Understanding What TRAK Gets Right and Wrong: A Theory of Data Attribution

Han Tong
Department of Statistics
Columbia University
New York, NY
ht2672@columbia.edu

Haolin Zou
Department of Statistics
Columbia University
New York, NY

Shubhangi Ghosh
Department of Statistics
Columbia University
New York, NY

Arian Maleki
Department of Statistics
Columbia University
New York, NY
mm4338@columbia.edu

Abstract

Data attribution, which traces a model's prediction back to specific training data, is an important tool for interpreting sophisticated AI models. The widely used TRAK algorithm addresses this challenge by first approximating the underlying model with a kernel machine and then leveraging techniques developed for approximating the leave-one-out (ALO) risk. Despite its strong empirical performance, the theoretical conditions under which TRAK approximations are accurate, as well as the regimes in which they break down, remain largely unexplored. In this paper, we provide a theoretical analysis of the TRAK algorithm, characterizing its performance and quantifying the errors introduced by the approximations on which the method relies. We show that although the approximations can incur significant errors, TRAK preserves the separation between highly influential and weakly influential data points. We corroborate our theoretical results through extensive simulations and empirical studies.

1 Introduction

1.1 Related work and contributions

Modern machine learning systems, including large language models, rely heavily on massive training datasets, where data quality significantly impacts model performance. This dependence has motivated growing interest in *data attribution*, the task of tracing a model's predictions back to individual training samples, to better understand model behavior, identify harmful or redundant data, and improve data efficiency.

A popular class of data attribution methods is based on counterfactual reasoning: asking how a model's prediction would change if a particular training point were removed or reweighted. Popular approaches rely on quantities such as the Shapley value [Jia et al., 2019, Ghorbani and Zou, 2019, Wang et al., 2025] and influence functions [Hampel, 1974, Sagun et al., 2018, Yeh et al., 2019, Han et al., 2020, Hammoudeh and Lowd, 2022, Ilyas et al., 2022, Fisher et al., 2023, Park et al., 2023, Guu et al., 2023, Hammoudeh and Lowd, 2024], whose computation is often prohibitively expensive or even intractable for modern AI models. Recently, the TRAK algorithm [Park et al., 2023] was proposed as a scalable data attribution method tailored to modern neural networks and has since

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

gained popularity [Ye et al., 2023, Xia et al., 2024, Hammoudeh and Lowd, 2024]. Its computational efficiency stems from three key approximations of the influence functions: linearization, random projection, and Approximate Leave-One-Out (ALO) corrections [Rad and Maleki, 2020, Auddy et al., 2024]. TRAK has demonstrated strong empirical performance on large models and datasets.

Despite its practical success, little is known about the theoretical conditions under which TRAK provides accurate attributions, or about the error introduced by its multiple approximation steps. This lack of theoretical understanding limits our ability to determine when and for what tasks TRAK can be reliably used, and what information is lost by its approximations.

This paper aims to close this gap. We provide the first systematic theoretical analysis of TRAK by evaluating the errors each approximation step of TRAK introduces. To give a high-level overview of our contributions, let $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ denote the true influence of data point $\mathbf{z}_i$ on a test data point $\mathbf{z}_{\mathrm{new}}$ [Koh and Liang, 2017]. This quantity will be defined more formally in (2). Furthermore, let $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ denote the TRAK approximation of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$.

Scope of our contributions. Our main theoretical results are proved in the underparameterized regime, i.e., $d < n$, where d denotes the number of parameters and n denotes the number of training samples; these quantities are defined formally below. Appendix A.2 provides a preliminary analysis of a ridge-regularized formulation for $d > n$, rather than a complete extension of our main theorems to the overparameterized regime.

Our analysis reveals the following facts about TRAK:

Negative result: While the error introduced by the ALO approximation is typically small and benign, both the linearization step and the projection step, especially when the number of projections is much smaller than the number of parameters, can introduce substantial errors. As a result, the exact ranking of training data points induced by $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ can differ significantly from that induced by $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$. Our empirical results are consistent with our theoretical results and confirm that TRAK does **not** preserve the exact rankings offered by $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$.

Positive result: Despite the large errors TRAK introduces, it can detect data points that have large influences from those that have small influences. Informally speaking, if the influence of data point $\mathbf{z}_i$, i.e. $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ is much larger than the $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_j, \mathbf{z}_{\mathrm{new}})$, then the $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ remains much larger than $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_j, \mathbf{z}_{\mathrm{new}})$ as well. Hence, when the objective is to identify the most influential data points, TRAK provides accurate and reliable information.

1.2 TRAK influence function

In this section, we formally present the three approximations employed by the TRAK algorithm. Consider the standard supervised learning setting, in which we are given a dataset

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\},$$

where $y_i \in \mathbb{R}$ is the response, $\mathbf{x}_i \in \mathbb{R}^p$ denotes the corresponding feature vector, and $\mathbf{z}_i := (y_i, \mathbf{x}_i)$. The goal is to learn a predictor that can accurately predict the response associated with a new feature vector $\mathbf{x}_{\mathrm{new}}$. Many successful approaches in AI and machine learning, including generalized linear models and neural networks, use predictors of the form $f(\mathbf{x}; \boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathbb{R}^d$ are parameters (to be learned from data), and $f(\cdot; \boldsymbol{\beta})$ is a linear or nonlinear function that produces the predictions. In most applications, the model parameters are estimated using empirical risk minimization:

$$\hat{\boldsymbol{\beta}} \triangleq \underset{\boldsymbol{\beta} \in \mathbb{R}^d}{\arg\min} \left\{ \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i; \boldsymbol{\beta})) \right\}, \tag{1}$$

where $\ell(y, f(\mathbf{x}; \boldsymbol{\beta}))$ denotes the loss function, such as the squared loss or cross-entropy loss, depending on the application and the nature of the response variable.

To interpret model predictions, many data attribution techniques seek to quantify the influence of individual training samples on a model's output. Specifically, the influence of a training data point $\mathbf{z}_i = (y_i, \mathbf{x}_i)$ on the model's prediction at a new data point $\mathbf{z}_{\mathrm{new}} = (y_{\mathrm{new}}, \mathbf{x}_{\mathrm{new}})$ is defined as

$$\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) \triangleq f(\mathbf{x}_{\mathrm{new}}; \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_{\mathrm{new}}; \hat{\boldsymbol{\beta}}). \tag{2}$$

where

$$\hat{\boldsymbol{\beta}}_{/i} \triangleq \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{j\neq i} \ell\big(y_j, f(\mathbf{x}_j;\boldsymbol{\beta})\big) \right\}. \tag{3}$$

A fundamental challenge in applying (2) for data valuation is its prohibitive computational cost. Computing the influence of a single training sample requires solving the leave-one-out optimization problem in (3). In modern machine learning and AI applications, where both the sample size n and the model dimension d are large, the exact computation of $\hat{\boldsymbol{\beta}}_{/i}$ becomes computationally prohibitive, especially when it must be performed repeatedly for many training data points. The **TRAK** algorithm mitigates this issue by approximating $\hat{\boldsymbol{\beta}}_{/i}$ using the following sequence of steps:

Step 1: Linearization. The first step in TRAK is to linearly approximate $f(\mathbf{x}, \boldsymbol{\beta})$ around $\hat{\boldsymbol{\beta}}$. Define

$$\boldsymbol{g}_j = \nabla f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}), \quad \boldsymbol{b}_j = f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}) - \nabla f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})^\top \hat{\boldsymbol{\beta}}.$$

Instead of solving the nonlinear optimization problem (3), TRAK solves the following linearized version:

$$\breve{\boldsymbol{\beta}} \triangleq \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{n} \ell\big(y_i, \mathbf{g}_i^\top \boldsymbol{\beta} + \mathbf{b}_i\big) \right\}, \quad \breve{\boldsymbol{\beta}}_{/i} \triangleq \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{j\neq i} \ell\big(y_j, \mathbf{g}_j^\top \boldsymbol{\beta} + \mathbf{b}_j\big) \right\}. \tag{4}$$

In the Appendix, we further show that under our assumptions, $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. The intuition is that the linear approximation of $f(\mathbf{x}, \boldsymbol{\beta})$ around $\hat{\boldsymbol{\beta}}$ provides a sufficiently accurate surrogate, so the corresponding optimization problem changes only slightly. As a result, when $\boldsymbol{\beta}$ is close to $\hat{\boldsymbol{\beta}}$, we can use the linear function of $\{\mathbf{g}_i\}_{i\in[n]}$ as an approximation to $f(\mathbf{x}_i, \boldsymbol{\beta})$.

Step 2: Random projection. To further reduce computational cost, TRAK applies random projection to the gradients $\mathbf{g}_i$. The idea is to reduce dimensionality while approximately preserving inner products. More specifically, we define the gradient matrix used in the TRAK algorithm as:

$$\mathbf{G}(\boldsymbol{\beta}) := \big(\nabla f(\mathbf{x}_1, \boldsymbol{\beta}), \ldots, \nabla f(\mathbf{x}_n, \boldsymbol{\beta})\big)^\top, \tag{5}$$

and for notational simplicity, we write $\mathbf{G} := \mathbf{G}(\hat{\boldsymbol{\beta}})$. Using a random projection matrix $\mathbf{S} \in \mathbb{R}^{d\times k}$ with $k \ll d$, TRAK defines the feature map $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{S}^\top \mathbf{g}(\mathbf{x})$, where $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x}, \hat{\boldsymbol{\beta}})$. Accordingly, we define $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)^\top \in \mathbb{R}^{n\times k}$, where $\boldsymbol{\phi}_i = \mathbf{S}^\top \mathbf{g}_i$ and $\boldsymbol{\phi}_{\text{new}} = \mathbf{S}^\top \mathbf{g}_{\text{new}}$.

Step 3: Approximate Leave-One-Out (ALO). Building on the ideas of approximate leave-one-out [Rad and Maleki, 2020, Beirami et al., 2017], TRAK estimates the influence via the ALO formula [1], which provides the following approximation for $\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_{\text{new}})$:

$$\mathcal{I}^{\text{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\text{new}}; k) \triangleq \frac{\dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \boldsymbol{\phi}_{\text{new}}^\top \big(\Phi^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\Phi\big)^{-1} \boldsymbol{\phi}_i}{1 - \ddot{\ell}_i(\breve{\boldsymbol{\beta}})\, \boldsymbol{\phi}_i^\top \big(\Phi^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\Phi\big)^{-1} \boldsymbol{\phi}_i}, \tag{6}$$

where k is the projection dimension, $\dot{\ell}(y, z)$ and $\ddot{\ell}(y, z)$ denote the first- and second-order derivatives of the loss function with respect to z, respectively; $\dot{\ell}_i(\breve{\boldsymbol{\beta}}) = \dot{\ell}(y_i, \mathbf{g}_i^\top \breve{\boldsymbol{\beta}} + \mathbf{b}_i)$, $\ddot{\ell}_i(\breve{\boldsymbol{\beta}}) = \ddot{\ell}(y_i, \mathbf{g}_i^\top \breve{\boldsymbol{\beta}} + \mathbf{b}_i)$, $\dot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}}) = [\dot{\ell}_1(\breve{\boldsymbol{\beta}}), \dot{\ell}_2(\breve{\boldsymbol{\beta}}), \ldots, \dot{\ell}_n(\breve{\boldsymbol{\beta}})]^\top$ and $\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}}) = [\ddot{\ell}_1(\breve{\boldsymbol{\beta}}), \ddot{\ell}_2(\breve{\boldsymbol{\beta}}), \ldots, \ddot{\ell}_n(\breve{\boldsymbol{\beta}})]^\top$.

We should also note that TRAK includes two additional steps: (i) ensemble averaging, which mitigates variability arising from random initialization and the inherent randomness of the optimization algorithms used to minimize the training loss, and (ii) soft thresholding of the influence estimates, which introduces shrinkage under the assumption that only a small subset of training points meaningfully affects the prediction for a given test example. Since our main analysis focuses on the accuracy of TRAK's core approximation steps, we discuss these two additional steps in Appendix A.3.

[1] In the TRAK paper, the authors note that replacing $\operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]$ with the identity matrix and ignoring the denominator in (6) does not significantly affect empirical performance. As a result, they use the simplified expression

$$\dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \boldsymbol{\phi}_{\text{new}}^\top \big(\Phi^\top \Phi\big)^{-1} \boldsymbol{\phi}_i.$$

However, for theoretical analysis, we will consider the complete form in (6).

2 Main theoretical contributions

2.1 Our theoretical framework

Throughout the theoretical part of the paper we assume that the elements of $\mathcal{D}$ are independent draws from the joint distribution

$$(y_i, \mathbf{x}_i) \sim q(y_i \mid f(\mathbf{x}_i; \boldsymbol{\beta}^*))\, p(\mathbf{x}_i). \tag{7}$$

Here, $\boldsymbol{\beta}^* \in \mathbb{R}^d$ denotes the true parameter vector. Unlike classical linear models, the parameter dimension d may differ from the feature dimension p; for example, in neural networks, d corresponds to the total number of weights. We define the loss function as the negative log-likelihood

$$\ell(y, f(\mathbf{x}; \boldsymbol{\beta})) = -\log q(y \mid f(\mathbf{x}; \boldsymbol{\beta})).$$

Motivated by modern machine learning and AI applications, where the models have billions or even trillions of parameters and are trained on gigabytes of data, in our theoretical studies we focus on the regime that the sample size n, the number of features p, and number of parameters d are all large. In other words, in our analysis we will ignore the errors that are negligible for large values of n, p and d.

In addition to this assumption, we impose several structural conditions to characterize the scaling behavior of the approximation errors introduced by TRAK. We give an informal summary below and defer the formal statements to Appendix B.1.2.

Assumption 2.1 (Data and model regularity)**.** We assume the following throughout the paper.

(A1) Sub-Gaussian design. The rows of $\boldsymbol{X}_n \in \mathbb{R}^{n \times p}$ are independent mean-zero sub-Gaussian random vectors with covariance $\boldsymbol{\Sigma}_p$. Moreover, for constants $c, C > 0$,

$$c\|\boldsymbol{\beta}^*\|^{-2} \leq \lambda_{\min}(\boldsymbol{\Sigma}_p) \leq \lambda_{\max}(\boldsymbol{\Sigma}_p) \leq C\|\boldsymbol{\beta}^*\|^{-2}.$$

Appendix B.1.2 provides further intuition for the scaling.

(A2) Model and loss regularity. We assume standard smoothness conditions on the prediction map f and the loss ℓ. Informally, we assume that $\|\nabla f(\mathbf{x}, \boldsymbol{\beta}^*)\|$ has sub-Gaussian tails with bounded second moment; the empirical Hessians associated with (1) and (4) are well-conditioned near $\hat{\boldsymbol{\beta}}$; and ℓ is strongly convex, with ℓ and $\nabla\ell$ being Lipschitz continuous. Formal statements are given in Appendix B.1.2.

In general, these are standard regularity conditions for high-dimensional M-estimation. Regarding the assumptions on the loss, Appendix B.5 verifies the loss regularity conditions for common objectives, including the squared loss, cross-entropy loss, and the Poisson loss with a softplus link. The regularity conditions on f also hold on a wide range of models. We mention two canonical examples below.

Example 2.2 (Linear model)**.** For generalized linear models with

$$f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^\top \boldsymbol{\beta},$$

the model assumptions in Appendix B.1.2 hold; see Appendix B.6.1 for the proof.

Example 2.3 (Neural network)**.** For a one-hidden-layer neural network

$$f(\mathbf{x}, \mathbf{W}, \mathbf{v}) = \mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x}),$$

where $\mathbf{W} \in \mathbb{R}^{h \times p}$, $\mathbf{v} \in \mathbb{R}^h$, and σ has bounded derivative, the model assumptions in Appendix B.1.2 hold; see Appendix B.6.2 for the proof.

2.2 Theoretical results

2.2.1 Overview of the results

Our theoretical contributions are presented in four sections:

• Section 2.2.2 characterizes the size of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ in two settings: (1) when $\mathbf{z}_i$ is related to $\mathbf{z}_{\mathrm{new}}$, in which case the influence of $\mathbf{z}_i$ on $\mathbf{z}_{\mathrm{new}}$ is expected to be large; and (2) when $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are unrelated, in which case the influence of $\mathbf{z}_i$ on $\mathbf{z}_{\mathrm{new}}$ is expected to be small. These order-of-magnitude characterizations will be used to interpret the properties of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$.

• Section 2.2.3 studies the error introduced by the linearization step in TRAK and highlights the key limitations of this approximation, and also its intriguing features.

• Section 2.2.4 studies the error that is introduced by the ALO approximation in TRAK, and shows that the error that is introduced in this step is benign and substantially smaller than the errors that are introduced in the linearization and projection steps of TRAK.

• Section 2.2.5 studies the error that is introduced by the projection step in TRAK and shows the important limitations of the projection step, especially when the dimension of the projection is much smaller than the number of parameters. Table 1 summarizes all the theoretical contributions of our

Table 1: Summary of the theoretical results of the paper regarding the size of the influence function and magnitudes of errors introduced at each step of TRAK. The displayed orders hold with high probability, with n^{ϵ} and $\mathrm{poly}(\log \mathrm{n})$ factors omitted.

Step	**Quantity / Result**	**Correlated case: $\mathbf{z}_{\mathrm{new}} = \mathbf{z}_i$**	**Uncorrelated case: $\mathbf{z}_{\mathrm{new}} \perp \mathbf{z}_i$**
True influence	$\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$	Proposition 2.4: $\frac{\|\boldsymbol{\beta}^*\|^2}{n}$	Proposition 2.5: $\frac{\|\boldsymbol{\beta}^*\|}{n}$
Linearization bound	$\mathcal{I}^{\mathrm{Linear}} - \mathcal{I}^{\mathrm{True}}$	Theorem 2.6: $\frac{\|\boldsymbol{\beta}^*\|^2}{n}$	Theorem 2.9: $\frac{\|\boldsymbol{\beta}^*\|}{n}$
ALO bound	$\mathcal{I}^{\mathrm{ALO}} - \mathcal{I}^{\mathrm{Linear}}$	Theorem 2.12: $\frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2}}$	Theorem 2.13: $\frac{\|\boldsymbol{\beta}^*\|}{n^{3/2}}$
Projection bound	$\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}; k)$	Theorem 2.15: $\frac{\|\boldsymbol{\beta}^*\|^2}{n}\frac{k}{d}$	Theorem 2.15: $\frac{\|\boldsymbol{\beta}^*\|}{n}\sqrt{\frac{k}{d}}$

paper.

2.2.2 Size of influence

In this section, we study the magnitude of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$. Our first result obtains an upper bound on the influence function and shows its sharpness.

Proposition 2.4. *Under Assumption 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant C^{True} such that, with probability tending to 1,*

$$\sup_{i\in[n]} \left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})\right| \;\le\; C^{\mathrm{True}} \frac{\|\boldsymbol{\beta}^*\|^2 \mathrm{poly}(\log \mathrm{n})}{n},$$

where $\mathrm{poly}(\log \mathrm{n})$ denotes a polynomial function of $\log(n)$. Furthermore, if Assumption B.8 holds, then there exists an absolute constant $\tilde{C}^{\mathrm{True}}$ such that, with probability tending to 1,

$$\inf_{i\in[n]} \left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_i)\right| \;\ge\; \tilde{C}^{\mathrm{True}} \frac{\|\boldsymbol{\beta}^*\|^2}{n}. \tag{8}$$

The proof of this proposition is presented in Appendix B.2.1. This theorem shows how large an influence of a data point on $\mathbf{z}_{\mathrm{new}}$ can be. But to understand this result better, we have to discuss how large $\left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})\right|$ can be when $\mathbf{z}_{\mathrm{new}}$ is independent of the training set. In such cases, we expect the influence to be small. Our next theorem addresses this question.

Proposition 2.5. *Suppose that $\mathbf{z}_{\mathrm{new}}$ is independent of the training set $\mathcal{D}$. Under Assumption 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant C^{True} such that, with probability tending to 1,*

$$\sup_{i\in[n]} \left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})\right| \;\le\; C^{\mathrm{True}} \frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}}.$$

The proof of Proposition 2.5 is shown in Appendix B.2.2.

Since n, p, and d are all assumed to be large, we expect $\|\boldsymbol{\beta}^*\|$ to be much larger than 1. Comparing the results of the two theorems, we observe a substantial gap between the influence of data points $\mathbf{z}_i$ in the correlated and independent cases. As we clarify in the next sections, this large separation is in fact one of the reasons behind the empirical success of the TRAK method.

2.2.3 Linearization step

In this section, we aim to understand the error introduced by the linearization step in TRAK. If projection and ALO steps are removed from TRAK, then the approximate influence function becomes

$$\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) \triangleq \mathbf{g}_{\mathrm{new}}^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}), \tag{9}$$

where $\mathbf{g}_{\mathrm{new}} = \nabla f(\mathbf{x}_{\mathrm{new}}, \hat{\boldsymbol{\beta}})$, and $\breve{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}_{/i}$ are defined in (4).

Following our approach from the previous section, we study $|\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) - \mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})|$ in two cases: (i) the correlated case, where $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$, and (ii) the uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent. We begin with the correlated case.

Theorem 2.6. *Under Assumptions 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant C^{Linear} such that, with probability tending to 1, the linearization error satisfies:*

$$\sup_{i\in[n]} \left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_i) - \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i)\right| \le C^{\mathrm{Linear}} \frac{\|\boldsymbol{\beta}^*\|^2}{n}.$$

Furthermore, if Assumption B.8 holds, then there exists an absolute constant c^{Linear} such that, with probability tending to 1,

$$\inf_{i\in[n]} |\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i)| \ge c^{\mathrm{Linear}} \frac{\|\boldsymbol{\beta}^*\|^2}{n}. \tag{10}$$

The proof can be found in Appendix B.2.3. We highlight several points concerning this theorem.

Remark 2.7. While this theorem obtains an upper bound on the difference $\left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_i) - \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i)\right|$, our arguments presented in Appendix C.2 show that this upper bound is sharp, and the order of the error is in fact $\Theta_p(\frac{\|\boldsymbol{\beta}^*\|^2}{n})$.

Remark 2.8. If we compare the error term $C^{\mathrm{Linear}} \frac{\|\boldsymbol{\beta}^*\|^2}{n}$ with the order of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_i)$ in Proposition 2.4, we notice that the error is very large and has the same order as the quantity that we wanted to approximate. This is consistent with the "Negative result" that we reported in the introduction.

Now to show some of the intriguing features of the linearization, our next theorem establishes the accuracy of $\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ for the case that $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent.

Theorem 2.9. *Suppose that $\mathbf{z}_{\mathrm{new}}$ is independent of the training set $\mathcal{D}$. Under Assumptions 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant $\tilde{C}^{\mathrm{Linear}}$ such that, with probability tending to 1, the linearization error satisfies*

$$\sup_{i\in[n]} \left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) - \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})\right| \le \tilde{C}^{\mathrm{Linear}} \frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}}.$$

The proof of Theorem 2.9 is shown in Appendix B.2.4.

Remark 2.10. Again, by comparing the above theorem with Proposition 2.5, we observe that the error

$$\left|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) - \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})\right|$$

is of the same order as $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ itself. Consequently, this approximation does not yield quantitatively accurate estimates of $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$. This observation is consistent with the negative result highlighted in the introduction.

Remark 2.11. Despite the negative results discussed in Remarks 2.8 and 2.10, the TRAK approximation still exhibits an intriguing positive property. Specifically, when $\mathbf{z}_{\mathrm{new}} = \mathbf{z}_i$, the quantity $\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ remains large, as shown by the second part of Proposition 2.4 and the second part of Theorem 2.6. In contrast, when a data point $\mathbf{z}_j$ is independent of $\mathbf{z}_{\mathrm{new}}$, Proposition 2.5 and Theorem 2.9 imply that $\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_j, \mathbf{z}_{\mathrm{new}})$ is of a much smaller order. As a result, even though the TRAK approximation may incur substantial absolute error, it can still reliably distinguish data points with large true influence $\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ from those with negligible influence. This is consistent with the "positive result" mentioned in the introduction.

2.2.4 ALO step

In this section, we characterize the error introduced by the ALO approximation. If we ignore the projection step of the TRAK method ($k = d$), the influence function produced by TRAK reduces to

$$\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) \triangleq \frac{\dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_{\mathrm{new}}^\top \mathbf{H}^{-1} \mathbf{g}_i}{1 - \ddot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_i^\top \mathbf{H}^{-1} \mathbf{g}_i}, \tag{11}$$

where $\mathbf{H} = \mathbf{G}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\mathbf{G}$, $\dot{\ell}_i(\boldsymbol{\beta}) = \dot{\ell}(y_i, \mathbf{g}_i^\top \boldsymbol{\beta} + \mathbf{b}_i)$ and $\ddot{\ell}_i(\boldsymbol{\beta}) = \ddot{\ell}(y_i, \mathbf{g}_i^\top \boldsymbol{\beta} + \mathbf{b}_i)$. In this section we would like to characterize the difference between $\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) - \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$. The next two theorems study this error for both correlated and independent cases.

