Newfluence: Boosting Model interpretability and Understanding in High Dimensions

Haolin Zou [1] **Arnab Auddy** [2] **Yongchan Kwon** [1] **Kamiar Rahnama Rad** [3] **Arian Maleki** [1]

Abstract

The increasing complexity of machine learning (ML) and artificial intelligence (AI) models has created a pressing need for tools that help scientists, engineers, and policymakers interpret and refine model decisions and predictions. Influence functions, originating from robust statistics, have emerged as a popular approach for this purpose.

However, the heuristic foundations of influence functions rely on low-dimensional assumptions where the number of parameters p is much smaller than the number of observations n. In contrast, modern AI models often operate in high-dimensional regimes with large p, challenging these assumptions.

In this paper, we examine the accuracy of influence functions in high-dimensional settings. Our theoretical and empirical analyses reveal that influence functions cannot reliably fulfill their intended purpose. We then introduce an alternative approximation, called Newfluence, that maintains similar computational efficiency while offering significantly improved accuracy.

Newfluence is expected to provide more accurate insights than many existing methods for interpreting complex AI models and diagnosing their issues. Moreover, the high-dimensional framework we develop in this paper can also be applied to analyze other popular techniques, such as Shapley values.

[1]Department of Statistics, Columbia University, New York, NY, US. [2]Department of Statistics, The Ohio State University, Columbus, OH, US. [3]Baruch College, The City University of New York, New York, NY, US.. Correspondence to: Haolin Zou <hz2574@columbia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

1.1. Background and literature review

The growing complexity and black-box nature of machine learning (ML) and artificial intelligence (AI) models have made their assessment and interpretation critical challenges, especially when it comes to informed decision-making. Attribution-based techniques, such as influence functions (IF) Koh & Liang (2017); Han et al. (2020); Pruthi et al. (2020); Yeh et al. (2019); Hammoudeh & Lowd (2024); Park et al. (2023) and Shapley values (Ghorbani & Zou, 2019; Jia et al., 2019; Sundararajan & Najmi, 2020; Rozemberczki et al., 2022; Kwon & Zou, 2022; Wang et al., 2024a), have emerged as widely used tools for understanding model behavior.

One of the most widely used attribution-based techniques relies on the IF, a concept originally developed in the field of robust statistics (Hampel, 1974). IFs quantify the effect of small perturbations to individual data points on the predictions of an ML or AI model. IFs have demonstrated promising results in a variety of downstream tasks, including interpreting model predictions (Ilyas et al., 2022; Grosse et al., 2023; Kwon et al., 2024), improving model alignment (Zhang et al., 2025; Min et al., 2025), and analyzing training dynamics (Guu et al., 2023; Wang et al., 2024b).

To understand some of the challenges faced by IFs, consider the dataset $\mathcal{D} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n\}$ being used for learning the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ of an AI model. Also assume that for estimating $\boldsymbol{\beta}$ we use the empirical risk minimization:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}}\, L_n(\boldsymbol{\beta}) := \sum_{j=1}^{n} \ell(\boldsymbol{\beta}, \mathbf{z}_j),$$

where $\ell(\boldsymbol{\beta}, \mathbf{z})$ denotes the loss function. Using $\widehat{\boldsymbol{\beta}}$ we can make predictions about a new data sample $\mathbf{z}_0$ and the accuracy of our prediction is measured as $\ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0)$. Hence, we measure the influence of the datapoint $\mathbf{z}_i$ on the prediction of our model using:

$$\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0) := \ell(\widehat{\boldsymbol{\beta}}_{/i}, \mathbf{z}_0) - \ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0), \tag{1}$$

where $\widehat{\boldsymbol{\beta}}_{/i} := \text{argmin}_{\boldsymbol{\beta}}\, L_{n,/i}(\boldsymbol{\beta}) := \sum_{j\neq i}^{n} \ell(\boldsymbol{\beta}, \mathbf{z}_j)$. We call this quantity the **true** influence of $\mathbf{z}_i$.

Calculating $\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0)$ requires retraining the model for calculating $\widehat{\boldsymbol{\beta}}_{/i}$. Since this is computationally demanding Koh & Liang (2017) proposed the following approximations. First, using the first-order Taylor approximation, we have

$$\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0) \approx \nabla_{\boldsymbol{\beta}} \ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0)^\top (\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}}) \tag{2}$$

The second step of approximation is used to calculate $(\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}})$ efficiently. Defining:

$$\widetilde{\boldsymbol{\beta}}(\epsilon) := \sum_{j=1}^{n} \ell(\boldsymbol{\beta}, \mathbf{z}_j) - \epsilon \ell(\beta, \mathbf{z}_i)$$

Note that $\widetilde{\boldsymbol{\beta}}(1) = \widehat{\boldsymbol{\beta}}_{/i}$. Since n is a large number, we have:

$$\widehat{\boldsymbol{\beta}}_{/i} - \widehat{\boldsymbol{\beta}} \approx - \left. \frac{d\widetilde{\boldsymbol{\beta}}(\epsilon)}{d\epsilon} \right|_{\epsilon=0}. \tag{3}$$

The derivative $\left. \frac{d\widetilde{\boldsymbol{\beta}}(\epsilon)}{d\epsilon} \right|_{\epsilon=0}$ can be calculated using the Hessian of the empirical risk, i.e. $\mathbf{G} = \sum_{i=1}^{n} \nabla^2_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \mathbf{z}_i)$:

$$\left. \frac{d\widehat{\boldsymbol{\beta}}}{d\epsilon} \right|_{\epsilon=0} = -\mathbf{G}^{-1} \nabla_{\boldsymbol{\beta}} \ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i). \tag{4}$$

Combining (2) and (4), Koh & Liang (2017) proposed the following approximation for $\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0)$:

$$\mathcal{I}^{\text{IF}}(\mathbf{z}_i, \mathbf{z}_0) = \nabla_{\boldsymbol{\beta}} \ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0)^\top \mathbf{G}^{-1} \nabla_{\boldsymbol{\beta}} \ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i). \tag{5}$$

Inspired by the analysis offered in Hampel (1974) for low dimensional settings, i.e. the setting, where the number of parameters p is much smaller than the number of observations n ($p \ll n$), it is widely believed that the approximations we mentioned in (2) and (3) are accurate. However, many modern AI and ML models have many parameters, that challenges the assumption $p \ll n$. Throughout the paper, we call the models in which p is not much smaller than n, high-dimensional models. Inspired by such models we would like to answer the following question:

$\mathcal{Q}_1$: Are the two approximations that led to (5) accurate under the high-dimensional settings?

