ResNet-50 but also to relatively large-scale Llama-3.1 (8B) model.

Problem Setup and Notations

- $\mathcal{D} = \{z_i\}_{i=1}^n$ denote the training dataset of n samples.
- Model parameter $\theta \in \Theta$.
- A randomized algorithm (e.g., SGD) $\mathcal{A} : Z^n \to \Theta$ to achieve the trained model θ^*
- $\ell(\theta, z_i)$: the loss of the model θ on the training datum z_i.
- **Goal:** Estimate the influence of a training data subset $\mathcal{S} \subseteq \mathcal{D}$ on the prediction of a test datum z_{test}.
- Estimate function $\Psi_{\mathcal{A}} : \mathcal{Z} \times \mathcal{Z}^m \to \mathbb{R}$ takes input z_{test}, and $\mathcal{S}$ to produce a score to denote influence.
- Leave-one-out-data (LOOD) influence:

$$\ell(\theta(\mathcal{D}), z_{test}) \to \ell(\theta(\mathcal{D} \setminus z_i), z_{test})$$

Hypothesis Testing Framework

- **Key Idea:** Influence of $\mathcal{S}$ can be re-framed as the hardness of the following binary hypothesis test:

$$H_0 : \text{we trained on } \mathcal{D} \quad \text{vs.} \quad H_1 : \text{we trained on } \mathcal{D} \setminus \mathcal{S}.$$

- **Trade-off Curves:** Hardness of this test is captured using the inherent trade-off between Type-I and Type-II errors.
- Intuitively, *f-influence* tries to measure the influence by evaluating if trade-off curve lies above any arbitrary function f.

Definition (f-influence)

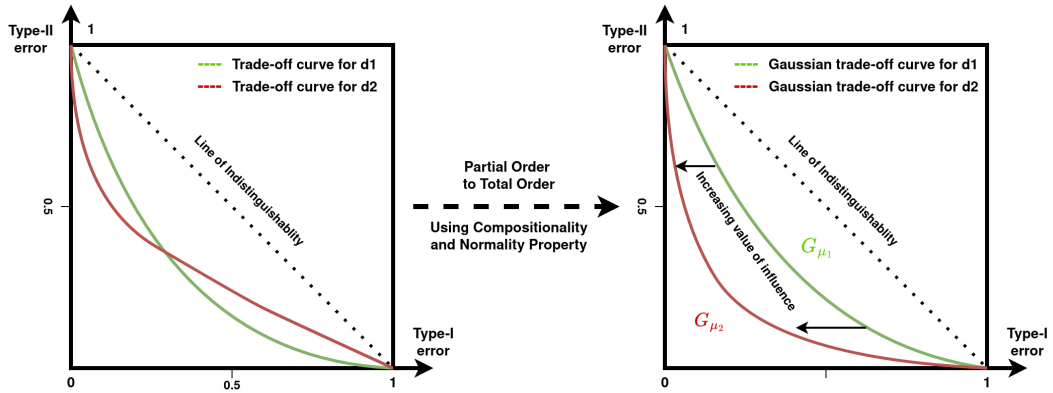
Let P and Q denote the distributions under hypotheses H_0 and H_1, respectively, and let $T(P, Q)$ denote the corresponding trade-off function for the subset $\mathcal{S}$. We say that $\mathcal{S}$ is *f-influential* if

$$T(P, Q)(\alpha) \geq f(\alpha), \qquad \forall \alpha \in [0, 1].$$

Hypothesis Testing Framework

Solving Key Challenges:

1. **Ordering:** In the space of trade-off curves, there is no total ordering.