Theorem 2.12. *Under Assumptions 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant $\tilde{C}^{\mathrm{ALO}}$ such that, with probability tending to 1, the ALO step error satisfies:*

$$\sup_{i \in [n]} \left| \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i) - \mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) \right| \le \tilde{C}^{\mathrm{ALO}} \frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}}.$$

The proof of Theorem 2.12 is postponed to Appendix B.2.5.

Theorem 2.13. *Suppose that $\mathbf{z}_{\mathrm{new}}$ is independent of the training set $\mathcal{D}$. Under Assumptions 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant C^{ALO} such that, with probability tending to 1, the ALO step error satisfies:*

$$\sup_{i \in [n]} \left| \mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) - \mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) \right| \le C^{\mathrm{ALO}} \frac{\|\boldsymbol{\beta}^*\|}{n^{1.5-\epsilon}}.$$

The proof of Theorem 2.13 is postponed to Appendix B.2.6.

Remark 2.14. Combining the above two theorems, we can conclude that in both cases, we have

$$\frac{\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})}{\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})} = 1 + o_p(1).$$

Hence, the error introduced by the ALO step is negligible compared to the magnitude of the influence.

2.2.5 Projection step

Finally, we study the impact of the projection step on the accuracy of the TRAK algorithm.

Theorem 2.15. *Suppose that $\mathbf{z}_{\mathrm{new}}$ is independent of the training set $\mathcal{D}$. Under Assumptions 2.1 (formally, Assumptions B.1–B.7 in the Appendix), for any $\epsilon > 0$ such that $\|\boldsymbol{\beta}^*\|^2 \le n^{1-\epsilon}$, there exists an absolute constant C^{TRAK} such that, with probability tending to 1, the projection error satisfies*

$$\sup_{i \in [n]} \left| \mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}; k) \right| \;\le\; C^{\mathrm{TRAK}} \frac{\|\boldsymbol{\beta}^*\| \sqrt{k}}{n^{1-\epsilon} \sqrt{d}}.$$

Furthermore, there is an absolute constant c^{TRAK} such that with probability tending to 1,

$$\inf_{i \in [n]} \left| \frac{\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i; k)}{\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \right| \;\ge\; c^{\mathrm{TRAK}} \frac{k}{d}.$$

The proof is presented in Appendix B.2.7. Figure 6 in Appendix C.2 confirms the sharpness of these orders.

Remark 2.16. The projection dimension k controls whether TRAK can preserve the separation between dependent and independent data points. By Theorem 2.6 and Theorem 2.12, in the dependent case $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$, the ALO influence is of the order $\gtrsim \frac{\|\boldsymbol{\beta}^*\|^2}{n^{1-\epsilon}}$. After random projection, Theorem 2.15 shows that this quantity is reduced by a factor of order k/d. Hence, $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i; k) \gtrsim \frac{\|\boldsymbol{\beta}^*\|^2 k}{n^{1-\epsilon} d}$.

The relevant comparison is with the independent case, where $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent. In that case, the projected influence is of a smaller intrinsic order before projection, but it is reduced only by

a factor of order $\sqrt{k/d}$ rather than k/d. Therefore, when k/d is very small, the dependent influence can be compressed more severely than the independent influence. As a result, the separation between the two regimes may disappear. In particular, Theorem 2.15 implies that if $\|\boldsymbol{\beta}^*\|\sqrt{k/d} \ll 1$, then the projected dependent influence $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i; k)$ can be no larger than the projected influence for an independent test point $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}; k)$. In this regime, TRAK may fail to distinguish truly influential training points from independent ones. Thus, to preserve the separation needed for reliable ranking, the projection dimension should satisfy $k \gg \frac{d}{\|\boldsymbol{\beta}^*\|^2}$.

2.3 Discussion of our theoretical results

Throughout this paper, we evaluate the accuracy of TRAK in two distinct settings: (i) Correlated case: the test data point is identical to one of the training samples, and (ii) Independent case: the test data point is independent of all training samples.

These represent two extreme scenarios, which naturally raises the question of how TRAK behaves in intermediate regimes. It turns out that more general cases can be analyzed as interpolations between the two extremes we have studied above. Appendix B.4 and Appendix C.4 provide further details on this point.

The main conclusion of Appendix B.4 is that, under mild conditions, if the correlation between $\mathbf{x}_{\mathrm{new}}$ and $\mathbf{x}_i$ is α_n, then the magnitude of the influence function scales by a factor of α_n. Note that this claim is true as long as $\alpha_n I^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_i) = \Omega(I^{\mathrm{True}}(\mathbf{z}_i, \tilde{\mathbf{z}}_{\mathrm{new}}))$ for independent $\tilde{\mathbf{z}}_{\mathrm{new}}$. Similarly, the errors introduced at each approximation step of TRAK are also scaled by the same factor. In other words, as long as correlations are strong enough, TRAK can distinguish correlated data points from independent ones.

3 Simulation results

In this section, we focus on nonlinear multi-class classification. We also perform simulations for binary classification and Poisson regression. However, we present the latter two experiments in Appendix C.1.1.

Our main goal is to evaluate whether $\mathcal{I}^{\mathrm{TRAK}}$ provides reliable rankings compared with $\mathcal{I}^{\mathrm{True}}$. We compare the uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent, with the correlated case, where $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$, and examine whether this order-level separation is preserved through the three TRAK approximation steps: linearization, ALO, and random projection. To save space, we present the full experimental setup in Appendix C.1.2. We should emphasize that the multi-class model is inherently nonlinear (See Appendix A.1 for more details).

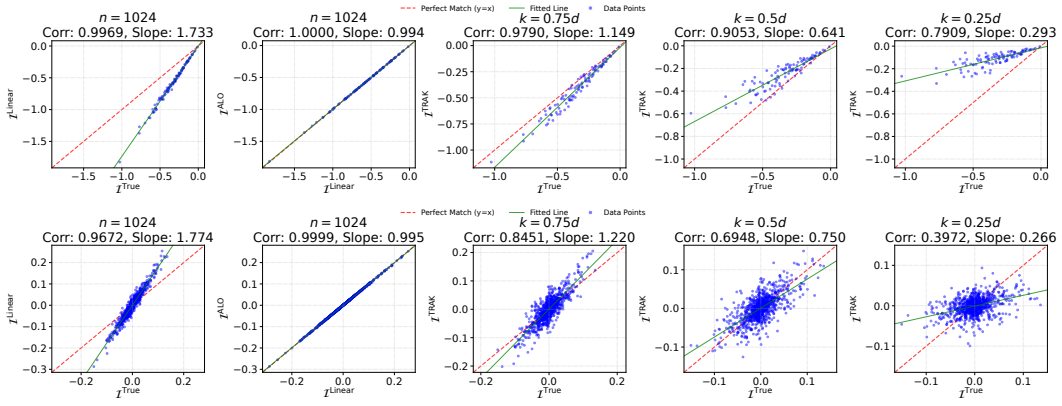

Figure 1: Experimental results for 3-class classification with $p = 100$ and $d = 200$. **Top row:** Correlated case, where $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$. **Bottom row:** Uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent. **First column:** $\mathcal{I}^{\mathrm{Linear}}$ versus $\mathcal{I}^{\mathrm{True}}$. **Second column:** $\mathcal{I}^{\mathrm{ALO}}$ versus $\mathcal{I}^{\mathrm{Linear}}$. **Right three columns:** $\mathcal{I}^{\mathrm{TRAK}}(k)$ versus $\mathcal{I}^{\mathrm{True}}$ with $k = 75\%d, 50\%d, 25\%d$ ($k = 150, 100, 50$).

Linearization step. The first column of Figure 1 plots $\mathcal{I}^{\mathrm{Linear}}$ against $\mathcal{I}^{\mathrm{True}}$ in the correlated and uncorrelated cases. The visible deviation from the red $y = x$ line indicates a substantial absolute discrepancy between the two quantities, consistent with Theorems 2.6 and 2.9. Nevertheless, $\mathcal{I}^{\mathrm{Linear}}$ remains strongly aligned with $\mathcal{I}^{\mathrm{True}}$: data points with larger true influence also tend to have larger linearized influence. Moreover, the correlated case has a much larger scale than the uncorrelated case, matching the order gap predicted by Propositions 2.4 and 2.5, as well as Theorems 2.6 and 2.9. Thus, although the linearization step introduces non-negligible magnitude error, it largely preserves the separation structure needed for reliable ranking.

ALO step. The second column of Figure 1 plots $\mathcal{I}^{\mathrm{ALO}}$ against $\mathcal{I}^{\mathrm{Linear}}$. The two quantities are nearly indistinguishable in both regimes, showing that the ALO step introduces only negligible error. This agrees with Theorems 2.12 and 2.13. Therefore, the ALO step preserves the separation already present after linearization and is not the main source of ranking distortion in this experiment.

Projection step. The right three columns of Figure 1 show $\mathcal{I}^{\mathrm{TRAK}}(k)$ plotted against $\mathcal{I}^{\mathrm{True}}$ for $k = 0.75d$, $k = 0.50d$, and $k = 0.25d$. As predicted by Theorem 2.15, the magnitude of $\mathcal{I}^{\mathrm{TRAK}}(k)$ decreases as k/d decreases. Furthermore, smaller projection dimensions reduce the separation between the two cases; this effect is further quantified in Table 2. When $k/d = 0.25$, the separation becomes much less pronounced, suggesting that small projection dimensions may fail to preserve the separation between influential and independent data points, as discussed in Remark 2.16. Moreover, as k decreases, the correlation between $\mathcal{I}^{\mathrm{True}}$ and $\mathcal{I}^{\mathrm{TRAK}}(k)$ drops quickly, further showing that projection can damage performance, especially when the projection dimension is small.

Quantitative check. Table 2 further quantifies how the separation between the correlated and uncorrelated cases changes across the TRAK approximation steps. For $\mathcal{I}^{\mathrm{True}}$, $\mathcal{I}^{\mathrm{Linear}}$, and $\mathcal{I}^{\mathrm{ALO}}$, the separation remains stable: the correlated 5% quantile is still about 1.5 times larger than the uncorrelated 95% quantile. Thus, the true influence, linearized influence, and ALO approximation all maintain a clear magnitude gap between the two regimes.

In contrast, this separation weakens as the projection dimension k decreases. The most stringent ratio, Corr. 5%/Uncorr. 95%, decreases from 1.42 at $k = 0.75d$ to 0.89 at $k = 0.25d$, and further to 0.42 at $k = 0.1d$. Hence, under aggressive random projection, the lower tail of the correlated magnitudes can overlap with, or even fall below, the upper tail of the uncorrelated magnitudes. This is consistent with Theorem 2.15: in the correlated case, $\mathcal{I}^{\mathrm{TRAK}}(k)$ scales as k/d, whereas in the uncorrelated case, it scales as $\sqrt{k/d}$. Since $\sqrt{k/d}$ decreases more slowly than k/d, small projection dimensions reduce the separation between the two regimes and can therefore harm ranking performance. For example, comparing $\mathcal{I}^{\mathrm{TRAK}}(0.5d)$ with $\mathcal{I}^{\mathrm{TRAK}}(0.25d)$, the projection dimension is reduced by a factor of two. Correspondingly, the correlated 5% and 95% quantiles both decrease by roughly a factor of two, whereas the uncorrelated quantiles decrease by a smaller factor, close to the predicted $1/\sqrt{2}$ scaling.

Table 2: The 5% and 95% quantiles of absolute influence values in the correlated (Corr.) and uncorrelated (Uncorr.) cases for 3-class classification with $p = 100$ and $d = 200$. The last three columns report separation ratios. The ratio Corr. 5%/Uncorr. 95% is the most stringent diagnostic: values above one indicate clear separation between the lower tail of the correlated magnitudes and the upper tail of the uncorrelated magnitudes, while values below one indicate overlap between the two regimes.

Quantity	Corr. 5%	Corr. 95%	Uncorr. 5%	Uncorr. 95%	Corr. 5%/Uncorr. 5%	Corr. 95%/Uncorr. 95%	Corr. 5%/Uncorr. 95%
$\mathcal{I}^{\mathrm{True}}$	1.1097×10^{-1}	6.0677×10^{-1}	7.2078×10^{-4}	6.6671×10^{-2}	153.96	9.10	1.66
$\mathcal{I}^{\mathrm{Linear}}$	1.8015×10^{-1}	1.0876	1.9678×10^{-3}	1.2082×10^{-1}	91.55	9.00	1.49
$\mathcal{I}^{\mathrm{ALO}}$	1.8033×10^{-1}	1.0860	1.7744×10^{-3}	1.2088×10^{-1}	101.63	8.98	1.49
$\mathcal{I}^{\mathrm{TRAK}}(0.75d)$	1.3540×10^{-1}	7.5510×10^{-1}	1.2054×10^{-3}	9.5666×10^{-2}	112.34	7.89	1.42
$\mathcal{I}^{\mathrm{TRAK}}(0.5d)$	8.1372×10^{-2}	4.8031×10^{-1}	1.3459×10^{-3}	7.3058×10^{-2}	60.46	6.57	1.11
$\mathcal{I}^{\mathrm{TRAK}}(0.25d)$	4.1744×10^{-2}	2.3990×10^{-1}	9.4412×10^{-4}	4.6958×10^{-2}	44.21	5.11	0.89
$\mathcal{I}^{\mathrm{TRAK}}(0.1d)$	1.2220×10^{-2}	1.1411×10^{-1}	6.2861×10^{-4}	2.9211×10^{-2}	19.44	3.91	0.42

Additional empirical validation. We also conduct experiments on CIFAR-10 [Krizhevsky et al., 2014] to assess whether the conclusions above remain robust on real-world data that may not satisfy our data-generation assumptions. The results, deferred to Appendix C.3, show qualitatively similar behavior: the linearization step introduces non-negligible magnitude distortion but retains substantial rank information, the ALO step is benign, and the projection step can weaken ranking performance when the projection dimension is small.

4 Conclusion and future directions

TRAK is a widely used data attribution method that attains scalability through three approximations: linearization, approximate leave-one-out (ALO), and projection. In this work, we theoretically analyze the errors introduced by each step. We show that the ALO approximation incurs negligible error, while the linearization step introduces substantial error but still preserves the ability to distinguish highly influential data points from uninfluential ones. In contrast, the projection step can introduce significant additional error, and using too few projections may prevent TRAK from reliably identifying influential points. Overall, our results clarify both the strengths and limitations of TRAK and provide guidance for its effective use at scale.

Our formal results are developed for the underparameterized regime $d < n$. Appendix A.2 outlines how ridge regularization modifies the TRAK formulation when $d > n$. Nevertheless, a complete theoretical treatment of the overparameterized regime remains an important direction for future work.

Acknowledgments and Disclosure of Funding

Arian Maleki would like to declare the following NSF funding: NSF-DMS 2515716 and NSF-DMS 2210506.

The authors declare no competing interests.

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1958.

Arnab Auddy, Haolin Zou, Kamiar Rahnama Rad, and Arian Maleki. Approximate leave-one-out cross validation for regression with ℓ_1 regularizers. *IEEE Transactions on Information Theory*, 70 (11):8040–8071, 2024.

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Jillian Fisher, Lang Liu, Krishna Pillutla, Yejin Choi, and Zaid Harchaoui. Influence diagnostics under self-concordance. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10028–10076. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/fisher23a.html`.

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.

Kelvin Guu, Albert Webson, Ellie Pavlick, Lucas Dixon, Ian Tenney, and Tolga Bolukbasi. Simfluence: Modeling the influence of individual training examples by simulating training runs. *arXiv preprint arXiv:2303.08114*, 2023.

Zayd Hammoudeh and Daniel Lowd. Identifying a training-set attack's target using renormalized influence estimation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1367–1381, 2022.

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

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

Xiaochuang Han, Byron C. Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5553–5563, Online, July 2020. Association for Computational Linguistics.

doi: 10.18653/v1/2020.acl-main.492. URL https://aclanthology.org/2020.acl-main.492/.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

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.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 2017.

Alex Krizhevsky, Vinod Nair, Geoffrey Hinton, et al. The cifar-10 dataset, 2014.

S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 27074–27113. PMLR, 2023.

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

L. Sagun, U. Evci, V. U. Güney, Y. N. Dauphin, and L. Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR) Workshop Track*, 2018.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Jiachen T. Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data Shapley in one training run. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning (ICML)*, 2024.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in)fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Appendix

Organization of the Appendix

- **Section A** summarizes additional methodological details. In particular, in Appendix A.1 we discuss how to transform a multi-class classification problem into a nonlinear scalar-output function $f(\cdot, \cdot)$.
- **Section B** summarizes all theoretical and technical details of the paper. Section B.1 collects all assumptions used throughout the paper. We further verify that these assumptions hold for common GLM loss functions in Section B.5, and show that they are satisfied by linear models and one-layer neural networks in Section B.6. All proofs are provided in Section B.2, and all technical lemmas are collected in Section B.3. A roadmap of our main theorems is summarized in Table 3.
- **Section C** presents additional simulation results that further verify our theoretical findings. Section C.1 focuses on correlation-based analysis, while Section C.2 focuses on magnitude-based analysis. Section C.3 presents results on real-world empirical datasets. Finally, Section C.4 presents results for more general correlated settings.

Table 3: Roadmap of theoretical results in the Appendix.

		Influence Function Approximation			
		True Influence	**Linearization Step**	**ALO Step**	**Projection Step**
Theorem	$\mathbf{z}_i = \mathbf{z}_{\text{new}}$	Proposition 2.4	Theorem 2.6	Theorem 2.12	Theorem 2.15
	Independent	Proposition 2.5	Theorem 2.9	Theorem 2.13	
Proof	$\mathbf{z}_i = \mathbf{z}_{\text{new}}$	Appendix B.2.1	Appendix B.2.3	Appendix B.2.5	Appendix B.2.7
	Independent	Appendix B.2.2	Appendix B.2.4	Appendix B.2.6	

A More details

A.1 Multi-class classification extension

The goal of this section is to extend our analysis to the multiclass classification setting. We consider a dataset with K classes and employ the standard cross-entropy loss,

$$\mathcal{L}_{\mathrm{CE}}(\mathbf{z}, \boldsymbol{\beta}) = -\log p_y(\mathbf{z}, \boldsymbol{\beta}),$$

where $p_y(\mathbf{z}, \boldsymbol{\beta})$ denotes the model's predicted probability for the true class label y associated with the data point $\mathbf{z}$. Since we consider a multiclass classification problem, the model must produce class probabilities for all K classes. Even in the simplest setting, the model parameters therefore take the form

$$\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_{K-1}, \mathbf{W}_K)^\top,$$

where each $\mathbf{W}_k \in \mathbb{R}^p$ corresponds to the weight vector for class k. For identifiability, we fix $\mathbf{W}_K = \mathbf{0}_p$. We then define

$$\boldsymbol{\beta} := \mathrm{vec}\big(\mathbf{W}_{[1:(K-1),:]}\big) \in \mathbb{R}^d,$$

where the parameter dimension is $d = (K-1)p > p$; thus, this model is no longer a simple linear model, unlike the binary classification case. To address this complexity, we employ a method similar to that used in Park et al. [2023]. Specifically, we define:

$$f(\mathbf{z}, \boldsymbol{\beta}) = \log\left(\frac{p(\mathbf{z}, \boldsymbol{\beta})}{1 - p(\mathbf{z}, \boldsymbol{\beta})}\right), \;\; p(\mathbf{z}, \boldsymbol{\beta}) = \frac{e^{\mathbf{W}_y^\top \mathbf{x}}}{\sum_{j=1}^K e^{\mathbf{W}_j^\top \mathbf{x}}},$$

where we set $\mathbf{W}_K = \mathbf{0}_p$. The loss function can then be expressed as:

$$\mathcal{L}_{CE}(\mathbf{z}, \boldsymbol{\beta}) = \log(1 + e^{-f(\mathbf{z}, \boldsymbol{\beta})}) = l(f(\mathbf{z}, \boldsymbol{\beta})),$$

where $l(s) = \log(1 + e^{-s})$. We can further express the gradient as:

$$\mathbf{g} = \nabla_{\boldsymbol{\beta}} f(\mathbf{z}, \hat{\boldsymbol{\beta}}) = \frac{(\mathbf{1}_y - \mathbf{p}(\mathbf{z}, \hat{\boldsymbol{\beta}}))}{1 - p(\mathbf{z}, \hat{\boldsymbol{\beta}})} \otimes \mathbf{x},$$

where $\mathbf{1}_y$ is a vector of length $K-1$ with entries $\mathbf{1}_{y[k]} = 1$ if $y = k$ and 0 otherwise (note that if $y = K$, $\mathbf{1}_y$ is a zero vector). Similarly, $\mathbf{p}(\mathbf{z}, \boldsymbol{\beta}) = \left(\frac{\exp(\mathbf{W}_1^\top \mathbf{x})}{\sum_{j=1}^K \exp(\mathbf{W}_j^\top \mathbf{x})}, \ldots, \frac{\exp(\mathbf{W}_{K-1}^\top \mathbf{x})}{\sum_{j=1}^K \exp(\mathbf{W}_j^\top \mathbf{x})}\right)^\top$ is also a vector of length $K-1$. Finally, we obtain $\mathbf{g} \in \mathbb{R}^d$ with $d = (K-1)p$.

A.2 Overparameterized regime with ridge regularization

Our main analysis considers the underparameterized regime $d < n$. Here, we outline a ridge-regularized formulation for $d > n$ and indicate which parts of the preceding analysis may continue to apply. Consider

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{j=1}^n \ell\big(y_j, f(\mathbf{x}_j, \boldsymbol{\beta})\big) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|_2^2 \right\}, \qquad \lambda > 0.$$

Using the unnormalized convention, define

$$\boldsymbol{H}_\lambda = \mathbf{G}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_\lambda)]\mathbf{G} + n\lambda \mathbb{I}_d, \qquad \boldsymbol{H}_{\lambda,/i} = \boldsymbol{H}_\lambda - \ddot{\ell}_i(\hat{\boldsymbol{\beta}}_\lambda)\mathbf{g}_i\mathbf{g}_i^\top.$$

Ridge regularization ensures that these Hessians are invertible under the usual convexity conditions.

The regularized ALO and projected-ALO scores are

$$\mathcal{I}_\lambda^{\mathrm{ALO}} = \dot{\ell}_i(\hat{\boldsymbol{\beta}}_\lambda)\mathbf{g}_{\mathrm{new}}^\top \boldsymbol{H}_{\lambda,/i}^{-1}\mathbf{g}_i$$

and

$$\mathcal{I}_{\lambda,k}^{\mathrm{TRAK}} = \dot{\ell}_i(\hat{\boldsymbol{\beta}}_\lambda)\mathbf{g}_{\mathrm{new}}^\top \mathbf{S}\big(\mathbf{S}^\top \boldsymbol{H}_{\lambda,/i}\mathbf{S}\big)^{-1}\mathbf{S}^\top \mathbf{g}_i.$$

In particular, the projected Hessian contains the projected regularizer:

$$\mathbf{S}^\top \boldsymbol{H}_{\lambda,/i}\mathbf{S} = \boldsymbol{\Phi}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_\lambda)]\boldsymbol{\Phi}_{/i} + n\lambda \mathbf{S}^\top\mathbf{S}.$$

Consequently,

$$\mathcal{I}_{\lambda,k}^{\mathrm{TRAK}} - \mathcal{I}_\lambda^{\mathrm{ALO}} = \dot{\ell}_i(\hat{\boldsymbol{\beta}}_\lambda)\mathbf{g}_{\mathrm{new}}^\top \left[\mathbf{S}\big(\mathbf{S}^\top \boldsymbol{H}_{\lambda,/i}\mathbf{S}\big)^{-1}\mathbf{S}^\top - \boldsymbol{H}_{\lambda,/i}^{-1}\right]\mathbf{g}_i.$$

Existing results show that ALO can consistently approximate LOO for a broad class of regularized estimators under suitable conditions [Rad and Maleki, 2020]. Moreover, under the same conditioning and concentration assumptions used in our main analysis, the projection argument continues to produce the k/d scaling in the correlated case and the $\sqrt{k/d}$ scaling in the independent case. Therefore, although ridge regularization makes the relevant Hessians invertible when $d > n$, an excessively small projection dimension can still introduce substantial magnitude error and weaken score separation.

