ON THE ACCURACY OF NEWTON STEP AND INFLUENCE FUNCTION DATA ATTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Data attribution aims to explain model predictions by estimating how they would change if certain training points were removed, supporting a wide range of applications, from interpretability and credit assignment to unlearning and privacy.

Even in the relatively simple case of linear models with normally distributed features, existing analyses of leading data attribution methods such as Influence Functions (IF) and single Newton Step (NS) remain limited. Current bounds require a strong dependence on the global strong convexity parameter λ, which is often very small in practice and scales like $O(d/n)$ for well-behaved regressions with dimension d and n samples. These bounds also scale poorly with the number of removed samples k and with d. Making the dependence on λ explicit reveals that existing results are very loose:

$$\text{Existing Bounds} = \Omega\left(\frac{k^2 d}{\lambda^3 n^2}\right) = \Omega\left(\frac{k^2 n}{d^2}\right).$$

We introduce new analytic tools for bounding the errors of NS, yielding substantially tighter results that do not depend on the global strong convexity λ. Moreover, we show that for logistic regressions with normally distributed features, our bounds also scale much more favorably with k and d:

$$\text{New Bound} = \tilde{O}\left(\frac{k}{n^2}\right).$$

We show that our bounds are tight up to poly-logarithmic factors, that they also yield similarly tight bounds on the accuracy of IF and provide the first theoretical explanation for the empirical observation that NS is more accurate than IF.

1 INTRODUCTION

Let $L = \ell_1 + \cdots + \ell_n : \Omega_\theta \to \mathbb{R}$ be an Empirical Risk Minimization (ERM) problem, where the contributions to the loss may represent the training samples of some supervised learning problem $\ell_i = \ell(f(\boldsymbol{\theta}, x_i), y_i)$. Define

$$\hat{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}\in\Omega_\theta} \left\{ \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}) \right\} \quad \text{and} \quad \hat{\boldsymbol{\theta}}_T \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta}\in\Omega_\theta} \left\{ \sum_{i\notin T} \ell_i(\boldsymbol{\theta}) \right\}.$$

Data attribution seeks to explain the dependence of $\hat{\boldsymbol{\theta}}$ on the training data by predicting its change when removing a subset $T \subset [n]$ of the samples ℓ_i Ilyas et al. (2022); Park et al. (2023). We can answer such a query by fully retraining the model $\hat{\boldsymbol{\theta}}$ with only a subset of the samples, but this is often computationally intensive and lacks a closed form solution which might be useful for downstream tasks and explainability Madry et al. (2024).

Therefore, data attribution often utilizes approximations known as *data models*, the most widespread of which are influence functions (IF) and single Newton steps (NS). This raises the natural question:

> *When do IF and NS give good approximations of the data-removal effect? How does the approximation error scale with the total number of samples n, the problem dimension d and number of samples removed $k = |T|$?*

Informally, IF data models use a first-order approximation to the sample-removal effect, and NS data models approximate by taking single Newton step on $\hat{\boldsymbol{\theta}}$ (see Section 1.1).

We analyze the accuracy of NS and IF in the analytically tractable setting of convex empirical risk minimization. Beyond the direct application for convex learning problems (such as last-layer fine-tuning through logistic regression), understanding the convex setting is curcial since leading data attributions for non-convex settings often rely on heuristic reductions to data attribution for convex problems (e.g., TRAK Park et al. (2023) uses an NTK assumption to approximate the attribution of a deep neural network). This is one motivation for our focus on uncovering the scaling laws for the convex setting – we discuss several further motivations later.

Existing analyses of IF and NS data attributions Rad & Maleki (2018); Giordano et al. (2019); Koh et al. (2019); Wilson et al. (2020), rely on a *global strong convexity* assumption that often does hold in practice[1]. Moreover, the resulting bounds scale poorly with the problem dimension d and with the number of samples removed k, resulting in very loose bounds Rubinstein & Hopkins (2025b).

For instance, it has been observed empirically that NS often provides a much better approximation of removal effects than IF Koh et al. (2019); Hu et al. (2024); Huang et al. (2024); Rubinstein & Hopkins (2025b), but to the best of our knowledge, no previous quantitative analysis comes close to explaining this phenomenon.

In this work, we provide the first theoretical analysis whose guarantees are tight enough to explain *why* and *when* the NS data attribution outperforms IF.

We tackle this question in two settings, with two main results. Our first main result concerns empirical risk minimization problems with a convex loss L. We place assumptions on the local strong convexity and local (higher-order) Lipschitzness of L in a small neighborhood of $\hat{\boldsymbol{\theta}}$, and show that under those assumptions the NS data model is quantitatively accurate. We offer a quantitative statement and comparison to prior work later.

Theorem 1.1 (NS Accuracy for Convex Losses (Informal, see Theorem 1.5))**.** *If $L_T = \sum_{i \notin T} \ell_i$ is strongly convex in a neighborhood of the first Newton step starting from $\hat{\boldsymbol{\theta}}$, and the Hessian of L_T is (mildly) Lipschitz along the first Newton step, then the output of the first Newton step is close $\hat{\boldsymbol{\theta}}_T$.*

Assumptions like local strong convexity or local Lipschitzness allow great generality across different loss functions and datasets, but can be difficult to interpret quantitatively, obscuring the significant differences observed in practice among various data models. Our second main result addresses a much more concrete setting, where we can hope to prove sharp quantitative bounds on the accuracy of data models, using Theorem 1.1. We focus on logistic regression with Gaussian features as a "model organism", and show that:

Theorem 1.2 (NS is more accurate than IF)**.** *Suppose L is the empirical logistic loss corresponding to n independent samples $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$ from a distribution (x, y) with $x \sim N(0, I)$, with population loss minimizer $\boldsymbol{\theta}^*$ having $\|\boldsymbol{\theta}^*\| = \Theta(1)$. For $T \subseteq [n]$, let $\hat{\boldsymbol{\theta}}_T^{\text{IF}}$ be the* IF *estimate of $\hat{\boldsymbol{\theta}}_T$ and let $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$ be the* NS *estimate of $\hat{\boldsymbol{\theta}}_T$. Let $k \leq n/\text{polylog}(n)$ and $n \geq d^3 > \text{polylog}(n)$. Using Theorem 1.1, we provide almost matching worst-case upper-bounds and average-case lower-bounds on the error of IF and NS, with high-probability over the training data*

$$\Omega\left(\frac{k}{n^2}\right) \leq \mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}}\right\|_2\right] \leq \max_{T \in \binom{[n]}{k}} \left\{\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}}\right\|_2\right\} = \widetilde{O}\left(\frac{k}{n^2}\right),$$

and

$$\widetilde{\Omega}\left(\frac{(k+d)\sqrt{kd}}{n^2}\right) \leq \mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{IF}}\right\|_2\right] \leq \max_{T \in \binom{[n]}{k}} \left\{\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{IF}}\right\|_2\right\} = \widetilde{O}\left(\frac{(k+d)\sqrt{kd}}{n^2}\right).$$

In particular, we conclude that the accuracy of NS is significantly better than the accuracy of IF.

[1] Some analyses require only that the loss is strongly convex in some subset of the optimization domain. (Rad & Maleki, 2018, Assumption 7) only require that the loss is strongly convex along the path between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_T$ and (Wilson et al., 2020, Assumption 1) only require that the loss is strongly convex in a ball around $\hat{\boldsymbol{\theta}}_T$. However, these papers do not give an argument for why the loss should be more convex in this area than anywhere else in the optimization domain. Therefore, without prior knowledge on $\hat{\boldsymbol{\theta}}_T$, it is not clear how one could get a concrete guarantee beyond the one in Theorem 1.4).

We work in the regime $n \geq d^3$ to simplify the scaling laws and their proofs (which are already quite technical). We expect that the upper bounds of Theorem 1.1 are similarly close to tight even up to $k, d = O(n)$.

1.1 Data Attribution Models

Influence Functions (IF) also known as infinitesimal jackknife (IJ) Jaeckel (1972) utilize a first order approximation to the effect of down weighting a sample on the model parameters. Defining $\boldsymbol{\theta}_{(\cdot)} : \mathbb{R}^n \to \mathbb{R}^d$ as the function that takes a set of weights $\mathbf{w} \in \mathbb{R}^n$, and optimizes the model on the weighted samples $\boldsymbol{\theta}_{\mathbf{w}} = \operatorname{argmin}_{\boldsymbol{\theta} \in \Omega} \{\sum_{i=1}^n w_i \ell_i(\boldsymbol{\theta})\}$, IF employs a first order Taylor series in $\mathbf{w}$. The resulting estimated change in the model parameters is given by the following formula (see e.g., Rousseeuw et al. (1986) for a derivation),

$$\boldsymbol{\theta}^{\text{IF}}_{\mathbf{w}} := \boldsymbol{\theta}_{\mathbf{w}=\mathbf{1}} + \frac{\partial \boldsymbol{\theta}_{\mathbf{w}}}{\partial \mathbf{w}} (\mathbf{w} - \mathbf{1}) = \hat{\boldsymbol{\theta}} + \mathbf{H}^{-1} \sum_{i=1}^n (1 - w_i)\, \mathbf{g}_i \,,$$

where $\mathbf{H}$ is the Hessian of the loss $L(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta})$ evaluated at $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_{\mathbf{w}=\mathbf{1}}$, and $\mathbf{g}_i$ is the gradient of ℓ_i at $\hat{\boldsymbol{\theta}}$.

Influence functions are ubiquitous due to their simplicity and applicability for downstream tasks Giordano et al. (2019); Broderick et al. (2020) and were made more applicable to modern machine learning with fast algorithms for estimating them without explicitly inverting the Hessian Koh & Liang (2017); Park et al. (2023).

However, while IF data attribution appears good at capturing the qualitative behavior of a model's change when dropping a data point, it often misses the scale of the change and significantly underestimates removal effects, especially in the high-dimensional regime Koh et al. (2019).

Single Newton Step (NS) data attribution, dating to Pregibon (1981) and explored recently in Koh & Liang (2017); Koh et al. (2019); Wilson et al. (2020); Huang et al. (2024) offers somewhat higher accuracy than the IF approach. Here we approximate $\hat{\boldsymbol{\theta}}_T$ using a single step of the Newton algorithm initialized at $\hat{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta}^{\text{NS}}_{\mathbf{w}} = \hat{\boldsymbol{\theta}} + \mathbf{H}^{-1}_{\mathbf{w}} \sum_{i=1}^n (1 - w_i)\, \mathbf{g}_i \,.$$

NS appears almost equivalent to the influence function, except for the key difference that it takes into account the change to the Hessian due to dropping the samples. Moreover, when the loss is a quadratic function of the model (e.g., in ordinary least squares regression), $\boldsymbol{\theta}^{\text{NS}}_{\mathbf{w}} = \hat{\boldsymbol{\theta}}_{\mathbf{w}}$, since Newton iteration converges in a single step.

While NS is typically much slower to compute than IF (because of the matrix inversion $\mathbf{H}^{-1}_{\mathbf{w}}$), spending the extra compute time appears to have a major accuracy benefit – Koh et al. Koh et al. (2019) showed empirically that NS data attribution is typically much more accurate than IF on real-world datasets. However, our theoretical understanding of this phenomenon is very limited.

1.2 Related Work and Previous State-of-the-Art

History and Applications IF and NS originated in early work on robust statistics Hampel (1974); Jaeckel (1972); Pregibon (1981). Both are now used in a wide variety of downstream applications. IFs are used for data attribution in neural networks Park et al. (2023); Basu et al. (2021); Bae et al. (2022); Engstrom et al. (2025), though when they are effective here remains a topic of current investigation. Beyond deep learning, data attribution via IF and NS is a core technique in *machine unlearning* for convex risk minimization problems Sekhari et al. (2021b); Neel et al. (2021); Suriyakumar & Wilson (2022), for robustness auditing/finding highly influential sets of samples Broderick et al. (2020); Rubinstein & Hopkins (2025a); Huang et al. (2024); Hu et al. (2024), and approximate cross-validation Wilson et al. (2020), and even evaluation of model fairness Ghosh et al. (2023). Recent works such as Lev & Wilson (2024); Rubinstein & Hopkins (2025b) propose refinements of IF and NS which can improve the tradeoffs between running time and approximation accuracy.

State-of-the-Art: Accuracy Guarantees under Global Strong Convexity Assumptions Motivated by the extensive downstream applications, several recent works analyze the accuracy IF and NS methods Rad & Maleki (2018); Koh et al. (2019); Giordano et al. (2019); Wilson et al. (2020) for convex ERMs. Each of these works uses slightly different assumptions and notations, but roughly speaking they all prove variations of the same high-level statement:

Theorem 1.3 (Existing Theoretical Guarantees (Informal))**.** *If the loss function is "sufficiently strongly convex" over the entire optimization domain $\Omega_{\boldsymbol{\theta}}$ and its Hessian is "sufficiently Lipschitz" in this domain, then the single Newton step is "close" to the global optimum.*

The details of Theorem 1.3 and its proof vary slightly from paper to paper, but mostly follow a similar thread. To make things concrete, we consider the specific instantiation of Theorem 1.3.

Denote the gradients of the individual samples by $\mathbf{g}_i = \nabla \ell_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ and the Hessian of the loss by $\mathbf{H}_{\boldsymbol{\theta}} = \sum_{i \notin T} \nabla^2 \ell_i|_{\boldsymbol{\theta}}$. Using these notations, previous analyses typically make Assumptions 3, 1 and 2.

Assumption 1 (Lipschitz Hessian)**.** *There exists a finite constant*[2] $C_{\text{Lip}} < \infty$ *such that*

$$\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Omega_\theta \quad \|\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\boldsymbol{\theta}'}\|_{\text{op}} \leq C_{\text{Lip}} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

Assumption 2 (Bounded Individual Gradients)**.**

$$C_\ell = \max_{i \in [n]} \{\|\mathbf{g}_i\|_2\}$$

Assumption 3 (Global Strong Convexity)**.** *There exists a finite constant $C_{\text{op}} < \infty$ such that*

$$\forall \boldsymbol{\theta} \in \Omega_\theta \quad \left\| \left(\sum_{i \notin T} \nabla^2 \ell_i|_{\boldsymbol{\theta}} \right)^{-1} \right\|_{\text{op}} \leq C_{\text{op}}$$

Assumptions 1 and 2 suffice to (loosely) bound the norm of the gradient after one Newton step, and Assumption 3 gives us a quantitative way of converting this into a bound in parameter space.