$\mathcal{Q}_2$: If the answer to $\mathcal{Q}_1$ is negative, can the formula presented in (5) be improved to yield an accurate approximation for $\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0)$?

A few empirical papers have reported the inaccuracies in the conclusions of the IF; Koh & Liang (2017) and Bae et al. (2022) empirically showed that IFs are often inaccurate in estimating leave-one-out scores, particularly when applied to deep neural network models. Basu et al. (2020) studied how IFs change across different model parameterizations and regularization techniques, showing they can be erroneous in some circumstances. Schioppa et al. (2023) examined the five major sources of inaccuracy and highlighted the potential pitfalls of the Taylor expansion that is commonly used in IFs.

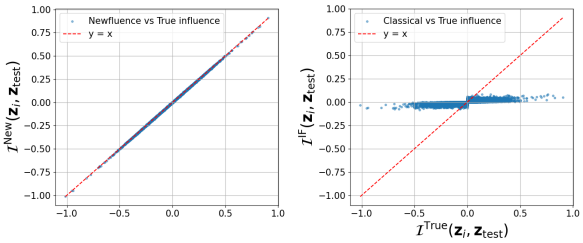

Figure 1. Comparison of Newfluence and $\mathcal{I}^{\text{IF}}(\mathbf{z}_i, \mathbf{z}_0)$. We use logistic ridge regression with $n = 500$, $p = 1000$, and $\lambda = 0.01$. The other details of the simulation are presented in Section 3 The figure shows results for all the influence of all the $n = 500$ training points on the prediction loss of $m = 100$ unseen new test points generated from the true logistic model. **Left:** Newfluence vs. true influence. **Right:** $\mathcal{I}^{\text{IF}}(\mathbf{z}_i, \mathbf{z}_0)$ vs. true influence. The Newfluence-based method offers a significantly better approximation.

Our goal is to develop a theoretical framework for analyzing the accuracy of IFs. We argue that the high dimensionality of modern AI and ML models undermines the approximations on which IFs are based. To support this claim, we adopt a high-dimensional asymptotic regime where $n, p \to \infty$ with $n/p \to \gamma$, for some fixed γ (Donoho et al., 2011; Zheng et al., 2017; Donoho & Montanari, 2016; El Karoui et al., 2013; Sur et al., 2019; Li & Wei, 2021). That is, both n and p are large, but their ratio remains bounded. Our theoretical results show that the answer to $\mathcal{Q}_1$ is negative, i.e. IFs can be inaccurate in high-dimensional settings. In response, we propose an alternative approximation of $\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_0)$, called Newfluence, that retains the computational simplicity of classical IFs but remain accurate under high-dimensional conditions. Empirical results further support our theoretical findings. Figure 1 compares the performance of Newfluence with that of $\mathcal{I}^{\text{IF}}(\mathbf{z}_i, \mathbf{z}_0)$.

1.2. Notations

Vectors and matrices are represented with boldfaced lower and upper case letters respectively, e.g. $\mathbf{a} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}$. For matrix $\mathbf{X}$, $\sigma_{\min}(\mathbf{X})$, $\sigma_{\max}(\mathbf{X})$, $\text{tr}(\mathbf{X})$ denote its minimum and maximum singular values and trace respectively. We denote any polynomials of $\log(n)$ by $\text{PolyLog}(n)$.

We use classic notations for deterministic and stochastic limit symbols such as $o(\cdot), O(\cdot), o_P(\cdot)$ and $O_P(\cdot)$. In addition, we use the notation $X = \Theta_P(1)$ if $X = O_P(1)$ but not $o_P(1)$, and $X = \Theta_P(a_n)$ if and onlf if $X/a_n = \Theta_P(1)$.

2. Theoretical results

2.1. Newfluence for High dimensional R-ERM

In this section, we formalize our ideas using the generalized linear model and propose our new measure of influence, called **Newfluence**. Consider the generalized linear model: $\mathbf{z}_i = (y_i, \mathbf{x}_i)$, where $y_i \in \mathbb{R}$ is the response and $\mathbf{x}_i \in \mathbb{R}^p$ is the feature. We further assume that a given dataset $D_n := \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ consists of n i.i.d. observations from a generalized linear model, i.e. $(y, \mathbf{x}) \sim p(\mathbf{x})q(y|\mathbf{x}^\top \boldsymbol{\beta}^*)$, and $\boldsymbol{\beta}^*$ represents the parameters that the ML system needs to learn.