A.3 Ensemble averaging and soft thresholding

For the m-th submodel, write the complete projected-ALO score as $q_{m,i}a_{m,i}$, where

$$q_{m,i} = \dot{\ell}_{m,i}(\breve{\boldsymbol{\beta}}_m), \qquad a_{m,i} = \frac{\boldsymbol{\phi}_{m,\mathrm{new}}^\top \boldsymbol{H}_m^{-1}\boldsymbol{\phi}_{m,i}}{1 - \ddot{\ell}_{m,i}(\breve{\boldsymbol{\beta}}_m)\boldsymbol{\phi}_{m,i}^\top \boldsymbol{H}_m^{-1}\boldsymbol{\phi}_{m,i}}.$$

Direct averaging reports $M^{-1}\sum_{m=1}^M q_{m,i}a_{m,i}$, whereas practical TRAK [Park et al., 2023] averages the two factors separately and reports $\bar{q}_i\bar{a}_i$. The difference between these two quantities is

$$\bar{q}_i\bar{a}_i - \frac{1}{M}\sum_{m=1}^M q_{m,i}a_{m,i} = -\widehat{\mathrm{Cov}}_M(q_i, a_i),$$

where

$$\widehat{\mathrm{Cov}}_M(q_i, a_i) = \mathrm{Cov}(q_i, a_i) + O_p(M^{-1/2}).$$

Thus, the covariance bias does not generally vanish with M.

For direct averaging over independent submodels and projections, the correlated score is of order

$$\frac{\|\boldsymbol{\beta}^*\|^2}{n}\frac{k}{d},$$

whereas the independent score is of order

$$\frac{\|\boldsymbol{\beta}^*\|}{n}\sqrt{\frac{k}{dM}}.$$

Therefore, direct averaging relaxes the sufficient projection-dimension condition from

$$k \gg \frac{d}{\|\boldsymbol{\beta}^*\|^2} \qquad \text{to} \qquad Mk \gg \frac{d}{\|\boldsymbol{\beta}^*\|^2}.$$

For factorized averaging, although the covariance bias may not vanish, it does not introduce a new order. Indeed, $a_{m,i}$ contains the same projected inner-product structure analyzed in our main results. Its leading order is

$$\frac{\|\boldsymbol{\beta}^*\|^2}{n}\frac{k}{d}$$

in the correlated case and

$$\frac{\|\boldsymbol{\beta}^*\|}{n}\sqrt{\frac{k}{d}}$$

in the independent case, up to the polylogarithmic factors contributed by $q_{m,i}$. The covariance term therefore inherits the same order difference. Consequently, factorized averaging may introduce a non-vanishing magnitude bias, but it does not alter the order separation between the correlated and independent cases. However, because its covariance term does not generally vanish with M,

increasing M reduces the random fluctuations but not this persistent magnitude bias. This further strengthens our conclusion that order separation can be maintained, whereas absolute magnitudes cannot generally be preserved.

Finally, define the soft-thresholding operator

$$\mathcal{T}_\tau(u) = \mathrm{sign}(u)(|u| - \tau)_+.$$

Our independent-case bound suggests choosing

$$\tau_n \asymp \frac{\|\boldsymbol{\beta}^*\|}{n}.$$

With a sufficiently large constant, this threshold removes independent scores satisfying our high-probability bound, while correlated scores that are asymptotically larger than τ_n remain nonzero. Thus, soft thresholding can strengthen order separation, although it does not preserve exact magnitudes or rankings.

B Proofs

B.1 Notations and assumptions

Inspired by recent advances in AI models with billions of parameters trained on massive datasets, we study regimes in which n, p, and d are all large. Hence, throughout the paper we will use the following standard notations. For deterministic sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $C > 0$ such that $|a_n| \le C|b_n|$ for all sufficiently large n, and $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. For random sequences $\{X_n\}$, we write $X_n = O_p(b_n)$ if for every $\varepsilon > 0$ there exist constants $C > 0$ and N such that $\mathbb{P}(|X_n| > Cb_n) < \varepsilon$ for all $n \ge N$, and $X_n = o_p(b_n)$ if X_n/b_n goes to zero in probability.

B.1.1 Notations and definitions

Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the minimum and maximum eigenvalues of a matrix A, respectively. A random vector $\mathbf{x} \in \mathbb{R}^d$ is called norm-subGaussian with parameter σ, denoted as $nSG(\sigma)$, if there exists $\sigma > 0$ such that

$$\mathbb{P}\big(\|\mathbf{x} - \mathbb{E}\mathbf{x}\| \ge t\big) \le 2e^{-\frac{t^2}{2\sigma^2}} \quad \text{for all } t \ge 0,$$

as defined in Jin et al. [2019].

B.1.2 Detailed modeling assumptions

In this section, we provide the formal statements of the assumptions summarized in Section 2.1. Throughout this section, let

$$\mathcal{T} := \big\{(t_1, t_2) \in [0,1]^2 : t_1 + t_2 \le 1\big\}.$$

For each $i \in [n]$ and $(t_1, t_2) \in \mathcal{T}$, define

$$\boldsymbol{\beta}_{1,i} = t_1\breve{\boldsymbol{\beta}} + t_2\breve{\boldsymbol{\beta}}_{/i} + (1 - t_1 - t_2)\boldsymbol{\beta}^*,$$
$$\boldsymbol{\beta}_{2,i} = t_1\hat{\boldsymbol{\beta}} + t_2\hat{\boldsymbol{\beta}}_{/i} + (1 - t_1 - t_2)\boldsymbol{\beta}^*.$$

Assumption B.1 (Sub-Gaussian design)**.** The rows of $\boldsymbol{X}_n \in \mathbb{R}^{n\times p}$ are independent, zero-mean sub-Gaussian random vectors with covariance matrix $\boldsymbol{\Sigma}_p$. We assume that $\boldsymbol{\Sigma}_p$ is well-conditioned, with

$$\rho_{\max} \asymp \rho_{\min} \asymp \|\boldsymbol{\beta}^*\|^{-2},$$

where $\rho_{\max}$ and $\rho_{\min}$ denote the largest and smallest eigenvalues of $\boldsymbol{\Sigma}_p$, respectively.

Justification of scaling. When $d > p$ and $\boldsymbol{\beta}^* = \mathrm{vec}(\mathbf{W})$ with $\mathbf{W} \in \mathbb{R}^{h\times p}$, we typically have

$$f(\mathbf{x}, \boldsymbol{\beta}^*) \asymp \|\mathbf{W}\mathbf{x}\| = O_p(1).$$

Assuming that both $\mathbf{W}^\top\mathbf{W}$ and $\boldsymbol{\Sigma}$ are well-conditioned, we obtain

$$\|\mathbf{W}\mathbf{x}\|^2 \asymp \frac{\|\mathbf{W}\|_F^2}{p}\|\mathbf{x}\|^2 \asymp \frac{\|\mathbf{W}\|_F^2}{p} \cdot p\|\boldsymbol{\Sigma}\| = \|\mathbf{W}\|_F^2\|\boldsymbol{\Sigma}\|,$$

which implies

$$\|\boldsymbol{\Sigma}\| \asymp \|\boldsymbol{\beta}^*\|^{-2}.$$

Assumption B.2 (Regularity conditions on ∇f)**.** There exists a constant $C_1 = O(1)$ such that, with probability at least $1 - q_n$, where $q_n \to 0$, the following holds for each $j \in \{1, 2\}$:

$$\sup_{\mathbf{x} \in \{\mathbf{x}_{\text{new}}, \mathbf{x}_1, \ldots, \mathbf{x}_n\}} \sup_{i \in [n]} \sup_{(t_1, t_2) \in \mathcal{T}} \|\nabla f(\mathbf{x}, \boldsymbol{\beta}_{j,i})\| \leq C_1 \operatorname{poly}(\log n).$$

Assumption B.3 (Regularity conditions on $\mathbf{G}$ and empirical Hessian)**.** Recall the definition of $\mathbf{G}(\boldsymbol{\beta})$ from Eq. (5). There exist constants $C_2, c_1 = \Theta(1)$ such that, with probability at least $1 - q_n$, the following conditions hold for each $j \in \{1, 2\}$:

$$\frac{1}{n} \sigma_{\max} \left(\mathbf{G}(\hat{\boldsymbol{\beta}})^\top \mathbf{G}(\hat{\boldsymbol{\beta}}) \right) \leq C_2 \|\boldsymbol{\beta}^*\|^{-2},$$

$$\inf_{i \in [n]} \inf_{(t_1, t_2) \in \mathcal{T}} \frac{1}{n} \sigma_{\min} \left(\mathbf{G}(\boldsymbol{\beta}_{j,i})^\top \mathbf{G}(\boldsymbol{\beta}_{j,i}) \right) \geq c_1 \|\boldsymbol{\beta}^*\|^{-2},$$

$$\sup_{i \in [n]} \sup_{(t_1, t_2) \in \mathcal{T}} \frac{1}{n} \sigma_{\max} \left(\sum_{r=1}^{n} \nabla^2 \ell(y_r, f(\mathbf{x}_r, \boldsymbol{\beta}_{j,i})) \right) \leq C_2 \|\boldsymbol{\beta}^*\|^{-2},$$

$$\inf_{i \in [n]} \inf_{(t_1, t_2) \in \mathcal{T}} \frac{1}{n} \sigma_{\min} \left(\sum_{r=1}^{n} \nabla^2 \ell(y_r, f(\mathbf{x}_r, \boldsymbol{\beta}_{j,i})) \right) \geq c_1 \|\boldsymbol{\beta}^*\|^{-2}.$$

Assumption B.4 (Norm-sub-Gaussian gradients)**.** Let $\mathbf{g}_{i,*} = \nabla f(\mathbf{x}_i, \boldsymbol{\beta}^*)$. We assume that $\{\mathbf{g}_{i,*}\}_{i=1}^n$ are i.i.d. norm-sub-Gaussian random vectors with parameter $\sigma > 0$, denoted by

$$\mathbf{g}_{i,*} \overset{\text{iid}}{\sim} \text{nSG}(\sigma).$$

Furthermore, let

$$\boldsymbol{\Sigma}_g^* = \mathbb{E}\left[\mathbf{g}_{i,*} \mathbf{g}_{i,*}^\top \right].$$

We assume that

$$\lambda_{\max}(\boldsymbol{\Sigma}_g^*) \asymp \lambda_{\min}(\boldsymbol{\Sigma}_g^*) = O(\|\boldsymbol{\beta}^*\|^{-2}).$$

Assumption B.5 (Regularity conditions on $\nabla^2 f$)**.** There exists a constant $C_2 = O(1)$ such that, with probability at least $1 - q_n$,

$$\sup_{\mathbf{x} \in \{\mathbf{x}_{\text{new}}, \mathbf{x}_1, \ldots, \mathbf{x}_n\}} \sup_{i \in [n]} \sup_{(t_1, t_2) \in \mathcal{T}} \left\| \nabla^2 f(\mathbf{x}, \boldsymbol{\beta}_{2,i}) \right\|_{\text{op}} \leq C_2 \|\boldsymbol{\beta}^*\|^{-2},$$

where $\boldsymbol{\beta}_{2,i}$ is defined at the beginning of this subsection.

Assumption B.6 (Regularity conditions on ℓ)**.** There exist constants $\tilde{C}_1, \tilde{C}_2 = O(1)$ and $\mu > 0$ such that all the following bounds hold simultaneously with probability at least $1 - q_n$:

$$\sup_{i \in [n]} \sup_{(t_1, t_2) \in \mathcal{T}} \left| \dot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}_{2,i})) \right| \leq \tilde{C}_1 \operatorname{poly}(\log n),$$

$$\sup_{i \in [n]} \sup_{(t_1, t_2) \in \mathcal{T}} \left| \ddot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}_{2,i})) \right| \leq \tilde{C}_1 \operatorname{poly}(\log n),$$

$$\inf_{i \in [n]} \inf_{(t_1, t_2) \in \mathcal{T}} \ddot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}_{2,i})) \geq \mu.$$

For each $i \in [n]$, define the nonlinear loss-curvature vector

$$\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}) := \left(\ddot{\ell}(y_j, f(\mathbf{x}_j, \boldsymbol{\beta})) \right)_{j \in [n] \setminus \{i\}} \in \mathbb{R}^{n-1}.$$

The following Lipschitz-like conditions hold:

$$\sup_{i \in [n]} \sup_{t \in [0,1]} \frac{\left\| \ddot{\boldsymbol{\ell}}_{/i}((1-t)\breve{\boldsymbol{\beta}}_{/i} + t\breve{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}}) \right\|_2}{\|\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i}\|_2} \leq \tilde{C}_2 \frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|} \sqrt{\operatorname{poly}(\log n)},$$

$$\sup_{i \in [n]} \sup_{t \in [0,1]} \frac{\left\| \ddot{\boldsymbol{\ell}}_{/i}((1-t)\breve{\boldsymbol{\beta}}_{/i} + t\breve{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}^*) \right\|_2}{\|\breve{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2} \leq \tilde{C}_2 \frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|} \sqrt{\operatorname{poly}(\log n)},$$

$$\sup_{i \in [n]} \sup_{t \in [0,1]} \frac{\left\| \ddot{\boldsymbol{\ell}}_{/i}((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}^*) \right\|_2}{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2} \leq \tilde{C}_2 \frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|} \sqrt{\operatorname{poly}(\log n)},$$

$$\sup_{i \in [n]} \frac{\left\| \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}) - \ddot{\boldsymbol{\ell}}_{/i}(\hat{\boldsymbol{\beta}}_{/i}) \right\|_2}{\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/i}\|_2} \leq \tilde{C}_2 \frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|} \sqrt{\operatorname{poly}(\log n)}.$$

For the frozen linearized model, define

$$\boldsymbol{d}_{/i}^{\mathrm{lin}}(\boldsymbol{\beta}) := \big(\ddot{\ell}(y_j, \mathbf{g}_j^\top \boldsymbol{\beta} + \mathbf{b}_j)\big)_{j\in[n]\setminus\{i\}} \in \mathbb{R}^{n-1}.$$

The corresponding curvature satisfies

$$\inf_{i\in[n]} \inf_{t\in[0,1]} \min_{j\in[n]\setminus\{i\}} \ddot{\ell}\left(y_j, \mathbf{g}_j^\top\left((1-t)\breve{\boldsymbol{\beta}}_{/i} + t\breve{\boldsymbol{\beta}}\right) + \mathbf{b}_j\right) \geq \mu,$$

and, simultaneously for every $i \in [n]$,

$$\sup_{t\in[0,1]} \left\| \boldsymbol{d}_{/i}^{\mathrm{lin}}((1-t)\breve{\boldsymbol{\beta}}_{/i} + t\breve{\boldsymbol{\beta}}) - \boldsymbol{d}_{/i}^{\mathrm{lin}}(\breve{\boldsymbol{\beta}}) \right\|_2 \leq \tilde{C}_2 \frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|} \sqrt{\mathrm{poly}(\log n)} \|\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}\|_2.$$

Assumption B.7 (Unique minimizer)**.** For $\mathbf{g}_j = \nabla f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})$, the function $\sum_{j=1}^n \ell\left(y_j, \mathbf{g}_j^\top \boldsymbol{\beta} + \mathbf{b}_j\right)$ has a unique minimizer.

The following assumption is only required for establishing the lower bounds in Proposition 2.4 and Theorem 2.6.

Assumption B.8 (Uniform lower bounds)**.** There exists a constant $D_1 > 0$ such that

$$\inf_{i\in[n]} \|\nabla f(\mathbf{x}_i; \boldsymbol{\beta}^*)\|_2^2 \geq D_1,$$

and

$$\inf_{i\in[n]} \left|\dot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}^*))\right| \geq D_1.$$

B.2 Proofs of theorems

B.2.1 Proof of Proposition 2.4

By using the mean value theorem we have

$$\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) \triangleq f(\mathbf{x}_{\mathrm{new}}; \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_{\mathrm{new}}; \hat{\boldsymbol{\beta}}) = \nabla^\top f(\mathbf{x}_{\mathrm{new}}, \tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}), \tag{12}$$

where $\tilde{\boldsymbol{\beta}}$ is a point on the line segment that connects $\hat{\boldsymbol{\beta}}_{/i}$ and $\hat{\boldsymbol{\beta}}$. Using the first order optimality conditions for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{/i}$ we have

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{j\neq i} \nabla l(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{/i})) \\ &= \frac{1}{n}\sum_{j\neq i} \nabla l(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})) + \frac{1}{n}\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})). \end{aligned} \tag{13}$$

Hence, by using the integral form of the mean value theorem we have

$$(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) = \Big(\int_{t=0}^1 \frac{1}{n}\sum_{j\neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}}))dt\Big)^{-1} \left(\frac{1}{n}\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))\right). \tag{14}$$

By combing (12) and (14) we obtain

$$\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \nabla^\top f(\mathbf{x}_{\mathrm{new}}, \tilde{\boldsymbol{\beta}})\Big(\int_{t=0}^1 \frac{1}{n}\sum_{j\neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}}))dt\Big)^{-1} \left(\frac{1}{n}\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))\right). \tag{15}$$

Therefore we can use the Cauchy-Schwartz inequality and Assumption B.3 to prove

$$|\mathcal{I}^{\mathrm{True}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})| \leq \frac{\|\boldsymbol{\beta}^*\|^2}{c_1 n} \|\nabla f(\mathbf{x}_{\mathrm{new}}, \tilde{\boldsymbol{\beta}})\| \|\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))\| = O_p\Big(\frac{\mathrm{poly}(\log n)\|\boldsymbol{\beta}^*\|^2}{n}\Big), \tag{16}$$

where to obtaint the last equality we have used Assumptions B.2 and B.6.

To prove the second part of the theorem, i.e. (8), again we use (15). Note that

$$\begin{aligned}
\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_i) &= \nabla^\top f(\mathbf{x}_{\text{new}}, \tilde{\boldsymbol{\beta}}) \Big(\int_{t=0}^1 \frac{1}{n} \sum_{j \neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})) dt \Big)^{-1} \left(\frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})) \right) \\
&= \nabla^\top f(\mathbf{x}_{\text{new}}, \boldsymbol{\beta}^*) \Big(\int_{t=0}^1 \frac{1}{n} \sum_{j \neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})) dt \Big)^{-1} \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})) \\
&+ \Big(\nabla^\top f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}}) - \nabla^\top f(\mathbf{x}_{\text{new}}, \boldsymbol{\beta}^*) \Big) \Big(\int_{t=0}^1 \frac{1}{n} \sum_{j \neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})) dt \Big)^{-1} \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})) \\
&= \nabla^\top f(\mathbf{x}_{\text{new}}, \boldsymbol{\beta}^*) \Big(\int_{t=0}^1 \frac{1}{n} \sum_{j \neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})) dt \Big)^{-1} \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})) + O_p \left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}} \right) \quad (17)
\end{aligned}$$

where to obtain the last inequality we have used the Cauchy-Schwartz inequality and Lemma B.10. With a similar reasoning we can further simplify (17) in the following way:

$$\begin{aligned}
\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_i) &= \nabla^\top f(\mathbf{x}_{\text{new}}, \boldsymbol{\beta}^*) \Big(\int_{t=0}^1 \frac{1}{n} \sum_{j \neq i} \nabla^2 l(y_j, f(\mathbf{x}_j, t\hat{\boldsymbol{\beta}}_{/i} + (1-t)\hat{\boldsymbol{\beta}})) dt \Big)^{-1} \frac{1}{n} \dot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}^*) \nabla f(\mathbf{x}_i, \boldsymbol{\beta}^*) \\
&+ O_p \left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}} \right). \quad (18)
\end{aligned}$$

Furthermore, using Assumption B.3 we have

$$|\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_i)| \geq \frac{\|\boldsymbol{\beta}^*\|^2}{\tilde{c}_2 n} |\nabla^\top f(\mathbf{x}_i, \boldsymbol{\beta}^*) \nabla f(\mathbf{x}_i, \boldsymbol{\beta}^*)| |\dot{\ell}(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}^*))| - O_p \left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}} \right). \quad (19)$$

Assumption B.8 then implies (8).

B.2.2 Proof of Proposition 2.5

First note that

$$\begin{aligned}
&|f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}})| \leq \sup_{t \in [0,1]} \|\mathbf{g}_{\text{new}}^\top (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})^\top \nabla^2 f(\mathbf{x}_{\text{new}}, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}) (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| \\
\leq & \sup_{t \in [0,1]} \|\nabla^2 f(\mathbf{x}_{\text{new}}, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}})\| \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{g}_{\text{new},*}^\top (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|(\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*})^\top (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| \\
\leq & C_2 \|\boldsymbol{\beta}^*\|^{-2} \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{g}_{\text{new},*}^\top (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*}\| \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|.
\end{aligned}$$

The first inequality follows from the mean value theorem. The second inequality uses the triangle inequality. The third inequality follows from Assumption B.5 and the Cauchy–Schwarz inequality. Thus, it suffices to bound the three terms in the last expression.

1. Bounding $\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|$. Using the mean value theorem and the first-order optimality conditions in (1) and (3), we have:

$$\begin{aligned}
0 &= \frac{1}{n} \sum_{j=1}^n \nabla l(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{/i})) - \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i})) \\
&= \frac{1}{n} \sum_{j=1}^n \nabla l(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})) + \frac{1}{n} \sum_{j=1}^n \nabla^2 l(y_j, f(\mathbf{x}_j, \tilde{\boldsymbol{\beta}}_{/i})) (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) - \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i})) \\
&= \frac{1}{n} \sum_{j=1}^n \nabla^2 l(y_j, f(\mathbf{x}_j, \tilde{\boldsymbol{\beta}}_{/i})) (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) - \frac{1}{n} \nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i})),
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_{/i} = (1-t_i)\hat{\boldsymbol{\beta}}_{/i} + t_i\hat{\boldsymbol{\beta}}$ for some $t_i \in [0,1]$. This implies:

$$\begin{aligned}
\sup_{i\in[n]} \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\| &= \sup_{i\in[n]} \Big\| \Big(\frac{1}{n}\sum_{j=1}^{n} \nabla^2 l(y_j, f(\mathbf{x}_j, \tilde{\boldsymbol{\beta}}_{/i}))\Big)^{-1} \frac{1}{n}\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i}))\Big\| \\
&\le \sup_{i\in[n]} \sigma_{\max}\Big(\frac{1}{n}\sum_{j=1}^{n} \nabla^2 l(y_j, f(\mathbf{x}_j, \tilde{\boldsymbol{\beta}}_{/i}))\Big)^{-1}\Big) \sup_{i\in[n]} \Big\| \frac{1}{n}\nabla l(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i}))\Big\| \\
&\le \frac{\tilde{C}_1\|\boldsymbol{\beta}^*\|^2}{c_1}\frac{\mathrm{poly}(\log \mathrm{n})}{n}, \qquad \text{with probability } \ge 1-q_n, \qquad (20)
\end{aligned}$$

where to obtain the last inequality we use Assumption B.2 and B.6.

2. Bounding $\|\mathbf{g}_{\mathrm{new}} - \mathbf{g}_{\mathrm{new},*}\|$. From Lemma B.10, for any $\epsilon > 0$, with probability at least $1 - q_n - n^{-\epsilon}$, there exists an absolute constant s_g^* such that

$$\|\mathbf{g}_{\mathrm{new}} - \mathbf{g}_{\mathrm{new},*}\| \le s_g^*(\log n)^{0.5} n^{\epsilon-0.5}. \qquad (21)$$

3. Bounding $\|\mathbf{g}_{\mathrm{new},*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\|$. By Assumption B.4, $\mathbf{g}_{\mathrm{new},*}$ is sub-Gaussian. Using (20) and the independence of $\mathbf{g}_{\mathrm{new},*}$ from $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}_{/i}$, $\breve{\boldsymbol{\beta}}$, and $\breve{\boldsymbol{\beta}}_{/i}$, there exists an absolute constant s_g such that $\mathbf{g}_{\mathrm{new},*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})$ is a $s_g\frac{\|\boldsymbol{\beta}^*\|\mathrm{poly}(\log \mathrm{n})}{n}$-sub-Gaussian random variable. Hence, for any $\epsilon > 0$, with probability at least $1 - 2n^{-\epsilon}$, there exists an absolute constant $\hat{s}_g$ such that

$$\|\mathbf{g}_{\mathrm{new},*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| \le \hat{s}_g\frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}}. \qquad (22)$$

Final bound. Combining the bounds above, there exists a constant s_1 such that, for any $\epsilon > 0$,

$$|f(\mathbf{x}_{\mathrm{new}}, \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_{\mathrm{new}}, \hat{\boldsymbol{\beta}})| \le s_1\Big(\frac{\|\boldsymbol{\beta}^*\|^2}{n^2}\mathrm{poly}(\log \mathrm{n}) + \frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}} + \frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}}\Big),$$

with probability at least $1 - 3(q_n + n^{-\epsilon})$. The final result follows from $\|\boldsymbol{\beta}^*\|^2 \ll n^{1-\epsilon}$.