Theorem 1.4 (Existing Theoretical Guarantees)**.** *Under Assumptions 1, 2 and 3, the error of the single Newton step data attribution is bounded by*

$$\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 = O(C_{\text{Lip}} C_{\text{op}}^3 k^2 C_\ell^2)\,.$$

For completeness, we prove Theorem 1.4 in Appendix B.

The biggest limitation of existing approaches is that they require a bound on the spectrum of the inverse Hessian on the entire optimization domain $\Omega_{\boldsymbol{\theta}}$ (Assumption 3). While some variants (e.g. Rad & Maleki (2018)) relax this assumption to one about the spectrum of the inverse Hessian (after samples are removed) on a subset of the domain large enough to include all of $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}^{\text{NS}}$, and $\hat{\boldsymbol{\theta}}_T$, this remains quite restrictive, unless one knows *a priori* that $\hat{\boldsymbol{\theta}}_T$ is close to $\hat{\boldsymbol{\theta}}$ in the first place, which is exactly what data attribution methods aim to discover.

As a simple running example, consider a logistic regression with feature vectors $\mathbf{x}_i \in \mathbb{R}^d$ and labels $y_i \in \{0, 1\}$. The Hessian of a logistic regression is given by $\mathbf{H} = \sum_{i \in [n]} \beta_i \mathbf{x}_i \mathbf{x}_i^\intercal$, where $\beta_i(\boldsymbol{\theta}) = \hat{y}_i(1 - \hat{y}_i)$ are the variances of this sample's prediction $\hat{y}_i = \text{softmax}(\boldsymbol{\theta}^\intercal \mathbf{x}_i)$.

In this case, $\left\|\mathbf{H}_{\boldsymbol{\theta}}^{-1}\right\|_{\text{op}}$ grows rapidly with the norm of the model $\|\boldsymbol{\theta}\|_2$. This is because $\beta_i = O(\exp(-|\boldsymbol{\theta}^\intercal \mathbf{x}_i|))$, so fixing the direction of $\boldsymbol{\theta}$ and taking its norm to infinity would almost surely cause the Hessian to decay exponentially. Therefore, if we perform our optimization over the domain $\Omega_{\boldsymbol{\theta}} = \mathbb{R}^d$, then $\left\|\mathbf{H}_{\boldsymbol{\theta}}^{-1}\right\|_{\text{op}}$ is not bounded and Assumption 3 does not hold.

[2]Keeping with existing nommenclature, we use the term "constant" to mean that C_{Lip} does not depend on $\boldsymbol{\theta}$, but it may still depend on n, k, d.

Global Strong Convexity via Regularization is Inadequate One approach to justifying this assumption Koh et al. (2019); Wilson et al. (2020) is by appeal to a L_2 regularization term to the loss, which gives a global lower bound on the spectrum of the Hessian. However, as Koh et al. note, the regularization coefficient used in practice tends to be very small, meaning that that the resulting bound on C_{op} is enormous, rendering the bound in Theorem C.1 rather weak – in particular, too weak to explain why NS still outperforms IF. Quoting Koh et al. (2019):

> Koh et al. (2019): The constraint in Proposition 3 implies that up to $O(1/\lambda^3)$ terms, influence underestimates the Newton approximation. [...] However, $\lambda/\sigma_{\text{max}}$ is quite small in our experiments [...] so the actual correlation of influence is better than predicted by this theory.

This "small λ" phenomenon is borne out in theory as well as experiments: it is a classical result in learning theory Dobriban & Wager (2018) that when optimizing for test-loss, the regularization coefficient λ scales with $\lambda = \Theta(\sigma^2 d/n) = O(d/n)$, where σ is the "signal-to-noise ratio" of the model. This introduces a "hidden" n^3-dependence into Theorem C.1 – even if we make favorable assumptions like $C_{\text{Lip}}, C_\ell, k, d \leq O(1)$ the bounds produced by theorems like Theorem C.1 would not even converge when $n \to \infty$:

$$\text{Existing Bounds} \simeq k^2 \times C_\ell^2 \times C_{\text{Lip}} \times C_{\text{op}}^3 \gtrsim_n 1 \times 1 \times n \times \frac{1}{\lambda^3} \gtrsim_n n\,.$$

1.3 Main Result on Convex Risk Minimization (Formal Version of Theorem 1.1)

We turn to the formal version of our main theorem for general convex risk minimization (Theorem 1.5). Comparing its informal version, Theorem 1.1, to the existing theory (see Theorem 1.3), the biggest change is that we assume only that the Hessian is well-behaved *in a small neighborhood of the first Newton step* and not over the entire optimization domain (or a subset of the domain whose radius is unknown without computing $\hat{\boldsymbol{\theta}}_T$ itself). At first this might seem like a small difference, but previous proof techniques break down without the global assumption which, as we have shown in Section 1.2, often does not hold in practice and may be hard to verify.

The second difference between Theorem 1.1 and existing theory is that we will require a much milder Lipschitz assumption on the change in the Hessian. Previous analyses require that the Hessian be Lipschitz in operator Koh et al. (2019); Wilson et al. (2020) or L_1 Giordano et al. (2019) norms, resulting in bounds that scale poorly with d and are harder to guarantee. For Theorem 1.1, it suffices to show that the Hessian is Lipschitz only in its change along a single direction, resulting in a tighter bound under a more easily verifiable assumption.

1.3.1 Assumptions and Theorem Statement

Let $T \subseteq [n]$ be a set of k samples to be removed. Let $\mathbf{g}_{\boldsymbol{\theta}} = \sum_{i \notin T} \nabla \ell_i|_{\boldsymbol{\theta}}$ and $\mathbf{H}_{\boldsymbol{\theta}} = \sum_{i \notin T} \nabla^2 \ell_i|_{\boldsymbol{\theta}}$ be the gradient and Hessian of the loss of the retained samples when evaluated at $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is not specified below, we will set $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ to be the global optimum of the loss L before samples are removed ($\mathbf{g} = \mathbf{g}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, $\mathbf{H} = \mathbf{H}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$).

Recall that the NS data attribution estimates that

$$\hat{\boldsymbol{\theta}}_T^{\text{NS}} \stackrel{\text{def}}{=} \hat{\boldsymbol{\theta}} - \mathbf{H}^{-1}\mathbf{g} \approx \hat{\boldsymbol{\theta}}_T\,.$$

Our first assumption will be that the Hessian is lower-bounded within a neighborhood of $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$. We will allow this neighborhood to assume the shape of any ellipsoid.

More concretely, let $\boldsymbol{\Sigma}$ be any positive-definite whitening matrix (natural choices could be the identity matrix $\boldsymbol{\Sigma} = \mathbf{I}$ or the Hessian $\boldsymbol{\Sigma} = \mathbf{H}$ at the original model $\theta = \hat{\boldsymbol{\theta}}$), and define $\|\mathbf{v}\|_{\boldsymbol{\Sigma}} := \mathbf{v}^\intercal \boldsymbol{\Sigma} \mathbf{v}$. Let $r > 0$ be any positive radius, and let $\mathcal{B}$ be the $\boldsymbol{\Sigma}$ ball of radius $r > 0$ around $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$

$$\mathcal{B}_{\boldsymbol{\Sigma},r} \stackrel{\text{def}}{=} \left\{ \hat{\boldsymbol{\theta}}_T^{\text{NS}} + \mathbf{e} \mid \|\mathbf{e}\|_{\boldsymbol{\Sigma}} \leq r \right\}\,.$$

Assumption 4 (Strong Convexity in $\mathcal{B}$). $\forall \boldsymbol{\theta} \in \mathcal{B} \quad \left\|\boldsymbol{\Sigma}^{1/2} \mathbf{H}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}^{1/2}\right\|_{\text{op}} \leq C_{\text{op}}$

Assumption 4 avoids the aforementioned issues with the previous analyses by limiting our evaluation of the Hessian to just a neighborhood of the Newton step (thus avoiding issues with potentially large changes to the Hessian at $\boldsymbol{\theta}$ far from $\hat{\boldsymbol{\theta}}$).

The next assumption tells us that Hessian changes slowly along the first Newton step:

Assumption 5 (Mildly Lipschitz Hessian)**.** $\forall \boldsymbol{\theta} = t\hat{\boldsymbol{\theta}} + (1-t)\hat{\boldsymbol{\theta}}_T^{\text{NS}} \quad \left\|\left(\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}\right)\mathbf{H}^{-1}\mathbf{g}\right\|_{\boldsymbol{\Sigma}^{-1}} \leq C_h$

Finally, similar to some previous analyses (e.g., Giordano et al. (2019)[Condition 1]), we also require a condition on the relationship between the parameters in our bound:

Condition 6. $C_h C_{\text{op}} < r$

Condition 6 encapsulates a non-trivial tradeoff, since on the one hand, increasing r makes the right-hand-side of this condition larger, helping us satisfy it, but on the other hand, this also increases the domain over which C_h and C_{op} are maximized.

Under these assumptions, we can bound the error of the Newton step approximation:

Theorem 1.5 (Main Result)**.** *Under Assumptions 4 and 5 and Condition 6, we have*

$$\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}}\right\|_{\boldsymbol{\Sigma}} \leq C_h C_{\text{op}}\,.$$

The proof of Theorem 1.5 is relatively simple and will follow by combining ideas from self-concordance theory Bach (2010); Hsu & Mazumdar (2024) and the analyses of Giordano et al. (2019); Wilson et al. (2020). Recall that Theorem 1.2 illustrates the power of Theorem 1.5 to analyze NS and IF data attribution.

2 Notation

General Notations We use lower-case Greek and Latin letters ($a, b, c, \alpha, \beta, \gamma$) to denote scalars and indices, bold lower lower-case letters ($\mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) to denote vectors and bold upper-case letters ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \boldsymbol{\Gamma}, \boldsymbol{\Pi}$) to denote matrices and higher order tensors.

Problem Settings Let $L : \mathbb{R}^d \to \mathbb{R}$ be a smooth convex loss function with gradient $\mathbf{g}_{\boldsymbol{\theta}} = \nabla L|_{\boldsymbol{\theta}}$ and Hessian $\mathbf{H}_{\boldsymbol{\theta}} = \nabla^{\otimes 2} L|_{\boldsymbol{\theta}}$, and let $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ be some initial point.

3 Proof of Theorem 1.5

In this section, we sketch the proof Theorem 1.5 which bounds the approximation error of NS data attributions. Recall that our assumptions for Theorem 1.5 were that the loss is strongly convex in a region $\mathcal{B}$ surrounding the NS data attribution $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$ and that the Hessian did not change rapidly along the NS path $[\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_T^{\text{NS}}]$, and our goal is to bound the distance between $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$.

Our proof of Theorem 1.5 will combine ideas from existing analyses of NS data attributions with ideas from self-concordance theory. We break the proof of Theorem 1.5 into 3 lemmas, which we state and motivate here, but defer the detailed proofs to Appendix A.1.

3.1 Bounded Gradient at $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$

Lemma 3.1 (Bounded Gradient at $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$)**.** *The loss gradient at $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$ is bounded by*

$$\left\|\mathbf{g}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T^{\text{NS}}}\right\|_{\Sigma^{-1}} \leq C_h\,.$$

Lemma 3.1 and its proof are similar to analogous bounds in many of the previous analyses of NS data attribution. The key difference between Lemma 3.1 and previous analyses is that through a few careful choices in our definition of Assumption 5, we are able to ensure that our bound on gradient is much tighter.

Previous analyses of this quantity scale like

$$\text{Previous Analyses} = \left\|\mathbf{H}^{-1}\right\|_{\text{op}}^2 \times C_{\text{Lip}} \times k^2 \times \max_{i\in[n]}\{\|\mathbf{x}_i\|_2\}^2 \approx \frac{1}{\lambda^2 n^2} \times (n + d^{3/2}) \times k^2 \times d\,.$$

Previous bounds are loose for a few reasons. First, since C_{Lip} is defined by the Lipschitz constraint of the Hessian (i.e., the rate at which the Hessian changes when moving in an adversarially selected direction), we show that even for normally distributed samples this is $\Theta(n)$ if we move along the direction of the current model $\boldsymbol{\theta}$. Simply by limiting our scope to models along the NS path and to changes in $\|\mathbf{Hg}\|$ (instead of $\|\mathbf{H}\|$), we are able to reduce the effective C_{Lip} from n to a $\widetilde{O}\left(\frac{n}{d}\right)$ scaling.

Moreover, instead of applying a triangle inequality for the ℓ_2 squared norm of the gradient at $\hat{\boldsymbol{\theta}}$ (which would scale like $k^2 d$), we utilize the fact that a "blessing of dimensionality" bounds this to being $\widetilde{O}\left(kd\right)$ Jin et al. (2019).

