Journal of Machine Learning Research 26 (2025) 1-58 Submitted 7/25; Revised 12/25; Published 12/25

Certified Machine Unlearning Under High Dimensional Regime

Haolin Zou HZ2574@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10025, USA

Arnab Auddy AUDDY.1@OSU.EDU
Department of Statistics
The Ohio State University
Columbus, OH 43210, USA

Yongchan Kwon YKWON@TOGETHER.AI
Together AI

Kamiar Rahnama Rad KAMIAR.RAHNAMARAD@BARUCH.CUNY.EDU
Baruch College
The City University of New York
New York, NY 10012, USA

Arian Maleki ARIAN@STAT.COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY 10025, USA

Editor: Samy Bengio

Abstract

Machine unlearning focuses on the computationally efficient removal of specific training data from trained models, ensuring that the influence of forgotten data is effectively eliminated without the need for full retraining. Despite advances in low-dimensional settings, where the number of parameters p is much smaller than the sample size n, extending similar theoretical guarantees to high-dimensional regimes remains challenging. We study an unlearning algorithm that starts from the original model parameters and performs a theory-guided sequence of Newton steps. After this update, carefully scaled isotropic Laplacian noise is added to the estimate to ensure that any (potential) residual influence of the deletion set is completely removed. We show that when both $n, p \to \infty$ with a fixed ratio n/p, significant theoretical and computational obstacles arise due to the interplay between the complexity of the model and the finite signal-to-noise ratio. Finally, we show that, unlike in low-dimensional settings where one Newton step suffices, in high-dimensional problems at least two Newton steps are required to effectively unlearn a fixed number of data points, and even more steps are required when the deletion set scales with n. We provide numerical experiments to support the theoretical claims of the paper.

Keywords: Certified removal, high dimensional statistics, Newton method, regularized empirical risk minimization, direct perturbation.

1. Introduction

Many real-world machine learning systems, including healthcare diagnostic tools and models like ChatGPT and DALL-E, rely on diverse user and entity data during training. If a user requests their data to be removed, it is reasonable to expect the responsible companies or entities to not only delete the data from their datasets but also eliminate its trace from the trained models. This process requires frequent and costly retraining of models. To address this challenge, the field of machine unlearning has emerged, focusing on efficient and less computationally intensive methods to remove dataset traces from models.

This field has made significant progress over the past few years Cao and Yang (2015); Bourtoule et al. (2021); Nguyen et al. (2022); Chundawat et al. (2023); Tarun et al. (2023); Gupta et al. (2021); Chen et al. (2021). Substantial empirical research, coupled with rigorous theoretical results, have established a strong foundation for this area.

As we will clarify in Section 4.2, existing theoretical results in the field of machine unlearning usually have an implicit focus on low-dimensional settings, where the number of model parameters p is much smaller than the number of observations n, i.e., $p \ll n$. However, in many real-world applications, the number of parameters is comparable to—or even exceeds—the number of observations. This discrepancy raises a fundamental and currently unresolved question in the field:

Are existing machine unlearning methods reliable in high-dimensional regimes as well?

The goal of this paper is to answer the above question for the machine unlearning algorithms that are based on the Newton method. More specifically, the paper makes the following contributions:

1. We study the performance of machine unlearning algorithms under proportional high-dimensional asymptotic settings (PHAS), where both the number of parameters p and the number of observations n are large, and their ratio $n/p \to \gamma_0 \in (0, \infty)$.

2. As we will clarify later, some of the notions introduced for evaluating the certifiability and accuracy of machine unlearning algorithms — such as ϵ-certifiability from Guo et al. (2019) and the excess risk considered in Sekhari et al. (2021) — are not well-suited to high-dimensional settings. To address this limitation, we refine some of these existing notions and propose new metrics tailored to evaluating the certifiability and accuracy of machine unlearning algorithms in high-dimensional regimes.

3. We consider the popular class of regularized empirical risk minimization (R-ERM) with a focus on generalized linear models, studied in previous works Guo et al. (2019); Sekhari et al. (2021); Neel et al. (2021), and analyze the performance of machine unlearning algorithms based on the Newton method under our high-dimensional setting. As a result, we show that:

   (a) Unlike the low-dimensional settings, machine unlearning algorithms based on a single Newton step (similar to those introduced in Guo et al. (2019) and Sekhari et al. (2021)) are not reliable, even when removing only one data point from the dataset.

(b) In contrast, a machine unlearning algorithm yields a reliable estimate with two Newton iterations when a number $m = O(1)$ datapoints are unlearned. Furthermore, we quantify the minimum number of steps needed when m grows with n.

2. Our Framework

2.1 General framework of approximate machine unlearning

In this paper, we consider a *generalized linear model* with an i.i.d. sample $\mathcal{D} = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\} \in \mathbb{R}^{n\times(p+1)}$ as an independent and identically distributed (i.i.d.) sample from some joint distribution

$$(y_i, \boldsymbol{x}_i) \sim q(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)p(\boldsymbol{x}_i),$$

where $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$ denote the response and feature vector respectively, and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the parameter of interest. An estimator of $\boldsymbol{\beta}^* \in \Theta$, denoted as $\hat{\boldsymbol{\beta}}$, may be obtained from a *learning algorithm* $A : \mathbb{R}^{n\times(p+1)} \to \Theta$. To formalize the idea of machine unlearning problem, let $\mathcal{M} \subset \{1, 2, ..., n\}$ be the subset of data indices to be removed, let $\mathcal{D}_\mathcal{M} := \{(y_i, \boldsymbol{x}_i) : i \in \mathcal{M}\}$ be the corresponding subset of $\mathcal{D}$, and let $\mathcal{D}_{\setminus\mathcal{M}} := \mathcal{D}\setminus\mathcal{D}_\mathcal{M}$. In order to remove $\mathcal{D}_\mathcal{M}$ and all traces of it from the model, we essentially need to obtain $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} = A(\mathcal{D}_{\setminus\mathcal{M}})$ [1], which is also called 'exact machine unlearning'. However, in many applications, it is considered impractical to exactly compute $A(\mathcal{D}_{\setminus\mathcal{M}})$, so **approximate** unlearning methods are more desirable. Such methods calculate an efficient approximation $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ of $A(\mathcal{D}_{\setminus\mathcal{M}})$.

To evaluate $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, inspired by Guo et al. (2019); Dwork (2006), we consider the following two principles that a 'good' machine unlearning algorithm should satisfy:

P1. [Certifiability] No information about the data points in $\mathcal{M}$ should be recoverable from the unlearning algorithm, as the users have requested their data to be entirely removed.

P2. [Accuracy] $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ should be "close" to $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ in terms of down stream tasks such as out-of-sample prediction.

The aforementioned principles serve as a conceptual framework for machine unlearning algorithms. However, to enable systematic evaluation and facilitate objective comparisons, it is essential that these principles be formalized into explicit, quantitative criteria and metrics.

Let us begin with the certifiability principle. Since $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ is only an approximation of $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, it is likely to inherently retain some information about $\mathcal{D}_\mathcal{M}$. To hide such residual information, random noise must be introduced into the estimate.[2] For example, independent and confidential noise can be directly added to $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$. Consequently, in studying the machine

1. Technically, A refers to a sequence of functions $\{A_{n,p} : \mathbb{R}^{n\times(p+1)} \to \Theta |\ n, p \in \mathbb{N}_+\}$ so $A(\mathcal{D}) = A_{n,p}(\mathcal{D})$ and $A(\mathcal{D}_{\setminus\mathcal{M}}) = A_{n-m,p}(\mathcal{D}_{\setminus\mathcal{M}})$ where $m := |\mathcal{M}|$, as the two functions are defined on different spaces thus cannot be identical. But we drop the subscripts for notational brevity.
2. For more information about this claim, the reader can refer to the literature of differential privacy Dwork (2006).

unlearning problem, we will focus on **randomized** approximations. We use the notation $\boldsymbol{b}$ for a random perturbation used to obtain such randomized approximation. In addition, inspired by Sekhari et al. (2021), we also allow the system to store and use a summary statistic $T(\mathcal{D})$ for the unlearning algorithm, so it can be formalized as:

$$\tilde{\boldsymbol{\beta}}^R_{\setminus\mathcal{M}} = \tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}).$$

where the supersript R in $\tilde{\boldsymbol{\beta}}^R_{\setminus\mathcal{M}}$ stands for 'randomized'. Inspired by differential privacy (Dwork (2006)) and the existing literature of machine unlearning, including Guo et al. (2019); Sekhari et al. (2021), we propose the following criterion for the certifiability principle, by comparing the distribution of the unlearned model with a randomized baseline of exact unlearning.

Definition 1 ((ϕ, ϵ)- Probabilistically certified approximate unlearning (PAU)) *For $\phi, \epsilon > 0$, a randomized unlearning algorithm $\tilde{A}$ is called a (ϕ, ϵ)-probablistically certified approximate machine unlearning (PAU) algorithm, if and only if $\exists \mathcal{X} \subset \mathbb{R}^{n\times(p+1)}$ with $\mathbb{P}(\mathcal{D} \in \mathcal{X}) \geq 1 - \phi$, such that $\forall \mathcal{D} \in \mathcal{X}$ and $\forall \mathcal{M} \subset [n]$ with $|\mathcal{M}| \leq m$, $\forall$ measurable $\mathcal{T} \subset \Theta$, we have*

$$\mathrm{e}^{-\epsilon} < \frac{\mathbb{P}\left(\tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}) \in \mathcal{T} \,\middle|\, \mathcal{D}\right)}{\mathbb{P}(\tilde{A}(\emptyset, A(\mathcal{D}_{\setminus\mathcal{M}}), T(\mathcal{D}_{\setminus\mathcal{M}}), \boldsymbol{b}) \in \mathcal{T} | \mathcal{D})} \leq \mathrm{e}^{\epsilon} \tag{1}$$

where $\boldsymbol{b}$ encodes external randomness introduced in the randomized unlearning algorithm $\tilde{A}$, such as noise injection.

Our criterion 1 differs slightly from similar definitions in the literature, such as the "ϵ-certified removal" in Guo et al. (2019), and the "(ϵ, δ)-unlearning" in Sekhari et al. (2021). Although a detailed comparison is postponed to Section 3.3 and Section 4.2, here we briefly outline the main differences and their motivations in the following remarks.

Remark 2 *The probabilities in* (1) *are conditional on $\mathcal{D}$. This is stronger than previous definitions based on marginal probabilities (Guo et al. (2019); Sekhari et al. (2021)), which can be regarded as an **average effect** over all possible datasets through the tower rule $\mathbb{P}(\cdot) = \mathbb{E}[\mathbb{P}(\cdot|\mathcal{D})]$, whereas our definition aims at guaranteeing the ratio in* (1) *close to 1 for **almost all realizations** of $\mathcal{D}$. This requires an additional source of randomness encoded by $\boldsymbol{b}$, resulting in **randomized** unlearning algorithms. Accordingly, the baseline estimator $\tilde{A}(\emptyset, A(\mathcal{D}_{\setminus\mathcal{M}}), T(\mathcal{D}_{\setminus\mathcal{M}}), \boldsymbol{b})$ in the denominator refers to the hypothetical outcome of the unlearning algorithm given the exact unlearning result $A(\mathcal{D}_{\setminus\mathcal{M}})$ and a null unlearning request. It should be interpreted as a perturbed version of the exact unlearned model $A(\mathcal{D}_{\setminus\mathcal{M}})$.*[3]

Remark 3 *The parameter ϕ characterizes the chance of a 'bad dataset' to break the inequality* (1)*. Within our framework, ϕ will depend on (m, n, p) and we expect it to be small*

3. Technically we can make this more clear by defining a perturbation function $R(\boldsymbol{\beta}, \boldsymbol{b})$ that returns a randomized version of $\boldsymbol{\beta}$ using a random perturbation $\boldsymbol{b}$, and define the non-randomized unlearning algorithm to be $\bar{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}))$, and define $\tilde{A} := \bar{A} \circ R$. If we assume $\bar{A}(\emptyset, \boldsymbol{\beta}, T) \equiv \boldsymbol{\beta}$, then we have $\tilde{A}(\emptyset, A(\mathcal{D}_{\setminus\mathcal{M}}), T(\mathcal{D}_{\setminus\mathcal{M}}), \boldsymbol{b}) = R(A(\mathcal{D}_{\setminus\mathcal{M}}))$, a randomized version of $A(\mathcal{D}_{\setminus\mathcal{M}})$. But for notational brevity, we do not introduce the notations above into the paper.

for large enough n, p and small enough m (compared to n), which will be made more clear later in Theorem 7. The key motivation of such change is that the existing ϵ or (ϵ, δ)-certifiability require worst-case bounds of certain quantities over all possible realization of $\mathcal{D}$, which can be prohibitively large or even unbounded, such as the "gradient residual norm" in Theorem 1 of Guo et al. (2019), and the global Lipschitzness constant of ℓ in Assumption 1 of Sekhari et al. (2021). To overcome this, we obtain high-probability bounds for corresponding quantities in high dimensions, for example Lemma 9, following which there is a more detailed discussion about this claim.

Intuitively, injecting a large amount of noise $\boldsymbol{b}$ into the estimates can conceal all residual information in $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ thus in favor of the certifiability principle and criterion 1, but it will harm the accuracy of it when used in downstream tasks such as prediction. To address this, we need a measure of accuracy:

Definition 4 (Generalization Error Divergence (GED)) *Let $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta})$ be a measure of error between y and $\boldsymbol{x}^\top\boldsymbol{\beta}$, and let $(y_0, \boldsymbol{x}_0)$ independent with $\mathcal{D}$. Then the **Generalization Error Divergence (GED)** of the learning and unlearning algorithms $A, \tilde{A}$ is defined as:*

$$\mathrm{GED}(A, \tilde{A}) := \mathbb{E}\left(|\ell(y_0|\boldsymbol{x}_0^\top A(\mathcal{D}_{\setminus\mathcal{M}})) - \ell(y_0|\boldsymbol{x}_0^\top \tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}))|\big|\mathcal{D}\right), \tag{2}$$

where the superscript ϵ in $\mathrm{GED}(A, \tilde{A})$ emphasizes on the fact that $\tilde{A}$ is (ϕ, ϵ)-PAU.

This metric measures the difference between the generalization error of the approximate unlearning algorithm $\tilde{A}$ and the exact unlearning $A(\mathcal{D}_{\setminus\mathcal{M}})$, which naturally arises when we want to compare the prediction performance of $\tilde{A}$ on a new data point compared to exact unlearning $A(\mathcal{D}_{\setminus\mathcal{M}})$. Note that other notions have been introduced in the literature for measuring the accuracy of machine unlearning algorithms. For instance, Sekhari et al. (2021) used the excess risk of $\tilde{A}$ against that of the true minimizer of population risk, but we will show in Section 4.2 that such notions are not useful in the high dimensional settings.

2.2 Regularized ERM and proportional high-dimensional asymptotic settings

To estimate $\boldsymbol{\beta}^*$ in the GLM model, researchers often use the following optimization problems known as regularized empirical risk minimization (R-ERM):[4]

$$\hat{\boldsymbol{\beta}} = A(\mathcal{D}) \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n} \ell(y_i|\boldsymbol{x}_i^\top\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}) \tag{3}$$

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} = A(\mathcal{D}_{\setminus\mathcal{M}}) \triangleq \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{j\notin\mathcal{M}} \ell(y_j|\boldsymbol{x}_j^\top\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}). \tag{4}$$

In this optimization problem, $\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta})$ is called the loss function, which is typically set to $-\log q(y|\boldsymbol{x}^\top\boldsymbol{\beta})$ when q is known, and $r(\boldsymbol{\beta})$ is called the regularizer, which is usually a convex function minimized at $\boldsymbol{\beta} = 0$. It aims at reducing the variance of the estimate and therefore, the value of $\lambda \in [0, \infty)$ controls the amount of regularization. R-ERM is used

4. One simple extension of these ideas is to include more than one regularizer with mutiple regularization strengths λ_1, λ_2, etc. For notational brevity we consider the simplest case, while generalization into multiple regularizers is straightforward.

in many classical and modern learning tasks, such as linear regression, matrix completion, Poisson and multinomial regressions, classification, and robust principal components.

Again, the objective is to find an unlearning algorithm $\tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b})$ that is (ϕ, ϵ)-PAU (Definition 1) and has small GED (Definition 4).

As described before, we aim to study high-dimensional settings in which both (n) and (p) are large. Towards this goal, we use one of the most widely-adopted high-dimensional asymptotic frameworks, proportional high-dimensional asymptotic setting (PHAS).

The proportional high-dimensional asymptotic setting (PHAS) has provided valuable insights into the optimality and practical effectiveness of various estimators over the past decade Maleki (2011); Donoho et al. (2009); Maleki and Montanari (2010); Bayati and Montanari (2011); Donoho et al. (2011); Bayati and Montanari (2012); Mousavi et al. (2017); El Karoui et al. (2013); Donoho and Montanari (2016); Oymak et al. (2013); Karoui and Purdom (2016); Amelunxen et al. (2013); Krzakala et al. (2012a,b); Celentano and Montanari (2024); Miolane and Montanari (2021); Wang et al. (2022, 2020); Li and Wei (2021); Liang and Sur (2022); Dudeja et al. (2023); Fan (2022); Dobriban and Wager (2018); Dobriban and Liu (2019); Wainwright (2019).

Definition 5 (Proportional High-dimensional Asymptotic Setting (PHAS)) *Assume that both n and p grow to infinity, and that $n/p \to \gamma_0 \in (0, \infty)$.*

While our theoretical goal is to derive finite-sample results applicable to any values of n and p, PHAS (Definition 5) will serve as a basis for simplifying and interpreting these results in high-dimensional settings. By default, all the following "big O" notations should be interpreted under the directional limit of PHAS, i.e. $n, p \to \infty, n/p \to \gamma_0 \in (0, \infty)$. In contrast, in classical, or low dimensional settings, the direction of limit is usually $n \to \infty, p \equiv p_0$, in which case $n/p \to \infty$.

2.3 Newton method and direct perturbation

Inspired by multiple algorithms in the literature of machine unlearning, such as the algorithms introduced in Guo et al. (2019); Sekhari et al. (2021), in this paper we use the **Newton method** with **direct perturbation** to construct a (ϕ, ϵ)-PAU algorithm with guaranteed prediction accuracy in terms of GED.

The Newton method, also called Newton-Raphson method, is essentially an iterative root-finding algorithm.

Definition 6 (Newton Method) *Suppose $\boldsymbol{f} : \mathbb{R}^p \to \mathbb{R}^p$ has an invertible Jacobian matrix $\boldsymbol{G}$ anywhere in an open set $\Theta \subset \mathbb{R}^p$, and $\boldsymbol{f}$ has a root $\boldsymbol{f}(\boldsymbol{\beta}^*) = \mathbf{0}$ in Θ. Starting from an initial point $\boldsymbol{x}^{(0)} \in \Theta$, the Newton method is the following iterative procedure: for step $t \geq 1$,*

$$\boldsymbol{x}^{(t)} := \boldsymbol{x}^{(t-1)} - \boldsymbol{G}^{-1}(\boldsymbol{x}^{(t-1)})\boldsymbol{f}(\boldsymbol{x}^{(t-1)}).$$

For more information about Newton method, please check Section 9.5 of Boyd and Vandenberghe (2004). Denote the objective function of $\hat{\boldsymbol{\beta}}$ as $L(\boldsymbol{\beta})$, i.e.,

$$L(\boldsymbol{\beta}) := \sum_{i=1}^{n} \ell(y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})$$

and that of $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ as $L_{\setminus\mathcal{M}}(\boldsymbol{\beta})$, meaning:

$$L_{\setminus\mathcal{M}}(\boldsymbol{\beta}) := \sum_{j\notin\mathcal{M}} \ell(y_j|\boldsymbol{x}_j^\top\boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}).$$

Notice that finding $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ is equivalent to solving the root of the gradient of $L_{\setminus\mathcal{M}}$ when it is smooth, and that $\hat{\boldsymbol{\beta}}$ is reasonably close to $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ if $m = |\mathcal{M}|$ is small. Hence we can initialize the Newton method at $\tilde{\boldsymbol{\beta}}^{(0)}_{\setminus\mathcal{M}} = \hat{\boldsymbol{\beta}}$ and iteratively compute:

$$\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} = \tilde{\boldsymbol{\beta}}^{(t-1)}_{\setminus\mathcal{M}} - \boldsymbol{G}^{-1}_{\setminus\mathcal{M}}(\tilde{\boldsymbol{\beta}}^{(t-1)}_{\setminus\mathcal{M}})\nabla L_{\setminus\mathcal{M}}(\tilde{\boldsymbol{\beta}}^{(t-1)}_{\setminus\mathcal{M}}), \quad t \geq 1$$

where $\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta})$ is the Hessian of $L_{\setminus\mathcal{M}}$, and $\nabla L_{\setminus\mathcal{M}}$ is its gradient.

Suppose that we stop the Newton method after T iterations. In order to obscure residual information of $\mathcal{D}_{\mathcal{M}}$, we introduce **direct perturbation**, resulting in the unlearning method we consider in this paper, **Perturbed Newton** estimator[5]:

$$\tilde{\boldsymbol{\beta}}^{R,T}_{\setminus\mathcal{M}} = \tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}) \triangleq \tilde{\boldsymbol{\beta}}^{(T)}_{\setminus\mathcal{M}} + \boldsymbol{b}.$$

Throughout this paper we consider Isotropic Laplacian distribution for $\boldsymbol{b}$, which has the density

$$p_{\boldsymbol{b}}(\boldsymbol{b}) = \frac{C^p\Gamma(\frac{p}{2})}{2\pi^{\frac{p}{2}}\Gamma(p)}\mathrm{e}^{-C\|\boldsymbol{b}\|} \propto \mathrm{e}^{-C\|\boldsymbol{b}\|}.$$

for some scale parameter $C > 0$. Drawing a sample from $p_{\boldsymbol{b}}(\boldsymbol{b})$ is equivalent to drawing $\|\boldsymbol{b}\|$ from a $Gamma(p, C^{-1})$ distribution, then sampling $\boldsymbol{b}$ uniformly on the sphere with radius $\|\boldsymbol{b}\|$ (see Lemma 16). The reason to consider this distribution is that its log density is Lipschitz in $\|\boldsymbol{b}\|$, thus naturally connected with the (ϕ, ϵ)-PAU criteria. This will be shown later in more details in Lemma 9.

It remains a question when to stop the Newton iteration. To find an 'exact' solution, it is usually run until certain convergence criterion is met. In contrast, it has been proposed in the literature that one Newton step suffices in the low dimensions, for example in Guo et al. (2019); Sekhari et al. (2021). However, as will be clarified later, under PHAS even to remove a single data point ($m = 1$), we need **at least two Newton steps** to guarantee good prediction accuracy (i.e. GED $\to 0$). This will be discussed later in Section 3.2.

Finally, regarding computational complexity, we note that the number of Newton iterations depends on the loss landscape. For generalized linear models, where the objective is convex, a finite number of Newton iterations is sufficient to reach the unique global optimum. In our experiments, as reported in Section 5, naive leave-data-out retraining with Newton's method required a constant number of iterations, typically between 4 and 20, depending on the initialization and the curvature of the loss. Thus, replacing full retraining with one or two Newton updates reduces computational cost by approximately 50%–90%. Exploring techniques that further reduce computational overhead is an interesting direction for future work.

5. Note that even though we do not distinguish the perturbations $\boldsymbol{b}$, it should be independently drawn each time it is used.