We use the following regularized empirical risk minimization (R-ERM) for estimate $\boldsymbol{\beta}^*$:

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} L_n(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell(y_j, \mathbf{x}_j^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}), \quad (6)$$

where with a slight abuse of notation, we have redefined the loss function $\ell(y, u)$ as a function of y and $u = \mathbf{x}_j^\top \boldsymbol{\beta}$. Examples of the loss include square loss $\frac{1}{2}(y-u)^2$, or negative log-likelihood $-\log q(y|\mathbf{x}^\top\boldsymbol{\beta} = u)$. Furthermore, $r : \mathbb{R}^p \to \mathbb{R}$ is the regularizer, e.g. LASSO: $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, ridge: $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$, and $\lambda > 0$ is the strength of regularization. In the rest of the paper, we use simplified notations: for $j \in \{0, 1, \ldots, n\}$, $\ell_j(\boldsymbol{\beta}) := \ell(y_j, \mathbf{x}_j^\top \boldsymbol{\beta})$, $\dot{\ell}_j(\boldsymbol{\beta}) = \frac{\partial}{\partial u}\ell(y_j, u)\big|_{u=\mathbf{x}_j^\top\boldsymbol{\beta}}$ and $\ddot{\ell}_j(\boldsymbol{\beta}) = \frac{\partial^2}{\partial u^2}\ell(y_j, u)\Big|_{u=\mathbf{x}_j^\top\boldsymbol{\beta}}$. Rewriting (1) and (5) under the setting of this section, we obtain:

$$\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_0) := \ell_0(\widehat{\boldsymbol{\beta}}_{/i}) - \ell_0(\widehat{\boldsymbol{\beta}}),$$

$$\mathcal{I}^{\mathrm{IF}}(\mathbf{z}_i, \mathbf{z}_0) := \dot{\ell}_0(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}).$$

In response to $\mathcal{Q}_1$ in Section 1.1, we study the error $|\mathcal{I}^{\mathrm{IF}}(\mathbf{z}_i, \mathbf{z}_0) - \mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_0)|$ under the asymptotic setting $n, p \to \infty$, while $n/p \to \gamma_0$, where γ_0 is a fixed (but arbitrary) number. Note that the assumption $n/p \to \gamma_0$ aims to cover high-dimensional problems. For instance, but choosing $\gamma < 1$, we it will even cover the situation where the number of features are less than the number of observations.

The classical expectation was for the IF approximation $\mathcal{I}^{\mathrm{IF}}(\mathbf{z}_i, \mathbf{z}_0)$ to closely match the true influence $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_0)$, making the difference negligible. However, our theoretical results in the next section show that, contrary to this expectation,

$$\mathcal{I}^{\mathrm{IF}}(\mathbf{z}_i, \mathbf{z}_0) = (1-H_{ii})\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_0) + O_p\left(\frac{\mathrm{PolyLog}(n)}{n}\right),$$

where $H_{ii} = (\mathbf{x}_i^\top \mathbf{G}^{-1}\mathbf{x}_i)\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})$. As will be clarified later, it follows from the definition of $\mathbf{G}$ and our high-dimensional framework that $H_{ii} = \Theta_P(1)$ and $0 \le H_{ii} \le 1$.

This implies that the commonly used $\mathcal{I}^{\mathrm{IF}}$ formula *underestimates* the true influence by a datapoint-dependent factor $(1 - H_{ii})$. Since $1 - H_{ii}$ varies with the datapoint $\mathbf{z}_i$, it is conceivable that for points with large true influence $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_0)$, the value of $1 - H_{ii}$ may be small—causing $\mathcal{I}^{\mathrm{IF}}(\mathbf{z}_i, \mathbf{z}_0)$ to be markedly lower and incorrectly suggesting that $\mathbf{z}_i$ is non-influential in the prediction.This underestimation phenomenon is also illustrated in Figure 1.

This motivates $\mathcal{Q}_2$ in Section 1.1: can we find a more accurate alternative? In response, we propose a modified influence function framework that addresses this limitation while retaining computational efficiency. First, note that the above calculation suggests the corrected estimator:

$$\tilde{\mathcal{I}}^{\mathrm{IF}} := \frac{1}{1 - H_{ii}}\mathcal{I}^{\mathrm{IF}}$$

which is in fact consistent for $\mathcal{I}^{\mathrm{True}}$. However, both $\mathcal{I}^{\mathrm{IF}}$ and $\tilde{\mathcal{I}}^{\mathrm{IF}}$ require computing the gradient $\nabla_{\boldsymbol{\beta}}\ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0)$ for each new test point $\mathbf{z}_0$. We avoid this in our proposal called **Newfluence**, which is given by

$$\mathcal{I}^{\mathrm{New}}(\mathbf{z}_i, \mathbf{z}_0) = \ell_0\left(\widehat{\boldsymbol{\beta}} + \frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}^{-1}\mathbf{x}_i}{1 - H_{ii}}\right) - \ell_0\left(\widehat{\boldsymbol{\beta}}\right). \quad (7)$$

Evaluating the loss function at any $\boldsymbol{\beta}$ and $\mathbf{z}_0$ should have similar computational demand as evaluating the gradient, so this method generally does not increase computational complexity. As we shall show in the next section, this estimator is consistent for the true influence $\mathcal{I}^{\mathrm{True}}$.

This alternative approach retains the interpretability advantages of IFs while eliminating unnecessary approximations and improving computational accuracy. This formula is inspired by the recent work on the literature of risk estimation in high-dimensional settings (Rahnama Rad & Maleki, 2020; Rahnama Rad et al., 2020; Auddy et al., 2024; Wang et al., 2018). We present the derivation of the formula in Appendix A.

2.2. Our main theoretical contributions: Smooth case

In this section, we formally state our main theoretical results. Before stating our results, we first review some of the assumptions we have made in our analysis. All these assumptions are mild, and have been shown to be satisfied by a large number of models (Rahnama Rad & Maleki, 2020; Auddy et al., 2024; El Karoui et al., 2013; Sur et al., 2019).