B.2.3 Proof of Theorem 2.6

By the first-order optimality conditions, we have:

$$\begin{aligned}
\sum_{j=1}^{n} \dot{\ell}(y_j, \mathbf{g}_j^\top\hat{\boldsymbol{\beta}} + \mathbf{b}_j)\mathbf{g}_j &= 0, \\
\sum_{j=1}^{n} \dot{\ell}(y_j, \mathbf{g}_j^\top\breve{\boldsymbol{\beta}} + \mathbf{b}_j)\mathbf{g}_j &= 0.
\end{aligned}$$

Under Assumption B.7, $\sum_{j=1}^{n} \ell(y_j, \mathbf{g}_j^\top\boldsymbol{\beta} + \mathbf{b}_j)$ has a unique minimizer. Therefore, we conclude that $\hat{\boldsymbol{\beta}} = \breve{\boldsymbol{\beta}}$. Next, we bound the difference:

$$\begin{aligned}
&|f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i}) - (\mathbf{g}_i^\top\breve{\boldsymbol{\beta}}_{/i} + \mathbf{b}_i)| \le |f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})| + |\mathbf{g}_i^\top(\hat{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}})| + |\mathbf{g}_i^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})| \\
&\le \sup_{t\in[0,1]} \|\mathbf{g}_i^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})^\top\Big(\int_{t=0}^{1} \nabla^2 f(\mathbf{x}_i, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}})dt\Big)(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| \\
&\quad + |\mathbf{g}_i^\top(\hat{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}})| + |\mathbf{g}_i^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})| \\
&\le \sup_{t\in[0,1]} \|\nabla^2 f(\mathbf{x}_i, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}})\|\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2 \\
&\quad + \|\mathbf{g}_{i,*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|\mathbf{g}_{i,*}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\| \\
&\quad + \|(\mathbf{g}_i - \mathbf{g}_{i,*})^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|(\mathbf{g}_i - \mathbf{g}_{i,*})^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\| \\
&\le C_2\|\boldsymbol{\beta}^*\|^{-2}\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2 \\
&\quad + \|\mathbf{g}_{i,*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|\mathbf{g}_{i,*}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\| \\
&\quad + \|\mathbf{g}_i - \mathbf{g}_{i,*}\|\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\| + \|\mathbf{g}_i - \mathbf{g}_{i,*}\|\|\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i}\|.
\end{aligned}$$

The first inequality follows from the mean value theorem and the fact that $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. The second inequality uses the triangle and Cauchy-Schwarz inequalities. The third inequality follows from Assumption B.5 and Cauchy-Schwarz. Thus, it suffices to bound the three terms in the last three lines. Using the same method as the one used in deriving (20) we have

$$\sup_{i\in[n]} \|\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}\| \quad = \quad O_p\Big(\frac{\|\boldsymbol{\beta}^*\|^2 \text{poly}(\log \text{n})}{n}\Big).$$

Furthermore, using the mean value theorem we have

$$\begin{aligned}
&\mathbf{g}_{i,*}^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_{i,*}^\top\Big(\int_{t=0}^1 \boldsymbol{G}_{/i}^\top \text{diag}[\boldsymbol{d}_{/i}^{\text{lin}}(\breve{\boldsymbol{\beta}}_{/i} - t\breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i} dt\Big)^{-1}\mathbf{g}_i \\
= \quad & \dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_{i,*}^\top\Big(\int_{t=0}^1 \boldsymbol{G}_{/i}^\top \text{diag}[\boldsymbol{d}_{/i}^{\text{lin}}(\breve{\boldsymbol{\beta}}_{/i} - t\breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i} dt\Big)^{-1}\mathbf{g}_{i,*} + O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}}\right), \qquad (23)
\end{aligned}$$

where to obtain the last equality we used an argument similar to the one we used for deriving (19). Hence, we have

$$|\mathbf{g}_{i,*}^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}})| \quad \leq \quad \frac{\|\boldsymbol{\beta}^*\|^2}{C_2 n}|\dot{\ell}_i(\breve{\boldsymbol{\beta}})|\mathbf{g}_{i,*}^\top\mathbf{g}_{i,*} + O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}}\right) = O_p\left(\frac{\text{poly}(\log \text{n})\|\boldsymbol{\beta}^*\|^2}{n}\right) \qquad (24)$$

Similarly, we have

$$|\mathbf{g}_{i,*}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})| = O_p\left(\frac{\text{poly}(\log \text{n})\|\boldsymbol{\beta}^*\|^2}{n}\right).$$

Now we aim to bound the term $\|\mathbf{g}_i - \mathbf{g}_{i,*}\|\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|$. According to (21) we have with high probability,

$$\|\mathbf{g}_i - \mathbf{g}_{i,*}\| \leq s_g^*(\log n)^{0.5} n^{\epsilon-0.5}. \qquad (25)$$

Therefore,

$$\|\mathbf{g}_i - \mathbf{g}_{i,*}\|\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\| = O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}}\right). \qquad (26)$$

Combining all the calculation above we can concluse that

$$|f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{/i}) - (\mathbf{g}_i^\top \breve{\boldsymbol{\beta}}_{/i} + \mathbf{b}_i)| = O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2 \text{poly}(\log \text{n})}{n}\right).$$

Now we turn our attention to proving (10). First note that using the same argument as the one presented above we have

$$\begin{aligned}
&\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_i) = \nabla f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}) = \nabla f(\mathbf{x}_i, \boldsymbol{\beta}^*)^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}) + O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2 \text{poly}(\log \text{n})}{n^{1.5-\epsilon}}\right) \\
= \quad & \dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_{i,*}^\top\Big(\int_{t=0}^1 \boldsymbol{G}_{/i}^\top \text{diag}[\boldsymbol{d}_{/i}^{\text{lin}}(\breve{\boldsymbol{\beta}}_{/i} - t\breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i} dt\Big)^{-1}\mathbf{g}_i + O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2 \text{poly}(\log \text{n})}{n^{1.5-\epsilon}}\right) \\
= \quad & \dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_{i,*}^\top\Big(\int_{t=0}^1 \boldsymbol{G}_{/i}^\top \text{diag}[\boldsymbol{d}_{/i}^{\text{lin}}(\breve{\boldsymbol{\beta}}_{/i} - t\breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i} dt\Big)^{-1}\mathbf{g}_{i,*} + O_p\left(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}}\right), \qquad (27)
\end{aligned}$$

Using (32) we can conclude that with high probability we have that

$$\begin{aligned}
&|\dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_{i,*}^\top\Big(\int_{t=0}^1 \boldsymbol{G}_{/i}^\top \text{diag}[\boldsymbol{d}_{/i}^{\text{lin}}(\breve{\boldsymbol{\beta}}_{/i} - t\breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i} dt\Big)^{-1}\mathbf{g}_{i,*}| \\
&\qquad \geq |\dot{\ell}_i(\breve{\boldsymbol{\beta}})|\mathbf{g}_{i,*}^\top\mathbf{g}_{i,*}(\frac{2\tilde{C}_1 C_2 n \text{poly}(\log \text{n})}{\|\boldsymbol{\beta}^*\|^2})^{-1}. \qquad (28)
\end{aligned}$$

Combing this results with Assumption B.8 establishes (10).

B.2.4 Proof of Theorem 2.9

By the first-order optimality conditions for $\hat{\boldsymbol{\beta}}$ and $\breve{\boldsymbol{\beta}}$, we have:

$$\begin{aligned}\sum_{j=1}^{n} \dot{\ell}(y_j, \mathbf{g}_j^\top \hat{\boldsymbol{\beta}} + \mathbf{b}_j)\mathbf{g}_j &= 0,\\ \sum_{j=1}^{n} \dot{\ell}(y_j, \mathbf{g}_j^\top \breve{\boldsymbol{\beta}} + \mathbf{b}_j)\mathbf{g}_j &= 0.\end{aligned}$$

Under Assumption B.7, $\sum_{j=1}^{n} \ell(y_j, \mathbf{g}_j^\top \boldsymbol{\beta} + \mathbf{b}_j)$ has a unique minimizer; therefore, we conclude that $\hat{\boldsymbol{\beta}} = \breve{\boldsymbol{\beta}}$. Next, we bound the difference:

$$\begin{aligned}&|f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}}_{/i}) - (\mathbf{g}_{\text{new}}^\top \breve{\boldsymbol{\beta}}_{/i} + \mathbf{b}_{\text{new}})| \le |f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}}_{/i}) - f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}})| + |\mathbf{g}_{\text{new}}^\top(\hat{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}})| + |\mathbf{g}_{\text{new}}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})|\\
&\le \sup_{t\in[0,1]} \|\mathbf{g}_{\text{new}}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})^\top \nabla^2 f(\mathbf{x}_{\text{new}}, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\|\\
&\quad + |\mathbf{g}_{\text{new}}^\top(\hat{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}})| + |\mathbf{g}_{\text{new}}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})|\\
&\le \sup_{t\in[0,1]} \|\nabla^2 f(\mathbf{x}_{\text{new}}, (1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}})\| \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2\\
&\quad + \|\mathbf{g}_{\text{new},*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|\mathbf{g}_{\text{new},*}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\|\\
&\quad + \|(\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*})^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|(\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*})^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\|\\
&\le C_2 \|\boldsymbol{\beta}^*\|^{-2} \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|^2\\
&\quad + \|\mathbf{g}_{\text{new},*}^\top(\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}})\| + \|\mathbf{g}_{\text{new},*}^\top(\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i})\|\\
&\quad + \|\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*}\| \|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\| + \|\mathbf{g}_{\text{new}} - \mathbf{g}_{\text{new},*}\| \|\breve{\boldsymbol{\beta}} - \breve{\boldsymbol{\beta}}_{/i}\|.\end{aligned}$$

The first inequality follows from the mean value theorem and the fact that $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. The second inequality uses the triangle and Cauchy-Schwarz inequalities. The third inequality follows from Assumption B.5 and Cauchy-Schwarz. Thus, it suffices to bound the three terms in the last three lines. Using the same method as (20) and (22), we can get the bounds

$$\sup_{i\in[n]} \|\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}\| = O\Big(\frac{\|\boldsymbol{\beta}^*\|^2 \mathrm{poly}(\log \mathrm{n})}{n}\Big), \text{ with probability } \ge 1 - q_n, \tag{29}$$

$$\|\mathbf{g}_{\text{new},*}^\top(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}})\| = O\Big(\frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}}\Big), \text{ with probability } \ge 1 - 2n^{-\epsilon}. \tag{30}$$

Combining (21), we conclude that there exists a constant s_1 such that, for any $\epsilon > 0$,

$$|f(\mathbf{x}_{\text{new}}, \hat{\boldsymbol{\beta}}_{/i}) - (\mathbf{g}_{\text{new}}^\top \breve{\boldsymbol{\beta}}_{/i} + \mathbf{b}_{\text{new}})| \le s_1 \Big(\frac{\|\boldsymbol{\beta}^*\|^2}{n^2} \mathrm{poly}(\log \mathrm{n}) + \frac{\|\boldsymbol{\beta}^*\|^2}{n^{1.5-\epsilon}} + \frac{\|\boldsymbol{\beta}^*\|}{n^{1-\epsilon}} \Big),$$

with probability at least $1 - 3(q_n + n^{-\epsilon})$. The final result follows from $\|\boldsymbol{\beta}^*\|^2 \ll n^{1-\epsilon}$.

B.2.5 Proof of Theorem 2.12

Proof. Throughout the proof, P_n denotes a deterministic fixed polynomial in $1 + \log n$, which may be enlarged from line to line.

Step 1: Conditioning and stability. For each $i \in [n]$, write

$$\boldsymbol{G}_{/i} := \mathbf{G}_{/i}(\hat{\boldsymbol{\beta}}), \qquad \boldsymbol{D}_i := \mathrm{diag}\big(\ddot{\ell}(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}))\big)_{j\neq i}, \qquad A_i := \boldsymbol{G}_{/i}^\top \boldsymbol{D}_i \boldsymbol{G}_{/i}.$$

Assumptions B.2 and B.6 give

$$\max_{i\in[n]} \Big\{ \|\mathbf{g}_i\|_2, \|\mathbf{g}_{i,*}\|_2, |\dot{\ell}_i(\breve{\boldsymbol{\beta}})| \Big\} \le P_n, \qquad \mu I_{n-1} \preceq \boldsymbol{D}_i \preceq B_n I_{n-1}, \quad B_n \le P_n.$$

Here $\dot{\ell}_i(\breve{\boldsymbol{\beta}}) = \dot{\ell}(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))$ because $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$.

Since deleting row i subtracts $\mathbf{g}_i\mathbf{g}_i^\top$ from $\mathbf{G}(\hat{\boldsymbol{\beta}})^\top\mathbf{G}(\hat{\boldsymbol{\beta}})$, Assumption B.3 implies

$$\frac{c_1 n}{2\|\boldsymbol{\beta}^*\|^2} \le \lambda_{\min}(\boldsymbol{G}_{/i}^\top \boldsymbol{G}_{/i}) \le \lambda_{\max}(\boldsymbol{G}_{/i}^\top \boldsymbol{G}_{/i}) \le \frac{C_2 n}{\|\boldsymbol{\beta}^*\|^2} \tag{31}$$

for all sufficiently large n. Indeed, after enlarging P_n, the removed row has squared norm at most P_n, whereas $n/\|\boldsymbol{\beta}^*\|^2 \ge n^\epsilon$. Consequently,

$$\frac{\|\boldsymbol{\beta}^*\|^2}{C_2 B_n n} I_d \preceq A_i^{-1} \preceq \frac{2\|\boldsymbol{\beta}^*\|^2}{\mu c_1 n} I_d. \tag{32}$$

Moreover, since $A_i \succeq \mu \boldsymbol{G}_{/i}^\top \boldsymbol{G}_{/i}$,

$$\|\boldsymbol{G}_{/i} A_i^{-1}\|_{\mathrm{op}}^2 = \|A_i^{-1}\boldsymbol{G}_{/i}^\top \boldsymbol{G}_{/i} A_i^{-1}\|_{\mathrm{op}} \le \frac{1}{\mu}\|A_i^{-1}\|_{\mathrm{op}} \le C\frac{\|\boldsymbol{\beta}^*\|^2}{n}.$$

Thus,

$$\|\boldsymbol{G}_{/i} A_i^{-1}\|_{\mathrm{op}} \le C\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}.$$

Step 2: A genuinely leave-one-out reference. Choose $\hat{\boldsymbol{\beta}}_{/i}$ as a measurable function of $\mathcal{D}_{/i} = \{\mathbf{z}_j : j \ne i\}$ and define

$$\boldsymbol{G}_{/i}^0 := \mathbf{G}_{/i}(\hat{\boldsymbol{\beta}}_{/i}), \qquad \boldsymbol{D}_i^0 := \mathrm{diag}\big(\ddot{\ell}(y_j, f(\mathbf{x}_j, \hat{\boldsymbol{\beta}}_{/i}))\big)_{j\ne i}, \qquad A_i^0 := (\boldsymbol{G}_{/i}^0)^\top \boldsymbol{D}_i^0 \boldsymbol{G}_{/i}^0.$$

These quantities are functions of $\mathcal{D}_{/i}$ alone. Therefore, $\mathbf{g}_{i,*}$ is independent of $(\boldsymbol{G}_{/i}^0, \boldsymbol{D}_i^0, A_i^0)$.

The uniform training-point Hessian bound in Assumption B.5, together with (20), gives

$$\rho_i := \|\boldsymbol{G}_{/i} - \boldsymbol{G}_{/i}^0\|_{\mathrm{op}} \le \|\boldsymbol{G}_{/i} - \boldsymbol{G}_{/i}^0\|_F \le C_2\sqrt{n}\,\|\boldsymbol{\beta}^*\|^{-2}\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/i}\|_2 \le \frac{P_n}{\sqrt{n}}.$$

The nonlinear endpoint Lipschitz-like condition in Assumption B.6 gives

$$\gamma_i := \|\boldsymbol{D}_i - \boldsymbol{D}_i^0\|_F \le P_n\frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|}\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/i}\|_2 \le P_n\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}} = o(1).$$

Hence, for sufficiently large n,

$$\frac{\mu}{2} I_{n-1} \preceq \boldsymbol{D}_i^0 \preceq (B_n + 1) I_{n-1}.$$

Singular-value perturbation and (31) imply

$$\|\boldsymbol{G}_{/i}^0\|_{\mathrm{op}} \le C\frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|}, \qquad \lambda_{\min}\big((\boldsymbol{G}_{/i}^0)^\top \boldsymbol{G}_{/i}^0\big) \ge c\frac{n}{\|\boldsymbol{\beta}^*\|^2}.$$

It follows that

$$\max\big\{\|A_i^{-1}\|_{\mathrm{op}}, \|(A_i^0)^{-1}\|_{\mathrm{op}}\big\} \le C\frac{\|\boldsymbol{\beta}^*\|^2}{n}.$$

Therefore, Lemma B.11 applies uniformly in i, with

$$a = C\frac{\|\boldsymbol{\beta}^*\|^2}{n}, \qquad \mu_0 = \frac{\mu}{2}, \qquad B \le P_n, \qquad \rho_i \le \frac{P_n}{\sqrt{n}}, \qquad \gamma_i \le P_n\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}.$$

Step 3: The infinity-norm bound. Define

$$\mathbf{u}_i^0 = \boldsymbol{G}_{/i}^0 (A_i^0)^{-1}\mathbf{g}_{i,*}, \qquad \mathbf{u}_{i,*} = \boldsymbol{G}_{/i} A_i^{-1}\mathbf{g}_{i,*}.$$

Conditional on $\mathcal{D}_{/i}$, the matrix $\boldsymbol{G}_{/i}^0(A_i^0)^{-1}$ is fixed and $\mathbf{g}_{i,*}$ retains its unconditional distribution. Then we get,

$$\|\boldsymbol{G}_{/i}^0(A_i^0)^{-1}\|_{\mathrm{op}} \le C\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}.$$

and

$$\mathbb{P}\left(\sup_{i\in[n]}\|\mathbf{u}_i^0\|_\infty \le \frac{P_n}{\sqrt{n}}\right) \longrightarrow 1. \tag{33}$$

On the event in (33), Lemma B.11 yields

$$\|\mathbf{u}_{i,*}\|_\infty \le \left(1+\frac{\gamma_i}{\mu_0}\right)\|\mathbf{u}_i^0\|_\infty + \left(1+\frac{2B}{\mu_0}\right)a\rho_i\|\mathbf{g}_{i,*}\|_2 \le \frac{P_n}{\sqrt{n}} + P_n\frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2}} \le \frac{P_n}{\sqrt{n}},$$

after enlarging P_n.

We next replace $\mathbf{g}_{i,*}$ by $\mathbf{g}_i$. Apply Lemma B.10 with $\eta = \epsilon/4$. Using the uniform training-point Hessian bound in Assumption B.5, we obtain

$$\sup_{i\in[n]}\|\mathbf{g}_i - \mathbf{g}_{i,*}\|_2 \le C_2\|\boldsymbol{\beta}^*\|^{-2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le P_n n^{-1/2+\eta}$$

with probability tending to one. Consequently,

$$\|\boldsymbol{G}_{/i}A_i^{-1}(\mathbf{g}_i - \mathbf{g}_{i,*})\|_\infty \le \|\boldsymbol{G}_{/i}A_i^{-1}\|_{\mathrm{op}}\|\mathbf{g}_i - \mathbf{g}_{i,*}\|_2 \le P_n\|\boldsymbol{\beta}^*\|n^{-1+\eta} \le P_n n^{-1/2-\epsilon/4}.$$

Combining these estimates gives

$$\sup_{i\in[n]}\|\boldsymbol{G}_{/i}A_i^{-1}\mathbf{g}_i\|_\infty \le \frac{P_n}{\sqrt{n}}. \tag{34}$$

Step 4: Exact representation of the ALO error. Set

$$\breve{\boldsymbol{\Delta}}_{/i} = \breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}, \qquad s_i = \dot{\ell}_i(\breve{\boldsymbol{\beta}}), \qquad \hat{\boldsymbol{\Delta}}_{/i} = s_i A_i^{-1}\mathbf{g}_i.$$

Using the frozen-gradient curvature vector from Assumption B.6, define

$$\boldsymbol{D}_i^{\mathrm{lin}}(t) := \mathrm{diag}\left(\boldsymbol{d}_{/i}^{\mathrm{lin}}(\breve{\boldsymbol{\beta}} + t\breve{\boldsymbol{\Delta}}_{/i})\right),$$

$$\overline{\boldsymbol{D}}_i := \int_0^1 \boldsymbol{D}_i^{\mathrm{lin}}(t)\,dt, \qquad \overline{A}_i := \boldsymbol{G}_{/i}^\top\overline{\boldsymbol{D}}_i\boldsymbol{G}_{/i}.$$

Since $\breve{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\mathbf{g}_j^\top\hat{\boldsymbol{\beta}} + \mathbf{b}_j = f(\mathbf{x}_j, \hat{\boldsymbol{\beta}})$, we have $\boldsymbol{D}_i^{\mathrm{lin}}(0) = \boldsymbol{D}_i$.

The first-order conditions for the full and leave-one-out linearized objectives, together with the integral Taylor formula, give the exact identity

$$\overline{A}_i\breve{\boldsymbol{\Delta}}_{/i} = s_i\mathbf{g}_i.$$

By the frozen-gradient curvature lower bound in Assumption B.6,

$$\overline{\boldsymbol{D}}_i \succeq \mu I_{n-1}.$$

Therefore, (31) and the same weighted-Gram argument used in Step 1 give

$$\|\overline{A}_i^{-1}\|_{\mathrm{op}} \le C\frac{\|\boldsymbol{\beta}^*\|^2}{n}, \qquad \|\overline{A}_i^{-1}\boldsymbol{G}_{/i}^\top\|_{\mathrm{op}} \le C\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}.$$

In particular,

$$\sup_{i\in[n]}\|\breve{\boldsymbol{\Delta}}_{/i}\|_2 \le P_n\frac{\|\boldsymbol{\beta}^*\|^2}{n}.$$

Define

$$\boldsymbol{\Gamma}_i := \overline{\boldsymbol{D}}_i - \boldsymbol{D}_i, \qquad \Delta_i := \boldsymbol{G}_{/i}^\top\boldsymbol{\Gamma}_i\boldsymbol{G}_{/i},$$

so that $\overline{A}_i = A_i + \Delta_i$. The frozen-gradient Lipschitz-like condition in Assumption B.6 implies

$$\|\boldsymbol{\Gamma}_i\|_F \le \int_0^1 \|\boldsymbol{D}_i^{\mathrm{lin}}(t) - \boldsymbol{D}_i\|_F\,dt \le P_n\frac{\sqrt{n}}{\|\boldsymbol{\beta}^*\|}\|\breve{\boldsymbol{\Delta}}_{/i}\|_2 \le P_n\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}. \tag{35}$$

Here the segment in Assumption B.6 is the same segment, with the parameterization reversed.