2.4 Notations

We adopt the following mathematical conventions. Scalars and scalar-valued functions are denoted by English or Greek letters (e.g. $C_1, a, \lambda > 0$)[6]. Caligraphic uppercase letters are used for sets, families or events (e.g. $\mathcal{T} \subset \mathbb{R}^p$) with an exception that $\mathcal{N}$ refers to the Gaussian distribution. $\mathbb{R}, \mathbb{R}_+$ denote the set of real, positive real numbers respectively. Vectors are represented by bold lowercase letters (e.g. $\boldsymbol{x} \in \mathbb{R}^p$), and matrices by bold uppercase (e.g. $\boldsymbol{X} \in \mathbb{R}^{n\times p}$). For a matrix $\boldsymbol{X}$, $\|\boldsymbol{X}\|$, $\|\boldsymbol{X}\|_{Fr}$, $\lambda_{\min}(\boldsymbol{X})$, $\lambda_{\max}(\boldsymbol{X})$, $\sigma_{\min}(\boldsymbol{X})$, $\sigma_{\max}(\boldsymbol{X})$ and $\mathrm{tr}(\boldsymbol{X})$ denote the (Euclidean) operator norm, Frobenius norm, minimal and maximal eigenvalues, minimal and maximal singular values and the trace of $\boldsymbol{X}$, respectively. Moreover,

$$\|\boldsymbol{X}\|_{p,q} := \sup_{\|\boldsymbol{w}\|_p \le 1} \|\boldsymbol{X}\|_q$$

is the operator norm induced by ℓ_p, ℓ_q norms, for $p, q \in \mathbb{R}_+ \cup \{+\infty\}$.

We denote $[n] := \{1, 2, \cdots, n\}$ for some $n \in \mathbb{N}_+$. For any index set $\mathcal{M} \subset [n]$, we use $\boldsymbol{X}_{\mathcal{M}}$ to denote the sub-matrix of $\boldsymbol{X}$ that consists of the rows indexed by $\mathcal{M}$. Similarly, $\boldsymbol{a}_{\mathcal{M}}$ represents the sub-vector of $\boldsymbol{a}$ containing the elements indexed by $\mathcal{M}$. More generally, we use the subscript $\cdot_{\mathcal{M}}$ to refer to a quantity corresponding to $\mathcal{D}_{\mathcal{M}} = \{(y_i, \boldsymbol{x}_i), i \in \mathcal{M}\}$, and the subscript $\cdot_{\setminus\mathcal{M}}$ to refer to the quantities corresponding to $\mathcal{D}_{\setminus\mathcal{M}} := \mathcal{D}\setminus\mathcal{D}_{\mathcal{M}}$.

We define $\dot{\ell}(y|z)$ and $\ddot{\ell}(y|z)$ as the first and second derivatives of the function ℓ with respect to z, respectively. Furthermore, we write $\ell_i(\boldsymbol{\beta})$ for short of $\ell(y_i|\boldsymbol{x}_i^\top\boldsymbol{\beta})$, and similarly $\dot{\ell}_i(\boldsymbol{\beta})$ and $\ddot{\ell}_i(\boldsymbol{\beta})$. Additionally, we introduce the vectors

$$\dot{\boldsymbol{\ell}} := \left[\dot{\ell}_1(\hat{\boldsymbol{\beta}}), \cdots, \dot{\ell}_n(\hat{\boldsymbol{\beta}})\right]^\top, \quad \ddot{\boldsymbol{\ell}} := \left[\ddot{\ell}_1(\hat{\boldsymbol{\beta}}), \cdots, \ddot{\ell}_n(\hat{\boldsymbol{\beta}})\right]^\top.$$

We define $\mathbf{diag}[\boldsymbol{a}]$ or $\mathbf{diag}[a_i]_{i\in[n]}$ as a diagonal matrix whose diagonal elements correspond to the entries of the vector $\boldsymbol{a} = (a_1, \cdots, a_n)^\top$.

We use the following notations for the limiting behavior of sequences. We use $\mathrm{polylog}(n)$ as a shorthand for finite degree polynomials of $\log(n)$. We use the conventional notations for limiting behavior of sequences: $a_n = o(b_n)$, $O(b_n)$, $\Omega(b_n)$, $\Theta(b_n)$ respectively mean that a_n/b_n is convergent (to 0), bounded, divergent, and asymptotically equivalent. We use similar notations for their stochastic analogies, e.g. $X_n = o_p(1)$ iff $X_n \to 0$ in probability, and so forth. Notably we use the notation $X_n \sim \Theta_p(1)$ iff X_n is $O_p(1)$ but not $o_p(1)$. Finally, the symbol "$\perp\!\!\!\perp$" means "independent" in probability.

3. Our Contributions

In this section, we present our main theoretical results on the certifiability and accuracy of the machine unlearning algorithms introduced in Section 2.3, under the proportional asymptotic regime. Before stating the results, we first provide a detailed overview of the assumptions underlying our analysis.

6. Usually we use $C(n)$ to indicate a quantity that grows at most at a speed of $\mathrm{polylog}(n)$, otherwise we tend to put n in the subscript.

3.1 Main assumptions

Our first series of assumptions are concerned with the structural properties of ℓ and r.

Assumption A1 (Separability) *The regularizer is separable:*

$$r(\boldsymbol{\beta}) = \sum_{k\in[p]} r_k(\beta_k).$$

This assumption can be generalized to include a linear transform: $r(\boldsymbol{\beta}) = \sum_{j\in[l]} r_j(\boldsymbol{a}_j^\top \boldsymbol{\beta})$, but we make the simplified assumption that $\boldsymbol{a}_j = \boldsymbol{e}_j$ where $\boldsymbol{e}_j$ is the jth canonical basis of $\mathbb{R}^p$ to avoid cumbersome notations. Generalizing the current proof to arbitrary $\boldsymbol{a}_j$ is trivial.

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

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

These assumptions ensure that the R-ERM estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ are unique, and the Newton method is applicable. While true for many applications, in other cases where certain structures such as sparsity of β is assumed, these assumptions can be violated. While a few papers have shown how the Newton method can be extended to non-differentiable settings (e.g. Auddy et al. (2024); Wang et al. (2018)), the theoretical study of such cases will not be the focus of this work, and are left for a future research. Note that we implicitly assumed ℓ and r to be non-negative without loss of generality. This can be achieved by subtracting their minimum (which are finite since ℓ, r are proper convex) from the function themselves.

In addition to the structural properties, we make several assumptions on the probablisic aspects of the data and model:

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

The Gaussianity assumption is prevalent in theoretical papers dealing with high-dimensional problems, for example Miolane and Montanari (2021); Weng et al. (2018); Rahnama Rad and Maleki (2020); Auddy et al. (2024). Although our proofs can be generalized to a broader class of distributions of $\boldsymbol{x}_i$ beyond Gaussianity, we do not discuss it in details in this paper.

The scaling we have adopted in the above assumption is based on the following rationale. First notice that since $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$,

$$\mathrm{var}(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*) = \mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)^2 \le \frac{C_X}{p}\|\boldsymbol{\beta}^*\|_2^2.$$

Heuristically speaking, uner PHAS and when the elements of $\boldsymbol{\beta}^*$ are $O(1)$, we have $\|\boldsymbol{\beta}^*\|_2 = O(\sqrt{p})$, and hence $\mathbb{E}(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)^2 = O(1)$. On the other hand, it is reasonable to assume that $y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}$ has $\Theta(1)$ variance. Therefore, under the settings of the paper we can see that the signal-to-noise ratio (SNR) of each data point, defined as $\frac{\mathrm{var}(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)}{\mathrm{var}(y_i|\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)}$, remains

bounded. We now introduce two more assumptions on the likelihood function ℓ and the response y. These are typically used in the analysis of high dimensional regression problems and are satisfied for a host of natural examples including linear and logistic regression. See, e.g., Zou et al. (2024).

Assumption B2 $\exists C, s > 0$ *such that*

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

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

This assumption requires that the derivatives of ℓ grows with y and z at most as fast as a polynomial function with order s, and the regularizer should be $O(\text{polylog}(n))$-Hessian-Lipschiz, with one example be the ridge penalty: $r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$ and $\nabla^2 r(\boldsymbol{\beta}) = \mathbb{I}_p$.

Assumption B3 $\mathbb{P}(|y_i| > C_y(n)) \leq q_n$ *and* $\mathbb{E}|y_i|^{2s} \leq C_{y,s}$ *for some* $C_y(n) = O(\text{polylog}(n))$*, a constant* $C_{y,s}$ *and* $q_n = o(n^{-1})$

This assmption essentially requires all $|y_i|$ to be stochastically bounded even when n, p increases.

Below are some examples where Assumptions B2 and B3 are satisfied. For simplicity we assume $\boldsymbol{x}_i \sim \mathcal{N}(0, \frac{1}{p}\mathbb{I}_p)$.

Example 1 (Linear regression) *Suppose* $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*, \sigma^2)$*, then we have* $y_i \sim \mathcal{N}(0, \tau^2)$ *with* $\tau^2 := \sigma^2 + \frac{1}{p}\|\boldsymbol{\beta}^*\|^2$*. Its negative log-likelihood is the* ℓ_2 *loss:*

$$\ell(y,z) = \frac{1}{2}(y-z)^2, \quad \dot{\ell}(y,z) = z-y, \quad \ddot{\ell}(y,z) = 1, \quad \dddot{\ell}(y,z) = 0.$$

And by Lemma 14 we have the following concentration for y_i*:*

$$\mathbb{P}(|y_i| \geq 6\tau\sqrt{\log(n)}) \leq \frac{1}{\sqrt{6\pi}} n^{-3}.$$

Example 2 (Logistic regression) *Suppose* $y_i \sim$ *Bernoulli*(p_i) *where* $p_i = (1+e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}^*})^{-1}$*. The negative log-likelihood is then*

$$\begin{aligned}
\ell(y,z) &= y\log(1+e^{-z}) + (1-y)\log(1+e^{z}), \quad y \in \{0,1\} \leq 2\log(2) + 2z, \\
|\dot{\ell}(y,z)| &= \left|\frac{e^z}{1+e^z} - y\right| \leq 1 + |y|, \\
|\ddot{\ell}(y,z)| &= \left|\frac{e^z}{(1+e^z)^2}\right| \leq 1 \\
|\dddot{\ell}(y,z)| &= \left|-\frac{1}{1+e^z} + \frac{3}{(1+e^z)^2} - \frac{2}{(1+e^z)^3}\right| \leq 6,
\end{aligned}$$

and obviously $|y_i| \leq 1$ *so that Assumption B3 is also satisfied.*

3.2 Main theorem and its implications

The main objective of this paper is to answer the following two questions:

$\mathcal{Q}_1$: Given a t-step Newton estimator $\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}}$, can we find a large enough perturbation $\boldsymbol{b}$ so that $\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}$ is (ϕ,ϵ)-PAU, for some $\phi \to 0$ under PHAS?

$\mathcal{Q}_2$: Given the perturbation level in $\mathcal{Q}_1$, can we find a sufficient number of Newton steps T such that , $\mathrm{GED}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, \tilde{\boldsymbol{\beta}}^{R,T}_{\setminus\mathcal{M}}) \to 0$ under PHAS?

The two theorems below are the two main theoretical results of our paper. They guarantee the **certifiability** and **accuracy** of the perturbed Newton-based estimators and answer $\mathcal{Q}_1$ and $\mathcal{Q}_2$ respectively.

Theorem 7 (Certifiability) *Under Assumptions A1-A3 and B1-B3, suppose that* $m = o(n^{\frac{1}{3}})$ *and that* $\boldsymbol{b}$ *has density* $p_{\boldsymbol{b}}(\boldsymbol{b}) \propto \mathrm{e}^{-\frac{\epsilon}{r_{t,n}}\|\boldsymbol{b}\|}$ *with*

$$r_{t,n} = [C_1(n)]^{2^{t-1}} \left(\frac{C_2(n) m^3}{2\lambda\nu n}\right)^{2^{t-2}},$$

for some $C_1(n), C_2(n) = O(\mathrm{polylog}(n))$ *and* $\epsilon > 0$. *Then* $\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}} = \tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}$ *achieves* (ϕ_n, ϵ)*-PAU with*

$$\phi_n = nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p} \to 0.$$

The proof, including the explicit expressions for $C_1(n), C_2(n)$, can be found in Section B.3. Theorem 7 shows that, with a perturbation $\boldsymbol{b}$ with a certain scale $r_{t,n}$, we can obtain a (ϕ_n, ϵ)-PAU algorithm from any steps of Newton iterations. However, it does not provide information on the accuracy of the approximations. Recall the metric of accuracy we defined:

$$\mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) = \mathbb{E}\left(\left|\ell(y_0, \boldsymbol{x}_0^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})) - \ell(y_0, \boldsymbol{x}_0^\top(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}))\right| | \mathcal{D}\right).$$

Our next theorem calculates the accuracy of the estimates that are obtained from the Newton method.

Theorem 8 (Accuracy) *Under Assumptions A1-A3 and B1-B3, with probability at least* $1 - (n+1)q_n - 14n^{1-c} - ne^{-p/2} - 2e^{-p} - e^{-(1-\log(2))p}$,

$$\max_{|\mathcal{M}|\le m} \mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \le \left(\frac{2\sqrt{p}}{\epsilon} + \frac{1}{\sqrt{p}}\right) r_{t,n}\sqrt{2m+2s}\cdot \mathrm{polylog}(n), \quad \forall t \ge 1.$$

where s *is the constant in Assumption B3. Moreover, let* $\alpha := \log(m+1)/\log(n)$. *If* $t > T = 1 + \log_2\left(\frac{\alpha+1}{1-3\alpha}\right)$, *then under PHAS, we have*

$$\mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) = o_p(1).$$

The proof of Theorem 8 can be found in Section B.3.

Theorems 7 and 8 answer the Questions $\mathcal{Q}_1$ and $\mathcal{Q}_2$ we raised at the beginning of this section:

$\mathcal{A}_1$: For any $t \geq 1$, if the perturbation scale $r_{t,n}$ is $\Omega\left((m^3/n)^{2^{t-2}}\text{polylog}(n)\right)$, then t steps of Newton is (ϕ, ϵ)-PAU with $\phi \to 0$ under PHAS[7].

$\mathcal{A}_2$: If the number of iterations t satisfies

$$t > T = 1 + \log_2\left(\frac{1+\alpha}{1-3\alpha}\right) \tag{5}$$

where $\alpha = \log(m+1)/\log(n)$, then $\text{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \to 0$ in probability under PHAS.

Before discussing the sharpness of the upper bound obtained in Theorem 7 and Theorem 8, we first provide their implication on the number of Newton steps required for effective unlearning. For now, suppose that the results of the theorems are sharp, in the sense that all $O_p(\cdot)$ cannot be reduced to $o_p(\cdot)$. Then, $\alpha = \log(m+1)/\log(n) > 0$ for $m \geq 1$, and hence by (5) we need at least $t > 1 + \log_2(1) = 1$ Newton steps. This is consistent with our claim at the end of Section 2.3 that **one Newton step is not enough**, even with $m = 1$. In other words, one Newton step will always make GED $\to \infty$ provided the bound in Theorem 8 is sharp.

On the other hand, two Newton steps suffice when $m = O(1)$, because in this case $\alpha = \log(m+1)/\log(n)$ can be arbitrarily small when n is large, in which case $1+\log_2(\frac{1+\alpha}{1-3\alpha}) < 2$. Furthermore, if $m = n^\gamma$ with some $\gamma < \frac{1}{3}$, then $\alpha \approx \gamma$ and the minimal number of step also increases according to (5). If m grows even faster than $n^{1/3}$, it essentially makes $\hat{\boldsymbol{\beta}}$, as the initial point of Newton iteration, out of the quadratic-converging area around the exact solution $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ (see Theorem 11 and 12 in Section 3.3.2).

Figure 1 provides an illustrative evidence for removing a single data point ($\mathcal{M} = \{i\}$), plotting $\ell_i(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\{i\}})$ against $\ell_i(\hat{\boldsymbol{\beta}}_{\setminus\{i\}})$ for $t = 1$(left), $t = 2$(middle) and plotting $\ell_i(\tilde{\boldsymbol{\beta}}^{\text{Guo}}_{\setminus\{i\}})$ (based on Algorithm 1 in Guo et al. (2019)) against $\ell_i(\hat{\boldsymbol{\beta}}_{\setminus\{i\}})$ (right).[8] Ideally, the points should lie on the $y = x$ line (dashed black) suggesting $\ell_i(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\{i\}}) \approx \ell_i(\hat{\boldsymbol{\beta}}_{\setminus\{i\}})$. This is true for two steps (middle), where the required noise is significantly smaller, enabling removal of the intended information while preserving the rest of the model. However, in the left and the right panel the points scatter away from $y = x$ line, which suggests that the amount of noise required for certified unlearning with a single Newton step is too large-it not only erases the targeted information but also corrupts parts of the model that should remain intact. [9]

7. The polylog(n) term here should match the form in Theorem 7, i.e. $[C_1(n)]^{2^{t-1}}\left(\frac{C_2(n)}{2\lambda\nu}\right)^{2^{t-2}}$. To reduce such cumbersomeness in practice, the polylog(n) term in can be replaced by $n^{\epsilon'}$ for arbitrary $\epsilon' > 0$, since $n^{\epsilon'} > \text{polylog}(n)$ for large enough n.
8. We remark that Algorithm 1 in Guo et al. (2019) appears to omit the $\lambda\mathbb{I}$ term in the Hessian. See line 12 of Algorithm 1: $H \leftarrow \sum_{i:i\notin B_1,B_2,\ldots,B_j} \nabla^2\ell(\mathbf{w}^\top\mathbf{x}_i, y_i)$. This point should be checked for accuracy. The figures in this paper use the corrected Hessian that includes the $\lambda\mathbb{I}$ term.
9. In Figure 1, the apparently smaller noise magnitude in required in Guo et al. (2019) compared to our one-step procedure may be attributable to two factors. First, the notions of certifiability are not directly comparable. In particular, Guo et al. (2019) adopts an $(\epsilon_{\text{GUO}}, \delta_{\text{GUO}})$-CR framework; in our comparison we use $\epsilon_{\text{GUO}} = 0.1$ and $\delta_{\text{GUO}} = 10^{-4}$, whereas our simulations are based on a different definition with $\epsilon = 0.1$. Since the underlying certifiability criteria differ, a direct quantitative comparison of the required noise levels is inherently difficult.

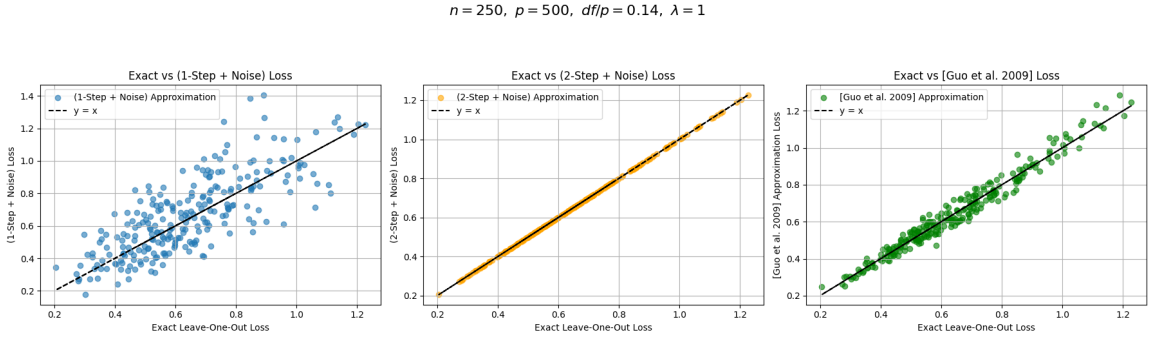

Figure 1: Comparison between $t = 1$ Newton step (left), $t = 2$ steps (right) and the Guo et al. (2019) for ridge logistic model. The quantities $\ell_i(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}})$, $\ell_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})$ and $\ell_i(\tilde{\boldsymbol{\beta}}^{\text{Guo}}_{\setminus\mathcal{M}})$ are plotted on the y and x axis for each $\mathcal{M} = \{i\}$ respectively, i.e. by leaving each observation out separately. The true unknown parameter vector is sampled as $\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p)$, feature vectors as $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_n/n)$, and responses as $y_i \sim \text{Bernoulli}\left(\sigma(\boldsymbol{x}_i^\top \boldsymbol{\beta}^*)\right)$ where $\sigma(.)$ is the sigmoid function.

3.3 Why a single Newton step is insufficient

In this section, we provide more theoretical discussions and results on why one Newton step is not sufficient in high dimensions. This section could also serve as a proof sketch for our main theorems (Theorem 7 and 8). Briefly speaking, for any approximate solution of $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ denoted by $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, if we want to achieve certifiability through adding noise $\boldsymbol{b}$, then both certifiability (Definition 1) and accuracy (Definition 4) are connected to the key quantity: the ℓ_2 error

$$\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2.$$

This relation will be discussed in more details in Section 3.3.1. Upper bounds for the ℓ_2 error of t-step Newton estimators are then provided in 3.3.2.

3.3.1 ℓ_2 error: a key for both certifiability and accuracy

Recall that we quantified certifiability by (ϕ, ϵ)-PAU (Definition 1) and prediction accuracy by GED (Definition 4). The next lemma connects (ϕ, ϵ)-PAU with $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, and its proof can be found in Appendix B.1.

Lemma 9 *Let $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} \in \mathbb{R}^p$ be any two estimators calculated from training data $\mathcal{D}$. Suppose $\boldsymbol{b} \in \mathbb{R}^p$ is a random vector independent of $\mathcal{D}$ and has a density $p_{\boldsymbol{b}}(\boldsymbol{b}) \propto \mathrm{e}^{-\frac{\epsilon}{r}\|\boldsymbol{b}\|}$. Define*

$$\mathcal{X}_r \triangleq \{\mathcal{D} : \max_{|\mathcal{M}| \le m} \|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \le r\},$$

Second, Guo et al. (2019) injects noise at the level of the empirical risk rather than directly perturbing the estimator. This distinction may be advantageous, as it can induce an effectively directional noise distribution at the estimator level, potentially reducing the apparent noise magnitude required to achieve certifiability.

then $\forall$ *measurable set* $\mathcal{T} \subset \mathbb{R}^p$, $\forall |\mathcal{M}| \leq m$,

$$\mathrm{e}^{-\epsilon} < \frac{\mathbb{P}\left(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b} \in \mathcal{T} | \mathcal{D}\right)}{\mathbb{P}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b} \in \mathcal{T} | \mathcal{D})} \leq \mathrm{e}^{\epsilon},$$

if and only if $\mathcal{D} \in \mathcal{X}_r$.

This lemma essentially states that we can make **any** estimator $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ certifiable, as long as we find a high-probability bound r for its ℓ_2 error $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|$ and use r as the scale parameter for the Laplace perturbation $\boldsymbol{b}$. Moreover, it provides a necessary and sufficient condition for (1) in the definition of PAU to hold: the set $\mathcal{X}_r$ precisely characterizes the collection of dataset $\mathcal{D}$ for which $\tilde{\boldsymbol{\beta}}^R_{\setminus\mathcal{M}} := \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$ satisfies (1). So the failure probablity in PAU is exactly $\phi_r = \mathbb{P}(\max_{|\mathcal{M}| \leq m} \|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 > r)$.