Finally we avoid a $\frac{1}{\lambda^2}$ factor simply from the fact that C_h depends only on the gradient at $\hat{\boldsymbol{\theta}}$, yielding a scaling of (see Appendix D.2).

$$C_h = \widetilde{O}\left(\frac{k}{n}\right)\,.$$

Proof of Lemma 3.1.

$$\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}} = \int_{\hat{\boldsymbol{\theta}}}^{\hat{\boldsymbol{\theta}}_T^{\text{NS}}} \mathbf{H}_{\boldsymbol{\theta}} \mathrm{d}\boldsymbol{\theta} - \mathbf{g}_{\hat{\boldsymbol{\theta}}} = \int_0^t \left(\mathbf{H}_{\boldsymbol{\theta}=t\hat{\boldsymbol{\theta}}_T^{\text{NS}}+(1-t)\hat{\boldsymbol{\theta}}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}}_T^{\text{NS}} - \hat{\boldsymbol{\theta}}\right)\mathrm{d}t + \underbrace{\mathbf{H}_{\hat{\boldsymbol{\theta}}}\left(\hat{\boldsymbol{\theta}}_T^{\text{NS}} - \hat{\boldsymbol{\theta}}\right) - \mathbf{g}_{\hat{\boldsymbol{\theta}}}}_{=0} =$$

$$= \int_0^t \left(\mathbf{H}_{\boldsymbol{\theta}=t\hat{\boldsymbol{\theta}}_T^{\text{NS}}+(1-t)\hat{\boldsymbol{\theta}}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right)\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}\mathrm{d}t$$

From Assumption 5, the integrand on the right hand side of this equation has bounded $\boldsymbol{\Sigma}^{-1}$ norm, so from the triangle inequality, we have

$$\left\|\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}}\right\|_{\boldsymbol{\Sigma}^{-1}} = \left\|\int_0^t \left(\mathbf{H}_{\boldsymbol{\theta}=t\hat{\boldsymbol{\theta}}_T^{\text{NS}}+(1-t)\hat{\boldsymbol{\theta}}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right)\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}\mathrm{d}t\right\|_{\boldsymbol{\Sigma}^{-1}} \leq C_h\,.$$

□

3.2 $\hat{\boldsymbol{\theta}}_T \in \mathcal{B}$

The second portion of our proof will be to show that $\hat{\boldsymbol{\theta}}_T$ lies within the region $\mathcal{B}$ around the first Newton step. We do this using ideas from self-concordant analysis Bach (2010); Hsu & Mazumdar (2024).

Lemma 3.2. $\hat{\boldsymbol{\theta}}_T \in \mathcal{B}$

Proof of Lemma 3.2. Assume for contradiction that $\hat{\boldsymbol{\theta}}_T \notin \mathcal{B}$. Let $\boldsymbol{\theta}_{\mathcal{B}}$ denote the intersection between the line segment from $\hat{\boldsymbol{\theta}}_T$ to $\hat{\boldsymbol{\theta}}_T^{\text{NS}}$ and $\partial\mathcal{B}$, and denote $\mathbf{d} \overset{\text{def}}{=} \hat{\boldsymbol{\theta}}_T^{\text{NS}} - \hat{\boldsymbol{\theta}}_T$

(A) Cauchy–Schwarz upper bound. By Cauchy–Schwarz,

$$\langle \mathbf{d},\, \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}}\rangle \;\leq\; \left\|\boldsymbol{\Sigma}^{1/2}\mathbf{d}\right\|_2 \left\|\boldsymbol{\Sigma}^{-1/2}\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}}\right\|_2 \;\leq\; C_h \left\|\boldsymbol{\Sigma}^{1/2}\mathbf{d}\right\|_2,$$

where the last step uses Lemma 3.1 .

(B) Fundamental theorem of calculus for g. Along the segment $\boldsymbol{\theta}(t) \overset{\text{def}}{=} \hat{\boldsymbol{\theta}}_T + t\mathbf{d}$ with $t \in [0,1]$,

$$\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}} \;=\; \int_0^1 \mathbf{H}_{\boldsymbol{\theta}(t)}\,\mathbf{d}\,\mathrm{d}t, \quad \text{so} \quad \langle \mathbf{d}, \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\text{NS}}}\rangle = \int_0^1 \mathbf{d}^\top \mathbf{H}_{\boldsymbol{\theta}(t)}\mathbf{d}\,\mathrm{d}t.$$

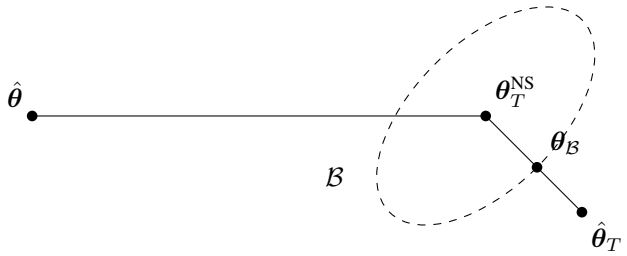

Figure 1: Diagram of proof of Lemma 3.2.

(C) Nonnegativity of the integrand. By weak convexity, $\mathbf{H}_{\boldsymbol{\theta}} \succeq 0$ everywhere, so $\mathbf{d}^\top \mathbf{H}_{\boldsymbol{\theta}(t)} \mathbf{d} \geq 0$ for all t.

(D) Geometry inside $\mathcal{B}$. Let $t_{\mathcal{B}} \in [0, 1)$ be such that $\boldsymbol{\theta}(t_{\mathcal{B}}) = \boldsymbol{\theta}_{\mathcal{B}}$ (the first time the segment enters $\mathcal{B}$ when moving from $\hat{\boldsymbol{\theta}}_T$ to $\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$). Then for all $t \in [t_{\mathcal{B}}, 1]$ we have $\boldsymbol{\theta}(t) \in \mathcal{B}$ and, by definition of r,

$$1 - t_{\mathcal{B}} \;=\; \frac{\left\| \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} - \boldsymbol{\theta}_{\mathcal{B}} \right\|_2}{\|\mathbf{d}\|_2} \;=\; \frac{r}{\|\mathbf{d}\|_2}.$$

(E) Strong convexity in $\mathcal{B}$. From Assumption 4, we have

$$\forall \boldsymbol{\theta} \in \mathcal{B} \left\| \boldsymbol{\Sigma}^{1/2} \mathbf{H}_{\boldsymbol{\theta}}^{-1} \boldsymbol{\Sigma}^{1/2} \right\|_{\mathrm{op}} \leq C_{\mathrm{op}} \Rightarrow \boldsymbol{\Sigma}^{-1/2} \mathbf{H}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}^{-1/2} \succeq C_{\mathrm{op}}^{-1} \mathbf{I}\,.$$

Therefore,

$$\langle \mathbf{d}, \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \rangle = \int_0^1 \mathbf{d}^\top \mathbf{H}_{\boldsymbol{\theta}(t)} \mathbf{d}\, \mathrm{d}t \;\geq\; \int_{t_{\mathcal{B}}}^1 \mathbf{d}^\top \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{-1/2} \mathbf{H}_{\boldsymbol{\theta}(t)} \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{1/2} \mathbf{d}\, \mathrm{d}t \;\geq\; \int_{t_{\mathcal{B}}}^1 C_{\mathrm{op}}^{-1} \left\| \boldsymbol{\Sigma}^{1/2} \mathbf{d} \right\|_2^2 \mathrm{d}t = C_{\mathrm{op}}^{-1}\, r\, \|\mathbf{d}\|_{\boldsymbol{\Sigma}}\,.$$

(F) Contradiction. Combining the upper bound from (A) and the lower bound from (E),

$$C_{\mathrm{op}}^{-1}\, r\, \|\mathbf{d}\|_{\boldsymbol{\Sigma}} \;\leq\; \langle \mathbf{d}, \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \rangle \;\leq\; C_h\, \|\mathbf{d}\|_{\boldsymbol{\Sigma}}\,.$$

Since $\hat{\boldsymbol{\theta}}_T \neq \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$ (otherwise it would be in $\mathcal{B}$), we have $\|\mathbf{d}\|_{\boldsymbol{\Sigma}} > 0$ and we can cancel it to obtain

$$r \;\leq\; C_h\, C_{\mathrm{op}}\,,$$

violating Condition 6, completing our proof. □

3.3 Concluding Theorem 1.5

Finally, we will use the fact that the ground truth lies within the region $\mathcal{B}$ where the loss is strongly convex to convert our bound on the norm of the gradient into a bound in parameter space.

Lemma 3.3. *Lemmas 3.1 and 3.2 imply Theorem 1.5.*

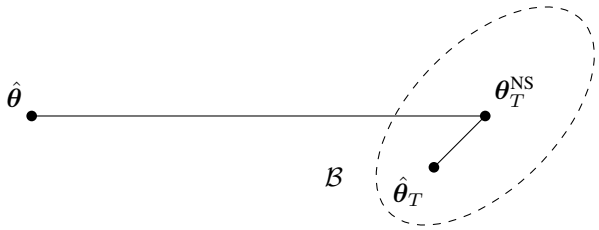

Figure 2: Diagram of proof of Lemma 3.3.

Proof of Lemma 3.3. Let $\mathbf{d} \stackrel{\mathrm{def}}{=} \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$ and consider the segment $\boldsymbol{\theta}(t) \stackrel{\mathrm{def}}{=} \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} + t\mathbf{d}$ for $t \in [0, 1]$.

Step 1: Bounded gradient at $\hat{\theta}_T^{\mathrm{NS}}$. By Lemma 3.1 and the dual pairing between $\|\cdot\|_{\mathbf{\Sigma}}$ and $\|\cdot\|_{\mathbf{\Sigma}^{-1}}$,

$$\langle \mathbf{d}, \mathbf{g}_{\hat{\theta}_T^{\mathrm{NS}}} \rangle \;\le\; \|\mathbf{d}\|_{\mathbf{\Sigma}} \left\| \mathbf{g}_{\hat{\theta}_T^{\mathrm{NS}}} \right\|_{\mathbf{\Sigma}^{-1}} \;\le\; C_h \, \|\mathbf{d}\|_{\mathbf{\Sigma}} \,. \tag{1}$$

Step 2: The path lies in $\mathcal{B}$. By Lemma 3.2 (using Condition 6), we have $\hat{\theta}_T \in \mathcal{B} \equiv \mathcal{B}_{\mathbf{\Sigma},r}$. Since $\mathcal{B}$ is a $\mathbf{\Sigma}$-ball centered at $\hat{\theta}_T^{\mathrm{NS}}$, it is convex. Therefore, $\boldsymbol{\theta}(t) \in \mathcal{B}$ for all $t \in [0,1]$.

Step 3: Curvature lower bound along the path. Assumption 4 gives, for all $t \in [0,1]$,

$$\mathbf{\Sigma}^{-1/2} \mathbf{H}_{\boldsymbol{\theta}(t)} \mathbf{\Sigma}^{-1/2} \succeq C_{\mathrm{op}}^{-1} \mathbf{I} \quad \Longrightarrow \quad \mathbf{d}^\top \mathbf{H}_{\boldsymbol{\theta}(t)} \mathbf{d} \;\ge\; C_{\mathrm{op}}^{-1} \, \mathbf{d}^\top \mathbf{\Sigma} \mathbf{d} \;=\; C_{\mathrm{op}}^{-1} \, \|\mathbf{d}\|_{\mathbf{\Sigma}}^2 \,.$$

Using the fundamental theorem of calculus for $\mathbf{g}$,

$$\mathbf{g}_{\hat{\theta}_T^{\mathrm{NS}}} \;=\; \int_0^1 \mathbf{H}_{\boldsymbol{\theta}(t)} \, \mathbf{d} \, \mathrm{d}t, \qquad \text{so} \qquad \langle \mathbf{d}, \mathbf{g}_{\hat{\theta}_T^{\mathrm{NS}}} \rangle \;=\; \int_0^1 \mathbf{d}^\top \mathbf{H}_{\boldsymbol{\theta}(t)} \mathbf{d} \, \mathrm{d}t \;\ge\; C_{\mathrm{op}}^{-1} \, \|\mathbf{d}\|_{\mathbf{\Sigma}}^2 \,. \tag{2}$$

Step 4: Combine (1) and (2). From (1) and (2),

$$C_{\mathrm{op}}^{-1} \, \|\mathbf{d}\|_{\mathbf{\Sigma}}^2 \;\le\; \langle \mathbf{d}, \mathbf{g}_{\hat{\theta}_T^{\mathrm{NS}}} \rangle \;\le\; C_h \, \|\mathbf{d}\|_{\mathbf{\Sigma}} \Rightarrow \|\mathbf{d}\|_{\mathbf{\Sigma}} = \left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_{\mathbf{\Sigma}} \;\le\; C_h \, C_{\mathrm{op}} \,.$$

□

4 Discussion and Limitations

Our work provides the first (to our knowledge) analysis of NS and IF data attribution methods which explains quantitatively the significant accuracy advantages of NS, in the context of a simple learning problem (such as logistic regression with Gaussian covariates). Under reasonable assumptions (see Theorem 1.2), this analysis even shows almost-matching upper and lower bounds on the scaling rate of the errors of these approximations with respect to k, n, and d. We do so via the more-general Theorem 1.5, which shows how to analyze the NS approximation without appeal to non-local strong convexity assumptions. Together, these results place IF and NS on firmer theoretical foundations.

Future Directions – Machine Unlearning and Beyond Quantitative accuracy bounds for IF and NS are of more than theoretical interest. One important application of IF and NS data attribution is machine unlearning, where the goal is to quickly "remove" samples from an already-learned model, e.g. to protect copyrighted material or respect "right to be forgotten" user requests Sekhari et al. (2021a); Neel et al. (2021); Suriyakumar & Wilson (2022). The dominant technique used in machine unlearning methods with provable guarantees is to first use a data attribution method to approximate the data-dropped model, then add noise to the resulting estimate to obtain a differential-privacy-like indistinguishability guarantee. Crucially, the magnitude of this noise – and hence the utility of the resulting model – scales with the best available bound on the accuracy of the data attribution method used. The $1/\lambda^3$ scaling of the best bounds prior to our work (where λ is a global ℓ_2 regularization parameter) is a significant bottleneck in machine unlearning Sekhari et al. (2021b). So, we hope that the better bounds on accuracy of data attribution methods we derive here will lead to much better unlearning algorithms.

Moreover, several recent works have recently proposed refinements to IF methods Lev & Wilson (2024); Rubinstein & Hopkins (2025b); Zou et al. (2025), retaining some of the computation speed of IF but seemingly improving accuracy. Do these methods close or narrow the quantitative gap we find here between IF and NS?

Limitations Our work is purely theoretical, aiming to explain a phenomenon observed in real-world data Koh et al. (2019). We focus on relatively simple settings – convex empirical risk minimization, logistic regression, Gaussian data – to sharpen the focus on the core phenomenon we study. Although data attribution for convex models sometimes forms a core component of data attribution for neural nets, our theorems do not directly speak to the effectiveness of IF methods applied to neural nets, nor do they allow for other "bells and whistles" such as non-smooth regularizers Suriyakumar & Wilson (2022). At a more detailed level, our characterization of NS and IF error scalings for Gaussian logistic regression problems in Theorem 1.2 applies only in the large-n regime, requiring $n \ge d^3$; characterizing the scaling rates allowing $n \ge d$ is an interesting direction for future work.