The resolvent identity now gives

$$\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i} = s_i(\overline{A}_i^{-1} - A_i^{-1})\mathbf{g}_i = -s_i\overline{A}_i^{-1}\boldsymbol{G}_{/i}^\top\boldsymbol{\Gamma}_i\boldsymbol{G}_{/i}A_i^{-1}\mathbf{g}_i.$$

Because $\boldsymbol{\Gamma}_i$ is diagonal,

$$\|\boldsymbol{\Gamma}_i\mathbf{v}\|_2 \le \|\boldsymbol{\Gamma}_i\|_F\|\mathbf{v}\|_\infty \qquad \text{for every } \mathbf{v} \in \mathbb{R}^{n-1}.$$

Consequently,

$$\|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\|_2 \le |s_i|\,\|\overline{A}_i^{-1}\boldsymbol{G}_{/i}^\top\|_{\mathrm{op}}\|\boldsymbol{\Gamma}_i\|_F\|\boldsymbol{G}_{/i}A_i^{-1}\mathbf{g}_i\|_\infty \le P_n\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{n}}\frac{1}{\sqrt{n}} = P_n\frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2}}.$$

Thus, simultaneously over all i,

$$\sup_{i\in[n]}\|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\|_2 \le P_n\frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2}}. \tag{36}$$

Step 5: From parameter error to influence error. By the definition of linearized influence,

$$\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i) = \mathbf{g}_i^\top\breve{\boldsymbol{\Delta}}_{/i}.$$

Write

$$H = \mathbf{G}^\top\operatorname{diag}\big(\ddot{\ell}_j(\breve{\boldsymbol{\beta}})\big)_{j=1}^n\mathbf{G}.$$

Since

$$A_i = H - \ddot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_i\mathbf{g}_i^\top,$$

the Sherman–Morrison identity yields

$$\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\mathbf{g}_i^\top A_i^{-1}\mathbf{g}_i = \mathbf{g}_i^\top\hat{\boldsymbol{\Delta}}_{/i}.$$

Therefore, Assumption B.2 and (36) imply

$$\sup_{i\in[n]}\big|\mathcal{I}^{\mathrm{Linear}}(\mathbf{z}_i, \mathbf{z}_i) - \mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)\big| \le \left(\sup_{i\in[n]}\|\mathbf{g}_i\|_2\right)\left(\sup_{i\in[n]}\|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\|_2\right) \le P_n\frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2}}.$$

Every fixed polynomial in $1 + \log n$ is bounded by n^ϵ for all sufficiently large n. Hence, the asserted bound follows with probability tending to one. □

B.2.6 Proof of Theorem 2.13

First, define $\breve{\boldsymbol{\Delta}}_{/i} := \breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\beta}}$ and the Jacobian compactly as

$$\boldsymbol{J}_{/i}(\boldsymbol{\theta}) = \boldsymbol{G}_{/i}^\top\operatorname{diag}(\boldsymbol{d}_{/i}^{\mathrm{lin}}(\boldsymbol{\theta}))\boldsymbol{G}_{/i}.$$

Also define the estimated and exact leave-i-out perturbation vectors as

$$\hat{\boldsymbol{\Delta}}_{/i} \triangleq \dot{\ell}_i(\breve{\boldsymbol{\beta}})\big[\boldsymbol{J}_{/i}(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\Delta}}_{/i})\big]^{-1}\mathbf{g}_i.$$

Applying the Woodbury matrix identity yields

$$\begin{aligned}\mathbf{g}_{\mathrm{new}}^\top\hat{\boldsymbol{\Delta}}_{/i} &= \dot{\ell}_i(\breve{\boldsymbol{\beta}})\,\mathbf{g}_{\mathrm{new}}^\top\Big(\boldsymbol{G}_{/i}^\top\operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}}_{/i} - \breve{\boldsymbol{\Delta}}_{/i})]\boldsymbol{G}_{/i}\Big)^{-1}\mathbf{g}_i \\ &= \frac{\dot{\ell}_i(\breve{\boldsymbol{\beta}})\,\mathbf{g}_{\mathrm{new}}^\top\big(\mathbf{G}^\top\operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\mathbf{G}\big)^{-1}\mathbf{g}_i}{1 - \ddot{\ell}_i(\breve{\boldsymbol{\beta}})\,\mathbf{g}_i^\top\big(\mathbf{G}^\top\operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\mathbf{G}\big)^{-1}\mathbf{g}_i}.\end{aligned}$$

From (29), we know that there exists an absolute constant s_Δ^* such that

$$\left\|\breve{\boldsymbol{\Delta}}_{/i}\right\|_2 \le s_\Delta^*\frac{\|\boldsymbol{\beta}^*\|^2\mathrm{poly}(\log \mathrm{n})}{n}, \qquad \text{with probability } \ge 1 - q_n. \tag{37}$$

By the first-order optimality conditions and the integral form of Taylor's theorem for the gradient of the linearized leave-one-out objective,

$$\breve{\boldsymbol{\Delta}}_{/i} = \dot{\ell}_i(\breve{\boldsymbol{\beta}}) \left[\int_0^1 \boldsymbol{J}_{/i}(\breve{\boldsymbol{\beta}}_{/i} - (1-t)\breve{\boldsymbol{\Delta}}_{/i})\, dt \right]^{-1} \mathbf{g}_i.$$

Introduce

$$\boldsymbol{\gamma}_{\boldsymbol{\delta}/i}(\boldsymbol{\theta}) \triangleq \boldsymbol{d}_{/i}^{\mathrm{lin}}(\boldsymbol{\theta}+\boldsymbol{\delta}) - \boldsymbol{d}_{/i}^{\mathrm{lin}}(\boldsymbol{\theta}),$$

so that

$$\boldsymbol{J}_{/i}(\boldsymbol{\theta}+\boldsymbol{\delta}) = \boldsymbol{J}_{/i}(\boldsymbol{\theta}) + \boldsymbol{G}_{/i}^\top \operatorname{diag}\big[\boldsymbol{\gamma}_{\boldsymbol{\delta}/i}(\boldsymbol{\theta})\big]\, \boldsymbol{G}_{/i}.$$

Note that $\boldsymbol{J}_{/i}(\boldsymbol{\theta}+\boldsymbol{\delta})$ remains positive definite for all $t \in [0,1]$, with $\boldsymbol{\theta} = \breve{\boldsymbol{\beta}}_{/i}$ and $\boldsymbol{\delta} = -(1-t)\breve{\boldsymbol{\Delta}}_{/i}$.

Using the notation above, we now bound the error. With high probability, we obtain

$$\begin{aligned}
& \left| \mathbf{g}_{\mathrm{new}}^\top \breve{\boldsymbol{\Delta}}_{/i} - \mathbf{g}_{\mathrm{new}}^\top \hat{\boldsymbol{\Delta}}_{/i} \right| \\
\leq\ & \left| \mathbf{g}_{\mathrm{new}*}^\top (\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}) \right| + \|\mathbf{g}_{\mathrm{new},*} - \mathbf{g}_{\mathrm{new}}\| \|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\| \\
\leq\ & (\|\boldsymbol{\beta}^*\|^{-1} + n^{\epsilon-0.5}) \|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\| \mathrm{poly}(\log \mathrm{n}),
\end{aligned} \tag{38}$$

where we use that $\mathbf{g}_{\mathrm{new},*}$ is independent of $\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}$, along with Assumption B.4 and Lemma B.10.

Next, we consider $\|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\|$. As we have get in (36),

$$\|\breve{\boldsymbol{\Delta}}_{/i} - \hat{\boldsymbol{\Delta}}_{/i}\| \preceq \frac{\|\boldsymbol{\beta}^*\|^2}{n^{3/2-\epsilon}}.$$

Returning to (38), we conclude that with high probability,

$$\left| \mathbf{g}_{\mathrm{new}}^\top \breve{\boldsymbol{\Delta}}_{/i} - \mathbf{g}_{\mathrm{new}}^\top \hat{\boldsymbol{\Delta}}_{/i} \right| = O\left(\frac{\|\boldsymbol{\beta}^*\|}{n^{3/2-\epsilon}} \right).$$

This completes the proof.

B.2.7 Proof of Theorem 2.15

By Equation (11), we have

$$\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \frac{\dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_{\mathrm{new}}^\top (\mathbf{G}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\mathbf{G})^{-1} \mathbf{g}_i}{1 - \ddot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_i^\top (\mathbf{G}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}(\breve{\boldsymbol{\beta}})]\mathbf{G})^{-1} \mathbf{g}_i}.$$

Using the Woodbury identity, we obtain

$$\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_{\mathrm{new}}^\top (\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i})^{-1} \mathbf{g}_i.$$

Similarly, from Equation (6), we have

$$\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \boldsymbol{\phi}_{\mathrm{new}}^\top (\Phi_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\Phi_{/i})^{-1} \boldsymbol{\phi}_i.$$

We ignore the k in $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}; k)$ for simplicity. Recall that for the projection matrix $S \in \mathbb{R}^{k \times d}$, TRAK defines the feature map $\boldsymbol{\phi}(\mathbf{x}) = S\mathbf{g}(\mathbf{x})$, where $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x}, \hat{\boldsymbol{\beta}})$. Accordingly, we define $\Phi = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n)^\top \in \mathbb{R}^{n \times k}$ with $\boldsymbol{\phi}_i = S\mathbf{g}_i$ and $\boldsymbol{\phi}_{\mathrm{new}} = S\mathbf{g}_{\mathrm{new}}$. The entries S_{ij} for $i \in [k], j \in [d]$ are i.i.d. $\mathcal{N}(0,1)$.

By 31 and Assumption B.6, with probability at least $1 - q_n - 2n^{-\epsilon}$,

$$\frac{n}{\|\boldsymbol{\beta}^*\|^2} \preceq \sigma_{\min}\big(\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}\big) \preceq \sigma_{\max}\big(\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}\big) \preceq \frac{n\mathrm{poly}(\log \mathrm{n})}{\|\boldsymbol{\beta}^*\|^2}. \tag{39}$$

We denote κ as the condition number of $\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}$ and $\kappa = O_p(\mathrm{poly}(\log \mathrm{n}))$ by Assumption B.6. Then we split $\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ and $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ as

$$\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \mathbf{g}_{\mathrm{new},*}^\top (\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i})^{-1} \mathbf{g}_i + \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, (\mathbf{g}_{\mathrm{new}} - \mathbf{g}_{\mathrm{new},*})^\top (\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i})^{-1} \mathbf{g}_i.$$

$\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}}) = \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, \boldsymbol{\phi}_{\mathrm{new},*}^\top (\Phi_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\Phi_{/i})^{-1}\boldsymbol{\phi}_i + \dot{\ell}_i(\breve{\boldsymbol{\beta}})\, (\boldsymbol{\phi}_{\mathrm{new}} - \boldsymbol{\phi}_{\mathrm{new},*})^\top (\Phi_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\Phi_{/i})^{-1}\boldsymbol{\phi}_i.$

We denote the first terms in $\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ and $\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ as $\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ and $\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$, and the last terms as $\mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$ and $\mathcal{D}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_{\mathrm{new}})$.

First, we consider the case when $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$. By our Theorem 2.13 and Lemma B.10, we get that

$$\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) \geq O_p\Big(\frac{\|\boldsymbol{\beta}^*\|^2}{n}\Big), \quad \mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) \leq O_p\Big(\frac{\|\boldsymbol{\beta}^*\|^3}{n^{1.5-\epsilon}}\Big).$$

Then, by the assumption $\|\boldsymbol{\beta}^*\| \leq n^{0.5-\epsilon}$, we obtain that with high probability, for sufficiently large n and p,

$$\frac{\mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \leq \frac{1}{2}.$$

Thus,

$$\begin{aligned}
&\frac{\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} = \frac{\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i) + \mathcal{D}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) + \mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \\
\geq\ & \frac{\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i) + \mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \geq \frac{\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \frac{1}{1 + \frac{\mathcal{D}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)}}.
\end{aligned}$$

We now focus on the ratio $\frac{\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)}$. From Lemma B.12, we get that with probability greater than $1 - C\exp(-c\epsilon^2 k)$,

$$\frac{1}{\kappa}(1-\epsilon)\frac{k}{d} \leq \frac{\mathcal{I}_*^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}_*^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \leq \min\Big(\kappa(1+\epsilon)\frac{k}{d}, 1\Big). \tag{40}$$

Hence, there exists a constant $c^{\mathrm{TRAK}} > 0$ such that with high probability,

$$\frac{\mathcal{I}^{\mathrm{TRAK}}(\mathbf{z}_i, \mathbf{z}_i)}{\mathcal{I}^{\mathrm{ALO}}(\mathbf{z}_i, \mathbf{z}_i)} \geq c^{\mathrm{TRAK}}\frac{k}{d}.$$

Next, we consider the case when $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent. Define

$$\mathbf{u}_i \coloneqq (\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i})^{-1}\mathbf{g}_i \quad \text{and} \quad \mathbf{u}_{i*} \coloneqq (\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i})^{-1}\mathbf{g}_{i,*}.$$

$$\mathbf{v}_i \coloneqq S^\top (S\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}S^\top)^{-1}S\mathbf{g}_i \quad \text{and} \quad \mathbf{v}_{i*} \coloneqq S^\top (S\boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}S^\top)^{-1}S\mathbf{g}_{i,*}.$$

Note that our goal is to calculate an upper bound for $\mathbf{g}_{\mathrm{new}}^\top \mathbf{v}_i$. Using Assumption B.4 and the independence of $\mathbf{g}_{\mathrm{new},*}$ and $\mathbf{v}_i$ we have $\mathbf{g}_{\mathrm{new},*}^\top \mathbf{v}_i$ is $\frac{\|\mathbf{v}_{i*}\|\mathrm{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}$-subGaussian. Thus,

$$|\mathbf{g}_{\mathrm{new},*}^\top \mathbf{v}_i| = O_p\Big(\frac{\|\mathbf{v}_{i*}\|\mathrm{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}\Big). \tag{41}$$

Now, by Lemma B.10, for any $\epsilon > 0$, $\mathbf{g}_i = \mathbf{g}_{i,*} + (\mathbf{g}_i - \mathbf{g}_{i,*}) = \mathbf{g}_{i,*} + O_{\mathrm{P}}\Big(\sqrt{\frac{\log n}{n^{1-2\epsilon}}}\Big)$. Consequently,

$$|(\mathbf{g}_{\mathrm{new}} - \mathbf{g}_{\mathrm{new},*})^\top \mathbf{v}_i| = O_p\Big(\frac{\mathrm{poly}(\log \mathrm{n})}{n^{0.5-\epsilon}}\|\mathbf{v}_{i*}\|\Big) = O_p\Big(\frac{\|\boldsymbol{\beta}^*\|}{n^{0.5-\epsilon}}\frac{\|\mathbf{v}_{i*}\|\mathrm{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}\Big). \tag{42}$$

Hence we have

$$|\mathbf{g}_{\mathrm{new}}^\top \mathbf{v}_i| = |\mathbf{g}_{\mathrm{new},*}^\top \mathbf{v}_i| + O_p\Big(\frac{\|\boldsymbol{\beta}^*\|}{n^{0.5-\epsilon}}\frac{\|\mathbf{v}_{i*}\|\mathrm{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}\Big) = O_p\Big(\frac{\|\mathbf{v}_{i*}\|\mathrm{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}\Big), \tag{43}$$

where to get the last equality we have used (41). To bound $\|\mathbf{v}_{i*}\|^2$ note that

$$\begin{aligned}
&\sigma_{\min}\big((SH_iS^\top)^{-0.5}SS^\top(SH_iS^\top)^{-0.5}\big)\mathbf{g}_{i,*}^\top S^\top (SH_iS^\top)^{-1}S\mathbf{g}_{i,*} \leq \|\mathbf{v}_{i*}\|^2 \\
\leq\ & \sigma_{\max}\big((SH_iS^\top)^{-0.5}SS^\top(SH_iS^\top)^{-0.5}\big)\mathbf{g}_{i,*}^\top S^\top (SH_iS^\top)^{-1}S\mathbf{g}_{i,*}
\end{aligned}$$

where we denote $H_i = \boldsymbol{G}_{/i}^\top \operatorname{diag}[\ddot{\boldsymbol{\ell}}_{/i}(\breve{\boldsymbol{\beta}})]\boldsymbol{G}_{/i}$. Observe that $\sigma_{\max}((SH_iS^\top)^{-0.5}SS^\top(SH_iS^\top)^{-0.5}) \leq \sigma_{\min}(H_i)^{-1}$ and $\sigma_{\min}((SH_iS^\top)^{-0.5}SS^\top(SH_iS^\top)^{-0.5}) \geq \sigma_{\max}(H_i)^{-1}$. Hence,

$$\frac{\mathbf{g}_{i,*}^\top S^\top (SH_iS^\top)^{-1}S\mathbf{g}_{i,*}}{\sigma_{\max}(H_i)} \leq \|\mathbf{v}_{i*}\|^2 \leq \frac{\mathbf{g}_{i,*}^\top S^\top (SH_iS^\top)^{-1}S\mathbf{g}_{i,*}}{\sigma_{\min}(H_i)}.$$

Further we can show that

$$\sigma_{\min}(H_i^{-1})\mathbf{g}_{i,*}^\top H_i^{-1}\mathbf{g}_{i,*} \leq \|\mathbf{u}_{i*}\|^2 \leq \sigma_{\max}(H_i^{-1})\mathbf{g}_{i,*}^\top H_i^{-1}\mathbf{g}_{i,*} \tag{44}$$

Combining these two equations with Lemma B.12, we finally get with high probability,

$$\frac{1}{\kappa^2}(1-\epsilon)\frac{k}{d} \leq \frac{\|\mathbf{v}_{i*}\|^2}{\|\mathbf{u}_{i*}\|^2} \leq \min\Big(\kappa(1+\epsilon)\frac{k}{d}, 1\Big)\kappa.$$

Combining this bound with (43) leads to

$$|\mathbf{g}_{\text{new}}^\top \mathbf{v}_i| = O_p\Big(\frac{\|\mathbf{u}_{i*}\|\text{poly}(\log(n))}{\|\boldsymbol{\beta}^*\|}\sqrt{\frac{k}{d}}\Big).$$

Since

$$\|\mathbf{u}_{i*}\| = O_p\Big(\frac{\|\boldsymbol{\beta}^*\|^2}{n^{1-\epsilon}}\Big),$$

and with $\kappa = O_p(\text{poly}(\log \text{n}))$, we obtain

$$|\mathbf{g}_{\text{new}}^\top \mathbf{v}_i| = O_p\Big(\frac{\|\boldsymbol{\beta}^*\|\sqrt{k}}{n^{1-\epsilon}\sqrt{d}}\text{poly}(\log(n))\Big).$$

B.3 Technical lemmas

Lemma B.9. *Assume that $\boldsymbol{G}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{G}$ and $\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G}$ are positive definite, and both $\boldsymbol{D}$ and $\boldsymbol{\Gamma}$ are diagonal matrices. Let $\mathbf{g}$ be a vector independent of $\mathbf{G}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{D}$. If $\|(\mathbf{G}^\top\boldsymbol{D}\mathbf{G})^{-1}(\mathbf{G}^\top\boldsymbol{\Gamma}\mathbf{G})\| < 1$, then there exists an absolute constant M_0 such that*

$$\left\|\left((\boldsymbol{G}^\top(\boldsymbol{D}+\boldsymbol{\Gamma})\boldsymbol{G})^{-1} - (\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G})^{-1}\right)\mathbf{g}\right\| \leq M_0\left\|(\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G})^{-1}\mathbf{G}^\top\boldsymbol{\Gamma}\mathbf{G}(\boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G})^{-1}\mathbf{g}\right\|.$$

Proof of Lemma B.9. Let $A = \boldsymbol{G}^\top\boldsymbol{D}\boldsymbol{G}$ and $\Delta = \boldsymbol{G}^\top\boldsymbol{\Gamma}\boldsymbol{G}$. By the Neumann series expansion, since $\|A^{-1}\Delta\| < 1$,

$$(A+\Delta)^{-1} = A^{-1} - A^{-1}\Delta A^{-1} + A^{-1}\Delta A^{-1}\Delta A^{-1} - \cdots.$$

Thus,

$$(A+\Delta)^{-1} - A^{-1} = -A^{-1}\Delta A^{-1} + A^{-1}\Delta A^{-1}\Delta A^{-1} - \cdots.$$

Multiplying by $\mathbf{g}$ and taking norms,

$$\begin{aligned}\|((A+\Delta)^{-1} - A^{-1})\mathbf{g}\| &\leq \|A^{-1}\Delta A^{-1}\mathbf{g}\| + \|A^{-1}\Delta A^{-1}\Delta A^{-1}\mathbf{g}\| + \cdots \\ &\leq \|A^{-1}\Delta A^{-1}\mathbf{g}\|\left(1 + \|A^{-1}\Delta\| + \|A^{-1}\Delta\|^2 + \cdots\right) \\ &= \|A^{-1}\Delta A^{-1}\mathbf{g}\| \cdot \frac{1}{1-\|A^{-1}\Delta\|}.\end{aligned}$$

Since $\|A^{-1}\Delta\| < 1$ by assumption, the geometric series converges. Taking $M_0 = \frac{1}{1-\|A^{-1}\Delta\|}$ completes the proof. □

Lemma B.10. *Suppose Assumptions B.4 hold. Then for all $\epsilon > 0$, with probability at least $1 - q_n - n^{-\epsilon}$, there exists a constant $C_{\boldsymbol{\beta}}, C_{\mathbf{g}} > 0$, independent of n, p and d, and*

$$\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\| \leq C_{\boldsymbol{\beta}}\|\boldsymbol{\beta}^*\|^2\sqrt{\frac{\log(n)}{n^{1-2\epsilon}}},$$

and

$$\|\mathbf{g}_i - \mathbf{g}_{i,*}\| \leq C_{\mathbf{g}}\sqrt{\frac{\log(n)}{n^{1-2\epsilon}}},$$

Proof of Lemma B.10. Denote $\dot{\ell}_{i*} = \dot{\ell}\big(y_i, f(\mathbf{x}_i, \boldsymbol{\beta}^*)\big)$. By Assumption B.6, we know that $\max_{i\in[n]} |\dot{\ell}_{i*}| \le \tilde{C}_1 \log n$. Also, using Lemma 2(1) in Jin et al. [2019], we know that $\mathbb{E}\|\mathbf{g}_{i,*}\|^2 \le 2\sigma^2$, then using $\mathbb{E}[\dot{\ell}_{i*}\mathbf{g}_{i,*}] = 0$ and $\{\dot{\ell}_{i*}\mathbf{g}_{i,*}\}_{i\in[n]}$ are independent, we get that

$$\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{i*}\mathbf{g}_{i,*}\|^2 = \frac{1}{n^2}\mathbb{E}\sum_{i=1}^n \dot{\ell}_{i*}^2 \mathbf{g}_{i,*}^\top \mathbf{g}_{i,*} \le \frac{2\tilde{C}_1\sigma^2 \log n}{n}.$$

Then using moment inequality, we get that

$$\mathbb{P}(\|\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{i*}\mathbf{g}_{i,*}\| \ge t) \le \frac{2\tilde{C}_1\sigma^2 \log n}{nt^2}.$$

Thus we can say that for all $\epsilon > 0$, $t = \frac{\sqrt{\log n}}{n^{\frac{1}{2}-\epsilon}}$, when n big enough, then with probability $\ge 1 - n^{-\epsilon}$, there exists a constant C_g such that

$$\|\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{i*}\mathbf{g}_{i,*}\| \le C_g \frac{\log^{\frac{1}{2}}(n)}{n^{\frac{1}{2}-\epsilon}}. \tag{45}$$

Using Mean-Value Theorem, denote $\dot{\ell}_i = \dot{\ell}\big(y_i, f(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\big)$, we know that there exists $\bar{\boldsymbol{\beta}} = t\boldsymbol{\beta}^* + (1-t)\hat{\boldsymbol{\beta}}$ where $t \in [0,1]$, such that

$$0 = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_i \mathbf{g}_i = \frac{1}{n}\sum_{i=1}^n \dot{\ell}_{i*}\mathbf{g}_{i,*} + \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(y_i, f(x_i, \bar{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

then

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \le \|(\frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(y_i, f(x_i, \bar{\boldsymbol{\beta}})))^{-1}\| \|\frac{1}{n}\sum_{i=1}^n \dot{\ell}_{i*}\mathbf{g}_{i,*}\|,$$

by Assumption B.3 , we know that with probability $\ge 1 - q_n$,

$$\sigma_{\min}(\frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(y_i, f(x_i, \bar{\boldsymbol{\beta}}))) \ge c_1 \|\boldsymbol{\beta}^*\|^{-2},$$

combining (45), we know that

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \le \frac{C_g}{c_1}\frac{\log^{\frac{1}{2}}(n)\|\boldsymbol{\beta}^*\|^2}{n^{\frac{1}{2}-\epsilon}} := C_{\boldsymbol{\beta}}\frac{\log^{\frac{1}{2}}(n)\|\boldsymbol{\beta}^*\|^2}{n^{\frac{1}{2}-\epsilon}}$$

Finally, using Assumption B.5, note that there also $\bar{\boldsymbol{\beta}}' = t'\boldsymbol{\beta}^* + (1-t')\hat{\boldsymbol{\beta}}$ where $t' \in [0,1]$, such that

$$\|\mathbf{g}_i - \mathbf{g}_{i,*}\| \le \|\nabla^2 f(\mathbf{x}_i, \bar{\boldsymbol{\beta}})\| \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \le C_2\|\boldsymbol{\beta}^*\|^{-2}\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\| \le C_2 C_{\boldsymbol{\beta}}\frac{\log^{\frac{1}{2}}(n)}{n^{\frac{1}{2}-\epsilon}} := C_{\mathbf{g}}\frac{\log^{\frac{1}{2}}(n)}{n^{\frac{1}{2}-\epsilon}}.$$

□

Lemma B.11 (Infinity-norm transfer for weighted Gram matrices)**.** *Let $\mathbf{G}, \mathbf{G}_0 \in \mathbb{R}^{m\times d}$ have full column rank, and let $\boldsymbol{D}, \boldsymbol{D}_0 \in \mathbb{R}^{m\times m}$ be diagonal matrices satisfying*

$$\mu I_m \preceq \boldsymbol{D} \preceq B I_m, \qquad \mu I_m \preceq \boldsymbol{D}_0 \preceq B I_m, \qquad 0 < \mu \le B.$$

Define

$$A = \mathbf{G}^\top \boldsymbol{D}\mathbf{G}, \qquad A_0 = \mathbf{G}_0^\top \boldsymbol{D}_0 \mathbf{G}_0,$$

and suppose that

$$\max\{\|A^{-1}\|_{\mathrm{op}}, \|A_0^{-1}\|_{\mathrm{op}}\} \le a.$$

Set

$$\rho = \|\mathbf{G} - \mathbf{G}_0\|_{\mathrm{op}}, \qquad \gamma = \|\boldsymbol{D} - \boldsymbol{D}_0\|_F.$$

Then, for every $\mathbf{g} \in \mathbb{R}^d$,

$$\|\mathbf{G}A^{-1}\mathbf{g}\|_\infty \le \left(1 + \frac{\gamma}{\mu}\right)\|\mathbf{G}_0 A_0^{-1}\mathbf{g}\|_\infty + \left(1 + \frac{2B}{\mu}\right) a\rho\|\mathbf{g}\|_2. \tag{46}$$

Proof. Write

$$E = \mathbf{G} - \mathbf{G}_0, \quad P = \mathbf{G}A^{-1}\mathbf{G}^\top, \quad \mathbf{v}_0 = A_0^{-1}\mathbf{g}, \quad \mathbf{u}_0 = \mathbf{G}_0\mathbf{v}_0, \quad \mathbf{u} = \mathbf{G}A^{-1}\mathbf{g}.$$

Since $A \succeq \mu\mathbf{G}^\top\mathbf{G}$, we have

$$\|\mathbf{G}A^{-1}\|_{\mathrm{op}}^2 = \|A^{-1}\mathbf{G}^\top\mathbf{G}A^{-1}\|_{\mathrm{op}} \le \frac{\|A^{-1}\|_{\mathrm{op}}}{\mu} \le \frac{a}{\mu}.$$

The same argument gives $\|\mathbf{G}_0A_0^{-1}\|_{\mathrm{op}} \le \sqrt{a/\mu}$. Furthermore,

$$P = \boldsymbol{D}^{-1/2}\Pi\boldsymbol{D}^{-1/2}, \qquad \Pi = \boldsymbol{D}^{1/2}\mathbf{G}A^{-1}\mathbf{G}^\top\boldsymbol{D}^{1/2},$$

where Π is an orthogonal projector. Thus $\|P\|_{\mathrm{op}} \le \mu^{-1}$.