Remark 10 *Note that requiring* ϕ_r *to be exactly zero—as was done, for instance in Guo et al. (2019), means* r *is the worst-case bound for* $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|$ *over all possible realization of* $\mathcal{D}$*, which would necessitate letting* $r \to \infty$ *under the PHAS condition, rendering the outcome of the machine unlearning procedure useless. This is one of the reasons we proposed the notion of* (ϕ, ϵ)*-PAU in this paper.*

Lemma 9 establishes the connection between (ϕ, ϵ)-PAU and the ℓ_2 error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Next we connect $\mathrm{GED}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, \tilde{\boldsymbol{\beta}}^R_{\setminus\mathcal{M}})$ with $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ by heuristically showing that[10]

$$\mathrm{GED} = \Theta_p(\frac{\sqrt{p}}{\epsilon}\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2).$$

Recall Definition 4:

$$\mathrm{GED}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, \tilde{\boldsymbol{\beta}}^R_{\setminus\mathcal{M}}) = \mathbb{E}\left(|\ell(y_0|\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \ell(y_0|\boldsymbol{x}_0^\top(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}))| \Big| \mathcal{D}\right).$$

By mean-value theorem,

$$|\ell(y_0|\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \ell(y_0|\boldsymbol{x}_0^\top(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}))| = |\dot{\ell}_0(\boldsymbol{\xi})\boldsymbol{x}_0^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \boldsymbol{b})|,$$

for some $\boldsymbol{\xi}$. In many cases, it can be shown that $|\dot{\ell}_0(\boldsymbol{\xi})| = \Theta_p(1)$ and for now we omit this term. Denote $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \boldsymbol{b}$ by $\boldsymbol{v}$ for brevity, and it suffices to show that $\max_{|\mathcal{M}| \leq m} |\boldsymbol{x}_0^\top \boldsymbol{v}| = \Theta_p(\frac{\sqrt{p}}{\epsilon}\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2)$. Assume the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}_0$ has eigenvalues of the order $\Theta_p(p^{-1})$ (see Assumption B1 and the discussions thereafter for justification on this scaling). Observe that $\boldsymbol{x}_0$ is independent from $\boldsymbol{v}$, so $\mathbb{E}[|\boldsymbol{x}_0^\top \boldsymbol{v}||\mathcal{D}, \boldsymbol{b}] \simeq \frac{1}{\sqrt{p}}\|\boldsymbol{v}\|_2$, and by triangular inequality we can bound $\|\boldsymbol{v}\|_2$ by

$$\|\boldsymbol{b}\|_2 - \|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \leq \|\boldsymbol{v}\|_2 \leq \|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 + \|\boldsymbol{b}\|_2.$$

10. Technically Theorem 8 is stronger by bounding $\max_{|\mathcal{M}| \leq m}$ GED, but the maximum doesn not affect the order of quantities concerned in the discussion here, so we omit it for now.

Actually both the upper and lower bounds are dominated by $\|\boldsymbol{b}\|_2$. This is because we choose $p(\boldsymbol{b}) \propto \mathrm{e}^{-\frac{\epsilon}{r}\|\boldsymbol{b}\|}$, and set r to be a high-probability bound for $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, according to Lemma 9. It can be shown that (Lemma 16) $\|\boldsymbol{b}\|_2 \sim \Gamma(p, \frac{\epsilon}{r})$, and it concentrates around its mean: $\|\boldsymbol{b}\|_2 \simeq \frac{pr}{\epsilon}$. So both the upper and lower bounds are $\Theta_p(\frac{pr}{\epsilon} \pm r) = \Theta_p(\frac{pr}{\epsilon})$. So we get the rate we claimed:

$$\mathrm{GED} = \Theta_p(\frac{1}{\sqrt{p}}\|\boldsymbol{v}\|_2) = \Theta_p(\frac{\sqrt{p}}{\epsilon}\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2).$$

Therefore we need to ensure that $\max_{|\mathcal{M}|\leq m} \|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = o_p(\frac{1}{\sqrt{p}}) = o_p(\frac{1}{\sqrt{n}})$ for a machine unlearning algorithm to be both certifiable and accurate under PHAS. However, this can only be achieved by more than one Newton steps, as will be discussed in the next subsection.

3.3.2 Understanding the ℓ_2 error $\|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|$

The discussion of the previous section implied that, if $\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}}$ is the outcome of the Newton algorithm after t steps, we can obtain a (ϕ_n, ϵ)-PAU, and accurate (meaning $\max_{|\mathcal{M}|\leq m} \mathrm{GED}(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \to 0$) machine unlearning algorithm by adding Laplace noise, if and only if $\|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = o_p(\frac{1}{\sqrt{n}})$. Hence, in this section we aim to address the following two questions:

- Does $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ satisfy $o_p(\frac{1}{\sqrt{n}})$?
- Does $\|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ satisfy $o_p(\frac{1}{\sqrt{n}})$ for $t \geq 2$?

We will start from $t = 1$ case. Define the set $\mathcal{X}_r^{(t)}$ in the same way as in Lemma 9:

$$\mathcal{X}_r^{(t)} := \{\mathcal{D} : \max_{|\mathcal{M}|\leq m} \|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \leq r\}.$$

Our first result obtains an upper bound for $\max_{|\mathcal{M}|\leq m} \|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$.

Theorem 11 *Under Assumptions A1-A3 and B1-B3, if we set* $r = \sqrt{\frac{m^3}{n}}\mathrm{polylog}(n)$*, then we have*

$$\mathbb{P}(\mathcal{D} \in \mathcal{X}_r^{(1)}) \geq 1 - nq_n - 8n^{1-c} - ne^{-p/2} - 2e^{-p},$$

where q_n *was defined in Assumption B3.*

The proof of this Theorem is prsented in Section B.2.1. The exact form of r, including the constant and polylog(n) terms, can be found in Lemma 26.

According to this theorem, $\max_{|\mathcal{M}|\leq m} \|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O_p\left(\sqrt{\frac{m^3}{n}}\mathrm{polylog}(n)\right)$, but in the last subsection we concluded that we need $\max_{|\mathcal{M}|\leq m} \|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = o_p(\frac{1}{\sqrt{p}})$ for accurate certified unlearning. Even for the case $m = 1$, $\frac{\mathrm{polylog}(n)}{\sqrt{n}} \gg \frac{1}{\sqrt{n}}$. If this bound is indeed unimprovable, then it indicates that one Newton step is not enough. We should note

that Theorem 11 does not show such sharpness of the bound, but numerical results support our findings. Figure 2 left panel plots $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ against p for single deletion ($m = 1$) case, and it indeed suggests a relationship of $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \simeq \frac{1}{\sqrt{n}}$.

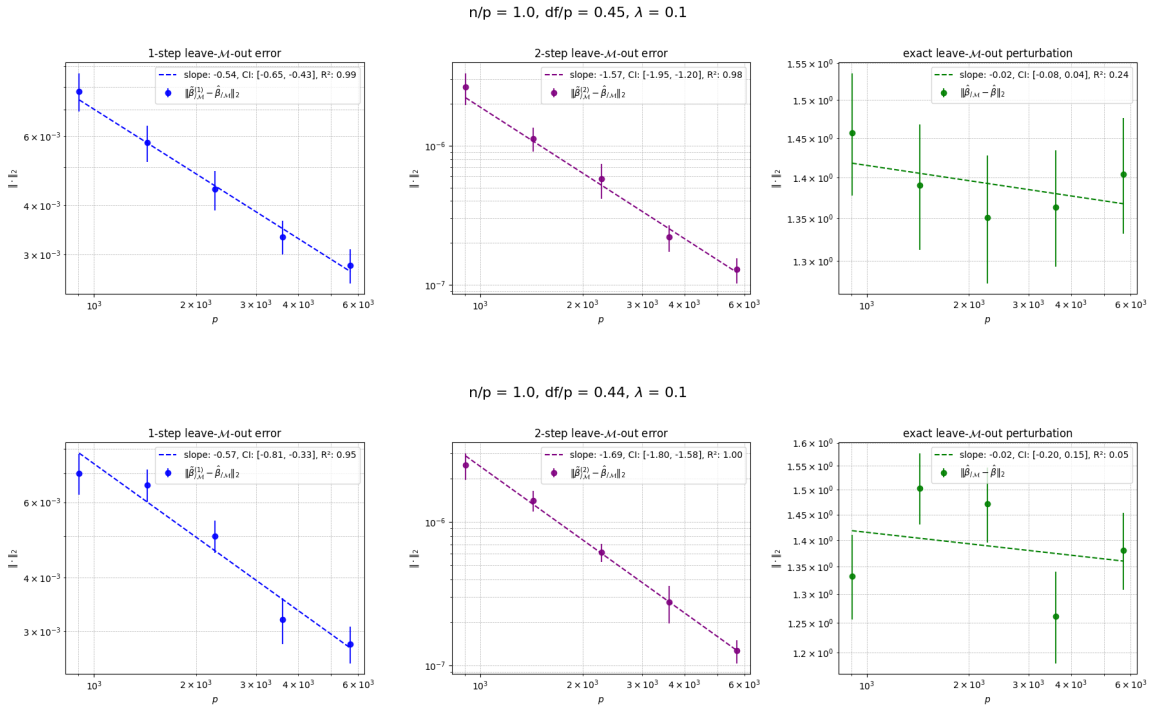

Figure 2: The approximation and exact unlearning error for ridge logistic regression as function of p for $n/p = 1$. Top: dense $\boldsymbol{X}$. Bottom: sparse $\boldsymbol{X}$. Left: The one Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Middle: The two Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Right: The exact unlearning error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$. Details of this simulation are described in Section 5.

Next we study multiple Newton steps:

Theorem 12 *Under Assumptions A1-A3, B1-B3, if* $m = o(n^{1/3})$, *then*

$$\mathbb{P}(\mathcal{D} \in \cap_{t=1}^{\infty} \mathcal{X}_{r_{n,t}}) \geq 1 - nq_n - 8n^{1-c} - ne^{-p/2} - 2e^{-p}$$

for some $r_{n,t} = \left(\frac{m^3}{n}\right)^{2^{t-2}} \text{polylog}(n)$.

The proof can be found in Section B.2.6 in the appendix, where we also provide the exact form of the polylog(n) term as well as constants. According to this theorem,

$$\max_{|\mathcal{M}| \leq m} \|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = o_p\left(\left(\frac{m^3}{n}\right)^{2^{t-2}} \text{polylog}(n)\right).$$

This for instance implies that as long as $m = o(n^{1/6})$,

$$\max_{|\mathcal{M}|\leq m} \|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O_p\left(\frac{m^3 \text{polylog}(n)}{n}\right) = o_p\left(\frac{1}{\sqrt{n}}\right).$$

This confirms that certifiable and accurate machine unlearning can be done with more than one Newton-Step.

Figure 2 displays the accuracy of one-step and two-step Newton approximations (without noise added) to the exact leave-$\mathcal{M}$-out estimator in logistic ridge regression. The empirical results (see the slopes) are consistent with the scalings predicted by Theorem 12: $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O_p(\frac{\text{polylog}(n)}{\sqrt{n}})$ and $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O_p(\frac{\text{polylog}(n)}{n})$. Moreover, the right pannel shows empirically that the exact unlearning error $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ has a constant scale regardless of problem dimensions. Simulation details are given in Section 5.

4. Related Work

4.1 Summary of the existing results

As we discussed earlier, the machine unlearning problem has received significant attention in recent years, both from theoretical and empirical perspectives Nguyen et al. (2022), Suriyakumar and Wilson (2022). Among the existing work, Guo et al. (2019); Sekhari et al. (2021); Neel et al. (2021); Izzo et al. (2021) are most closely related to our contributions, as they focus on theoretical aspects of the machine unlearning algorithms. Therefore, we provide a more detailed comparison between their contributions and ours.

In Guo et al. (2019), the authors introduced the concepts of ϵ-certified and (ϵ, δ)-certified machine unlearning. Our notion of (ϕ, ϵ)-probabilistically certified approximate machine unlearning, introduced in Definition 1, is inspired by the ϵ-certified machine unlearning notion of Guo et al. (2019). However, our definition is more flexible, allowing the unlearning algorithm to be non-private on datasets that occur with low probability. Furthermore, Guo et al. (2019) analyzed the level of Laplacian noise that must be added to the objective function to ensure that the output of a single-step Newton method satisfies either ϵ-certifiability or (ϵ, δ)-certifiability. In these studies, the authors assumed that n is large, while p is fixed.

The work of Sekhari et al. (2021) builds upon and improves the results of Guo et al. (2019) in several directions: (1) they incorporate the dependence on the number of parameters or features p in their analysis, and (2) they introduce a notion of excess risk to evaluate the accuracy of the approximations produced by machine unlearning algorithms. Their main conclusion is that a single Newton step suffices to yield an accurate machine unlearning algorithm— a conclusion that stands in contrast to the message of our paper. We argue that the analysis presented in Sekhari et al. (2021) lacks sharpness, and as a result, the bounds they derive are not useful in many high-dimensional settings. In fact, under the high-dimensional regime considered in our work, many of the bounds in Sekhari et al. (2021) diverge as $n, p \to \infty$. Since a thorough clarification of this point requires several pages, we defer the detailed discussion to Section 4.2.

The authors of Neel et al. (2021) have studied gradient-based methods initialized with the pre-trained models for machine unlearning, and established their theoretical performance—particularly in terms of the number of gradient descent iterations required. The

discussions we present in Section 4.2 can be used for interpreting the results of Neel et al. (2021) in high-dimensional settings as well.

The authors of Izzo et al. (2021) proposed a projection-based update method called projective residual update (PRU), applicable to linear and logistic regression. It reduces the general $O(mp^2)$ time complexity of one Newton step to $O(m^2p)$, as it considers only the projection onto the $m-$dimensional subspace spanned by $\boldsymbol{X}_{\mathcal{M}}$. However, it lacks performance guarantee even in the low dimensional settings.

In parallel with the theoretical advances in machine unlearning, many empirical methods have also been studied, particularly for deep neural network models. A standard approach is to perform the gradient ascent algorithm on a forget set or the gradient descent algorithm on a remaining set (Graves et al., 2021; Goel et al., 2022).

We note that in unlearning complex learning paradigms including large language models, second order methods or their approximations are popularly used. In particular, SOUL in Jia et al. (2024) uses gradient momentum as well as Hessian diagonals in the second order unlearning step, while requiring multiple second order updates. In that work the authors show that SOUL has superior performance over first order methods on several real data experiments, including LLMs.

In contrast, in the current work we show that two Netwon iterations suffice when using the full Hessian matrix on strongly convex regression problems. Our work thus distinguishes from SOUL in a number of ways:

1. We use the full Hessian matrix while SOUL uses diagonal elements for faster computation.

2. We show certifiability of unlearning in strongly convex regression problems, while SOUL focuses on empirical validation on large language models.

3. We show that two full Hessian Newton iterations suffice while SOUL uses multiple updates with partial Hessian.

4. We do not focus on the decrease of accuracy on the forget set as a measure of unlearning. Instead our measure of accuracy is generalization error divergence (GED) which compares the accuracy on new test data when using the unlearned and the retrained estimator.

While standard methods typically rely on either a forget set or a remaining set, Kurmanji et al. (2023) proposed a novel loss function that leverages both datasets. Specifically, they proposed a new loss function that encourages an unlearned model to remain similar to the original on the remaining data, while diverging on the forget set. As an alternative approach, Foster et al. (2024) proposed a training-free machine unlearning algorithm, demonstrating solid performance at limited computational cost. Compared to our work, all these aforementioned approaches have shown promising empirical results, but they rely on heuristics and lack theoretical guarantees on machine unlearning. Along these lines, Pawelczyk et al. (2024) demonstrated that many available machine unlearning algorithms are not effective in removing the effect of poisoned data points across various settings, which calls for many principled research works in this field. The readers may refer to Xu et al. (2023); Li et al. (2025) for a comprehensive survey of machine unlearning.

4.2 Detailed comparison with Sekhari et al. (2021)

To clarify some of the subtleties that influence the analysis in Sekhari et al. (2021), we revisit the assumptions and results of that work within the context of the setting and assumptions outlined in Section 3.1 of our paper. To make our problem similar to the one studied in Sekhari et al. (2021), define:

$$f(\boldsymbol{\beta}; \boldsymbol{x}, y) = \ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}) + \frac{\lambda}{n} r(\boldsymbol{\beta}).$$

We further define

$$\hat{F}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{\beta}; \boldsymbol{x}_i, y_i),$$

and

$$F(\boldsymbol{\beta}) = \mathbb{E}\ell(y|\boldsymbol{x}^\top \boldsymbol{\beta}),$$

where the expected value is with respect to a new sample $(y, \boldsymbol{x})$. Assumption (1) of Sekhari et al. (2021) states that:

Assumption 1 of Sekhari et al. (2021). For any $(y, \boldsymbol{x})$, $f(\boldsymbol{\beta}; \boldsymbol{x}, y)$ as a function of $\boldsymbol{\beta}$, is ν-strongly convex, L-Lipschitz, and M-Hessian Lipschitz, meaning: $\forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$,

- ν strongly convex:

$$f(\boldsymbol{\beta}_2; \boldsymbol{x}, y) \geq f(\boldsymbol{\beta}_1; \boldsymbol{x}, y) + \nabla f(\boldsymbol{\beta}_1; \boldsymbol{x}, y)(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) + \frac{\nu}{2} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2.$$

- Lipschitzness:

$$|f(\boldsymbol{\beta}_2; \boldsymbol{x}, y) - f(\boldsymbol{\beta}_1; \boldsymbol{x}, y)| \leq L \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

- M-Hessian Lipschitzness:

$$\|\nabla^2 f(\boldsymbol{\beta}_1; \boldsymbol{x}, y) - \nabla^2 f(\boldsymbol{\beta}_2; \boldsymbol{x}, y)\|_2 \leq M \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2.$$

Also, Sekhari et al. (2021) considered a slightly relaxed version of certifiability that they call (ϵ, δ)-certifiability which is defined in the following way. Again as before, we present their definition in our notations:

Definition 2 of Sekhari et al. (2021) For all delete requests $\mathcal{M}$ of size at most m, and any $\mathcal{T} \subset \Theta$, the learning algorithm A and unlearning algorithm $\tilde{A}$ satisfy (ϵ, δ)-unlearning property if and only if:

$$\mathbb{P}\left(\tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}) \in \mathcal{T}\right) \leq \mathrm{e}^{\epsilon} \mathbb{P}\left(\tilde{A}(\emptyset, A(\mathcal{D}_{\setminus \mathcal{M}}), T(\mathcal{D}_{\setminus \mathcal{M}}), \boldsymbol{b}) \in \mathcal{T}\right) + \delta,$$

and

$$\mathbb{P}\left(\tilde{A}(\emptyset, A(\mathcal{D}_{\setminus \mathcal{M}}), T(\mathcal{D}_{\setminus \mathcal{M}}), \boldsymbol{b}) \in \mathcal{T}\right) \leq \mathrm{e}^{\epsilon} \mathbb{P}\left(\tilde{A}(\mathcal{D}_{\mathcal{M}}, A(\mathcal{D}), T(\mathcal{D}), \boldsymbol{b}) \in \mathcal{T}\right) + \delta.$$

Based on the assumptions above, Sekhari et al. (2021) has proved the following theorem.

Theorem 13 *Sekhari et al. (2021) Consider a machine unlearning estimate obtained by adding i.i.d. Gaussian noise (with variance specified in Algorithm 1 of the paper) to the estimate $\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}}$ produced by a single step of the Newton method. Under Assumption 1 of Sekhari et al. (2021) mentioned above, and assuming that the elements of the dataset are i.i.d., we have*

1. *Approximation accuracy of a single Newton method:*

$$\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \leq \frac{2ML^2m^2}{\nu^3 n^2}.$$

2. *The unlearning algorithm satisfies (ϵ, δ)-unlearning.*

3. *For any subset $\mathcal{M}$ of size less than or equal to m:*

$$\mathbb{E}|F(\tilde{\boldsymbol{\beta}}^{R,1}_{\setminus\mathcal{M}}) - \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta})| = O\left(\frac{\sqrt{p}Mm^2L^3}{\nu^3 n^2 \epsilon}\sqrt{\log\left(\frac{1}{\epsilon}\right)} + \frac{4mL^2}{\nu n}\right).$$

If one assumes that $L = O(1)$, $M = O(1)$, and $\nu = O(1)$, as is implicitly assumed in most of the conclusions in Sekhari et al. (2021) it seems that a single step of the Newton method is sufficient for ensuring that:

$$\mathbb{E}|F(\tilde{\boldsymbol{\beta}}^{R,1}_{\setminus\mathcal{M}}) - \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta})| \to 0,$$

as $p, n \to \infty$, as long as $m = o(\frac{n}{p^{1/4}})$. This conclusion is in fact mentioned in the abstract of Sekhari et al. (2021). Our claim is that

- One cannot assume that L, M and ν are $O(1)$ in high-dimensional settings. These quantities are, in fact, expected to depend on n, p. It is therefore important to account for such dependencies in the theoretical analysis.

- Once we obtain the correct order of these three parameters, we will notice that the bounds of Sekhari et al. (2021) are not sharp for high-dimensional settings.

In the remainder of this section, we aim to incorporate these considerations into a refined analysis. Below we provide a detailed description of Assumption 1 of Sekhari et al. (2021) mentioned above. To make our discussion clear, similar to Guo et al. (2019) we focus on the ridge regularizer,

$$r(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2.$$

1. Strong convexity assumption: The authors of Sekhari et al. (2021) assume f to be ν-strongly convex in $\boldsymbol{\beta}$, which means the empirical loss function $\hat{F}_n(\boldsymbol{\beta})$ is ν-strongly convex. Note that

$$\nabla^2 f(\boldsymbol{\beta}; \boldsymbol{x}, y) = \ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top + \frac{\lambda}{n}I.$$

   Furthermore, $\ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top$ is a rank-one matrix. Hence, it is straightforward to see that $\nu = O(\frac{1}{n})$.

2. Lipschitzness of the loss function: Below we perform some heuristic calculations to suggest what the order of the Lipschitz constant can be as a function of n, p. By using the mean value theorem

$$\begin{aligned}|\ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}_1) - \ell(y|\boldsymbol{x}^\top\boldsymbol{\beta}_2)| &= |\dot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\xi})\boldsymbol{x}^\top(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)| \\ &\leq |\dot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\xi})|\|\boldsymbol{x}\|\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,\end{aligned}$$

where $\boldsymbol{\xi}$ is a point on the line that connects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

Hence, in order to understand the order of the Lipschitz constant we should understand the orders of $\|\boldsymbol{x}_i\|$ and $|\dot{\ell}(y_i|\boldsymbol{x}_i^\top\boldsymbol{\xi})|$. Since we want the Lipschitz property to hold for every $(\boldsymbol{x}, y)$, using Cauchy-Schwartz inequality doesn't compromise the sharpness as equality can be attained for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 + t\boldsymbol{x}$ for some $t \in \mathbb{R}$.

Note that according to Assumption B1 we conclude that $\|\boldsymbol{x}_i\| = O_p(1)$. Moreover, it can be shown that $|\dot{\ell}(y_i|\boldsymbol{x}_i^\top\boldsymbol{\xi})| = O_p(1)$ (details can be found in Appendix 28). Hence, assuming that the Lipschitz constant does not grow is an acceptable assumption. So far, we have ignored the regularizer. Note that

$$\left|\frac{\lambda}{n}\|\boldsymbol{\beta}_1\|^2 - \frac{\lambda}{n}\|\boldsymbol{\beta}_2\|^2\right| \leq \frac{\lambda}{n}\|\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\|\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|.$$

If we assume that each elements of $\boldsymbol{\beta}$ is bounded by a constant, then $\frac{\lambda}{n}\|\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2\| = O(\frac{\sqrt{p}}{n})$. Hense even under PHAS, we can assume that the Lipschitz constant of $f(\boldsymbol{\beta}; y, \boldsymbol{x})$ remains $O(1)$ for all values of $(y, \boldsymbol{x}, \boldsymbol{\beta})$ of interest.