REFERENCES

scribe Aleksandr Podkopaev and lecturer Alessandro Rinaldo. Lecture 5: Sub-exponential random variables and orlicz norms. Scribed lecture notes 36-710 / 36-709, Spring 2019, Carnegie Mellon University, Department of Statistics & Data Science, Pittsburgh, PA, USA, February 2019. URL https://www.stat.cmu.edu/~arinaldo/Teaching/36709/S19/Scribed_Lectures/Feb5_Aleksandr.pdf. Scribed by Aleksandr Podkopaev.

Francis Bach. Self-concordant analysis for logistic regression. 2010.

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

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

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

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

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

Soumya Ghosh, Prasanna Sattigeri, Inkit Padhi, Manish Nagireddy, and Jie Chen. Influence based approaches to algorithmic fairness: A closer look. In *XAI in action: past, present, and future applications*, 2023.

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

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

Daniel Hsu and Arya Mazumdar. On the sample complexity of parameter estimation in logistic regression with normal design. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2418–2437. PMLR, 2024.

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

Jenny Y Huang, David R Burt, Tin D Nguyen, Yunyi Shen, and Tamara Broderick. Approximations to worst-case data dropping: unmasking failure modes. *arXiv preprint arXiv:2408.09008*, 2024.

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

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

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael 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 (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

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

Omri Lev and Ashia C Wilson. The approximate fisher influence function: Faster estimation of data influence in statistical models. *arXiv preprint arXiv:2407.08169*, 2024.

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

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

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

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

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

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

Ittai Rubinstein and Samuel B. Hopkins. Robustness auditing for linear regression: To singularity and beyond. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025a. URL https://openreview.net/forum?id=V5ns6uvRZ9.

Ittai Rubinstein and Samuel B Hopkins. Rescaled influence functions: Accurate data attribution in high dimension. *arXiv preprint arXiv:2506.06656*, 2025b.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 18075–18086, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/9627c45df543c816a3ddf2d8ea686a99-Abstract.html.

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

Vinith M. Suriyakumar and Ashia C. Wilson. Algorithms that approximate data removal: New results and limitations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/77c7faab15002432ba1151e8d5cc389a-Abstract-Conference.html.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Huiming Zhang and Haoyu Wei. Sharper sub-weibull concentrations. *Mathematics*, 10(13):2252, 2022.

Haolin Zou, Arnab Auddy, Yongchan Kwon, Kamiar Rahnama Rad, and Arian Maleki. Newfluence: Boosting model interpretability and understanding in high dimensions. *arXiv preprint arXiv:2507.11895*, 2025.

A Deferred Proofs

A.1 Proofs of Lemmas 3.1, 3.2, 3.3

B Proof Theorem 1.4

A common approach to proving Theorem 1.4 Koh et al. (2019) is to bound the size of the loss gradient at the end of this first Newton step. Let

$$\mathbf{g} = \sum_{i \notin T} \nabla \ell_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\sum_{i \in T} \nabla \ell_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$$

be the gradient of the loss at the starting point $\hat{\boldsymbol{\theta}}$, and let

$$\mathbf{g}^{\mathrm{NS}} = \sum_{i \notin T} \nabla \ell_i|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}}$$

be the gradient of the loss at the end of the first Newton step. By design of the Newton step, as long as the Hessian does not change too much, we expect $\mathbf{g}^{\mathrm{NS}}$ to be small. More concretely, we have

$$\mathbf{g}^{\mathrm{NS}} = \mathbf{g} + \int_{\hat{\boldsymbol{\theta}}}^{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \mathbf{H}_{\boldsymbol{\theta}} \mathrm{d}\boldsymbol{\theta} = \underbrace{\mathbf{g} + \int_{\hat{\boldsymbol{\theta}}}^{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \mathbf{H}_{\hat{\boldsymbol{\theta}}} \mathrm{d}\boldsymbol{\theta}}_{=\mathbf{0}} + \int_{\hat{\boldsymbol{\theta}}}^{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \left(\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right) \mathrm{d}\boldsymbol{\theta}\,,$$

where the first term cancels precisely by design of the Newton step, and the second term is bounded by

$$\begin{aligned}
\left\| \int_{\hat{\boldsymbol{\theta}}}^{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \left(\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right) \mathrm{d}\boldsymbol{\theta} \right\|_2 &= \left\| \int_0^1 \left(\mathbf{H}_{\boldsymbol{\theta}=(1-t)\hat{\boldsymbol{\theta}}+t\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right) \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}\right) \mathrm{d}t \right\|_2 \leq \text{ (triangle inequality)} \\
&\leq \int_0^1 \left\| \left(\mathbf{H}_{\boldsymbol{\theta}=(1-t)\hat{\boldsymbol{\theta}}+t\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}}\right) \right\|_{\mathrm{op}} \left\| \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \mathrm{d}t \leq \text{ (Lipschitz assumption)} \\
&\leq \int_0^1 C_{\mathrm{Lip}} t \left\| \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2^2 \mathrm{d}t = \frac{1}{2} C_{\mathrm{Lip}} \left\| \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2^2 = \frac{1}{2} C_{\mathrm{Lip}} \left\| \mathbf{H}^{-1} \mathbf{g} \right\|_2^2
\end{aligned}$$

Given a bound on the norm of the gradient at $\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$, we can deduce that $\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$ is close to the optimum by utilizing our global strong convexity assumption:

$$\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \leq \max_{\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}} \left\{ \left\| \mathbf{H}_{\boldsymbol{\theta}}^{-1} \right\|_{\mathrm{op}} \right\} \left\| \mathbf{g}^{\mathrm{NS}} \right\| \leq \frac{1}{2} C_{\mathrm{Lip}} C_{\mathrm{op}} \left\| \mathbf{H}^{-1} \mathbf{g} \right\|_2^2 \leq \frac{1}{2} C_{\mathrm{Lip}} C_{\mathrm{op}}^3 \left\| \mathbf{g} \right\|_2^2 \leq \frac{C_{\mathrm{Lip}} C_{\mathrm{op}}^3 k^2 C_\ell^2}{2} \tag{3}$$

where $C_\ell := \max_{i \in [n]} \left\{ \left\| \nabla \ell_i|_{\hat{\boldsymbol{\theta}}} \right\|_2 \right\}$ and the last two steps utilized the fact that $\left\| \mathbf{H}^{-1} \right\|_{\mathrm{op}} = \left\| \mathbf{H}_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{-1}} \right\|_{\mathrm{op}} \leq C_{\mathrm{op}}$ and the triangle inequality.

C Asymptotic Analysis - Setting and Theorem Statements

Existing analyses of NS and IF are often parametrized as a function of complex quantities like C_{op} and C_h above, making it difficult to compare their results. To get a sense for the asymptotic behavior of Theorem 1.5 and how it compares to previous results, we analyze their respective asymptotic behavior for well-behaved logistic regressions.

In particular, we analyze the existing bounds on the NS approximation error (Theorem 1.4), our new bounds (Theorem 1.5) and the distance between IF and NS in "under-parametrized agnostic learning with normally distributed features". More concretely, we assume that the features and labels are drawn i.i.d from some underlying distribution $\mathbf{x}, y \sim \mathcal{X} \times \mathcal{Y}$ such that:

- The problem is under-parametrized ($n \geq d^2$) and the number of samples being removed is at most $k \leq n/\text{polylog}(n)$.
- The features are iid normally distributed $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
- The optimal model
$$\boldsymbol{\theta}^\star \stackrel{\text{def}}{=} \underset{\boldsymbol{\theta} \in \Omega_\theta}{\text{argmin}} \left\{ \underset{\mathbf{x}, y \sim \mathcal{X} \times \mathcal{Y}}{\mathbb{E}} \left[\ell(y, \langle \boldsymbol{\theta}, \mathbf{x} \rangle \right] \right\},$$
(where ℓ is the logistic loss) has norm $\|\boldsymbol{\theta}^\star\|_2 = \Theta(1)$, though we do not assume that the labels were generated by this model.
- For our upper bounds, we will show that the given quantity is bounded for adversarial drop sets T of size k, and for our lower bounds, we will show that the expectation of this quantity is large for random drops sets T of this size.

Under the assumptions above, we can derive nearly matching upper and lower bounds for the existing bounds on NS accuracy and the distance between IF and NS estimates

Theorem C.1 (Asymptotic Analysis of Existing Bounds)**.** *Under the assumptions above, with high probability (over the randomness of* $\mathbf{X}, \mathbf{y}$*),*

$$\max_{T \in \binom{[n]}{k}} \left\{ \left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 \right\} \leq \max_{T \in \binom{[n]}{k}} \{\text{Existing Bounds}\} = \widetilde{O}\left(\frac{k^2 d}{n^2 \lambda^3} \right),$$

and

$$\mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} [\text{Existing Bounds}] = \Omega\left(\frac{k^2 d}{n^2 \lambda^3} \right).$$

Theorem C.2 (Asymptotic Analysis of Influence Functions)**.** *Under the assumptions above, with high probability*

$$\max_{T \in \binom{[n]}{k}} \left\{ \left\| \hat{\boldsymbol{\theta}}_T^{\text{IF}} - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 \right\} = \widetilde{O}\left(\frac{(k+d)\sqrt{kd}}{n^2} \right),$$

and

$$\mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\| \hat{\boldsymbol{\theta}}_T^{\text{IF}} - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 \right] = \Omega\left(\frac{(k+d)\sqrt{kd}}{n^2} \right).$$

Theorem C.3 (Asymptotic Analysis of Theorem 1.5)**.** *Under the assumptions above, with high probability Theorem 1.5 yields a bound of order*

$$\max_{T \in \binom{[n]}{k}} \left\{ \left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 \right\} \leq \max_{T \in \binom{[n]}{k}} \{\text{New Bounds}\} = \widetilde{O}\left(\frac{k}{n^2} \right),$$

and these bounds are tight up to poly-logarithmic factors

$$\mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \right\|_2 \right] = \Omega\left(\frac{k}{n^2} \right).$$

D Asymptotic Scaling of Newton Step Accuracy

In this appendix we prove Theorem C.3, which gives tight asymptotic bounds on the accuracy of the Newton step estimator under random or adversarial sample removal. Our goal is to establish that

$$\max_{T \in \binom{[n]}{k}} \left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 = \widetilde{O}\left(\frac{k}{n^2}\right), \qquad \mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \right] = \Omega\left(\frac{k}{n^2}\right),$$

under the assumptions stated in Section C.

D.1 Model Gradient Inner Product

A key intermediate step is to control the alignment between iterates along the path between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}^{\mathrm{NS}}$, namely to show that

Lemma D.1. *Under the assumptions of Section C, with high probability,*

$$\sup_{\boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}^{\mathrm{NS}}]} \left| \langle \boldsymbol{\theta}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle \right| = \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Lemma D.2 (Warm-up at $\boldsymbol{\theta}^\star$)**.** *Under the assumptions of Section C, with high probability,*

$$\left| \langle \boldsymbol{\theta}^\star, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1} \mathbf{g}_{\boldsymbol{\theta}^\star} \rangle \right| = \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Lemma D.2 is essentially equivalent to Rubinstein & Hopkins (2025a)[Claim 7.10], and follows from a similar proof.

D.1.1 Claim: $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}^{\mathrm{NS}}(\boldsymbol{\theta}^\star)$

The first step is to relate the empirical MLE $\hat{\boldsymbol{\theta}}$ to a Newton step starting from the population optimum $\boldsymbol{\theta}^\star$. At first sight this may appear strange: $\hat{\boldsymbol{\theta}}$ is designed to approximate $\boldsymbol{\theta}^\star$, so why apply a Newton step from $\boldsymbol{\theta}^\star$? The reason is analytical convenience. Unlike $\hat{\boldsymbol{\theta}}$, the true parameter $\boldsymbol{\theta}^\star$ is independent of the sample randomness. This independence makes it possible to apply powerful concentration inequalities when bounding deviations of gradients and Hessians. Therefore, it is very useful to have a quantitative control of $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{NS}}(\boldsymbol{\theta}^\star)\|_2$ even though $\boldsymbol{\theta}^{\mathrm{NS}}(\boldsymbol{\theta}^\star)$ is not a model we would actually compute in practice.

Claim D.3. *Let $\boldsymbol{\theta}^{\mathrm{NS}}(\boldsymbol{\theta}^\star)$ denote the Newton step starting from the population optimum $\boldsymbol{\theta}^\star$. Then under the assumptions of Section C, with high probability*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{NS}}(\boldsymbol{\theta}^\star)\|_2 \le \widetilde{O}\left(\frac{d}{n}\right).$$

Proof. We apply Theorem 1.5 in the special case $\boldsymbol{\Sigma} = \mathbf{I}$ and radius $r = 1$:

Step 1: Verifying Assumption 4. Lemma G.7 implies that with high probability the empirical Hessian $\nabla^2 L_n(\boldsymbol{\theta})$ is uniformly close to its expectation across the relevant domain. In particular, $\nabla^2 L_n(\boldsymbol{\theta})$ is positive definite with condition number $\Omega(n)$, thereby satisfying Assumption 4.

Step 2: Verifying Assumption 5. We need to control

$$\|\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\boldsymbol{\theta}^\star}\|_{\mathrm{op}}.$$

The Hessian is Lipschitz in $\boldsymbol{\theta}$ with Lipschitz constant bounded by

$$C_{\mathrm{Lip}} = \max_{\boldsymbol{\theta} \in \Omega_\theta} \left\{ \|\mathbf{T}\|_{\mathrm{op}} \right\} \le \max_{\mathbf{e} \in \mathbb{S}^{d-1}} \sum_{i=1}^{n} |\langle \mathbf{x}_i, \mathbf{e} \rangle|^3.$$

By Lemma G.2, this quantity is at most $O(n + d^{3/2}) = O(n)$ since $n \geq d^2$. Combining with the fact that $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 = O(\sqrt{d/n})$ (G.5), we obtain

$$\|\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\boldsymbol{\theta}^\star}\|_{\text{op}} = O\left(n \cdot \sqrt{\frac{d}{n}}\right) = O(\sqrt{nd}).$$

Multiplying by $\left\|\mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\text{op}} = O(1/n)$ and $\|\mathbf{g}_{\boldsymbol{\theta}^\star}\| = O(\sqrt{nd})$ from Lemma G.10, the overall contribution to Assumption 5 is bounded by $O(d)$.