Assumption A1. *The regularizer is separable:*

$$r(\boldsymbol{\beta}) = \sum_{k=1}^{p} r_k(\beta_k).$$

Assumption A2. *Both the loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ and the regularizer $r : \mathbb{R}^p \to \mathbb{R}_+$ are twice differentiable.*

Assumption A3. *Both ℓ and r are proper convex, and r is ν-strongly convex in $\boldsymbol{\beta}$ for some constant $\nu > 0$.*

Assumption A4. *$\exists C, s > 0$ such that*

$$\max\{\ell(y,u), |\dot{\ell}(y,z)|, |\dddot{\ell}(y,z)|\} \leq C(1+|y|^s+|z|^s)$$

and that $\nabla^2 r(\boldsymbol{\beta}) = \operatorname{diag}[\ddot{r}_k(\beta_k)]_{k\in[p]}$ is $C_{rr}(n)$-Lipschitz (in Frobenius norm) in $\boldsymbol{\beta}$ for some $C_{rr}(n) = O(\mathrm{PolyLog}(n))$.

We also adopt the following assumptions on the data generation mechanism:

Assumption B1. *The feature vectors $\mathbf{x}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Furthermore, $\lambda_{\max}(\boldsymbol{\Sigma}) \leq \frac{C_X}{p}$, for some constant $C_X > 0$.*

Assumption B2. *$\mathbb{P}(|y_i| > C_y(n)) \leq q_n^{(y)}$ for some $C_y(n) = O(\mathrm{PolyLog}(n))$ and $q_n^{(y)} = o(n^{-1})$.*

Theorem 2.1. *Under assumptions A1,... A4, and B1,B2, we have*

1. $|\mathcal{I}^{\mathrm{New}} - \mathcal{I}^{\mathrm{True}}| = o_P(\frac{1}{n}\mathrm{PolyLog}(n))$

2. $|\mathcal{I}^{\mathrm{True}}| = O_P(\frac{1}{\sqrt{n}}\mathrm{PolyLog}(n))$,

3. $\mathcal{I}^{\mathrm{IF}} = (1-H_{ii})\mathcal{I}^{\mathrm{True}} + o_P(\frac{1}{n}\mathrm{PolyLog}(n))$.

We present the proof in Appendix B. We now present a few remarks on the above theorem.

Remark 2.2. *For many common choices of loss functions and regularization parameters, $\dot{\ell}_i(\widehat{\boldsymbol{\beta}}) = \Theta_P(1)$, and $\max_j \ddot{r}_j(\widehat{\boldsymbol{\beta}}_j) = O_p(1)$. Then it follows from the definition of H_{ii} and our assumptions that $H_{ii} = \Theta_P(1/(1+\lambda))$. Thus the classical estimator $\mathcal{I}^{\mathrm{IF}}$ incurs considerable bias in such situations, when λ is not very large. This is reflected in our simulation experiments in the next section.*

Remark 2.3. *Comparing $\mathcal{I}^{\mathrm{True}}$ and $\mathcal{I}^{\mathrm{IF}}$, we find that the approximation error in $\mathcal{I}^{\mathrm{IF}}$ is $H_{ii} \cdot \mathcal{I}^{\mathrm{True}}$, which, for moderate regularization strength λ, is of the same order as the true influence itself. As a result, a highly influential data point may appear non-influential due to approximation errors involved in obtaining $|\mathcal{I}^{\mathrm{IF}}|$, potentially leading to misleading conclusions—an issue also noted in prior empirical studies (Basu et al., 2020; Bae et al., 2022).*

Remark 2.4. *In contrast to $\mathcal{I}^{\mathrm{IF}}$, the error $|\mathcal{I}^{\mathrm{New}} - \mathcal{I}^{\mathrm{True}}|$ is much smaller than $|\mathcal{I}^{\mathrm{True}}|$, indicating that $\mathcal{I}^{\mathrm{New}}$ provides a reliable approximation of $\mathcal{I}^{\mathrm{True}}$ when n and p are large.*

3. Numerical Experiments

We evaluate the accuracy of Newfluence and classical influence function in ℓ_2-regularized logistic regression. We generate synthetic binary classification datasets with feature dimensions $p \in \{500, 1000, 2000\}$ and set the sample size to maintain a fixed ratio $n/p = 0.5$. The true model coefficients $\boldsymbol{\beta}^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ are drawn from a standard normal distribution, and input features are sampled as $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p/n)$. Labels $y \in \{0,1\}^n$ are generated according to a Bernoulli model with success probabilities given by $\sigma(\mathbf{x}^\top \boldsymbol{\beta}^\star)$, where σ denotes the sigmoid function. The code used to produce the numerical results is available online[1].

We fit a logistic ridge model using Newton's method, with regularization parameter $\lambda \in \{0.01, 10\}$. For each model, we compute the true influence function $\mathcal{I}^{\mathrm{True}}$ for $m = 100$ unseen test examples. These true influence values are compared against two first-order approximations: (i) **Newfluence**, and (ii) the classical influence function, $\mathcal{I}^{\mathrm{IF}}$. For each test point, we evaluate the rank correlation between true and approximated influence across training points using Kendall's τ.

The value of τ lies in the interval $[-1, 1]$, with $\tau = 1$ indicating perfect agreement between rankings, $\tau = -1$ indicating perfect disagreement (reversed order), and $\tau = 0$ indicating no correlation in pairwise orderings.

Results are summarized in Tables 1 and 2. When $\lambda = 0.01$, corresponding to an effective degrees of freedom ratio $\mathrm{df}/p \approx 0.344$ (where $\mathrm{df} := \sum_{i=1}^n H_{ii}$), Newfluence estimates exhibit nearly perfect rank agreement with the true influence ranking ($\tau \approx 1.00$), while classical IFs are notably less accurate ($\tau \approx 0.88$). In contrast, when $\lambda = 10$, the regularization is strong and the effective degrees of freedom is substantially smaller ($\mathrm{df}/p \approx 0.023$), placing the model in a low-dimensional regime. In this setting, both Newfluence and the classical influence estimates achieve perfect agreement with the true influence ranking.

These results illustrate that classical IFs can be accurate in the low-dimensional regime—i.e., when $\mathrm{df}/p \ll 1$—but fail to match the fidelity of Newfluence in high-dimensional settings where the model complexity (as measured by df/p) is non-negligible.