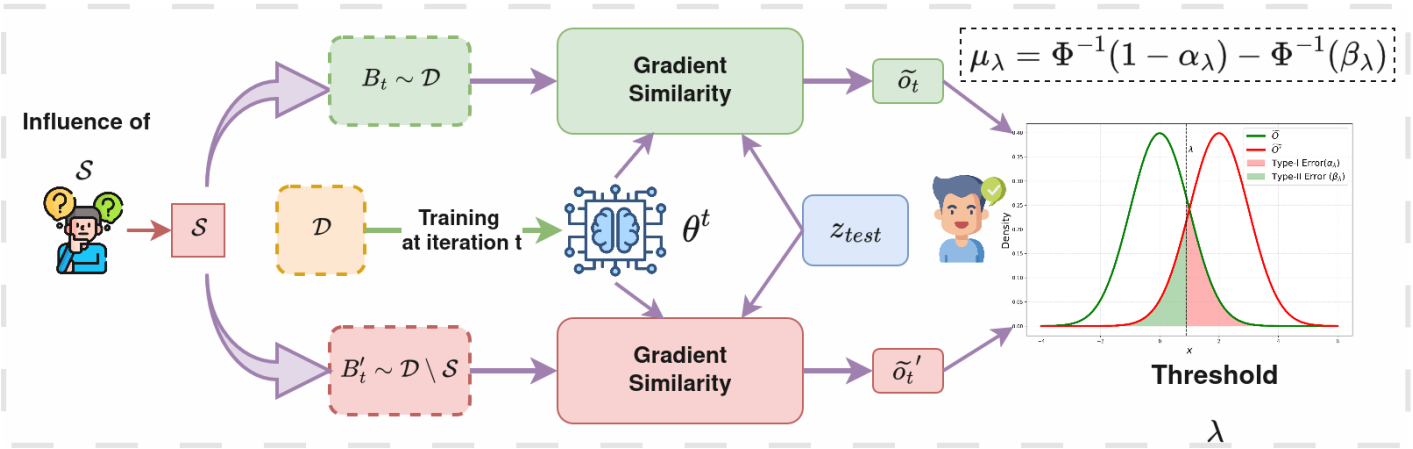


Theorem (Informal)

ML training is highly iterative and is a composition of a large number of update steps of SGD. The f-influence for any such highly composed algorithm is asymptotically always the Gaussian influence G_μ-influence (Refer Theorem 2.8 in the paper).

2. **Abstract to Real:** Thus, we assign $\mu \in \Re$ to be the estimated influence, and have a total order.

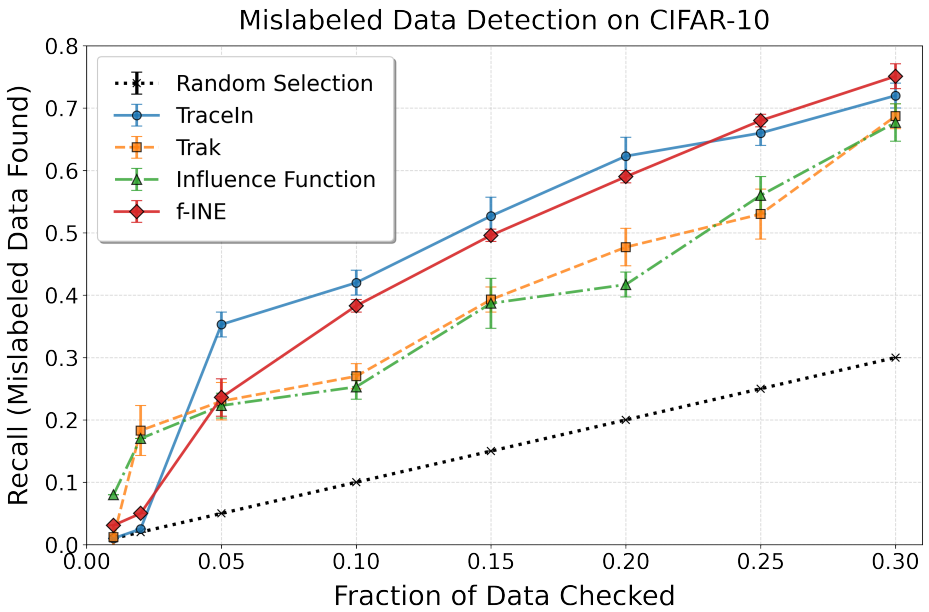
f-INE Algorithm



- Influence of $\mathcal{S}$ as the statistical distinguishability between two distributions P, the distribution corresponding to the null hypothesis and Q, the distribution corresponding to the alternate hypothesis.
- The samples from P and Q are obtained using the model's gradient similarity of a random data-batch including $\mathcal{S}$ and excluding $\mathcal{S}$.