The exact identity

$$A - A_0 = \mathbf{G}^\top(\boldsymbol{D} - \boldsymbol{D}_0)\mathbf{G}_0 + E^\top\boldsymbol{D}_0\mathbf{G}_0 + \mathbf{G}^\top\boldsymbol{D}E$$

together with $A^{-1} - A_0^{-1} = -A^{-1}(A - A_0)A_0^{-1}$ yields

$$\begin{aligned}\mathbf{u} = \mathbf{u}_0 &- P(\boldsymbol{D} - \boldsymbol{D}_0)\mathbf{u}_0 + (I_m - P\boldsymbol{D})E\mathbf{v}_0 \\ &- \mathbf{G}A^{-1}E^\top\boldsymbol{D}_0\mathbf{u}_0.\end{aligned}$$

Because $\boldsymbol{D} - \boldsymbol{D}_0$ is diagonal,

$$\|P(\boldsymbol{D} - \boldsymbol{D}_0)\mathbf{u}_0\|_\infty \le \frac{1}{\mu}\|(\boldsymbol{D} - \boldsymbol{D}_0)\mathbf{u}_0\|_2 \le \frac{\gamma}{\mu}\|\mathbf{u}_0\|_\infty.$$

The remaining terms satisfy

$$\|(I_m - P\boldsymbol{D})E\mathbf{v}_0\|_\infty \le \left(1 + \frac{B}{\mu}\right)\rho a\|\mathbf{g}\|_2$$

and

$$\|\mathbf{G}A^{-1}E^\top\boldsymbol{D}_0\mathbf{u}_0\|_\infty \le \sqrt{\frac{a}{\mu}}\,\rho B\sqrt{\frac{a}{\mu}}\,\|\mathbf{g}\|_2 = \frac{B}{\mu}\,a\rho\|\mathbf{g}\|_2.$$

Combining these inequalities proves (46). □

Lemma B.12. *Let $S \in \mathbb{R}^{k\times d}$ be such that $\forall i \in [k], j \in [d]$, $S_{ij} \overset{i.i.d.}{\sim} N(0,1)$. Let A in $\mathbb{R}^{d\times d}$ be a symmetric positive definite matrix with condition number $\left(\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right)$ upper bounded by $\kappa > 0$. Here $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the extremal eigenvalues of A. Let $u \in \mathbb{R}^d$. Then with probability greater than $1 - C\exp(-c\epsilon^2 k)$, where $C, c > 0$ are absolute constants,*

$$\frac{1}{\kappa}(1-\epsilon)\frac{k}{d} \le \frac{u^TS^\top(SAS^\top)^{-1}Su}{u^TA^{-1}u} \le \min(\kappa(1+\epsilon)\frac{k}{d}, 1) \tag{47}$$

Proof of Lemma B.12. Define,

$$\begin{aligned}P &\coloneqq A^{1/2}S^\top(SAS^\top)^{-1}SA^{1/2} \\ \hat{A}^{-1} &\coloneqq S^\top(SAS^\top)^{-1}S \\ w &\coloneqq A^{-1/2}u\end{aligned}$$

See that P is an orthogonal projection matrix, i.e. for any $l \in \mathbb{R}^d$,

$$\|l\|^2 = \|Pl\|^2 + \|(I-P)l\|^2 \tag{48}$$

Also since $P^2 = P$, $\|Pw\|^2 = w^TPw$. See also now,

$$\begin{aligned}u^TS^\top(SAS^\top)^{-1}Su = u^\top\hat{A}^{-1}u = w^TPw = \|Pw\|^2 \\ u^TA^{-1}u = \|w\|^2\end{aligned}$$

Thus, in Lemma B.12, we are now trying to derive a concentration for $\frac{\|Pw\|^2}{\|w\|^2}$. Consider the eigenvalue decomposition of A, $A = Q\Lambda Q^\top$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$. Define $V \coloneqq Q^T S^\top$. Since Q is orthogonal, we again have $\forall i \in [d], j \in [k], V_{ij} \overset{i.i.d.}{\sim} N(0,1)$, and $V^T V$ is almost surely positive definite. See,

$$\begin{aligned} P &= A^{1/2} S^\top (SAS^\top)^{-1} SA^{1/2} \\ &= Q\Lambda^{1/2} V (V^\top \Lambda V)^{-1} V^\top \Lambda^{1/2} Q^\top = Q P_V Q^\top, \end{aligned}$$

where $P_V \coloneqq \Lambda^{1/2} V (V^\top \Lambda V)^{-1} V^\top \Lambda^{1/2}$. It is easy to see that P_V is an orthogonal projection matrix satisfying Equation (48) if we replace P by P_V. Define,

$$\Pi \coloneqq V(V^T V)^{-1} V^\top; \qquad h \coloneqq \Lambda^{1/2} w$$

We will first show that, almost surely,

$$\frac{1}{\kappa} \frac{\|\Pi h\|^2}{\|h\|^2} \leq \frac{\|P_V w\|^2}{\|w\|^2} \leq \kappa \frac{\|\Pi h\|^2}{\|h\|^2} \tag{49}$$

By definition of Λ, we have,

$$\begin{aligned} &\lambda_d I \preceq \Lambda \preceq \lambda_1 I \\ &\overset{(i)}{\implies} \lambda_d V^T V \preceq V^\top \Lambda V \preceq \lambda_1 V^T V \\ &\implies \frac{1}{\lambda_1} (V^T V)^{-1} \preceq (V^\top \Lambda V)^{-1} \preceq \frac{1}{\lambda_d} (V^T V)^{-1}, \end{aligned}$$

where (i) is since for $B, C, D \in \mathbb{R}^{d \times d}$ and any $j \in \mathbb{R}^d$, $B \preceq C \implies j^\top (C - B) j \geq 0 \implies j^T D^\top (C - B) D j \geq 0$. Thus we have almost surely,

$$\frac{1}{\lambda_1} \|\Pi h\|^2 \leq \|P_V w\|^2 \leq \frac{1}{\lambda_d} \|\Pi h\|^2 \tag{50}$$

Also it is clear that we have the following bounds for $\|h\|^2 = w^\top \Lambda w$, $\lambda_d \|w\|^2 \leq \|h\|^2 \leq \lambda_1 \|w\|^2$. Combining with Equation (50), we have shown Equation (49) □

Using the fact that the quadratic forms of a Gaussian vector projected onto an independent random subspace follow a Beta distribution, Chapter 7 Anderson [1958], we have $\frac{\|\Pi h\|^2}{\|h\|^2} \sim \text{Beta}\,(k/2, (d-k)/2)$. By concentration of Beta distribution (using standard concentration inequalities from Vershynin [2018]), we have,

$$P\left(\left| \frac{\|\Pi h\|^2}{\|h\|^2} - \frac{k}{d} \right| > \epsilon \right) \leq C \exp(-ck\epsilon^2) \tag{51}$$

By Equation (49), we have almost surely $\frac{1}{\kappa} \frac{\|\Pi Q^T h\|^2}{\|Q^T h\|^2} \leq \frac{\|P_V Q^T w\|^2}{\|Q^T w\|^2} \leq \kappa \frac{\|\Pi Q^T h\|^2}{\|Q^T h\|^2}$. Thus since $P = Q P_V Q^\top$ and by Equation (51), we have shown Lemma B.12.

B.4 Discussion of more general correlated cases

Throughout this paper, we evaluate the accuracy of TRAK in two distinct settings:

- **Correlated case:** The test data point is identical to one of the training samples.
- **Independent case:** The test data point is independent of all training samples.

These represent two extreme scenarios that may arise in practice, which naturally raises the question of how TRAK behaves in intermediate regimes. The aim of this section is to address this question. For brevity, we present only the main results along with brief proof sketches. The formal presentation of the results and full proofs will be presented in a longer version of the paper.

Suppose we wish to evaluate the influence of $\mathbf{x}_i$ on a test point $\mathbf{x}_{\text{new}}$, where both vectors are zero-mean and satisfy

$$\mathbb{E}(\mathbf{x}_i^\top \mathbf{x}_{\text{new}}) = \alpha_n \mathbb{E}(\mathbf{x}_i^\top \mathbf{x}_i).$$

To simplify the analysis, assume that $\mathbf{x}_i, \mathbf{x}_{\text{new}} \sim \mathcal{N}(0, \mathbf{\Sigma})$. Then $\mathbf{x}_{\text{new}}$ can be represented as

$$\mathbf{x}_{\text{new}} = \alpha_n \mathbf{x}_i + \sqrt{1-\alpha_n^2}\, \mathbf{x}'_{\text{new}}, \tag{52}$$

where $\mathbf{x}'_{\text{new}}$ is an independent copy of $\mathbf{x}_i$.

We begin by analyzing the magnitude of the TRAK influence function $\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_{\text{new}})$. Define

$$\begin{aligned} \mathbf{g}_{i,*} &= \nabla f(\mathbf{x}_i; \boldsymbol{\beta}_*), \\ \tilde{\mathbf{g}}_{i,*} &= \nabla f\big(\alpha_n \mathbf{x}_i + \sqrt{1-\alpha_n^2}\, \mathbf{x}'_{\text{new}}; \boldsymbol{\beta}_*\big). \end{aligned} \tag{53}$$

Following calculations similar to (24) and (27), we have

$$\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_{\text{new}}) = \Theta_p\left(\tilde{\mathbf{g}}_{i,*}^\top \mathbf{g}_{i,*} \left(\frac{n \operatorname{poly}(\log n)}{\|\boldsymbol{\beta}_*\|^2}\right)^{-1}\right).$$

Thus, the analysis reduces to understanding the effect of α_n on the inner product $\tilde{\mathbf{g}}_{i,*}^\top \mathbf{g}_{i,*}$. Using Lemma B.13, we obtain

$$|\mathbb{E}\tilde{\mathbf{g}}_{i,*}^\top \mathbf{g}_{i,*}| \le \alpha_n |\mathbb{E}\mathbf{g}_{i,*}^\top \mathbf{g}_{i,*}|$$

Moreover, this upper bound is sharp for a broad class of functions $f(\cdot;\cdot)$. In particular, if the projection of f onto the first Hermite polynomial has non-negligible energy, then

$$|\mathbb{E}\tilde{\mathbf{g}}_{i,*}^\top \mathbf{g}_{i,*}| = \Omega_p(\alpha_n |\mathbb{E}\mathbf{g}_{i,*}^\top \mathbf{g}_{i,*}|)$$

Hence, we can conclude that under reasonably general conditions we have

$$\left|\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_{\text{new}})\right| = \Theta_p\left(\alpha_n\, \mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_i)\right). \tag{54}$$

Finally, the other results in this paper concerning $\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_i)$ can be extended in a similar manner to the setting where the test data has correlation α_n with $\mathbf{x}_i$. Overall, the magnitude of the influence scales by a factor of α_n, and the corresponding error terms scale proportionally as well.

Lemma B.13. *Let $\mathbf{u}, \mathbf{v} \sim \mathcal{N}(0, \mathbf{\Sigma})$ be independent, and define*

$$\boldsymbol{w}_\alpha = \alpha \boldsymbol{u} + \sqrt{1-\alpha^2}\, \boldsymbol{v}, \qquad 0 \le \alpha \le 1.$$

Let $h : \mathbb{R}^n \to \mathbb{R}$ satisfy $\mathbb{E}[h(\boldsymbol{u})^2] < \infty$ and $|\mathbb{E}h(\boldsymbol{u})| = 0$. Then

$$\mathbb{E}\big[h(\boldsymbol{w}_\alpha) h(\mathbf{u})\big] \le \alpha\, \mathbb{E}[h(\mathbf{u})^2].$$

Proof. Write

$$\boldsymbol{u} = \mathbf{\Sigma}^{1/2} \tilde{\boldsymbol{u}}, \qquad v = \mathbf{\Sigma}^{1/2} \tilde{\boldsymbol{v}},$$

where $\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}} \sim \mathcal{N}(0, I_n)$ are independent. Define

$$\phi(\tilde{\boldsymbol{u}}) = h(\mathbf{\Sigma}^{1/2} \tilde{\boldsymbol{u}}).$$

Then

$$h(\boldsymbol{w}_\alpha) = \phi\big(\alpha \tilde{\boldsymbol{u}} + \sqrt{1-\alpha^2}\tilde{\boldsymbol{v}}\big).$$

Expand ϕ in multivariate Hermite polynomials:

$$\phi = \phi_0 + \phi_1 + \phi_2 + \cdots,$$

where ϕ_m is the degree-m Hermite component. Then

$$\mathbb{E}\big[\phi(\tilde{\boldsymbol{u}})\phi(\alpha\tilde{\boldsymbol{u}} + \sqrt{1-\alpha^2}\tilde{\boldsymbol{v}})\big] = \sum_{m \ge 0} \alpha^m \|\phi_m\|_{L^2}^2.$$

The degree-zero term is

$$\|\phi_0\|_{L^2}^2 = (\mathbb{E}h(\mathbf{u}))^2 = 0.$$

Hence

$$\mathbb{E}\big[h(\boldsymbol{w}_\alpha) h(\boldsymbol{u})\big] = \sum_{m \ge 1} \alpha^m \|\phi_m\|_{L^2}^2.$$

The proof follows by noting that $0 \leq \alpha \leq 1$,

$$\alpha^m \leq \alpha \quad \text{for all } m \geq 1.$$

Therefore,

$$\mathbb{E}\big[h(\boldsymbol{w}_\alpha)h(\boldsymbol{u})\big] \leq \alpha \sum_{m\geq 1} \|\phi_m\|_{L^2}^2.$$

But

$$\sum_{m\geq 1} \|\phi_m\|_{L^2}^2 = \mathrm{Var}(h(\boldsymbol{u})).$$

Thus,

$$\mathbb{E}\big[h(\boldsymbol{w}_\alpha)h(\boldsymbol{u})\big] \leq \alpha\, \mathrm{Var}(h(\boldsymbol{u})).$$

□

B.5 Confirming our assumptions on the loss function

In this section, we examine Assumption B.6 in the context of logistic regression, linear regression, and Poisson regression. Our goal is to justify that the orders are reasonable.

B.5.1 Logistic Regression

Since

$$\dot{\ell}_i(\boldsymbol{\beta}) = -y_i + \frac{e^{\mathbf{g}_i^\top \boldsymbol{\beta}}}{1+e^{\mathbf{g}_i^\top \boldsymbol{\beta}}}, \quad \ddot{\ell}_i(\boldsymbol{\beta}) = \frac{e^{\mathbf{g}_i^\top \boldsymbol{\beta}}}{\left(1+e^{\mathbf{g}_i^\top \boldsymbol{\beta}}\right)^2}, \quad \dddot{\ell}_i(\boldsymbol{\beta}) = \frac{e^{\mathbf{g}_i^\top \boldsymbol{\beta}}\left(1-e^{\mathbf{g}_i^\top \boldsymbol{\beta}}\right)}{\left(1+e^{\mathbf{g}_i^\top \boldsymbol{\beta}}\right)^3},$$

using simple algebra, it is straightforward to show that for any $\boldsymbol{\beta}$, we have

$$\|\dot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_\infty \leq 1, \quad \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_\infty \leq \tfrac{1}{4}, \quad \|\dddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_\infty \leq \tfrac{1}{10}.$$

Hence, the first inequality of Assumption B.6 holds. Note that we can assume $\mathbf{g}_i^\top \boldsymbol{\beta} = O_p(1)$, so the second inequality of Assumption B.6 also holds. oreover,

$$\|\ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta}+\delta) - \ddot{\boldsymbol{\ell}}_{/i}(\boldsymbol{\beta})\|_2 \leq \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}+\delta) - \ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})\|_2 = \sqrt{\sum_i \left(\ddot{\ell}_i(\boldsymbol{\beta}+\delta) - \ddot{\ell}_i(\boldsymbol{\beta})\right)^2}.$$

Using the mean-value theorem, where $\epsilon_i \in [0, \delta_i]$, we have

$$= \sqrt{\sum_i \dddot{\ell}_i(\boldsymbol{\beta}+\epsilon_i)^2 \left(\mathbf{g}_i^\top \delta\right)^2} \leq \sqrt{\delta^\top \mathbf{G}^\top \mathbf{G} \delta} \leq \sqrt{\sigma_{\max}(\mathbf{G}^\top \mathbf{G})}\, \|\delta\|_2.$$

Based on the inequality above, we obtain

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}\Big((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}\Big) - \ddot{\boldsymbol{\ell}}_{/i}\Big(\hat{\boldsymbol{\beta}}\Big)\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq \sup_{t\in[0,1]} \frac{(1-t)\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2 \sqrt{\sigma_{\max}(\mathbf{G}^\top \mathbf{G})}}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \leq \sqrt{\sigma_{\max}(\mathbf{G}^\top \mathbf{G})}.$$

By Assumption B.3, we know that $\sigma_{\max}\big(\mathbf{G}^\top \mathbf{G}\big) = O_p\big(n/\|\boldsymbol{\beta}^*\|^2\big)$, thus the third inequality in Assumption B.6 holds.

B.5.2 Linear Regression

Considering $\ell\big(y \mid \boldsymbol{x}^\top \boldsymbol{\beta}\big) = \frac{1}{2}\big(y - \boldsymbol{x}^\top \boldsymbol{\beta}\big)^2$, we have $\dot{\ell}_i(\boldsymbol{\beta}) = -\big(y_i - \mathbf{g}_i^\top \boldsymbol{\beta}\big)$ and $\ddot{\ell}_i(\boldsymbol{\beta}) = 1$. Therefore, the second and third inequalities in Assumption B.6 hold directly. Moreover, since $y \sim \mathcal{N}\big(\boldsymbol{x}^\top \boldsymbol{\beta}^*, \sigma_\epsilon^2\big)$, by properties of sub-Gaussian random variables we obtain

$$\max_{i\in[n]} \big|\dot{\ell}_i(\boldsymbol{\beta})\big| \preceq \sqrt{\log n}.$$

Hence, the first inequality in Assumption B.6 holds.

B.5.3 Poisson Regression

Consider

$$\ell\big(y \mid \boldsymbol{x}^\top\boldsymbol{\beta}\big) = h\big(\boldsymbol{x}^\top\boldsymbol{\beta}\big) - y\log h\big(\boldsymbol{x}^\top\boldsymbol{\beta}\big) + \log y!,$$

with the conditional mean $\mathbb{E}[y \mid \boldsymbol{x}, \boldsymbol{\beta}] = h\big(\boldsymbol{x}^\top\boldsymbol{\beta}\big)$ where $h(s) = \log(1+e^s)$. Then the first- and second-order derivatives of the loss with respect to $s := \mathbf{x}^\top\boldsymbol{\beta}$ are

$$\dot{\ell}(y,s) = h'(s)\Big(1 - \frac{y}{h(s)}\Big) = \sigma(s)\Big(1 - \frac{y}{h(s)}\Big),$$

and

$$\ddot{\ell}(y,s) = \sigma(s)\big(1-\sigma(s)\big)\Big(1 - \frac{y}{h(s)}\Big) + y\frac{\sigma(s)^2}{h(s)^2}.$$

By Lemma 8 of Rad and Maleki [2020], we have

$$\sup_{t\in[0,1]} \frac{\|\ddot{\boldsymbol{\ell}}_{/i}\Big((1-t)\hat{\boldsymbol{\beta}}_{/i} + t\hat{\boldsymbol{\beta}}\Big) - \ddot{\boldsymbol{\ell}}_{/i}\Big(\hat{\boldsymbol{\beta}}\Big)\,\|_2}{\|\hat{\boldsymbol{\beta}}_{/i} - \hat{\boldsymbol{\beta}}\|_2} \;\;\leq\;\; (1+6\|y\|_\infty)\sqrt{\sigma_{\max}\Big(\mathbf{G}^\top\mathbf{G}\Big)},$$

and

$$\|\dot{\boldsymbol{\ell}}(\boldsymbol{\beta})\,\|_\infty \;\;\leq\;\; 1 + \|y\|_\infty.$$

where σ is sigmoid function. Since $\boldsymbol{x}_i^\top\boldsymbol{\beta} \sim \mathcal{N}\big(0, \boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta}\big)$, we have $\max_{i\in[n]}\big|\boldsymbol{x}_i^\top\boldsymbol{\beta}\big| = O(\log n)$ and $\|y\|_\infty = O(\mathrm{poly}\log n)$, both holding with high probability. Therefore, using Assumption B.3, the three inequalities in Assumption B.6 hold.

B.6 Confirming our assumptions on the nonlinear function $f(\cdot,\cdot)$

B.6.1 Verification of Example 2.2.

Consider the GLM where $f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^\top\boldsymbol{\beta}$. We verify Assumptions B.2, B.3, B.4 and B.5.

First, note that $p = d$ and $\nabla f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}$. Since we assume $|\beta_i^*| = O(1)$, we have $p \asymp \|\boldsymbol{\beta}^*\|^2$. Combined with Assumption B.1, this implies that Assumptions B.2 and B.4 hold, due to standard properties of sub-Gaussian vectors. Next, because $\nabla^2 f = 0$, Assumption B.5 holds automatically.

Furthermore, since

$$\mathbf{G}(\boldsymbol{\beta})^\top\mathbf{G}(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{X},$$

and $\sigma_{\max}(\boldsymbol{\Sigma}) \asymp \sigma_{\min}(\boldsymbol{\Sigma}) \asymp O_p(1/p)$ with $p \asymp \|\boldsymbol{\beta}^*\|^2$, and the vectors $\mathbf{x}_i,\ i \in [n]$, are independent sub-Gaussian, we obtain

$$\sigma_{\max}(\boldsymbol{X}^\top\boldsymbol{X}) \asymp \sigma_{\min}(\boldsymbol{X}^\top\boldsymbol{X}) \asymp O_p\left(\frac{1}{\|\boldsymbol{\beta}^*\|^2}\right),$$

which verifies Assumption B.3.

B.6.2 Verification of Example 2.3.

Setting and Assumptions. We verify Assumptions B.2, B.3, B.4 and B.5, in a linear-layer setting. Let the parameter vector be

$$\boldsymbol{\beta} = (\mathrm{vec}(\mathbf{W})^\top, \mathbf{v}^\top)^\top,$$

where $\mathbf{W} \in \mathbb{R}^{h\times p}$ satisfies

$$\sigma_{\min}(\mathbf{W}^\top\mathbf{W}) \asymp \sigma_{\max}(\mathbf{W}^\top\mathbf{W}).$$

We further assume $|v_l| = O(1)$; that is, there exist constants $v, V > 0$ such that

$$v \leq |v_l| \leq V, \qquad \forall l \in [h].$$

We also assume the initialization satisfies

$$p\|\mathbf{v}\|^2 \asymp \|\mathbf{W}\|_F^2,$$

which holds for common isotropic initializations.