4. Concluding Remarks

Interpreting today's black-box systems—ranging from foundation models to the latent "world models" that power model-based reinforcement learning—requires attribution tools whose guarantees scale with model dimensionality. In this work we have shown that the classical influence-function approximation of (Koh & Liang, 2017) incurs a systematic, datapoint-specific bias in high dimensions: it underestimates the true leave-one-out effect by a factor $1 - H_{ii}$. To remedy this, we proposed NEWFLUENCE, a single-Newton-step correction that preserves the computational economy of influence functions while eliminating

[1] https://anonymous.4open.science/r/corrected-influence-functions-2F7E/

Table 1. Kendall's τ (std in parentheses) for Newfluence and classical influence-based estimates across different training set sizes computed over $m = 100$ unseen test data points for logistic ridge when $\lambda = 0.01$, leading to $\mathrm{df}/p = 0.344$.

n	p	τ (Newfluence)	τ (IF)
250	500	0.99 (0.00)	0.88 (0.01)
500	1000	1.00 (0.00)	0.88 (0.01)
1000	2000	1.00 (0.00)	0.88 (0.00)

Table 2. Kendall's τ (std in parentheses) for Newfluence and classical influence-based estimates across different training set sizes computed over $m = 100$ unseen test data points for logistic ridge when $\lambda = 10.00$, leading to $\mathrm{df}/p = 0.023$.

n	p	τ (Newfluence)	τ (IF)
250	500	1.00 (0.00)	1.00 (0.00)
500	1000	1.00 (0.00)	1.00 (0.00)
1000	2000	1.00 (0.00)	1.00 (0.00)

their high-dimensional bias. Our theory establishes consistency under the proportional asymptotics $p \asymp n$, and experiments with high-dimensional logistic ridge regression confirm near-perfect rank agreement between NEWFLUENCE and ground truth, whereas the classical estimator degrades markedly.

Although our theoretical analysis presently targets generalized linear models with strongly convex, twice-differentiable objectives, the Newton-step correction is model-agnostic. We therefore view NEWFLUENCE as a first step toward principled influence estimation in the non-convex, non-smooth, and sequential settings that characterize modern AI models.

We hope that recognizing and correcting the high-dimensional bias documented here will encourage the community to reassess existing interpretability tools and to design influence-aware debugging, data-valuation, and alignment pipelines that scale with the complexity of contemporary AI systems.

References

Auddy, A., Zou, H., Rahnama Rad, K., and Maleki, A. Approximate leave-one-out cross validation for regression with l1 regularizers. *IEEE Transactions on Information Theory*, 70(11):8040–8071, 2024.

Bae, J., Ng, N., Lo, A., Ghassemi, M., and Grosse, R. B. If influence functions are the answer, then what is the question? *Advances in Neural Information Processing Systems*, 35:17953–17967, 2022.

Basu, S., Pope, P., and Feizi, S. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

Donoho, D. and Montanari, A. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166: 935–969, 2016.

Donoho, D. L., Maleki, A., and Montanari, A. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 110 (36):14557–14562, 2013.

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

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Guu, K., Webson, A., Pavlick, E., Dixon, L., Tenney, I., and Bolukbasi, T. Simfluence: Modeling the influence of individual training examples by simulating training runs. *arXiv preprint arXiv:2303.08114*, 2023.

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

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

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Jalali, S. and Maleki, A. New approach to bayesian high-dimensional linear regression. *Information and Inference: A Journal of the IMA*, 7, 07 2016.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 06–11 Aug 2017.

Kwon, Y. and Zou, J. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *International Conference on AI and Statistics*, 2022.

Kwon, Y., Wu, E., Wu, K., and Zou, J. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models, 2024. URL https://arxiv.org/abs/2310.00902.

Li, Y. and Wei, Y. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

Min, T., Lee, H., Ryu, H., Kwon, Y., and Lee, K. Understanding impact of human feedback via influence functions. *arXiv preprint arXiv:2501.05790*, 2025.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113. PMLR, 2023.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

Rahnama Rad, K. and Maleki, A. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4):965–996, 2020.

Rahnama Rad, K., Zhou, W., and Maleki, A. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, pp. 4067–4077. PMLR, 2020.

Rozemberczki, B., Watson, L., Bayer, P., Yang, H.-T., Kiss, O., Nilsson, S., and Sarkar, R. The shapley value in machine learning. In *The 31st International Joint Conference on Artificial Intelligence and the 25th European Conference on Artificial Intelligence*, pp. 5572–5579. International Joint Conferences on Artificial Intelligence Organization, 2022.

Schioppa, A., Filippova, K., Titov, I., and Zablotskaia, P. Theoretical and practical perspectives on what influence functions do. *Advances in Neural Information Processing Systems*, 36:27560–27581, 2023.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Sur, P., Chen, Y., and Candès, E. J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability theory and related fields*, 175:487–558, 2019.

Wang, J. T., Mittal, P., Song, D., and Jia, R. Data shapley in one training run. *arXiv preprint arXiv:2406.11011*, 2024a.

Wang, J. T., Song, D., Zou, J., Mittal, P., and Jia, R. Capturing the temporal dependence of training data influence. *arXiv preprint arXiv:2412.09538*, 2024b.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*, pp. 5228–5237. PMLR, 2018.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Zhang, H., Zhang, Z., Zhang, Y., Zhai, Y., Peng, H., Lei, Y., Yu, Y., Wang, H., Liang, B., Gui, L., et al. Correcting large language model behavior via influence function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14477–14485, 2025.

Zheng, L., Maleki, A., Weng, H., Wang, X., and Long, T. Does ℓ_p-minimization outperform ℓ_1-minimization? *IEEE Transactions on Information Theory*, 63(11):6896–6935, 2017.

Zou, H., Auddy, A., Kwon, Y., Rahnama Rad, K., and Maleki, A. Certified data removal under high-dimensional settings. *arXiv preprint arXiv:2505.07640*, 2025. URL https://arxiv.org/abs/2505.07640.