3. Hessian-Lipschitzness of the loss function: By straightforward calculations we obtain that

$$\nabla^2 f(\boldsymbol{\beta}; y, \boldsymbol{x}) = \ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta})\boldsymbol{x}\boldsymbol{x}^\top + \frac{\lambda}{n}I.$$

Then, we have

$$\begin{aligned}\|\nabla^2 f(\boldsymbol{\beta}_1; y, \boldsymbol{x}) - \nabla^2 f(\boldsymbol{\beta}_2; y, \boldsymbol{x})\| &= \|\ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta}_1)\boldsymbol{x}\boldsymbol{x}^\top - \ddot{\ell}(y|\boldsymbol{x}^\top\boldsymbol{\beta}_2)\boldsymbol{x}\boldsymbol{x}^\top\| \\ &\simeq |\dddot{\ell}(y|\xi)|\|\boldsymbol{x}\|)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,\end{aligned}$$

so it is acceptable to assume f is $O(1)$-Hessian-Lipschitz.

Putting all the scalings above together, we have:

$$\nu = O\left(\frac{1}{n}\right)\ ;\ M = O(1)\,;\ L = O(1).$$

Therefore the bound in Theorem 13 becomes

$$\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \leq \frac{2ML^2m^2n}{\nu^3} = O(m^2n).$$

In contrast, our Theorem 11 shows that

$$\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O\left(\sqrt{\frac{m^3}{n}}\right),$$

which goes to zero even when m growns with n slowly enough.

Similarly, the excess risk of Theorem 13 becomes:

$$\mathbb{E}(F(\tilde{\boldsymbol{\beta}}^{R,1}_{\setminus\mathcal{M}}) - \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta})) = O\left(\frac{\sqrt{p}m^2 n}{\epsilon}\sqrt{\log\left(\frac{1}{\epsilon}\right)}\right).$$

Note that the bounds on the excess risk is proportional to $n^{3/2}$ even when we set $m = 1$, hence it does not provide useful information about the accuracy of the approximations. We should mention that in this paper we did not work with the excess risk, and instead we worked with the GED (Definition 4), since the excess risk does not converge to zero in high dimensional R-ERM even for exact unlearning $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$ without perturbation. To see this, consider a simple model of linear regression with ℓ_2 loss, and $y_i|\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}_i^\top\boldsymbol{\beta}^*, \sigma^2)$, $\boldsymbol{x}_i \sim \mathcal{N}(0, \frac{1}{n}\mathbb{I}_p)$. Then

$$\begin{aligned} f(y|\boldsymbol{x}^\top\boldsymbol{\beta}) &= (y - \boldsymbol{x}^\top\boldsymbol{\beta})^2 \\ F(\boldsymbol{\beta}) &= \mathbb{E}f(y|\boldsymbol{x}^\top\boldsymbol{\beta}) = \sigma^2 + \frac{1}{n}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 \\ \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta}) &= \sigma^2 \\ \mathbb{E}|F(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \min_{\boldsymbol{\beta}} F(\boldsymbol{\beta})| &= \frac{1}{n}\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \boldsymbol{\beta}^*\|^2. \end{aligned}$$

It is known that in high dimensions, for most R-ERM estimators including the MLE, $\frac{1}{n}\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \boldsymbol{\beta}^*\|^2 \to \alpha_* > 0$ (e.g. Donoho et al. (2011); Wang et al. (2020); Thrampoulidis et al. (2018)), so in general the excess risk does not converge to zero under PHAS. In contrast, the Generalization Error Divergence (GED) we defined still converges to 0 under PHAS by Theorem 8.

4.3 Literature on approximate leave-one-out cross validation

Another line of work related to our paper focuses on efficient approximations of the leave-one-out cross-validation (LO) estimate of the risk. LO is widely known to provide an accurate estimate of out-of-sample prediction error (Rahnama Rad et al. (2020)). However, a major limitation of LO is its high computational cost. As a result, recent studies have explored methods for approximating LO more efficiently (Beirami et al., 2017; Stephenson and Broderick, 2020; Rahnama Rad and Maleki, 2020; Giordano et al., 2019b,a; Wang et al., 2018; Rahnama Rad et al., 2020; Patil et al., 2021, 2022).

Among this body of work, the contributions of Rahnama Rad and Maleki (2020); Rahnama Rad et al. (2020) are most closely related to our own. Specifically, Rahnama Rad and Maleki (2020) demonstrated that approximating the LO solution using a single Newton step yields an estimate that falls within the statistical error of the true LO estimate.

While there are some technical parallels, our work differs from theirs in several key ways: (1) The criteria considered in this paper differ fundamentally from those in Rahnama Rad and Maleki (2020), where the focus is solely on risk estimation and not on privacy-related concerns. Consequently, the notion of certifiability, which is central to our analysis, was not considered in that line of work. (2) We consider a more general setting involving the removal

of m data points, rather than a single leave-one-out sample. (3) As we show in Theorem 8, due to the stricter requirements of our certifiability criterion, a single Newton step is no longer sufficient, and multiple steps are necessary to achieve a reliable approximation.

5. Numerical Experiments

We present numerical experiments to test our theoretical findings regarding the perturbation scale $r_{t,n}$, as well as the accuracy of the one-step and two-step Newton approximations, as functions of n, p, and m.

In all numerical experiments in this paper, we set $\epsilon = 0.1$, which guarantees that the certifiability condition in Lemma 9 is satisfied.

We emphasize that our notion of certifiability is a composite condition: in addition to the likelihood-based requirement, it also demands that the GED vanish, ensuring that the perturbed estimator remains accurate after noise injection. A central objective of our simulations is to assess whether the noise level can be chosen sufficiently small so that the GED remains negligible. This is precisely what our numerical results demonstrate.

In Section 5.1 we numerically study the scaling behavior of the errors of the one-step Newton $\|\tilde{\boldsymbol{\beta}}^{(1)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, two-step Newton $\|\tilde{\boldsymbol{\beta}}^{(2)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, and $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$ with respect to p (and n) for a fixed $m = 1$.

In Section 5.2, we numerically examine how the errors $\|\tilde{\boldsymbol{\beta}}^{(1)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, $\|\tilde{\boldsymbol{\beta}}^{(2)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, and $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$ scale with m for fixed (p, n). Both synthetic and real datasets are considered.

In Section 5.3, we evaluate our certified unlearning procedure by adding the required noise to obscure the contribution of $\mathcal{D}_{\setminus\mathcal{M}}$ in the estimate $\tilde{\boldsymbol{\beta}}^{(t)}$.

5.1 Scaling of l_2 errors of Newton estimators

As for the synthetic data, we randomly generate the true unknown parameter vector $\boldsymbol{\beta}^* \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p)$.

- **Dense design matrix.** Each row $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_p/n)$.
- **Sparse design matrix.** Each row has ρp nonzero entries (with $\rho = 0.1$). The nonzero coordinates are chosen uniformly at random, and each selected entry is drawn independently from a zero-mean Gaussian distribution with variance $1/(\rho n)$.

Both constructions satisfy

$$\mathrm{var}(\boldsymbol{x}^\top \boldsymbol{\beta}^*) = \frac{p}{n},$$

and are consistent with the finite-signal-to-noise, high-dimensional regime considered in this paper. We sample the responses as

$$y_i \sim \text{Bernoulli}\left(\sigma(\boldsymbol{x}^\top \boldsymbol{\beta}^*)\right),$$

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

We used 100 MCMC samples to compute means and standard errors. The normalized degrees of freedom

$$\mathrm{df}/p \;:=\; \mathrm{tr}(\mathbf{H})/p \tag{6}$$

where the generalized hat matrix is defined as

$$\boldsymbol{H} := \boldsymbol{X}\boldsymbol{G}(\hat{\boldsymbol{\beta}})^{-1}\boldsymbol{X}^\top \mathrm{diag}[\ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}})].$$

In all the figures in this section, the plots are on a log-log scale, the error bars denote standard errors across 100 trials, and dashed lines represent least squares fits in log-log scale, with the slope, 95% confidence interval, and R^2 displayed in each legend. The code for reproducing our experimental results is available at [project repository link].

In Figures 2, 3, and 4 we empirically examine the scaling behavior of the errors of one-step Newton $\|\tilde{\boldsymbol{\beta}}^{(1)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, two-step Newton $\|\tilde{\boldsymbol{\beta}}^{(2)}_{/\mathcal{M}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$, and $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$ with respect to $p \in \{904, 1433, 2271, 3600, 5705\}$[11] for $m = 1$ and $n/p \in \{0.5, 1, 2\}$. These figures present a comprehensive three-panel log-log plot analyzing the scaling behavior (in terms of p and equivalently n) of different unlearning approximation methods in logistic ridge regression with $\lambda = 0.1$. The first row uses dense design matrices, and the second row uses sparse design matrices. Each panel displays error measurements versus dimension p, and fitted regression lines showing the power-law relationships. The left panel examines the 1-step Newton approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, measuring the difference between the single Newton step unlearned model and the exact unlearned model. The middle panel shows the 2-step Newton approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, quantifying how well the two-step approximation recovers the true unlearned parameters. The right panel presents the baseline perturbation magnitude $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, measuring the difference between the original full-data model and the exact leave-$\mathcal{M}$-out model, which represents the fundamental scale of parameter changes when removing data points. Each panel includes a least squares regression line, with the slope coefficients, confidence intervals, and R^2 values displayed in the legends to characterize the power-law scaling relationships. Based on these slopes, we can conjecture the following scalings (for fixed m):

$$\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \;=\; O\left(n^{-0.5}\right), \tag{7}$$

$$\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \;=\; O\left(n^{-1.5}\right). \tag{8}$$

We should emphasize that (7) is consistent with the conclusion of Theorem 11, which suggests the sharpness of Theorem 11 in term of n. However, (8) conjectures rate $n^{-1.5}$ for $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$, while Theorem 12 has proved n^{-1} for this quantity.

The nature of this gap is very interesting yet not fully understood currently. One conjecture is that the extra factor of $O(n^{-0.5})$ observed from numerical experiments might be applicable to broader model class beyond logistic. Taking a single removal datapoint as an example, recall that by Theorem 3.5, the first Newton step **does** reduce the l_2 error from $||\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus i}|| = O_p(1)$ to $O_p(n^{-0.5})$. This $O(n^{-0.5})$ actually comes from event F_5 (equation 13,

11. This is an equispace grid in the logscale.

appendix B.2.1), which essentially states that

$$\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}\|_{2,\infty} = O_p(\sqrt{\frac{m}{n}}).$$

The key to this $O(n^{-0.5})$ factor here is the independence between $\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})$ and $\boldsymbol{X}_{\mathcal{M}}$, since they are functions of $\mathcal{D}_{\setminus\mathcal{M}}$ and $\mathcal{D}_{\setminus\mathcal{M}}$ respectively.

For $t \geq 2$ Newton steps, however, we didn't use the inf norm method, but an l_2 norm bound, resulting in the quadratic convergence in Lemma B.6 without an extra $O(n^{-0.5})$ factor. However, if we could establish the following results:

$$\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t+1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_\infty \simeq \|\boldsymbol{X}_{\setminus\mathcal{M}}(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t+1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_\infty \lesssim \frac{1}{\sqrt{n}}\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t+1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2,$$

then we could use the same l_∞ norm trick used in the first Newton step for all subsequent steps and the extra $n^{-0.5}$ factors would appear. But establishing the result above might need additional assumptions (e.g. ruling out very sparse signals), and we would like to explore this in future research.

In addition, there is no clear trend of $\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{/\mathcal{M}}\|_2$ against n in the right panels, which is also consistent with our theory (Equation (16), where the bound depends only on m but not on n, p).

5.2 Impact of $m = |\mathcal{M}|$ on the perturbation

Our next set of synthetic and real data experiment aims to fix n, p and explore the dependence of $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ on m, the number of samples that are removed.

As for the synthetic data, we study the performance of the ridge-regularized logistic regression, under the setting $n = 1800$, $p = 2000$, and $\lambda = 1$, leading to $\frac{\mathrm{df}}{p} = 0.16$.

As for the real data experiment, we used the IMDB movie-review dataset for binary classification and evaluated ridge-regularized logistic regression under the setting $n = 25{,}000$, $p = 12{,}610$, and $\lambda = 100$, yielding $\mathrm{df}/p = 0.12$. The reviews were converted into features via a bag-of-words representation, producing an extremely sparse design matrix. This example highlights that our theoretical predictions appear to extend beyond the original assumptions, since the method continues to perform well even in this highly sparse real-data regime.

In Figures 5 and 6, we consider $m \in \{1, 3, 7, 14, 27, 54, 105, 205, 400\}$, and calculate the following quantities: $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ (left panel), $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(2)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$ (middle panel), $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$ (right panel). As is clear from the three graphs, the estimates of $\log(\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2)$, $\log(\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(2)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2)$ and $\log(\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2)$ show linear behavior in terms of m. Based on these slopes (as presented in the legends of Figures 5 and 6) we can conjecture the following scalings (for fixed p):

$$\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O(m) \tag{9}$$

$$\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(2)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O(m^2) \tag{10}$$

$$\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O(\sqrt{m}). \tag{11}$$

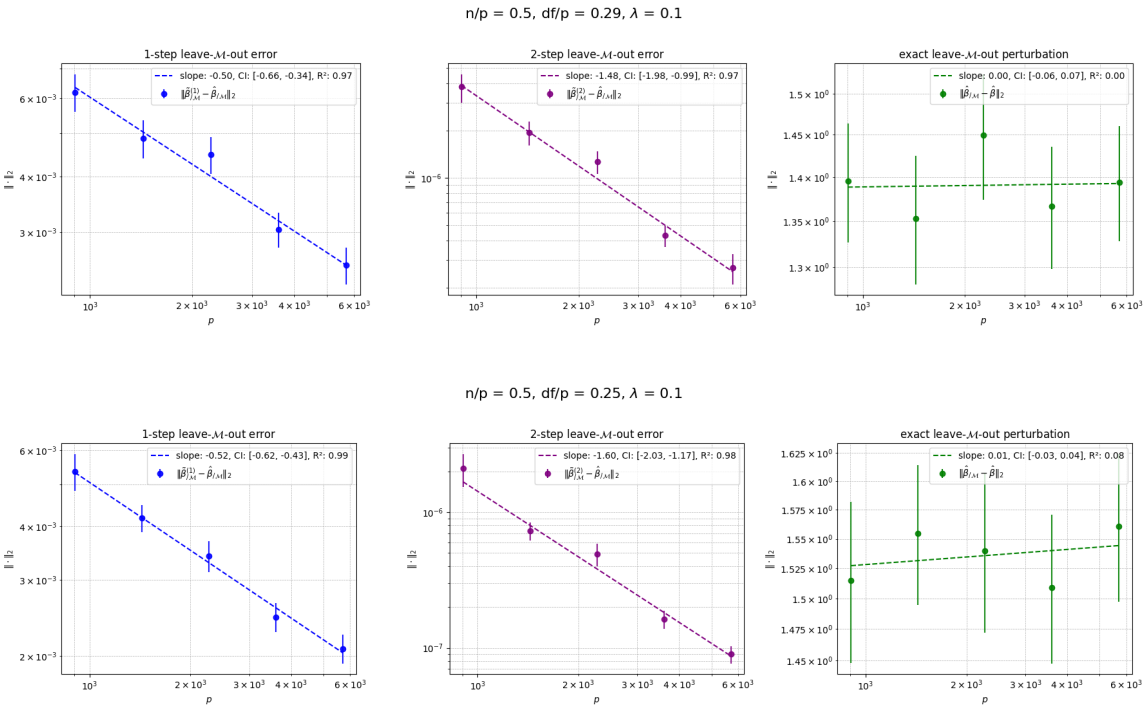

Figure 3: The approximation and exact unlearning error for ridge logistic regression as function of p for $n/p = 0.5$. Top: dense $\boldsymbol{X}$. Buttom: sparse $\boldsymbol{X}$. Left: The one Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Middle: The two Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Right: The exact unlearning error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$.

Note that among the conjectures above, (11) is consistent with our theory (Lemma 17 in conjunction with Equation (16)). However, we obtained in Theorem 11 that $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O(m^{1.5})$, in contrast with the $O(m)$ rate in (9). Similarly, Theorem 12 suggests $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O(m^3)$ whereas (10) suggests $O(m^2)$ rate for the same quantity. Again, this discrepancy is left for future study.

5.3 Performance of the certified Newton approximation plus noise

As previously discussed, in certified machine unlearning, noise is injected to obscure the remainder of $\mathcal{D}_{\setminus\mathcal{M}}$ in the estimate $\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}}$. In the algorithms we have studied in this paper, noise is added via a random vector $\boldsymbol{b}_t \sim \text{Gamma}(p, \epsilon/r_{t,n})$, where $\epsilon = 0.1$ and $r_{t,n}$ is an upper bound for

$$\max_{|\mathcal{M}|\le m} \left\| \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} \right\|_2 .$$

However, the implicit constants in Theorem 11 and 12 can sometimes be intractable or unfavorably large. On the other hand, computing $r_{t,n}$ requires evaluating all $\binom{n}{|\mathcal{M}|}$ possible

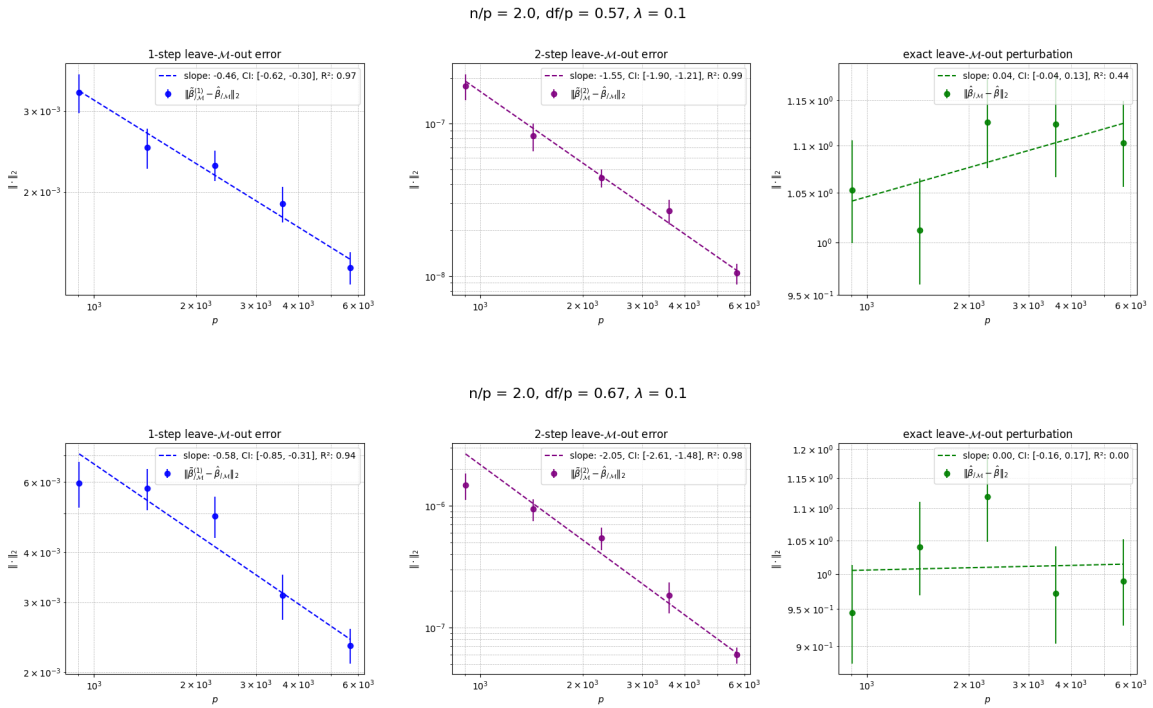

Figure 4: The approximation and exact unlearning error for ridge logistic regression as function of p for $n/p = 2$. Top: dense $\boldsymbol{X}$. Buttom: sparse $\boldsymbol{X}$. Left: The one Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Middle: The two Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Right: The exact unlearning error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$.

subsets, which is computationally infeasible even for moderate n. Hence, instead of exhaustively enumerating all $\binom{n}{|\mathcal{M}|}$ configurations, we select a random subset of size $m_0 = 100$ and compute the maximum $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}}\|_2$ over this subset. To approximate the global maximum, we rescale the result by a factor of $\log\binom{n}{|\mathcal{M}|} / \log m_0$.

In Figures 8, 9, and 7, we empirically investigate the scaling behavior of three quantities: the absolute difference between the exact loss and the single-step Newton approximation (first column), the absolute difference between the exact loss and the single-step Newton approximation plus noise (second column), and the absolute difference between the exact loss and the two-step Newton approximation plus noise (third column), with respect to $p \in \{904, 1433, 2271, 3600, 5705\}$ for $|\mathcal{M}| = m = 1$ and $n/p \in \{0.5, 1, 2\}$. The first row of each figure shows in-sample results, while the second row presents out-of-sample behavior. Together, these figures provide a comprehensive set of three-panel log-log plots that illustrate how different unlearning approximation methods scale (with respect to p, and equivalently n) in logistic ridge regression with $\lambda = 0.1$. As in all the figures in this section, the plots are on a log-log scale, the scattered points are 100 distinct datapoint to be

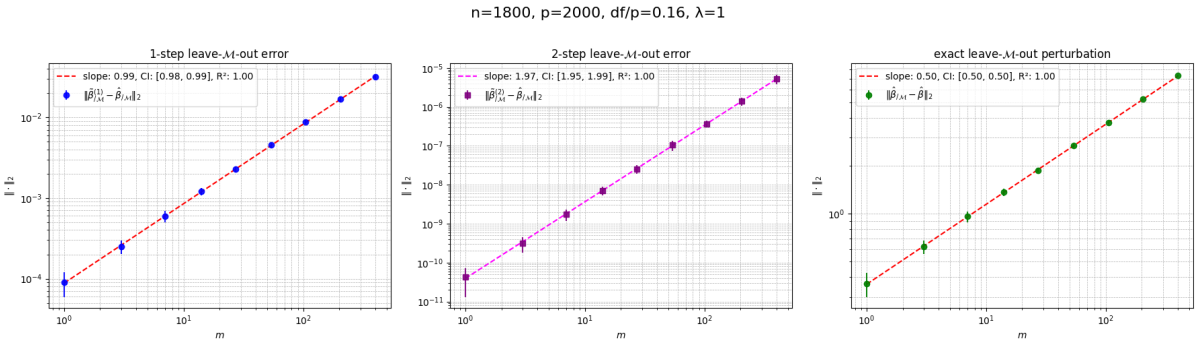

Figure 5: The approximation and exact unlearning error error for ridge logistic regression as function of $m = |\mathcal{M}|$. Left: The one Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Middle: The two Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Right: The exact unlearning error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$.

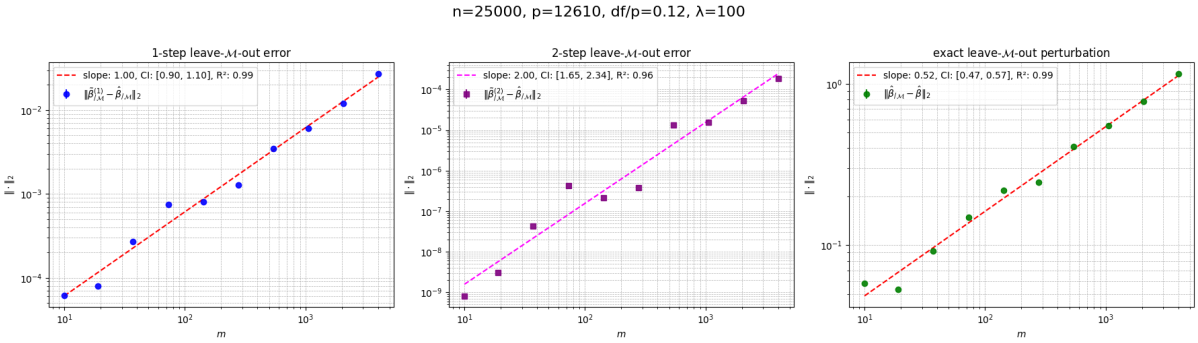

Figure 6: The approximation and exact unlearning error error for ridge logistic regression as function of $m = |\mathcal{M}|$ for the IMDB dataset. The design matrix is constructed using a bag-of-words representation, resulting in an extremely sparse matrix. Left: The one Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Middle: The two Newton step approximation error $\|\tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2$. Right: The exact unlearning error $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}\|_2$.

unlearned, and dashed lines represent least-squares fits in log-log scale, with the slope, 95% confidence interval, and R^2 displayed in each legend.