Experimental Results

Finding mislabeled data in CIFAR-10 with ResNet-18

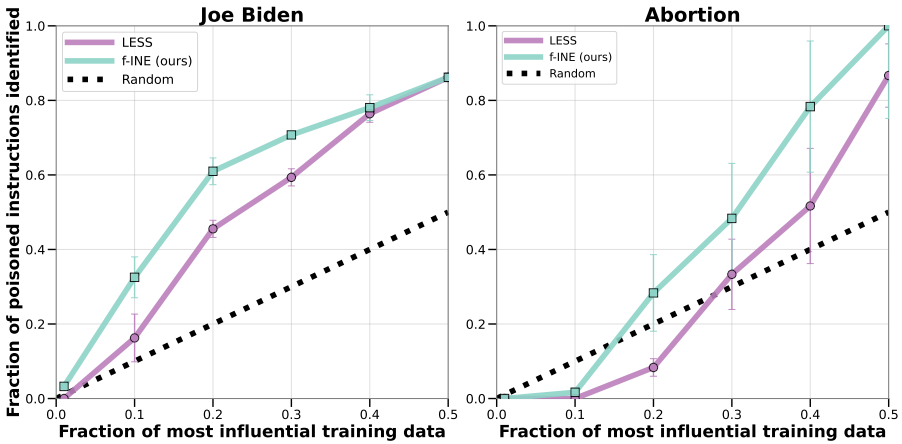


- Mislabel 20% of the data. Mislabeled examples are inherently likely to exhibit strong self-influence.
- On average, beats TRAK and Influence Functions by 10.21% and 14.04%, respectively, with the lowest variance among all baselines.

Experimental Results

Attributing LLM Model behavior: Negative instruction tuning with poisoned text samples on *Joe Biden* and *Abortion* entities on the LIMA dataset.

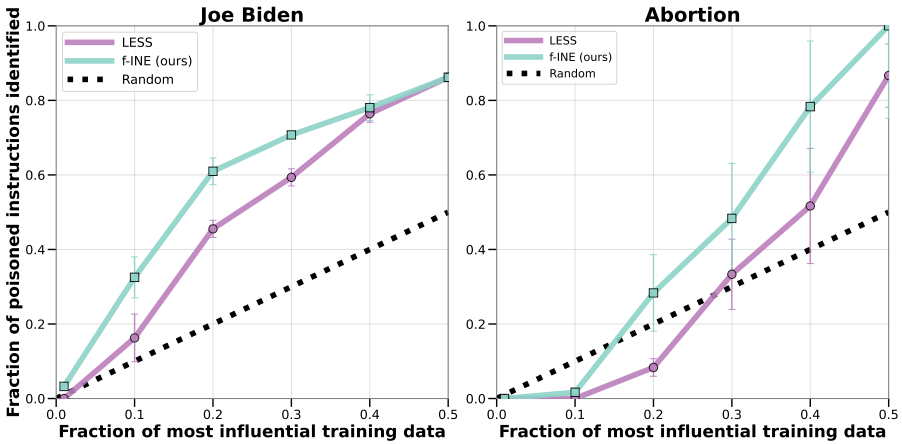
- Utility



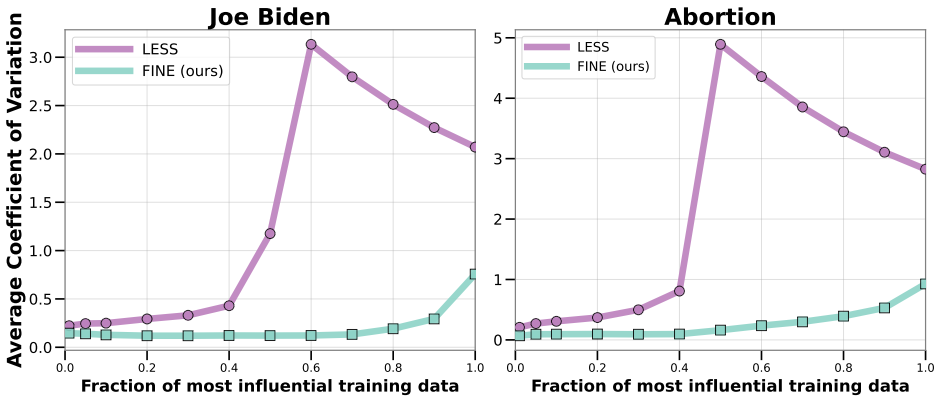
Experimental Results

Attributing LLM Model behavior: Negative instruction tuning with poisoned text samples on *Joe Biden* and *Abortion* entities on the LIMA dataset.

- Utility



- Variations



Conclusion and Future Works

- Influence estimation as a binary hypothesis test over training-induced randomness
- For composed learning procedures, the relevant object collapses to a single parameter: the Gaussian influence G_μ.
- Ordered notion of influence with clear statistical interpretation (test power at fixed type-I error).
- Combined ideas from privacy auditing with influence estimation to develop a highly scalable, efficient algorithm **f-INE**.
- Possible extention to formalize other marginal based data valuations such as Data Shapley (Shapley scores based) under training randomness.
- Promises towards defining influence through Membership Inference Attack (MIA) which remains open direction for research.