A. Derivation of the Newfluence for smooth problems

Here we provide the detailed derivation of $\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}$. The key is to use one step of Newton method to get an approximation of $\widehat{\boldsymbol{\beta}}_{/i}$, and using Woodbury formula to reduce computational complexity of the leave-one-out Hessian.

Recall that the Newton method, or Newton-Raphson method, is an iterative algorithm to find the root of a function.

Definition A.1 (Newton-Raphson Method)**.** *Given a function $\mathbf{f} : \mathbb{R}^p \to \mathbb{R}^p$ with a unique root $\mathbf{x}^* \in \mathbb{R}^p$, the Newton-Raphson Method finds $\mathbf{x}^*$ iteratively, starting from an initial point $\mathbf{x}^{(0)}$:*

$$\mathbf{x}^{(t)} := \mathbf{x}^{(t-1)} - \mathbf{G}^{-1}(\mathbf{x}^{(t-1)})\mathbf{f}(\mathbf{x}^{(t-1)}),$$

where $\mathbf{G}(\mathbf{x})$ is the Jacobian of f, which is assumed to exist and invertible.

Note that when the objective function $L_{n,/i}(\boldsymbol{\beta})$ is smooth, $\widehat{\boldsymbol{\beta}}_{/i}$ is the root of its gradient:

$$\mathbf{0} = \nabla L_{n,/i}(\widehat{\boldsymbol{\beta}}_{/i}) = \sum_{j\neq i} \dot{\ell}_j(\widehat{\boldsymbol{\beta}}_{/i})\mathbf{x}_j + \lambda \nabla r(\widehat{\boldsymbol{\beta}}_{/i}).$$

If we replace $\mathbf{f}$ in Definition A.1, then its Jacobian is just the Hessian of $L_{n,/i}$:

$$\mathbf{G}_{/i}(\widehat{\boldsymbol{\beta}}) = \sum_{j\neq i} \mathbf{x}_j \mathbf{x}_j^\top \ddot{\ell}_j(\widehat{\boldsymbol{\beta}}) + \lambda \nabla^2 r(\widehat{\boldsymbol{\beta}}). \tag{8}$$

Moreover, since $L_{n,/i} = L_n - \ell_i$ and $\nabla L_n(\widehat{\boldsymbol{\beta}}) = 0$, we have

$$\nabla L_{n,/i}(\widehat{\boldsymbol{\beta}}) = \nabla L_n(\widehat{\boldsymbol{\beta}}) - \nabla \ell_i(\widehat{\boldsymbol{\beta}}) = -\ell_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_i.$$

Inspired by the corresponding literature (e.g. (Rahnama Rad & Maleki, 2020; Rahnama Rad et al., 2020; Auddy et al., 2024)), we initiate the Newton iteration at the full model parameter $\widehat{\boldsymbol{\beta}}$ and apply **a single update**, and call the result $\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}$:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} &:= \widehat{\boldsymbol{\beta}} - \mathbf{G}_{/i}(\widehat{\boldsymbol{\beta}})\nabla L_{n,/i}(\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{\beta}} + \dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}_{/i}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i. \end{aligned} \tag{9}$$

Furthermore, we use Lemma C.1 to simplify the calculation of $\mathbf{G}_{/i}$ without repeatedly taking inverse for each i. For now we write $\mathbf{G}, \mathbf{G}_{/i}$ and drop their dependence on $\widehat{\boldsymbol{\beta}}$:

$$\begin{aligned}\mathbf{G}_{/i}^{-1} &= (\mathbf{G} - \nabla^2 \ell_i(\widehat{\boldsymbol{\beta}}))^{-1} \\ &= (\mathbf{G} - \mathbf{x}_i \mathbf{x}_i^\top \ddot{\ell}_i(\widehat{\boldsymbol{\beta}}))^{-1} \\ (\text{ Woodbury Formula }) \quad &= \mathbf{G}^{-1} + \frac{\mathbf{G}^{-1}\mathbf{x}_i^\top \mathbf{x}_i \mathbf{G}^{-1} \ddot{\ell}_i(\widehat{\boldsymbol{\beta}})}{1 - \mathbf{x}_i^\top \mathbf{G}^{-1}\mathbf{x}_i \ddot{\ell}_i(\widehat{\boldsymbol{\beta}})}. \end{aligned}$$

Inserting this back into (9):

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} &= \widehat{\boldsymbol{\beta}} + \dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}_{/i}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i \\ &= \widehat{\boldsymbol{\beta}} + \dot{\ell}_i(\widehat{\boldsymbol{\beta}}) \left[\mathbf{G}^{-1} + \frac{\mathbf{G}^{-1}\mathbf{x}_i^\top \mathbf{x}_i \mathbf{G}^{-1} \ddot{\ell}_i(\widehat{\boldsymbol{\beta}})}{1 - \mathbf{x}_i^\top \mathbf{G}^{-1}\mathbf{x}_i \ddot{\ell}_i(\widehat{\boldsymbol{\beta}})}\right] \mathbf{x}_i \\ &= \widehat{\boldsymbol{\beta}} + \dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}^{-1}\mathbf{x}_i \frac{1}{1 - H_{ii}} \end{aligned} \tag{10}$$

where H_{ii} is the (i,i) element of

$$\mathbf{H} = \mathbf{X}\mathbf{G}^{-1}\mathbf{X}^\top \operatorname{diag}[\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})]_{i=1}^n,$$

and $\mathrm{diag}[\ddot{\ell}_i(\widehat{\boldsymbol{\beta}})]_{i=1}^n$ is the diagonal matrix with diagonal elements being $\{\ddot{\ell}_i(\widehat{\boldsymbol{\beta}}), i = 1, 2, ..., n\}$. Replacing $\widehat{\boldsymbol{\beta}}_{/i}$ by $\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}$ in the definition of the true influence $\mathcal{I}^{\text{True}}$ yields our definition of $\mathcal{I}^{\text{New}}$:

$$\begin{aligned}\mathcal{I}^{\text{New}} &= \ell_0(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}) - \ell_0(\widehat{\boldsymbol{\beta}}) \\ &= \ell_0\left(\widehat{\boldsymbol{\beta}} + \dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}_{/i}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i\frac{1}{1-H_{ii}}\right) - \ell_0\left(\widehat{\boldsymbol{\beta}}\right)\end{aligned}$$