The out-sample error is defined as

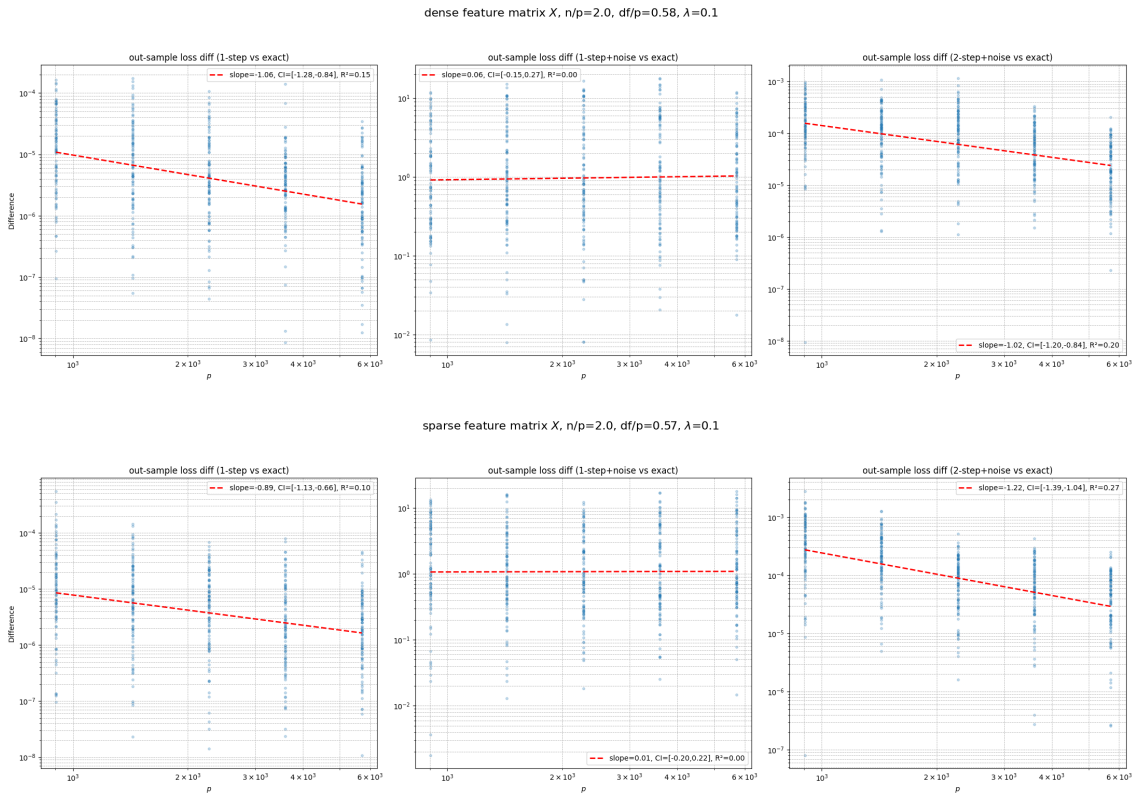

Figure 7: Comparison of the loss of the exactly unlearned model with its first-order (noiseless and noisy) and second-order Newton-step approximations in ridge logistic regression, plotted as a function of p under the setting $n = 2p$. The top panel corresponds to the *dense* design matrix X, and the bottom panel corresponds to the *sparse* design matrix X. Within each panel: Left shows the loss of the absolute error between the exact unlearned model and the one-step Newton approximation; middle shows the absolute error between the exact unlearned model and the one-step Newton approximation plus noise; right shows the absolute error between the exact unlearned model and the two-step Newton approximation plus noise.

- the absolute error between the loss of the exact unlearned model and the one-step Newton approximation (uncertified) $\left|\ell\left(y_0 \mid \boldsymbol{x}_0^\top \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)}\right) - \ell\left(y_0 \mid \boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\right)\right|$; the first column of Figures 8, 9, and 7

- the absolute error between the loss of the exact unlearned model and the one-step Newton approximation plus noise $\left|\ell\left(y_0 \mid \boldsymbol{x}_0^\top (\boldsymbol{b}_1 + \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)})\right) - \ell\left(y_0 \mid \boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\right)\right|$; the second column of Figures 8, 9, and 7,

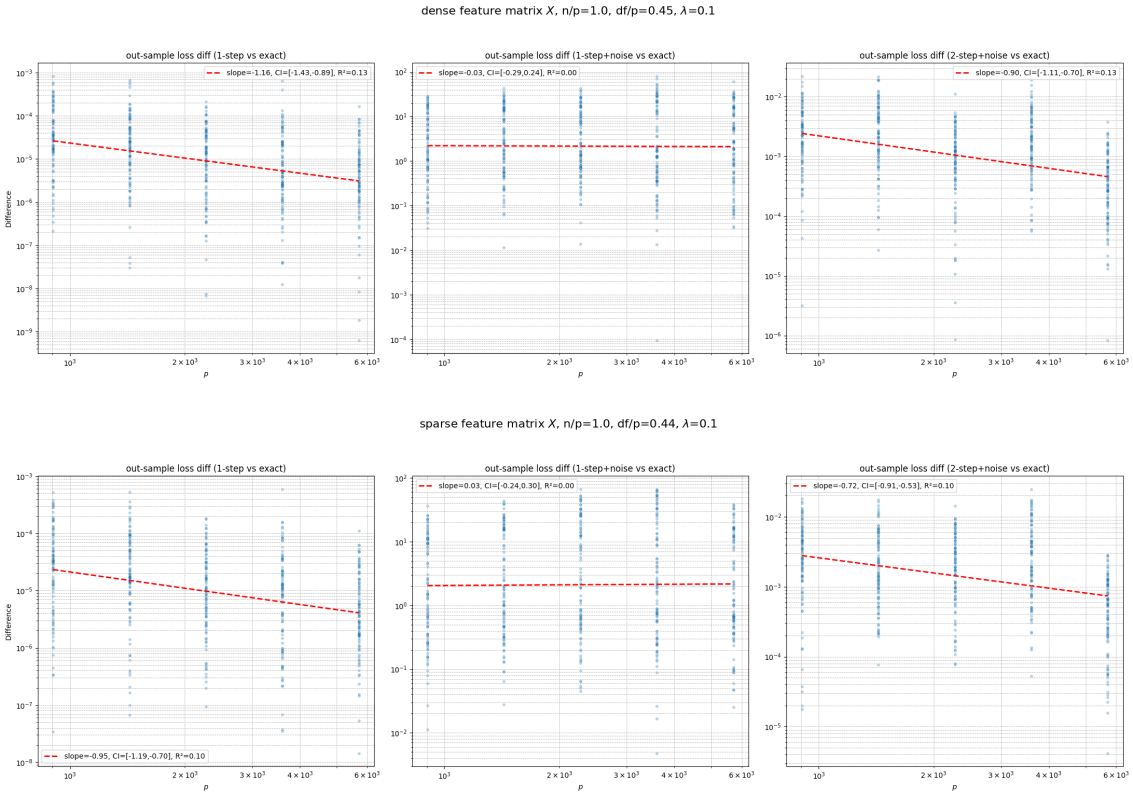

Figure 8: Comparison of the loss of the exactly unlearned model with its first-order (noiseless and noisy) and second-order Newton-step approximations in ridge logistic regression, plotted as a function of p under the setting $n = p$. The top panel corresponds to the *dense* design matrix X, and the bottom panel corresponds to the *sparse* design matrix X. Within each panel: Left shows the loss of the absolute error between the exact unlearned model and the one-step Newton approximation; middle shows the absolute error between the exact unlearned model and the one-step Newton approximation plus noise; right shows the absolute error between the exact unlearned model and the two-step Newton approximation plus noise.

- the absolute error between the loss of the exact unlearned model and the two-step Newton approximation plus noise $\left|\ell\left(y_0 \mid \boldsymbol{x}_0^\top(\boldsymbol{b}_2 + \tilde{\boldsymbol{\beta}}^{(2)}_{\setminus\mathcal{M}})\right) - \ell\left(y_0 \mid \boldsymbol{x}_0^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\right)\right|$; the third column of Figures 8, 9, and 7,

for $i \in \mathcal{M}$, where $\boldsymbol{x}_0$ is a dense or sparse Gaussian vector (as described at the beginning of this section) and $y_0 \sim \text{Binomial}\left(\sigma(\boldsymbol{x}_0^\top \boldsymbol{\beta}^*)\right)$ is a new unseen data point.

Figures 8, 9, and 7 demonstrate that the amount of noise required for certified unlearning in the single-step Newton model is so large that it not only removes the targeted information but also adversely affects parts of the model that should be preserved.

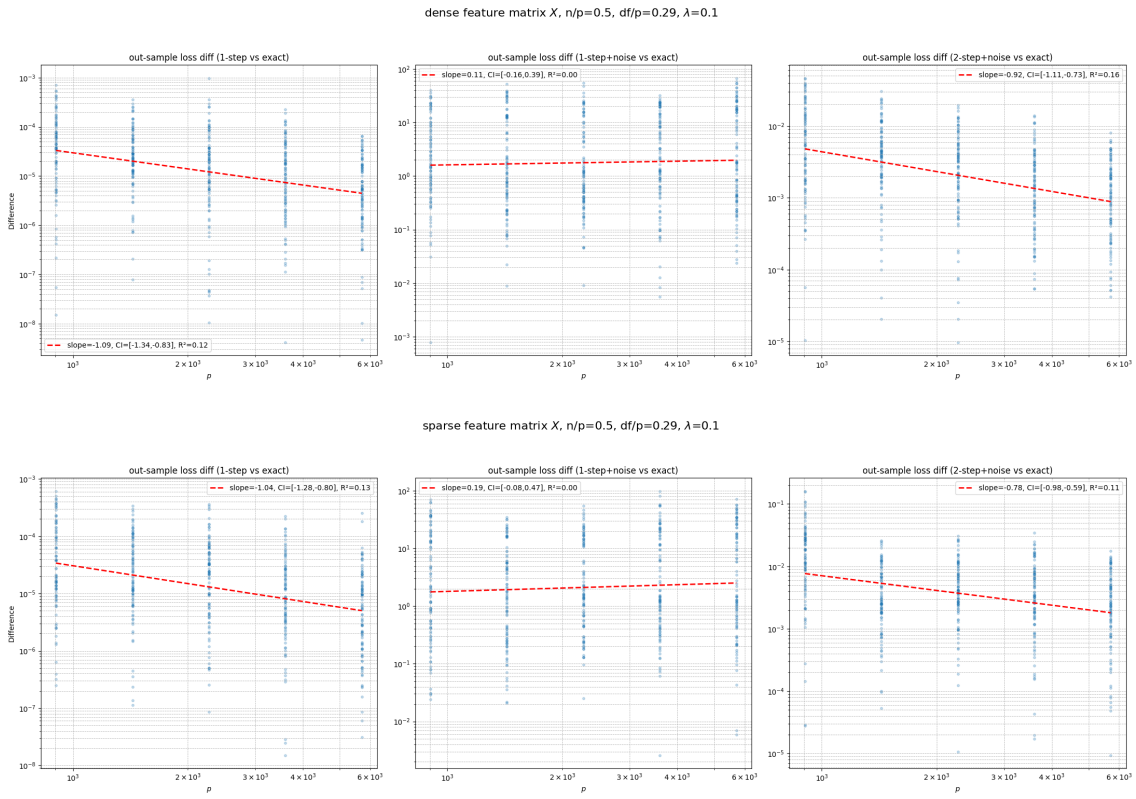

Figure 9: Comparison of the loss of the exactly unlearned model with its first-order (noiseless and noisy) and second-order Newton-step approximations in ridge logistic regression, plotted as a function of p under the setting $n = 0.5p$. The top panel corresponds to the *dense* design matrix X, and the bottom panel corresponds to the *sparse* design matrix X. Within each panel: Left shows the loss of the absolute error between the exact unlearned model and the one-step Newton approximation; middle shows the absolute error between the exact unlearned model and the one-step Newton approximation plus noise; right shows the absolute error between the exact unlearned model and the two-step Newton approximation plus noise.

In contrast, the third column shows that the error in the loss introduced by the two-step Newton model decreases "approximately" as a power law with the model dimension p (and equivalently n). The two-step Newton model requires significantly less noise, enabling it to effectively remove only the targeted information while preserving the rest of the model's learned patterns.

6. Conclusion

In this paper, we introduced a framework for analyzing the performance of machine unlearning algorithms in high-dimensional settings. Focusing on algorithms that are based

on a small number of Newton updates, we demonstrated that the behavior of certifiability and accuracy in high dimensions is significantly more nuanced than in low-dimensional regimes. For example, our theoretical and empirical results show that one-step Newton-based unlearning may fail to provide sufficient accuracy for certified unlearning. However, performing just one additional Newton step can lead to successful unlearning—provided the unlearning set is not too large. Exploring alternative certifiability conditions and other machine unlearning algorithms under the high-dimensional settings is an important direction for future research, aimed at developing a more comprehensive understanding of the complexities inherent in high-dimensional settings.

Acknowledgments

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

References

R. J. Adler and J. E. Taylor. *Random Fields and Geometry*. Springer New York, NY, 2007.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*, 2013.

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

M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory*, 57(2):764–785, 2011.

M. Bayati and A. Montanari. The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory*, 58(4):1997–2017, 2012.

A. Beirami, M. Razaviyayn, S. Shahrampour, and V. Tarokh. On optimal generalizability in parametric learning. *Advances in neural information processing systems*, 30, 2017.

L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

M. Celentano and A. Montanari. Correlation adjusted debiased lasso: debiasing the lasso with inaccurate covariate model. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1455–1482, 2024.

S. Chatterjee. A new perspective on least squares under convex constraint. *The Annals of Statistics*, 42(6):2340–2381, 2014.

M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert, and Y. Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911, 2021.

V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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

D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields*, 166(3-4):935–969, 2016.

D. L. Donoho, A. Maleki, and A. Montanari. Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci.*, 106(45):18914–18919, Sep. 2009.

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

R. Dudeja, Y. M. Lu, and S. Sen. Universality of approximate message passing with semi-random matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

C. Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.

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

Z. Fan. Approximate message passing algorithms for rotationally invariant matrices. *The Annals of Statistics*, 50(1):197–224, 2022.

J. Foster, S. Schoepf, and A. Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12043–12051, 2024.

D. Fourdrinier, W. E. Strawderman, and M. T. Wells. *Shrinkage estimation*. Springer, 2018.

R. Giordano, M. I. Jordan, and T. Broderick. A higher-order swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*, 2019a.

R. Giordano, W. Stephenson, R. Liu, M. Jordan, and T. Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019b.

S. Goel, A. Prabhu, A. Sanyal, S.-N. Lim, P. Torr, and P. Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.

L. Graves, V. Nagisetty, and V. Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021.

Z. Izzo, M. Anne Smart, K. Chaudhuri, and J. Zou. Approximate data deletion from machine learning models. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/izzo21a.html`.

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

J. Jia, Y. Zhang, Y. Zhang, J. Liu, B. Runwal, J. Diffenderfer, B. Kailkhura, and S. Liu. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*, 2024.

N. E. Karoui and E. Purdom. Can we trust the bootstrap in high-dimension? *arXiv preprint arXiv:1608.00696*, 2016.

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012a.

F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mechanics: Theory and Experiment*, 2012(08):P08009, 2012b.

M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

N. Li, C. Zhou, Y. Gao, H. Chen, Z. Zhang, B. Kuang, and A. Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2025.

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

T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and minimum ℓ_1-norm interpolated classifiers. *The Annals of Statistics*, 50(3):1669–1695, 2022.

A. Maleki. Approximate message passing algorithm for compressed sensing. *Stanford University Ph.D. Thesis*, 2011.

A. Maleki and A. Montanari. Analysis of approximate message passing algorithm. In *Proc. IEEE Conf. Inform. Science and Systems (CISS)*, 2010.

L. Miolane and A. Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4), 2021.

A. Mousavi, A. Maleki, and R. G. Baraniuk. Consistent parameter estimation for LASSO and approximate message passing. *Annals of Statistics*, 45(6):2427–2454, 2017.

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

T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *Proc. Annual Allerton Conference on Communication, Control, and Computing*, pages 1002–1009. IEEE, 2013.

P. Patil, Y. Wei, A. Rinaldo, and R. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

P. Patil, A. Rinaldo, and R. Tibshirani. Estimating Functionals of the Out-of-Sample Error Distribution in High-Dimensional Ridge Regression. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6087–6120. PMLR, 2022.

M. Pawelczyk, J. Z. Di, Y. Lu, A. Sekhari, G. Kamath, and S. Neel. Machine unlearning fails to remove data poisoning attacks. *arXiv preprint arXiv:2406.17216*, 2024.

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

K. Rahnama Rad, W. Zhou, and A. Maleki. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4067–4077. PMLR, 26–28 Aug 2020.

A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

W. Stephenson and T. Broderick. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2424–2434. PMLR, 2020.

V. Suriyakumar and A. C. Wilson. Algorithms that approximate data removal: New results and limitations. *Advances in Neural Information Processing Systems*, 35:18892–18903, 2022.

A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

C. Thrampoulidis, E. Abbasi, and B. Hassibi. Precise error analysis of regularized m-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

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

S. Wang, H. Weng, and A. Maleki. Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823, 2020.

S. Wang, H. Weng, and A. Maleki. Does slope outperform bridge regression? *Information and Inference: A Journal of the IMA*, 11(1):1–54, 2022.

H. Weng, A. Maleki, and L. Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *The Annals of Statistics*, 46(6A):3099 – 3129, 2018.

H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), Aug. 2023. ISSN 0360-0300. URL `https://doi.org/10.1145/3603620`.

H. Zou, A. Auddy, K. . Rahnama Rad, and A. Maleki. Theoretical analysis of leave-one-out cross validation for non-differentiable penalties under high-dimensional settings. *arXiv preprint arXiv:2402.08543*, 2024.

Appendix A. Technical Lemmas

In this section we provide some auxilliary lemmas to be used in the proofs of our lemmas and theorems.

Lemma 14 $\mathbb{P}(\|\boldsymbol{x}\| > 2c\rho_{\max}\sqrt{\log(n)}) \leq \frac{1}{\sqrt{2\pi c}} n^{-c}$ *for* $n \geq 3$.

Lemma 15 *Let* Z_i *be* N *dependent* $\mathcal{N}(0, \sigma_i^2)$ *random variables. Suppose* $\sigma_i^2 \leq \sigma_{\max}^2 < \infty$. *Then* $\forall n \geq 1, c > 0$*:*

$$\mathbb{P}(\max_{i\in[N]} |Z_i| > 2\sigma_{\max}\sqrt{\log(N) + c\log(n)}) \leq 2n^{-c}.$$

Proof Let $Z := \max_{i\in[N]} Z_i$. By Lemma 23,

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{-\frac{t^2}{2\sigma_{\max}^2}}.$$

By Lemma 22, $\mathbb{E}Z \leq \sigma_{\max}\sqrt{2\log(N)}$. Setting $t = \sigma_{\max}\sqrt{2c\log(n)}$ yields

$$\mathbb{P}(Z > \sigma_{\max}(\sqrt{2\log(N)} + \sqrt{2c\log(n)})) \leq n^{-c}.$$

Finally notice that

$$\max_{i\in[N]} |Z_i| = \max\left\{\max_{i\in[N]} Z_i, \max_{i\in[N]}(-Z_i)\right\},$$

so by a union bound and the fact that $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ we have

$$\mathbb{P}(\max_{i\in[N]} |Z_i| > 2\sigma_{\max}\sqrt{\log(N) + c\log(n)}) \leq 2n^{-c}.$$

■

Lemma 16 *Suppose* $\boldsymbol{b} \in \mathbb{R}^p$ *is a random vector with density*

$$p_{\boldsymbol{b}}(\boldsymbol{b}) \propto e^{-C\|\boldsymbol{b}\|},$$

then $\|\boldsymbol{b}\| \sim \Gamma(p, C)$ *with density*

$$p_{\|\boldsymbol{b}\|}(r) = \frac{C^p}{\Gamma(p)} r^{p-1} e^{-cr}, \text{ and } \mathbb{P}\left(\|\boldsymbol{b}\| > \frac{2p}{C}\right) \leq e^{-(1-\log(2))p}.$$

Proof

Note that the pdf of $\boldsymbol{b}$ depends only on its ℓ_2 norm $\|\boldsymbol{b}\|$. Thus, $\boldsymbol{b}$ has a spherically symmetric distribution. By Theorem 4.2 of Fourdrinier et al. (2018) the pdf of $\|\boldsymbol{b}\|$ is given by:

$$p_{\|\boldsymbol{b}\|}(r) \propto \frac{2\pi^{p/2}}{\Gamma(p/2)} r^{p-1} \mathrm{e}^{-Cr}$$

and thus $\|\boldsymbol{b}\| \sim \Gamma(p, C)$. This finishes the proof of the first part. For the second part we use the moment generating function of the Gamma distribution to write:

$$\mathbb{P}(\|\boldsymbol{b}\| > \frac{2p}{C}) \leq \inf_{0<t<C} \mathbb{E}(\mathrm{e}^{t\|\boldsymbol{b}\|})\mathrm{e}^{-\frac{2tp}{C}} = \inf_{0<t<C} \left(1 - \frac{t}{C}\right)^{-p} \mathrm{e}^{-\frac{2tp}{C}} = 2^p \mathrm{e}^{-p}.$$

The last equality follows since the infimum in the previous line is achieved by $t = \frac{C}{2}$. ■

Lemma 17 *Under Assumptions A1, A2, and A3,*

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}} = \bar{\boldsymbol{G}}^{-1} \left(\sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \boldsymbol{x}_i \right) = \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1} \left(\sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i \right)$$

where

$$\begin{aligned}
\bar{\boldsymbol{G}} &:= \int_0^1 \boldsymbol{G}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) dt \\
\bar{\boldsymbol{G}}_{\setminus\mathcal{M}} &:= \int_0^1 \boldsymbol{G}_{\setminus\mathcal{M}}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) dt \\
\boldsymbol{G}(\boldsymbol{\beta}) &:= \boldsymbol{X}^\top \mathbf{diag}[\ddot{\ell}_i(\boldsymbol{\beta})]_{i\in[n]} \boldsymbol{X} + \lambda \mathbf{diag}[\ddot{r}_k(\boldsymbol{\beta})]_{k\in[p]} \\
\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) &:= \boldsymbol{X}_{\setminus\mathcal{M}}^\top \mathbf{diag}[\ddot{\ell}_i(\boldsymbol{\beta})]_{i\notin\mathcal{M}} \boldsymbol{X}_{\setminus\mathcal{M}} + \lambda \mathbf{diag}[\ddot{r}_k(\boldsymbol{\beta})]_{k\in[p]}
\end{aligned}$$

Proof Consider the optimality conditions of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$:

$$\begin{aligned}
\sum_{i \notin \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \boldsymbol{x}_i + \lambda \nabla r(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) = 0 \\
\sum_{i \in [n]} \dot{\ell}_i(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i + \lambda \nabla r(\hat{\boldsymbol{\beta}}) = 0
\end{aligned}$$

Subtracting one from another and applying the mean value theorem, we get

$$\bar{\boldsymbol{G}} \cdot (\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}) = \sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \boldsymbol{x}_i. \tag{12}$$

Multiplying $\bar{\boldsymbol{G}}^{-1}$:

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}} = \bar{\boldsymbol{G}}^{-1} \left(\sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \boldsymbol{x}_i \right).$$