Let σ be an activation function with

$$0 \le \sigma'(x) \le E.$$

For $\mathbf{x} \in \mathbb{R}^p$, define $\sigma(\mathbf{x})$ element-wise. Define the function

$$f(\mathbf{x}, \mathbf{W}, \mathbf{v}) = \mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x}).$$

The gradients are given by

$$\nabla_{\mathbf{v}} f(\mathbf{x}, \mathbf{W}, \mathbf{v}) = \sigma(\mathbf{W}\mathbf{x}) \in \mathbb{R}^h,$$
$$\nabla_{\mathbf{W}} f(\mathbf{x}, \mathbf{W}, \mathbf{v}) = \operatorname{diag}\left(\sigma'(\mathbf{W}\mathbf{x})\right) \mathbf{v}\, \mathbf{x}^\top \in \mathbb{R}^{h \times p}.$$

For notational convenience, for $l \in [h]$ and $i \in B \subset [n]$, define

$$A^B_{li} = \sigma'(\mathbf{W}_l^\top \mathbf{x}_i), \qquad D^B_{li} = v_l \sigma'(\mathbf{W}_l^\top \mathbf{x}_i).$$

Then the relationship between the Gram matrices is

$$D^B (D^B)^\top = \operatorname{diag}(\mathbf{v})\, A^B (A^B)^\top \operatorname{diag}(\mathbf{v}).$$

Assume that for any $B \subset [n]$ with $|B| \gg h$, with high probability we can find constants $\underline{\alpha}, \bar{\alpha} > 0$ such that

$$\bar{\alpha}|B| \;\ge\; \lambda_{\max}(A^B A^{B,\top}) \;\ge\; \lambda_{\min}(A^B A^{B,\top}) \;\ge\; \underline{\alpha}|B|.$$

Intuitively, if for all $l \in [h]$,

$$\sum_{i \in B} \sigma'^2(\mathbf{W}_l^\top \mathbf{x}_i) \asymp |B|,$$

then

$$\operatorname{tr}(A^B A^{B,\top}) = \sum_{l=1}^{h} \sum_{i \in B} \sigma'^2(\mathbf{W}_l^\top \mathbf{x}_i) \asymp h|B|.$$

If $A^B A^{B,\top}$ is well-conditioned, this implies the eigenvalue bounds above. Consequently, utilizing the bounds on v_l,

$$V^2\, \bar{\alpha}\, |B| \;\ge\; \lambda_{\max}(D^B D^{B,\top}) \;\ge\; \lambda_{\min}(D^B D^{B,\top}) \;\ge\; v^2\, \underline{\alpha}\, |B|.$$

We denote $A^{[n]} = A$ and $D^{[n]} = D$ for simplicity.

We further assume $n > d$ and $n \gg h^2$. Intuitively, if $n < d$ where d is the parameter dimension, the empirical Fisher

$$\mathbf{G}^\top \mathbf{G} = \sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i^\top$$

cannot be invertible. Similarly, in very wide networks, the sample size must dominate the width to obtain stable approximations.

Verification of Assumption B.2. First, we verify Assumption B.2. Note that

$$\|\nabla_{\mathbf{v}} f(\mathbf{x}, \mathbf{W}, \mathbf{v})\| = \|\sigma(\mathbf{W}\mathbf{x})\| \le E\|\mathbf{W}\mathbf{x}\| = O_{\text{HP}}(\sqrt{\operatorname{tr}(\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}^\top)}) = O_{\text{HP}}(\sqrt{\|\mathbf{W}\|_F^2 / \|\boldsymbol{\beta}^*\|^2}) = O_{\text{HP}}(1).$$

$$\|\nabla_{\mathbf{W}} f(\mathbf{x}, \mathbf{W}, \mathbf{v})\|_F^2 \le E^2 \|\mathbf{v}\|^2 \|\mathbf{x}\|^2 = O_{\text{HP}}\left(\frac{p\|\mathbf{v}\|^2}{\|\boldsymbol{\beta}^*\|^2}\right) = O_{\text{HP}}(1),$$

using the scaling assumption $p\|\mathbf{v}\|^2 \asymp \|\mathbf{W}\|_F^2$.

Verification of Assumption B.3. Next, we verify Assumption B.3. For simplicity, we consider

$$\mathbf{g}_i = \operatorname{vec}(\nabla_{\mathbf{W}} f(\mathbf{x}_i, \mathbf{W}, \mathbf{v})) = c_i \otimes \mathbf{x}_i,$$

where $c_i = D_{:,i}$ is the i-th column of D. Denoting $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_n)$, we know that

$$M := \frac{1}{n}\mathbf{G}^\top \mathbf{G} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i \mathbf{g}_i^\top = \frac{1}{n}\sum_{i=1}^{n} c_i c_i^\top \otimes \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n}\sum_{i=1}^{n} (c_i \otimes \mathbf{x}_i)(c_i \otimes \mathbf{x}_i)^\top.$$

To analyze the eigenvalues of M, consider any $q \in \mathbb{R}^{ph}$ such that $q = \mathrm{vec}(Q)$ and $\|q\| = \|Q\|_F = 1$. We only need to bound

$$q^\top M q = q^\top \left(\frac{1}{n} \sum_{i=1}^n (c_i \otimes \mathbf{x}_i)(c_i \otimes \mathbf{x}_i)^\top \right) q = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top Q c_i)^2.$$

Note that $Q \in \mathbb{R}^{p \times h}$. We consider two cases separately: $\mathrm{rank}(Q) = 1$ and $\mathrm{rank}(Q) > 1$.

First, consider the case when $\mathrm{rank}(Q) = 1$. There exist $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^h$ such that $Q = ab^\top$. WLOG, assume $\|a\|^2 = \|b\|^2 = 1$. Then we have:

$$\begin{aligned} q^\top M q &= a^\top \left[\tfrac{1}{n} (\mathbf{x}_1, \ldots, \mathbf{x}_n) \operatorname{diag}\left((b^\top c_1)^2, \ldots, (b^\top c_n)^2 \right) (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \right] a \\ &\geq \|a\|^2 \lambda_{\min}\left(\tfrac{1}{n} (\mathbf{x}_1, \ldots, \mathbf{x}_n) \operatorname{diag}\left((b^\top c_1)^2, \ldots, (b^\top c_n)^2 \right) (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \right). \end{aligned} \tag{55}$$

For any $b \in \mathbb{R}^h$ with $\|b\| = 1$, denote $S_b = \{i : \ (b^\top c_i)^2 \geq \frac{\underline{\alpha} v^2}{2}\}$ and let $S_b^c = [n] \setminus S_b$. We know that

$$\begin{aligned} \|b\|^2 \underline{\alpha} n v^2 \leq b^\top D D^\top b = \sum_{i=1}^n (b^\top c_i)^2 &= \sum_{i \in S_b} (b^\top c_i)^2 + \sum_{i \notin S_b} (b^\top c_i)^2 \\ &\leq \|b\|^2 E^2 V^2 h |S_b| + (n - |S_b|) \frac{\underline{\alpha} v^2}{2} \\ &= \left(E^2 V^2 h - \frac{\underline{\alpha} v^2}{2} \right) |S_b| + \frac{\underline{\alpha} v^2}{2} n, \end{aligned}$$

where in the second inequality we used $|c_i|^2 \leq E^2 V^2 h$. Thus, we get

$$|S_b| \geq \frac{\underline{\alpha} v^2}{2E^2 V^2 h - \underline{\alpha} v^2} n \gg h,$$

using $n \gg h^2$. We can further refine the inequality:

$$\begin{aligned} \|b\|^2 \underline{\alpha} n v^2 \leq b^\top D D^\top b = \sum_{i=1}^n (b^\top c_i)^2 &= \sum_{i \in S_b} (b^\top c_i)^2 + \sum_{i \notin S_b} (b^\top c_i)^2 \\ &\leq \|b\|^2 \bar{\alpha} V^2 |S_b| + (n - |S_b|) \frac{\underline{\alpha} v^2}{2} \\ &= \frac{\underline{\alpha} v^2}{2} n + |S_b| \left(\bar{\alpha} V^2 - \frac{\underline{\alpha} v^2}{2} \right), \end{aligned}$$

where we used $b^\top D^{S_b} D^{S_b, \top} b \leq \|b\|^2 |S_b| \bar{\alpha} V^2$ for $|S_b| \gg h$. Then we obtain

$$|S_b| \geq \frac{\underline{\alpha} v^2}{2\bar{\alpha} V^2 - \underline{\alpha} v^2} n := C(\underline{\alpha}, \bar{\alpha}, v, V) n.$$

Returning to (55), we have

$$\begin{aligned} q^\top M q &\geq \lambda_{\min}\left(\tfrac{1}{n} (\mathbf{x}_1, \ldots, \mathbf{x}_n) \operatorname{diag}\left((b^\top c_1)^2, \ldots, (b^\top c_n)^2 \right) (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \right) \\ &\geq \lambda_{\min}\left(\frac{1}{n} \sum_{i \in S_b} \mathbf{x}_i \mathbf{x}_i^\top (b^\top c_i)^2 \right) \geq \lambda_{\min}\left(\frac{1}{n} \sum_{i \in S_b} \mathbf{x}_i \mathbf{x}_i^\top \frac{\underline{\alpha} v^2}{2} \right) \\ &= \frac{\underline{\alpha} v^2}{2} \frac{|S_b|}{n} \lambda_{\min}\left(\frac{1}{|S_b|} \sum_{i \in S_b} \mathbf{x}_i \mathbf{x}_i^\top \right). \end{aligned}$$

By the sub-Gaussianity of i.i.d $\mathbf{x}_i$, we have

$$\lambda_{\min}\left(\frac{1}{|S_b|} \sum_{i \in S_b} \mathbf{x}_i \mathbf{x}_i^\top \right) - \lambda_{\min}(\mathbf{\Sigma}) \left(1 - \sqrt{\frac{p}{|S_b|}} \right)^2 \to 0, \quad \text{a.s.},$$

where $\lambda_{\min}(\mathbf{\Sigma})(1-\sqrt{\frac{p}{|S_b|}})^2 \asymp \|\boldsymbol{\beta}^*\|^{-2}$. Finally, we get

$$q^\top M q \succeq \|\boldsymbol{\beta}^*\|^{-2}, \quad \text{a.s..}$$

Next, we consider the upper bound in the rank-1 case. Similar to (55), we get

$$\begin{aligned} q^\top M q &= a^\top \Big[\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b^\top c_1)^2,\dots,(b^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top\Big] a \\ &\le \|a\|^2 \lambda_{\max}\big(\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b^\top c_1)^2,\dots,(b^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top\big). \end{aligned} \tag{56}$$

Define $T_b = \{i : (b^\top c_i)^2 \ge ph\}$. We know that

$$\|b\|^2 \bar{\alpha} n V^2 \ge b^\top D D^\top b = \sum_{i=1}^n (b^\top c_i)^2 = \sum_{i\in T_b}(b^\top c_i)^2 + \sum_{i\notin T_b}(b^\top c_i)^2 \ge |T_b| ph.$$

So we get $|T_b| \le \frac{\bar{\alpha}V^2}{ph} n$. Note that

$$\begin{aligned} &\lambda_{\max}\big(\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b^\top c_1)^2,\dots,(b^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top\big) \\ &\le \lambda_{\max}\left(\tfrac{1}{n}\sum_{i\in T_b}\mathbf{x}_i\mathbf{x}_i^\top (b^\top c_i)^2\right) + \lambda_{\max}\left(\tfrac{1}{n}\sum_{i\notin T_b}\mathbf{x}_i\mathbf{x}_i^\top (b^\top c_i)^2\right). \end{aligned}$$

We consider these two parts. For the first part:

$$\lambda_{\max}\left(\tfrac{1}{n}\sum_{i\in T_b}\mathbf{x}_i\mathbf{x}_i^\top (b^\top c_i)^2\right) \preceq \frac{|T_b|}{n}\frac{p}{\|\boldsymbol{\beta}^*\|^2}E^2V^2 h\|b\|^2 \le \frac{\bar{\alpha}V^2}{ph}\frac{phE^2V^2}{\|\boldsymbol{\beta}\|^2} \asymp \|\boldsymbol{\beta}^*\|^{-2},$$

with high probability. For the second part:

$$\begin{aligned} \lambda_{\max}\left(\tfrac{1}{n}\sum_{i\notin T_b}\mathbf{x}_i\mathbf{x}_i^\top (b^\top c_i)^2\right) &\le \frac{n-|T_b|}{n}\lambda_{\max}\left(\frac{1}{n-|T_b|}\sum_{i\notin T_b}\mathbf{x}_i\mathbf{x}_i^\top (b^\top c_i)^2\right) \\ &\le \lambda_{\max}(\mathbf{\Sigma})\left(1+\sqrt{\frac{p}{n-|T_b|}}\right)^2 \asymp \|\boldsymbol{\beta}^*\|^{-2}, \end{aligned}$$

with high probability, where we used that $\|\mathbf{x}_i\mathbf{x}_i^\top\|_2 = \mathbf{x}_i^\top\mathbf{x}_i = O_{\text{HP}}(\frac{p}{\|\boldsymbol{\beta}^*\|^2})$ and $|T_b| \le \frac{\bar{\alpha}V^2}{ph}n$. Thus,

$$q^\top M q \preceq \|\boldsymbol{\beta}^*\|^{-2}, \quad \text{with high probability.}$$

This concludes the rank-1 case.

Next, we consider the case when $\operatorname{rank}(Q) = R > 1$. We can write $Q = \sum_{r=1}^R a_r b_r^\top$, where $\{a_r\}_{r\in[R]}$ and $\{b_r\}_{r\in[R]}$ are orthogonal sets respectively. WLOG, assume $\|b_r\| = 1$ for $r = 1,\dots,R$. Since $\|Q\|_F^2 = 1$, we have $\sum_{r=1}^R \|a_r\|^2 = 1$. Note that

$$q^\top M q \approx \sum_{r=1}^R a_r^\top \left[\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b_r^\top c_1)^2,\dots,(b_r^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top\right] a_r.$$

Using results from Case 1, we already know that for each r:

$$\begin{aligned} &\lambda_{\min}(\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b_r^\top c_1)^2,\dots,(b_r^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top) \\ \asymp &\lambda_{\max}\big(\tfrac{1}{n}(\mathbf{x}_1,\dots,\mathbf{x}_n)\operatorname{diag}\big((b_r^\top c_1)^2,\dots,(b_r^\top c_n)^2\big)(\mathbf{x}_1,\dots,\mathbf{x}_n)^\top\big) \asymp \|\boldsymbol{\beta}^*\|^{-2}. \end{aligned}$$

We then obtain

$$q^\top M q \asymp \sum_{r=1}^R \|a_r\|^2 \|\boldsymbol{\beta}^*\|^{-2} \asymp \|\boldsymbol{\beta}^*\|^{-2}.$$

Verification of Assumption B.4. For simplicity, we consider $\sigma(x) = x$. We have

$$\mathbf{g}_{i,*} = \begin{pmatrix} \mathbf{W} \\ \mathbf{v} \otimes \mathbb{I}_p \end{pmatrix} \mathbf{x}_i,$$

and

$$\left\| \begin{pmatrix} \mathbf{W} \\ \mathbf{v} \otimes \mathbb{I}_p \end{pmatrix} \right\| = \sqrt{\lambda_{\max}(\mathbf{W}^\top \mathbf{W}) + \|\mathbf{v}\|^2} \leq \sqrt{\frac{2\|\mathbf{W}\|_F^2}{p}} \asymp \frac{\|\boldsymbol{\beta}^*\|}{\sqrt{p}},$$

where for the first inequality we use the assumption that $\lambda_{\min}(\mathbf{W}^\top \mathbf{W}) \asymp \lambda_{\max}(\mathbf{W}^\top \mathbf{W})$.

Using $\mathbf{x}_i \sim$ zero-mean sub-Gaussian with covariance $\boldsymbol{\Sigma}$, where $\|\boldsymbol{\Sigma}\| \asymp \|\boldsymbol{\beta}^*\|^{-2}$, we know that $\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{p}} \mathbf{x}_i$ is zero-mean sub-Gaussian with covariance $\boldsymbol{\Sigma}'$ such that $\|\boldsymbol{\Sigma}'\| \asymp \frac{1}{p}$. Thus $\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{p}} \mathbf{x}_i \sim \text{nSG}(\sigma)$ for some $\sigma > 0$. Then

$$\begin{aligned} \mathbb{P}\left(\left\| \begin{pmatrix} \mathbf{W} \\ \mathbf{v} \otimes \mathbb{I}_p \end{pmatrix} \mathbf{x}_i - \mathbb{E} \begin{pmatrix} \mathbf{W} \\ \mathbf{v} \otimes \mathbb{I}_p \end{pmatrix} \mathbf{x}_i \right\| \geq t \right) &\leq \mathbb{P}\left(\left\| \begin{pmatrix} \mathbf{W} \\ \mathbf{v} \otimes \mathbb{I}_p \end{pmatrix} \right\| \|\mathbf{x}_i - \mathbb{E}\mathbf{x}_i\| \geq t \right) \\ &\asymp \mathbb{P}\left(\frac{\|\boldsymbol{\beta}^*\|}{\sqrt{p}} \|\mathbf{x}_i - \mathbb{E}\mathbf{x}_i\| \geq t \right) \\ &\leq 2\exp\left(-\frac{t^2}{2\sigma^2} \right). \end{aligned}$$

Therefore,

$$\mathbf{g}_{i,*} \overset{i.i.d.}{\sim} \text{nSG}(\sigma).$$

C Simulations and empirical studies

C.1 Correlation-based experiments

C.1.1 Vanilla linear models

We first consider the linear case, where $f(\mathbf{x}, \boldsymbol{\beta}^*) = \mathbf{x}^\top \boldsymbol{\beta}^*$, and focus on binary classification and Poisson regression as representative examples. In this setting, the linearization error is exactly zero; therefore, only the ALO step and the projection step need to be evaluated.

We generate the rows $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ of the design matrix $\boldsymbol{X}$ independently from a mean-zero multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$. Here, $\boldsymbol{\Sigma}$ is a Toeplitz matrix satisfying

$$\text{corr}(X_{ij}, X_{ij'}) = 0.1^{|j-j'|}.$$

We rescale $\boldsymbol{\beta}^*$ and $\boldsymbol{\Sigma}$ such that $\|\boldsymbol{\beta}^*\|^2 = p$ and $\boldsymbol{\beta}^{*\top} \boldsymbol{\Sigma} \boldsymbol{\beta}^* = 1$, ensuring that our theoretical assumptions are satisfied.

Binary classification. For binary classification, we randomly select 100 training points to remove. To reduce variability, we independently sample 10 new test points $\mathbf{z}_{\text{new}}$ from the same distribution. For each test point, we compute its influence with respect to each removed training point. This procedure yields 1,000 realizations of $\mathcal{I}^{\text{True}}$, $\mathcal{I}^{\text{ALO}}$, and $\mathcal{I}^{\text{TRAK}}(k)$, where k denotes the projection dimension.

ALO Step. Figure 2 validates Theorem 2.13. As shown, $\mathcal{I}^{\text{True}}$ and $\mathcal{I}^{\text{ALO}}$ are nearly identical, with fitted slopes close to one. This confirms that the ALO step introduces only negligible magnitude error, consistent with our theoretical guarantee.

Projection Step. Figure 3 compares $\mathcal{I}^{\text{True}}$ and $\mathcal{I}^{\text{TRAK}}$ for $k = 0.75d$, $0.50d$, and $0.25d$, where $d = p = 100$. The slopes decrease as k decreases, validating Theorem 2.15. This shows that the projection step systematically reduces the magnitude of the TRAK influence. Nevertheless, for sufficiently large k, the projected influence still preserves the qualitative separation between influential and non-influential data points.

Poisson regression. We next consider Poisson regression with a softplus mean function. As shown in Figures 2 and 4, we observe trends similar to those in binary classification, further validating Theorems 2.13 and 2.15.

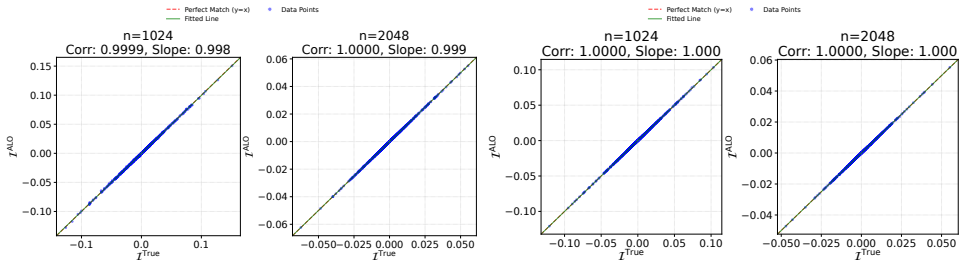

Figure 2: **Left two panels:** Binary logistic regression with $p = 100$, $n = 1024$, and $n = 2048$. The x-axis shows $\mathcal{I}^{\mathrm{True}}$, and the y-axis shows $\mathcal{I}^{\mathrm{ALO}}$. **Right two panels:** Poisson regression under the same p and n settings.

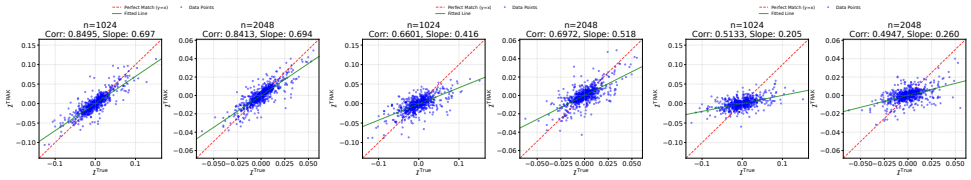

Figure 3: Binary logistic regression with projection ($p = 100$, $n = 1024$, and $n = 2048$). The x-axis shows $\mathcal{I}^{\mathrm{True}}$, and the y-axis shows $\mathcal{I}^{\mathrm{TRAK}}$. From left to right, the projection dimensions are $k = 75$, 50, and 25.

C.1.2 Nonlinear case

Simulation setting. We next consider nonlinear multi-class classification. The feature vectors $\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top$ are sampled independently from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a Toeplitz covariance matrix satisfying

$$\mathrm{corr}(X_{ij}, X_{ij'}) = 0.1^{|j-j'|}.$$

We consider multi-class classification with $K = 3$ in Figure 1, results for $K = 5$ are shown in Appendix Figure 5. The simulation results are qualitatively consistent across both settings.

We rescale $\boldsymbol{\beta}^*$ and $\mathbf{\Sigma}$ such that $\|\boldsymbol{\beta}^*\|^2 = p$ and $\|\mathbf{\Sigma}\| = \|\boldsymbol{\beta}^*\|^{-2}$. For each trial t, we generate $(\boldsymbol{X}^{(t)}, y^{(t)})$ with

$$y^{(t)} \sim \mathcal{M}(\mathrm{softmax}(\boldsymbol{X}^{(t)}\mathbf{W}^{*(t)\top})),$$

where $\mathcal{M}$ denotes the multinomial distribution and $\mathbf{W}^{*(t)} \in \mathbb{R}^{K\times p}$. To ensure identifiability, we set $\mathbf{W}_K^{*(t)} = \mathbf{0}_p$. We write the free model parameters as

$$\boldsymbol{\beta}^* \triangleq \mathrm{vec}(\mathbf{W}^{*(t)}_{[1:(K-1),:]}) \in \mathbb{R}^d, \quad d = (K-1)\times p.$$

For the uncorrelated case, we randomly remove 100 training points and independently sample 10 new test points $\mathbf{z}_{\mathrm{new}}$, yielding 1,000 realizations of the influence quantities. For the correlated case, we set $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$, yielding 100 realizations. These two settings allow us to test whether the approximations preserve the separation between dependent and independent data points, and the rankings of $\mathcal{I}^{\mathrm{True}}$.

C.2 Magnitude-based simulations

In this section, we examine how the magnitudes of the influence quantities and approximation errors vary with the sample size n, the feature dimension p, and the projection dimension k. We focus on the nonlinear setting of 3-class classification as a representative example. The data-generating mechanism and other experimental details are the same as in Appendix C.1. We consider both the correlated case, where $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$, and the uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\mathrm{new}}$ are independent.

Figure 6 confirms the scaling behavior predicted by our theory. We first consider the magnitude of the true influence. In both the correlated and uncorrelated cases, $|\mathcal{I}^{\mathrm{True}}|$ decreases at rate $1/n$ as the sample size increases. The dependence on p, however, differs across the two regimes. In the

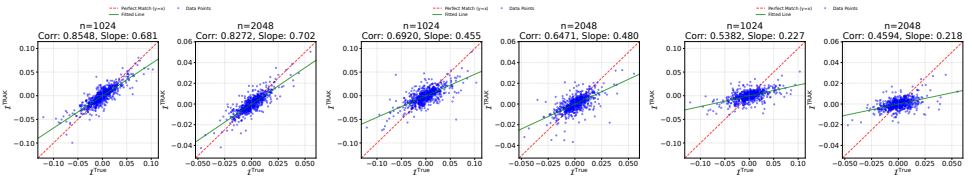

Figure 4: Poisson regression with projection. The settings are identical to those in Figure 3.