A.1. A Heuristic Lower Bound on H_{ii}

We show that H_{ii} is bounded below by a constant with non-vanishing probability in the case of linear regression, i.e., an ERM with squared loss.

$$\begin{aligned}H_{ii} &= \mathbf{x}_i^\top \mathbf{G}^{-1}(\widehat{\boldsymbol{\beta}})\mathbf{x}_i \\ &\geq \|\mathbf{x}_i\|^2 \sigma_{\min}(\mathbf{G}^{-1}(\widehat{\boldsymbol{\beta}})) \\ &\geq \|\mathbf{x}_i\|^2 (\sigma_{\max}(\mathbf{G}(\widehat{\boldsymbol{\beta}})))^{-1}.\end{aligned}$$

By standard concentration results for χ^2 distribution (Lemma C.2), we know that $\mathbb{P}(\|\mathbf{x}_i\|^2 > \frac{1}{2}) \geq 1 - e^{-\frac{1}{16}p}$. Also, we have

$$\begin{aligned}\sigma_{\max}(\mathbf{G}(\widehat{\boldsymbol{\beta}})) &= \|\mathbf{X}^\top\mathbf{X} + \lambda\nabla^2 r(\boldsymbol{\beta})\| \\ &\leq \|\mathbf{X}\|^2 + \lambda\|\nabla^2 r(\boldsymbol{\beta})\|.\end{aligned}$$

By Lemma C.4, $\|\mathbf{X}\| = O_P(1)$. We claim without proof that, for most commonly used regularizers we have $\|\nabla^2 r(\boldsymbol{\beta})\| = O_P(1)$. For example, for the ridge penalty $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$ we have $\|\nabla^2 r(\boldsymbol{\beta})\| = 2$. Therefore $H_{ii} > C$ w.p $\to 1$.

B. Proof of Theorem 2.1

B.1. Proof of Part 1

It follows from Lemma 3.3 of (Zou et al., 2025) (with $m = t = 1$) that $\|\widehat{\boldsymbol{\beta}}_{/i} - \tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}\| = o_P(\frac{1}{\sqrt{n}}\text{PolyLog}(n))$. Thus

$$\begin{aligned}\mathcal{I}^{\text{New}} - \mathcal{I}^{\text{True}} &= \ell_0(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}) - \ell_0(\widehat{\boldsymbol{\beta}}_{/i}) \\ &= \ell(y_0, \mathbf{x}_0^\top\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}) - \ell(y_0, \mathbf{x}_0^\top\widehat{\boldsymbol{\beta}}_{/i}) \\ &= \left(\int_0^1 \nabla_{\boldsymbol{\beta}}\ell(y_0, \mathbf{x}_0^\top(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} + t(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i})))dt\right)^\top (\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i}) \\ &= \left(\int_0^1 \dot{\ell}_0(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} + t(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i}))dt\right)\mathbf{x}_0^\top(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i}) \\ &\leq C\left(1 + |y_0|^s + |\mathbf{x}_0^\top\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}|^s + |\mathbf{x}_0^\top\widehat{\boldsymbol{\beta}}_{/i}|^s\right)\mathbf{x}_0^\top(\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i}) \\ &\leq 4C\sqrt{C_X}\left(1 + (C_y(n))^s + (3\|\widehat{\boldsymbol{\beta}}_{/i}\|\sqrt{C_X\log(n)/p})^s\right)\|\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}} - \widehat{\boldsymbol{\beta}}_{/i}\|\sqrt{\log(n)/p} \\ &= O_p\left(\frac{\text{PolyLog}(n)}{n}\right) \qquad (11)\end{aligned}$$

with high probability. Here the first inequality follows from Assumption A4. The second inequality holds with high probability by Assumption B2 and by Lemma C.2, since $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\lambda_{\max}(\boldsymbol{\Sigma}) \leq C_X/p$, and $\mathbf{x}_0$ is independent of $\{(y_i, \mathbf{x}_i) : 1 \leq i \leq n\}$, and hence of $\widehat{\boldsymbol{\beta}}_{/i}, \tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}$. In the last step we use the fact that $\|\widehat{\boldsymbol{\beta}}_{/i}\| = O_P(\sqrt{p})$, and finally the above quoted bound on the error of the Newton step, i.e., $\|\widehat{\boldsymbol{\beta}}_{/i} - \tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}\| = o_P(\frac{1}{\sqrt{n}}\text{PolyLog}(n))$.