Step 3: Apply Theorem 1.5. Thus the conditions of Theorem 1.5 are satisfied with $C_h = O(d)$ and $C_{\text{op}} = O(1/n)$. The theorem then yields

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\text{NS}}(\boldsymbol{\theta}^\star)\|_2 \;\leq\; C_h C_{\text{op}} = \widetilde{O}\left(\frac{d}{n}\right).$$

□

D.1.2 $\left|\langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle\right| = \widetilde{O}\left(\frac{\sqrt{k}}{n}\right)$

Claim D.4. *Under the assumptions of Section C, with high probability,*

$$\left|\langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle\right| \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Proof. Start from the warm-up bound at $\boldsymbol{\theta}^\star$ (Lemma D.2):

$$\left|\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1} \mathbf{g}_{\boldsymbol{\theta}^\star} \rangle\right| \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right). \tag{4}$$

We telescope the difference between the target quantity at $\hat{\boldsymbol{\theta}}$ and the warm-up quantity at $\boldsymbol{\theta}^\star$:

$$\langle \hat{\boldsymbol{\theta}}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle = \langle \boldsymbol{\theta}^\star, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1} \mathbf{g}_{\boldsymbol{\theta}^\star} \rangle \tag{5}$$

$$+ \underbrace{\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star,\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle}_{(I)} + \underbrace{\langle \boldsymbol{\theta}^\star,\, (\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} - \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1})\, \mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle}_{(II)} + \underbrace{\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1} (\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\boldsymbol{\theta}^\star}) \rangle}_{(III)}. \tag{6}$$

We will show that each of (I)–(III) is $\widetilde{O}\left(\sqrt{k}/n\right)$, leaving the remaining "Newton–step correction"

$$\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1} (\mathbf{g}_{\text{NS}}(\boldsymbol{\theta}^\star) - \mathbf{g}_{\boldsymbol{\theta}^\star}) \rangle$$

to be handled next.

Auxiliary bounds. We use the following high-probability estimates:

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;=\; \widetilde{O}\left(\sqrt{\frac{d}{n}}\right), \tag{7}$$

$$\left\|\mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\text{op}} \;=\; O\left(\frac{1}{n}\right), \qquad \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} \;=\; O\left(\frac{1}{n}\right), \tag{8}$$

$$\left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}} - \mathbf{H}_{\boldsymbol{\theta}^\star}\right\|_{\text{op}} \;\leq\; L_H \cdot \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \;=\; \widetilde{O}\left(\sqrt{nd}\right), \tag{9}$$

$$\left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} - \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\text{op}} \;\leq\; \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}} - \mathbf{H}_{\boldsymbol{\theta}^\star}\right\|_{\text{op}} \left\|\mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\text{op}} \;=\; \widetilde{O}\left(\frac{\sqrt{nd}}{n^2}\right), \tag{10}$$

$$\|\mathbf{g}_{\hat{\boldsymbol{\theta}}}\|_2 \;=\; \widetilde{O}\left(\sqrt{kd}\right), \qquad \|\mathbf{g}_{\boldsymbol{\theta}^\star}\|_2 \;=\; \widetilde{O}\left(\sqrt{kd}\right). \tag{11}$$

Here (9) uses Lipschitz continuity of the Hessian with constant $L_H = \widetilde{O}\,(n)$ (Lemma G.2).

Term (I). By Cauchy–Schwarz and (7), (8), (11),

$$|(I)| \;\le\; \|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^\star\|_2 \cdot \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\mathrm{op}} \cdot \|\mathbf{g}_{\hat{\boldsymbol{\theta}}}\|_2 \;=\; \widetilde{O}\left(\sqrt{\frac{d}{n}}\right) \cdot O\left(\frac{1}{n}\right) \cdot \widetilde{O}\left(\sqrt{kd}\right) \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n} \cdot \frac{d}{\sqrt{n}}\right) \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right),$$

since $n \ge d^2$ implies $d/\sqrt{n} \le 1$.

Term (II). By (10), (11), and $\|\boldsymbol{\theta}^\star\|_2 = \Theta(1)$,

$$|(II)| \;\le\; \|\boldsymbol{\theta}^\star\|_2 \cdot \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} - \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\mathrm{op}} \cdot \|\mathbf{g}_{\hat{\boldsymbol{\theta}}}\|_2 \;=\; \widetilde{O}\left(\frac{\sqrt{nd}}{n^2} \cdot \sqrt{kd}\right) \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n} \cdot \frac{d}{\sqrt{n}}\right) \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Bounding term (III). Recall

$$\text{(III)} \;=\; \langle \boldsymbol{\theta}^\star,\; \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\left(\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\boldsymbol{\theta}^\star}\right)\rangle.$$

We first handle the regime $k > n/d > d$ by inserting $\mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}$ and splitting into three pieces:

$$\begin{aligned}\text{(III)} = \underbrace{\langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\big(\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}\big)\rangle}_{\text{(III.a)}} + \underbrace{\langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}\rangle - \langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}\rangle}_{\text{(III.b)}} \\ + \underbrace{\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}\rangle - \langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{g}_{\boldsymbol{\theta}^\star}\rangle}_{\text{(III.c)}}.\end{aligned}$$

(III.a). As shown above,

$$|\text{(III.a)}| \;\le\; \|\hat{\boldsymbol{\theta}}\|_2 \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\mathrm{op}} \left\|\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\mathrm{NS}(\boldsymbol{\theta}^\star)}\right\|_2 \;\le\; O(1) \cdot O\big(\tfrac{1}{n}\big) \cdot (k+d) \left\|\hat{\boldsymbol{\theta}} - \mathrm{NS}(\boldsymbol{\theta}^\star)\right\|_2.$$

By Claim D.3, $\|\hat{\boldsymbol{\theta}} - \mathrm{NS}(\boldsymbol{\theta}^\star)\|_2 \le \widetilde{O}\left(\frac{d}{n}\right)$, hence

$$|\text{(III.a)}| \;\le\; \widetilde{O}\left(\tfrac{(k+d)\,d}{n^2}\right) \;\le\; \widetilde{O}\left(\tfrac{\sqrt{k}}{n}\right),$$

using $n \ge d^2$ and $k \le n/\mathrm{polylog}(n)$.

(III.b)–(III.c): integral representation. Using the closed form $\mathrm{NS}(\boldsymbol{\theta}^\star) = \boldsymbol{\theta}^\star + \mathbf{H}_{\mathrm{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\mathrm{all}}(\boldsymbol{\theta}^\star)$, the fundamental theorem of calculus along the segment $\boldsymbol{\theta}(t) = \boldsymbol{\theta}^\star + t\big(\mathrm{NS}(\boldsymbol{\theta}^\star) - \boldsymbol{\theta}^\star\big)$, $t \in [0,1]$, yields

$$\mathbf{g}_T(\boldsymbol{\theta}^\star) - \mathbf{g}_T(\mathrm{NS}(\boldsymbol{\theta}^\star)) = \int_0^1 \mathbf{H}_T(\boldsymbol{\theta}(t))\,\mathrm{d}\boldsymbol{\theta}(t) = \int_0^1 \mathbf{H}_T(\boldsymbol{\theta}(t))\,\mathbf{H}_{\mathrm{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\mathrm{all}}(\boldsymbol{\theta}^\star)\,\mathrm{d}t.$$

Split the integrand as

$$\mathbf{H}_T(\boldsymbol{\theta}(t)) = \underbrace{\big(\mathbf{H}_T(\boldsymbol{\theta}(t)) - \mathbf{H}_T(\boldsymbol{\theta}^\star)\big)}_{\text{(b-part)}} + \underbrace{\mathbf{H}_T(\boldsymbol{\theta}^\star)}_{\text{(c-part)}},$$

and map these to (III.b) and (III.c) respectively:

$$\text{(III.b)} \;=\; \langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \int_0^1 \big(\mathbf{H}_T(\boldsymbol{\theta}(t)) - \mathbf{H}_T(\boldsymbol{\theta}^\star)\big)\,\mathbf{H}_{\mathrm{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\mathrm{all}}(\boldsymbol{\theta}^\star)\,\mathrm{d}t\rangle,$$

$$\text{(III.c)} \;=\; \langle \hat{\boldsymbol{\theta}},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \int_0^1 \mathbf{H}_T(\boldsymbol{\theta}^\star)\,\mathbf{H}_{\mathrm{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\mathrm{all}}(\boldsymbol{\theta}^\star)\,\mathrm{d}t\rangle - \langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\mathbf{g}_T(\boldsymbol{\theta}^\star)\rangle.$$

Bound for (III.b). By Cauchy–Schwarz and submultiplicativity,

$$|\text{(III.b)}| \;\le\; \|\hat{\boldsymbol{\theta}}\|_2 \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} \int_0^1 \|\mathbf{H}_T(\boldsymbol{\theta}(t)) - \mathbf{H}_T(\boldsymbol{\theta}^\star)\|_{\text{op}} \left\|\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\right\|_{\text{op}} \|\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\|_2 \, \mathrm{d}t.$$

Along the path, $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}^\star\|_2 \le \|\text{NS}(\boldsymbol{\theta}^\star) - \boldsymbol{\theta}^\star\|_2$. Using Lemma G.2 and the under-parametrized assumption:

$$\|\mathbf{H}_T(\boldsymbol{\theta}) - \mathbf{H}_T(\boldsymbol{\theta}^\star)\|_{\text{op}} \;\le\; \widetilde{O}\left(k \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_2\right) + \widetilde{O}\left(\tfrac{d^2}{\sqrt{n}}\right) \;\le\; \widetilde{O}\left(k\sqrt{\tfrac{d}{n}}\right) + \widetilde{O}\left(\tfrac{d^2}{\sqrt{n}}\right) \;=\; \widetilde{O}\left(\sqrt{nd}\right),$$

since $n > d^2$, $k \le n/\text{polylog}(n)$) Moreover, $\left\|\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\right\|_{\text{op}} = O(\frac{1}{n})$ and $\|\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\|_2 = \widetilde{O}\left(\sqrt{nd}\right)$. With $\|\hat{\boldsymbol{\theta}}\|_2 = O(1)$, $\left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} = O(\frac{1}{n})$, we obtain

$$|\text{(III.b)}| \;\le\; O(1) \cdot O\left(\tfrac{1}{n}\right) \cdot \widetilde{O}\left(\sqrt{nd}\right) \cdot O\left(\tfrac{1}{n}\right) \cdot \widetilde{O}\left(\sqrt{nd}\right) \;=\; \widetilde{O}\left(\tfrac{nd}{n^2}\right) \;=\; \widetilde{O}\left(\tfrac{d}{n}\right) \;\le\; \widetilde{O}\left(\tfrac{\sqrt{k}}{n}\right),$$

since $k > d$

Bound for (III.c). Write

$$\text{(III.c)} = \langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{H}_T(\boldsymbol{\theta}^\star)\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\rangle + \langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star,\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{H}_T(\boldsymbol{\theta}^\star)\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\rangle.$$

For the leading part,

$$\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{H}_T(\boldsymbol{\theta}^\star)\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\rangle = \sum_{i\in T} \beta_i \,\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{x}_i\rangle \,\langle \mathbf{x}_i,\, \mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\rangle,$$

for curvature weights $\beta_i \in (0, \beta_{\max}]$. By independence and subgaussianity, the vector $\sum_{i\in T} \beta_i \,\langle \boldsymbol{\theta}^\star, \mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{x}_i\rangle \,\mathbf{x}_i$ has mean $\frac{k}{n}\,\boldsymbol{\theta}^\star$ and fluctuation of size $\widetilde{O}\left(\frac{\sqrt{kd}}{n}\right)$. Contracting with $\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)$, whose norm is $\widetilde{O}\left(\sqrt{d/n}\right)$, yields

$$\begin{aligned}|\text{(III.c)}| \;\le\; &\underbrace{\tfrac{k}{n}\,\|\boldsymbol{\theta}^\star\|_2 \cdot \widetilde{O}\left(\sqrt{\tfrac{d}{n}}\right)}_{\text{mean term}} + \underbrace{\widetilde{O}\left(\tfrac{\sqrt{kd}}{n}\right) \cdot \widetilde{O}\left(\sqrt{\tfrac{d}{n}}\right)}_{\text{variance term}} + \\ &+ \underbrace{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 \cdot \left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} \cdot \|\mathbf{H}_T(\boldsymbol{\theta}^\star)\|_{\text{op}} \cdot \left\|\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\right\|_{\text{op}} \cdot \|\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\|_2}_{\text{residual}}\end{aligned}$$

Each piece is $\widetilde{O}\left(\frac{\sqrt{k}}{n}\right)$: the mean term uses Lemma D.2 to bound the contraction with $\widetilde{O}\left(1/\sqrt{n}\right)$, the variance term gives $\widetilde{O}\left(\frac{\sqrt{k}}{n} \cdot \frac{d}{\sqrt{n}}\right) \le \widetilde{O}\left(\frac{\sqrt{k}}{n}\right)$ since $n > d^2$, and the residual uses $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2 = \widetilde{O}\left(\sqrt{d/n}\right)$, $\left\|\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\right\|_{\text{op}} = O(1/n)$, $\|\mathbf{H}_T(\boldsymbol{\theta}^\star)\|_{\text{op}} = \widetilde{O}\left(k\right)$, $\left\|\mathbf{H}_{\text{all}}(\boldsymbol{\theta}^\star)^{-1}\right\|_{\text{op}} = O(1/n)$, $\|\mathbf{g}_{\text{all}}(\boldsymbol{\theta}^\star)\|_2 = \widetilde{O}\left(\sqrt{nd}\right)$, to give $\widetilde{O}\left(\sqrt{d/n}\right) \cdot \frac{1}{n} \cdot k \cdot \frac{1}{n} \cdot \sqrt{nd} = \widetilde{O}\left(\frac{k\sqrt{d}}{n^{3/2}}\right) \le \widetilde{O}\left(\frac{\sqrt{k}}{n}\right)$ since $k > n/d$.