For the other version, rearranging the terms in we get (12):

$$\bar{\boldsymbol{G}}_{\setminus\mathcal{M}} (\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}) = \sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i,$$

therefore

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}} = \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1} \left(\sum_{i \in \mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_i \right).$$

■

Lemma 18 *Let $\boldsymbol{X}^\top = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ be a $p \times m$ matrix, where $\boldsymbol{Z} \in \mathbb{R}^{p\times m}$ has elements $Z_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\boldsymbol{\Sigma}$ is positive semi-definite with $\sigma_{\max}(\boldsymbol{\Sigma}) := \rho_{\max}$. Then, for any $c \geq 1$*

$$\mathbb{P}\left(\|\boldsymbol{X}^\top\|_{2,\infty} \geq \sqrt{m\rho_{\max}(1+4c\log(n))}\right) \leq \exp\left(-cm\log(n)\right).$$

Proof [Proof of Lemma 18]

Note that by a well-known equality for the $2 \to \infty$ matrix norm, we have

$$\|\boldsymbol{X}^\top\|_{2,\infty} = \max_{1\leq k\leq p} \|\boldsymbol{e}_k^\top \boldsymbol{X}^\top\|_2 = \max_{1\leq k\leq p} \|\boldsymbol{e}_k^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}\|_2.$$

Note that $\boldsymbol{e}_k^\top \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} \sim \mathcal{N}(0, \sigma_{kk}\mathbb{I}_m)$ for $1 \leq k \leq p$, where σ_{kk} is the (k,k) element of $\boldsymbol{\Sigma}$ and thus,

$$\mathbb{P}(\|\boldsymbol{X}^\top\|_{2,\infty} \geq \sqrt{m\rho_{\max}(1+4c\log(n))}) \leq \exp(-cm\log(n))$$

by concentration inequalities for χ^2 distributed random variables (see, e.g., Laurent and Massart, 2000). ■

A.1 Adopted Lemmata

Lemma 19 (Lemma 14 of Auddy et al. (2024)) *For $1 < s \leq N$ and $N > 2$, we have*

1. $$\binom{N}{s} \leq e^{s\log\frac{eN}{s}}$$

2. *For $m \leq (n+1)/3$,*
$$\sum_{s=0}^{m}\binom{n}{s} \leq 2\binom{n}{m}$$

Proof We prove each part separately:

- The first part is exactly Lemma 14 of Auddy et al. (2024).
- Notice that for $s < m < n$,

$$\frac{\binom{n}{s}}{\binom{n}{m}} = \frac{\frac{n!}{s!(n-s)!}}{\frac{n!}{m!(n-m)!}} = \frac{m(m-1)\cdots(s+1)}{(n-s)(n-s-1)\cdots(n-m+1)} \leq \left(\frac{m}{n-m+1}\right)^{m-s},$$

so

$$\begin{aligned}\frac{\sum_{s=0}^{m}\binom{n}{s}}{\binom{n}{m}} &\leq 1 + \sum_{s=0}^{m-1}\left(\frac{m}{n-m+1}\right)^{m-s} \leq \sum_{t=0}^{\infty}\left(\frac{m}{n-m+1}\right)^{t} = \frac{n-m+1}{n-2m+1} \\ &\leq 1 + \frac{m}{n-2m+1} \leq 2\end{aligned}$$

provided $\frac{m}{n-2m+1} \leq 1 \Leftrightarrow m \leq \frac{n+1}{3}$. ■

Lemma 20 (Lemma 19 in Auddy et al. (2024)) *Suppose the rows of $\boldsymbol{X}$ satisfy Assumption B1, then*

$$\mathbb{P}(\|\boldsymbol{X}^\top \boldsymbol{X}\| \geq (\sqrt{\gamma_0}+3)^2 C_X) \leq \mathrm{e}^{-p}.$$

Lemma 21 (Lemma 17 in Auddy et al. (2024)) *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{iid}{\sim} N(0, \boldsymbol{\Sigma}) \in \mathbb{R}^p$ and suppose $\rho_{\max}(\boldsymbol{\Sigma}) \leq p^{-1} C_X$ for some constant $C_X > 0$, then*

$$\mathbb{P}(\max_{1\leq i\leq n} \|\boldsymbol{x}_i\| \geq 2\sqrt{C_X}) \leq n e^{-p/2}.$$

Lemma 22 (Lemma 4.10 of Chatterjee (2014)) *Let Z_i be N dependent $\mathcal{N}(0, \sigma_i^2)$ random variables with $\sigma_i \leq \sigma_{\max} < \infty$, then*

$$\mathbb{E} \max_{i\in[N]} Z_i \leq \sigma_{\max}\sqrt{2\log(N)}.$$

Proof The proof is essentially the same as Lemma 4.10 of Chatterjee (2014), except for the absolute value. Let $Z := \max_{i\in[N]} Z_i$. By Jensen's inequality, $\forall t > 0$:

$$e^{t\mathbb{E}Z} \leq \mathbb{E}e^{tZ} = \mathbb{E}e^{t\max_{i\in[N]} Z_i} \leq \sum_{i\in[N]} \mathbb{E}e^{tZ_i} = \sum_{i\in[N]} e^{\frac{1}{2}t^2\sigma_i^2} \leq N e^{\frac{1}{2}t^2\sigma_{\max}^2},$$

which implies

$$\mathbb{E}Z \leq \frac{1}{t}[\log(N) + \frac{1}{2}t^2\sigma_{\max}^2].$$

The right hand side minimizes at $t^2 = \frac{2\log(N)}{\sigma_{\max}^2}$, and hence we have $\mathbb{E}Z \leq \sigma_{\max}\sqrt{2\log(N)}$. ■

Lemma 23 (Borel-TIS inequality, Theorem 2.1.1 of Adler and Taylor (2007)) *Let Z_i be N dependent $\mathcal{N}(0, \sigma_i^2)$ random variables with $\sigma_i \leq \sigma_{\max} < \infty$, and $Z := \max_{i\in[N]} Z_i$, then $\forall t > 0$,*

$$\mathbb{P}(Z - \mathbb{E}Z > t) \leq e^{-\frac{t^2}{2\sigma_{\max}^2}}.$$

Lemma 24 (Lemma 6 of Jalali and Maleki (2016)) *Let $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_p)$, then*

$$\mathbb{P}(\boldsymbol{x}^\top \boldsymbol{x} \geq p + pt) \leq \mathrm{e}^{-\frac{p}{2}(t-\log(1+t))}.$$

Appendix B. Detailed Proofs

B.1 Proof of Lemma 9

Proof We first prove sufficiency. Conditional on the data $\mathcal{D}$, $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$ is nothing but a translation of $\boldsymbol{b}$, so its conditional density is

$$p_{\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}|\mathcal{D}}(\boldsymbol{\beta}) = p_{\boldsymbol{b}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}),$$

and it is similar for $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$. Notice that $\log(p_{\boldsymbol{b}}(\boldsymbol{b})) = -\frac{\epsilon}{r}\|\boldsymbol{b}\|$ is $\frac{\epsilon}{r}$-Lipschitz in $\boldsymbol{b}$, therefore $\forall\ \mathcal{D} \in \mathcal{X}_r$,

$$\begin{aligned}
&|\log(p_{\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})) - \log(p_{\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta}))| \\
=\ &|\log(p_{\boldsymbol{b}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}})) - \log(p_{\boldsymbol{b}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}))| \\
\leq\ &\frac{\epsilon}{r}\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \\
\leq\ &\epsilon
\end{aligned}$$

which is equivalent to

$$\mathrm{e}^{-\epsilon} \leq \frac{p_{\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})}{p_{\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})} \leq \mathrm{e}^{\epsilon}.$$

The result then follows by integrating the densities over the set $\mathcal{T}$.

For the necessity part, for $\mathcal{D}_0 \notin \mathcal{X}_r$, by definition $\exists\mathcal{M} \subset [n], \exists\delta > 0$, $\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \geq r + \delta$. The proof is then straightforward since $-\frac{\epsilon}{r}\|\boldsymbol{b}\|$ is not $\frac{\epsilon}{r+\delta}$-Lipschitz. To be more specific, define

$$\begin{aligned}
\mathcal{T}_+ &:= \{\boldsymbol{\beta} \in \mathbb{R}^p | \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| - \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \geq r + \frac{\delta}{2}\}, \\
\mathcal{T}_- &:= \{\boldsymbol{\beta} \in \mathbb{R}^p | \|\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| - \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \geq r + \frac{\delta}{2}\}.
\end{aligned}$$

By basic geometry this is the areas within two pieces of a hyperboloid. Then $\forall T \subset \mathcal{T}_+$,

$$\frac{p_{\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})}{p_{\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})} = \mathrm{e}^{\frac{\epsilon}{r}(\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| - \|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|)} \geq \mathrm{e}^{\epsilon(1+\frac{\delta}{2r})},$$

and similarly $\forall T \subset \mathcal{T}_-$,

$$\frac{p_{\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})}{p_{\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}+\boldsymbol{b}}(\boldsymbol{\beta})} \leq \mathrm{e}^{-\epsilon(1+\frac{\delta}{2r})}.$$

■

B.2 ℓ_2 error of t-step Newton estimators

In this section, we prove Theorem 11 and Theorem 12. In fact we prove some statements slightly stonger than what we discussed in the beginning of the Appendix: not only can we bound the ℓ_2 errors of t-step Newton estimators, but also we can bound them **simutaneously**. To be more specific, we define a "failure event" F for the dataset $\mathcal{D}$, such that:

1. On event F,

$$\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O\left(\sqrt{\frac{m^3}{n}}\mathrm{polylog}(n)\right)$$

(Lemma 26),

2. $\mathbb{P}(F) = \phi_n \to 0$ (Lemma 27), and
3. On event F, the Hessian $\boldsymbol{G}_{\setminus\mathcal{M}}$ is polylog(n)-Lipschitz in $\boldsymbol{\beta}$, so that the Newton estimators satisfy quadratic convergence (Lemma 30) and we get

$$\|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 = O\left(\left(\frac{m^3}{n}\right)^{2^{t-2}} \text{polylog}(n)\right), \quad \forall\, t \geq 1$$

(Theorem 12).

In the rest of the section, we adopt the following notations for brevity:

$$\begin{aligned}
\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}) &= [\ddot{\ell}_i(\boldsymbol{\beta})]_{i\in[n]}, \ddot{\boldsymbol{L}}(\boldsymbol{\beta}) = \mathbf{diag}[\ddot{\boldsymbol{\ell}}(\boldsymbol{\beta})] \\
\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) &= [\ddot{\ell}_i(\boldsymbol{\beta})]_{i\notin\mathcal{M}}, \ddot{\boldsymbol{L}}_{\setminus\mathcal{M}} = \mathbf{diag}[\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta})] \\
\ddot{\boldsymbol{r}}(\boldsymbol{\beta}) &= [\ddot{r}_k(\beta_k)]_{k\in[p]}, \ddot{\boldsymbol{R}}(\boldsymbol{\beta}) = \mathbf{diag}[\ddot{\boldsymbol{r}}(\boldsymbol{\beta})] \\
\boldsymbol{G}(\boldsymbol{\beta}) &= \boldsymbol{X}^\top \ddot{\boldsymbol{L}}(\boldsymbol{\beta})\boldsymbol{X} + \lambda\ddot{\boldsymbol{R}}(\boldsymbol{\beta}) \\
\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) &= \boldsymbol{X}^\top_{\setminus\mathcal{M}} \ddot{\boldsymbol{L}}(\boldsymbol{\beta})\boldsymbol{X}_{\setminus\mathcal{M}} + \lambda\ddot{\boldsymbol{R}}(\boldsymbol{\beta}).
\end{aligned}$$

B.2.1 ℓ_2 error of one Newton step

We first provide a finer characterization of $\mathcal{X}_r^{(1)} = \{\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \leq r\}$ in terms of some smaller "failure events" $F_i, i = 1, 2, ..., 5$ that are easier to verify:

$$\begin{aligned}
F_1 &:= \{\|\boldsymbol{X}\| > C_1\}, \\
F_2 &:= \{\max_{i\in[n]} |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| > C_\ell(n)\}, \\
F_3 &:= \{\exists \mathcal{M} \subset \mathcal{D} \text{ with } |\mathcal{M}| \leq m, \boldsymbol{\beta} \in \mathcal{B}_{1,\mathcal{M}}, \\
&\qquad \|\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\| > C_{\ell\ell}(n)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \\
&\qquad \text{or } \|\nabla^2 r(\boldsymbol{\beta}) - \nabla^2 r(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\| > C_{rr}(n)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|\}
\end{aligned}$$

where $\boldsymbol{\xi}(t) := t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, $\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) := [\ddot{\ell}_i(\boldsymbol{\beta})]_{i\notin\mathcal{M}}$,
$\mathcal{B}_{1,\mathcal{M}} := \{\boldsymbol{\beta} : \boldsymbol{\beta} = t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, t \in [0,1]\}$.

$$\begin{aligned}
F_4 &:= \{\exists \mathcal{M} \subset \mathcal{D} \text{ with } |\mathcal{M}| \leq m, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}_{2,\mathcal{M}}, \\
&\qquad \|\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_1) - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_2)\| > C_{\ell\ell}(n)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|, \\
&\qquad \text{or } \|\nabla^2 r(\boldsymbol{\beta}) - \nabla^2 r(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\| > C_{rr}(n)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|\}
\end{aligned}$$

where $\mathcal{B}_{2,\mathcal{M}} := \mathcal{B}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, 1)$

$$\begin{aligned}
F_5 &:= \{\max_{|\mathcal{M}|\leq m} \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}} \boldsymbol{G}^{-1}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}^\top_{\mathcal{M}}\|_{2,\infty} > C_{xx}(n)\sqrt{\frac{m}{n}}\}, \\
&\text{where } \bar{\boldsymbol{X}}_{\setminus\mathcal{M}} = \begin{pmatrix} \boldsymbol{X}_{\setminus\mathcal{M}} \\ \mathbb{I}_p \end{pmatrix}, \ \boldsymbol{X}^\top_{\mathcal{M}} = (\boldsymbol{x}_i)_{i\in\mathcal{M}} \\
F &:= \cup_{i=1}^5 F_i
\end{aligned} \tag{13}$$

Remark 25 *Technically, F_4 will not be used in this section, but will be used in the proof of Theorem 12 in Section B.2.6. However, we enlist it among other events as it is similar to F_3 and the proof is similar too.*

Note that the undetermined constants, namely C_1, $C_\ell(n)$, $C_{\ell\ell}(n)$, $C_{rr}(n)$ and $C_{xx}(n)$, will be decided later when we analyze their probabilities. Roughly speaking, these constants will be at most $O(\text{polylog}(n))$ under Assumptions B1-B3. The next lemma shows that $\left(\mathcal{X}_r^{(1)}\right)^c \subset F := \cup_{i=1}^5 F_i$ for a specific r:

Lemma 26 *under Assumptions A1-A3, Let $F = \cup_{i=1}^5 F_i$ be the failure event, then $\left(\mathcal{X}_r^{(1)}\right)^c \subset F$ with*

$$r = \left[\frac{2\sqrt{3}}{3\lambda^2\nu^2}[C_{\ell\ell}(n) + \lambda C_{rr}(n)]C_1(C_1+1)C_\ell^2(n)C_{xx}(n)\right]\frac{m^{3/2}}{\sqrt{n}} := C_1(n)\frac{m^{3/2}}{\sqrt{n}}$$

i.e., under F^c,

$$\max_{|\mathcal{M}|\le m} \|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \le C_1(n)\frac{m^{3/2}}{\sqrt{n}}.$$

The proof of this Lemma is postponed to Section B.2.2.

Now that we have $\left(\mathcal{X}_r^{(1)}\right)^c \subset F = \cup_{i=1}^5 F_i$, we can bound $\phi = \mathbb{P}(\mathcal{D} \notin \mathcal{X}_r^{(1)})$ by bounding $\mathbb{P}(F)$ instead:

Lemma 27 *Under Assumptions A1-A3 and B1-B3,*

$$\mathbb{P}(F) \le nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p}$$

where the constants $C_\ell(n), C_{\ell\ell}(n), C_{rr}(n)$ and $C_{xx}(n)$ in (13) are all $O(\text{polylog}(n))$.

The proof is in Section B.2.3 and the exact form of the constants can be found in the proof of Lemma 28 and 29. Theorem 11 is then a direct application of Lemma 26 and Lemma 27:

Proof [Proof of Theorem 11] By Lemma 26, $(\mathcal{X}_r^{(1)})^c \subset F$ with $r = C_1(n)\frac{m^{\frac{3}{2}}}{\sqrt{n}}$. By Lemma 27, $\mathbb{P}(F) \le nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p}$. Combining the two lemmas we immediately have

$$\mathbb{P}(\mathcal{D} \in \mathcal{X}_r^{(1)}) \le nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p},$$

where $r = C_1(n)\frac{m^{\frac{3}{2}}}{\sqrt{n}}$ and $C_1(n) = O(\text{polylog}(n))$ by Lemma 27, which concludes the proof of Theorem 11. ■

B.2.2 Proof of Lemma 26

Proof Recall that we denote $\boldsymbol{G}(\boldsymbol{\beta})$ and $\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta})$ to be the Hessian of the loss functions of the full model and the unlearned model respectively, and

$$\bar{\boldsymbol{G}} := \int_0^1 \boldsymbol{G}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})dt$$

$$\bar{\boldsymbol{G}}_{\setminus\mathcal{M}} := \int_0^1 \boldsymbol{G}_{\setminus\mathcal{M}}(t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})dt$$

By Lemma 17, we have, using the notations above and in (13),

$$\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}} = \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\left(\sum_{i\in\mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\right) = \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{X}_{\mathcal{M}}^\top \dot{\boldsymbol{\ell}}_{\mathcal{M}}$$

and by Definition 6 we have

$$\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} - \hat{\boldsymbol{\beta}} = \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}})\left(\sum_{i\in\mathcal{M}} \dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\right) = \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{X}_{\mathcal{M}}^\top \dot{\boldsymbol{\ell}}_{\mathcal{M}}.$$

If we define $\boldsymbol{v}_{\mathcal{M}} := \boldsymbol{X}_{\mathcal{M}}^\top \dot{\boldsymbol{\ell}}_{\mathcal{M}}$, then by subtracting the two equations above, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)} &= \left[\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1} - \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}})\right]\boldsymbol{v}_{\mathcal{M}} \\
&= \left[\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1} - \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\right]\boldsymbol{v}_{\mathcal{M}} + \left[\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}})\right]\boldsymbol{v}_{\mathcal{M}} \\
&:= \boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}} + \boldsymbol{M}_2\boldsymbol{v}_{\mathcal{M}},
\end{aligned}$$

thus we have

$$\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(1)}\| \le \|\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}\| + \|\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}\|.$$

Since $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ possess similar properties, we will bound $\|\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}\|$ in the following, while $\|\boldsymbol{M}_2\boldsymbol{v}_{\mathcal{M}}\|$ can be bounded using the same method.

$$\begin{aligned}
\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}} &= \left[\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1} - \boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\right]\boldsymbol{v}_{\mathcal{M}} \\
&= \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\left[\boldsymbol{G}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}\right]\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{v}_{\mathcal{M}}
\end{aligned}$$

Notice that we can write $\boldsymbol{G}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}$ in a more compact form:

$$\begin{aligned}
\boldsymbol{G}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \bar{\boldsymbol{G}}_{\setminus\mathcal{M}} &= \boldsymbol{X}_{\setminus\mathcal{M}}^\top[\ddot{\boldsymbol{L}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \bar{\ddot{\boldsymbol{L}}}]\boldsymbol{X}_{\setminus\mathcal{M}} + \lambda[\ddot{\boldsymbol{R}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) - \bar{\ddot{\boldsymbol{R}}}] \\
&:= \bar{\boldsymbol{X}}_{\setminus\mathcal{M}}^\top \boldsymbol{\Gamma}\bar{\boldsymbol{X}}_{\setminus\mathcal{M}},
\end{aligned}$$

where in the last step we define

$$\begin{aligned}
\bar{\boldsymbol{X}}_{\setminus\mathcal{M}} &:= \begin{pmatrix} \boldsymbol{X}_{\setminus\mathcal{M}} \\ \boldsymbol{I}_p \end{pmatrix}, \\
\boldsymbol{\Gamma} &:= \begin{pmatrix} \mathbf{diag}[\int_0^1 \ddot{\ell}_i(\boldsymbol{\xi}(t)) - \ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})dt]_{i\notin\mathcal{M}}, & \boldsymbol{0} \\ \boldsymbol{0}, & \mathbf{diag}[\int_0^1 \ddot{r}(\xi_k(t)) - \ddot{r}(\hat{\beta}_{\lambda\setminus\mathcal{M},k})]_{k\in[p]} \end{pmatrix}, \\
\boldsymbol{\xi}(t) &:= t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}.
\end{aligned} \tag{14}$$

We then have

$$\begin{aligned}
&\|\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}\| \\
=& \sup_{\|\boldsymbol{w}\|=1} |\boldsymbol{w}^\top \boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}| \\
=& \sup_{\|\boldsymbol{w}\|=1} \left|\boldsymbol{w}^\top \bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}^\top \boldsymbol{\Gamma}\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{v}_{\mathcal{M}}\right| \\
\leq& \sup_{\|\boldsymbol{w}\|=1} \|\boldsymbol{\Gamma}\|_{Fr}\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{w}\|_2 \left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{v}_{\mathcal{M}}\right\|_\infty \\
=& \|\boldsymbol{\Gamma}\|_{Fr} \cdot \sup_{\|\boldsymbol{w}\|=1} \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{w}\|_2 \cdot \left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{v}_{\mathcal{M}}\right\|_\infty, \qquad (15)
\end{aligned}$$

where in the penultimate line we use Cauchy Schwarz inequality: for two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ and diagonal matrix $\boldsymbol{D} = \mathbf{diag}[d_k]_{k\in[p]}$,

$$\begin{aligned}
\boldsymbol{u}^\top \boldsymbol{D}\boldsymbol{v} = \sum_{k\in[p]} d_k u_k v_k \leq& \left(\sum_{k\in[p]} d_k^2\right)^{\frac{1}{2}} \left(\sum_{k\in[p]} u_k^2 v_k^2\right)^{\frac{1}{2}} \leq \left(\sum_{k\in[p]} d_k^2\right)^{\frac{1}{2}} \left(\sum_{k\in[p]} u_k^2\right)^{\frac{1}{2}} \max_{1\leq k\leq p} |v_k| \\
=& \|\boldsymbol{D}\|_{\mathrm{Fr}}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_\infty.
\end{aligned}$$

Now we bound the three terms in (15) separately.