B.2. Proof of Part 2

Next, note that by definition of $\mathcal{I}^{\text{New}}$, we have by a Taylor series expansion that

$$\begin{aligned}\mathcal{I}^{\text{New}} - \frac{\dot{\ell}_0(\widehat{\boldsymbol{\beta}})\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}} &= \ell_0\left(\widehat{\boldsymbol{\beta}} + \frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}}\right) - \ell_0(\widehat{\boldsymbol{\beta}}) - \frac{\dot{\ell}_0(\widehat{\boldsymbol{\beta}})\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}} \\ &= \left(\frac{\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}}\right)^2 \ddot{\ell}_0(\xi) \\ &= O_p(\text{PolyLog}(n)) \cdot (\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i)^2 = O_P\left(\frac{\text{PolyLog}(n)}{n}\right) \end{aligned} \tag{12}$$

Here $\xi = t\widehat{\boldsymbol{\beta}} + (1-t)\tilde{\boldsymbol{\beta}}_{/i}^{\text{Newton}}$ for some $t \in [0,1]$. To get the second last equality we use Lemma C.4 to conclude that $\|\mathbf{G}^{-1}\mathbf{x}_i\| = O_P(1)$, Assumptions A4 and B2 to arrive at $(1-H_{ii})^{-1} = O_P(1)$, $\dot{\ell}_0(\widehat{\boldsymbol{\beta}}), \dot{\ell}_i(\widehat{\boldsymbol{\beta}}), \ddot{\ell}_i(\xi) = O_P(\text{PolyLog}(n))$. To get the last equality we again use the conclusion that $\|\mathbf{G}^{-1}\mathbf{x}_i\| = O_P(1)$, and finally that $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\lambda_{\max}(\boldsymbol{\Sigma}) = C_X/p$, independent of $\{(y_i, \mathbf{x}_i) : 1 \le i \le n\}$. Note also that $n/p \to \gamma_0$.

From (11) and (12) we then have

$$\mathcal{I}^{\text{True}} - \frac{\dot{\ell}_0(\widehat{\boldsymbol{\beta}})\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}} = O_P\left(\frac{\text{PolyLog}(n)}{n}\right). \tag{13}$$

By arguments identical to (12) we have $(1-H_{ii})^{-1} = O_P(1)$, $\dot{\ell}_0(\widehat{\boldsymbol{\beta}}), \dot{\ell}_i(\widehat{\boldsymbol{\beta}}), \ddot{\ell}_i(\xi) = O_P(\text{PolyLog}(n))$, and $\|\mathbf{G}^{-1}\mathbf{x}_i\| = O_P(1)$, and thus

$$\frac{\dot{\ell}_0(\widehat{\boldsymbol{\beta}})\dot{\ell}_i(\widehat{\boldsymbol{\beta}})\mathbf{x}_0^\top \mathbf{G}^{-1}\mathbf{x}_i}{1-H_{ii}} = O_P\left(\frac{\text{PolyLog}(n)}{\sqrt{n}}\right).$$

This completes the proof of part 2. In fact, for many choices of loss functions, such as squared loss, logistic loss, or Poisson negative log likelihood the above quantity is in fact $\Theta_P\left(\frac{\text{PolyLog}(n)}{\sqrt{n}}\right)$. We omit a more detailed analysis here.

B.3. Proof of Part 3

Since we are in the setup of generalized linear models,

$$\nabla_{\boldsymbol{\beta}}\ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_0) = \mathbf{x}_0\dot{\ell}_0(\widehat{\boldsymbol{\beta}}), \quad \nabla_{\boldsymbol{\beta}}\ell(\widehat{\boldsymbol{\beta}}, \mathbf{z}_i) = \mathbf{x}_i\dot{\ell}_i(\widehat{\boldsymbol{\beta}}).$$

Thus, definition of $\mathcal{I}^{\text{IF}}$ and (13) together imply that

$$\mathcal{I}^{\text{True}} - \frac{\mathcal{I}^{\text{IF}}}{1-H_{ii}} = O_P\left(\frac{\text{PolyLog}(n)}{n}\right).$$

from where the conclusion of Part 2 follows immediately since $0 < 1 - H_{ii} \le 1$.

C. Auxiliary Lemmata

Lemma C.1 (Woodbury Inversion Formula)**.** *Suppose* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is nonsingular, and* $\mathbf{M} = \mathbf{A} + \mathbf{UBV}$*, then*

$$\mathbf{M}^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{B}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

provided that all relevant inverse matrices exist.

Lemma C.2 (Lemma 6 of Jalali & Maleki (2016))**.** *Let* $\mathbf{z} \sim N(0, \mathbb{I}_p)$*, then*

$$\mathbb{P}(\mathbf{z}^\top\mathbf{z} \ge p + pt) \le e^{-\frac{p}{2}(t - \log(1+t))}$$

Lemma C.3 (Lemma 12 in (Rahnama Rad & Maleki, 2020))**.** $\mathbf{X} \in \mathbb{R}^{p\times p}$ *is composed of independently distributed* $N(0, \boldsymbol{\Sigma})$ *rows, with* $\rho_{\max} = \sigma_{\max}(\Sigma)$*, where* $\Sigma \in \mathbb{R}^{p\times p}$*. Then*

$$\mathbb{P}(\|\mathbf{X}^\top\mathbf{X}\| \ge (\sqrt{n} + 3\sqrt{p})^2\rho_{\max}) \le \mathrm{e}^{-p}.$$

Lemma C.4. *Suppose* $\mathbf{X}_{n\times p}$ *has iid rows* $\mathbf{x}_i \sim N(0, p^{-1}\mathbb{I}_p)$*, then*

1. $\mathbb{P}(\max_{1\le i\le n} \|\mathbf{x}_i\| \ge 2) \le ne^{-p/2}$
2. $\mathbb{P}(\|\mathbf{X}^\top\mathbf{X}\| \ge (\sqrt{\gamma_0}+3)^2) \le e^{-p}$

Proof. 1. By Lemma C.2 and let $\mathbf{z} = n^{-1/2}\mathbf{x}_i$ we have $\mathbf{z} \sim N(0, \mathbb{I}_p)$ so that

$$\begin{aligned}\mathbb{P}(\|\mathbf{x}_i\| \ge 2) &= \mathbb{P}(\mathbf{x}_i^\top \mathbf{x}_i \ge 4)\\ &\le \mathbb{P}(\mathbf{z}^\top\mathbf{z} \ge 4p)\\ &\le e^{-\frac{p}{2}(3-\log(4))} \le e^{-p/2}\end{aligned}$$

The rest follows from a union bound over all i.

2. It is a direct application of C.3 with $\rho_{\max} = p^{-1}$.

□