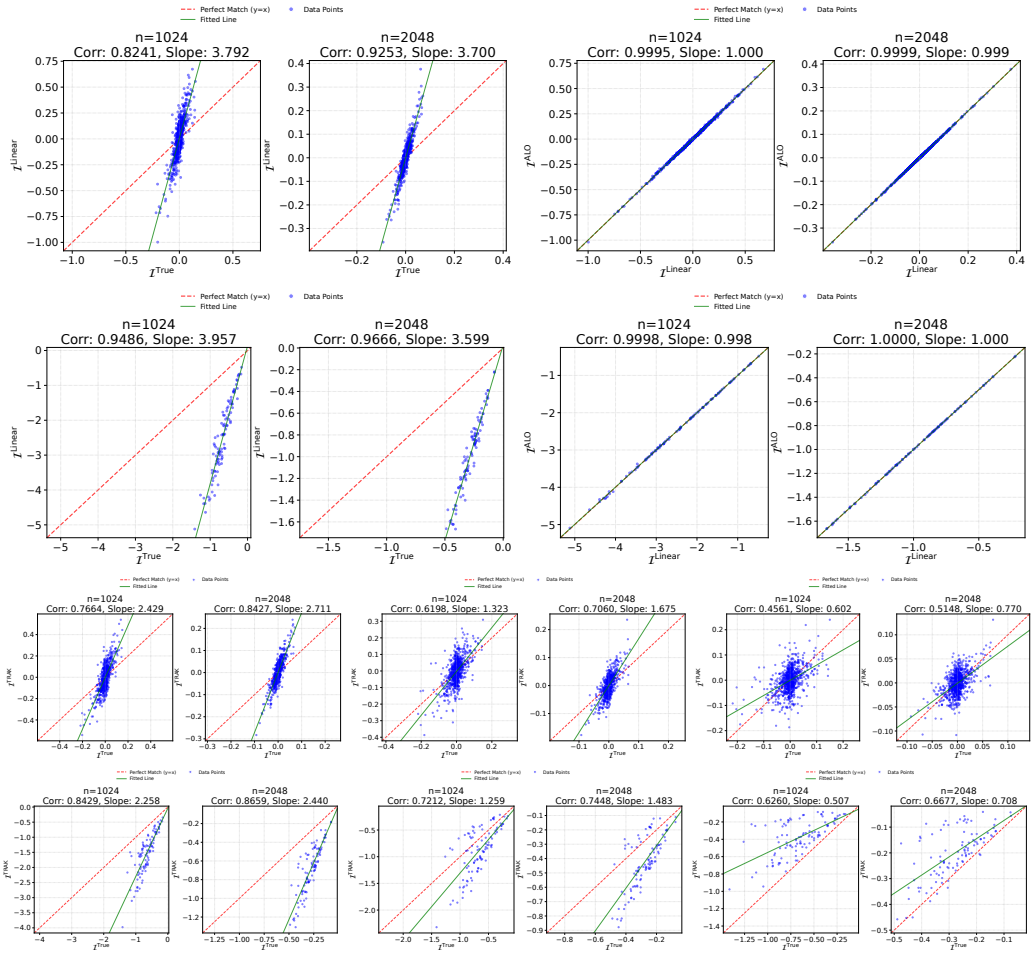

Figure 5: Experimental results for 5-class classification with $p = 100$ and $d = 400$. **First and third rows:** Results for the uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\text{new}}$ are independent. **Second and fourth rows:** Results for the correlated case, where $\mathbf{z}_i = \mathbf{z}_{\text{new}}$. **First and second rows:** $\mathcal{I}^{\text{Linear}}$ versus $\mathcal{I}^{\text{True}}$ and $\mathcal{I}^{\text{ALO}}$ versus $\mathcal{I}^{\text{Linear}}$. **Third and fourth rows:** $\mathcal{I}^{\text{TRAK}}(k)$ versus $\mathcal{I}^{\text{True}}$ with $k = 75\%d, 50\%d, 25\%d$, corresponding to $k = 300, 200, 100$.

correlated case, $|\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_i)|$ grows approximately as p/n, consistent with Proposition 2.4. In the uncorrelated case, $|\mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_{\text{new}})|$ grows approximately as $\sqrt{p}/n$, consistent with Proposition 2.5. This difference is precisely the magnitude separation between dependent and independent data points predicted by our theory.

We next examine the linearization error. In the correlated case, the empirical order of

$$|\mathcal{I}^{\text{Linear}}(\mathbf{z}_i, \mathbf{z}_i) - \mathcal{I}^{\text{True}}(\mathbf{z}_i, \mathbf{z}_i)|$$

is approximately p/n, matching the order of the true influence itself and validating Theorem 2.6. In the uncorrelated case, the corresponding error scales approximately as $\sqrt{p}/n$, consistent with

Theorem 2.9. These results confirm that the linearization step can introduce a non-negligible absolute error, while its order still reflects the separation between the correlated and uncorrelated regimes.

The ALO step behaves differently. Across both regimes, the empirical magnitude of

$$|\mathcal{I}^{\mathrm{ALO}} - \mathcal{I}^{\mathrm{Linear}}|$$

is substantially smaller than the corresponding influence magnitude. In both the correlated and uncorrelated cases, the observed order is approximately p/n^2, which is smaller than our theoretical upper bounds: $p/n^{3/2}$ for the correlated case in Theorem 2.12, and $\sqrt{p}/n^{3/2}$ for the uncorrelated case in Theorem 2.13. Although these bounds may not be tight, the conclusion is unchanged: the ALO approximation is benign in magnitude and does not destroy the separation between dependent and independent data points.

Finally, we examine the projection step. In the uncorrelated case, the magnitude of $\mathcal{I}^{\mathrm{TRAK}}(k)$ scales as $\sqrt{p}/n$ and grows approximately as $\sqrt{k/d}$ as the relative projection dimension increases. In contrast, in the correlated case, $\mathcal{I}^{\mathrm{TRAK}}(k)$ scales as p/n and grows approximately linearly in k/d. These observations are consistent with Theorem 2.15. They also illustrate the main limitation of the projection step: when k/d is too small, the projection can substantially reduce the separation between the correlated and uncorrelated influence magnitudes.

C.3 Empirical validation on CIFAR-10

In this section, we conduct experiments on CIFAR-10 to assess whether our conclusions remain robust on real-world data that may not satisfy the data-generation assumptions used in our theory. The goal is not to maximize predictive accuracy, but to examine whether the qualitative behavior observed in the simulations persists in a more realistic multi-class classification problem.

Experimental setup. We use CIFAR-10, which contains 10 classes and $50{,}000$ training images. Following the multi-class formulation in Appendix A.1, we train a multinomial logistic regression model using the standard $K - 1$ parameterization. For computational feasibility, each image is first flattened after standard normalization, and we then apply a fixed Gaussian random projection to obtain a $p = 200$-dimensional feature vector. The projected features are standardized using the empirical mean and standard deviation computed from the selected training set.

Since $K = 10$ and $p = 200$, the parameter dimension is

$$d = p(K - 1) = 200 \times 9 = 1800.$$

The trained model achieves a training accuracy of 39.78% and a test accuracy of 38.68%. This accuracy is lower than that of modern deep learning baselines, but it is typical for a simple linear classifier built on low-dimensional random-projection features.

We sample 100 training points and 100 test points, giving $10{,}000$ uncorrelated pairs. We also evaluate the corresponding correlated case, where $\mathbf{z}_i = \mathbf{z}_{\mathrm{new}}$. For each pair, we compute $\mathcal{I}^{\mathrm{True}}$, $\mathcal{I}^{\mathrm{Linear}}$, $\mathcal{I}^{\mathrm{ALO}}$, and $\mathcal{I}^{\mathrm{TRAK}}(k)$ for several projection dimensions.

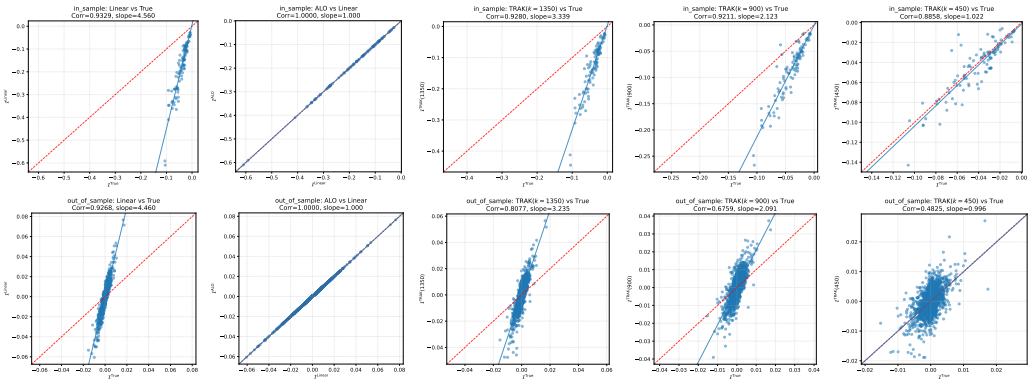

Figure 7: CIFAR-10 experimental results with $p = 200$ and $d = 1800$. **Top row:** Correlated case, where $\mathbf{z}_i = \mathbf{z}_{\text{new}}$. **Bottom row:** Uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\text{new}}$ are independent. **First column:** $\mathcal{I}^{\text{Linear}}$ versus $\mathcal{I}^{\text{True}}$. **Second column:** $\mathcal{I}^{\text{ALO}}$ versus $\mathcal{I}^{\text{Linear}}$. **Right three columns:** $\mathcal{I}^{\text{TRAK}}(k)$ versus $\mathcal{I}^{\text{True}}$ with $k = 75\%d, 50\%d, 25\%d$, corresponding to $k = 1350, 900, 450$.

Approximation behavior. Figure 7 visualizes how the influence estimates behave after the three TRAK approximation steps in both the correlated and uncorrelated cases. The conclusions are qualitatively similar to those in Figure 1. The linearization step introduces visible magnitude distortion, but the approximate scores remain aligned with the true influence scores. The ALO step is nearly indistinguishable from the linearized influence, suggesting that ALO introduces little additional error. The projection step reduces the magnitude of the scores, and this reduction becomes more pronounced as the projection dimension decreases.

These observations are consistent with the main simulation results. In particular, the projection step is the main source of degradation when k is small: smaller projection dimensions compress the influence scores more aggressively and can weaken the separation between the correlated and uncorrelated regimes.

Rank alignment. We next examine whether the approximate influence scores preserve the rankings induced by $\mathcal{I}^{\text{True}}$. For each test point, we compare the Top-m and Bottom-m training examples selected by each approximation with those selected by $\mathcal{I}^{\text{True}}$. We evaluate the Top-$m$ most positive influence scores, and the Bottom-m most negative influence scores.

We report two ranking metrics. The first is the **Exact Match Count**, defined as the number of test points, out of 100, for which the retrieved ranked list exactly matches the ranked list obtained from $\mathcal{I}^{\text{True}}$, including the ordering. The second is the **Overlap Ratio**, defined as the fraction of overlap between the retrieved set and the exact set, averaged across test points.

Table 4: Rank alignment between exact and approximate influence rankings on CIFAR-10 in the uncorrelated case. Metrics are reported for both Top-m positive influence scores, and Bottom-m negative influence scores. Exact Matches count the number of test points, out of 100, for which the retrieved ranked list exactly matches the list obtained from $\mathcal{I}^{\text{True}}$. Overlap Ratio is the average set overlap with the corresponding exact Top-m or Bottom-m set.

Method	**Metric**	**Top-m Positive**						**Bottom-m Negative**					
		1	**3**	**5**	**10**	**20**	**50**	**1**	**3**	**5**	**10**	**20**	**50**
$\mathcal{I}^{\text{Linear}}$	Exact Matches	74	24	6	0	0	0	73	20	1	0	0	0
	Overlap Ratio	0.740	0.767	0.790	0.772	0.760	0.805	0.730	0.777	0.750	0.762	0.781	0.805
$\mathcal{I}^{\text{ALO}}$	Exact Matches	75	24	6	0	0	0	73	20	1	0	0	0
	Overlap Ratio	0.750	0.767	0.788	0.772	0.759	0.805	0.730	0.777	0.750	0.763	0.781	0.805
$\mathcal{I}^{\text{TRAK}}(k = 1350)$	Exact Matches	68	13	0	0	0	0	66	15	1	0	0	0
	Overlap Ratio	0.680	0.627	0.648	0.576	0.575	0.674	0.660	0.670	0.620	0.607	0.581	0.674
$\mathcal{I}^{\text{TRAK}}(k = 900)$	Exact Matches	44	8	0	0	0	0	54	2	0	0	0	0
	Overlap Ratio	0.440	0.503	0.508	0.473	0.474	0.628	0.540	0.537	0.476	0.474	0.480	0.628
$\mathcal{I}^{\text{TRAK}}(k = 450)$	Exact Matches	26	2	0	0	0	0	30	0	0	0	0	0
	Overlap Ratio	0.260	0.323	0.352	0.343	0.390	0.586	0.300	0.383	0.354	0.376	0.409	0.586

Table 4 shows that the linearization and ALO approximations preserve substantial ranking information. Although exact list matches become rare for larger values of k, the overlap ratios for $\mathcal{I}^{\mathrm{Linear}}$ and $\mathcal{I}^{\mathrm{ALO}}$ remain around 0.75–0.80 for both Top-m and Bottom-m. This indicates that these two approximations often identify similar influential training points, even if the exact order differs.

In contrast, the TRAK projection step leads to a clear degradation as the projection dimension decreases. For example, for Top-1 Positive, the exact match count decreases from 68% at $k = 1350$ to 44% at $k = 900$ and 26% at $k = 450$. The overlap ratios also decrease substantially across most Top-m and Bottom-m settings. This supports the same conclusion as the synthetic simulations: random projection can preserve useful ranking information when k is sufficiently large, but aggressive projection can weaken the separation structure and harm rank agreement.

C.4 Results of more general correlated cases

In this section, we show some simulations as defined in Section B.4. We still consider 3-class classification with $p = 100$ and $d = 200$, using the same data-generating mechanism as in Figure 1 to generate our training data. But for the new test data, we generate them using (52). We randomly select 100 data points to remove and generate 10 new data points for each removed data point, using α_n uniformly distributed from -1 to 1. We then compare the influence between $\mathbf{z}_{\mathrm{new}}$ and $\mathbf{z}_i$ with the influence of $\mathbf{z}_i$ on $\mathbf{z}_i$.

We plot their ratio using different influence definitions. As we can see from Figure 8, the x-axis shows the different α values of each new data point, and the y-axis shows the influence ratio. The result is quite intuitive and is indicated in Section B.4: although the exact relationship cannot be preserved due to the randomness of the independent part of $\mathbf{x}_i$, the ratio is still of order $\Theta(1)$, meaning that the prediction in (54) holds.

Also, some interesting and consistent results can be seen from Figure 8. As we can see, the error introduced in the linearization and ALO steps is mild. However, as the projection dimension k becomes smaller, for example from 150 to 75, the width of the band becomes obviously wider, which again shows that too aggressive projection may deteriorate the performance of TRAK, consistent with our previous message.

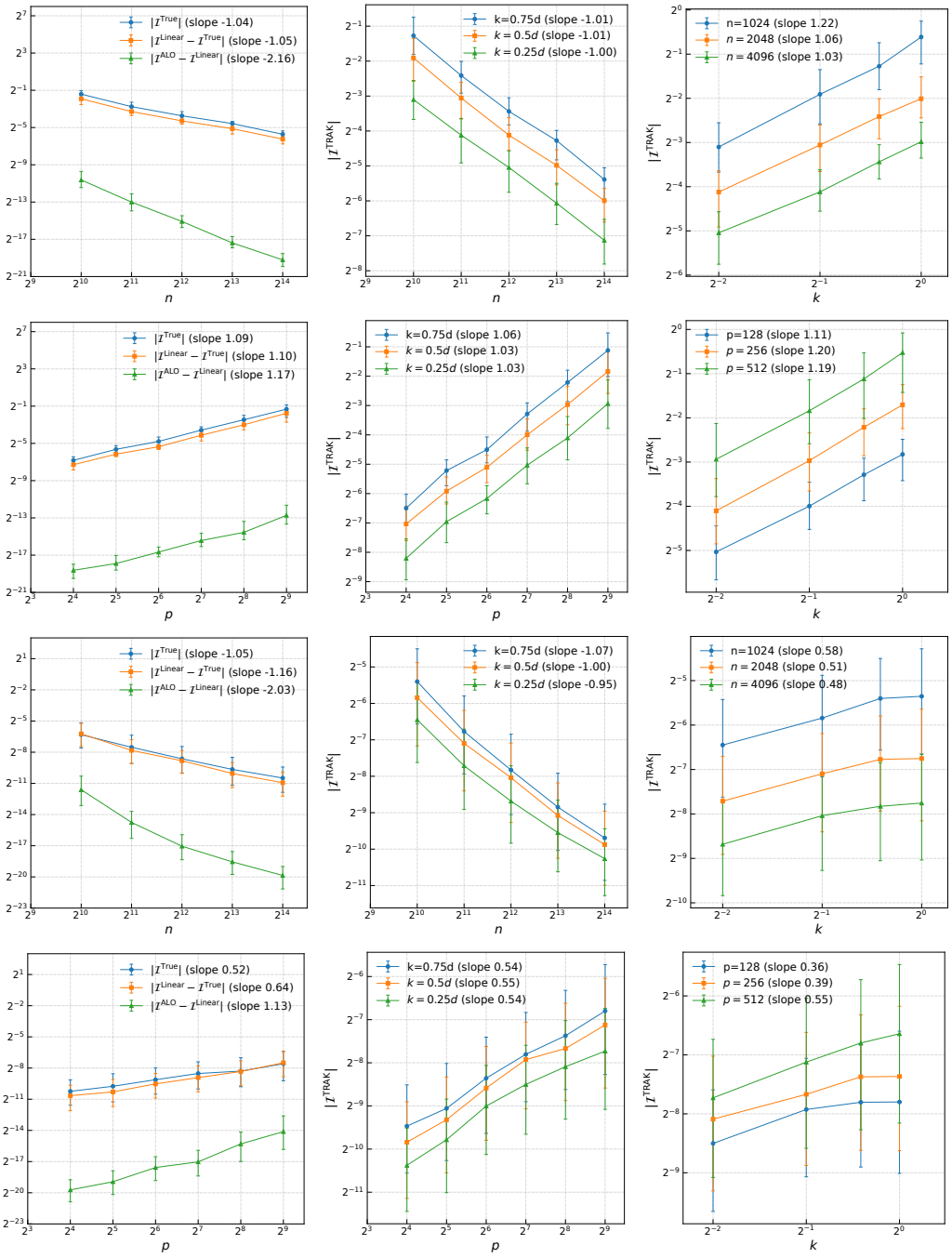

Figure 6: 3-class classification: magnitude-order checks. **First and second rows:** Correlated case, where $\mathbf{z}_i = \mathbf{z}_{\text{new}}$. **Third and fourth rows:** Uncorrelated case, where $\mathbf{z}_i$ and $\mathbf{z}_{\text{new}}$ are independent. **First and third rows:** p is fixed at 100 and n varies. **Second and fourth rows:** n is fixed at 5,000 and p varies. In each row, the first panel shows the magnitudes of $\mathcal{I}^{\text{True}}$, the linearization error, and the ALO error; the second panel shows $\mathcal{I}^{\text{TRAK}}(k)$ for different projection dimensions; and the third panel shows how $\mathcal{I}^{\text{TRAK}}(k)$ varies with k/d.

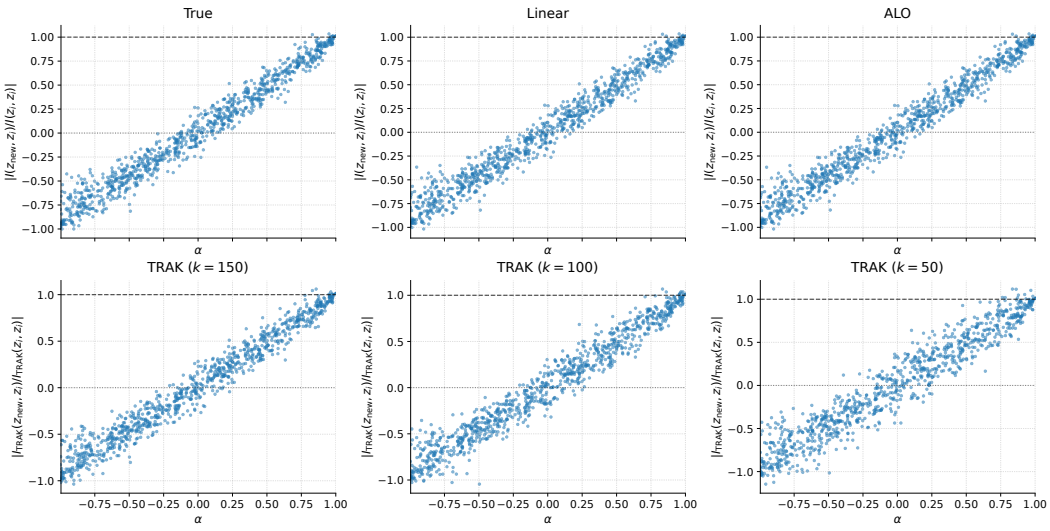

Figure 8: The setting is the same as in Figure 1. We show the ratio of the influence between $\mathbf{z}_{\text{new}}$ and $\mathbf{z}_i$ to the influence of $\mathbf{z}_i$ on $\mathbf{z}_i$. The first row shows the ratios for $\mathcal{I}^{\text{True}}$, $\mathcal{I}^{\text{Linear}}$, and $\mathcal{I}^{\text{ALO}}$; the second row shows the ratios for $\mathcal{I}^{\text{TRAK}}(k)$, with $k = 75\%, 50\%, 25\%$ of d, corresponding to $k = 150, 100, 50$.

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: The abstract and introduction state the paper's main theoretical contribution: characterizing the approximation errors in TRAK and showing that, under the stated assumptions, the approximations can preserve meaningful separation between influential and weakly influential data points. The empirical claims are limited to simulations and CIFAR-10 validation reported in the main text and appendix.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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 add a paragraph in conclusion explaining the limitations of our work.

   Guidelines:

   - The answer [N/A] 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: We state the full set of assumptions in the appendix and verify that these assumptions are satisfied by multiple model classes, including GLMs and MLPs.

Guidelines:

- The answer [N/A] 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: We clearly state the experimental settings so that the experiments can be reproduced.

Guidelines:

- The answer [N/A] 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: [No].

   Justification: We do not provide code as part of the anonymous submission. The main contribution of the paper is theoretical, and the simulation and empirical studies are described in detail in the main text and appendix, including the data-generating mechanisms, model classes, sample sizes, projection dimensions, and evaluation metrics. The empirical study uses CIFAR-10, which is publicly available. We plan to release code after publication.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the experimental settings in the main text and appendix, including the data-generating mechanisms, model classes, sample sizes, feature dimensions, projection dimensions, and the in-sample and out-of-sample evaluation regimes. For the CIFAR-10 experiments, we also describe the preprocessing steps, random-projection feature construction, model parameterization, and evaluation metrics.

   Guidelines:

   - The answer [N/A] 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: [Yes].

   Justification: For the magnitude-based results, such as Appendix Figure 6, we report error bars and define how they are computed. For other results, such as the ranking table and scatter plots, we do not report error bars or statistical significance tests. The main claims

of the paper are theoretical, and the experiments are designed to illustrate and verify the predicted approximation behavior, scaling patterns, and separation effects rather than to compare competing methods through repeated randomized trials.

Guidelines:

- The answer [N/A] 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: [No].

   Justification: We do not provide detailed compute-resource information. The experiments are moderate-scale simulations and multinomial logistic-regression experiments, and each reported experiment can be run within about one hour on standard computing resources.

   Guidelines:

   - The answer [N/A] 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: The research is theoretical and empirical work on data attribution and influence approximation. It uses synthetic data and the public CIFAR-10 dataset only.

   Guidelines:

   - The answer [N/A] 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: [Yes].

    Justification: The paper studies theoretical properties of data attribution methods. Potential positive impacts include improving transparency, debugging, and data valuation. Potential negative impacts include possible misuse of attribution scores for overconfident data ownership or compensation decisions when the assumptions of the theory are violated. We therefore emphasize that our results concern separation under stated assumptions, not exact ranking or exact monetary valuation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release high-risk models, scraped datasets, generative models, or datasets with sensitive content. The empirical study uses CIFAR-10, a standard public benchmark dataset, and the work poses no such misuse risks requiring safeguards.

    Guidelines:

    - The answer [N/A] 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: The paper uses CIFAR-10 as an existing public benchmark dataset and properly cites the original dataset source. We do not redistribute CIFAR-10 or introduce a derived dataset as a new asset.

    Guidelines:

    - The answer [N/A] 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: [N/A].

    Justification: The paper does not release any new datasets, model checkpoints, or software packages as new assets.

    Guidelines:

    - The answer [N/A] 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: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects.

    Guidelines:

    - The answer [N/A] 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: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or research with human subjects, so IRB approval or equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] 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 rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The core method development, theoretical analysis, simulations, and empirical evaluation do not involve LLMs as an important, original, or non-standard component. Any use of LLMs for writing, editing, or formatting does not affect the methodology, scientific rigor, or originality of the research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.