1. Define

$$\begin{aligned}
\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) &:= [\ddot{\ell}_i(\boldsymbol{\beta})]_{i\notin\mathcal{M}} \\
\bar{\ddot{\boldsymbol{\ell}}}_{\setminus\mathcal{M}} &:= \left[\int_0^1 \ddot{\ell}_i(\boldsymbol{\xi}(t))dt\right]_{i\notin\mathcal{M}} \text{ where } \boldsymbol{\xi}(t) = t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} \\
\ddot{\boldsymbol{r}}(\boldsymbol{\beta}) &:= [\ddot{r}(\beta_k)]_{k\in[p]} \\
\bar{\ddot{\boldsymbol{r}}} &:= \left[\int_0^1 \ddot{\boldsymbol{r}}(\boldsymbol{\xi}(t))dt\right]_{k\in[p]},
\end{aligned}$$

   Then

$$\|\boldsymbol{\Gamma}\|_{Fr} \leq \|\bar{\ddot{\boldsymbol{\ell}}}_{\setminus\mathcal{M}} - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2 + \lambda\|\bar{\ddot{\boldsymbol{r}}} - \ddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2.$$

   where

$$\begin{aligned}
&\|\bar{\ddot{\boldsymbol{\ell}}}_{\setminus\mathcal{M}} - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2 \\
=& \sqrt{\sum_{i\notin\mathcal{M}} \left[\int_0^1 [\ddot{\ell}_i(\boldsymbol{\xi}(t) - \ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}))]\right]^2} \leq \sqrt{\int_0^1 \sum_{i\notin\mathcal{M}} [\ddot{\ell}_i(\boldsymbol{\xi}(t) - \ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}))]^2 dt} \\
\leq& \sqrt{\int_0^1 \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\xi}(t)) - \ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|^2 dt} \leq \sqrt{\int_0^1 C_{\ell\ell}^2(n)\|\boldsymbol{\xi}(t) - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^2 dt} \\
=& C_{\ell\ell}(n)\sqrt{\int_0^1 t^2\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^2 dt} \\
=& \frac{\sqrt{3}}{3} C_{\ell\ell}(n)\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|.
\end{aligned}$$

By Lemma 17, under event F:

$$\begin{aligned}\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| &= \|\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{X}_{\mathcal{M}}^{\top}\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\| \le \frac{1}{\lambda\nu}\|\boldsymbol{X}_{\mathcal{M}}^{\top}\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\| \\ &\le \frac{1}{\lambda\nu}\|\boldsymbol{X}\| \cdot \|\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\| \le \frac{1}{\lambda\nu}\|\boldsymbol{X}\|\sqrt{m}\max_{i\in[n]}|\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \\ &\le \frac{\sqrt{m}}{\lambda\nu}C_1 C_\ell(n), \end{aligned} \tag{16}$$

where the first line uses Lemma 17, the second uses strong convexity of the risk function, and the last line uses the definition of events F_1, F_2. Therefore we have

$$\|\dddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}} - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2 \le \frac{\sqrt{3m}}{3\lambda\nu}C_1 C_\ell(n) C_{\ell\ell}(n).$$

Similarly we have

$$\|\dddot{\boldsymbol{r}}_{\setminus\mathcal{M}} - \ddot{\boldsymbol{r}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2 \le \frac{\sqrt{3m}}{3\lambda\nu}C_1 C_\ell(n) C_{rr}(n),$$

so we have

$$\|\boldsymbol{\Gamma}\|_{Fr} \le \frac{\sqrt{3m}}{3\lambda\nu}[C_{\ell\ell}(n) + \lambda C_{rr}(n)]C_1 C_\ell(n).$$

2.

$$\begin{aligned}\sup_{\|\boldsymbol{w}\|=1}\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{w}\|_2 &\le \sup_{\|\boldsymbol{w}\|=1}\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\|\|\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\|\|\boldsymbol{w}\| \\ &\le \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\|\|\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\|.\end{aligned}$$

Notice that $\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\| \le \|\boldsymbol{X}_{\setminus\mathcal{M}}\| + 1 \le \|\boldsymbol{X}\| + 1 \le C_1 + 1$, so we have

$$\sup_{\|\boldsymbol{w}\|=1}\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\bar{\boldsymbol{G}}_{\setminus\mathcal{M}}^{-1}\boldsymbol{w}\|_2 \le \frac{C_1+1}{\lambda\nu}$$

3.

$$\begin{aligned}&\left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{v}_{\mathcal{M}}\right\|_\infty \\ =&\left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^{\top}\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\right\|_\infty \\ \le&\left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^{\top}\right\|_{2,\infty}\left\|\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\right\|_2,\end{aligned}$$

where in the last line we use the definition of the $(2,\infty)$ norm of a matrix $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_{2,\infty} := \sup_{\|\boldsymbol{w}\|_2\le 1}\|\boldsymbol{A}\boldsymbol{w}\|_\infty.$$

We know $\left\|\dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}})\right\|_2 \le \sqrt{m}\max_{i\in\mathcal{M}}|\dot{\ell}_i(\hat{\boldsymbol{\beta}})| \le \sqrt{m}C_\ell(n)$, so by the definition of event F_5,

$$\left\|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\left(\sum_{i\in\mathcal{M}}\dot{\ell}_i(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\right)\right\|_\infty \le C_\ell(n)C_{xx}(n)\frac{m}{\sqrt{n}}.$$

Combining all the results above we have

$$\|\boldsymbol{M}_1\boldsymbol{v}_{\mathcal{M}}\| \leq \frac{\sqrt{3}}{3\lambda^2\nu^2}[C_{\ell\ell}(n)+\lambda C_{rr}(n)]C_1(C_1+1)C_\ell^2(n)C_{xx}(n)\frac{m^{3/2}}{\sqrt{n}}.$$

Similar arguments lead to the same bound for $\|\boldsymbol{M}_2\boldsymbol{v}_{\mathcal{M}}\|$. So we finally have, under event F^c,

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| &\leq \frac{2\sqrt{3}}{3\lambda^2\nu^2}[C_{\ell\ell}(n)+\lambda C_{rr}(n)]C_1(C_1+1)C_\ell^2(n)C_{xx}(n)\frac{m^{3/2}}{\sqrt{n}}\\ &:= C_1(n)\frac{m^{3/2}}{\sqrt{n}}.\end{aligned}$$

■

B.2.3 Proof of Lemma 27

Lemma 28 *Under Assumptions A1-A3 and B1-B3, if we set*

$$\begin{aligned}C_1 &= (\sqrt{\gamma_0}+3)\sqrt{C_X},\\ C_\ell(n) &= \mathrm{polylog}_3(n)\\ C_{\ell\ell}(n) &= \max\{\mathrm{polylog}_{10}(n), \mathrm{polylog}_{11}(n)\},\end{aligned}$$

in the definition of F_2, F_3 in (13), then we have

$$\mathbb{P}(F_1\cup F_2\cup F_3) \leq nq_n + 4n^{1-c} + ne^{-p/2} + e^{-p},$$

The proof can be found in Section B.2.4, and the definition of the polylog(n) terms are summarized in (21).

Lemma 29 *Under Assumption B1, for any $c\geq 0$ we have $\mathbb{P}(F_5)\leq 4n^{-c}+e^{-p}$ with*

$$C_{xx}(n) = \frac{(\sqrt{\gamma_0}+3)C_X\vee 1}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}}$$

The proof of Lemma 29 can be found in Section B.2.5. By combining the above two lemmata we have

Proof [Proof of Lemma 27] By Lemma 20,

$$\mathbb{P}(\|\boldsymbol{X}\| > (\sqrt{\gamma_0}+3)\sqrt{C_X}) \leq e^{-p}.$$

By Lemma 28, $\mathbb{P}(F_2\cup F_3)\leq 2n^{1-c}+2nq_n+e^{-p}+ne^{-p/2}$ with

$$\begin{aligned}C_\ell(n) = \mathrm{polylog}_3(n) &= \mathrm{polylog}_1(n) + \frac{4C_X}{\lambda\nu}[1+C_y^s(n)+\mathrm{polylog}_1^s(n)]\\ C_{\ell\ell}(n) &= \max\{\mathrm{polylog}_{10}(n), \mathrm{polylog}_{11}(n)\}.\end{aligned}$$

Finally by Lemma 29, $\mathbb{P}(F_5) \le 4n^{-c} + e^{-p}$ with

$$C_{xx}(n) = \frac{(\sqrt{\gamma_0}+3)C_X \vee 1}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}}.$$

Using a union bound over the events above we have

$$\mathbb{P}(F) \le nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p}.$$

■

B.2.4 Proof of Lemma 28

Proof The proof can be divided into 4 steps. In the following we denote $\hat{\boldsymbol{\beta}}_{\setminus\cdot}$ to be the model trained by excluding the observations indicated in $\cdot$, which can be the indices like $\hat{\boldsymbol{\beta}}_{\setminus i,j}$, or a subset $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, or both, e.g. $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i}$. We allow the content in $\cdot$ to overlap, in which case we simply take a union, for example $\hat{\boldsymbol{\beta}}_{\setminus i,i} := \hat{\boldsymbol{\beta}}_{\setminus i}$. By slight abuse of notation, we use the index "0" as a placeholder to denote no observations being excluded, so that $\max_{0\le i\le n}\hat{\boldsymbol{\beta}}_{\setminus i}$ means the maximum among all $\hat{\boldsymbol{\beta}}_{\setminus i}$ and also $\hat{\boldsymbol{\beta}}$. These additional definitions help us to keep the notations unified and simple.

Step 1

Define event $E_1 := \{\max_i |y_i| \le C_y(n)\}$ with probability at least $1 - nq_n$ by Assumption B3 and a union bound. Under E_1, $\forall\mathcal{M} \subset \mathcal{D}$ (including $\mathcal{M} = \emptyset$ case where $\hat{\boldsymbol{\beta}}_{\lambda,\emptyset} := \hat{\boldsymbol{\beta}}$). We have

$$\begin{aligned}\lambda\nu\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^2 &\overset{(a)}{\le} \sum_{i\notin\mathcal{M}} \ell_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) + \lambda r(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \overset{(b)}{\le} \sum_{i\in[n]} \ell_i(\mathbf{0}) \\ &\overset{(c)}{\le} \sum_{i\in[n]} 1 + |y_i|^s \le n(1 + C_y^s(n)),\end{aligned}$$

where (a) uses the ν-strong convexity of r (Assumption A2), (b) uses $r(\mathbf{0}) = 0$ and $\ell(\cdot) \ge 0$, (c) uses Asumption B2, and the last inequality uses the definition of event E_1. By rearranging the terms we have under E_1 that, $\forall\mathcal{M}$

$$\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \le \sqrt{(\lambda\nu)^{-1}(1 + C_y^s(n))n}. \tag{17}$$

Step 2

Define event $E_2 := \{\max_{i,j\in[n]} |\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{\setminus i,j}| \le \mathrm{polylog}_1(n)\}$ with $\hat{\boldsymbol{\beta}}_{\setminus i,i} := \hat{\boldsymbol{\beta}}_{\setminus i}$ and

$$\mathrm{polylog}_1(n) := 2\sqrt{(\lambda\nu)^{-1}\gamma_0 C_X(1 + C_y^s(n))(1+c)\log(n)}$$

for any arbitrary $c > 0$ that will appear in the tail probability.

$$\begin{aligned}
&\mathbb{P}(E_1^c \cup E_2^c) \\
&= \mathbb{P}(E_1^c) + \mathbb{P}(E_2^c \cap E_1) \\
&\le nq_n + \mathbb{P}\left(\max_{i,j\in[n]} |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i,j}| > 2\sqrt{\frac{C_X}{p}} \|\hat{\boldsymbol{\beta}}_{\setminus i,j}\| \sqrt{(1+c)\log(n)},\ E_1 \right) \\
&\le nq_n + \sum_{i\in[n]} \mathbb{E}\mathbb{P}\left(\max_{j\in[n]} |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i,j}| > 2\sqrt{\frac{C_X}{p}} \|\hat{\boldsymbol{\beta}}_{\setminus i,j}\| \sqrt{(1+c)\log(n)} | \mathcal{D}_{\setminus i} \right) \\
&\le nq_n + 2n^{1-c},
\end{aligned}$$

where the second line used $\|\hat{\boldsymbol{\beta}}_{\setminus i,j}\| \le \sqrt{(\lambda\nu)^{-1}(1 + C_y^s(n))n}$ under E_1, the third line uses a union bound over i and the tower rule, and the last line uses Lemma 15 by observing that conditional on $\mathcal{D}_{\setminus i}$, $\max_{j\in[n]} |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i,j}|$ is the maximum of n Gaussians with $\sigma_i^2 = \hat{\boldsymbol{\beta}}_{\setminus i,j}^\top \boldsymbol{\Sigma} \hat{\boldsymbol{\beta}}_{\setminus i,j} \le \frac{C_X}{p} \|\hat{\boldsymbol{\beta}}_{\setminus i,j}\|^2$.

Step 3

Define $E_3 := \{\max_{i\in[n]} \|\boldsymbol{x}_i\| \le 2\sqrt{C_X}, \|\boldsymbol{X}\| \le (\sqrt{\gamma_0}+3)\sqrt{C_X}\}$ then $\mathbb{P}(E_3^c) \le ne^{-p/2} + e^{-p}$ by Lemma 20 and Lemma 21.

Under $\cap_{i=1}^3 E_i$, we have the following results: $\forall 1 \le i \le n, 0 \le j \le n$,

$$\begin{aligned}
|\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus i,j})| &\le 1 + |y_i|^s + |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i,j}|^s \\
&\le 1 + C_y^s(n) + \mathrm{polylog}_1^s(n) := \mathrm{polylog}_2(n) \\
\|\hat{\boldsymbol{\beta}}_{\setminus j} - \hat{\boldsymbol{\beta}}_{\setminus i,j}\| &\le \frac{1}{\lambda\nu} |\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus i,j})| \|\boldsymbol{x}_i\| \le \frac{2\sqrt{C_X}}{\lambda\nu} \mathrm{polylog}_2(n) \\
|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus j}| &\le |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i,j}| + |\boldsymbol{x}_i^\top (\hat{\boldsymbol{\beta}}_{\setminus j} - \hat{\boldsymbol{\beta}}_{\setminus i,j})| \\
&\le \mathrm{polylog}_1(n) + \frac{4C_X}{\lambda\nu} \mathrm{polylog}_2(n) \\
&:= \mathrm{polylog}_3(n) \qquad (18) \\
|\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus j})| &\le 1 + C_y^s(n) + \mathrm{polylog}_3^s(n) := \mathrm{polylog}_4(n).
\end{aligned}$$

Step 4

Under $\cap_{i=1}^3 E_i$, $\forall \mathcal{M} \subset \mathcal{D}, |\mathcal{M}| \le m, \forall j \notin \mathcal{M}$: by Lemma 17,

$$\begin{aligned}
\|\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j}\| &\le \frac{1}{\lambda\nu} \|\boldsymbol{X}_{\mathcal{M}} \dot{\boldsymbol{\ell}}_{\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus j})\| \\
&\le \frac{\sqrt{m}}{\lambda\nu} \|\boldsymbol{X}\| \max_{i\in\mathcal{M}} |\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus j})| \le \frac{\sqrt{m}}{\lambda\nu} (\sqrt{\gamma_0}+3)\sqrt{C_X} \mathrm{polylog}_4(n). \qquad (19)
\end{aligned}$$

Define

$$E_4 := \{\max_{j\in[n]} \max_{\substack{|\mathcal{M}|\le m \\ j\notin\mathcal{M}}} |\boldsymbol{x}_i^\top (\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j})| \le \mathrm{polylog}_5(n)\}$$

where $\mathrm{polylog}_5(n) := \frac{2(\sqrt{\gamma_0}+3)C_X}{\lambda\nu} \sqrt{m(2m+c)p^{-1}\log(n)} \mathrm{polylog}_4(n)$. Note that it is indeed $O(\mathrm{polylog}(n))$ when $m = o(\sqrt{p})$. Then

$$\mathbb{P}(\cup_{i=1}^4 E_i^c) \le \mathbb{P}(\cup_{i=1}^3 E_i^c) + \mathbb{P}(E_4^c \cap (\cap_{i=1}^3 E_i)),$$

and now we work on the second term since the first is already known. Let $N := \sum_{s=0}^{m-1} \binom{n-1}{s-1} \leq 2\binom{n-1}{m-1} = \frac{2m}{n}\binom{n}{m}$, then

$$\log(N) \leq \log(2) + \log(m) - \log(n) + m\log(em/n) \leq 2m\log(n)$$

for $n \geq 8$, so that we have

$$\sqrt{(2m+c)\log(n)} \geq \sqrt{\log(N) + c\log(n)}.$$

Therefore we have

$$\begin{aligned}
&\mathbb{P}(E_4^c \cap (\cap_{i=1}^3 E_i)) \\
\overset{(a)}{\leq} &\mathbb{P}(\max_{j\in[n]} \max_{\substack{|\mathcal{M}|\leq m \\ j\notin\mathcal{M}}} |\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j})| > 2\sqrt{C_X/p}\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j}\|\sqrt{\log(N) + c\log(n)},\ \cap_{i=1}^3 E_i) \\
\leq &\sum_{j\in[n]} \mathbb{E}\mathbb{P}(\max_{\substack{|\mathcal{M}|\leq m \\ j\notin\mathcal{M}}} |\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j})| > 2\sqrt{C_X/p}\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},j} - \hat{\boldsymbol{\beta}}_{\setminus j}\|\sqrt{\log(N) + c\log(n)}\ |\mathcal{D}_{\setminus j}) \\
\overset{(b)}{\leq} &2n^{1-c},
\end{aligned}$$

where in (a) we used (19) under $\cap_{i=1}^3 E_i$, and in (b) we used Lemma 15 again.

We therefore have

$$\mathbb{P}(\cup_{i=1}^4 E_i^c) \leq nq_n + 4n^{1-c} + ne^{-p/2} + e^{-p}.$$

Under event $\cap_{i=1}^4 E_i$, $\forall i \in [n], \forall |\mathcal{M}| \leq m, i \notin \mathcal{M}$:

$$\begin{aligned}
|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i}| &\leq |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus i}| + |\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i} - \hat{\boldsymbol{\beta}}_{\setminus i})| \\
&\leq \mathrm{polylog}_1(n) + \mathrm{polylog}_5(n) \\
&:= \mathrm{polylog}_6(n) \\
|\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i})| &\leq 1 + C_y^s(n) + \mathrm{polylog}_6^s(n) := \mathrm{polylog}_7(n) \\
\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i}\| &\leq \frac{1}{\lambda\nu}\|\boldsymbol{x}_i\| \cdot |\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i})| \leq \frac{2\sqrt{C_X}}{\lambda\nu}\mathrm{polylog}_7(n) \\
|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}| &\leq |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i}| + |\boldsymbol{x}_i^\top(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M},i})| \\
&\leq \mathrm{polylog}_6(n) + \frac{4C_X}{\lambda\nu}\mathrm{polylog}_7(n) \\
&:= \mathrm{polylog}_8(n) \\
|\dot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| &\leq 1 + C_y^s(n) + \mathrm{polylog}_8^s(n) := \mathrm{polylog}_9(n).
\end{aligned}$$

Step 5

Now we are ready to bound events F_2 and F_3 as stated in this lemma. For the convenience of the readers we re-state the definitions of these two events:

$$\begin{aligned}
F_2 &= \{\max_{i\in[n]} |\dot{\ell}_i(\hat{\boldsymbol{\beta}})| > C_\ell(n)\}, \\
F_3 &= \{\exists t \in [0,1], \mathcal{M} \subset \mathcal{D} \text{ with } |\mathcal{M}| \leq m, \|\ddot{\boldsymbol{\ell}}(\boldsymbol{\xi}(t)) - \ddot{\boldsymbol{\ell}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\| > C_{\ell\ell}(n)\|\boldsymbol{\xi}(t) - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|\}, \\
&\quad \text{where } \boldsymbol{\xi}(t) = t\hat{\boldsymbol{\beta}} + (1-t)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}},\ \ddot{\boldsymbol{\ell}}(\boldsymbol{\beta}) := [\ddot{\ell}_i(\boldsymbol{\beta})]_{i\in[n]}.
\end{aligned}$$

Event F_2 was actually bounded in (18) with $C_\ell(n) = \text{polylog}_3(n)$. Note that we allow $j = 0$ in (18), which indicates

$$\mathbb{P}(F_2) \le \mathbb{P}(\cup_{i=1}^3 E_i^c) \le nq_n + 2n^{1-c} + ne^{-p/2} + e^{-p}.$$

This proves part (a) of the lemma. For part (b), if $\boldsymbol{\beta} \in \mathcal{B}_{1,\mathcal{M}}$, then it is a convex combination of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}$, so

$$|\boldsymbol{x}_i^\top \boldsymbol{\beta}| \le \max\{|\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}|, |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}|\} \le \max\{\text{polylog}_3(n), \text{polylog}_8(n)\} := \text{polylog}_9(n).$$

Therefore

$$|\dddot{\ell}_i(\boldsymbol{\beta})| \le 1 + C_y^s(n) + \text{polylog}_9^s(n) := \text{polylog}_{10}(n). \tag{20}$$

In addition, notice that $\forall \boldsymbol{\beta} \in \mathcal{B}_{1,\mathcal{M}}, \forall s \in [0,1], \boldsymbol{\eta}(s) := s\boldsymbol{\beta} + (1-s)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} \in \mathcal{B}_{1,\mathcal{M}}$ as well. Therefore under $\cup_{i=1}^4$, $\forall |\mathcal{M}| \le m, \forall \boldsymbol{\beta} \in \mathcal{B}_{1,\mathcal{M}}$,

$$\|\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2^2 = \sum_{i\notin\mathcal{M}} [\ddot{\ell}_i(\boldsymbol{\beta}) - \ddot{\ell}_i(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})]^2 = \sum_{i\notin\mathcal{M}} [\dddot{\ell}_i \boldsymbol{x}_i^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})]^2,$$

where $\dddot{\ell}_i := \int_0^1 \dddot{\ell}_i(s\boldsymbol{\beta} + (1-s)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})ds$, so

$$|\dddot{\ell}_i| \le \int_0^1 |\dddot{\ell}_i(s\boldsymbol{\beta} + (1-s)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})|ds \le \text{polylog}_{10}(n),$$

so

$$\begin{aligned}
\|\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}) - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\|_2^2 &\le \sum_{i\notin\mathcal{M}} [\dddot{\ell}_i \boldsymbol{x}_i^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})]^2 \\
&\le \text{polylog}_{10}^2(n)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})^\top \boldsymbol{X}_{\setminus\mathcal{M}}^\top \boldsymbol{X}_{\setminus\mathcal{M}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \\
&\le \text{polylog}_{10}^2(n)(\sqrt{\gamma_0} + 3)^2 C_X \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^2
\end{aligned}$$

Also, notice that $\|\dddot{\boldsymbol{r}}(\boldsymbol{\beta}) - \dddot{\boldsymbol{r}}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\| \le C_{rr}(n)\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|$ is trivially satisfied by Assumption B2.