Conclusion for $k > n/d > d$. Combining the bounds for (III.a), (III.b), and (III.c),

$$|\text{(III)}| \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Case $k \le n/d$. This regime is simpler; we bound (III) directly via Cauchy–Schwarz:

$$|\text{(III)}| = \left|\langle \boldsymbol{\theta}^\star,\, \mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\,(\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\boldsymbol{\theta}^\star})\rangle\right| \;\le\; \|\boldsymbol{\theta}^\star\|_2 \left\|\mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\right\|_{\text{op}} \|\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\boldsymbol{\theta}^\star}\|_2.$$

The gradient difference over the k affected samples satisfies $\|\mathbf{g}_{\hat{\boldsymbol{\theta}}} - \mathbf{g}_{\boldsymbol{\theta}^\star}\|_2 \le \widetilde{O}\left(k\sqrt{d/n}\right)$, hence

$$|\text{(III)}| \;\le\; O(1) \cdot O\left(\tfrac{1}{n}\right) \cdot \widetilde{O}\left(k\sqrt{\tfrac{d}{n}}\right) \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n} \cdot \sqrt{\tfrac{kd}{n}}\right) \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right),$$

since $k \le n/d$. □

D.1.3 FINAL CLAIM (NEWTON STEP GEOMETRY AT $\hat{\boldsymbol{\theta}}$).

Let $\mathbf{H} \overset{\text{def}}{=} \mathbf{H}_{\hat{\boldsymbol{\theta}}}$ and $\mathbf{g} \overset{\text{def}}{=} \mathbf{g}_{\hat{\boldsymbol{\theta}}}$. By definition of the Newton step at $\hat{\boldsymbol{\theta}}$ (with subset T),

$$\hat{\boldsymbol{\theta}}_T^{\text{NS}} \;=\; \hat{\boldsymbol{\theta}} - \mathbf{H}^{-1}\mathbf{g} \quad\Longrightarrow\quad \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\text{NS}} \;=\; \mathbf{H}^{-1}\mathbf{g}.$$

Therefore

$$\langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\text{NS}},\, \mathbf{H}^{-1}\mathbf{g}\rangle \;=\; \langle \mathbf{H}^{-1}\mathbf{g},\, \mathbf{H}^{-1}\mathbf{g}\rangle \;=\; \|\mathbf{H}^{-1}\mathbf{g}\|_2^2 \;\le\; \left\|\mathbf{H}^{-1}\right\|_{\text{op}}^2 \|\mathbf{g}\|_2^2. \tag{12}$$

By strong convexity at $\hat{\boldsymbol{\theta}}$, $\left\|\mathbf{H}^{-1}\right\|_{\text{op}} = O(1/n)$; by our gradient bounds at $\hat{\boldsymbol{\theta}}$, $\|\mathbf{g}\|_2 = \widetilde{O}\left(\sqrt{kd}\right)$. Thus

$$\|\mathbf{H}^{-1}\mathbf{g}\|_2^2 \;\le\; \widetilde{O}\left(\frac{kd}{n^2}\right) \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right), \tag{13}$$

since $\frac{kd}{n^2} = \frac{\sqrt{k}}{n} \cdot \frac{\sqrt{k}\,d}{n}$ and $\frac{\sqrt{k}\,d}{n} \le 1/\sqrt{\text{polylog}(n)}$ under $n \ge d^2$ and $k \le n/\text{polylog}(n)$. □

D.1.4 EXTEND TO ALL $\boldsymbol{\theta}$ ON THE SEGMENT.

For any $s \in [0,1]$, define the point on the segment $\boldsymbol{\theta}(s) \overset{\text{def}}{=} \hat{\boldsymbol{\theta}} + s\big(\hat{\boldsymbol{\theta}}_T^{\text{NS}} - \hat{\boldsymbol{\theta}}\big)$. Then

$$\langle \boldsymbol{\theta}(s),\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}\rangle = \langle \hat{\boldsymbol{\theta}},\, \mathbf{H}^{-1}\mathbf{g}\rangle \;+\; s\,\langle \hat{\boldsymbol{\theta}}_T^{\text{NS}} - \hat{\boldsymbol{\theta}},\, \mathbf{H}^{-1}\mathbf{g}\rangle.$$

Taking absolute values and using the triangle inequality together with Claim D.4 (which gave $\left|\langle \hat{\boldsymbol{\theta}},\, \mathbf{H}^{-1}\mathbf{g}\rangle\right| = \widetilde{O}\left(\frac{\sqrt{k}}{n}\right)$) and (12)–(13),

$$\left|\langle \boldsymbol{\theta}(s),\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}\rangle\right| \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right) \;+\; s\,\|\mathbf{H}^{-1}\mathbf{g}\|_2^2 \;\le\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right) \;+\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right) \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

Because this holds for every $s \in [0,1]$, we conclude

$$\sup_{\boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}},\, \hat{\boldsymbol{\theta}}_T^{\text{NS}}]} \left|\langle \boldsymbol{\theta},\, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}\rangle\right| \;=\; \widetilde{O}\left(\frac{\sqrt{k}}{n}\right).$$

This completes the proof of Lemma D.1.

D.2 CONCLUDING THE UPPER BOUND IN THEOREM C.3

From Lemma G.11, we know that with high probability, uniformly for all $\boldsymbol{\theta}$ of length at most 10, it holds that the third order derivative of the loss concentrates. In particular, recall that the third order derivative of a logistic regression loss is given by

$$\mathbf{T}_{\boldsymbol{\theta}} \overset{\text{def}}{=} \nabla^{\otimes 3} L = \sum_{i \in [n]} \gamma_i(\langle \boldsymbol{\theta}, \mathbf{x}_i\rangle \mathbf{x}_i^{\otimes 3}$$

From Lemma G.11, with high probability

$$\forall \boldsymbol{\theta} \text{ s.t. } \|\boldsymbol{\theta}\|_2 < 10, \quad \left\|\mathbf{T}_{\boldsymbol{\theta}} - \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{y} \sim \mathcal{X} \times \mathcal{Y}}[\mathbf{T}_{\boldsymbol{\theta}}]\right\|_{\text{op}} = \widetilde{O}\left(d^{3/2} + \sqrt{nd}\right).$$

Moreover, from our assumption that the features are normally distributed, it is easy to see that

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{y} \sim \mathcal{X} \times \mathcal{Y}}[\mathbf{T}_{\boldsymbol{\theta}}] = \Theta(n) \cdot \boldsymbol{\theta}^{\otimes 3}.$$

Therefore, plugging the Newton step from $\hat{\boldsymbol{\theta}}$ into two of the inputs of $\mathbf{T}_{\boldsymbol{\theta}}$ yields a vector whose L_2 norm is close to that their inner product with $\boldsymbol{\theta}$ squared

$$\left\|\mathbf{T}_{\boldsymbol{\theta}}\left(\mathbf{H}^{-1}\mathbf{g}, \mathbf{H}^{-1}\mathbf{g}\right)\right\|_2 \le \underbrace{\left\|\mathbb{E}\left[\mathbf{T}_{\boldsymbol{\theta}}\right]\left(\mathbf{H}^{-1}\mathbf{g}, \mathbf{H}^{-1}\mathbf{g}\right)\right\|_2}_{=\Theta(n\langle \boldsymbol{\theta}, \mathbf{H}^{-1}\mathbf{g}\rangle^2)} + \underbrace{\left\|\mathbf{H}^{-1}\mathbf{g}\right\|_2^2 \left\|\mathbf{T}_{\boldsymbol{\theta}} - \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{y} \sim \mathcal{X} \times \mathcal{Y}}[\mathbf{T}_{\boldsymbol{\theta}}]\right\|_{\text{op}}}_{=\widetilde{O}\left(d^{3/2}\frac{kd}{n^2} + \frac{kd^{3/2}}{n^{3/2}}\right) = \widetilde{O}\left(\frac{k}{n}\right)}$$

To conclude the upper bound in Theorem C.3, we simply note that for any $\boldsymbol{\theta} \in [\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}]$, combining the fundamental theorem of calculus and the triangle inequality yields

$$\left\| \left(\mathbf{H}_{\boldsymbol{\theta}} - \mathbf{H}_{\hat{\boldsymbol{\theta}}} \right) \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathbf{g}_{\hat{\boldsymbol{\theta}}} \right\|_2 = \left\| \int_0^t \mathbf{T}_{\hat{\boldsymbol{\theta}} + \tau(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}})} (\mathbf{H}^{-1}\mathbf{g}, \mathbf{H}^{-1}\mathbf{g}) \mathrm{d}\tau \right\|_2 \leq \widetilde{O}\left(\frac{k}{n}\right).$$

Therefore, the assumptions to Theorem 1.5 hold with $C_h = \widetilde{O}\left(\frac{k}{n}\right)$, $C_{\mathrm{op}} = O\left(\frac{1}{n}\right)$, $\boldsymbol{\Sigma} = \mathbf{I}$ and $r = 1$, yielding the desired scaling.

D.3 Lower Bound

Finally, to prove Theorem C.3, we need to lower-bound the expected error of the NS estimate over random choices of the drop-set T.

Lemma D.5. *In the setting of Theorem C.3, with high probability over the training set,*

$$\mathop{\mathbb{E}}_{T \in \binom{[n]}{k}} \left[\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \right] = \Omega\left(\frac{k}{n^2}\right).$$

D.3.1 Proof Sketch

The key idea is that the error of Newton step (NS) attribution is governed by the second moment of the dropped gradient.

For a given drop set $T \subseteq [n]$, the discrepancy between the retrained model $\hat{\boldsymbol{\theta}}_T$ and its NS approximation $\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}$ is controlled by the local curvature:

$$\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \gtrsim \frac{1}{\|\mathbf{H}_{\boldsymbol{\theta}}\|_{\mathrm{op}}} \left\| \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \right\|_2 = \Theta\left(\frac{1}{n}\right) \left\| \mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} \right\|_2.$$

Expanding $\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}}$ via Taylor's theorem around $\hat{\boldsymbol{\theta}}$ yields a third-order remainder term:

$$\mathbf{g}_{\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}} = \int_0^1 \left(\mathbf{H}_{\hat{\boldsymbol{\theta}} + t(\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} - \hat{\boldsymbol{\theta}})} - \mathbf{H}_{\hat{\boldsymbol{\theta}}} \right) (\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} - \hat{\boldsymbol{\theta}}) \, \mathrm{d}t \approx \mathbf{T}_{\hat{\boldsymbol{\theta}}}(\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}),$$

where $\mathbf{T}_{\hat{\boldsymbol{\theta}}}$ is the third-order derivative tensor of the loss at $\hat{\boldsymbol{\theta}}$.

From Lemma G.11, $\mathbf{T}_{\hat{\boldsymbol{\theta}}}$ concentrates around its expectation. For Gaussian covariates, this expectation has the simple form

$$\mathbf{T}_{\hat{\boldsymbol{\theta}}} \simeq \Theta(n) \, \hat{\boldsymbol{\theta}}^{\otimes 3} \quad \text{whenever} \quad \left\| \hat{\boldsymbol{\theta}} \right\|_2 = O(1).$$

Therefore, along the short NS trajectory from $\hat{\boldsymbol{\theta}}$, the dominant contribution is

$$\mathbf{T}_{\hat{\boldsymbol{\theta}}}(\mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}}) \approx \hat{\boldsymbol{\theta}} \, \langle \hat{\boldsymbol{\theta}}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle^2.$$

Taking expectations over random drop sets T then reduces to analyzing the second moment of the dropped gradient:

$$\mathop{\mathbb{E}}_{T} \left[\langle \hat{\boldsymbol{\theta}}, \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1}\mathbf{g}_{\hat{\boldsymbol{\theta}}} \rangle^2 \right] = \hat{\boldsymbol{\theta}}^\top \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \mathop{\mathbb{E}}_{T} \left[\sum_{i \in T} \alpha_i \mathbf{x}_i \mathbf{x}_i^\top \right] \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \hat{\boldsymbol{\theta}}.$$

Since each term in the expectation contributes roughly k/n in variance, and $\left\| \mathbf{H}_{\hat{\boldsymbol{\theta}}}^{-1} \right\|_{\mathrm{op}} = \Theta(1/n)$, this evaluates to

$$\mathop{\mathbb{E}}_{T} \left[\left\| \hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\mathrm{NS}} \right\|_2 \right] = \Omega\left(\frac{k}{n^2}\right),$$

which matches the claimed lower bound.

E Asymptotic Analysis of Influence Functions

In this section we will analyze the asymptotic behavior of the difference between the IF and NS estimates, proving Theorem C.2.

E.1 Upper Bound

We begin by proving the upper bound that with high probability

$$\max_{T\in\binom{[n]}{k}}\left\{\left\|\hat{\boldsymbol{\theta}}_T^{\mathrm{IF}}-\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}\right\|\right\}=\widetilde{O}\left(\frac{(k+d)\sqrt{kd}}{n}\right)$$

Proof. This will follow from applying the CS inequality to the concentration bounds proven in Appendix G.

In particular, we note that

$$\hat{\boldsymbol{\theta}}_T^{\mathrm{IF}}-\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}=\left(\mathbf{H}^{\mathrm{all}}\right)^{-1}\left(\mathbf{H}^{T}\right)\left(\mathbf{H}^{\setminus T}\right)^{-1}\mathbf{g}^{T}$$

Combining Lemmas G.10, G.5 and Vershynin (2010), Theorem 5.39, we have

$$\begin{aligned}\left\|\hat{\boldsymbol{\theta}}_T^{\mathrm{IF}}-\hat{\boldsymbol{\theta}}_T^{\mathrm{NS}}\right\|_2 &\le \left\|\left(\mathbf{H}^{\mathrm{all}}\right)^{-1}\right\|_{\mathrm{op}}\left\|\left(\mathbf{H}^{T}\right)\right\|_{\mathrm{op}}\left\|\left(\mathbf{H}^{\setminus T}\right)^{-1}\right\|_{\mathrm{op}}\left\|\mathbf{g}^{T}\right\|_2 = \\ &= \widetilde{O}\left(\frac{1}{n}\times(k+d)\times\frac{1}{n}\times\sqrt{kd}\right)=\widetilde{O}\left(\frac{(k+d)\sqrt{kd}}{n^2}\right),\end{aligned}$$

as desired.

□

E.2 Lower Bounds

We now proceed to prove the lower bound portion of Theorem C.2.

Here, our goal is to show that when dropping a random subset of the train set, the IF estimate is at least $\widetilde{\Omega}\left(\frac{(k+d)\sqrt{kd}}{n^2}\right)$ from the NS estimate.

We analyze this norm using the identity that

$$\boldsymbol{\theta}^{\mathrm{IF}}-\boldsymbol{\theta}^{\mathrm{NS}}=\mathbf{H}_{\setminus T}^{-1}\mathbf{H}_T\mathbf{H}^{-1}\mathbf{g}_T\,.$$

From Lemmas G.11 and G.5, we know that with high probability the third order derivative of the loss is globally bounded and the learned model $\hat{\boldsymbol{\theta}}$ is close to population optimum $\boldsymbol{\theta}^\star$. Combined these yield a bound on the difference between $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ and $\mathbf{H}_{\boldsymbol{\theta}^\star}$. Moreover, from Lemma G.7, we know that with high probability, the Hessian converges uniformly, so that

$$\mathbf{H}=\mathbf{H}_{\hat{\boldsymbol{\theta}}}\approx\mathbf{H}_{\boldsymbol{\theta}^\star}\approx\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]=\Theta(n)\mathbf{I}\,,$$

where each of these approximations yields an error that has operator norm at most $\widetilde{O}\left(\sqrt{nd}\right)$.

Therefore, $\left\|\mathbf{H}^{-1}-\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\right\|_{\mathrm{op}}=\widetilde{O}\left(\frac{\sqrt{nd}}{n^2}\right)$.

From Lemma G.6, we know that with high probability the Hessian on the retained samples $\mathbf{H}_{\setminus T}$ has spectrum $\Theta(n)$, yielding the scaling

$$\left\|\boldsymbol{\theta}^{\mathrm{IF}}-\boldsymbol{\theta}^{\mathrm{NS}}\right\|_2=\Theta\left(\frac{1}{n}\right)\times\left(\left\|\mathbf{H}_T\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{g}_T\right\|_2\pm\widetilde{O}\left(\frac{(k+d)\sqrt{kd}}{n}\times\sqrt{\frac{d}{n}}\right)\right)$$

Therefore, all that is left is to analyze the expectation of $\left\|\mathbf{H}_T\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{g}_T\right\|_2$ over drop-sets T. We do this in 3 regimes:

E.2.1 $k \gg d$

In this regime, Lemma G.7 suffices to guarantee that with high probability $\sigma\left(\mathbf{H}_T\right) = \Theta(k)$, so

$$\left\|\mathbf{H}_T\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{g}_T\right\|_2 = \Theta(k) \times \left\|\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{g}_T\right\|_2 = \Theta\left(\frac{k}{n}\right) \times \|\mathbf{g}_T\| = \widetilde{\Theta}\left(\frac{k\sqrt{kd}}{n}\right),$$

where the last step utilized Lemma G.10.

E.2.2 $k \ll d$

In this case, we have

$$\mathbf{H}_T\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{g}_T = \sum_{i,j\in T}\beta_i\mathbf{x}_i\mathbf{x}_i^\intercal\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{x}_j\alpha_j$$

Breaking this summation up into the $i = j$ and the $i \neq j$ contributions, we have

$$\sum_{i\in T}\beta_i\mathbf{x}_i\mathbf{x}_i^\intercal\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{x}_i\alpha_i = \sum_{i\in T}\mathbf{x}_i \times \Omega\left(\frac{d}{n}\right),$$

so the sum over these vectors has norm

$$\left\|\sum_{i\in T}\beta_i\mathbf{x}_i\mathbf{x}_i^\intercal\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{x}_i\alpha_i\right\|_2 = \Omega\left(\frac{d\sqrt{kd}}{n}\right).$$

To conclude our statement for this case, we need to show that the contribution of the cross terms ($i \neq j$) cannot cancel out this main term.

Fix some index i, and consider

$$\sum_{j\neq i}\mathbf{x}_i\mathbf{x}_i^\intercal\mathbb{E}\left[\mathbf{H}_{\hat{\boldsymbol{\theta}}}\right]^{-1}\mathbf{x}_j = \left(\sum_{j\neq i}G_{i,j}\right)\mathbf{x}_i\,.$$

Clearly the $G_{i,j} = \widetilde{O}\left(\frac{\sqrt{d}}{n}\right)$ elements are bounded with high probability over the training set and their expectation is small (simply from our assumption that $\mathbf{x}_i$ are drawn from a centered distribution), so standard concentration bounds suffice to show that with high probability

$$\left|\sum_{j\neq i}G_{i,j}\right| = \widetilde{O}\left(\frac{\sqrt{kd}}{n}\right) \ll \frac{d}{n}$$

for all $i \in T$.

E.2.3 $k \approx d$

In this regime we have a hard time applying the arguments for either the k or the d scalings above, simply because it is difficult to rule out the possibility of the two contributions canceling out.

Therefore, we simply note that setting $\mathbf{\Pi}$ to be a random projection to a subspace of dimension $\dim(V) = \frac{d}{\text{polylog}(d)}$, the norm of $\mathbf{\Pi}\mathbf{H}_T\mathbf{H}^{-1}\mathbf{g}_T$ clealy lower-bounds the norm of $\mathbf{H}_T\mathbf{H}^{-1}\mathbf{g}_T$, and by applying our proof from the $k \gg d$ regime, we can show that with high probability $\left\|\mathbf{\Pi}\mathbf{H}_T\mathbf{H}^{-1}\mathbf{g}_T\right\|_2 = \widetilde{\Omega}\left(\frac{k\sqrt{kd}}{n}\right)$, also yielding the desired scaling.

This concludes the proof of Theorem C.2.

F Asymptotic Analysis of Previous Results

In this section, we will prove Theorem C.1. In particular, we will show that with high probability over the training set, for all $T \in \binom{[n]}{k}$, Theorem 1.4 yields a bound that scales like

$$\text{Existing Bounds} = \widetilde{\Theta}\left(\frac{k^2 d}{n^2 \lambda^3}\right)$$

To prove this, we show that

Lemma F.1. *In the setting of Theorem C.1, with high probability over the training set, we have*

- $C_{\text{Lip}} = \Theta(n)$
- $C_{\text{op}} = \Theta\left(\frac{1}{\lambda n}\right)$
- $C_\ell = \Theta\left(\sqrt{d}\right)$

Since the bound in Theorem 1.4 scales like

$$\left\|\hat{\boldsymbol{\theta}}_T - \hat{\boldsymbol{\theta}}_T^{\text{NS}}\right\|_2 = O(C_{\text{Lip}} C_{\text{op}}^3 k^2 C_\ell^2)\,,$$

Theorem C.1 follows immediately from Lemma F.1.

Proof of Lemma F.1. C_{Lip} measures the Lipschitzness of the Hessian. By definition, which is in turn given by the third derivative of the loss $\mathbf{T}$. Lemma G.11 tells us that with high probability over this training set, this third moment converges to its expectation uniformly, and from our assumption that the features are normally distributed, we have

$$\mathbb{E}\left[\mathbf{T}_{\boldsymbol{\theta}}\right] \simeq n\boldsymbol{\theta}^{\otimes 3}\,.$$

Therefore, with high probability $C_{\text{Lip}} = \Theta(n)$ regardless of the choice of T.

Because our optimization domain $\Omega_\theta = \mathbb{R}^d$ contains limits where the Hessian of the unregularized logistic loss decays to 0, the spectrum of the Hessian is globally lower-bounded only by the regularization term, yielding the scaling

$$C_{\text{op}} = \Theta\left(\frac{1}{\lambda n}\right)$$

Finally, C_ℓ does not depend on the set of samples being removed and clearly concentrates around

$$C_\ell = \Theta\left(\sqrt{d}\right)$$

□

G Useful Concentration Bounds

G.1 Concentration of Norms of Subgaussian Variables

Lemma G.1 ((Jin et al., 2019, Lemma 1))**.** *Let $X \in \mathbb{R}^d$ be a random vector that is $\frac{\sigma}{\sqrt{d}}$-subGaussian in the usual sense. Then X is norm-subGaussian with parameter $c\sigma$, i.e. $X \sim nSG(c\sigma)$ for an absolute constant c.*

G.2 Concentration of Higher Order Moments of Subgaussian Random Variables

We prove a bound on the operator norm of the sum of fourth moments of independent subgaussian random variables.

Lemma G.2 (Higher Order Empirical Moments of a Sub-Gaussian)**.** *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{X}$ be iid samples drawn from a subgaussian distribution on $\mathbb{R}^d$ with mean $\mathbf{0}$ and bounded covariance $\|\mathbf{\Sigma}\|_{\text{op}} = O(1)$. Then, for any fixed power $t \geq 2$, with probability $1 - e^{-\Omega(d)}$*

$$\max_{\mathbf{e} \in \mathbb{S}^{d-1}} \left\{ \sum_{i=1}^{n} \left| \langle \mathbf{x}_i, \mathbf{e} \rangle^t \right| \right\} = O\left(n + d^{t/2}\right) .$$

This result is a key technical component in our analysis of the Newton step data attribution method.

Theorem G.3. *Let $x_1, \ldots, x_n$ be independent samples from a subgaussian distribution on $\mathbb{R}^d$ with mean zero and covariance $\mathbf{\Sigma}$ such that $\|\mathbf{\Sigma}\|_{\text{op}} \leq 1$. Let $n \geq d(\log d)^{O(1)}$. Then with probability at least $1 - n^{-\omega(1)}$,*

$$\left\| \sum_{i=1}^{n} (x_i \otimes x_i)(x_i \otimes x_i)^\top \right\|_{\text{op}} = O(nd) .$$

To prove the theorem, we first seek to bound the operator norm of the expectation of a single term in the sum. Let X be a random vector drawn from the same distribution as the x_i. We are interested in $\left\|\mathbb{E}(X \otimes X)(X \otimes X)^\top\right\|_{\text{op}}$. This is equivalent to finding the maximum of $\mathbb{E}\langle u, X \otimes X \rangle^2$ over all unit vectors $u \in \mathbb{R}^{d^2}$.

Lemma G.4. *Let X be a subgaussian random vector in $\mathbb{R}^d$ with $\mathbb{E}\, X = 0$ and $\mathbb{E}\, XX^\top = \mathbf{\Sigma}$ with $\|\mathbf{\Sigma}\|_{\text{op}} \leq 1$. Then*

$$\left\|\mathbb{E}(X \otimes X)(X \otimes X)^\top\right\|_{\text{op}} = O(d) .$$

Proof. Let $u \in \mathbb{R}^{d^2}$ be a unit vector. We can view u as a $d \times d$ matrix U such that $\|U\|_F^2 = \sum_{i,j} U_{ij}^2 = 1$. The expression $\langle u, X \otimes X \rangle$ is equivalent to the quadratic form $X^\top U X$. Let the singular value decomposition of U be $U = \sum_{a=1}^{d} \sigma_a u_a v_a^\top$, where σ_a are the singular values and u_a, v_a are the left and right singular vectors, respectively. Since $\|U\|_F^2 = 1$, we have $\sum_{a=1}^{d} \sigma_a^2 = 1$. The quadratic form can now be written as:

$$X^\top U X = \sum_{a=1}^{d} \sigma_a (X^\top u_a)(v_a^\top X) .$$

We want to bound $\mathbb{E}[(X^\top U X)^2]$:

$$\begin{aligned}
\mathbb{E}[(X^\top U X)^2] &= \mathbb{E}\left[\left(\sum_{a=1}^{d} \sigma_a (X^\top u_a)(v_a^\top X) \right)^2 \right] \\
&= \sum_{a,b=1}^{d} \sigma_a \sigma_b \, \mathbb{E}[(X^\top u_a)(v_a^\top X)(X^\top u_b)(v_b^\top X)] \\
&\leq \sum_{a,b=1}^{d} \sigma_a \sigma_b (\mathbb{E}(X^\top u_a)^4)^{1/4} (\mathbb{E}(X^\top v_a)^4)^{1/4} (\mathbb{E}(X \top u_b)^4)^{1/4} (\mathbb{E}(X \top v_b)^4)^{1/4} \\
&\leq O(1) \cdot \sum_{a,b=1}^{d} \sigma_a \sigma_b \\
&\leq O(d) \cdot \sum_{a=1}^{d} \sigma_a^2 \\
&= O(d) .
\end{aligned}$$

□

Now we can prove Theorem G.3.

Proof. Given Lemma G.4, we can prove Theorem G.3 by applying the matrix Bernstein inequality. We bound the deviation of the sum from its expectation. Let $Z_i = (x_i \otimes x_i)(x_i \otimes x_i)^\top$. We want to bound $\|\sum_i (Z_i - \mathbb{E}\, Z_i)\|_{\text{op}}$. We will use the matrix Bernstein inequality. A key challenge is that the Z_i are not bounded. We address this by truncation.

Let $C = O(\sqrt{d}\log n)$. For a subgaussian vector x_i, we have $\Pr(\|x_i\|_2 > t\sqrt{d}) \le \exp(-\Omega(t^2))$ Jin et al. (2019). Let $\mathcal{E}_i$ be the event that $\|x_i\|_2 \le C$. By a union bound, $\Pr(\forall i, \mathcal{E}_i) \ge 1 - n \cdot n^{-\omega(1)} = 1 - n^{-\omega(1)}$. Let $\tilde{x}_i = x_i \cdot \mathbb{I}(\mathcal{E}_i)$ and $\tilde{Z}_i = (\tilde{x}_i \otimes \tilde{x}_i)(\tilde{x}_i \otimes \tilde{x}_i)^\top$. With high probability, $\sum_i Z_i = \sum_i \tilde{Z}_i$. It suffices to bound $\left\|\sum_i (\tilde{Z}_i - \mathbb{E}\, \tilde{Z}_i)\right\|_{\text{op}}$.