Simialrly, for F_4, under $\cup_{i=1}^4 E_i^c$, $\forall |\mathcal{M}| \le m, \forall i \notin \mathcal{M}, \forall \boldsymbol{\beta} \in \mathcal{B}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, 1)$,

$$\begin{aligned}
|\boldsymbol{x}_i^\top \boldsymbol{\beta}| &\le |\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}| + |\boldsymbol{x}_i^\top (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| \le \text{polylog}_8(n) + 2\sqrt{C_X} \\
|\dddot{\ell}_i(\boldsymbol{\beta})| &\le 1 + C_y^s(n) + |\text{polylog}_8(n) + 2\sqrt{C_X}|^s := \text{polylog}_{11}(n)
\end{aligned}$$

It then follows that

$$\|\ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_1) - \ddot{\boldsymbol{\ell}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_2)\|_2^2 \le \sum_{i\notin\mathcal{M}} \dddot{\ell}_i [\boldsymbol{x}_i^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)]^2 \le \text{polylog}_{11}^2(n)(\sqrt{\gamma_0} + 3)^2 C_X \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|^2.$$

Finally, if we define $C_\ell(n) := \text{polylog}_3(n)$, $C_{\ell\ell}(n) = \max\{\text{polylog}_{10}(n), \text{polylog}_{11}(n)\}$, we have

$$\mathbb{P}(F_1 \cup F_2 \cup F_3 \cup F_4) \le \mathbb{P}(\cup_{i=1}^4 E_i^c) \le nq_n + 4n^{1-c} + ne^{-p/2} + e^{-p}.$$

To make the $\text{polylog}_k(n)$ terms explicit, we repeat them here:

$$\begin{aligned}
\text{polylog}_1(n) &= 2\sqrt{(\lambda\nu)^{-1}\gamma_0 C_X(1+C_y^s(n))(1+c)\log(n)} \\
\text{polylog}_2(n) &= 1 + C_y^s(n) + \text{polylog}_1^s(n) \\
\text{polylog}_3(n) &= \text{polylog}_1(n) + \frac{C_X}{\lambda\nu}\text{polylog}_2(n) \\
\text{polylog}_4(n) &= 1 + C_y^s(n) + \text{polylog}_3^s(n) \\
\text{polylog}_5(n) &= \frac{2(\sqrt{\gamma_0}+3)C_X}{\lambda\nu}\sqrt{m(2m+c)p^{-1}\log(n)}\text{polylog}_4(n) \\
\text{polylog}_6(n) &= \text{polylog}_1(n) + \text{polylog}_5(n) \\
\text{polylog}_7(n) &= 1 + C_y^s(n) + \text{polylog}_6^s(n) \\
\text{polylog}_8(n) &= \text{polylog}_6(n) + \frac{4C_X}{\lambda\nu}\text{polylog}_7(n) \\
\text{polylog}_9(n) &= 1 + C_y^s(n) + \text{polylog}_8^s(n) \\
\text{polylog}_{10}(n) &= 1 + C_y^s(n) + \text{polylog}_9^s(n) \\
\text{polylog}_{11}(n) &= 1 + C_y^s(n) + |\text{polylog}_8(n) + 2\sqrt{C_X}|^s. \qquad (21)
\end{aligned}$$

■

B.2.5 Proof of Lemma 29

Proof Recall that the goal is to bound

$$\mathbb{P}(F_5) = \mathbb{P}\left(\max_{|\mathcal{M}|\le m} \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty} > \frac{(\sqrt{\gamma_0}+3)C_X \vee 1}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}}\right).$$

Notice that

$$\begin{aligned}
&\max_{|\mathcal{M}|\le m} \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty} \\
=&\max\left\{\max_{|\mathcal{M}|\le m} \|\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty},\ \max_{|\mathcal{M}|\le m} \|\boldsymbol{X}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty}\right\},
\end{aligned}$$

and we will bound the two terms separately.

Let us fix a subset $\mathcal{M}_0 \subset [n]$ of size $|\mathcal{M}_0| = s \le m$. Then $\boldsymbol{X}_{\mathcal{M}_0}$ is independent of $\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})$. That is,

$$\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top = \boldsymbol{\Sigma}_*^{1/2}\boldsymbol{Z}_{\mathcal{M}_0}^\top$$

where $\boldsymbol{\Sigma}_* = \boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{\Sigma}\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})$ satisfies

$$\sigma_{\max}(\boldsymbol{\Sigma}_*) \le \|\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\|^2 \times \|\boldsymbol{\Sigma}\| \le \frac{C}{p(\lambda\nu)^2}$$

almost surely, due to the ν-strong convexity of the penalty function. Thus by Lemma 18, since $|\mathcal{M}_0| = s$, we have for any $c \geq 0$ that

$$\mathbb{P}\left(\|\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top\|_{2,\infty} \geq \frac{1}{\lambda\nu}\sqrt{\frac{s(1+4(c+1)\log(n))}{p}} \right) \leq \exp(-(c+1)s\log(n)).$$

Now taking a union bound over all possible choices of $\mathcal{M}_0$ such that $|\mathcal{M}_0| \leq s$, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{|\mathcal{M}|\leq m} \|\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty} \geq \frac{1}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}} \right) \\
\leq &\sum_{s=1}^{m} \binom{n}{s} \exp(-(c+1)s(\log(n))) \\
\leq &\sum_{s=1}^{m} \exp(-(c+1)s\log(n)) \leq 2n^{-c}.
\end{aligned}$$

The proof technique for the second assertion is almost identical to the first part, with the following differences. We fix $\mathcal{M}_0$ with $|\mathcal{M}_0| = s \leq m$, and just as in part i), we obtain:

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{X}_{\setminus\mathcal{M}_0}\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top\|_{2,\infty} \geq \frac{C\|\boldsymbol{X}_{\mathcal{M}_0}\|}{\lambda\nu}\sqrt{\frac{s(1+4(c+1)\log(n))}{p}} \big| \boldsymbol{X}_{\mathcal{M}_0} \right) \\
\leq &\exp(-(c+1)s\log(n)).
\end{aligned}$$

Now by the independence of $\boldsymbol{X}_{\mathcal{M}_0}$ and $\boldsymbol{X}_{\setminus\mathcal{M}_0}$ for any fixed $\mathcal{M}_0$, we can remove the conditioning on $\boldsymbol{X}_{\setminus\mathcal{M}_0}$ to write:

$$\begin{aligned}
&\mathbb{P}\left(\|\boldsymbol{X}_{\setminus\mathcal{M}_0}\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top\|_{2,\infty} \geq \frac{\|\boldsymbol{X}_{\mathcal{M}_0}\|}{\lambda\nu}\sqrt{\frac{s(1+4(c+1)\log(n))}{p}} \right) \\
= &\mathbb{E}(\mathbb{P}(\ldots|\boldsymbol{X}_{\mathcal{M}_0})) \leq \exp(-(c+1)s\log(n)).
\end{aligned}$$

Then taking the union bound over all possible $\mathcal{M}_0$ as before, we have:

$$\mathbb{P}\left(\|\boldsymbol{X}_{\setminus\mathcal{M}_0}\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top\|_{2,\infty} \geq \frac{\|\boldsymbol{X}\|}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}} \right) \leq 2n^{-c}.$$

Now by Lemma 20, the final bound becomes

$$\mathbb{P}\left(\|\boldsymbol{X}_{\setminus\mathcal{M}_0}\boldsymbol{G}_{\setminus\mathcal{M}_0}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}_0})\boldsymbol{X}_{\mathcal{M}_0}^\top\|_{2,\infty} \geq \frac{(\sqrt{\gamma_0}+3)C_X}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}} \right) \leq 2n^{-c} + \mathrm{e}^{-p}.$$

Finally, by taking the maximum of the bounds obtained in the two parts above and a union bound, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{|\mathcal{M}|\leq m} \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\boldsymbol{G}_{\setminus\mathcal{M}}^{-1}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})\boldsymbol{X}_{\mathcal{M}}^\top\|_{2,\infty} > \frac{(\sqrt{\gamma_0}+3)C_X \vee 1}{\lambda\nu}\sqrt{\frac{m(1+4(c+1)\log(n))}{p}} \right) \\
\leq &4n^{-c} + e^{-p}.
\end{aligned}$$

■

B.2.6 ℓ_2 error of multiple Newton steps

To study multiple Newton steps, we first prove quadratic convergence for the Newton method in a general setting:

Lemma 30 *Suppose $\boldsymbol{f}(\boldsymbol{\beta}) : \mathbb{R}^p \to \mathbb{R}^p$ has Jacobian $\boldsymbol{G}(\boldsymbol{\beta})$ with $\lambda_{\min}(\boldsymbol{G}(\boldsymbol{\beta})) \geq \nu$ for all $\boldsymbol{\beta}$, and suppose $\boldsymbol{f}(\boldsymbol{\beta}) = \mathbf{0}$ has a unique solution $\boldsymbol{\beta}^*$. Suppose $\{\boldsymbol{\beta}^{(t)}\}_{t\geq 1}$ is the path of Newton method in searching for $\boldsymbol{\beta}^*$, i.e. $\forall t \geq 2$,*

$$\boldsymbol{\beta}^{(t)} := \boldsymbol{\beta}^{(t-1)} - \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{f}(\boldsymbol{\beta}^{(t-1)}).$$

Let $r_{t,n} := \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\|$ be the l_2 error of the t^{th} step. If $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in B(\boldsymbol{x}^*, r_1)$, $\|\boldsymbol{G}(\boldsymbol{x}_1) - \boldsymbol{G}(\boldsymbol{x}_2)\| \leq C\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, then

$$r_{t,n} \leq \frac{C}{2\nu} r_{t-1,n}^2.$$

Consequently,

$$r_{t,n} \leq \left(\frac{C}{2\nu}\right)^{2^{t-2}} r_1^{2^{t-1}}.$$

Proof By the definition of Newton steps,

$$\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^* = \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^* - \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{f}(\boldsymbol{\beta}^{(t-1)}).$$

Notice that we can add $\boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{f}(\boldsymbol{\beta}^*)$ to the right hand side because $\boldsymbol{f}(\boldsymbol{\beta}^*) = 0$, so we have

$$\begin{aligned}
\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^* &= \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^* - \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{f}(\boldsymbol{\beta}^{(t-1)}) + \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{f}(\boldsymbol{\beta}^*) \\
&= \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^* - \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})[\boldsymbol{f}(\boldsymbol{\beta}^{(t-1)}) - \boldsymbol{f}(\boldsymbol{\beta}^*)] \\
&= \boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^* - \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\bar{\boldsymbol{G}}(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*) \\
&= \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})[\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)}) - \bar{\boldsymbol{G}}](\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*),
\end{aligned}$$

where in the penultimate step we used Taylor expansion with

$$\bar{\boldsymbol{G}} := \int_0^1 \boldsymbol{G}(a\boldsymbol{\beta}^{(t-1)} + (1-a)\boldsymbol{\beta}^*)da,$$

and the last step uses the trick that $\boldsymbol{\beta}^{(t-1)} = \boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)})\boldsymbol{\beta}^{(t-1)}$. Notice that

$$\begin{aligned}
\|\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)}) - \bar{\boldsymbol{G}}\| &= \|\int_0^1 [\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)}) - \boldsymbol{G}(a\boldsymbol{\beta}^{(t-1)} + (1-a)\boldsymbol{\beta}^*)]da\| \\
&\leq \int_0^1 \|\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)}) - \boldsymbol{G}(a\boldsymbol{\beta}^{(t-1)} + (1-a)\boldsymbol{\beta}^*)\|da \\
&\leq \int_0^1 C\|(1-a)(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*)\|da \\
&= C\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\| \int_0^1 (1-a)da \\
&= \frac{C}{2}\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|
\end{aligned}$$

Therefore we have

$$r_{t,n} = \|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*\| \le \|\boldsymbol{G}^{-1}(\boldsymbol{\beta}^{(t-1)})\| \cdot \|\boldsymbol{G}(\boldsymbol{\beta}^{(t-1)}) - \bar{\boldsymbol{G}}\| \cdot \|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\| \\ \le \frac{1}{\nu}\frac{C}{2}\|\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\beta}^*\|^2 = \frac{C}{2\nu} r_{t-1,n}^2.$$

We then immediately have

$$r_{t,n} = \left(\frac{C}{2\nu}\right)^{2^{t-2}} r_{1,n}^{2^{t-1}}.$$

where $r_{1,n} = C_1(n)\sqrt{\frac{m^3}{n}}$ is the bound we obtained in Lemma 26. ■

It is then a direct application of Lemma 30 to prove Theorem 12:

Proof [Proof of Theorem 12] It is a direct application of Lemma 30 together with several previous results.

First, under event F^c, by Lemma 17, $r_1 := \|\tilde{\boldsymbol{\beta}}^{(1)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \le C_1(n) m^{\frac{3}{2}} n^{-\frac{1}{2}} = o(1)$ since $m = o(n^{1/3})$. Then for the Lipschiz condition of $\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta})$, notice that under event F_4^c, for large enough n, $\forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{B}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, r_1) \subset \mathcal{B}(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}, 1)$,

$$\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_1) - \boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_2) = \boldsymbol{X}_{\setminus\mathcal{M}}^\top [\ddot{\boldsymbol{L}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_1) - \ddot{\boldsymbol{L}}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_2)] \boldsymbol{X}_{\setminus\mathcal{M}} + \lambda r[\ddot{\boldsymbol{R}}(\boldsymbol{\beta}_1) - \ddot{\boldsymbol{R}}(\boldsymbol{\beta}_2)] \\ = \bar{\boldsymbol{X}}_{\setminus\mathcal{M}}^\top \boldsymbol{\Gamma} \bar{\boldsymbol{X}}_{\setminus\mathcal{M}}$$

where the last line is similar to (14). Under F_4^c, we have $\|\boldsymbol{\Gamma}\|_2 \le \|\boldsymbol{\Gamma}\|_{Fr} \le (C_{\ell\ell}(n) + \lambda C_{rr}(n))\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$, so

$$\|\boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_1) - \boldsymbol{G}_{\setminus\mathcal{M}}(\boldsymbol{\beta}_2)\| \le \|\bar{\boldsymbol{X}}_{\setminus\mathcal{M}}\|^2 (C_{\ell\ell}(n) + \lambda C_{rr}(n))\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \\ \le (1 + (\sqrt{\gamma_0} + 3)\sqrt{C_X})^2 (C_{\ell\ell}(n) + \lambda C_{rr}(n))\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \\ := C_2(n)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|.$$

We then apply Lemma 30 with the Lipschitz constant $C = C_2(n) = O(\text{polylog}(n))$ and the strong convexity constant to be $\lambda\nu$. ■

B.3 Proof of Theorem 7 and Theorem 8

Proof [Proof of Theorem 7] Recall that we defined

$$\mathcal{X}^{(t)}_{r_{t,n}} := \{\mathcal{D} : \max_{|\mathcal{M}| \le m} \|\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|_2 \le r_{t,n}\}.$$

By Lemma 9, $\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}} = \tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}$ is (ϕ_n, ϵ)-PAU if $\mathbb{P}(\mathcal{X}^{(t)}_{r_{t,n}}) \ge 1 - \phi_n$.

By Lemma 26, $\left(\mathcal{X}^{(t)}_{r_{t,n}}\right)^c \subset F = \cup_{i=1}^5 F_i$ defined in (13).

By Lemma 27,

$$\mathbb{P}(F) \le nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p} := \phi_n,$$

as long as we set the constants C_1, $C_\ell(n)$, $C_{\ell\ell}(n)$, $C_{rr}(n)$ in the definion of F (13) as in Lemma 27, so that

$$r_{t,n} = [C_1(n)]^{2^{t-1}} \left(\frac{C_2(n)m^3}{2\lambda\nu n}\right)^{2^{t-2}},$$
$$C_1(n) = \frac{2\sqrt{3}}{3\lambda^2\nu^2}[C_{\ell\ell}(n) + \lambda C_{rr}(n)]C_1(C_1+1)C_\ell^2(n)C_{xx}(n),$$
$$C_2(n) = (1 + (\sqrt{\gamma_0}+3)\sqrt{C_X})^2(C_{\ell\ell}(n) + \lambda C_{rr}(n)).$$

Therefore we conclude that $\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}} = \tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}$ achieves (ϕ_n, ϵ)-PAU with

$$\phi_n = nq_n + 8n^{1-c} + ne^{-p/2} + 2e^{-p},$$

if $\boldsymbol{b}$ has density $p(\boldsymbol{b}) \propto e^{-\frac{\epsilon}{r_{t,n}}\|\boldsymbol{b}\|}$. ■

Proof [Proof of Theorem 8] First notice that by Taylor expansion,

$$|\ell_0(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}) - \ell_0(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| \le |\bar{\dot{\ell}}_0\|\boldsymbol{x}_0^\top(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b})|.$$

Let F be the failure event defined in (13). Recall that by Equation (17) and (19), under F^c,

$$\|\hat{\boldsymbol{\beta}}\| \le \sqrt{(\lambda\nu)^{-1}(1 + C_y^s(n))n}$$
$$\max_{|\mathcal{M}|\le m} \|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \le \frac{\sqrt{m}}{\lambda\nu}(\sqrt{\gamma_0}+3)\sqrt{C_X}\mathrm{polylog}_4(n).$$

Define events

$$E_5 := \left\{|\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}| \le \sqrt{(\lambda\nu)^{-1}C_X\gamma_0(1 + C_y^s(n)2c\log(n))}\right\}$$
$$E_6 := \left\{\max_{|\mathcal{M}|\le m} |\boldsymbol{x}_0^\top(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| \le \frac{2C_X}{\lambda\nu}(\sqrt{\gamma_0}+3)\sqrt{\frac{m(2m+c)}{p}\log(n)}\mathrm{polylog}_4(n)\right\}.$$

Then for any $\mathcal{D} \in F^c$, we have

$$\mathbb{P}(E_5^c|\mathcal{D}) \le \mathbb{P}\left(|\boldsymbol{x}_0^\top\hat{\boldsymbol{\beta}}| \le \sqrt{\frac{C_X}{p}}\|\hat{\boldsymbol{\beta}}\|\sqrt{2c\log(n)}|\mathcal{D}\right) \le 2n^{-c}$$
$$\mathbb{P}(E_6^c|\mathcal{D}) \le \mathbb{P}\left(\exists|\mathcal{M}| \le m, |\boldsymbol{x}_0^\top(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| > \sqrt{\frac{C_X}{p}}\|\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \cdot 2\sqrt{\log(N) + c\log(n)}\middle|\mathcal{D}\right)$$
$$\le 2n^{-c},$$

where $N = \sum_{s=0}^m \binom{n}{s} \le 2\binom{n}{m}$, so

$$\log(N) \le \log(2) + m\log(en/m) \le 2m\log(n).$$

Under $E_5 \cap E_6$,

$$|\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}| \leq |\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}| + |\boldsymbol{x}_0^\top (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| \leq \text{polylog}_{12}(n)$$

if $m = o(\sqrt{n})$.

Define

$$E_7 := \left\{ \forall |\mathcal{M}| \leq m, |\boldsymbol{x}_0^\top (\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b})| \leq 2\sqrt{C_X} \left(\frac{2\sqrt{p}}{\epsilon} + \frac{1}{\sqrt{p}} \right) r_{t,n} \cdot \sqrt{(2m+c)\log(n)} \right\},$$

$$E_8 := \{ \|\boldsymbol{b}\| \leq \frac{2p}{\epsilon} r_{t,n}, |y_0| \leq C_y(n) \}.$$

Under F^c, $\|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\| \leq r_{t,n}$. Then for any $\mathcal{D} \in F^c$

$$\begin{aligned}
\mathbb{P}(E_7^c \cap E_8 | \mathcal{D}) &\leq \mathbb{P}(\exists |\mathcal{M}| \leq m : |\boldsymbol{x}_0^\top (\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b})| > \\
&\qquad \sqrt{\frac{C_X}{p}} \|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}\| \cdot 2\sqrt{\log(N) + c\log(n)} | \mathcal{D}) \\
&\leq 2n^{-c}.
\end{aligned}$$

Under $(\cap_{i=5}^8 E_i)$, $\forall a \in [0,1]$,

$$\begin{aligned}
&|\boldsymbol{x}_0^\top [a(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} + \boldsymbol{b}) + (1-a)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}]| \\
&\leq |\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}| + a|\boldsymbol{x}_0^\top (\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b})| \\
&\leq \text{polylog}_{12}(n) + 2\sqrt{C_X} \left(\frac{2\sqrt{p}}{\epsilon} + \frac{1}{\sqrt{p}} \right) r_{t,n} \cdot \sqrt{(2m+c)\log(n)} \\
&\leq \text{polylog}_{13}(n),
\end{aligned}$$

provided $r_{t,n} = o\left(\frac{\epsilon}{\sqrt{mp\text{polylog}(n)}} \right)$ so that the second term is $O(\text{polylog}(n))$. Thus, for any $\mathcal{D} \in F^c$,

$$\begin{aligned}
&\mathbb{E}\left[\int |\boldsymbol{x}_0^\top [a(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} + \boldsymbol{b}) + (1-a)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}]|^{2s} da | \mathcal{D} \right] \\
&\leq (\text{polylog}_{13}(n))^{2s} \\
&\quad + C_s (\mathbb{E}\|\boldsymbol{x}_0\|^{4s})^{1/2} (\|\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^{4s} + \|\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}\|^{4s} + \mathbb{E}|\boldsymbol{b}|^{4s})^{1/2} \mathbb{P}((\cap_{i=5}^8 E_i)^c | \mathcal{D}) \\
&\leq (\text{polylog}_{13}(n))^{2s} + C_s C_X \left(\frac{2p}{\epsilon} + 1 \right)^{2s} r_{t,n}^{2s} \cdot (\sqrt{(2m+2s)\log(n)})^{2s} \cdot n^{-2s} \\
&\leq \text{polylog}_{14}(n) \qquad (22)
\end{aligned}$$

using events $\cap_{i=5}^8 E_i$ for some $c \geq 2s$. Similarly we obtain

$$\mathbb{E}\left[|\boldsymbol{x}_0^\top [(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}}^{(t)} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) + \boldsymbol{b}]|^2 da | \mathcal{D} \in F^c \right] \leq 2C_X \left(\frac{2\sqrt{p}}{\epsilon} + \frac{1}{\sqrt{p}} \right)^2 r_{t,n}^2 \cdot (2m+2)(\log(n)) \quad (23)$$

using events $\cap_{i=5}^{8} E_i$ for some $c \geq 2$. Finally we have that,

$$\begin{aligned}\mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) &= \mathbb{E}\left(|\ell_0(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}) - \ell_0(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})||\mathcal{D}\right) \\ &\leq \left(\mathbb{E}(|\bar{\dot{\ell}}_0|^2|\mathcal{D})\right)^{1/2}\left(\mathbb{E}(|\boldsymbol{x}_0^\top \boldsymbol{b} + \boldsymbol{x}_0^\top(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})|^2|\mathcal{D})|\right)^{1/2}\end{aligned}$$

Now using the previous inequalities, we will bound each of the quantities on the right hand side of the latest display. First notice that by (22), for any $\mathcal{D} \in F^c$

$$\begin{aligned}&\mathbb{E}(|\bar{\dot{\ell}}_0|^2|\mathcal{D}) \\ &= \mathbb{E}\left(|\int_0^1 \dot{\ell}_0((\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}) + (1-a)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})da|^2|\mathcal{D}\right) \\ &\leq 3C^2\left(1 + \mathbb{E}|y_0|^{2s} + \mathbb{E}\left[\int |\boldsymbol{x}_0^\top[a(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}) + (1-a)\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}]|^{2s}da|\mathcal{D}\right]\right) \\ &\leq 3C^2(1 + C_{y,s} + \mathrm{polylog}_{14}(n))\end{aligned}$$

where we use Assumption B3 for the last two inequalities. Similarly using (23) we have for any $\mathcal{D} \in F^c$ that

$$\begin{aligned}&\mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) \\ &= \mathbb{E}\left(|\ell_0(\tilde{\boldsymbol{\beta}}^{(t)}_{\setminus\mathcal{M}} + \boldsymbol{b}) - \ell_0(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})||\mathcal{D}\right) \\ &\leq 3C(1 + C_{y,s} + \mathrm{polylog}_{14}(n))^{1/2}\sqrt{C_X}\left(\frac{2\sqrt{p}}{\epsilon} + \frac{1}{\sqrt{p}}\right) r_{t,n} \cdot \sqrt{(2m+2)(\log(n))}.\end{aligned}$$

This proves the first part of the theorem.

To prove the second part, notice that for any $\mathcal{D} \in F^c$, $\mathrm{GED}(\tilde{\boldsymbol{\beta}}^{R,t}_{\setminus\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}}) = O(1)$ if $r_{t,n} = o\left(\frac{\epsilon}{\sqrt{mp\mathrm{polylog}(n)}}\right)$. To find the smallest t such that this holds true, define $\alpha = \log(m+1)/\log(n)$, then we have

$$r_{t,n} = [C_1(n)]^{2^{t-1}}\left(\frac{C_2(n)(m+1)^3}{2\lambda\nu n}\right)^{2^{t-2}} = \mathrm{polylog}(n)n^{(3\alpha-1)2^{t-2}}.$$

We need

$$r_{t,n} = o\left(\frac{\epsilon}{\sqrt{mp\mathrm{polylog}(n)}}\right) = o(n^{-\frac{1}{2}(\alpha+1)}),$$

which is satisfied when

$$t > T = 1 + \log_2\left(\frac{\alpha+1}{1-3\alpha}\right).$$

■