The truncated variables $\tilde{Z}_i$ are bounded: $\left\|\tilde{Z}_i\right\|_{\text{op}} \le \|\tilde{x}_i\|_2^4 \le C^4 = O(d^2 \log^4 n)$. This is our parameter R in the matrix Bernstein inequality.

Next, we need to bound the variance parameter $\sigma^2 = \left\|\sum_i \mathbb{E}(\tilde{Z}_i - \mathbb{E}\, \tilde{Z}_i)^2\right\|_{\text{op}} \le \left\|\sum_i \mathbb{E}\, \tilde{Z}_i^2\right\|_{\text{op}}$.

$$\mathbb{E}\, \tilde{Z}_i^2 = \mathbb{E}[(x_i \otimes x_i)(x_i \otimes x_i)^\top (x_i \otimes x_i)(x_i \otimes x_i)^\top \cdot \mathbb{I}(\mathcal{E}_i)] \preceq \mathbb{E}[\|x_i\|_2^4 (x_i \otimes x_i)(x_i \otimes x_i)^\top].$$

We need to bound the operator norm of this matrix. As in Lemma G.4, we test it with a unit vector $u \in \mathbb{R}^{d^2}$, which we view as a matrix U with $\|U\| = 1$.

$$\mathbb{E}[\|x_i\|_2^4 (x_i^\top U x_i)^2] \le \sqrt{\mathbb{E}\, \|x_i\|_2^8\, \mathbb{E}(x_i^\top U x_i)^4}.$$

Since x_i is subgaussian, $\mathbb{E}\, \|x_i\|_2^p = O(d^{p/2})$. So $\mathbb{E}\, \|x_i\|_2^8 = O(d^4)$. Also, $\mathbb{E}(x_i^\top U x_i)^4 = O(d^2)$ by extending the logic of Lemma G.4. This gives a variance parameter for a single term of order $O(d^3)$. Summing over n terms, $\sigma^2 = O(nd^3)$.

The matrix Bernstein inequality states that for $t > 0$,

$$\Pr\left(\left\|\sum_i (\tilde{Z}_i - \mathbb{E}\, \tilde{Z}_i)\right\|_{\text{op}} \ge t\right) \le 2d^2 \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

Plugging in $R = O(d^2 \text{polylog}(n))$ and $\sigma^2 = O(nd^3 \text{polylog}(n))$, and setting $t = O(nd)$ while using $n \ge d(\log d)^{O(1)}$, we find that the deviation is bounded by t with probability $1 - n^{-\omega(1)}$. The lemma follows by combining the bound on the expectation and the deviation. □

G.3 Parameter Learning for Logistic Regression and Local Strong Convexity

Lemma G.5. *Let X be a subgaussian random vector with covariance $\mathbf{I}$ on $\mathbb{R}^d$. Let $\theta \in \mathbb{R}^d$ with $\|\theta\|_2 = 1$. Let y be a $\{\pm 1\}$ random variable with $\Pr(y = 1) = \text{softmax}(\theta^\intercal X)$. Let $\hat{\theta}$ be the maximum likelihood estimator of θ given independent draws $(X_1, y_1), \ldots, (X_n, y_n)$. Suppose $n \ge d(\log d)^{O(1)}$. Then with very high probability,*

$$\left\|\hat{\theta} - \theta\right\|_2 \le \tilde{O}\left(\sqrt{\frac{d}{n}}\right).$$

We will use this lemma to derive another useful one:

Lemma G.6. *Under the same assumptions as Lemma G.5, if $\hat{\theta}$ is the MLE for θ, then with very high probability,*

$$\nabla^2 L_n(\hat{\theta}) \succeq \Omega(n) \cdot \mathbf{I}$$

so long as $n \ge d(\log d)^{O(1)}$. Furthermore, the same is true if we consider the Hessian only on a subset of samples: there is a constant $c > 0$ such that with high probability all $S \subseteq [n]$ with $|S| \le cn/\log n$,

$$\nabla^2 L_{[n]\setminus S}(\hat{\theta}) \succeq \Omega(n) \cdot \mathbf{I}.$$

We give both proofs at the end of this section after building the requisite lemmas.

G.3.1 Uniform Convergence of the Hessian

We establish uniform convergence of the Hessian. While this uses standard techniques, we include it for completeness. To prove the lemma we start with uniform convergence of the Hessian.

Lemma G.7. *Under the assumptions of Lemma G.5, if $n \geq d(\log d)^{O(1)}$, then with very high probability for all $\alpha \in \mathbb{R}^d$ with $\|\alpha\| \leq 10$, the Hessian satisfies*

$$\left\|\nabla^2 L_n(\alpha) - \mathbb{E}\,\nabla^2 L_n(\alpha)\right\|_{op} \leq \tilde{O}(\sqrt{dn})\,.$$

Proof. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be the covariates. We aim to bound

$$\max_{\alpha \in \mathbb{R}^d} \left\|\nabla^2 L_n(\alpha) - \mathbb{E}\,\nabla^2 L_n(\alpha)\right\|_{op}\,.$$

By a symmetrization argument, losing a constant factor it will be enough to prove a very-high-probability bound on

$$\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n} \max_{\|\alpha\| \leq 10} \left\| \sum_{i \leq n} \varepsilon_i \beta_i(\alpha) x_i x_i^\top \right\|$$

where $\beta_i(\alpha)$ is the variance of the prediction of the logistic model α on x_i and $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables. This in turn is precisely

$$\mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n} \sup_{\|\alpha\| \leq 10, \|u\| \leq 1} \sum_{i \leq n} \varepsilon_i \beta_i(\alpha) \langle x_i, u\rangle^2$$

Let $S \subseteq \mathbb{R}^{d+d}$ be a δ-net of the set $\{(a,u) : \|a\| \leq 10, \|u\| \leq 1\}$; we may assume $|S| \leq \delta^{-O(d)}$. Consider a fixed $(a,u) \in S$. Each $\varepsilon_i \beta_i(a) \langle x_i, u\rangle^2$ is a mean-zero, $O(1)$-subexponential random variable with variance $O(1)$. By composition of subexponential random variables (Aleksandr Podkopaev & Alessandro Rinaldo (2019)), $\sum_{i \leq n} \varepsilon_i \beta_i(a) \langle x_i, u\rangle^2$ is $O(1)$-subexponential with variance $O(n)$. Taking the supremum over all $|S|$ such random variables, we get that with very high probability,

$$\sup_{(a,u) \in S} \left| \sum_{i \leq n} \varepsilon_i \beta_i(a) \langle x_i, u\rangle^2 \right| \leq O(\sqrt{nd \log(n/\delta)^2} + d \log(n/\delta)^2)\,.$$

Now let (a,u) be arbitrary, and let $(a_0, u_0) \in S$ such that $\delta_a = a_0 - a, \delta_u = u_0 - u$ with $\|\delta_a\|, \|\delta_u\| \leq \delta$. We have

$$\begin{aligned}
\sum_{i \leq n} \varepsilon_i \beta_i(a_0 + \delta_a) \langle x_i, u_0 + \delta_u\rangle^2 &= \sum_{i \leq n} \varepsilon_i (\beta_i(a_0) \pm O(\|x_i\|\delta_a) \cdot \langle x_i, u_0 + \delta_u\rangle^2 \\
&= \sum_{i \leq n} \varepsilon_i (\beta_i(a_0) \pm O(\|x_i\|\delta))(\langle x_i, u_0\rangle^2 \pm O(\|x_i\|^2 \delta)) \\
&= \sum_{i \leq n} \varepsilon_i \beta_i(a_0) \langle x_i, u_0\rangle^2 \pm O(\|x_i\|^3 \delta)\,.
\end{aligned}$$

via Lipschitzness of the sigmoid function. Hence for any $\varepsilon_1, \ldots, \varepsilon_n$ and $x_1, \ldots, x_n$,

$$\sup_{\|a\| \leq 10, \|u\| \leq 1} \left| \sum_{i \leq n} \varepsilon_i \beta_i(a) \langle x_i, u\rangle^2 \right| \leq \sup_{(a,u) \in S} \left| \sum_{i \leq n} \varepsilon_i \beta_i(a) \langle x_i, u\rangle^2 \right| + O(\delta) \sum_{i \leq n} \|x_i\|^3$$

With very high probability, the latter is at most $O(\sqrt{nd \log(n/\delta)^2} + d \log(n/\delta)^2) + \tilde{O}(\delta d^{3/2} n)$. Picking $\delta = (nd)^{-O(1)}$ finishes the proof. □

Next we need a lower bound on the population Hessian.

Lemma G.8. *Under the assumptions of Lemma G.5, the population Hessian satisfies*

$$\mathbb{E}\,\nabla^2 L_n(\alpha) \succeq n \cdot e^{-O(\|\alpha\|)} \cdot \mathbf{I}\,.$$

Proof. Let x be a single sample from the covariate distribution. Let $\beta(\alpha)$ be the variance of the prediction of the logistic model α on x. Let v be a fixed unit vector. It suffices to prove a lower bound on $\mathbb{E}\,\beta(\alpha)\cdot\langle v,x\rangle^2$. Note that $\langle v,x\rangle$ is 1-subgaussian with mean 0 and variance 1. By Paley-Zygmund, we have for any $\varepsilon\in(0,1)$ that $\Pr(|\langle v,x\rangle|\geq\varepsilon\|v\|)\geq 1-O(\varepsilon^2)$ So there exists a constant C such that $\Pr(|\langle v,x\rangle|\geq\|v\|/C)\geq 0.99$. Now, $\langle x,\alpha\rangle$ is also mean zero and $\|\alpha\|$-subgaussian. So, with probability at least 0.99, we have $|\langle x,\alpha\rangle|\leq\|\alpha\|/C$. Putting these together, we have that

$$\mathbb{E}\,\beta(\alpha)\langle v,x\rangle^2\geq e^{-O(\|\alpha\|)}\,.$$

□

Putting together Lemma G.8 and Lemma G.7 immediately proves the following lemma.

Lemma G.9. *Under the assumptions of Lemma G.5, with very high probability, for every $\alpha\in\mathbb{R}^d$ we have $\nabla^2 L_n(\alpha)\succeq(n\cdot e^{-O(\|\alpha\|)}-\tilde{O}(\sqrt{dn}))\cdot\mathbf{I}$.*

G.3.2 Gradient at θ

We also need to show that the gradient of L_n has small norm at θ.

Lemma G.10. *Under the assumptions of Lemma G.5, with very high probability,*

$$\|\nabla L_n(\theta)\|\leq\tilde{O}\left(\sqrt{dn}\right)$$

Proof. We start by expanding the gradient of the logistic loss explicitly. For the logistic regression loss the gradient at θ is

$$\nabla L_n(\theta)=\sum_{i=1}^{n}(\sigma(\langle\theta,x_i\rangle)-y_i)x_i,$$

where $\sigma(z)=\frac{1}{1+e^{-z}}$ is the sigmoid function.

Let $\alpha_i=\sigma(\langle\theta,x_i\rangle)-y_i$. Since y_i is generated according to the true model $\Pr(y_i=1|x_i)=\sigma(\langle\theta,x_i\rangle)$, we have $\mathbb{E}[\alpha_i|x_i]=0$.

Therefore, $\nabla L_n(\theta)=\sum_{i=1}^n\alpha_i x_i$ where each $\alpha_i x_i$ is a mean-zero random vector conditional on x_i.

Since $\mathbb{E}[\alpha_i^2|x_i]\leq O(1)$, by the vector Bernstein inequality, we have

$$\Pr(\|\nabla L_n(\theta)\|_2\geq t)\leq 2\exp\left(-\frac{t^2/2}{\sigma^2+Rt/3}\right),$$

where $\sigma^2=\sum_{i=1}^n\mathbb{E}[\|\alpha_i x_i\|_2^2]\leq O(d)$ and $R=O(\sqrt{d\log n})$ with high probability for subgaussian x_i. Setting $t=\tilde{O}(\sqrt{dn})$ finishes the proof. □

G.3.3 Proofs of Lemmas G.6 and G.5

Now we can prove our parameter learning statement.

Proof of Lemma G.5. Suppose the very high probability events of Lemma G.7 and Lemma G.10 hold. Then $\nabla^2 L_n(\alpha)\succeq\Omega(n)\cdot\mathbf{I}$ for all α such that $\|\alpha-\theta\|\leq 1$, and $\|\nabla L_n(\theta)\|\leq\tilde{O}(\sqrt{dn})$. Hence, $\|\hat{\theta}-\theta\|\leq\tilde{O}(\sqrt{dn})/n=\tilde{O}(\sqrt{d/n})$. □

Proof of Lemma G.6. The first part follows immediately from Lemma G.7 and Lemma G.5 using that $\|\hat{\theta}\|\leq 2$.

For the second statement, it will be enough to show that with high probability all $S\subseteq[n]$ with $|S|\leq O(n/\log n)$ satisfy $\sum_{i\in S}x_i x_i^\top\leq O(|S|\log n+d)$. Putting together Vershynin (2010), Theorem 5.39, and a union bound over all S of size $O(n/\log n)$ completes the result. □

G.4 Concentration of Third-Order Derivatives

Lemma G.11. *Under the assumptions of Lemma G.5, with high probability, for every* $\|\alpha\| \leq 10$,

$$\sup_{\|e\|=1} |\langle \nabla^3 L_n(\alpha), e \otimes e \otimes e \rangle - \mathbb{E}\langle \nabla^3 L_n(\alpha), e \otimes e \otimes e \rangle| \leq \tilde{O}(\sqrt{nd} + d^{3/2}) .$$

We omit the proof of Lemma G.11 because it follows the same outline as Lemma G.7, substituting the sub-Weibull concentration bound of Zhang & Wei (2022) for composition of subexponential random variables, as in the proof of Lemma G.2.

H LLM Usage

LLMs (ChatGPT) were used to a limited extent for converting existing proofs into latex and for generating tikz diagrams. These were all verified by the authors and match the original text and / or diagrams designed by the authors.

Moreover, LLMs (ChatGPT) were used to expand on our literature survey to find additional related works (as a supplementary tool in addition to standard literature survey techniques).